Bayesian learning: Challenges, limitations and pragmatics

Heide, R. de

Bayesian Learning: Challenges, Limitations and Pragmatics

Proefschrift
ter verkrijging van
de graad van Doctor aan de Universiteit Leiden
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 26 januari 2021
klokke 16:15 uur

door

Rianne de Heide
geboren te Rotterdam in 1989

Promotores:

Prof. dr. P.D. Grünwald (Universiteit Leiden en Centrum Wiskunde & Informatica, Amsterdam)

Prof. dr. J.J. Meulman

Co-promotor:

Dr. W.M. Koolen (Centrum Wiskunde & Informatica, Amsterdam)

Samenstelling van de promotiecommissie:

Prof. dr. E.R. Eliel

Prof. dr. R.M. van Luijk

Prof. dr. A. Carpentier (Otto von Guericke Universität Magdeburg)

Dr. A. Ramdas (Carnegie Mellon University)

Dr. D.M. Roy (University of Toronto)

The author's PhD position at the Mathematical Institute was supported by the Leiden IBM-SPSS Fund. The research was performed at the Centrum Wiskunde & Informatica (CWI). Part of the work was done while the author was visiting Inria Lille, partly funded by Leids Universiteits Fonds / Drs. J.R.D. Kuikenga Fonds voor Mathematici travel grant number W19204-1-35.

Cover design by Chantal Bekker
Printing by Drukkerij Haveka

Universiteit Leiden
Mathematisch Instituut

CWI
Centrum Wiskunde & Informatica

Origin of the material

This dissertation is based on the following papers. The author of this dissertation contributed substantially to each of these papers.

Chapter 2 is based on the paper that is under review as

> Tom Sterkenburg and Rianne de Heide. On the truth-convergence of open-minded Bayesianism.

Chapter 3 is accepted for publication in *Psychonomic Bulletin & Review*, and is available as the technical report

> Rianne de Heide and Peter Grünwald. Why optional stopping can be a problem for Bayesians. *arXiv 1708.08278*. August 2017.

Chapter 4 is published as

> Allard Hendriksen, Rianne de Heide and Peter Grünwald. Optional Stopping with Bayes Factors: a categorization and extension of folklore results, with an application to invariant situations. Bayesian Analysis, advance publication, 18 August 2020. doi:10.1214/20-BA1234.

Chapter 5 is based on the technical report

> Peter Grünwald, Rianne de Heide and Wouter Koolen. Safe Testing. *arXiv 1906.07801*. June 2019.

Chapter 6 is published as

> Rianne de Heide, Alisa Kirichenko, Nishant Mehta and Peter Grünwald. Safe-Bayesian Generalized Linear Regression. AISTATS 2020, PMLR 108:2623-2633.

The software for this chapter is partly available as

> Rianne de Heide (2016). SafeBayes: Generalized and Safe-Bayesian Ridge and Lasso Regression. R package version 1.1. https://cran.r-project.org/src/contrib/Archive/SafeBayes/

Chapter 7 is published as

> Xuedong Shang, Rianne de Heide, Emilie Kaufman, Pierre Ménard and Michal Valko. Fixed-Confidence Guarantees for Bayesian Best-Arm Identification. AISTATS 2020, PMLR 108:1823-1832.

Contents

Chapter 1

Introduction

This dissertation is about Bayesian learning from data. How can humans and computers learn from data? This question is at the core of both statistics and — as its name already suggests — machine learning. Bayesian methods are widely used in these fields, yet they have certain limitations and problems of interpretation. In two chapters of this dissertation, we examine such a limitation, and overcome it by extending the standard Bayesian framework. In two other chapters, we discuss how different philosophical interpretations of Bayesianism affect mathematical definitions and theorems about Bayesian methods and their use in practise. While some researchers see the Bayesian framework as normative (all statistics should be based on Bayesian methods), in the two remaining chapters, we apply Bayesian methods in a pragmatic way: merely as *tool* for interesting learning problems (that could also have been addressed by non-Bayesian methods). In this introductory chapter, I first explain Bayesian learning by means of a coin tossing example. Thereafter, I review how different scientists view Bayesian learning, and in Section 1.3 the limitations and challenges of Bayesian inference that are addressed in this dissertation are discussed. In Sections 1.4 through 1.7, I give a brief introduction to the topics of this dissertation.

1.1 Bayesian learning

Learning A *learner*, which can be a human or a computer, interacts with the world she wants to learn about via *data*, also called *observations*, *examples* or *samples*. We can view the data as finite initial segments $Z^t := Z_1, \ldots, Z_t$ of an infinite *data stream*, denoted with Z^ω. The learner's task is *inductive inference*: inference that progresses from given examples to hitherto unknown examples and to general observational statements. The learner needs to start with background assumptions that restrict the space of possible outcomes. This is called *prior knowledge* or *inductive bias*. We assume that there is some collection of *hypotheses* that the learner can propose or investigate. We can view an *hypothesis* as a general statement about the world. In our context, the fields of machine learning and statistics, hypotheses are often expressed by a probability distribution over a sample space. We call those *statistical hypotheses*. A set of

statistical hypotheses is a (statistical) *model*. A model captures the background assumptions mathematically: It is a simplified description of the part of the world we consider relevant. In some chapters of this dissertation, we examine the behaviour of standard methods under *misspecification*, which means that the true world is not in the set of ways the world could be that would make the assumptions true. In other words: the model is wrong.

Example 1.1 (Coin tossing)**.** Suppose we toss a coin with unknown bias. If it lands heads, we denote a one, if it lands tails, we denote a zero. The learner sees a finite string z^t of zeros and ones. We can model the coin tosses by Bernoulli random variables with parameter $\theta \in [0,1]$. A possible hypothesis is: 'The coin is fair', and the corresponding statistical hypothesis is that the data, i.e. the outcomes $z^t = z_1, \ldots, z_t$, are independently distributed according to a Bernoulli distribution with parameter $\theta = 1/2$.

Learning objectives The task of the learner is inductive inference, which can have three distinct objectives. The first objective is *estimation*, for example: estimating a regression coefficient. Another objective is to predict or classify future data, e.g. predicting how well a patient will respond to a certain medicine, given patient characteristics such as white blood cell count, age, gender, etc. A third objective, which is the focus of several chapters of this dissertation, is *testing*. The learner is handed an hypothesis and some finite data sequence, and is requested to conjecture an assessment, often binary valued: {true, false} or {accept, reject}. There is also a dichotomy between *exploratory* and *confirmatory* research. In exploratory research the learner is given some data, and asked to produce an hypothesis about the origin of the data. We might for example be interested in understanding a possible genetic basis for a disease. Paraphrasing Tukey (1980): Exploratory research is about finding the question. In confirmatory research the validity of an existing hypothesis is tested.

Example 1.1 (continued)**.** In the coin tossing example, we can *estimate* the bias of the coin, or we can *predict* the next outcome, or we can *test* whether the coin is fair or not.

Bayesian inference With the model in place and the data to our disposal, we need one more ingredient for induction: a *method*, or *rule* for inference. In this dissertation, the focus is on (variations on) *Bayesian* inference. The essence of Bayesian inference is that it employs probability distributions both over statistical hypotheses as well as over data. Following Ghosh, Delampady and Samanta (2007), we denote with θ a quantity of interest. The learner starts with specifying a *prior distribution* $\pi(\theta)$, which quantifies her uncertainty about θ before seeing the data Z . Then she calculates the *posterior* $\pi(\theta \mid z)$, the conditional density of θ given $Z = z$, by Bayes theorem

$$\pi(\theta \mid z) = \frac{\pi(\theta) f(z \mid \theta)}{\int_\Theta \pi(\theta') f(z \mid \theta') \, \mathrm{d}\theta'}. \tag{1.1}$$

The numerator consists of the prior $\pi(\theta)$ and the likelihood $f(z \mid \theta)$, the denominator is the marginal density of Z, also called *Bayes marginal (likelihood)* or *model evidence*. The posterior distribution represents the learner's uncertainty regarding θ conditioned on the data. It is a trade-off between the prior and data distributions, determined by the strength of the prior information and the amount of data available.

A property that many find attractive of Bayesian methods, is that all inference goes via the posterior distribution. In the situation of parameter estimation the learner could for example report the posterior mean and variance

$$\mathbf{E}(\theta \mid z) = \int_{-\infty}^{\infty} \theta \pi(\theta \mid z)\, \mathrm{d}\theta \quad ; \quad \mathrm{Var}(\theta \mid z) = \int_{-\infty}^{\infty} (\theta - \mathbf{E}(\theta \mid z))^2\, \pi(\theta \mid z)\, \mathrm{d}\theta. \tag{1.2}$$

In case of hypothesis testing, she could compute the posterior odds or Bayes factor, see Section 1.5.

Computation For a long time Bayesian inference was mostly limited to *conjugate* families of distributions: specific choices of the model and prior distribution that give a closed-form expression for the posterior. The development of Markov Chain Monte Carlo (MCMC) methods in the 1990s (Gelfand and Smith, 1990) revolutionised Bayesian statistics. MCMC methods are algorithms that generate samples from a probability distribution, by constructing a reversible Markov chain that has the target distribution as its equilibrium distribution. In Chapter 6 we develop some MCMC algorithms.

Let us return to our coin tossing example.

Example 1.1 (continued)**.** Suppose a learner wants to learn the bias of the coin, i.e. the parameter θ of a Bernoulli distribution. She first needs to specify a prior distribution on the parameter space: the interval $[0, 1]$. At this point, it is unclear how she should choose the prior; we will get back on this issue in Section 1.2.1. Already back in 1814, Laplace suggested that, if one is ignorant about the bias of the coin, one should choose a uniform distribution over the parameter space (Laplace, 1814), although the idea to translate *ignorance* to *uniform* was later challenged (see Section 1.2.1). Let us follow Laplace for now: the learner chooses a uniform distribution, which corresponds to a $\mathrm{Beta}(1, 1)$ distribution. As the Beta distribution is conjugate to the Bernoulli family, quantities such as in (1.2) can be easily computed analytically. Specifically, the coin is tossed t times and she observes the sequence z^t consisting of n_1 ones and n_0 zeros. The likelihood is $f(z \mid \theta) = \theta^{n_1}(1-\theta)^{n_0}$. Due to the Beta-Bernoulli conjugacy, she can easily compute the posterior distribution (1.1), which has the form of a $\mathrm{Beta}(1 + n_1, 1 + n_0)$ distribution. To give an estimate of the parameter θ, she can take the posterior mean $\mathbf{E}(\theta \mid z) = (n_1 + 1)/(n_1 + n_0 + 2)$. Alternatively, she can report the posterior mode: $\arg\max_\theta \pi(\theta \mid z^t) = n_1/(n_1 + n_0)$.

With modern MCMC methods, Bayesian analyses are not restricted to conjugate families anymore, and models with many parameters can be handled, even non-parametric (roughly: *infinite-dimensional*) models. These problems can also be addressed with non-Bayesian, often called *classical* methods, see Section 1.2.3. There exist however philosophers and statisticians who believe that all learning problems should be addressed in a Bayesian way, I will loosely call them *Bayesians*.

In the example, we saw how Bayesian inference is done in practise. However, we already encountered a potential problem: How should the learner choose the prior? There are different views on this, and choice of prior is only one of many quarrels among Bayesians. To cite the famous mathematician I.J. Good: "There are 46656 varieties of Bayesians" (Good, 1971); in other words, there is no unique Bayesian theory of inference. *Bayesianism* extends far beyond the field of statistics: There is Bayesian epistemology, Bayesian confirmation theory (in philosophy

of science), Bayesian learning theory (in psychology), Bayesian decision theory, and more. Discussions about the foundations of Bayesianism are mostly held by philosophers, yet these certainly affect (statistical) practise: Adherents to different varieties of Bayesianism choose different priors, and present different mathematical definitions and theorems. The implications of the philosophical discussions about Bayesianism for statistical practise are the subject of Chapters 3 and 4.

In the next section I explain the common ground of most of the varieties of Bayesianism. This is followed by an exposition of the main differences and disputes between Bayesians, in particular, the *subjectivists* and the *objectivists*, yet I also introduce a third category that encompasses many Bayesian statisticians: the *pragmatists*.

Since this dissertation is about Bayesian methods, an obvious question is: Why do people use a Bayesian approach? For some (who perhaps may be called the *true Bayesians*) the main reasons are philosophical, for others the fact that all inference is based on the posterior distributions is attractive, and many find it intuitively appealing. Others have a more pragmatic view: There exists an interesting problem, and Bayesian inference is a good way to solve it. In Section 1.2.2 I discuss some of those arguments for the use of Bayesian methods, and also some against.

Section 1.2.3 briefly describes 'the other' main theory of statistics: *classical* or *frequentist* statistics. In Chapters 5 and 7, we use Bayesian methods, but we want them to have certain frequentist properties and guarantees.

1.2 Views on Bayesianism

As I mentioned above quoting I.J. Good, there is no unified Bayesian movement, or theory of inference, yet, there are some common foundations. Notable Bayesians and texts presenting some influential interpretations are: Ramsey (1926), Savage (1954), Jeffreys (1961), De Finetti e.g. (1974), Jeffrey (1992), Howson and Urbach (2006), and, from a more statistical perspective: Bernardo and Smith (1994), Gelman et al. (2003), and Ghosh, Delampady and Samanta (2007).

Central to Bayesian statistics, epistemology and confirmation theory — the interests of this dissertation — is the *epistemic* interpretation[1] of probability as *degrees of belief*. Most Bayesians further agree (Romeijn, 2005a; Easwaran, 2011) that these degrees of belief should obey rationality conditions in two respects. In the first place, these concern the degrees of belief at a certain point in time: Kolmogorov's 1933 axioms of probability theory. Secondly, these concern how degrees of belief should change over time: this should be done by conditionalisation. We have seen in the previous section and Example 1.1 how this is done: Formally, let S be some statement, then we start with a *prior* probability $P_{\text{old}}(S)$ — our prior belief in S. Upon acquiring new evidence[2] E, we transform our prior probability to generate a *posterior* probability by

[1] One can also interpret a (mathematical) probability as *physical* probability: a relative frequency or propensity, often termed *chance*. Some also called this *objective* probability, however, I find that an unfortunate wording, because of possible confusion with what follows next in the main text: subjective and objective probability, which can both apply to physical and epistemic probabilities. See also Hacking (2006), who discusses the concept of probability historically and philosophically.

[2] Assume for simplicity here that E comprises every statement we became certain of and had positive prior probability.

conditionalising on E, that is, $P_{\text{new}}(S) = P_{\text{old}}(S|E)$. This is called *Bayes' rule*.

But this is where the agreement among Bayesians ends. The first issue that is at the heart of many disputes among Bayesians — the interpretation of epistemic probability — is closely related to the issue of the origin of priors. I now describe the views on these two issues held by two central categories of Bayesians: the subjectivists and the objectivists. After that, I add a third category: the pragmatists.

1.2.1 The origin of priors

Subjectivism At one end of the spectrum of Bayesians, the subjectivists (Ramsey, De Finetti, Savage) take probability to be the expression of personal opinion. Probabilities can be related to betting contracts (see Section 1.2.2)), and the most extreme subjectivists impose no rationality constraints on prior probabilities other than probabilistic coherence, i.e. respecting Kolmogorov's probability axioms (De Finetti, 1937; Savage, 1954). For some subjectivists (e.g. Jeffrey (1965)), there can be some further constraints, but they exclude little, and in general, the prior probability assignments may originate from non-rational factors.

Objectivism At the other tail of the spectrum, the objectivists (Jeffreys, Jaynes) feel that prior probabilities should be rationally constrained, for example by physical probabilities or symmetry principles. Ideally such rationality constraints would uniquely determine a prior for every specific case, making prior probabilities *logical probabilities*. The objective program was already started by Sir Harold Jeffreys in 1939 (Jeffreys, 1939), and he advanced his *theory of invariants* in 1948 (Jeffreys, 1946; Jeffreys, 1948). His invariance principle leads to a rule to identify distributions that represent 'ignorance' about a quantity of interest, considering the statistical model. This distribution is now known as *Jeffreys' prior*[3]. Assuming regularity conditions (see Grünwald (2007), p.234ff.), it is proportional to the square root of the determinant of the Fisher information, and it is invariant under 1-1 differentiable transformations of the parameter space. Jeffreys' invariance principle is modified by Jaynes into his *maximum entropy* principle (Jaynes, 1957). However, no principles exist that uniquely determine rational priors in all cases (which is, besides, not claimed by any self-declared objective Bayesian either). This is by no means the only problem with objectivism, see Seidenfeld (1979). Still, some authors advocate its use in practice (Berger, 2006).

Example 1.1 (continued)**.** Jeffreys' prior for the coin tossing example is

$$\begin{aligned}\pi(\theta) &\propto \sqrt{I(\theta)} \\ &= \sqrt{\mathbf{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(z \mid \theta)\right)^2\right]} \\ &= \frac{1}{\sqrt{\theta(1-\theta)}},\end{aligned}$$

which corresponds to a $\text{Beta}(1/2, 1/2)$ distribution.

[3] Related are *reference priors* for higher dimensional models (Bernardo, 1979), Jaynes' maximum entropy priors (see the main text), and MDL-type priors (Grünwald, 2007).

Pragmatism Nowadays, many if not most statisticians using Bayesian methods do not adhere to a particular philosophy, but choose their priors for *pragmatic* reasons: for mathematical or computational convenience, because of their effects (e.g. shrinkage priors, see Chapter 6), to provide applied researchers with a *default* Bayesian method (see Chapter 3 and 4), or to construct methods that satisfy specific criteria (such as the GROW in Chapter 5). Often, these priors exhibit a mix of subjective and objective elements, but the reasons for using these priors and Bayesian methods in general are practical rather than philosophical. This is what I call *pragmatic Bayesianism*. Pragmatic Bayesians do not view probabilities as degrees of belief; they call them for example *weights*. This view is eloquently described by Gelman and Shalizi (2012).

Besides the interpretation of degrees of belief and the origin of priors, philosophers disagree about many other aspects of Bayesianism, such as whether probability should be treated as countably or finitely additive (see Seidenfeld and Schervish (1983), Kadane, Schervish and Seidenfeld (1999), Williamson (1999) and Elliot (2014)), whether conditionalisation can be generalised to situations in which the observations are themselves probabilistic statements (see Jeffrey (1965)), and more.

1.2.2 Arguments for Bayesianism and criticism

There are various arguments for (types of) Bayesianism. The most well-known are probably the *Dutch Book arguments*, introduced by Ramsey (1926) and De Finetti (1937). They relate probability, as degrees of belief, to a willingness to bet. If a bookmaker does not respect the axioms of probability theory, a clever gambler can make a Dutch book: He can propose a set of bets that wins him some amount of money no matter what the outcomes may be. There exist versions with finite and countable additivity, see e.g. Freedman (2003). Related arguments are exchangeability and De Finetti's (1937) representation theorem, see e.g. Bernardo (1996), Easwaran (2011) and Romeijn (2017).

In Bayesian decision theory, there are *complete class theorems*, originally due to Wald (1947) (see e.g. Robert (2007)), which provide a very pragmatic argument for Bayesianism. They basically state that for every method for learning from data, there exists a method that is at least as good, and that is *Bayesian* in the sense that it is based on updating beliefs using Bayes' theorem with a particular prior. A drawback of this argument is the limited applicability of these theorems, it holds for compact parameter spaces and convex loss functions, and besides that, there is still considerable room for manoeuvre in the choice of the prior. In particular, the choice of prior may depend on e.g. the sample size and the choice of loss function, which may be unnatural to many non-pragmatists.

Bayesian statistics can be justified in other 'non-Bayesian' ways too. Some find Bayesian analysis attractive because it does not rely on counterfactuals, whereas some non-Bayesian methods do: they rely on integration over the sample space, hence on data that could have but have not realised (Dawid and Vovk, 1999). Others like Bayesian methods because all inference is based on the posterior only, which leads to straightforward uncertainty quantification — for example, separate 'confidence intervals' are not needed. Other reasons are more practical. Bayesian inference often works very well in practise. For example in clinical trials, researchers often

have to deal with missing data because of the intention-to-treat policy. Here Bayesian ways of dealing with the missing data because of drop-outs often outperform other, classical methods (Asendorpf et al., 2014). Another example of a practical motivation is the success of *shrinkage priors*, which are chosen to produce a sparse estimate of a regression parameter vector; these are discussed in Chapter 6.

Criticism

How to specify the prior? This question both divides subjective and objective Bayesians, and lies at the root of the main criticisms from non-Bayesians. Several issues can be filed under *the problem of priors*. Subjectivists and objectivists debate whether there should be constraints on prior probabilities, other than the laws of probability theory. In the case of objective Bayes, there are no principles that uniquely determine objective priors in all cases. In particular, it is unclear how a prior should represent ignorance. Subjective Bayesianism is criticised for the idea that prior and posterior represent the learner's subjective belief, while scientists are expected to be concerned with objective knowledge (Gelman, 2008).

Another objection to Bayesianism is the *problem of old evidence* (Glymour, 1981): suppose a new hypothesis is proposed, and it turns out to explain old evidence very well. How can the old evidence be used to confirm this hypothesis? Related is the *problem of new theories* (Earman, 1992): the standard Bayesian framework does not provide a way to incorporate new hypotheses in course of the learning process. This problem is addressed in Chapter 2.

1.2.3 Classical statistics and frequentism

The major alternative to Bayesian statistics is *classical statistics*. It is really a hotchpotch of many different methods, philosophical views, and interpretations of *probability* (see e.g. Section 1.5 and Hájek (2019)). The common factor is that it only considers probability assignments over the sample space and not over parameters that themselves represent probability distributions. The most important interpretation of the concept of probability in classical statistics, developed by Von Mises (1939), is that it can be identified with a relative frequency: we can describe the probability of a coin landing 'tails', with the number of tails in a (very long) sequence of coin tosses, divided by the total number of tosses. This is called *frequentism*. Since this is the predominant view, classical statistics is often called *frequentist statistics*, but methods based on other physical interpretations of probability, such as propensity, are considered classical as well.

1.3 The topics of this dissertation: challenges, limitations, and pragmatics

I now give a high-level description of the main topics of this dissertation. This is followed by a brief, specific introduction for every chapter.

Bayesian inference under model misspecification

The Bayesian framework as described above provides us with a way to change our degrees of belief over time when new evidence obtains. A Bayesian learner starts with specifying a model, and assigning prior probabilities to its elements. If the model is appropriate, i.e. if the true data generating process is in the model and the prior does not exclude it from the start, consistency is guaranteed: the learner will converge on the truth as more and more data are obtained. However, it might happen that the model is *misspecified*: the true data generating process is not part of the model (or is assigned zero prior probability), which can be problematic in different ways, and in this dissertation, Bayesianism is extended in two different ways to face the problem.

First, it might happen that in the course of the learning process, the learner wants to incorporate an hypothesis that did not occur to her before. The standard Bayesian framework does not offer a way how to incorporate new hypotheses, it seems that the learner has to throw away her data and start from scratch by specifying the larger model and assigning prior probabilities to its elements. In Chapter 2, further introduced in Section 1.4, we consider an *open-minded Bayesian logic*, to allow for dynamically incorporating new hypotheses.

Secondly, it could be that we want Bayes to concentrate on the *best* element in the model, instead of the truth, which is outside the model, where the *best* is the element that is closest to the truth. In Chapter 6, we show that standard Bayesian inference can fail to concentrate on this best element in the model. We subsequently modify Bayes theorem (1.1) by equipping the likelihood with an exponent, called the *learning rate*, and call this *generalised Bayes*. When the learning rate is chosen appropriately, generalised Bayes concentrates on the best element in the model. In Section 1.6 this problem is presented further.

Bayes factor hypothesis testing under optional stopping

Bayes factor hypothesis testing is a Bayesian approach to hypothesis testing based on the ratio of two Bayes marginal likelihoods. In Chapters 3 and 4, we study *optional stopping*, which informally means 'looking at the results so far to decide whether or not to gather more data'. Different authors make claims about whether or not Bayes factor hypothesis testing is robust under optional stopping, but it turns out that one can give three different mathematical definitions of what *robustness under optional stopping* actually means. We see in Chapters 3 and 4 that adhering to one of the varieties of Bayesianism has implications for the claims one can make in practise. For example, in Chapter 3 we elucidate claims about optional stopping which are only meaningful from a purely subjective Bayesian perspective, yet the suggestion is made as if those claims apply to pragmatic inference. In Section 1.5 I give an overview of current practise in hypothesis testing, with P-values, and with Bayes factors.

A new theory for hypothesis testing with a Bayesian interpretation

In Chapter 5 we introduce a new theory for hypothesis testing. The central concept of this theory is the E-variable, a random variable similar to, but in many cases an improvement of the P-value. We introduce an optimality criterion, called GROW, for designing E-variables,

and it turns out that these GROW E-variables have an interpretation as a Bayes factor, yet with special priors, which are very different from those currently used by Bayesians. This is an example of radical pragmatism: we do not choose these priors based on any philosophical considerations, but these special priors are designed so that the resulting method satisfies some practically motivated criterion — namely, the GROW. One could even state that the Bayesian interpretation of GROW E-variables is merely a by-product, yet a convenient one, because it provides a common language for adherents of different frequentist and Bayesian testing philosophies. In Section 1.5 these schools of hypothesis testing (Fisherian, Neyman-Pearsonian, the commonly used hybrid form with P-values, and Bayesian) are briefly discussed.

Best-arm identification with a Bayesian-flavoured algorithm

Another example of radical pragmatic Bayesianism can be found in Chapter 7. There, we want to identify from a sequence of probability distributions the one with the highest mean. We can assign prior probabilities to distributions ν_j, $j = 1, \ldots, K$ of having the highest mean, and update these with Bayes' theorem when we obtain a sample. We can construct a rule which distribution to sample at time t based on the posterior distribution, but in order to meet certain frequentist (and Bayesian) criteria, we do not always pick the distribution with the highest posterior probability of having the highest mean. The setting of Chapter 7, which is called *Best-arm identification*, is introduced in Section 1.7.

1.4 Chapter 2: Merging

In Chapter 2, we consider the problem of dynamically incorporating hypotheses during the Bayesian learning process. Here, successful learning means that if the true data generating process is added to our model at some point, the learner almost-surely converges to the truth as more and more data becomes available.

Setting Let the sample space be the set of all infinite sequences, denoted by $\mathcal{X}^\infty$, and consider a σ-algebra $\mathcal{F}_\infty$ containing all Borel sets[4]. We can for example look at the space of all binary infinite sequences, 2^ω (Cantor space). Now let H^* and P be two probability measures over this measurable space $(\mathcal{X}^\infty, \mathcal{F}_\infty)$ of infinite sequences, and denote with $A \in \mathcal{F}_\infty$ a *proposition*[5]. An example of such a proposition is: 'the frequency of ones is equal to 0.4', or 'every other bit is the next bit of π'. We think of H^* as the *truth*, i.e. the distribution generating the data, and we can view P as the learner's belief distribution.

The learner starts with a number of propositions A_i, $i \in \mathbb{N}$, to which she assigns a prior belief $P(A_i)$. At each time step t she observes an evidence item $x_t \in \mathcal{X}$, and she updates her belief in the Bayesian way: her posterior belief in proposition A_i is

$$P(A_i|x^t) = \frac{P(A_i \cap x^t)}{P(x^t)}.$$

[4] For a more detailed exposition, see Chapter 2.

[5] Many authors call this an *hypothesis*, but to keep the introduction simple and to avoid confusion with statistical hypotheses, I call it a proposition here, following e.g. Huttegger (2015).

The learning goal If H^* is the true distribution that governs the generation of the data, the learner should use the data coming from H^* to change her beliefs P towards H^*. Eventually, if she sees enough data, we want P to come *close* to H^*. There are many notions for this closeness, and an obvious one would be concentration of the learner's posterior distribution on the true distribution. However, this is too strong for our purposes, as we do not want to exclude the possibility of different distributions that are from some point on empirically equivalent (see Lehrer and Smorodinsky (1996)). Thus, we will use the notion of *truth-merger*, which comes in two variants. The first is called *strong merger* (Kalai and Lehrer, 1993; Lehrer and Smorodinsky, 1996; Leike, 2016)), which is still reasonably strong, as discussed in Chapter 2.

Definition 1.1 (Strong truth-merger)**.** P merges with the truth H^* if H^*-almost surely

$$\sup_{A\in\mathcal{F}_\infty} \left|P(A|x^t) - H^*(A|x^t)\right| \to 0 \quad \text{as} \quad t\to\infty.$$

In words, with true probability 1, the learner's probabilities conditional on the past will asymptotically coincide with the true probabilities. Truth-merger is thus concerned with learning the probabilities of future outcomes. In Chapter 2, we are mainly concerned with the predictive probabilities up to a finite point in time, which is captured in the notion of *weak merger* (Lehrer and Smorodinsky, 1996):

Definition 1.2 (Weak truth-merger)**.** We say that P weakly merges with the truth H^* if and only if for $\ell\in\mathbb{N}$ we have H^*-almost surely

$$\sup_{A\in\mathcal{F}_{t+\ell}} \left|P(A|x^t) - H^*(A|x^t)\right| \to 0 \quad \text{as} \quad t\to\infty,$$

where $\mathcal{F}_{t+\ell}$ denotes the σ-algebra generated by the first $t+\ell$ outcomes.

Strong merger implies weak merger, as follows directly from the definitions.

Contribution In the standard form, a Bayesian learner starts with specifying her prior distribution P, and learns by conditionalisation on the data. The prior specifies a particular model (set of hypotheses to which positive probability is assigned), and if the truth H^* is in this model and H^* is absolutely continuous with respect to P, then she will almost surely merge with the truth (Blackwell and Dubins, 1962). However, she cannot include every hypothesis from the start (see Chapter 2), she needs to commit to restrictions on her model (inductive assumptions), and there is no room to adapt the model later on in the standard form of Bayesianism as described in Section 1.2. In particular, she can not expand the model to incorporate new hypotheses (the Bayesian *problem of new theory*) that might be more in accordance with the data than the hypotheses in the initially formulated model. For example, somebody might come along and tell her about a new hypothesis that is eminently reasonable but which she simply did not think of. Thus, the challenge is to come up with an *open-minded* Bayesian inductive logic that can dynamically incorporate new hypotheses. Wenmackers and Romeijn (2016) formalise this idea, but in Chapter 2 we show that their proposal does not preserve merger with the true hypothesis. We then diagnose the problem, and offer two versions of a *forward-looking* open-minded Bayesian that do weakly merge with the truth when it is formulated.

1.5 Chapters 3, 4 and 5: Hypothesis testing

A large part of this dissertation is about hypothesis testing. Here I first introduce this topic in a simplified setting; for a more general treatment, see Chapter 4, Section 4.4. I then summarise our contributions of Chapters 3, 4 and 5.

Setting Let the *null hypothesis* $\mathcal{H}_0$ and the *alternative hypothesis* $\mathcal{H}_1$ be statistical hypotheses, i.e. sets of probability distributions on a measurable space $(\Omega, \mathcal{F})$. Let $X^n \coloneqq X_1, \ldots, X_n$ be random variables taking values in the outcome space Ω.

The learning goal We wish to test the veracity of $\mathcal{H}_0$, possibly in contrast with some alternative $\mathcal{H}_1$, based on a sample X^n that may or may not be generated according to an element of $\mathcal{H}_0$ or $\mathcal{H}_1$. There are several paradigms for testing, based on different philosophies and also with different objectives. The most commonly used framework for hypothesis testing in the applied sciences, often referred to as *classical* or *frequentist*, is that of the p-value based null hypothesis significance testing (NHST).

Definition 1.3 (P-value)**.** A p-value is a random variable P such that for all $0 \leq \alpha \leq 1$ and all $P_0 \in \mathcal{H}_0$, we have $P_0(P \leq \alpha) = \alpha$.

P-values were advocated by Sir Ronald Fisher to measure the strength of evidence against the null hypothesis, a smaller p-value indicating greater evidence (Fisher, 1934). In his framework of *significance testing*, the learner comes up with a null hypothesis that the sample comes from an infinite population with known (hypothetical) distribution, so if the data are unusual under $\mathcal{H}_0$, it constitutes evidence against the null. The *level of significance* is simply a convention[6] to use as a cut-off level for rejecting $\mathcal{H}_0$. In his later work (Fisher, 1955; Fisher, 1956), he refined this and prescribed to report the exact level of significance, which is thus a property of the data.

Jerzy Neyman and Egon Pearson developed an alternative theory of null hypothesis testing where the main concern is to limit the false positive rate of the test, and a second hypothesis, the *alternative hypothesis* needs to be specified. As opposed to the Fisherian framework in which the p-value is a measure of evidence, the outcome of the Neyman-Pearson test is acceptance or rejection of the null hypothesis. The probability α of falsely rejecting the null hypothesis when it is true is called the *Type I error*, the probability β of falsely accepting the null hypothesis is called the *Type II error*, and the complement $1 - \beta$ is called the *power* of a test. If we fix the significance level α, a *most powerful* test is the one that minimises the Type II error β, and Neyman and Pearson proved in the famous lemma named after them, that such a most powerful test for simple $\mathcal{H}_0$ and $\mathcal{H}_1$ has the form of a likelihood ratio threshold test. Note that in this framework the significance level is a property of the test. Whereas in Fisher's framework, p-values from single experiments provide evidence against $\mathcal{H}_0$, in the Neyman-Pearsonian framework the behaviour of the test in the long run is considered, and we can view the significance level α as a relative frequency of the Type I errors over many repeated experiments. As such, a test does not

[6] According to some authors (Hubbard, 2004; Gigerenzer and Marewski, 2014), the 5% level was taken just because 5% tables were available to Fisher at the time he wrote his earlier works.

provide evidence for the truth or falsehood of a particular hypothesis (Neyman and Pearson, 1933).

The current practise of the p-value based NHST is, remarkably, a hybrid of the methods proposed by Fisher on the one hand, and Neyman and Pearson on the other hand, despite their utter disagreement about hypothesis testing (see Hubbard (2004) who quotes their reciprocal reproaches), and the conflicting aspects of their theories of inference. Typically, a significance level α is pre-specified (often 0.05), then an experiment is designed so that it achieves a certain power $1-\beta$, and after the data are obtained a p-value is calculated. When the p-value is smaller than α, the null hypothesis is rejected, and in many journals, the p-value is reported as well, often with a superscript of one or more stars[7] indicating whether $p < 0.05$, $p < 0.01$, or $p < 0.001$. As early as the 1960s (e.g. Edwards, Lindman and Savage (1963)), many papers have been published in which the p-value based NHST is criticised. Besides that it is a combination of the two (incompatible) frameworks described above, it is criticised because of the widespread misinterpretations of p-values (for example, they are thought to be equal to the Type I error rate, or to the probability of an hypothesis being true given the data), their dependence on counterfactuals and the need of the full experimental protocol to be determined upfront. For articles debating the use of p-value based NHST, see e.g. Berger and Sellke (1987), Wagenmakers (2007), Gigerenzer and Marewski (2014), Grünwald (2016), Wasserstein, Lazar et al. (2016) and Benjamin et al. (2018). For work on Fisherian versus Neyman-Pearsonian views, see e.g. Gigerenzer et al. (1990), Gigerenzer (1993) and Hubbard (2004), and an interesting investigation into why many are unaware of these different views and their incompatibility is Huberty (1993).

Another framework for hypothesis testing is based on *Bayes factors* (Jeffreys, 1961; Kass and Raftery, 1995). Since the last decade this framework has been advocated by several researchers as an alternative for the p-value based NHST (see e.g. Wagenmakers (2007)). Here, $\mathcal{H}_0$ and $\mathcal{H}_1$ are represented by measures P_0 and P_1 that are taken to be Bayesian marginal distributions. Denote $\mathcal{H}_j = \{P_{\theta|j}; \theta \in \Theta_j\}$, with (possibly infinite) parameter spaces Θ_j, and define prior distributions π_0 and π_1 on Θ_0 and Θ_1 respectively. The Bayes marginals then are, for any set $A \subset \Omega$

$$P_0(A) = \int_{\Theta_0} P_{\theta|0}(A)\mathrm{d}\pi_0(\theta) \quad ; \quad P_1(A) = \int_{\Theta_1} P_{\theta|1}(A)\mathrm{d}\pi_1(\theta). \tag{1.3}$$

The Bayes factor is defined as the ratio of these Bayes marginals (for simple $\mathcal{H}_0$ and $\mathcal{H}_1$ this is simply a likelihood ratio). Sometimes we want to allow for improper prior distributions (integrating to infinity). For this case, we give a more general definition in Chapter 4, in terms of versions of the Radon-Nikodym derivatives of P_0 and P_1 w.r.t. some underlying measure. A large Bayes factor corresponds to evidence against the null hypothesis. Sometimes, one can also obtain frequentist Type-I error guarantees with Bayes factors. The probability under (an element of) the null hypothesis that a Bayes factor based on a sample with a fixed size n is larger than $1/\alpha$ for $\alpha \in (0,1)$ is by Markov's inequality bounded by α. Thus, one can use Bayes factors together with a frequentist Type I error guarantee by choosing a threshold of $1/\alpha$, and rejecting the null if the Bayes factor exceeds that threshold.

[7] See Gigerenzer and Marewski (2014) for an excellent critique of, as they call it, *the null ritual*.

Contribution (1) When a researcher wants to use p-value based NHST, the experimental protocol must be completely determined upfront. In practise, researchers often adjust the protocol due to unforeseen circumstances, or collect data until a point has been proven. This is often referred to as *optional stopping*. Informally, this means: 'looking at the results so far to decide whether or not to gather more data'. With (standard) p-value based NHST (aiming to control the Type I error) this is not possible: one can prove that even if the null hypothesis is the true data generating process, one is guaranteed to reject it upon collecting and testing more and more data. Bayes factor hypothesis testing on the other hand has been claimed by several authors to continue to be valid under optional stopping. But what does it mean for a test to *remain valid under optional stopping*? It turns out that different authors mean quite different things by 'Bayesian methods can handle optional stopping', and such claims are often only made informally, or in restricted settings. We can discern three main mathematical concepts of *handling optional stopping*, which we identify and formally define in Chapter 4: *τ-independence, calibration and (semi-)frequentist*. We also mathematically prove that Bayesian methods can indeed handle optional stopping in many (but not all!) ways, in many (but not all!) settings. While Chapter 4 is written to untangle the optional stopping confusion by giving rigorous mathematical definitions and theorems, Chapter 3 is written for practitioners and methodologists who want to work with *default* Bayes factors introduced by the self-named Bayesian psychology community (Rouder et al., 2009; Jamil et al., 2016; Ly, Verhagen and Wagenmakers, 2016). That chapter is mainly a response to the paper *Optional stopping: no problem for Bayesians* (Rouder, 2014), and we explain for a non-mathematical audience why there is more nuance to this issue than Rouder's title suggests, and why his claims (which are actually about *calibration*, which we formally define in Chapter 4.4.2, and *not* about Type I error control) are relevant only under a subjective interpretation of priors. *Default* priors do not have such an interpretation, making the relevance of Rouder's claims for practise doubtful. In Chapter 4 we prove that Rouder's intuitions about calibration are correct, but they do not carry over to other notions of optional stopping than calibration; and therefore they do not apply to most practically relevant issues with optional stopping with Bayes factor hypothesis testing.

Contribution (2) Many agree that the p-value based NHST paradigm is inappropriate (or at least suboptimal) for scientific research, yet the dispute about its replacement continues to be unresolved. Some propose a Bayesian revolution (yet sometimes overlook the limitations of Bayesian approaches, see Chapter 3 and 4), others adhere to more Fisherian or Neyman-Pearsonian views. Finally, some are more pragmatic and just want to use an appropriate test for their situation that gives them certain guarantees. Wouldn't it be nice to have a common language for adherents to those different testing schools that expresses strength of evidence, that allows for evidence from experiments originating from those different paradigms to be freely combined, and that resolves some of the main problems with p-values, such as interpretability issues for practitioners? In Chapter 5 we introduce a theory for hypothesis testing based on *E-test statistics* (we call them E-variables) that achieve just that. The definition of an E-variable is simple:

Definition 1.4 (E-test statistic)**.** An E-test statistic is a non-negative random variable E satisfying

$$\text{for all } P \in \mathcal{H}_0\text{: } \mathbf{E}_P[E] \leq 1.$$

E-variables are flexible: they can be based on Fisherian, Neyman-Pearsonian and Bayesian testing philosophies; E-variables resultant from those different paradigms can be freely combined while preserving Type I error guarantees, and they allow for a clear interpretation in terms of money or gambling. In Chapter 5 we develop this theory of E-variables; this includes the development of optimal, 'GROW' E-variables.

1.6 Chapter 6: Generalised linear regression

In Chapter 6, we consider Bayesian generalised linear regression under model misspecification. Here, successful learning means that the (generalised) posterior distribution concentrates on an element in the model that is in some sense *optimal*, although it is not the true data generating distribution (which is not in the model). I start by explaining linear regression in the well-specified case. I then introduce the (more general) learning goal and summarise our contributions.

Setup In *linear regression*, we wish to find a relationship between a regressor variable $X \in \mathcal{X}$ and a regression variable $Y \in \mathbb{R}$, where $\mathcal{X}$ is some set. We want to learn a function $g : \mathcal{X} \to \mathbb{R}$ from the data, and we assume Gaussian noise on Y, that is $Y_i = g^*(X_i) + \varepsilon_i$, where $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, and g^* is the true function we want to learn. We can thus formulate the conditional density of Y^n given X^n as

$$p_{g,\sigma^2}(Y^n|X^n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (Y_i - g(X_i))^2}{2\sigma^2}\right).$$

In *linear* regression, we search for a function g for our problem among linear combinations of *basis functions*: $g_\beta(X) = \sum_{j=1}^p \beta_j g_j(X)$. This can be further extended to *generalised linear models* (GLMs), where the dependent variable Y is not necessarily continuous-valued any more (but from some set $\mathcal{Y}$), and the noise is not necessarily Gaussian. An example that we encounter in Chapter 6 is the *logistic regression model* $\{f_\beta : \beta \in \mathbb{R}^p\}$, where the outcomes $Y_i \in \{0, 1\}$ are binary random variables, the independent variables are p-dimensional vectors $X_i \in \mathbb{R}^p$, with the conditional density

$$p_{f_\beta}(Y_i = 1|X_i) := \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}}.$$

Learning goal We are given an i.i.d. sample $Z^n \sim P$ from a distribution P on the sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and we want to do inference with the generalised Bayesian posterior Π_n on our model $\mathcal{F}$, defined by its density

$$\pi_n(f) := \frac{\exp\left(-\sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f)}{\int_{\mathcal{F}} \exp\left(-\sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f)\, \mathrm{d}\rho(f)}. \tag{1.4}$$

Here $\ell_f(z_i)$ is the loss of f, an element of our model $\mathcal{F}$, on outcome $z_i \in \mathcal{Z}$, and π_0 is the density of a prior distribution on $\mathcal{F}$ relative to some underlying measure ρ. For GLMs, (1.4) can equivalently be interpreted in terms of the standard Bayesian posterior based on the conditional likelihood $p_f(y|x)$, i.e.

$$\pi_n(f) \propto \prod_{i=1}^{n} (p_f(y_i|x_i))\pi_0(f). \tag{1.5}$$

For example, consider standard linear regression with square loss $\ell_f(x, y) = (y - f(x))^2$ and fixed learning rate η. Then (1.4) induces the same posterior $\pi_n(f)$ over $\mathcal{F}$ as does (1.5) with $p_f(y|x) \propto \exp(-(y - f(x))^2)$, which is the same as (1.4) with ℓ_f replaced by the conditional log-loss $\ell'_f(x, y) := -\log p_f(y|x)$. All examples of GLMs in Chapter 6 can be interpreted in terms of (1.5) for a misspecified model, that is, the density $P(Y|X)$ is not equal to p_f for any $f \in \mathcal{F}$.

We do not assume that the model is well-specified, however, we do assume that there exists an optimal element in our model $f^* \in \mathcal{F}$ that achieves the smallest *risk* (expected loss) $\mathbf{E}[\ell_{f^*}(Z)] = \inf_{f\in\mathcal{F}} \mathbf{E}[\ell_f(Z)]$. For GLMs this has additional interpretations: it means that if $\mathcal{F}$ contains the true regression function g^*, then $f^* = g^*$, and also, f^* is the element in $\mathcal{F}$ closest to P in KL divergence $\inf_{f\in\mathcal{F}} \mathbf{E}_{X,Y\sim P}[\log(p(Y|X)/p_f(Y|X)]$. As more and more data becomes available, we want the Bayesian posterior (1.4) to concentrate in neighbourhoods of f^*.

Contribution In the last decade it has become clear that standard Bayesian inference can behave badly under *model misspecification*, that is, when the true distribution P is not in the model $\mathcal{F}$. Grünwald and Van Ommen (2017) give a simple linear regression example in which Bayesian model selection, model averaging and ridge regression severely overfit: Bayes learns the noise of the sample in stead of (or in addition to) the signal. For small sample sizes, the posterior does not concentrate on element $f^* \in \mathcal{F}$ closest in KL divergence to the true distribution P, even if the true *regression function* is in the model (in their example, only the noise is misspecified). They also provide a remedy for this problem: using the appropriate *generalised* Bayesian posterior, defined analogously to (1.4) by its density

$$\pi_n(f) := \frac{\exp\left(-\eta \sum_{i=1}^{n} \ell_f(z_i)\right) \cdot \pi_0(f)}{\int_{\mathcal{F}} \exp\left(-\eta \sum_{i=1}^{n} \ell_f(z_i)\right) \cdot \pi_0(f)\, \mathrm{d}\rho(f)},$$

where $\eta > 0$ is the *learning rate*, and $\eta = 1$ corresponds to standard Bayesian inference. They show with simulations that for *small enough* η (which can be found by the *Safe-Bayesian algorithm*), this results in excellent performance. In Chapter 6 we show that failure of standard Bayes ($\eta = 1$) and empirical success of generalised Bayes (with small enough η) on similar toy problems extends to more general priors (lasso, horseshoe) than considered by Grünwald and Van Ommen (2017) and more general models (GLMs). Additionally, we show real-world examples on which generalised Bayes outperforms standard Bayes. Grünwald and Mehta (2019) showed concentration with high probability of generalised Bayes with learning rate $\overline{\eta}$ in the neighbourhood of f^* under the $\overline{\eta}$-*central condition*. In Chapter 6 we show under what circumstances this central condition holds for GLMs. Furthermore, we provide MCMC algorithms for generalised Bayesian lasso and logistic regression.

1.7 Chapter 7: Best-arm identification

In this chapter we consider a Bayesian-flavoured *anytime* best-arm identification strategy. We show that it is in some sense optimal, meaning that it (asymptotically) uses as few samples as possible to indicate with a certain confidence which of a sequence of probability distributions has the highest mean.

Setup A finite stochastic *multi-armed bandit* model is a sequence of K probability distributions $\nu = (\nu_1, \dots, \nu_K)$, which we call *arms*. With μ_i we denote the expectation of distribution ν_i of arm i (assumed it exists), and we denote the optimal arm $I^\star$ to be the arm with mean $\mu^\star \coloneqq \max_{i \in [K]} \mu_i$, assuming it to be unique. The learner, who does not know about the distributions ν, interacts with the model by choosing at each time $t = 1, 2, \dots$ an arm I_t to sample, and she observes an evidence item, called *reward*, $Y_{t,I_t} \sim \nu_{I_t}$. The learner chooses arm I_t based on the history $(I_1, Y_{1,I_1}, \dots, I_{t-1}, Y_{t-1,I_{t-1}})$, and possibly some side information or exogenous randomness, denoted by U_{t-1}. Let $\mathcal{F}_t$ be the σ-algebra generated by $(U_0, I_1, Y_{1,I_1}, U_1, \dots, I_t, Y_{t,I_t}, U_t)$, then I_t is $\mathcal{F}_{t-1}$-measurable. The sequence of random variables $(I_t)_{t \in \mathbb{N}}$ is called the *strategy* of the learner or a *bandit algorithm*.

Learning goal We consider a setting called *best-arm identification* (BAI), the name says it all: we want to *explore* the arms to make an informed guess which one has the highest mean. A BAI strategy consists of three components. The *sampling rule* selects an arm I_t at round t. The *recommendation rule* returns a guess for the best arm at time t (it is $\mathcal{F}_t$-measurable), and thirdly, the *stopping rule* τ, a stopping time with respect to $(\mathcal{F}_t)_{t \in \mathbb{N}}$, decides when the exploration is over. Two main mathematical frameworks for BAI exist. One is the *fixed-budget* setting, where the stopping time τ is fixed to some (known) maximal budget, and the goal is to minimise the probability of returning a suboptimal arm (Audibert and Bubeck, 2010). The other, which we consider in Chapter 7, is the *fixed-confidence* setting, in which given a risk parameter δ, the goal is to ensure that the probability to stop and recommend a suboptimal arm is smaller than δ, while minimizing the total number of samples $\mathbb{E}[\tau]$ to make this δ-correct recommendation (Even-dar, Mannor and Mansour, 2003). There exist several sampling rules for the fixed-confidence setting, most of them depend on the risk parameter δ, but one that does not is the *tracking* rule of Garivier and Kaufmann (2016), which also asymptotically achieves the minimal sample complexity combined with the Chernoff stopping rule (see ibid. and Chapter 7). A sampling rule that does not depend on the risk parameter δ or a budget is called *anytime* by Jun and Nowak (2016), and is appealing for many (future) applications in machine learning, such as hyper-parameter optimisation.

Contribution We consider a *Bayesian-flavoured* anytime sampling rule introduced by Russo (2016), called Top-Two Thompson Sampling (TTTS). So far, there has been no theoretical support for the employment of TTTS for fixed-confidence BAI, and Russo (2016) proves posterior consistency (with optimal rates) under restrictive assumptions on the models and priors, excluding two settings mostly used in practise: Gaussian and Beta-Bernoulli bandits. In Chapter 7 we address the following: We (1) propose a new Bayesian sampling rule (T3C), computationally superior to TTTS; (2) establish δ-correctness of two new Bayesian stopping and recommend-

ation rules; (3) provide sample complexity analyses of TTTS and T3C under our proposed stopping rule; (4) prove optimal posterior convergence rates for Gaussian and Beta-Bernoulli bandits.

1.8 This dissertation

Each of the chapters of this dissertation corresponds to one of the papers listed on page i, therefore, the chapters are self-contained. However, they are written for different audiences, hence they differ greatly in style, in technical level, and in background knowledge required to be able to read them.

Chapter 2 (Sterkenburg and De Heide, 2019) is written for a readership of mathematical philosophers. Mathematical philosophy is a field in which philosophical questions are treated with tools and methodology from mathematics: with definitions, theorems and proofs, and with precision and rigour. Mathematical philosophers often have a strong understanding of some fields in mathematics, such as measure theoretic probability, set theory, and of course logic.

Chapter 3 (De Heide and Grünwald, 2018) is written for statisticians and methodologists. These are often researchers in a methodology department associated to a faculty of applied research (psychology, biology, etc.), who study and develop statistical methods for their field. Only elementary probability theory and statistics is needed to read this chapter.

The subsequent four chapters (Hendriksen, De Heide and Grünwald, 2020; Grünwald, De Heide and Koolen, 2019; De Heide et al., 2020; Shang et al., 2020) are aimed at mathematical statisticians and machine learning theorists with a solid mathematical background (in particular, obviously, mathematical statistics and probability theory). For Chapter 4 some familiarity with group theory is useful to fully appreciate it, and for Chapter 6 the same holds for statistical learning theory, although in both chapters all necessary preliminaries are (concisely) provided.

Chapter 2

On the Truth-Convergence of Open-Minded Bayesianism

Abstract

Wenmackers and Romeijn (2016) formalize ideas going back to Shimony (1970) and Putnam (1963) into an *open-minded* Bayesian inductive logic, that can dynamically incorporate statistical hypotheses proposed in the course of the learning process. In this paper, we show that Wenmackers and Romeijn's proposal does not preserve the classical Bayesian consistency guarantee of almost-sure merger with the true hypothesis. We diagnose the problem, and offer a *forward-looking* open-minded Bayesians that does preserve a version of this guarantee.

2.1 Introduction

On the standard philosophical conception of Bayesian learning, an agent starts out with a particular prior distribution and learns by conditionalizing on the data it receives. Well-known results on the merger of opinion show that the specific prior does not matter too much, as long as there is agreement on what is possible at all. These same results can also be taken to show that the agent converges to the truth, as long as its prior does not exclude this truth from the start (Earman, 1992, 141ff; Huttegger, 2015).

However, a Bayesian agent cannot include in its prior *every* possible truth from the start; not in practice, and not even in theory (Putnam, 1963; Dawid, 1985; Belot, 2013; Sterkenburg, 2019). A Bayesian agent must commit to restrictive *inductive assumptions* in its initial choice of prior (Howson, 2000; Romeijn, 2004). Standard results about convergence to the truth only apply if these initial assumptions are actually valid in the learning situation at hand. But there is, on the standard conception, no room for the agent to readjust (Levi, 1980); not even if these assumptions start looking faulty.

In more explicitly statistical terms, a Bayesian agent's prior can be seen to specify a particular *model*, or set of hypotheses. If the model is appropriate, if one of the hypotheses is *true*, there is—at least for a countable model—a guarantee of *consistency* that the agent with probability 1 (almost surely, a.s.) converges on this truth. But if it is not, the agent's beliefs can with positive probability always and forever remain off the mark. On the standard conception, there is, again, no room for the agent to later adapt this model (Dawid, 1982); there is, in particular, no room to expand the model, to incorporate new hypotheses that might be more in accord with the data (Gillies, 2001; Gelman and Shalizi, 2013).

The question of how to open up the standard conception to make room for incorporating new hypotheses is the Bayesian *problem of new theory* (Chihara, 1987, 556ff; Earman, 1992, 132f; Romeijn, 2005b). An early account that engages with the problem of new theory is the *tempered personalism* due to Shimony (1970). Central to Shimony's account is an idea he traces back to Putnam (1963; see Shimony, 1970, p. 89; 1969), and in more veiled form to Jeffreys (1961; see Shimony, 1970, 97ff; also see Howson, 1988). This is the idea that, rather than taking as starting point an hypothesis set that is as wide as possible, Bayesian inference is relative to a limited set of "seriously proposed hypotheses," that is dynamically expanded as new such hypotheses are proposed. In this context Shimony introduced the notion of a *catch-all hypothesis* that is the complement of all seriously proposed hypotheses at any given time.

Recently, Wenmackers and Romeijn (2016) have worked out these ideas in a statistical setting, into what they brand an *open-minded* Bayesianism. In a number of different versions they propose a Bayesian inductive logic that allows for an agent to adopt newly formulated statistical hypotheses during the learning process.

One important question that they leave untouched, however, is whether these formalizations actually preserve the consistency guarantee of truth-convergence. That is, if the *true* hypothesis is one of the actually formulated hypotheses, thus becomes part of the open-minded Bayesian's hypothesis set, is the agent from that point on still guaranteed to almost surely converge on this truth? That is the question we investigate in this paper.

We proceed as follows. First, in section 7.5, we introduce the statistical framework of Bayesian learning that Wenmackers and Romeijn employ, and discuss their different versions of open-minded Bayesians. Then, in section 2.3, we investigate the guarantee of convergence to the truth. We focus on the property of *weak merger* with the true hypothesis, whenever part of the hypothesis set, and show that all the proposed versions of open-minded Bayesianism, unlike the standard Bayesian, *fail* to guarantee this property. In section 2.4 we diagnose the problem and the exact nature of the convergence we could possibly attain, in the course of which we introduce the notions of an *hypothesis* and *posterior scheme* and that of a *completed agent measure*. We then set out for a version of open-minded Bayesianism for which we can show, for every hypothesis and posterior scheme, strong merger of the completed agent function, from which weak merger of the agent follows. This leads us, finally, to our proposal of a *forward-looking* open-minded Bayesian. The general threat to truth-convergence lies in the possibility of an endless stream of overfitting hypotheses: our forward-looking proposal meets this threat by neutralizing the role of old evidence. In an initial proto-version this is achieved by a constraint on the *posteriors* assigned to new hypotheses; in the final version this is achieved by combining a constraint on new hypotheses' *priors* (instantiating the idea of the catch-all) with the stipulation

that new hypotheses' likelihoods on old evidence are equal to the *agent's own past probability assignment*.

We should emphasize that Wenmackers and Romeijn in their paper (and we in this paper) are concerned with the question of how to *incorporate* externally proposed new hypotheses: their proposals are attempts to make this aspect part of a Bayesian logic of inductive inference. They are in their paper (and we are here) *not* concerned with *when* new hypotheses should be taken into consideration, let alone with *how* new hypotheses are conceived. To paraphrase Lindley (2000, p. 303) paraphrasing de Finetti: if you have your statistical model, reasoning is mere calculation, but constructing your model actually requires *thinking*. We are here only concerned with the former, but presume, with Wenmackers and Romeijn, that the scope of mere calculation may be slightly extended, to the procedure of incorporating given new hypotheses into your model.

2.2 The open-minded Bayesians

In this section, we first set out the presupposed formal framework (sect. 2.2.1), and then discuss the standard Bayesian (sect. 2.2.2), the *vocal* open-minded Bayesian (sect. 2.2.3), the *silent* open-minded Bayesian (sect. 2.2.4) as well as its *retroactive* variant (sect. 2.2.5), and finally the *hybrid* open-minded Bayesian (sect. 2.2.6).

2.2.1 Formal framework: outcomes and hypotheses

In the statistical set-up employed by Wenmackers and Romeijn,[1] the domain of a Bayesian agent's probability function is the Cartesian product $\Omega \times \Theta$ of an *outcome space* Ω and a *statistical hypothesis space* Θ.

The outcome space

In all of the following, we assume the simple scenario of repeatedly sampling from two possible elementary outcomes, 0 and 1. Formally, the outcome space Ω is the space $\{0,1\}^\omega$ of all infinite binary sequences E^ω. It is convenient for our purposes to treat a probability measure over this space as a function P over the *finite* sequences, that satisfies $P(\varnothing) = 1$, where $\varnothing$ is the empty outcome sequence, and $P(E^t) = P(E^t0) + P(E^t1)$ for all finite outcome sequences E^t, where E^tE denotes outcome sequence E^t of length t followed by elementary outcome $E \in \{0,1\}$. Formally, the set of *cones* $[\![E^t]\!] := \{E^\omega \in \Omega : E^\omega \text{ extends } E^t\}$ for all finite sequences E^t generates a σ-algebra $\mathfrak{F}$ over Ω containing all the Borel sets, and an assignment P as above induces a unique measure μ on $(\Omega, \mathfrak{F})$ with $\mu([\![E^t]\!]) = P(E^t)$ for all finite E^t.

The hypothesis space

We consider *statistical* hypotheses that are given by likelihood functions over the possible outcomes. That is, we take hypotheses H to be themselves probability measures over the

[1] For a recent alternative proposal for open-minded Bayesianism in a framework that does not explicitly deal with statistical hypotheses, see Raidl (2020).

outcome space.

As a basic example, the i.i.d. or *Bernoulli* hypothesis H_θ with parameter $\theta \in [0,1]$ assigns each length-t data sequence E^t a probability $H_\theta(E^t) = \theta^{t_1} \cdot (1-\theta)^{t-t_1}$ with t_1 the number of 1's in E^t. This induces one-step conditional probabilities $H_\theta(1 \mid E^t) = \theta$ at each time point t, i.e., no matter the past sequence E^t. Thus H_θ formalizes the data-generating process where the same elementary outcome is always produced with the same probability; for instance, the process of repeatedly tossing a coin (heads is 1, tails is 0) with bias θ.

Other hypotheses can express various dependencies of current probabilities on the structure of the past data. At the extreme end are *deterministic* hypotheses, that at each point in time only allow for one particular next outcome. This corresponds to a function assigning probability 1 to each initial segment of one particular infinite outcome stream E^ω.

We will assume that at any time there are only a finite number of explicitly formulated hypothesis. These N hypotheses $H_0, \ldots, H_{N-1}$ are collected in a hypothesis set $\Theta_N := \{H_i\}_{i<N}$.

Below we will consider expanding sequences of hypotheses sets, for which the following notation will be useful. Let $N(t)$ denote the number of hypotheses formulated before time t, so that the hypothesis formulated at time t (if it exists) is $H_{N(t)}$. We often write $t_0 < t_1 < t_2 < \ldots$ for the time points at which new hypotheses are formulated. In that case we abbreviate $N_i := N(t_i) = N(t_0) + i$, so that H_{N_i} is the hypothesis formulated at t_i and $\Theta_{N_i+1} = \{H_i\}_{i\leq N_i}$ is the hypothesis set right after the formulation of H_{N_i}. Note, again, that we do not make any assumptions on the origin of the new hypotheses; all we suppose is that the inquiry prompts some (plausibly data-dependent!) stream of incoming hypotheses. We will say more about this in our analysis in sect. 2.4.

Full probability functions from marginal over Θ_N

Choose some distribution over Θ_N for an agent's marginal probability function over the formulated hypotheses. Since hypotheses are likelihood functions, we can define the agent's marginal likelihood function over the outcomes, *conditional* on hypothesis H_i, by

$$P(E \mid H_i) := H_i(E).$$

Then by the law of total probability we obtain the unconditional marginal likelihood over the outcomes by

$$P(E) = \sum_{i<N} P(H_i) \cdot H_i(E). \tag{2.1}$$

Thus stipulating the marginal over Θ_N defines a probability function P over all of $\Omega \times \Theta_N$.[2]

2.2.2 The standard Bayesian

A Bayesian agent starts with a set Θ_N of N hypotheses, and a probability function P_0, or *prior*, over Θ_N and hence over $\Omega \times \Theta_N$.[3] When the agent receives a new outcome E_t at time $t > 0$, it must update its probability function P_{t-1} to a new probability function or *posterior* P_t.

The orthodox Bayesian way of updating on the evidence is by use of *Bayes's rule*,

$$P_t(\cdot) := P_0(\cdot \mid E^t),$$

with E^t the outcome sequence up to time t. In particular, for the agent's *predictive probabilities*, or its marginal probability function over finite-length future outcomes,

$$P_t(E^s) = P_0(E^s \mid E^t) = \frac{P_0(E^t E^s)}{P_0(E^t)}.$$

Equivalently but more in line with the procedure in sect. 2.2.1, the agent first updates the marginal posterior over the hypotheses, again by Bayes's rule and by Bayes's theorem:

$$P_t(H_i) := P_0(H_i \mid E^t) = \frac{P_0(H_i) \cdot H_i(E^t)}{P_0(E^t)}. \tag{2.2}$$

Then, by the law of total probability on the conditional marginal likelihood,

$$\begin{aligned} P_t(E^s) = P_0(E^s \mid E^t) &= \sum_{i<N} P_0(H_i \mid E^t) \cdot H_i(E^s \mid E^t) \\ &= \sum_{i<N} P_t(H_i) \cdot H_i(E^s \mid E^t). \end{aligned}$$

In summary, the **standard Bayesian** proceeds as follows.

($t = 0$) N **hypotheses**

At the start each explicitly formulated hypothesis H_i in Θ_N receives a prior $P_0(H_i) > 0$, such that $\sum_{i<N} P_0(H_i) = 1$.

[2] Our account of hypotheses is a slightly simplified version of Wenmackers and Romeijn's. They take as hypotheses *sets* of probability functions, so that there is a difference between the "theoretical context" $T_N = \{H_i\}_{i<N}$, the set of hypotheses, and $\Theta_N = \cup_{i<N} H_i$, the set of all probability functions that constitute the hypotheses. Furthermore, an hypothesis's likelihood is then only settled with the aid of a subprior over the hypothesis's elements. While this additional complexity arguably does more justice to the actual shape of hypotheses in scientific or statistical inference, nothing in the following should hinge on the simpler formulation we have chosen to adopt. (Also note that Wenmackers and Romeijn's running example of the food inspection only figures "elementary" hypotheses that are singleton sets, i.e., single probability functions as in our framework.) That said, a natural further development of the current work would allow for representing 'hypotheses' as models in the form of continuous distributions over parametric hypothesis spaces, so at to be able to explicitly analyze, for instance, adding (continuously many) new parameters to an already included model.

[3] We always assume that the prior for given hypothesis set Θ_N is *regular*, meaning that it assigns nonzero probability to each element in Θ_N.

($t > 0$) **Evidence** E^t

Updating on evidence at a later point in time proceeds by

$$P_t(H_i) := P_0(H_i \mid E^t) = \frac{P_0(H_i) \cdot H_i(E^t)}{P_0(E^t)}.$$

($t > 0$) **New hypothesis** H_N

An hypothesis formulated at a later point in time is not an element of the set Θ_N of hypotheses. This hypothesis's prior and posterior probability is and will always remain 0.

2.2.3 The vocal open-minded Bayesian

Wenmackers and Romeijn's proposal of an open-minded Bayesianism starts with postulating, alongside the set Θ_N of explicitly formulated hypotheses, a *catch-all hypothesis* (2016, an idea presented in but preceding Shimony, 1970, p. 95; e.g., Savage in a discussion edited by Barnard and Cox, 1962, p. 70). This catch-all hypothesis $\overline{\Theta}_N$ comprises all (yet) unformulated hypotheses; Wenmackers and Romeijn explicitly define it as the complement of Θ_N within the class of all possible hypotheses.

Their *vocal* variant of open-minded Bayesianism (Wenmackers and Romeijn, 2016, 1234f, 1238ff) derives its name from the fact that the catch-all hypothesis comes with a symbolic prior and likelihood function that figures in all calculations. This in contrast to the *silent* version (sect. 2.2.4 below), where no such prior or likelihood is formulated.

Specification

Thus the vocal open-minded Bayesian starts with an hypothesis set Θ_N of N explicitly formulated hypotheses, and in addition a catch-all hypothesis $\overline{\Theta}_N$. Each explicit hypothesis is assigned a numerical prior probability, summing to 1; and in addition the catch-all hypothesis is assigned an "indefinite" or "merely symbolic" prior τ_N. The numerical probability assigned to an $H \in \Theta_N$ specifies the prior probability value $P_0(H \mid \Theta_N)$, *conditional* on the hypothesis set; the *unconditional* or absolute prior is given by the normalization $P_0(H) := (1 - \tau_N) \cdot P_0(H \mid \Theta_N)$, which is also indefinite because it involves τ_N. While the catch-all thus receives an explicit yet indefinite prior value $P_0(\overline{\Theta}_N) = \tau_N$, the prior probability values $P_0(H')$ of the (yet) unformulated hypotheses $H' \in \overline{\Theta}_N$ are left fully unspecified.

In addition to the indefinite prior, the catch-all comes with a symbolic likelihood function $x_N(\cdot) := P_0(\cdot \mid \overline{\Theta}_N)$. Thus the unconditional marginal likelihood function, analogous to (2.1)

but now not even conditional on Θ_N, is given by the indefinite term

$$\begin{aligned} P_0(E) &= \sum_{i<N} P_0(H_i) \cdot H_i(E) + \tau_N \cdot x_N(E) \\ &= (1-\tau_N) \sum_{i<N} P_0(H_i \mid \Theta_N) \cdot H_i(E) + \tau_N \cdot x_N(E). \end{aligned}$$

The calculation of an explicit hypothesis's posterior on receiving evidence E proceeds by Bayes's rule and theorem in accordance with (2.2), but now also results in an indefinite term because it involves $P_0(E)$.

Finally and crucially, at each point in time the open-minded Bayesian may receive a *newly formulated hypothesis*. This new hypothesis, in terminology due to Earman (1992, p. 196), is *shaved off* from the catch-all. Formally, the vocal agent extends its current hypothesis set Θ_N to the new set $\Theta_{N+1} = \Theta_N \cup \{H_N\}$ to include the newly formulated hypothesis H_N, leaving a cleanly shaven catch-all $\overline{\Theta}_{N+1} = \overline{\Theta}_N \setminus \{H_N\}$. To specify the new hypothesis's prior $P_0(H_N)$ the agent then chooses a prior probability value p that it takes from the prior τ_N, leaving the indefinite remainder $\tau_{N+1} := \tau_N - p$ for the new catch-all $\overline{\Theta}_{N+1}$. Writing $x_{N+1}(\cdot) = P_0(\cdot \mid \overline{\Theta}_{N+1})$ for the new catch-all's indefinite likelihood function, expressions for the marginal likelihoods and posteriors that explicitly contain H_N can be calculated as above.

In summary, the **vocal open-minded Bayesian** proceeds as follows.

($t = 0$) N **explicit hypotheses**

Each explicit hypothesis H_i in Θ_N receives a prior $P_0(H_i \mid \Theta_N) > 0$ conditional on Θ_N, such that $\sum_{i<N} P_0(H_i \mid \Theta_N) = 1$. Moreover, the catch-all hypothesis $\overline{\Theta}_N = \Theta \setminus \Theta_N$ receives an indefinite unconditional prior $P_0(\overline{\Theta}_N) := \tau_N$, and the unconditional priors of the explicit hypothesis are given by $P_0(H_i) := (1-\tau_N) \cdot P_0(H_i \mid \Theta_N)$.

($t > 0$) **Evidence** E^t

Updating proceeds in the standard fashion, although involving an indefinite prior and likelihood of the catch-all:

$$P_t(H_i) := P_0(H_i \mid E^t) = \frac{P_0(H_i) \cdot H_i(E^t)}{\sum_{j=0}^{N-1} P_0(H_j) \cdot H_j(E^t) + \tau_N \cdot x_N(E^t)}.$$

($t > 0$) **New hypothesis** H_N

When a new explicit hypothesis H_N is formulated, extending the hypothesis set to $\Theta_{N+1} = \Theta_N \cup \{H_N\}$, the prior τ_N of the earlier catch-all is decomposed into a value $p < \tau_N$ for the prior $P_0(H_N)$ of the new hypothesis and a remainder $\tau_{N+1} = \tau_N - p$ for the prior $P_0(\overline{\Theta}_{N+1})$ of the new catch-all.

Discussion

The obvious drawback of this proposal is the introduction of purely symbolic terms for the priors and likelihoods of the catch-alls. Apart from the pain of doing actual calculations with these terms, it is quite unclear how to understand them.

Wenmackers and Romeijn variously refer to these terms as "unknown," "undefined," "indefinite," or "unspecified." But even if we grant that these terms can be considered unknown to the agent (leaving aside worries about the notion, not just of an unknown probability, but of an unknown *epistemic* probability), it seems to us that there is a difference between terms that are unknown yet *definite*, and terms that are not. Only in the first case is there an actual matter to the fact of, say, $\tau_N < c$ for a numerical constant c. Thus it is only in the first case that it is clear that the shaving off from the catch-all actually imposes a *limitation* on how much prior the agent can still assign to a newly formulated hypothesis.[4] In contrast, it is less clear whether an *indefinite* probability value allows for shaving off *any desired* definite prior. This might not be a problem to Wenmackers and Romeijn; indeed this would fit their suggestion that the unconditional probability of the catch-all's complement is always *infinitesimally* small (ibid., 1248). However, for our purposes it will prove to be important to impose such constraints on the agent, which is why we will not further pursue the idea of indefinite or infinitesimal priors.

2.2.4 The silent open-minded Bayesian

The motivation for the *silent* version of open-minded Bayesianism (Wenmackers and Romeijn, 2016, 1234f, 1241f) is to evade the difficulties surrounding a symbolic assignment of prior and likelihood to the catch-all. This is achieved by doing away with this assignment altogether, namely, by always only considering *conditional* probability evaluations, conditional on the current hypothesis set. The corresponding Bayesian agent is simply silent about the *absolute* probability values.

Specification

The silent open-minded Bayesian starts out, as before, with an hypothesis set Θ_N of explicitly formulated hypotheses, assigning each $H \in \Theta_N$ a conditional probability value $P_0(H \mid \Theta_N)$. As opposed to the vocal Bayesian, there is no bookkeeping of the catch-all or the unconditional prior P_0.

Since all probability terms are conditional on the current hypothesis set, updating on evidence proceeds fully conditional on Θ_N. That is, the term $P_t(H_i \mid \Theta_N)$ is evaluated via the usual Bayesian updating (2.2), conditional on Θ_N.

If a new hypothesis H_N is formulated, the silent open-minded Bayesian again extends its current hypothesis set Θ_N to the new set $\Theta_{N+1} = \Theta_N \cup \{H_N\}$ to include the newly formulated hypothesis H_N. It then assigns the new hypothesis conditional on the new hypothesis set a

[4] For instance, Wenmackers and Romeijn (2016, p. 1240) mention the possibility of assigning a uniform prior to a new hypothesis. If τ_N has an (unknown yet) definite value, then that would only be possible if this value is in fact greater than $\frac{1}{N+1}$.

posterior value of choice, i.e., a value for $P_t(H_N \mid \Theta_{N+1})$. The new posterior values of the earlier hypotheses are calculated by renormalization, thus preserving the probability ratios.

In summary, the **silent open-minded Bayesian** proceeds as follows.

$(t = 0)$ N **explicit hypotheses**

Each explicit hypothesis in Θ_N receives a prior $P_0(H_i \mid \Theta_N)$ conditional on the initial hypothesis set.

$(t > 0)$ **Evidence** E^t

Updating proceeds in the usual way, conditional on the current context Θ_N:

$$P_t(H_i \mid \Theta_N) := P_0(H_i \mid E^t, \Theta_N) = \frac{P_0(H_i \mid \Theta_N) \cdot H_i(E^t)}{P_0(E^t \mid \Theta_N)}.$$

$(t > 0)$ **New hypothesis** H_N

When a new hypothesis H_N is formulated, extending the hypothesis set to $\Theta_{N+1} = \Theta_N \cup \{H_N\}$, the posterior $P_t(H_N \mid \Theta_{N+1})$ is set to a value $p \in (0,1)$, and the posteriors of the remaining explicit hypotheses conditional on the new hypothesis set are renormalized by

$$P_t(H_i \mid \Theta_{N+1}) := (1 - p) \cdot P_t(H_i \mid \Theta_N).$$

Discussion

In the silent version Wenmackers and Romeijn do away with the explicit monitoring of the catch-all hypothesis by simply always "hiding behind the conditionalization stroke" (2016, p. 1243). As they themselves point out, one might feel uneasy about thus still leaving unspecified the agent's unconditional, *absolute* convictions. One might indeed feel that this threatens to sufficiently compromise *coherence* that this is no *Bayesian* account anymore (cf. Glymour, 2016, p. 1282). What is certainly lost, in moving to larger models, is the guarantee of *dynamic* coherence (see sect. 2.4.1 below for more details).

However, it is surely more in line with statistical practice that probabilities are always evaluated under the tentative assumption of a particular model, without any pledge to the truth of this model. The discussion by Sprenger (2020) (also see Sprenger and Hartmann, 2019, ch. 12, Vassend, 2019) is a recent example of several earlier expressions of this view in the Bayesian literature (e.g., Lindley, 2000, p. 334; 1982), that tends to go together with a commitment to coherence only for as long as the model does not change (see indeed Shimony, 1970, 103f). Perhaps most outspoken in this latter respect is Howson's account of Bayesianism, "a theory of valid inductive inference from pre-test to post-test distributions," that offers the worry of an "inconsistent assignment over time" a simple reply: "so what?" (1988, p. 81).

Moreover, Wenmackers and Romeijn stay far from the latter extreme: both versions of their open-minded Bayesian are "conservative extensions" where the probabilities conditional on an expanded model cohere with those conditional on the original model (2016, 1235f). Bayes's rule amounts to restricting the subalgebra on the outcome space (to the subtree of the outcome space that extends the evidence) while preserving all probability ratios within; the rule for incorporating new hypotheses *enlarges* the subalgebra on the *hypothesis space* (to the larger hypothesis set) while likewise preserving all probability ratios within the original (ibid.).

We conclude that the silent version holds a conceptual advantage over the vocal version. The main *formal* difference, for our purposes, is that in the vocal version, a new hypothesis is assigned a certain prior value that is constrained by the catch-all's prior; whereas in the silent version, a new hypothesis is assigned a *posterior* value, the choice of which is *unconstrained*.

Wenmackers and Romeijn indeed worry that "[t]he silent proposal allows too much freedom in the assignment of a posterior to the new hypothesis—so much freedom, that it is not clear that the old evidence has any impact" (ibid., 1245). This prompts them to propose a *hybrid* variant of the vocal and the silent versions (sect. 2.2.6 below). Before we turn to this version, we will take a quick look at a more direct tweak of the silent version that replaces the choice of posterior by the choice of prior, so that the calculation of the former requires some "reconstructive work" that does take old evidence into account (ibid., 1242).

2.2.5 The silent open-minded Bayesian: retroactive variant

Thus the alternative variant of the silent version is where we 'retroactively' assign a *prior* to a new hypothesis, i.e., a value p_0 to $P_0(H_N \mid \Theta_{N+1})$. After renormalizing the priors of the other hypotheses,

$$P_0(H_i \mid \Theta_{N+1}) := (1 - p_0) \cdot P_0(H_i \mid \Theta_N) \tag{2.3}$$

for all $H \in \Theta_N$, we can with the help of Bayes's rule (using the the new likelihood $H_N(E^t)$), calculate $P_t(H_N \mid \Theta_{N+1})$ from there.

Formally, however, it does not make a difference whether we choose a prior and then calculate the posterior, or the other way around. (Provided, that is, that H_N's likelihood on E^t is positive, or its posterior can only be 0.) For any desired posterior p_t for a new hypothesis, we can uniquely reconstruct a prior p_0 that in combination with the new hypothesis's likelihood, will result at time t in *that* posterior. After all, there are, unlike in the vocal version, no constraints on choosing a prior p_0.

2.2.6 The hybrid open-minded Bayesian

The vocal and the silent version are combined in the *hybrid* version (Wenmackers and Romeijn, 2016, 1245f) as follows. The agent starts out, as in the vocal version, with an explicit yet symbolic assignment to the catch-all hypothesis. During the normal learning process of updating on the evidence, it stays in the "silent phase," in which it evaluates all probabilities conditional on the current hypothesis set. Only when a new hypothesis is formulated does it enter the "vocal phase," in which it, like in the vocal version, retroactively shaves off a prior for the new hypothesis

from the catch-all's prior, after which it, like in the retroactive silent version, recalculates the priors and posteriors (again conditional, but on the *new* hypothesis set) from there.

In summary, the **hybrid open-minded Bayesian** proceeds as follows.

$(t = 0)$ N **explicit hypotheses**

Each explicit hypothesis H_i in Θ_N receives a prior $P_0(H_i \mid \Theta_N) > 0$ conditional on Θ_N, such that $\sum_{i<N} P_0(H_i \mid \Theta_N) = 1$. Moreover, as in the vocal version, the catch-all hypothesis $\overline{\Theta}_N = \Theta \setminus \Theta_N$ receives an unconditional prior $P_0(\overline{\Theta}_N) := \tau_N$, and the unconditional priors of the explicit hypothesis are given by $P_0(H_i) := (1 - \tau_N) \cdot P_0(H_i \mid \Theta_N)$.

$(t > 0)$ **Evidence** E^t

Updating proceeds as in the silent version, conditional on the current context Θ_N:

$$P_t(H_i \mid \Theta_N) := P_0(H_i \mid E^t, \Theta_N) = \frac{P_0(H_i \mid \Theta_N) \cdot H_i(E^t)}{P_0(E^t \mid \Theta_N)}.$$

$(t > 0)$ **New hypothesis** H_N

When a new explicit hypothesis H_N is formulated, extending the hypothesis set to $\Theta_{N+1} = \Theta_N \cup \{H_N\}$, as in the vocal version the unconditional prior τ_N of the earlier catch-all is decomposed into a value $p < \tau_N$ for the unconditional prior $P_0(H_N)$ of the new hypothesis and a remainder $\tau_{N+1} = \tau_N - p$ for the unconditional prior $P_0(\overline{\Theta}_{N+1})$ of the new catch-all. The priors conditional on the new hypothesis set are obtained by renormalization,

$$P_0(H_i \mid \Theta_{N+1}) = \left(1 - \frac{p}{1 - \tau_{N+1}}\right) \cdot P_0(H_i \mid \Theta_N),$$

from which the conditional posteriors are obtained by the usual updating,

$$P_t(H_i \mid \Theta_{N+1}) := P_0(H_i \mid E^t, \Theta_{N+1}) = \frac{P_0(H_i \mid \Theta_{N+1}) \cdot H_i(E^t)}{P_0(E^t \mid \Theta_{N+1})}.$$

Thus the hybrid version combines the conceptually more pleasing conditional reasoning of the silent version with the constraint on new priors introduced by the catch-all in the vocal version. This constraint proves important for our concern in this paper, the guarantee of truth-merging.

2.3 The open-minded Bayesians' truth-convergence

We start by introducing the formal property of convergence to the truth, as satisfied by the standard Bayesian (sect. 2.3.1). After some preliminary remarks about the meaning and the

promise of this property in the open-minded case (sect. 2.3.2), we demonstrate and diagnose its failure for the silent (sect. 2.3.3) and the hybrid (sect. 2.3.4) version.

2.3.1 The standard Bayesian

Suppose the standard, 'closed-minded' Bayesian starts with a hypothesis set that includes the hypothesis H^* that is actually *true*, meaning that the probabilities given by H^* are the true probabilities that govern the generation of the data. In that case, one can prove a strong statement about the agent's convergence to this truth. Namely, one can prove that, H^*-almost surely, the *total variational distance*

$$\sup_{A\in\mathfrak{F}} |P_t(A) - H^*(A \mid E^t)| \tag{2.4}$$

between the agent's probabilities and the H^*-probabilities on future events goes to 0 as $t \to \infty$. That is, *with true probability 1* (as given by H^*), the agent's probabilities conditional on the past will convergence on all events' *true* probabilities. We say that the agent *strongly merges* with the truth.

Definition 1. For probability measures P and Q on $(\Omega, \mathfrak{F})$, we say that P strongly merges with Q if Q-a.s.

$$\sup_{A\in\mathfrak{F}} |P(A \mid E^t) - Q(A \mid E^t)| \xrightarrow{t\to\infty} 0. \tag{2.5}$$

A standard Bayesian's strong merger with the truth follows directly from a fundamental result due to Blackwell and Dubins.

Theorem 1 (Blackwell and Dubins, 1962)**.** *For probability measures P and Q on $(\Omega, \mathfrak{F})$ such that the latter is absolutely continuous with respect to the former, i.e., $Q(A) > 0$ implies $P(A) > 0$ for all events A in the σ-algebra $\mathfrak{F}$ on Ω, it holds that Q-a.s. P strongly merges with Q.*

Namely, if the Bayesian agent's hypothesis set contains H^*, meaning that its regular prior probability $P(H^*) > 0$, then, in terminology due to Kalai and Lehrer (1993, p. 1037), *P holds a grain* of H^*, or *P holds a grain of the truth*. That is to say, there is an $a \in (0,1)$, namely $a = P(H^*)$, such that the marginal prior *P on the outcome space* equals $a \cdot H^* + (1-a) \cdot P'$, for some probability measure P'. More precisely still, from the fact that $P(H^*) > 0$, we have that P *dominates* H^*, meaning that $P(E^t) \geq a \cdot H^*(E^t)$ for all finite outcome sequences E^t, but that implies that also $P(A) \geq a \cdot H^*(A)$ for *all* events $A \in \mathfrak{F}$ generated from the finite sequences. But that means that H^* is absolutely continuous with respect to P.

Corollary 2. *If P holds a grain of the truth H^*, then P strongly merges with H^*.*

Strong merger is indeed a very strong notion, as it includes all tail events A, the occurrence of which cannot be verified in finite time. A more down-to-earth notion of truth-convergence is *weak merger* (Kalai and Lehrer, 1994), that only concerns the special case of the next outcome. This is the notion we will be focusing on in this paper.

Definition 2. For probability measures P and Q on $(\Omega, \mathfrak{F})$, we say that P weakly merges with Q if Q-a.s.

$$\sup_{E_{t+1}\in\{0,1\}} |P_t(E_{t+1}) - H^*(E_{t+1} \mid E^t)| \xrightarrow{t\to\infty} 0. \tag{2.6}$$

In fact, weak merger of two probability measures is equivalent, for every $d \in \mathbb{N}$, to merger where the supremum ranges over all future outcomes of length up to d (ibid.). Nevertheless, as we will explain in more detail in our analysis in sect. 2.4, we will in this paper focus on the case $d = 1$. Moreover, as we will still explain too, despite the fact that this is already a sufficient condition for strong merger, the notion of holding a grain of the truth will be central to our analysis. When in the following we refer to "truth-convergence" without further qualification, we mean weak merger as in definition 2.[5]

2.3.2 The open-minded Bayesians

The question we shall investigate is whether Wenmackers and Romeijn's proposals can retain this conception of convergence to the truth, *whenever the true hypothesis H^* is formulated*. More precisely, the question is whether we can show that, if H^* is indeed formulated at some time t_0, the agent function $P_t(\cdot \mid \Theta_{N(t)})$, as $t > t_0$ goes to infinity, weakly merges with H^*. The question is whether we can show that, after H^* has been formulated,

$$\sup_{E_{t+1}\in\{0,1\}} |P_t(E_{t+1} \mid \Theta_{N(t)}) - H^*(E_{t+1} \mid E^t)| \xrightarrow{t\to\infty} 0 \text{ with } H^*\text{-probability 1.} \tag{2.7}$$

One might already object here that we should rather consider merging of the *unconditional* agent function $P_t(\cdot) = P_t(\cdot \mid \Theta_{N(t)} \cup \overline{\Theta}_{N(t)})$. For an adherent to the vocal variant, the agent's beliefs are constituted by a function over all hypotheses, including those in the catch-all, and so, from this perspective, an agent's truth-merging should be taken to mean merging of *that* function. However, we already argued in favour of the conditional perspective of the silent or hybrid version; and the question of convergence of a measure that is partly unspecified introduces problems of interpretation that look unsurmountable.

This is not to say that the truth-merging of $P_t(\cdot \mid \Theta_{N(t)})$ is unproblematic in its interpretation. Indeed, we will below be much concerned with meeting two challenges in squaring the semi-formal expression (2.7) with our intuitive demand of truth-convergence. *Semi*-formal, because

[5] There exist other notions of truth-convergence one could consider. Note, first of all, that the presupposition of a true *statistical hypothesis* can be distinguished from what is perhaps the more usual setting in philosophy, where truth-values are attached to events or elements of the outcome space (Gaifman and Snir, 1982, Earman, 1992). Note, further, that the notion of merging is concerned with learning the probabilities *of future outcomes*. This can be distinguished from learning *the correct hypothesis* ('learning the parameter' in a statistical model), which would correspond to the agent's posterior concentrating on the correct element in the hypothesis set. One reason why we do not consider this notion here is that such posterior-concentration is rather trivially impossible unless we exclude the possibility of different hypotheses that nevertheless from some point on are 'empirically equivalent' in that they give the same predictive probabilities (cf. Lehrer and Smorodinsky, 1996, 148f). Finally, there are still less powerful notions of truth-merging, including *almost weak merging*. See Lehrer and Smorodinsky (1996), Leike (2016, ch. 3) for overviews of learning notions and necessary and sufficient conditions.

we are not yet clear, first of all, about the exact nature of the probability-1 qualification. Second, we are not yet fully clear, certainly not until the first is resolved, about the exact nature of the agent measure that we seek merging for.

Nevertheless, the intuitive demand that (2.7) is supposed to capture is already sufficiently precise to isolate a straightforward case in which truth-convergence *is* guaranteed (sect. 2.3.2). This will then also already point us to the general case that might be problematic (sect. 2.3.2). In fact, this is already enough to show that this case *is* problematic: all the variants of open-minded Bayesianism are *not* in general guaranteed to preserve truth-convergence (sects. 2.3.3–2.3.4). Only in the discussion leading up to our diagnosis of this failure and our proposal of a *forward-looking* open-minded Bayesian, in sect. 2.4, will we finally face the aforementioned challenges head-on.

Finitely many new hypotheses

The answer to our question is a clear *yes* if we can be sure that, after H^* is formulated, *no further new hypotheses will ever be formulated*. For each of the different versions of open-minded Bayesianism, the agent with function $P_t(\cdot \mid \Theta_{N(t)})$ after formulation of H^* can then be treated as a standard Bayesian that starts its investigation at t with a fixed hypothesis set $\Theta_{N(t)}$. Thus, as $H^* \in \Theta_{N(t)}$, the agent then holds a grain of the truth and we can simply apply corollary 2 to $P_t(\cdot \mid \Theta_{N(t)})$ to indeed obtain not just weak but strong merger with the truth from there.

This observation easily extends to the more general case where we can be sure that after some finite point in time there will no longer be new hypotheses formulated. So suppose H^* is formulated at $t_0 \leq t$, say in response to data E^{t_0}. Then, to put it graphically, from each of the possible nodes E^t in the outcome tree extending E^{t_0}, we can run corollary 2 on the fixed agent function to obtain, with probability 1, truth-merger from *there*; but that means we already have the guarantee of truth-merger from *here*, at E^{t_0}. Hence, under the assumption that no more hypotheses are formulated after some finite time t, we have strong merger whenever the truth H^* is formulated. This assumption can be reformulated as saying that, on any infinite outcome stream, only a finite number of new hypotheses will ever be formulated.

Fact 1. *All open-minded Bayesians are guaranteed to strongly merge with the truth whenever the truth is formulated, if there is a finite bound on the number of new hypotheses that will be formulated.*

Infinitely many new hypotheses

The previous assumption, in entailing that from some point on the open-minded Bayesian reduces to a standard, fixed-minded, Bayesian, thereby also neutralizes a good part of the distinctive interest of the former. It is, more importantly, an assumption that we do not generally want to make: we certainly do not want to assume that, when the true hypothesis is formulated, who or whatever is responsible for designing new hypotheses *knows* that it can stop now.

On the other hand, it also sounds unrealistic that in an actual scientific inquiry, certainly after the true hypothesis has already been found, one would mindlessly keep incorporating newly arriving hypotheses indefinitely. One would presumably only look out for new hypotheses if

the currently available ones do not seem to do: if there is some misfit between the data and the current hypotheses. Incorporating this element, possibly in the shape of a formal model verification procedure, would still not render the scenario of an unending stream of false hypotheses insignificant: there is now a tension to be resolved between risking sticking to suboptimal hypotheses and risking incorporating false ones.

Important as this element is, it is beyond the scope of the current paper. We are here first concerned with the consistency requirement of truth-convergence in the most general case where the agent might forever keep receiving new (and false) hypotheses, which it has to incorporate irrespective of the past outcomes and current hypothesis set.

This general case is potentially problematic because if the agent keeps having to distribute some of its posterior to these new and false hypotheses (and so keeps having to incorporate these in its predictions), this could get in the way of its converging on the true hypothesis's *true* predictive probabilities. In fact, this *is* problematic, for all the versions of open-minded Bayesianism. We now first look at the silent variants (sect. 2.3.3), where this shows very directly; and then at the more interesting hybrid variant (sect. 2.3.4).

2.3.3 The silent open-minded Bayesian

This version is the least constrained of the open-minded Bayesianisms, which makes it most straightforwardly fail to guarantee truth-convergence. We first show this for the standard open-minded version of sect. 2.2.4, and then for the retroactive variant of sect. 2.2.5.

The silent open-minded version: original variant

The reason for the failure of truth-convergence is that we cannot exclude infinite streams of false hypotheses that keep occupying a specific share of the posterior probability and in this way keep distorting the predictive probabilities.

Fact 2. *The original variant of the silent open-minded Bayesian is not guaranteed to weakly merge with the truth whenever the truth is formulated.*

Example 2.1. Consider the scenario where the data is generated by some Bernoulli distribution H_{θ^*}. Suppose for concreteness that $\theta^* = 9/10$, and that this correct hypothesis $H^* = H_{\theta^*}$ is indeed formulated at some stage t_0. Now consider the possibility that infinitely often (i.e., for each stage $t' > t_0$ there is a still later stage $t > t'$ at which) a new hypothesis $H_{N(t)}$ is formulated that issues a predictive probability $H_{N(t)}(1 \mid E^t) = 0$. Since there are no restrictions on the posterior which the silent open-minded Bayesian can assign to these newly formulated hypotheses, it can choose to keep assigning a value $P_t(H_{N(t)} \mid \Theta_{N(t)+1}) \geq 1/10 + \varepsilon$ for positive

ε. In that case there will be infinitely many stages t at which the predictive probability

$$\begin{aligned} P_t(0 \mid \Theta_{N(t)+1}) &= \sum_{H \in \Theta_{N(t)+1}} P_t(H \mid \Theta_{N(t)+1}) \cdot H(0 \mid E^t) \\ &> \left(\frac{1}{10} + \varepsilon\right) \cdot H_{N(t)}(0 \mid E^t) \\ &= \frac{1}{10} + \varepsilon, \end{aligned}$$

blocking convergence to the correct predictive probability $H^*(0 \mid \cdot) = 1/10$. ◇

This example can be adapted at will to show that for any true H^* there are hypothesis streams and posterior assignments that block convergence. The essential trait is that the newly formulated hypotheses receive—keep receiving—too much posterior. This leads us to an obvious diagnosis: the silent open-minded Bayesian is allowed too much freedom in assigning posteriors to newly formulated hypotheses.

The silent open-minded version: retroactive variant

Following up on the previous diagnosis, one way in which it might *seem* we can constrain the freedom of the open-minded Bayesian is to insist that the posterior must be informed by the old evidence. This is the retroactive variant of the silent open-minded Bayesian, sect. 2.2.5 above; but as we explained there already, there is, barring the case where the new hypothesis's likelihood is 0, actually no formal difference between the two versions. That is, any choice of posterior can be modeled as a retroactive choice of prior. This means that any counterexample to the silent open-minded version also yields a counterexample to the retroactive variant, including the previous example 2.1.

Fact 3. *The retroactive variant of the silent open-minded Bayesian is not guaranteed to weakly merge with the truth whenever the truth is formulated.*

Example 2.2. Recall from the reconstruction of p_0 from p_t in sect. 2.2.5 that the exact calculations now do depend on the likelihoods of all hypotheses on the past data, something that was not specified in example 2.1. The most straightforward circumstance is where the new hypothesis's likelihood on E^t actually *equals* the probability of E^t on Θ_N,

$$H_N(E^t) = P_0(E^t \mid \Theta_N), \tag{2.8}$$

in which case a prior assignment $P_0(H_N \mid \Theta_{N+1}) := p$ translates into a posterior $P_t(H_N \mid \Theta_{N+1}) = p$. In that case, a prior choice of $p \geq 1/10 + \varepsilon$ recovers the previous example. If the new hypothesis's likelihood on the past data is *lower* than $P_0(E^t \mid \Theta_N)$, the prior must be set higher to retrieve the same posterior. As an illustration, if $H_N(E^t) = 1/3 \cdot P_0(E^t \mid \Theta_N)$, then a posterior $p_t > 1/10$ requires a choice of prior $p_0 > 1/4$.

Arguably, however, the more plausible circumstance is for newly proposed hypotheses to have *higher* likelihood than the earlier hypotheses. Plausibly, new hypotheses (formulated after we have already *seen* the past data) rather *overfit* the data: in the most extreme case, actually have a likelihood 1. In that case, of course, the same posterior p_t requires a smaller prior p_0. To

illustrate again, suppose indeed $H_N(E^t) = 1$; then in general to obtain posterior p_t we need to set

$$p_0 = \frac{P_0(E^t \mid \Theta_N)}{P_0(E^t \mid \Theta_N) + \frac{1}{p_t} - 1}. \tag{2.9}$$

But if the data is actually generated by H_{θ^*} with $\theta^* = 0.9$, then $P_0(E^t \mid \Theta_N)$, with high probability, will not exceed H_{θ^*}'s likelihood on the past data E^t, which for typical data is about $0.9^{0.9t} \cdot 0.1^{0.1t}$. In that case, the same posterior only requires an exponentially smaller prior: already for $t = 10$, for instance, it suffices for posterior $p_t > 1/10$ to set $p_0 > 1/200$. ◇

The arguably most natural circumstance of new hypotheses that overfit is thus also the most difficult case for our purposes. An extremely modest choice of prior here already suffices for a substantial posterior, and the threat to truth-convergence is precisely such substantial posterior assignments to new and false hypotheses.

One can defend the retroactive approach on the grounds that it accommodates how old evidence confirms new theories (Wenmackers and Romeijn, 2016, 1244f); or one can disown it on the grounds that it involves a "double counting" of the old evidence, since the hypothesis and presumably its prior was already formulated in response to the evidence (cf. Earman, 1992, 132f). We point out here that for the above reason of overfitting hypotheses, a retroactive procedure appears more challenging for the aim of truth-convergence. Of course, in the silent version, this cannot make an *essential* difference: both variants are formally equivalent, and the challenge above is limited to a moderate choice of prior in the retroactive variant that does not correspond to a moderate choice of posterior in the original variant. But our analysis below reveals that in the hybrid case, the difference between prior and posterior assignments will be crucial for the guarantee of truth-convergence.

2.3.4 The hybrid open-minded Bayesian

The diagnosis from the previous section was clear: the (retroactive) silent open-minded Bayesian is allowed too much freedom in assigning posteriors (priors) to newly formulated hypotheses. Given this diagnosis, one might expect the hybrid version to do better. After all, here there is an explicit constraint on priors: there is only so much the agent can shave off from the catch-all!

Again, this is only so because we interpret the catch-all's prior as at least having some determinate value. This does not quite exclude that this is "a number extremely close to unity," but it does exclude a conception where it is some indeterminate value arbitrarily close to 1, perhaps made precise as "unity minus an infinitesimal" (Wenmackers and Romeijn, 2016, p. 1244). Perhaps the latter is the more natural conception. When it comes to truth-convergence, however, this renders the hybrid version on a par with the silent version: both put no constraints on the choice of prior (posterior), wherefore convergence cannot be guaranteed.[6]

[6] Wenmackers and Romeijn (ibid.) evoke Earman's worry that the procedure of shaving-off from the catch-all "leads to the assignment of ever smaller initial probabilities to successive waves of new theories until a point is reached where the new theory has such a low initial probability as to stand not much of a fighting chance" (1992, p. 196). On our analysis, the danger is rather that new theories keep amassing *too much* probability.

We will for this reason proceed with supposing that the hybrid version is characterized by putting definite constraints on the choices of priors. Specifically, we imagine that there is a certain limited reservoir of prior probability, from which the probability for new hypotheses must be taken. We can think of this constraint as simply that, a constraint; we are not committed to understanding this constraint in terms of a catch-all. Nevertheless, we see it as a conceptual plus that it *can* be understood in this way, and this carries over to our own proposal in sect. 2.4.

Failure of truth-convergence

Unfortunately, the constraint introduced in the hybrid version does not suffice: we can even produce a scenario where convergence to the true predictive probabilities is *guaranteed to fail*. This scenario again exploits the possibility of a stream of overfitting hypotheses, that despite the constraint on new prior assignments still keep taking up too much posterior. More precisely, on every possible outcome stream we can repeat the following: wait while all current probabilistic hypotheses have lower and lower likelihood on the unfolding sequence of outcomes, until the difference with the maximal likelihood of a new overfitting hypothesis is large enough for such a new hypothesis to have a sufficient impact, despite its necessarily constrained prior, on the agent's predictive probabilities.

Proposition 3. *The hybrid open-minded Bayesian is not guaranteed to weakly merge with the truth whenever the truth is formulated.*

Example 2.3. Suppose that the true hypothesis is the Bernoulli $H^* = H_{\theta^*}$ with $\theta^* = 1/2$, and that this hypothesis is indeed formulated at a point in time t_0. Thus H^* is assigned some unconditional prior value $p^* =: P_0(H^*)$, leaving the catch-all $\overline{\Theta}_{N_0+1}$ with some unconditional prior $\tau_{N_0+1} = \tau_{N_0} - p^*$.

Consider a history with $t_0 < t_1 < t_2 < \ldots$ infinitely many later points in time at which a new hypothesis is formulated. The vocal open-minded Bayesian is restricted by the prior held by the catch-all in how much prior it can shave off and assign to these new hypotheses; but it can choose to assign each H_{N_i} an unconditional prior

$$P_0(H_{N_i}) = 2^{-i} \cdot \tau_{N_0+1}, \tag{2.10}$$

since $\sum_{i=1}^{\infty} 2^{-i} \cdot \tau_{N_0+1} = \tau_{N_0+1}$.

Now consider such a history where the newly proposed hypotheses all maximally overfit the past data at their time of formulation, i.e., $H_{N_i}(E^{t_i}) = 1$ for each i, and then make some biased prediction $H_{N_i}(0 \mid E^{t_i}) = p_i$, with $|p_i - 1/2| > \varepsilon$ for some pre-set $\varepsilon > 0$.

Suppose, further, that all hypotheses formulated before the true hypothesis, and all the new hypotheses after their formulation, issue predictive probabilities that are bounded away from 1: there is some $\delta > 0$ such that all predictive probabilities are smaller than $1 - \delta$ (equivalently, all predictive probabilities are *greater* than δ). The idea is that, whatever the subsequent data, the hypotheses in play will each point in time leak some of their likelihood, so that, when a new overfitting hypotheses H_{N_i} comes in, *after the stretch of time between t_{i-1} and t_i has been large*

enough, its relative likelihood is so large that its biased prediction will sufficiently distort the overall predictive probability.

Specifically, fix some $\varepsilon' < \varepsilon$, and let

$$r = \frac{\frac{1}{2} + \varepsilon'}{\frac{1}{2} + \varepsilon}, \tag{2.11}$$

which itself lies in the interval $\left(\frac{1}{2}, 1\right)$. Now if at each t_i we have

$$P_{t_i}(H_{N_i} \mid \Theta_{N_i+1}) > r, \tag{2.12}$$

then we have for E with $H_{N_i}(E \mid E^{t_i}) > \frac{1}{2} + \varepsilon$ that

$$\begin{aligned} P_{t_i}(E \mid \Theta_{N_i+1}) &= \sum_{H \in \Theta_{N_i+1}} P_{t_i}(H \mid \Theta_{N_i+1}) \cdot H(E \mid E^{t_i}) \\ &> P_{t_i}(H_{N_i} \mid \Theta_{N_i+1}) \cdot H_{N_i}(E \mid E^{t_i}) \\ &> \frac{\frac{1}{2} + \varepsilon'}{\frac{1}{2} + \varepsilon} \cdot \left(\frac{1}{2} + \varepsilon\right) \\ &= \frac{1}{2} + \varepsilon', \end{aligned}$$

blocking convergence.

As worked out in appendix 2.A.2, inequality (2.12) is guaranteed if each

$$t_i - t_{i-1} > \frac{-\log(1-r) - (-\log r) + i - \log \tau_{N_0+1}}{-\log(1-\delta)}. \tag{2.13}$$

To break (2.13) down a little, note that if ε is reasonably large, and ε' chosen very small, then r is relatively close to 1/2 and has a minor influence on the bound. For instance, if $r < 2/3$, which would follow from $\varepsilon > 1/4$ and $\varepsilon' \approx 0$, then $-\log(1-r) - (-\log r) < 1$, so that (2.14) is already implied by

$$t_i - t_{i-1} > \frac{1 + i - \log \tau_{N_0+1}}{-\log(1-\delta)}. \tag{2.14}$$

Furthermore, we have $\delta = 1/2$ and (2.14) reduces to

$$t_i - t_{i-1} > 1 + i - \log \tau_{N_0+1} \tag{2.15}$$

in the extreme case where all hypotheses except H_{N_i} after t_{i-1} always give predictive probabilities $(1/2, 1/2)$. ◇

Discussion

The failure of truth-convergence of the hybrid open-minded agent may strike one as surprising. It is, after all, characteristic of the hybrid procedure that the true hypothesis, once formulated, holds an explicitly assigned share $p^* > 0$ of the absolute prior. As soon as the true hypothesis

is formulated, the unconditional agent function P_0 holds a grain p^* of this truth, *no matter what hypotheses with what priors are still added later*. This carries over to the retroactive prior measures conditional on any hypothesis set after the truth is formulated: $P_0(H^* \mid \Theta_N) \geq p^*$ for all hypothesis sets Θ_N after the formulation of H^*. But does this not suggest that the agent function holds a grain of the truth, and was this not already enough for strong truth-merger?

A complete answer to what is wrong with this intuition requires us to make perfectly precise the desideratum of an open-minded agent's truth-convergence. We will here first briefly make the above intuition precise in a particular way, a way that is clearly *faulty*, but that allows us to highlight the challenges we face in formalizing our desideratum of an open-minded agent's truth-convergence. In the next section we proceed to meet these challenges and formalize our desideratum, to subsequently propose a version of an open-minded Bayesian that does satisfy a version of truth-convergence.

Thus let us for a moment consider the measure $P_0(\cdot \mid \Theta_\infty)$, induced by the actually generated hypotheses and prior assignments *in the limit*. This measure must also hold a grain p^* of the truth. What, exactly, is unsatisfying about proclaiming truth-convergence of the open-minded agent, from the fact that we can always derive, with corollary 2, strong truth-merger of *this* measure?

The straightforward answer is that this formal almost-sure strong merger must be unsatisfying because, as we already know from example 2.3, it can go together with a *guaranteed failure of weak merger*. But how can this be? Here it is important to note that, in example 2.3, the hypothesis stream emphatically depends on the actually generated data stream. While the agent function $P_0(\cdot \mid \Theta_\infty)$ induced by this particular data and hence hypotheses stream can be shown to a.s. merge with H^* (as it contains a grain of H^*), this is still consistent with it *failing to merge on the actual data stream that induced it*. (The latter is consistent with truth-merger, because, in our example, any particular outcome stream that is actually generated is an H^*-probability-0 event.)

This provides an illustration of the two challenges we already identified in sect. 2.3.2. First, since we have an hypothesis stream as a moving part, we have to be very careful with the interpretation of probability-1 statements on the data space. The agent function $P_0(\cdot \mid \Theta_\infty)$ was only put in place, so to speak, after already fixing the actually generated data stream, and the a.s. merger only derived after that. In contrast, intuitively, the 'almost sure' should range over the possible data *and all that depends on it*, including the possible hypotheses (hence possible shapes of the agent function) that are formulated in response to it. The challenge is to attain a formal a.s. merger that is also still meaningful in this sense. This is intertwined with the second challenge, which is to make precise which agent function we actually seek merger for. The obvious diagnosis is that the functions $P_0(\cdot \mid \Theta_\infty)$, having this "after the fact" quality of being dependent on a particular data and hence hypothesis stream, and indeed of then having available this hypothesis set *from the start*, are not what we are after.

We now proceed to look for an answer to these two challenges, towards reclaiming a property of truth-convergence.

2.4 The forward-looking Bayesians and their truth-convergence

We further analyze the goal of truth-convergence, introducing the assumption of a scheme for hypothesis and posterior generation and the notion of a completed agent measure (sect. 2.4.1). We then propose a *forward-looking* open-minded Bayesian, the competed agent measure of which *does* retain a grain of the truth, from which weak merger follows. We first propose a proto-variant of this version, which is a variant of the silent open-minded Bayesian with a limited posterior reservoir (sect. 2.4.2), before we introduce the final version, that is a variant of the hybrid open-minded Bayesian with a restriction on new hypotheses' likelihoods (sect. 2.4.3).

2.4.1 Towards regaining truth-convergence

Fixing the hypothesis scheme

We start with the first challenge in drawing up the desired convergence statement: how should we think about the 'almost surely'? In the following, we suppose for simplicity of presentation that the agent possesses the true hypothesis H^* from the start, $H^* \in \Theta_0$.[7]

We first observe that it is impossible to derive a statement of the following form.

(i) For every H^*, there is an H^*-measure-1 class of infinite output streams on which the open-minded agent converges to H^*, independent of the stream of newly formulated hypotheses.

Already in the case of the standard Bayesian agent, the H^*-measure-1 class of output streams on which the agent converges cannot generally be independent of the other elements in the agent's hypothesis class. Consider for the true H^* again the Bernoulli-1/2 measure: it is not hard to see that for each possible infinite outcome stream, there exist hypothesis sets that contain H^* yet are such that the agent does not converge on *this* outcome stream. As an extreme case, the agent will not converge on outcome stream E^ω if the hypothesis set contains an hypothesis that assigns probability 1 to this exact sequence E^ω: the agent will converge, not on the true predictive probabilities 1/2, but on predictive probabilities 1 for the correct next outcomes. This example concerns the initial hypothesis set of a standard (or indeed open-minded) agent, but easily transfers to the streams of newly formulated hypotheses given to any plausible version of an open-minded agent.[8] Thus a statement of form (i) is too strong.

This leads us to the following statement, where we have shifted the quantifiers to allow the exact measure-1 class to depend on the hypothesis stream.

[7] For the general case where the truth is formulated after some finite time t, or more specifically, after some finite sequence E^t, mentions of 'an H^*-measure-1 class of infinite outcome streams' should be replaced by 'an $H^*(\cdot \mid E^t)$-measure-1 class of infinite outcome streams extending E^t,' and the 'stream (scheme) of newly formulated hypotheses' by the 'stream (scheme) of newly formulated hypotheses after E^t.'

[8] We only need to assume that the agent's posteriors will indeed converge on the predictions of hypotheses that perform *perfectly*, which is a minimal condition for a version that will in fact have the capacity to converge to the truth.

(ii) For every H^*, every hypothesis stream, there is an H^*-measure-1 class of infinite outcome streams on which the open-minded agent converges to H^*.

In order to demonstrate a statement of the form (ii), we must prove, for any given hypothesis stream, a.s. convergence on the presupposition of *this* stream. Formally, we conceive of $\Theta_{N(\cdot)}$ as a function that maps each time t to an hypothesis set Θ. Of course, this function must also return hypothesis sets that actually correspond to some possible open-minded agent. For instance, for each t there can be at most one hypothesis in $\Theta_{N(t+1)} \setminus \Theta_{N(t)}$.

There is a clear sense, however, in which a statement of form (ii) is too weak. The main challenge for establishing truth-convergence is, recall example 2.3, the possibility of overfitting hypotheses *in reaction to each possible outcome stream*. In light of such scenarios, presupposing a particular hypothesis stream, irrespective of the generated data, is obviously unsatisfying.

But we can just as well assume that the generation of hypotheses is given by a function that links hypothesis sets, not simply to the possible points in time, but to all *possible finite outcome sequences*. That is, we presuppose some *data-dependent* (what we shall call) *scheme* for generating hypotheses, or simply *hypothesis scheme*, that is a function $\Theta_{(\cdot)}$ that maps each finite data sequence E^t to an hypothesis set Θ_{E^t}. Again, this function must also be constrained by the open-minded agent's specification.

This then leads us to aim for a convergence statement of the following form.

(iii) For every H^*, every hypothesis scheme, there is an H^*-measure-1 class of infinite outcome streams on which the open-minded agent converges to H^*.

Note that the assumption of a particular H^* in conjunction with an hypothesis scheme comes down to treating hypothesis streams as *random* quantities, as they are given by a function on the outcome streams governed by probability measure H^*. One could take this further and consider for the true measure more elaborate probabilistic models that also directly range over the class of possible hypothesis streams. We do not go this way here: we stick here to a true measure H^* that is a function over outcome sequences only, and work towards a convergence statement where the H^* measure-1 class can depend on the hypothesis scheme. Of course, there is more to say about the conceptual status of a convergence statement of the form (iii), and we will say a bit more below.

We first observe, however, that there is still something left implicit in statement (iii). This is the agent's actual choice of posteriors (or, depending on the version, retroactive choice of priors resulting in posteriors) for the incoming hypotheses.

Fixing the posterior scheme

But given a particular hypothesis scheme, perhaps we could always derive convergence for a particular H^*-measure-1 class of outcome streams, that *is* independent of the exact (positive) posterior values the agent chooses to assign to these incoming hypotheses?

Unfortunately, this is again not attainable in general. Again we indeed already have for the standard Bayesian agent that a different choice of prior distribution over the exact same hypothesis

set (more exactly, a different *regular* prior distribution that assigns each element positive probability) can result in a different H^*-measure-1 class of outcome sequences on which it converges to H^*. In fact, we can show that there are single hypotheses sets such that for *every* individual stream we can tweak the priors in such a way that convergence fails on *this* stream.

Proposition 4. *There exist countable hypothesis sets Θ and hypotheses $H^* \in \Theta$ such that for every infinite outcome stream E^ω, there is a regular prior distribution P over Θ such that the Bayesian agent P's predictive probabilities do not converge to H^* on E^ω.*

Proof. See Appendix 2.A.3. □

This result pertains to the initial hypothesis set of a standard (or indeed open-minded) agent, but the initial set is already part of an open-minded agent's hypothesis scheme, and the result could also again readily be modified to pertain to the posterior assignments to a scheme's newly formulated hypotheses. Thus the result implies that we must allow the measure-1 class to also depend on the *posterior scheme*, that specifies what numerical posterior values are assigned to each (incoming) hypothesis. Formally, the combination of the hypothesis and the posterior scheme is now codified in a function $P_{(\cdot)}$ that maps each finite data sequence E^t to a posterior distribution P_{E^t} over the hypothesis set Θ_{E^t}. Again, this function must also return distributions that actually correspond to some possible open-minded agent; that is to say, these distributions must be consistent with the specifications of the version of the open-minded agent in question. For instance, in case of the hybrid agent (sec. 2.2.6 above), the distribution P_{E^t} is the distribution $P_t(\cdot \mid \Theta_N)$ after having observed E^t and with $\Theta_N = \Theta_{E^t}$. By the specification of the hybrid agent, this distribution $P_t(\cdot \mid \Theta_N) = P_0(\cdot \mid E^t, \Theta_N)$ is derived from some prior distribution P_0 over Θ_N. This latter distribution must cohere with the priors $P_0(\cdot \mid \Theta_{N'})$ for earlier and later hypothesis sets $\Theta_{N'}$, which likewise constrain the distributions $P_{E^s}(\cdot) = P_s(\cdot \mid \Theta_{N'})$ for E^s that extend or are extended by E^t. Whenever we invoke hypothesis and posterior schemes in the following, we implicitly limit our attention to schemes that actually correspond to open-minded agents of the version we are then considering.[9]

This then leads us, finally, to aim for a convergence statement of the following form.

(iv) For every H^*, every hypothesis and posterior scheme, there is an H^*-measure-1 class of infinite outcome streams on which the open-minded agent converges to H^*.

Having thus derived the formal structure of the strongest convergence statement we can hope for, let us expand a little bit on its conceptual status. One possible interpretation is that this statement corresponds to an assumption that prior to the inquiry, both the future hypotheses and the posteriors that will be assigned to them are, albeit still dependent on the random data and unknown the agent, already fixed. There is at least a superficial tension between such an interpretation and a crucial motivation for investigating open-minded agents, namely that

[9] Some care is required in deriving relations between the functions $P_{E^t}(\cdot \mid \Theta_{E^t})$ from the agent specifications, which also involves matching the original notation for agent functions ("$P_t(\cdot \mid \Theta_N)$") with the $P_{E^t}(\cdot \mid \Theta_{E^t})$. The former notation leaves implicit what exactly are the past data that have resulted in the posteriors and hypothesis sets, which becomes especially risky when analyzing retroactive assignments (what future hypothesis set and posteriors is $P_0(\cdot \mid \Theta_N)$ actually reconstrued from?). This will mostly matter for the proofs to follow: see appendix 2.A.1 on notation used there for details.

hypotheses and their priors are not forever fixed in advance, and the agent has the freedom to change its mind. How problematic this is, would then conceivably depend on one's view on the external process where the hypotheses and posteriors come from: is there some mechanical procedure that delivers them, or is this rather some process of creative and fundamentally unalgorithmic scientific discovery? On the other hand, we think it is actually not so clear that the mathematical structure of (proving) a statement of form (iv), "fix arbitrary x, we now show…," commits one to a conceptual view of the kind, "assuming that x is fixed in advance, we have that…," let alone what it exactly means for an hypothesis scheme to be (unknown to the agent but) determined in advance. These are philosophically murky waters, and we will here limit ourselves to noting that mathematically, this is the best we can aim for. Indeed, if already for the standard Bayesian agent the precise measure-1 class must depend on the other hypotheses and exact priors, it is only natural to aim for the analogous statement for the open-minded agent—in general. This does not exclude the possibility of deriving statements of form (i) with certain restrictions on the possible hypotheses, say a restriction of effective computability. But this lies out of the scope of the current paper.

With this conceptual provision, we are now clear on the nature of the 'a.s.' qualification. In fact, we have also already touched on the second challenge: what, exactly, is the agent function that we seek convergence for? We will now make this precise.

The completed agent measure

Given an hypothesis and a posterior scheme, an open-minded Bayesian's probability assignments after each possible finite outcome sequence are fully determined. For all finite E^t, the agent's assignment to any event A is fixed and given by

$$P_{E^t}(A) = P_{E^t}(A \mid \Theta_{E^t}). \tag{2.16}$$

The corresponding convergence statement of form (iii), for *strong* merger, is that for each hypothesis and posterior scheme, we have for an H^*-measure-1 class of infinite outcome sequences that

$$\sup_{A\in\mathfrak{F}} \left|P_{E^t}(A) - H^*(A \mid E^t)\right| \xrightarrow{t\to\infty} 0. \tag{2.17}$$

Here we still adhered to the simplifying assumption made at the beginning of sect. 2.4.1, that the truth H^* is contained in the initial hypothesis class. The general case is covered by adding the formulation of H^* on the outcome stream as an condition for the convergence. That is, for an H^*-measure-1 class of infinite outcome sequences,

$$H^* \text{ is formulated} \Longrightarrow \sup_{A\in\mathfrak{F}} \left|P_{E^t}(A) - H^*(A \mid E^t)\right| \xrightarrow{t\to\infty} 0. \tag{2.18}$$

For *weak* merger, this comes down to

$$H^* \text{ is formulated} \Longrightarrow \sup_{E_{t+1}\in\{0,1\}} \left|P_{E^t}(E_{t+1}) - H^*(E_{t+1} \mid E^t)\right| \xrightarrow{t\to\infty} 0. \tag{2.19}$$

A circumstance that makes convergence of the terms (2.16) hard to analyze is that, even under the assumption of a given hypothesis and posterior scheme, *they may not correspond to a single probability measure*. That is to say, the assignments $P_{E^t}(\cdot)$ cannot in general be reconstrued as the conditional probabilities of a particular measure: there need not be a single measure P such that $P(\cdot \mid E^t) = P_{E^t}(\cdot)$ for each E^t. This stems from the fact that an open-minded agent's assignments can be dynamically incoherent, in the sense that for finite sequences E^{t_1}, E^{t_2}, the second extending the first,

$$P_{E^{t_1}}(A \mid E^{t_2}) \neq P_{E^{t_2}}(A). \tag{2.20}$$

In words, the agent's assignment to event A at time t_1, conditional on the extended outcome sequence E^{t_2}, may not equal the agent's assignment to A at time t_2, after having in fact seen E^{t_2}. To make this slightly more concrete, consider again the hybrid open-minded agent. From its specification, there is some prior distribution P_0 such that $P_{E^{t_1}}(A \mid E^{t_2}) = P_0(A \mid E^{t_2}, \Theta_{E^{t_1}})$ and $P_{E^{t_2}}(A) = P_0(A \mid E^{t_2}, \Theta_{E^{t_2}})$. But there is no reason why the terms $P_0(A \mid \Theta_{E^{t_1}})$ and $P_0(A \mid \Theta_{E^{t_2}})$, conditional on different hypotheses, should be equal.

Nevertheless, the agent's *one-step* predictive probabilities, given a particular hypothesis and posterior scheme, *do* induce a coherent set of probability assignments. The predictive probabilities $P_{E^t}(E_{t+1})$ induce a probability assignment P^∞ on all finite evidence sequences, by

$$P^\infty(E^t) := \prod_{i=0}^{t-1} P_{E^i}(E_{i+1}), \tag{2.21}$$

and this induces a measure on all outcome streams. We will call this measure P^∞ the *completed agent measure*.

If we are able to show that, for any given hypothesis and posterior scheme, this measure retains a grain of the truth H^*, then a statement of form (iii), for *strong* merger, follows from corollary 2. That is, for any given hypothesis and posterior scheme, we can conclude that for an H^*-measure-1 class of outcome streams,

$$H^* \text{ is formulated} \Longrightarrow \sup_{A \in \mathfrak{F}} \left| P^\infty(A \mid E^t) - H^*(A \mid E^t) \right| \xrightarrow{t \to \infty} 0. \tag{2.22}$$

However, this statement concerns the completed agent measure P^∞, and not the open-minded agent's actual assignments at each time, given by (2.16). These assignments $P^\infty(A \mid E^t)$ and $P_{E^t}(A)$ may not coincide. The potential disagreement lies in the fact that $P^\infty(A \mid E^t)$ is already influenced by what *future* hypotheses, formulated after E^t but before A, say about A; whereas $P_{E^t}(A)$ only depends on the hypothesis set Θ_{E^t}.

Still, we do have by definition that these functions coincide on the one-step predictive probabilities. We have that $P^\infty(E_{t+1} \mid E^t) = P_{E^t}(E_{t+1})$ for each outcome sequence E^t and single outcome E_{t+1}, so that convergence statement (2.22) does imply convergence statement (2.19).[10]

[10] In fact, for any t, measures $P^\infty(\cdot \mid E^t)$ and P_{E^t} coincide up to the smallest time ahead at which a new hypothesis will be formulated; though this only implies weak convergence of the latter for $d > 1$ if this time horizon will eventually always be at least d.

Thus, if we can show, for any given hypothesis and posterior scheme, that the open-minded agent's completed agent measure holds a grain of the truth, then we can derive a convergence statement of form (iii) for *weak* merger of the agent functions. Consequently, in the following, we will work towards ensuring this property, that the completed agent measure holds a grain of the truth, whenever the truth is formulated.

The failure of holding a truth-grain

Consider again the hybrid open-minded agent. Connecting back to the discussion of sect. 2.3.4, it might seem that the completed agent measure should hold a grain of the truth as soon as for every single E^t, the retroactive prior function $P_0(\cdot \mid \Theta_{E^t})$ holds at least a grain p^* of H^*; that is, whenever all these $P_0(\cdot \mid \Theta_{E^t})$ *uniformly* retain at least the same grain of the truth. This, however, is *not* so.

That this cannot be so is again already implied by example 2.3. This example in fact features a (partially specified) hypothesis and posterior scheme for overfitting hypothesis generation, where every $P_0(\cdot \mid \Theta_{E^t})$ for $t \geq t^*$ holds at least a grain p^* of the truth. Yet we saw that the agent (the completed agent measure) in that example fails to merge with H^*, which by the contraposition of corollary 2 entails that the completed agent measure cannot hold a grain of H^*.

Proposition 5. *For the hybrid open-minded Bayesian, there are hypothesis schemes with $H^* \in \Theta_0$ such that nevertheless the completed agent measure fails to hold a grain of the truth: there is no $a \in (0,1)$ with $P^\infty(E^t) \geq a \cdot H^*(E^t)$ for all E^t.*

Proof. Such a scheme is given by example 2.3; see appendix 2.A.4 for details. □

What, intuitively, explains this fact, that each $P_0(\cdot \mid \Theta_{E^t})$ can uniformly hold a grain of the truth, yet P^∞ does not? The difference between each of the former functions and P^∞ is that in the latter, overfitting hypotheses are not represented in the predictive probabilities issued by the agent until this hypothesis actually comes in. But by definition these overfitting hypotheses have high likelihood (and thus issue high predictive probabilities) on these initial segments; so taking them out will deflate the agents' predictive probabilities on these initial segments. The counterexample shows that this effect can be so strong that it destroys the grain of the truth.

In our proposal of a *forward-looking* open-minded Bayesian, that we turn to now, we focus on making sure that the completed agent measure does retain a grain of the truth, whenever the truth is formulated, in order to derive a guarantee of truth-convergence.

2.4.2 The forward-looking open-minded Bayesian, proto-version

We first consider a version of an open-minded Bayesian, a proto-version of the forward-looking open-minded Bayesian that we propose in sect. 2.4.3 below, that rests on the following simple idea. Instead of a limited reservoir of probability for assigning *priors* to new hypotheses, the agent has a limited reservoir of *posterior* mass to assign to new hypotheses.

Specification

The forward-looking open-minded agent, in this proto-version, is like the silent open-minded agent, in that we do not stipulate a catch-all or a limited absolute reservoir of prior probability. However, we do stipulate a limited absolute reservoir of *posterior* probability: unlike the silent open-minded Bayesian, that can assign any posterior to a new hypothesis, the agent must shave off a new posterior from this reservoir, thereby shrinking the reservoir for posterior assignments to future new hypotheses. We assume that the starting reservoir holds a certain real-valued mass $d > 0$ (we do not need to assume that this mass is bounded by 1). In addition, as a minimal restriction that facilitates the proof of truth-convergence, we assume that there is a constant $c < 1$ such that agent is not allowed to assign a posterior greater than c to any single new hypothesis.

In summary, the **proto-version** of the **forward-looking open-minded Bayesian** proceeds as follows.

$(t = 0)$ N **explicit hypotheses**

As in the silent version, each explicit hypothesis H_i in Θ_N receives a prior $P_0(H_i \mid \Theta_N) > 0$ conditional on Θ_N, such that $\sum_{i<N} P_0(H_i \mid \Theta_N) = 1$. In addition, there is assumed a reservoir $\tau_N = d > 0$ of posterior probability, and a maximal one-time probability $c < 1$.

$(t > 0)$ **Evidence** E^t

Updating proceeds in the usual way, conditional on the current hypothesis set Θ_N.

$(t > 0)$ **New hypothesis** H_N

As in the silent version, when a new hypothesis H_N is formulated, extending the hypothesis set to $\Theta_{N+1} = \Theta_N \cup \{H_N\}$, the posterior $P_t(H_N \mid \Theta_{N+1})$ is directly set to a value p_N; but now this value $p_N \leq c$ must be obtained from decomposing the posterior reservoir τ_N into p_N and a remainder $\tau_{N+1} = \tau_N - p_N$ that is the new posterior reservoir.

Verification

The forward-looking open-minded Bayesian's constraints in attributing posterior mass to newly formulated hypotheses rules out a scenario like example 2.3, where constrained prior assignments still lead to high posterior values. As a matter of fact, the restriction on posterior values results in a completed agent measure that *does* retain a grain of the truth, whenever it is proposed.

Theorem 6. *For the proto-version of the forward-looking open-minded Bayesian, for any hypothesis and posterior scheme, the completed agent measure conditional on any E^t with $H^* \in \Theta_{E^t}$ holds a grain of H^*.*

Proof. See appendix 2.A.5. □

Corollary 7. *For the proto-version of the forward-looking open-minded Bayesian, for any hypothesis and posterior scheme, we have that H^*-a.s.*

$$H^* \text{ is formulated} \Longrightarrow \sup_{E_{t+1}\in\{0,1\}} \left|P_{E^t}(E_{t+1}) - H^*(E_{t+1} \mid E^t)\right| \xrightarrow{t\to\infty} 0.$$

Discussion

As mentioned, this proto-version of a forward-looking Bayesian is a constrained version of the silent open-minded Bayesian. More precisely, it is a constrained version, not of the retroactive, but of the *standard* variant of the silent Bayesian. The posteriors of new hypotheses are chosen directly; and however this is done (within the constraint of the posterior reservoir), it is not required to be (not part of the agent's specification to be) an explicit calculation of the posterior from a chosen prior and the hypothesis's likelihood on the past outcome sequence.

Again, the choice of posterior *can* always proceed like this: formally, any choice of posterior corresponds, via the likelihood on the past data, to a choice of prior. But the *constraint* on the posteriors does not translate into a simple constraint on the priors, depending as it does on the contingent fact of the actually formulated hypotheses' likelihoods, and so a *retroactive* variant of the forward-looking Bayesian does not appear a natural option—as, of course, its name is intended to suggest.

That said, the idea of an absolute reservoir of posterior probability is not a terribly natural conception. Unlike the idea of an absolute reservoir of prior probability, it cannot be coupled to a conception of a prior assignment to a catch-all hypothesis, from which new hypotheses may be shaven off. Perhaps the best way to understand this is simply as a pragmatic device, that is easy to understand and does the job of regaining the guarantee of truth-convergence.

However, we think there is yet a conceptually more pleasing option, that is formally very similar to the current version but that has a more natural interpretation. In fact, this version, our actual forward-looking Bayesian, *does* regain the idea of shaving prior mass from a catch-all, while still looking forward.

2.4.3 The forward-looking open-minded Bayesian

An alternative way of defusing the threat of extreme posteriors of incoming hypotheses is to place restrictions, not directly on the posteriors, but on the *likelihoods* of new hypotheses. Our proposal is to introduce the stipulation that new hypotheses have some default likelihood on past outcomes.

We will focus on an idea that we borrowed from the theory of competitive online learning[11], and that has important technical and conceptual advantages. This idea is to identify the likelihood of

[11] See Cesa-Bianchi and Lugosi, 2006 for a general account of competitive online learning or prediction with expert advice. The idea that we refer to, first proposed, within the setting of *specialists* (Freund et al., 1997), by Chernov and Vovk (2009), is known as the *specialist* or *abstention trick*; also see Koolen, Adamskiy and Warmuth, 2012; Mourtada and Maillard, 2017. An instance of this idea also appears in Romeijn (2004, p. 349).

new hypotheses on past data with the *agent's* probability assignment to this data, induced from its past predictive probabilities. That is, a new hypothesis H_N's likelihood $H_N(E^t)$ on the data sequence E^t generated by its time t of formulation is set equal to the product $\prod_{s=0}^{t-1} P_0(E_{s+1} \mid E^s, \Theta_{N(s)})$ of predictive probabilities. Note that this is precisely the completed agent measure's assignment $P^\infty(E^t)$.

This is a natural way of modeling that a new hypothesis is only evaluated *after* its formulation; or that with respect to this new hypotheses, the old evidence does not count. The new hypotheses is, to put it differently, at its time of formulation treated in a *neutral* fashion, in that it is supposed to have had the same predictive success on the past data as the agent itself. This also translates in this new hypothesis having, for any chosen prior $P_0(H_N \mid \Theta_{N+1})$, at its time of formulation t a *posterior* $P_0(H_N \mid E^t, \Theta_{N+1})$ that simply *equals the prior*.

Moreover, this allows us to recover the picture of a catch-all, or more precisely, the fixed well of prior probability from which the agent must draw in its assignment to (new) hypotheses. In combination with the restriction on prior assignments that this entails, this version of a forward-looking Bayesian indeed regains truth-convergence.

Specification

The forward-looking open-minded Bayesian, in its current version, proceeds exactly as the hybrid-open minded Bayesian, except for the crucial stipulation that each new hypothesis N_i formulated at time t_i satisfies

$$H_{N_i}(E^t) := P^\infty(E^t) \text{ for all } t \leq t_i. \tag{2.23}$$

In summary, the **forward-looking open-minded Bayesian** proceeds as follows.

$(t = 0)$ N **explicit hypotheses**

As in the hybrid version, each explicit hypothesis H_i in Θ_N receives a prior $P_0(H_i \mid \Theta_N) > 0$ conditional on Θ_N, such that $\sum_{i<N} P_0(H_i \mid \Theta_N) = 1$; and the catch-all hypothesis $\overline{\Theta}_N = \Theta \setminus \Theta_N$ receives an unconditional prior $P_0(\overline{\Theta}_N) := \tau_N$, so that the unconditional priors of the explicit hypothesis are given by $P_0(H_i) := (1 - \tau_N) \cdot P_0(H_i \mid \Theta_N)$.

$(t > 0)$ **Evidence** E^t

Updating proceeds in the usual way, conditional on the current hypothesis set Θ_N.

$(t > 0)$ **New hypothesis** H_N

As in the hybrid version, when a new explicit hypothesis H_N is formulated, extending the hypothesis set to $\Theta_{N+1} = \Theta_N \cup \{H_N\}$, the unconditional prior τ_N of the earlier catch-all is decomposed into a value $p < \tau_N$ for the unconditional prior $P_0(H_N)$ of the new hypothesis and a remainder $\tau_{N+1} = \tau_N - p$ for the unconditional prior $P_0(\overline{\Theta}_{N+1})$ of the new catch-all. The priors conditional on the new hypothesis set are obtained by renormalization, from

which the conditional posteriors are obtained by the usual updating on their likelihoods, where the new hypothesis's likelihood $H_N(E^t)$ is stipulated to equal $P^\infty(E^t)$.

Verification

Although they differ in their interpretation and also slightly in the precise shape of the constraints they impose, the forward-looking Bayesian and its proto-version share the formal property of a constraint on new posterior assignments. In appendix 2.A.5 we give a general proof that for both types of constraints shows that a completed agent measure will hold a grain of the truth, whenever it is formulated, from which weak merger of the agent follows.[12]

Theorem 8. *For the forward-looking open-minded Bayesian, for any hypothesis and posterior scheme, the completed agent measure conditional on any E^t with $H^* \in \Theta_{E^t}$ holds a grain of H^*.*

Proof. See appendix 2.A.5. □

Corollary 9. *For the forward-looking open-minded Bayesian, for any hypothesis and posterior scheme, we have that H^*-a.s.*

$$H^* \text{ is formulated} \Longrightarrow \sup_{E_{t+1}\in\{0,1\}} \left|P_{E^t}(E_{t+1}) - H^*(E_{t+1} \mid E^t)\right| \xrightarrow{t\to\infty} 0.$$

Beyond weak merger

Corollary 9 states, for the forward-looking agent, and as a consequence of the strong truth-merger of the completed agent measure, the weak truth-merger (with $d = 1$) of the agent measures P_{E^t}. The obvious further question is whether we also have strong merger, or at least weak merger for any finite d, for the agent measures P_{E^t}. We conjecture that already strong merger does hold, but unfortunately we have no proof, and must leave this as an open

[12] An alternative proof proceeds by deriving from the abstention stipulation (2.23) that the forward-looking agent's probability $P^\infty(E^t)$ must *coincide* with the retroactive prior probability $P_0(E^t \mid \Theta_{N_i})$ for every Θ_{N_i} with $t_{i+1} > t$. The additional stipulation of a fixed amount of prior mass guarantees again that these $P_0(E^t \mid \Theta_{N_i})$ indeed uniformly retain a grain of the truth, so that truth-merger follows. Recall from sect. 2.4.1 that the hybrid open-minded Bayesian's completed agent measure can fail to retain a grain of the truth even if every $P_0(\cdot \mid \Theta_{N_i})$ for $i \geq i^*$ uniformly does so: stipulation (2.23) thus rules out this possibility.

question.[13]

2.5 Conclusion

We investigated the failure of truth-convergence for Wenmackers and Romeijn's versions of open-minded Bayesianism, and, towards reclaiming this property, proposed a *forward-looking* open-minded Bayesian. The general threat to convergence to the truth is the possibility of new and false hypotheses that keep receiving too much posterior: either by direct assignment or by retroactive calculation from a high likelihood on the past evidence. The proto-version and the final version of our forward-looking Bayesian implement the two respective ways of meeting this threat: by restricting the posteriors, or by restricting the priors and likelihoods.

We think that the final version of our forward-looking agent, which is based on an idea from the theory of competitive online learning, indeed provides an elegant account of how a Bayesian agent should deal with newly formulated hypotheses. The idea of identifying a new hypothesis's likelihood with the agent's probability assignment on the past data is a graceful way of neutralizing the impact of old evidence. Moreover, this idea has the pleasant consequence that the stipulation of a limited reservoir of prior probability (with the associated interpretation of a catch-all hypothesis) is sufficient to guarantee truth-convergence. Unlike the proto-version, that we ourselves feel is mainly a technical device geared towards the aim of truth-convergence, we think the final version makes intuitive sense quite independent of this aim.

There are a number of avenues for further investigation. Firstly, we proved, more precisely, the forward-looking agent's weak truth-merger, or convergence to the true one-step predictive probabilities. We leave as an open question whether this may be extended to an arbitrary finite-length horizon, or even to strong merger, that includes all tail events. Secondly, a possible lingering doubt is that in our convergence statement the measure-1 class of sequences is dependent on the hypothesis and posterior scheme. This at least suggests an interpretation where the latter quantities are somehow fixed prior to the inquiry, which, one might feel, does not sit well with the original motivation for investigating an open-minded agent. Whether or not this is so, we showed that in general we cannot avoid this dependence, as an analogue in fact already holds in the case of the standard Bayesian. Nevertheless, it might be avoided as further refinements are added to our proposal. Perhaps, finally, the main peculiarity about our approach is that in the course of an inquiry hypotheses are not (should not be) introduced haphazardly. There will

[13] For any infinite E^ω in the measure-1 class of infinite streams on which we, for given hypothesis and posterior scheme, have strong merger with H^* of the completed agent measure, it might seem that strong truth-merger of the agent functions $P_{E^t}(\cdot \mid \Theta_{E^t})$ on this E^ω should follow, too: as the posterior reservoir is used up the measures $P^\infty(\cdot \mid E^t)$ and $P_{E^t}(\cdot \mid \Theta_{E^t})$ can differ less and less. However, on any individual E^ω, it is possible that the posterior reservoir is *not* fully used up: this allows for a counterexample, on this particular stream, where the same constant posterior keeps being assigned to new hypotheses on side-branches of E^ω to force a difference between $P^\infty(\cdot \mid E^t)$ and $P_{E^t}(\cdot \mid \Theta_{E^t})$. Now one could push further and consider the measure-1 class that is the countable intersection of the previous class and, for *every* length s, the measure-1 class of streams on which every measure $P_{E^s}(\cdot \mid E^s)$, from that point treated as a standard Bayesian, strongly merges with H^*. But even for a stream E^ω in this class, it is still consistent that the agent measures $P_t(\cdot \mid E^t)$ do not strongly merge with H^* on this particular E^ω; at the same time, such a scenario is now so bizarre that it does not seem feasible to turn it into an actual counterexample, for which this must actually happen with positive probability. This invites the hope for some (martingale) argument that such scenarios must indeed have probability 0.

normally only arise a need for formulating a new hypothesis if some misfit between the data and the current model is observed, which may indeed be regulated via a formal model verification procedure. This raises the question how (our version of) an open-minded Bayesian inductive logic may be extended beyond just *how* to incorporate externally proposed hypotheses, to also include *when* to accept such new hypotheses, and how this interacts with the guarantee of truth-convergence.

2.A Calculations and proofs

2.A.1 Notation

We introduce additional notation for use in the appendices.

For sequences E^t and E^s we write $E^t \preccurlyeq E^s$ if E^t is an initial segment of E^t, and $E^t \prec E^s$ if $E^t \preccurlyeq E^s$ and $E^t \neq E^s$. We write $E^t \mid E^s$ if neither $E^t \preccurlyeq E^s$ nor $E^s \preccurlyeq E^t$. For the concatenation of sequences E^t and E^s we write $E^{t+s} = E^t E^s$. For sequences $E^t \preccurlyeq E^s$ we write $E^{t:s}$ for the sequence E^s minus its initial segment E^t.

Recall that an hypothesis and posterior scheme are given by a function $P_{(\cdot)}$ that for given sequence E^t returns a distribution $P_{E^t} = P_{E^t}(\cdot \mid \Theta_{E^t})$ over hypothesis set Θ_{E^t}. This induces the distribution $P_{E^t}(\cdot) = \sum_{H \in \Theta_{E^t}} P_{E^t}(H) \cdot H(\cdot \mid E^t)$ over events in the outcome space.

The conditional distributions $P_{E^t}(\cdot \mid \Theta)$ for $\Theta \subseteq \Theta_{E^t}$ are clearly well-defined. One can also derive from the specifications of any of the open-minded versions we discussed that for $E^s \succ E^t$

$$P_{E^s}(\cdot \mid \Theta_{E^t}) = P_{E^t}(\cdot \mid E^{t:s}, \Theta_{E^t}), \tag{2.24}$$

a fact that we will rely on in the proofs of lemma 4 and corollary 10, in 2.A.5 below.

The conditional distributions $P_{E^t}(\cdot \mid \Theta)$ for $\Theta \supset \Theta_{E^t}$ are *not* well-defined, because the posteriors of the elements in $\Theta \setminus \Theta_{E^t}$ are not defined. Nevertheless, for the purpose of analyzing an open-minded agent's procedure of retro-actively setting a prior (as in the proof of lemma 6 in 2.A.5 below), it will be useful to agree on the following. For $E^s \succ E^t$, the probability $P_{E^t}(H \mid \Theta_{E^s})$ is the posterior probability of $H \in \Theta_{E^s}$ after E^t, retroactively calculated from the posterior probability $P_{E^s}(H \mid \Theta_{E^s})$ after E^s. More precisely, we can define for all $H \in \Theta_{E^s}$,

$$P_{E^t}(H \mid E^{t:s}, \Theta_{E^s}) := P_{E^s}(H \mid \Theta_{E^s}), \tag{2.25}$$

from which the function $P_{E^t}(\cdot \mid \Theta_{E^s})$, by using the likelihoods of all $H \in \Theta_{E^s}$ on $E^{t:s}$, can unambiguously be retrieved.

2.A.2 Calculations for example 2.3

We want to ensure (2.12), that is,

$$\frac{P_0(H_{N_i} \mid \Theta_{N_i+1}) \cdot H_{N_i}(E^{t_i})}{\sum_{H \in \Theta_{N_i+1}} P_0(H \mid \Theta_{N_i+1}) \cdot H(E^{t_i})} > r. \tag{2.26}$$

Write $q := P_0(H_{N_i} \mid \Theta_{N_i+1})$ for the conditional prior, that by (2.10) equals

$$\frac{P_0(H_i)}{1 - \tau_{N_i+1}} = \frac{2^{-i} \cdot \tau_{N_0+1}}{1 - (1 - \sum_{j=1}^{i} 2^{-j}) \cdot \tau_{N_0+1}} = \frac{2^{-i} \cdot \tau_{N_0+1}}{1 - 2^{-i} \cdot \tau_{N_0+1}}. \tag{2.27}$$

Since $H_{N_i}(E^{t_i}) = 1$, (2.26) translates into

$$q > r \cdot \left(q + \sum_{H \in \Theta_{N_i+1} \setminus \{H_{N_i}\}} P_0(H \mid \Theta_{N_i+1}) \cdot H(E^{t_i}) \right), \tag{2.28}$$

that is,

$$\frac{1-r}{r}\cdot q > \sum_{H\in\Theta_{N_i+1}\setminus\{H_{N_i}\}} P_0(H \mid \Theta_{N_i+1})\cdot H(E^{t_i}). \tag{2.29}$$

Now assuming that there is positive δ such that all other hypotheses' predictive probabilities are no more than $1-\delta$ for each possible outcome from t_{i-1} up to t_i, so that

$$\sum_{H\in\Theta_{N_i+1}\setminus\{H_{N_i}\}} P_0(H \mid \Theta_{N_i+1})\cdot H(E^{t_i}) < (1-q)\cdot(1-\delta)^{t_i-t_{i-1}}, \tag{2.30}$$

it suffices for (2.29) that

$$\frac{1-r}{r}\cdot\frac{q}{1-q} > (1-\delta)^{t_i-t_{i-1}}. \tag{2.31}$$

Writing out

$$\frac{q}{1-q} = \frac{\left(\frac{2^{-i}\cdot\tau_{N_0+1}}{1-2^{-i}\cdot\tau_{N_0+1}}\right)}{\left(1-\frac{2^{-i}\cdot\tau_{N_0+1}}{1-2^{-i}\cdot\tau_{N_0+1}}\right)} = \frac{\left(\frac{2^{-i}\cdot\tau_{N_0+1}}{1-2^{-i}\cdot\tau_{N_0+1}}\right)}{\left(\frac{1}{1-2^{-i}\cdot\tau_{N_0+1}}\right)} = 2^{-i}\cdot\tau_{N_0+1}, \tag{2.32}$$

we thus require

$$\frac{1-r}{r}\cdot 2^{-i}\cdot\tau_{N_0+1} > (1-\delta)^{t_i-t_{i-1}}, \tag{2.33}$$

that is,

$$t_i - t_{i-1} > \frac{-\log(1-r) - (-\log r) + i - \log\tau_{N_0+1}}{-\log(1-\delta)}. \tag{2.34}$$

2.A.3 Proof of proposition 4

Let the truth $H^* \in \Theta$ be Bernoulli-1/2, and put $P(H^*) = 1/2$. Define an infinite series of times $t_0, t_1, t_2, \ldots$ by $t_0 = 0$, $t_{i+1} = t_i + i + 3$. For each time t_i, let $E^{t_i}_j$ be the j-th ($0 < j \leq 2^{t_i}$) outcome sequence of length t_i. We will now define a countable collection of hypotheses $H_{i,j}$ that each overfit one particular sequence between two successive times t_{i-1} and t_i, and follow H^* elsewhere. More precisely, we define for each i, for each positive $j \leq 2^{t_i}$ and the corresponding j' such that $E^{t_{i-1}}_{j'} \prec E^{t_i}_j$, the hypothesis $H_{i,j}$ by

$$H_{i,j}(E^s) = \begin{cases} 2^{-t_{i-1}} & \text{if } E^{t_{i-1}}_{j'} \preccurlyeq E^s \preccurlyeq E^{t_i}_j \\ 0 & \text{if } E^{t_{i-1}}_{j'} \preccurlyeq E^s \text{ but } E^s \mid E^{t_i}_j \\ H^*(E^s)\cdot 2^{t_i-t_{i-1}} & \text{if } E^{t_i}_j \prec E^s \\ H^*(E^s) & \text{otherwise.} \end{cases} \tag{2.35}$$

Given an infinite outcome stream E^ω. We can now assign positive prior to each of these hypotheses as follows. Denote by $(E^{t_i}_j)^{\mathrm{C}}$ the sequence $E^{t_i}_j$ with the very last outcome inverted, 0 for 1 or vice versa. For each i, for each $j \leq 2^{t_i}$, let

$$P(H_{i,j}) = \begin{cases} 2^{-i-2} & \text{if } (E^{t_i}_j)^{\mathrm{C}} \prec E^\omega \\ 2^{-i-2}\cdot(2^{t_i}-1)^{-1} & \text{otherwise.} \end{cases} \tag{2.36}$$

This is a valid prior assignment because $\sum_{H\in\Theta} P(H) = 2^{-1} + \sum_{i>0}(2^{-i-1}) = 1$.

Now we consider, on the stream E^ω, for arbitrary i and the j such that $E_j^{t_i} \prec E^\omega$, the error in the agent's predictive probability $P(0 \mid E_j^{t_i-1})$ after having observed all of $E_j^{t_i}$ but the very last outcome. That is, we consider the distance

$$\left| P(0 \mid E_j^{t_i-1}) - H^*(0 \mid E_j^{t_i-1}) \right|. \tag{2.37}$$

To this end, write $\Theta' := \Theta \setminus \{H_{i,j}\}$ and first consider the posterior ratio of $P(H_{i,j} \mid E_j^{t_i-1})$, write α, and $P(\Theta' \mid E_j^{t_i-1}) = 1 - \alpha$,

$$\frac{\alpha}{1-\alpha} = \frac{P(H_{i,j} \mid E_j^{t_i-1})}{P(\Theta' \mid E_j^{t_i-1})} = \frac{P(H_{i,j}) \cdot H_{i,j}(E_j^{t_i-1})}{P(\Theta') \cdot P(E_j^{t_i-1} \mid \Theta')}. \tag{2.38}$$

It follows from specification (2.35) that all hypotheses in Θ' assign true probability $H^*(E_j^{t_i-1})$ to $E_j^{t_i-1}$, except for the overfitting hypotheses $H_{i',j'}$ for $i' \leq i$ and j' such that there is j'' with $E_{j''}^{t_{i'}-1} \prec E_{j'}^{t_{i'}}, E^\omega$. But for each $i' < i$, among these hypotheses $H_{i',j'}$ there is only one $H_{i',k'}$ that does *not* give probability 0 to $E_j^{t_i-1}$, and with assignment (2.36) each member of the majority already holds at least as much prior as the single exception $H_{i',k'}$. Similarly, for i, it is, among these $H_{i,j'}$ and apart from $H_{i,j}$, only the hypothesis $H_{i,k}$ for $E_k^{t_i} \prec E^\omega$ that does not assign probability 0 to $E_j^{t_i-1}$, and each other $H_{i,j'}$ already holds at least as much prior as $H_{i,k}$. We thus have that the likelihood of hypothesis set Θ' satisfies

$$P(E_j^{t_i-1} \mid \Theta') = \sum_{H\in\Theta'} P(H \mid \Theta') \cdot H(E_j^{t_i-1}) < H^*(E_j^{t_i-1}) = 2^{-t_i+1}, \tag{2.39}$$

wherefore

$$\begin{aligned} \frac{\alpha}{1-\alpha} &> \frac{2^{-i-2} \cdot 2^{-t_{i-1}}}{(1-2^{-i-2}) \cdot 2^{-t_i+1}} \\ &= \frac{2^{-i-3}}{(1-2^{-i-2}) \cdot 2^{-(t_i-t_{i-1})}} \\ &= \frac{2^{-i-3}}{(1-2^{-i-2}) \cdot 2^{-i-3}} \\ &> 1, \end{aligned}$$

meaning that $\alpha > 1/2$.

Finally, apart from $H_{i,j}$, it is only the hypothesis $H_{i,k}$ for $E_k^{t_i} \prec E^\omega$ that is still included in the posterior over Θ conditional on $E_j^{t_i-1}$ (that did not assign probability 0 to $E_j^{t_i-1}$) *and* that gives a predictive probability $H_{i,k}(0 \mid E_j^{t_i-1})$ different from $H^*(0 \mid E_j^{t_i-1}) = 1/2$. Write $\alpha' := P(H_{i,k} \mid E_j^{t_i-1})$ for the posterior of $H_{i,k}$, and abbreviate $\Theta_{i;j,k} := \{H_{i,j}, H_{i,k}\}$. Since

indeed $H_{i,k}(0 \mid E_j^{t_i-1}) = 1 - H_{i,j}(0 \mid E_j^{t_i-1})$,

$$P(0 \mid E_j^{t_i-1}, \Theta_{i;j,k}) = \frac{\alpha}{\alpha+\alpha'} \cdot H_{i,j}(0 \mid E_j^{t_i-1}) + \frac{\alpha'}{\alpha+\alpha'} \cdot H_{i,k}(0 \mid E_j^{t_i-1}) \tag{2.40}$$

evaluates to either $\frac{\alpha}{\alpha+\alpha'} = 1 - \frac{\alpha'}{\alpha+\alpha'}$ or $\frac{\alpha'}{\alpha+\alpha'}$. Using $\alpha' < 1/2 < \alpha$, it follows that

$$\left|P(0 \mid E_j^{t_i-1}, \Theta_{i;j,k}) - H^*(0 \mid E_j^{t_i-1})\right| = 1/2 - \frac{\alpha'}{\alpha+\alpha'}. \tag{2.41}$$

We can then rewrite (2.37) as

$$\left|(\alpha+\alpha') \cdot P(0 \mid E_j^{t_i-1}, \Theta_{i;j,k}) + (1-(\alpha+\alpha')) \cdot H^*(0 \mid E_j^{t_i-1}) - H^*(0 \mid E_j^{t_i-1})\right|, \tag{2.42}$$

which simplifies to

$$\begin{aligned}(\alpha+\alpha') \cdot \left|P(0 \mid E_j^{t_i-1}, \Theta_{i;j,k}) - H^*(0 \mid E_j^{t_i-1})\right| &= (\alpha+\alpha') \cdot \left(1/2 - \frac{\alpha'}{\alpha+\alpha'}\right) \\ &= \frac{\alpha+\alpha'}{2} - \alpha' \\ &> 1/4 - 1/2 \cdot \alpha'.\end{aligned}$$

But note that $H_{i,j}$ and $H_{i,k}$ have the same likelihood $H_{i,j}(E_j^{t_i-1}) = H_{i,k}(E_j^{t_i-1})$, so that by assignment (2.36) the ratio

$$\frac{\alpha}{\alpha'} = \frac{P(H_{i,j})}{P(H_{i,k})} = 2^{t_i} - 1, \tag{2.43}$$

which implies that $\alpha' < (2^{t_i}-1)^{-1}$ is arbitrarily small for large enough i. That means that indeed for any choice of $\varepsilon > 0$, we have for infinitely many i that

$$\left|P(0 \mid E_j^{t_i-1}) - H^*(0 \mid E_j^{t_i-1})\right| > 1/4 - \varepsilon,$$

blocking convergence on the stream E^ω. □

2.A.4 Proof of proposition 5

Consider example 2.3 with $t_0 = 0$, $\varepsilon' > 1/4$, and where after each t_i all hypotheses H_{N_j} for $j \leq i$ always give predictive probabilities $(1/2, 1/2)$. Let the sequence of time points $t_0 < t_1 < t_2 \ldots$ at which overfitting hypotheses are introduced satisfy (2.13), with prior assignments given by (2.10). This defines a hypothesis and posterior scheme, and thus induces a completed agent measure.

Next, take an infinite outcome stream E^ω that is constructed as follows. For any $i \geq 0$, take for the subsequence $E^{t_i+2:t_{i+1}}$ any sequence of length $t_{i+1} - t_i - 1$, and let E_{t_i+1} be the outcome with $P_{t_i}(E_{t_i+1} \mid \Theta_{E^{t_i}}) < 1/2 - \varepsilon' = 1/4$ (for E_1 take either 0 or 1). Now the completed agent measure

P^∞ fails to hold a grain of H^* on any such sequence E^ω. Namely, for such a sequence E^ω we have by construction that for each t, with j maximal such that $t_j < t$, that

$$P^\infty(E^t) < \left(2^{-1}\right)^{t-j} \cdot \left(2^{-2}\right)^j = 2^{-t-j}. \tag{2.44}$$

But since $2^{-t-j}/2^{-t} = 2^{-j}$ goes to 0 as t hence j goes to infinity, there is no positive a such that $P^\infty(E^t) \geq a \cdot H^*(E^t)$ for all t. □

2.A.5 Proof of theorems 6 and 8

We show for both the forward-looking open-minded Bayesian agent and its proto-version that for any hypothesis and posterior scheme, any finite outcome sequence E^{t_0}, for any hypothesis $H \in \Theta_{E^{t_0}}$, there is a constant $a \in (0,1)$ such that for every outcome sequence $E^t \succcurlyeq E^{t_0}$ it holds that

$$P^\infty(E^{t_0:t} \mid E^{t_0}) \geq a \cdot H(E^{t_0:t} \mid E^{t_0}). \tag{2.45}$$

In words, for any outcome sequence E^{t_0}, the completed agent measure conditional on E^{t_0} holds a positive grain of every hypothesis H in the hypothesis set $\Theta_{E^{t_0}}$. In particular, the completed agent measure conditional on E^{t_0} holds a grain of the truth H^*, if H^* is in $\Theta_{E^{t_0}}$.

Our proof consists of two main steps. First, we show that for any open-minded agent the completed agent measure conditional on E^{t_0} dominates the agent function $P_{E^{t_0}}$ with a factor that involves the posteriors assigned to new hypotheses (lemma 4 and corollary 10). Second, we show for (the proto-version of) the forward-looking open-minded Bayesian that this latter factor is indeed at least a positive constant (lemma 5 and 6, respectively).

In all of the following statements we quantify over all E^{t_0} and $E^t \succcurlyeq E^{t_0}$, and in the accompanying proofs we start by presupposing any such two sequences. This allows for the following simplified notation, that unambiguously pertains to a particular instantiated E^t and initial segment E^{t_0}. We abbreviate $P_s := P_{E^s}$ and $\Theta_s := \Theta_{E^s}$ for all $E^s \preccurlyeq E^t$. Moreover, we always let $i \geq 0$ denote the number of new hypotheses that are formulated along the sequence $E^{t_0+1:t}$, and we write $p_j := P_{t_j}(H_{E^{t_j}} \mid \Theta_{t_j})$ for the conditional posterior assigned to the j-th ($j \leq i$) such hypothesis $H_{E^{t_j}} \in \Theta_{t_j} \setminus \Theta_{t_{j-1}}$, incoming at time t_j.

Lemma 4. *For an open-minded agent, we have that for any hypothesis and posterior scheme, for every E^{t_0}, every $E^t \succ E^{t_0}$, every $0 \leq j \leq i$,*

$$P_{t_j}(E^{t_j:t} \mid \Theta_{t_j}) \geq \frac{\prod_{k=0}^{j-1}(1-p_{k+1}) \cdot P_{t_0}(E^{t_0:t} \mid \Theta_{t_0})}{\prod_{k=0}^{j-1} P_{t_k}(E^{t_k:t_{k+1}} \mid \Theta_{t_k})}. \tag{2.46}$$

Proof. We proceed by induction. The base case, $j = 0$, follows trivially from empty products.

Next, assuming as induction hypothesis that (2.46) holds for given $j < i$, we derive for $j+1$ that

$$\begin{aligned}
P_{t_{j+1}}(E^{t_{j+1}:t} \mid \Theta_{t_{j+1}}) &= \sum_{H\in\Theta_{t_{j+1}}} P_{t_{j+1}}(H \mid \Theta_{t_{j+1}}) \cdot H(E^{t_{j+1}:t} \mid E^{t_{j+1}}) \\
&\geq (1-p_{j+1}) \sum_{H\in\Theta_{t_j}} P_{t_j}(H \mid E^{t_j:t_{j+1}}, \Theta_{t_j}) \cdot H(E^{t_{j+1}:t} \mid E^{t_{j+1}}) \\
&= (1-p_{j+1}) \sum_{H\in\Theta_{t_j}} \frac{P_{t_j}(H \mid \Theta_{t_j}) \cdot H(E^{t_j:t_{j+1}} \mid E^{t_j})}{P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j})} \cdot H(E^{t_{j+1}:t} \mid E^{t_{j+1}}) \\
&= (1-p_{j+1}) \cdot \frac{\sum_{H\in\Theta_{t_j}} P_{t_j}(H \mid \Theta_{t_j}) \cdot H(E^{t_j:t} \mid E^{t_j})}{P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j})} \\
&= \frac{(1-p_{j+1}) \cdot P_{t_j}(E^{t_j:t} \mid \Theta_{t_j})}{P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j})} \\
&\geq \frac{(1-p_{j+1}) \cdot \prod_{k=0}^{j-1}(1-p_{k+1}) \cdot P_{t_0}(E^{t_0:t} \mid \Theta_{t_0})}{P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j}) \cdot \prod_{k=0}^{j-1} P_{t_k}(E^{t_k:t_{k+1}} \mid \Theta_{t_k})} \\
&= \frac{\prod_{k=0}^{j}(1-p_{k+1}) \cdot P_{t_0}(E^{t_0:t} \mid \Theta_{t_0})}{\prod_{k=0}^{j} P_{t_k}(E^{t_k:t_{k+1}} \mid \Theta_{t_k})}. \qquad \square
\end{aligned}$$

Corollary 10. *For an open-minded agent, we have that for any hypothesis and posterior scheme, for every E^{t_0}, every $E^t \succ E^{t_0}$,*

$$P^\infty(E^{t_0:t} \mid E^{t_0}) \geq \prod_{j=0}^{i-1}(1-p_{j+1}) \cdot P_{t_0}(E^{t_0:t} \mid \Theta_{t_0}). \tag{2.47}$$

Proof. We write out

$$\begin{aligned}
P^\infty(E^{t_0:t} \mid E^{t_0}) &= \prod_{s=t_0}^{t-1} P_s(E_{s+1} \mid \Theta_s) \\
&= \left(\prod_{j=0}^{i-1} \prod_{s=t_j}^{t_{j+1}-1} P_s(E_{s+1} \mid \Theta_s)\right) \prod_{s=t_i}^{t-1} P_s(E_{s+1} \mid \Theta_{t_i}) \\
&= \left(\prod_{j=0}^{i-1} P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j})\right) \cdot P_{t_i}(E^{t_i:t} \mid \Theta_{t_i}),
\end{aligned}$$

where the latter equality follows from the fact that for each j and $t_j \leq t'_j < t_{j+1}$ we have

$$\begin{aligned}\prod_{s=t_j}^{t'_j} P_s(E_{s+1} \mid \Theta_s) &= \prod_{s=t_j}^{t'_j} P_{t_j}(E_{s+1} \mid E^{t_j:s}, \Theta_{t_j}) \\ &= \prod_{s=t_j}^{t'_j} \frac{P_{t_j}(E^{t_j:s+1} \mid \Theta_{t_j})}{P_{t_j}(E^{t_j:s} \mid \Theta_{t_j})} \\ &= \frac{P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j})}{P_{t_j}(E^{t_j:t_j} \mid \Theta_{t_j})} \\ &= P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j}).\end{aligned}$$

But applying lemma (2.46) for $i = j$ then yields

$$\begin{aligned}P^\infty(E^t \mid E^{t_0}) &\geq \left(\prod_{j=0}^{i-1} P_{t_j}(E^{t_j:t_{j+1}} \mid \Theta_{t_j})\right) \cdot \frac{\prod_{j=0}^{i-1}(1-p_{j+1}) \cdot P_{t_0}(E^t \mid \Theta_{t_0})}{\prod_{j=0}^{i-1} P_{t_j}(E^{t_{j+1}} \mid \Theta_{t_j})} \\ &= \prod_{j=0}^{i-1}(1-p_{j+1}) \cdot P_{t_0}(E^t \mid \Theta_{t_0}). \quad \square\end{aligned}$$

Lemma 5. *For the proto-version of the forward-looking open-minded agent, we have that for every hypothesis and posterior scheme, there is a constant $b \in (0,1)$ such that for every E^{t_0}, every $E^t \succ E^{t_0}$,*

$$\prod_{j=1}^{i}(1-p_j) \geq b. \tag{2.48}$$

Proof. We have by specification that $0 < p_j \leq c$ for each j and a positive constant $c < 1$, and that $\sum_{j=1}^{i} p_j \leq d$ for some positive constant d. Using the standard inequality $\frac{x-1}{x} \leq \ln x$ for $x > 0$, this allows us to derive

$$\begin{aligned}-\ln \prod_{j=1}^{i}(1-p_j) &= \sum_{j=1}^{i} -\ln(1-p_j) \\ &\leq \sum_{j=1}^{i} \frac{p_j}{1-p_j} \\ &\leq \frac{1}{1-c} \sum_{j=1}^{i} p_j \\ &\leq \frac{d}{1-c},\end{aligned}$$

where the second inequality follows from the fact that $1-c \leq 1-p_j$ for all j. Thus we have

$$\prod_{j=1}^{i}(1-p_j) \geq \exp\left(-\frac{d}{1-c}\right), \tag{2.49}$$

yielding the desired statement with constant $b = \exp\left(-\frac{d}{1-c}\right)$ independent of E^t. □

Lemma 6. *For the forward-looking open-minded agent, we have that for every hypothesis and posterior scheme, there is a constant $b \in (0,1)$ such that for every E^{t_0}, every $E^t \succ E^{t_0}$,*

$$\prod_{j=1}^{i}(1-p_j) \geq b. \tag{2.50}$$

Proof. By specification, and in particular the abstention trick (2.23), for each j the posterior $p_j = P_{t_j}(H_{t_j} \mid \Theta_{t_j})$ conditional on Θ_{t_j} equals the prior $P_0(H_{t_j} \mid \Theta_{t_j})$ conditional on Θ_{t_j}. But the latter is calculated from a choice of *absolute* prior, denoted p'_j, by

$$p_j = \frac{p'_j}{1-\tau_j} = \frac{\tau_{j-1}-\tau_j}{1-\tau_j}, \tag{2.51}$$

where τ_j is the probability of the catch-all after formulation of H_{t_j}. We thus have

$$\begin{aligned}\prod_{j=1}^{i}\left(1-p_j\right) &= \prod_{j=1}^{i}\left(1-\frac{\tau_{j-1}-\tau_j}{1-\tau_j}\right)\\ &= \prod_{j=1}^{i}\left(\frac{1-\tau_{j-1}}{1-\tau_j}\right)\\ &= \frac{1-\tau_0}{1-\tau_i}\\ &\geq 1-\tau_0,\end{aligned}$$

yielding the desired statement with constant $b = 1-\tau_0$ independent of E^t. □

Finally, combining the previous results, we obtain that for the (proto-version of) the forward-looking open-minded Bayesian, for any hypothesis and posterior scheme, any E^{t_0}, any hypothesis $H \in \Theta_{E^{t_0}}$, any $E^t \succcurlyeq E^{t_0}$, it holds that

$$\begin{aligned}P^{\infty}(E^{t_0:t} \mid E^{t_0}) &\geq \prod_{j=0}^{i-1}(1-p_{j+1}) \cdot P_{t_0}(E^{t_0:t} \mid \Theta_{t_0})\\ &\geq b \cdot P_{t_0}(E^{t_0:t} \mid \Theta_{t_0})\\ &\geq b \cdot P_{t_0}(H \mid \Theta_{t_0}) \cdot H(E^{t_0:t} \mid E^{t_0}),\end{aligned}$$

yielding the desired statement (2.45) with constant $a = b \cdot P_{t_0}(H \mid \Theta_{t_0})$ independent of $E^{t_0:t}$. □

Chapter 3

Why optional stopping is a problem for Bayesians

Abstract

Recently, optional stopping has been a subject of debate in the Bayesian psychology community. Rouder (2014) argues that optional stopping is no problem for Bayesians, and even recommends the use of optional stopping in practice, as do Wagenmakers et al. (2012). This article addresses the question whether optional stopping is problematic for Bayesian methods, and specifies under which circumstances and in which sense it is and is not. By slightly varying and extending Rouder's (2014) experiments, we illustrate that, as soon as the parameters of interest are equipped with default or pragmatic priors — which means, in most practical applications of Bayes factor hypothesis testing — resilience to optional stopping can break down. We distinguish between three types of default priors, each having their own specific issues with optional stopping, ranging from no-problem-at-all (Type 0 priors) to quite severe (Type II priors).

3.1 Introduction

P-value based null-hypothesis significance testing (NHST) is widely used in the life and behavioral sciences, even though the use of p-values has been severely criticized for at least the last 50 years. During the last decade, within the field of psychology, several authors have advocated the Bayes factor as the most principled alternative to resolve the problems with p-values. Subsequently, these authors have made an admirable effort to provide practitioners with *default Bayes factors* for common hypothesis tests (Rouder et al. (2009), Jamil et al. (2016) and Rouder et al. (2012) and many others).

We agree with the objections against the use of p-value based NHST and the view that this paradigm is inappropriate (or at least far from optimal) for scientific research, and we agree that the Bayes factor has many advantages. However, as also noted by Gigerenzer and Marewski,

2014, it is not the panacea for hypothesis testing that a lot of articles make it appear. The Bayes factor has its limitations (cf. also (Tendeiro and Kiers, 2019)), and it seems that the subtleties of when those limitations apply sometimes get lost in the overwhelming effort to provide a solution to the pervasive problems of p-values.

In this article we elucidate the intricacies of handling optional stopping with Bayes factors, primarily in response to Rouder (2014). *Optional stopping* refers to 'looking at the results so far to decide whether or not to gather more data', and it is a desirable property of a hypothesis test to be able to *handle optional stopping*. The key question is whether Bayes factors can or cannot handle optional stopping. Yu et al. (2014), Sanborn and Hills (2014) and Rouder (2014) tried to answer this question from different perspectives and with different interpretations of the notion of handling optional stopping. Rouder (2014) illustrates, using computer simulations, that optional stopping is not a problem for Bayesians, also citing Lindley (1957) and Edwards, Lindman and Savage (1963) who provide mathematical results to a similar (but not exactly the same) effect. Rouder used the simulations to concretely illustrate more abstract mathematical theorems; these theorems are indeed formally proven by Deng, Lu and Chen (2016) and, in a more general setting, by Hendriksen, De Heide and Grünwald (2020). Other early work indicating that optional stopping is not a problem for Bayesians includes Savage (1954) and Good (1991). We briefly return to all of these in Section 3.5.

All this earlier work notwithstanding, we maintain that optional stopping can be a problem for Bayesians — at least for *pragmatic Bayesians* who are either willing to use so-called 'default', or 'convenience' priors, or otherwise are willing to admit that their priors are imperfect and are willing to subject them to robustness analyses. In practice, nearly all statisticians who use Bayesian methods are 'pragmatic' in this sense.

Rouder (2014) was written mainly in response to Yu et al. (2014), and his main goal was to show that Bayesian procedures retain a clear interpretation under optional stopping. He presents a criterion which, if it holds for a given Bayesian method, indicates that, in some specific sense, it performs as one would hope under optional stopping. The main content of this article is to investigate this criterion, which one may call *prior-based calibration*, for common testing scenarios involving default priors. We shall encounter two types of default priors, and we shall see that Rouder's calibration criterion — while indeed providing a clear *interpretation* to Bayesian optional stopping whenever defined — is in many cases either of limited *relevance* (Type I priors) or *undefined* (Type II priors).

We consider a strengthening of Rouder's check which we call *strong calibration*, and which remains meaningful for all default priors. Then, however, we shall see that strong calibration fails to hold under optional stopping for all default priors except, interestingly, for a special type of priors (which we call "Type 0 priors") on a special (but common) type of nuisance parameters. Since these are rarely the only parameters incurring in one's models, one has to conclude that optional stopping is usually a problem for pragmatic Bayesians — at least under Rouder's calibration criterion of handling optional stopping. There exist (at least) two other reasonable definitions of 'handling optional stopping', which we provide in Section 3.5. There we also discuss how, under these alternative definitions, Type I priors are sometimes less problematic, but Type II priors still are. As explained in the conclusion (Section 3.6), the overall crux is that default and pragmatic priors represent *tools* for inference just as much or even more

than *beliefs* about the world, and should thus be equipped with a precise prescription as to what type of inferences they can and cannot be used for. A first step towards implementing this radical idea is given by one of us in the recent paper *Safe Probability* (Grünwald, 2018).

Readers who are familiar with Bayesian theory will not be too surprised by our conclusions: It is well-known that what we call Type II priors violate the *likelihood principle* (Berger and Wolpert, 1988) and/or lead to (mild) forms of *incoherence* (Seidenfeld, 1979) and, because of the close connection between these two concepts and optional stopping, it should not be too surprising that issues arise. Yet it is still useful to show how these issues pan out in simple computer simulations, especially given the apparently common belief that optional stopping is *never* a problem for Bayesians. The simulations will also serve to illustrate the difference between the subjective, pragmatic and objective views of Bayesian inference, a distinction which matters a lot and which, we feel, has been underemphasized in the psychology literature — our simulations may in fact serve to help the reader decide what viewpoint he or she likes best.

In Section 3.2 we explain important concepts of Bayesianism and Bayes factors. Section 3.3 explains Rouder's calibration criterion and repeats and extends Rouder's illustrative experiments, showing the sense in which optional stopping is indeed not a problem for Bayesians. Section 3.4 then contains additional simulations indicating the problems with default priors as summarized above. In Section 3.5 we discuss conceptualizations of 'handling optional stopping' that are different from Rouder's; this includes an explication of the purely subjective Bayesian viewpoint as well as an explication of a frequentist treatment of handling optional stopping, which only concerns sampling under the null hypothesis. We illustrate that some (not all!) Bayes factor methods can handle optional stopping in this frequentist sense. We conclude with a discussion of our findings in Section 3.6.

3.2 Bayesian probability and Bayes factors

Bayesianism is about a certain interpretation of the concept *probability*: as *degrees of belief*. Wagenmakers (2007) and Rouder (2014) give an intuitive explanation for the different views of frequentists and Bayesians in statistics, on the basis of coin flips. The frequentists interpret probability as a limiting frequency. Suppose we flip a coin many times, if the probability of heads is $3/4$, we see a proportion of $3/4$ of all those coin flips with heads up. Bayesians interpret probability as a degree of belief. If an agent believes the probability of heads is $3/4$, she believes that it will be 3 times more likely that the next coin flip will result in heads than tails; we return to the operational meaning of such a 'belief' in terms of betting in Section 3.5.

A Bayesian first expresses this belief as a probability function. In our coin flipping example, it might be that the agent believes that it is more likely that the coin is biased towards heads, which the probability function thus reflects. We call this the *prior distribution*, and we denote[1] it by $\mathbb{P}(\theta)$, where θ is the parameter (or several parameters) of the model. In our example, θ

[1] With some abuse of notation, we use $\mathbb{P}$ both to denote a generic probability distribution (defined on sets), and to denote its associated probability mass function and a probability density function (defined on elements of sets); whenever in this article we write $\mathbb{P}(z)$ where z takes values in a real-valued scalar or vector space, this should be read as $f(z)$ where f is the density of $\mathbb{P}$.

expresses the bias of the coin, and is a real number between 0 and 1. After the specification of the prior, we conduct the experiment and obtain the data D and the likelihood $\mathbb{P}(D|\theta)$. Now we can compute the *posterior distribution* $\mathbb{P}(\theta|D)$ with the help of *Bayes' theorem*:

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(D|\theta)\mathbb{P}(\theta)}{\mathbb{P}(D)}. \tag{3.1}$$

Rouder (2014) and Wagenmakers (2007) provide a clear explanation of Bayesian hypothesis testing with Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995), which we repeat here for completeness. Suppose we want to test a null hypothesis $\mathcal{H}_0$ against an alternative hypothesis $\mathcal{H}_1$. A hypothesis can consist of a single distribution, for example: 'the coin is fair'. We call this a *simple hypothesis*. A hypothesis can also consist of two or more, or even infinitely many hypotheses, which we call a *composite hypothesis*. An example is: 'the coin is biased towards heads', so the probability of heads can be any number between 0.5 and 1, and there are infinitely many of those numbers. Suppose again that we want to test $\mathcal{H}_0$ against $\mathcal{H}_1$. We start with the so called *prior odds*: $\mathbb{P}(\mathcal{H}_1)/\mathbb{P}(\mathcal{H}_0)$, our belief before seeing the data. Let's say we believe that both hypotheses are equally probable, then our prior odds are 1-to-1. Next we gather data D, and update our odds with the new knowledge, using Bayes' theorem (Eq. 3.1):

$$\texttt{post-odds}|D = \frac{\mathbb{P}(\mathcal{H}_1|D)}{\mathbb{P}(\mathcal{H}_0|D)} = \frac{\mathbb{P}(\mathcal{H}_1)}{\mathbb{P}(\mathcal{H}_0)}\frac{\mathbb{P}(D|\mathcal{H}_1)}{\mathbb{P}(D|\mathcal{H}_0)}. \tag{3.2}$$

The left term is called *posterior odds*, it is our updated belief about which hypothesis is more likely. Right of the prior odds, we see the *Bayes factor*, the term that describes how the beliefs (prior odds) are updated via the data. If we have no preference for one hypothesis and set the prior odds to 1-to-1, we see that the posterior odds are just the Bayes factor. If we test a composite $\mathcal{H}_0$ against a composite $\mathcal{H}_1$, the Bayes factor is a ratio of two likelihoods in which we have two or more possible values of our parameter θ. Bayesian inference tells us how to calculate $\mathbb{P}(D \mid \mathcal{H}_j)$: we integrate out the parameter with help of a prior distribution $\mathbb{P}(\theta)$, and we write Eq. (3.2) as:

$$\texttt{post-odds}|D = \frac{\mathbb{P}(\mathcal{H}_1|D)}{\mathbb{P}(\mathcal{H}_0|D)} = \frac{\mathbb{P}(\mathcal{H}_1)}{\mathbb{P}(\mathcal{H}_0)}\frac{\int_{\theta_1}\mathbb{P}(D|\theta_1)\mathbb{P}(\theta_1)\,\mathrm{d}\theta_1}{\int_{\theta_0}\mathbb{P}(D|\theta_0)\mathbb{P}(\theta_0)\,\mathrm{d}\theta_0} \tag{3.3}$$

where θ_0 denotes the parameter of the null hypothesis $\mathcal{H}_0$, and similarly, θ_1 is the parameter of the alternative hypothesis $\mathcal{H}_1$. If we observe a Bayes factor of 10, it means that the *change* in odds from prior to posterior in favor of the alternative hypothesis $\mathcal{H}_1$ is a factor 10. Intuitively, the Bayes factor provides a measure of whether the data have increased or decreased the odds on $\mathcal{H}_1$ relative to $\mathcal{H}_0$.

3.3 Handling Optional stopping in the Calibration Sense

Validity under optional stopping is a desirable property of hypothesis testing: we gather some data, look at the results, and decide whether we stop or gather some additional data. Informally we call 'peeking at the results to decide whether to collect more data' *optional stopping*, but if

we want to make more precise what it means if we say that a test can handle optional stopping, it turns out that different approaches (frequentist, subjective Bayesian and objective Bayesian) lead to different interpretations or definitions. In this section we adopt the definition of handling optional stopping that was used by Rouder, and show, by repeating and extending Rouder's original simulation, that Bayesian methods do handle optional stopping in this sense. In the next section, we shall then see that for 'default' and 'pragmatic' priors used in practice, Rouder's original definition may not always be appropriate — indicating there are problems with optional stopping after all.

3.3.1 Example 0: Rouder's example

We start by repeating Rouder's (2014) second example, so as to explain his ideas and re-state his results. Suppose a researcher wants to test the null hypothesis $\mathcal{H}_0$ that the mean of a normal distribution is equal to 0, against the alternative hypothesis $\mathcal{H}_1$ that the mean is not 0: we are really testing whether $\mu = 0$ or not. In Bayesian statistics, the composite alternative $\mathcal{H}_1 : \mu \neq 0$ is incomplete without specifying a prior on μ; like in Rouder's example, we take the prior on the mean to be a standard normal, which is a fairly standard (though by no means the only common) choice (Berger, 1985; Bernardo and Smith, 1994). This expresses a belief that small effect sizes are possible (though the prior probability of the mean being *exactly* 0 is 0), while a mean as large as 1.0 is neither typical nor exceedingly rare. We take the variance to be 1, such that the mean equals the effect size. We set our prior odds to 1-to-1: This expresses a priori indifference between the hypotheses, or a belief that both hypotheses are really equally probable. To give a first example, suppose we observe $n = 10$ observations Now we can observe the data and update our prior beliefs. We calculate the posterior odds, in our case equal to the Bayes factor, via Eq. (3.2) for data $D = (x_1, \ldots, x_n)$:

$$\texttt{post-odds}|x_1, \ldots, x_n = \frac{1}{1} \cdot \frac{\exp\left\{\frac{n^2\overline{x}^2}{2(n+1)}\right\}}{\sqrt{n+1}} \tag{3.4}$$

where n is the sample size (10 in our case), and $\overline{x}$ is the sample mean. Suppose we observe posterior odds of 3.5-to-1 in favor of the null.

Calibration, Mathematically As Rouder writes: 'If a replicate experiment yielded a posterior odds of 3.5-to-1 in favor of the null, then we expect that the null was 3.5 times as probable as the alternative to have produced the data.' In mathematical language, this can be expressed as

$$\texttt{post-odds}|\text{“}\texttt{post-odds}|x_1, \ldots, x_n = a\text{”} \;=\; a, \tag{3.5}$$

for the specific case $n = 10$ and $a = 1/3.5$; of course we would expect this to hold for general n and a. The quotation marks indicate that we condition on an event, i.e. a set of different data realizations; in our case this is the set of all data $x_1, \ldots, x_n$ for which the posterior odds are a. We say that (3.5) expresses *calibration of the posterior odds*. To explain further, we draw the analogy to weather forecasting: consider a weather forecaster who, on each day, announces the probability that it will rain the next day at a certain location. It is standard terminology to call such a weather forecaster *calibrated* if, on average on those days for which he predicts

'probability of rain is 30%', it rains about 30% of the time, on those days for which he predicts 40%, it rains 40% of the time, and so on. Thus, although his predictions presumably depend on a lot of data such as temperature, air pressure at various locations etc., given *only* the fact that this data was such that he predicts a, the actual probability is a. Similarly, given only the fact the posterior odds based on the full data are a (but not given the full data itself), the posterior odds should still be a (readers who find (3.5) hard to interpret are urged to study the simulations below).

Indeed, it turns out that (3.5) is the case. This can be shown either as a mathematical theorem, or, as Rouder does, by computer simulation. At this point, the result is merely a sanity check, telling us that Bayesian updating is not crazy, and is not really surprising. Now, instead of a fixed n, let us consider optional stopping: we keep adding data points until the the posterior odds are at least 10-to-1 for either hypothesis, unless a maximum of 25 data points was reached. Let τ be the sample size (which is now data-dependent) at which we stop; note that $\tau \leq 25$. Remarkably, it turns out that we still have

$$\texttt{post-odds} | \text{“}\texttt{post-odds} | x_1, \ldots, x_\tau = a\text{”} \; = \; a, \tag{3.6}$$

for this (and in fact any other data-dependent) stopping time τ. In words, *the posterior odds remain calibrated under optional stopping*. Again, this can be shown formally, as a mathematical theorem (we do so in Hendriksen, De Heide and Grünwald, 2020; see also Deng, Lu and Chen, 2016).

Calibration, Proof by Simulation Following Yu et al. (2014) and Sanborn and Hills (2014), Rouder uses computer simulations, rather than mathematical derivation, to elucidate the properties of analytic methods. In Rouder's words 'this choice is wise for a readership of experimental psychologists. Simulation results have a tangible, experimental feel; moreover, if something is true mathematically, we should be able to see it in simulation as well'. Rouder illustrates both (3.5) and (3.6) by a simulation which we now describe.

Again we draw data from the null hypothesis: say $n = 10$ observations from a normal distribution with mean 0 and variance 1. But now we repeat this procedure 20,000 times, and we see the distribution of the posterior odds plotted as the blue histogram on the log scale in Figure 3.1a. We also sample data from the alternative distribution $\mathcal{H}_1$: first we sample a mean from a standard normal distribution (readers that consider this 'sampling from the prior' to be strange are urged to read on), and then we sample 10 observations from a normal distribution with this just obtained mean, and variance 1. Next, we calculate the posterior odds from Eq. (3.4). Again, we perform 20,000 replicate experiments of 10 data points each, and we obtain the pink histogram in Figure 3.1a. We see that for the null hypothesis, most samples favor the null (the values of the Bayes factor are smaller than 1), for the alternative hypothesis we see that the bins for higher values of the posterior odds are higher.

In terms of this simulation, Rouder's claim that, 'If a replicate experiment yielded a posterior odds of 3.5-to-1 in favor of the null, then we expect that the null was 3.5 times as probable as the alternative to have produced the data', as formalized by (3.5), now says the following: if we look at a specific bin of the histogram, say at 3.5, i.e. the number of all the replicate experiments that yielded approximately a posterior odds of 3.5, then the bin from $\mathcal{H}_1$ should be about 3.5

times as high as the bin from $\mathcal{H}_0$. Rouder calls the ratio of the two histograms the *observed posterior odds*: the ratio of the binned posterior odds counts we observe from the simulation experiments we did. What we expect the ratio to be for a certain value of the posterior odds, is what he calls the *nominal posterior odds*. We can plot the observed posterior odds as a function of the nominal posterior odds, and we see the result in Figure 3.1b. The observed values agree closely with the nominal values: all points lie within simulation error on the identity line, which can be considered as a 'proof of (3.5) by simulation'.

Rouder (2014) repeats this experiment under optional stopping: he ran a simulation experiment with exactly the same setup, except that in each of the 40,000 simulations, sampling occurred until the posterior odds were at least 10-to-1 for either hypothesis, unless a maximum of 25 observations was reached. This yielded a figure indistinguishable from Figure 3.1b, from which Rouder concluded that 'the interpretation of the posterior odds holds with optional stopping'; in our language, *the posterior odds remain calibrated under optional stopping* — it is a proof, by simulation, that (3.6) holds. From this and similar experiments, Rouder concluded that Bayes factors still have a clear interpretation under optional stopping (we agree with this for what we call below Type 0 and I priors, not Type II), leading to the claim/title 'optional stopping is no problem for Bayesians' (for which we only agree for Type 0 and purely subjective priors).

Is sampling from the prior meaningful? When presenting Rouder's simulations to other researchers, a common concern is: 'how can sampling a parameter from the prior in $\mathcal{H}_1$ be meaningful? In any real-life experiment, there is just one, fixed population value, i.e. one fixed value of the parameter that governs the data.' This is indeed true, and not in contradiction with Bayesian ideas: Bayesian statisticians put a distribution on parameters in $\mathcal{H}_1$ that expresses their uncertainty about the parameter, and that should not be interpreted as something that is 'sampled' from. Nevertheless, Bayesian posterior odds calculations are done by calculating weighted averages via integrals, and the results are *mathematically equivalent* to what one gets if, as above, one samples a parameter from the prior, and the data from the parameter, and then takes averages over many repetitions. We (and Rouder) really want to establish (3.5) and (3.6) (which can be interpreted without resorting to sampling a parameter from a prior), and we note that it is equivalent to the curve in Figure 3.1b coinciding with the diagonal.

Some readers of an earlier draft of this paper concluded that, given its equivalence to an experiment involving sampling from the prior, which feels meaningless to them, (3.6) is itself invariably meaningless. Instead, they claim, because in real-life the parameter often has one specific fixed value, one should look at what happens under sampling under fixed parameter values. Below we shall see that if we look at such *strong calibration*, we sometimes (Example 1) still get calibration, but usually (Example 2) we do not; so such readers will likely agree with our conclusion that 'optional stopping can be a problem for Bayesians', even though they would disagree with us on some details, because we do think that (3.6) can be a meaningful statement for some, but not all priors. To us, the importance of the simulations is simply to verify (3.6) and, later on (Example 2), to show that (3.8), the stronger analogue of (3.6) that we would like to hold for default priors, does not always hold.

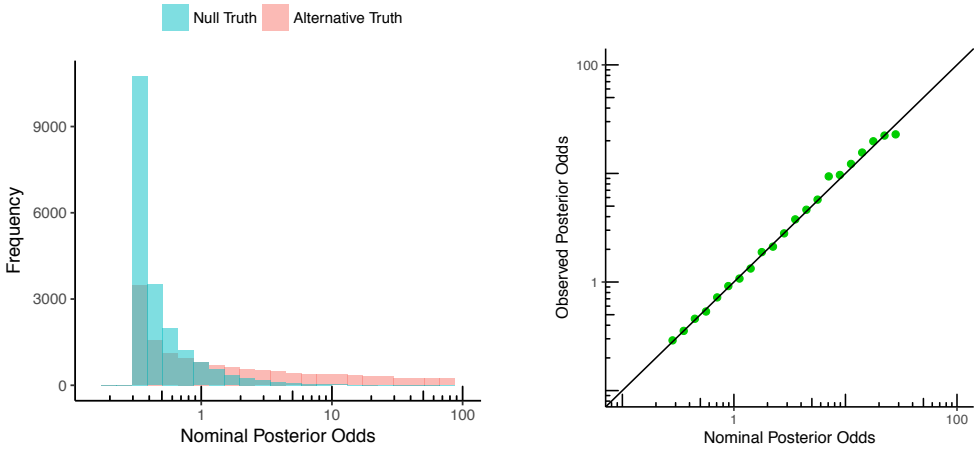


Figure 3.1: The interpretation of the posterior odds in Rouder's experiment, from 20,000 replicate experiments. (a) The empirical sampling distribution of the posterior odds as a histogram under $\mathcal{H}_0$ and $\mathcal{H}_1$. (b) Calibration plot: the observed posterior odds as a function of the nominal posterior odds.

3.3.2 Example 1: Rouder's example with a nuisance parameter

We now adjust Rouder's example to a case where we still want to test whether $\mu = 0$, but the variance σ^2 is unknown. Posterior calibration will still be obtained under optional stopping; the example mainly serves to gently introduce the notions of *improper prior* and *strong vs. prior calibration*, that will play a central role later on. So, $\mathcal{H}_0$ now expresses that the data are independently normally distributed with mean 0 and some unknown variance σ^2, and $\mathcal{H}_1$ expresses that the data are normal with variance σ^2, and some mean μ, where the uncertainty about μ is once again captured by a normal prior: the mean is distributed according to a normal with mean zero and variance (again) σ^2 (this corresponds to a standard normal distribution on the effect size). If $\sigma^2 = 1$, this reduces to Rouder's example; but we now allow for arbitrary σ^2. We call σ^2 a *nuisance parameter*: a parameter that occurs in both models, is not directly of interest, but that needs to be accounted for in the analysis. The setup is analogous to the standard 1-sample frequentist t-test, where we also want to test whether a mean is 0 or not, without knowing the variance; in the Bayesian approach, such a test only becomes defined once we have a prior for the parameters. For μ we choose a normal,[2] for the nuisance parameter σ we will make the standard choice of Jeffreys' prior for the variance: $\mathbb{P}_J(\sigma) := 1/\sigma$ (Rouder et al., 2009). To obtain the Bayes factor for this problem, we integrate out the parameter σ cf.

[2] The advantage of a normal is that it makes calculations relatively easy. A more common and perhaps more defensible choice is a Cauchy distribution, used in the 'default Bayesian t-test', which we consider further below.

Eq. (3.3). Again, we assign prior odds of 1-to-1, and obtain the posterior odds:

$$\begin{aligned}\texttt{post-odds}|D &= \frac{1}{1}\frac{\int_0^\infty \frac{1}{\sigma}\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x_i^2}{2\sigma^2}\right)\mathrm{d}\sigma}{\int_0^\infty \frac{1}{\sigma}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mu}{2\sigma^2}\right)\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\mathrm{d}\mu\,\mathrm{d}\sigma} \\ &= \frac{1}{\sqrt{n+1}}\left(1-\frac{\left(\frac{1}{n+1}\sum_{i=1}^n x_i\right)^2}{\frac{1}{n+1}\sum_{i=1}^n x_i^2}\right)^{-\frac{n}{2}}\end{aligned}$$

Formally, Jeffreys' prior on σ is a 'measure' rather than a distribution, since it does not integrate to 1: clearly

$$\int_0^\infty \mathbb{P}_J(\sigma)\,\mathrm{d}\sigma = \int_0^\infty \frac{1}{\sigma}\,\mathrm{d}\sigma = \infty, \tag{3.7}$$

Priors that integrate to infinity are often called *improper*. Use of such priors for nuisance parameters is not really a problem for Bayesian inference, since one can typically plug such priors into Bayes' theorem anyway, and this leads to proper posteriors, i.e. posteriors that do integrate to one, and then the Bayesian machinery can go ahead. Since Jeffreys' prior is meant to express that we have no clear prior knowledge about the variance, we would hope that Bayes would remain interpretable under optional stopping, no matter what the (unobservable) variance in our sampling distribution actually is. Remarkably, this is indeed the case: for all $\sigma_0^2 > 0$, we have the following analogue of (3.6):

$$\texttt{post-odds}|\sigma^2 = \sigma_0^2, \text{“}\texttt{post-odds}|x_1, \ldots, x_\tau = a\text{”} \;=\; a, \tag{3.8}$$

In words, this means that, given that the posterior odds (calculated based on Jeffreys' prior, i.e. without knowing the variance) are equal to a *and* that the actual variance is σ_0^2, the posterior odds are still a, irrespective of what σ_0^2 actually is. This statement may be quite hard to interpret, so we proceed to illustrate it by simulation again.

To repeat Rouder's experiment, we have to simulate data under both $\mathcal{H}_0$ and $\mathcal{H}_1$. To do this we need to specify the variance σ^2 of the normal distribution(s) from which we sample. Whereas, as in the previous experiment, we can sample the mean in $\mathcal{H}_1$ from the prior, for the variance we seem to run into a problem: it is not clear how one should sample from an improper prior. θ. But we cannot directly sample σ from an improper prior. As an alternative, we can pick any particular fixed σ^2 to sample from, as we now illustrate. Let us first try $\sigma^2 = 1$. Like Rouder's example, we sample the mean of the alternative hypothesis $\mathcal{H}_1$ from the aforementioned normal distribution. Then, we sample 10 data points from a normal distribution with the just sampled mean and the variance that we picked. For the null hypothesis $\mathcal{H}_0$ we sample the data from a normal distribution with mean zero and the same variance. We continue the experiment just as Rouder did: we calculate the posterior odds from 20,000 replicate experiments of 10 generated observations for each hypothesis, and construct the histograms and the plot of the ratio of the counts to see if calibration is violated. In Figure 3.2a we see the calibration plot for the experiment described above. In Figure 3.2b we see the results for the same experiment, except that we performed optional stopping: we sampled until the posterior odds were at least 10-to-1 for $\mathcal{H}_1$, or the maximum of 25 observations was reached. We see that the posterior odds in this experiment with optional stopping are calibrated as well.

Prior Calibration vs. Strong Calibration Importantly, the same conclusion remains valid whether we sample data using $\sigma^2 = 1$, or $\sigma^2 = 2$, or any other value — in simulation terms (3.8) simply expresses that we get calibration (i.e. all points on the diagonal) no matter what σ^2 we actually sample from: even though calculation of the posterior odds given a sample makes use of the prior $\mathbb{P}_J(\sigma) = 1/\sigma$ and does not know the 'true' σ, calibration is retained under sampling under arbitrary 'true' σ. We say that the posterior odds are *prior-calibrated* for parameter μ and *strongly calibrated* for σ^2. More generally and formally, consider general hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ (not necessarily expressing that data are normal) that share parameters γ_0, γ_1 and suppose that (3.8) holds with γ_1 in the role of σ^2. Then we say that γ_0 is prior-calibrated (to get calibration in simulations we need to draw it from the prior) and γ_1 is strongly calibrated (calibration is obtained when drawing data under all possible γ_1).

Notably, strong calibration is a special property of the chosen prior. If we had chosen another proper or improper prior to calculate the posterior odds (for example, the improper prior $\mathbb{P}'(\sigma) \propto \sigma^{-2}$ has sometimes been used in this context) then the property that calibration under optional stopping is retained under any choice of σ^2 will cease to hold; we will see examples below. The reason that $\mathbb{P}_J(\sigma) \propto 1/\sigma$ has this nice property is that σ is a special type of nuisance parameter for which there exists a suitable group structure, relative to which both models are invariant (Eaton, 1989; Berger, Pericchi and Varshavsky, 1998; Dass and Berger, 2003). This sounds more complicated than it is — in our example, the invariance is scale invariance: if we divide all outcomes by any fixed σ (multiply by $1/\sigma$), then the Bayes factor remains unchanged; similarly, one may have for example location invariances.

If such group structure parameters are equipped with a special prior (which, for reasons to become clear, we shall term *Type o prior*), then we obtain strong calibration, both for fixed sample sizes and under optional stopping, relative to these parameters.[3] Jeffreys' prior for the variance $\mathbb{P}_J(\sigma)$ is the Type o prior for the variance nuisance parameter. Dass and Berger (2003) show that such priors can be defined for a large class of nuisance parameters — we will see the example of a prior on a common mean rather than a variance in Example 3 below; but there also exist cases with parameters that (at least intuitively) are nuisance parameters, for which Type o priors do not exist; we give an example in Appendix 3.A. For parameters of interest, including e.g. any parameter that does not occur in both models, Type o priors never exist.

3.4 When Problems arise: Subjective versus Pragmatic and Default Priors

Bayesians view probabilities as degree of belief. The degree of belief an agent has before conducting the experiment, is expressed as a probability function. This *prior* is then updated with data from experiments, and the resulting *posterior* can be used to base decisions on. For one pole of the spectrum of Bayesians, the pure *subjectivists*, this is the full story (De Finetti, 1937; Savage, 1954): any prior capturing the belief of the agent is allowed, but it should always be

[3] Technically, the Type o prior for a given group structure is defined as the right-Haar prior for the group (Berger, Pericchi and Varshavsky, 1998): a unique (up to a constant) probability measure induced on the parameter space by the right Haar measure on the related group. Strong calibration is proven in general by Hendriksen, De Heide and Grünwald, 2020, and Hendriksen, 2017 for the special case of the 1-sample t-test.

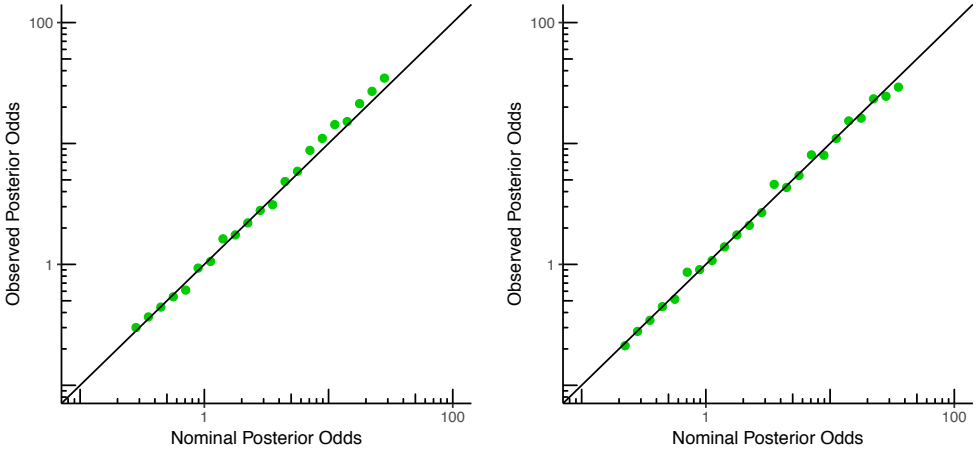


Figure 3.2: Calibration of the experiment of Section 3.3.2 from 20,000 replicate experiments. (a) The observed posterior odds as a function of the nominal posterior odds. (b) The observed posterior odds as a function of the nominal posterior odds with optional stopping.

interpreted as the agent's personal degree of belief; in Section 3.5 we explain what such a 'belief' really means. On the other end of the spectrum, the *objective Bayesians* (Jeffreys, 1961; Berger, 2006) argue that degrees of belief should be restricted, ideally in such a way that they do not depend on the agent, and in the extreme case boil down to a single, rational, probability function, where a priori distributions represent indifference rather than subjective belief and a posteriori distributions represent 'rational degrees of confirmation' rather than subjective belief. Ideally, in any given situation there should then just be a single appropriate prior. Most objective Bayesians do not take such an extreme stance, recommending instead *default* priors to be used whenever only very little a priori knowledge is available. These make a *default* choice for the functional form of a distribution (e.g. Cauchy) but often have one or two parameters that can be specified in a subjective way. These may then be replaced by more informative priors when more knowledge becomes available after all. We will see several examples of such default priors below.

So what category of priors is used in practice? Recent papers that advocate the use of Bayesian methods within psychology such as Rouder et al. (2009), Rouder et al. (2012) and Jamil et al. (2016) are mostly based on default priors. Within the statistics community, nowadays a pragmatic stance is by far the most common, in which priors are used that mix 'default' and 'subjective' aspects (Gelman, 2017) and that are also chosen to allow for computationally feasible inference. Very broadly speaking, we may say that there is a scale ranging from completely 'objective' (and hardly used) via 'default' (with a few, say 1 or 2 parameters to be filled in subjectively, i.e. based on prior knowledge) and 'pragmatic' (with functional forms of the prior based partly on prior knowledge, partly by defaults, and partly by convenience) to the fully subjective. Within the pragmatic stance, one explicitly acknowledges that one's prior distribution may have some arbitrary aspects to it (e.g. chosen to make computations easier rather than

reflecting true prior knowledge). It then becomes important to do sensitivity analyses: studying what happens if a modified prior is used or if data are sampled not by first sampling parameters θ from the prior and then data from $\mathbb{P}(\cdot \mid \theta)$ but rather directly from a fixed θ within a region that does not have overly small prior probability.[4]

The point of this article is that Rouder's view on what constitutes 'handling optional stopping' is tailored towards a fully subjective interpretation of Bayes; as soon as one allows default and pragmatic priors, problems with optional stopping do occur (except for what we call Type 0 priors). We can distinguish between three types of problems, depending on the type of prior that is used. We now give an overview of type of prior and problem, giving concrete examples later.

1. *Type 0 Priors:* these are priors on parameters freely occurring in both hypotheses for which strong calibration (as with σ^2 in (3.8)) holds under optional stopping. This includes all right Haar priors on parameters that satisfy a group structure; Hendriksen, De Heide and Grünwald (2020) give a formal definition; Dass and Berger (2003) and Berger, Pericchi and Varshavsky (1998) give an overview of such priors. We conjecture, but have no proof, that such right Haar priors on group structure parameters are the *only* priors allowing for strong calibration under optional stopping, i.e. the only Type 0 Priors. Some, but not all so-called 'nuisance parameters' admit group structure/right Haar priors. For example, the variance in the t-test setting does, but the mean in 2×2 contingency tables (Appendix 3.A) does not.

2. *Type I Priors:* these are default or pragmatic priors that do *not* depend on any aspects of the experimental setup (such as the sample size) or the data (such as the values of covariates) and are not of Type 0 above. Thus, strong calibration under optional stopping is violated with such priors — an example is the Cauchy prior in Example 2 of Section 3.4.1 below.

3. *Type II Priors:* these are default and pragmatic priors that are not of Type 0 or I: the priors may themselves depend on the experimental setup, such as the sample size, the covariates (design), or the stopping time itself, or other aspects of the data. Such priors are quite common in the Bayesian literature. Here the problem is more serious: as we shall see, prior calibration is ill-defined, and correspondingly Rouder's experiments cannot be performed for such priors, and 'handling optional stopping' is in a sense impossible in principle. An example is the g-prior for regression as in Example 3 below or Jeffreys' prior for the Bernoulli model as in Section 3.4.3 below.

We illustrate the problems with Type I and Type II priors by further extending Rouder's experiment to two extensions of our earlier setting, namely the Bayesian t-test, going back to Jeffreys (1961) and advocated by Rouder et al. (2009), and objective Bayesian linear regression, following Liang et al. (2008). Both methods are quite popular and use default Bayes factors based on default priors, to be used when no clear or very little prior knowledge is readily available.

[4] To witness, one of us recently spoke at the bi-annual OBAYES (Objective Bayes) conference, and noticed that a substantial fraction of the talks featured such fixed θ-analyses and/or used priors of Type II below.

3.4.1 Example 2: Bayesian t-test — The Problem with Type I Priors

Suppose a researcher wants to test the effect of a new fertilizer on the growth of some wheat variety. The null hypothesis $\mathcal{H}_0$ states that there is no difference between the old and the new fertilizer, and the alternative hypothesis $\mathcal{H}_1$ states that the fertilizers have a different effect on the growth of the wheat. We assume that the length of the wheat is normally distributed with the same (unknown) variance under both fertilizers, and that with the old fertilizer, the mean is known to be $\mu_0 = 1$ meter. We now take a number of seeds and apply the new fertilizer to each of them. We let the wheat grow for a couple of weeks, and we measure the lengths. The null hypothesis $\mathcal{H}_0$ is thus: $\mu = \mu_0 = 1$, and the alternative hypothesis $\mathcal{H}_1$ is that the mean of the group with the new fertilizer is different from 1 meter: $\mu \neq 1$.

Again we follow Rouder's calibration check; again, the end goal is to illustrate a mathematical result, (3.9) below, which will be contrasted with (3.6). And again, to make the result concrete, we will first perform a simulation, generating data from both models and updating our prior beliefs from this data as before. We do this using the *Bayesian t-test*, where Jeffreys' prior $\mathbb{P}_J(\sigma) = 1/\sigma$ is placed on the standard deviation σ within both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. Within $\mathcal{H}_0$ we set the mean to $\mu_0 = 1$ and within $\mathcal{H}_1$, a standard Cauchy prior is placed on the effect size $(\mu - \mu_0)/\sigma$; details are provided by Rouder et al. (2009). Once again, the nuisance parameter σ is equipped with an improper Jeffreys' prior, so, like in Experiment 1 above and for the reasons detailed there, for simulating our data, we will choose a fixed value for σ; the experiments will give the same result regardless of the value we choose.

We generate 10 observations for each fertilizer under both models: for $\mathcal{H}_0$ we sample data from a normal distribution with mean $\mu_0 = 1$ meter and we pick the variance $\sigma^2 = 1$. For $\mathcal{H}_1$ we sample data from a normal distribution where the variance is 1 as well, and the mean is determined by the effect size above. We adopt a Cauchy prior to express our beliefs about what values of the effect size are likely, which is mathematically equivalent to the effect size being sampled from a standard Cauchy distribution. We follow Rouder's experiment further, and set our prior odds on $\mathcal{H}_0$ and $\mathcal{H}_1$, before observing the data, to 1-to-1. We sample 10 data points from each of the hypotheses, and we calculate the Bayes factors. We repeat this procedure 20000 times. Then, we bin the 20000 resulting Bayes factors and construct a histogram. In Figure 3.3a we see the distribution of the posterior odds when either the null or the alternative are true in one figure. In Figure 3.3b we see the calibration plot for this data from which Rouder checks the interpretation of the posterior odds: the observed posterior odds is the ratio of the two histograms, where the width of the bins is 0.1 on the log scale. The posterior odds are calibrated, in accordance with Rouder's experiments. We repeated the experiment with the difference that in each of the 40,000 experiments we sampled more data points until the posterior odds were at least 10-to-1, or the maximum number of 25 data points was reached. The histograms for this experiment are in Figure 3.3c. In Figure 3.3d we can see that, as expected, the posterior odds are calibrated under optional stopping as well.

Since σ^2 is a nuisance parameter equipped with its Type o prior, it does not matter what value we take when sampling data. We may ask ourselves what happens if, similarly, we fix particular values of the mean and sample from them, rather than from the prior; for sampling from $\mathcal{H}_0$, this does not change anything since the prior is concentrated on the single point $\mu_0 = 1$; in $\mathcal{H}_1$, this means we can basically pick any μ and sample from it. In other words, we will check

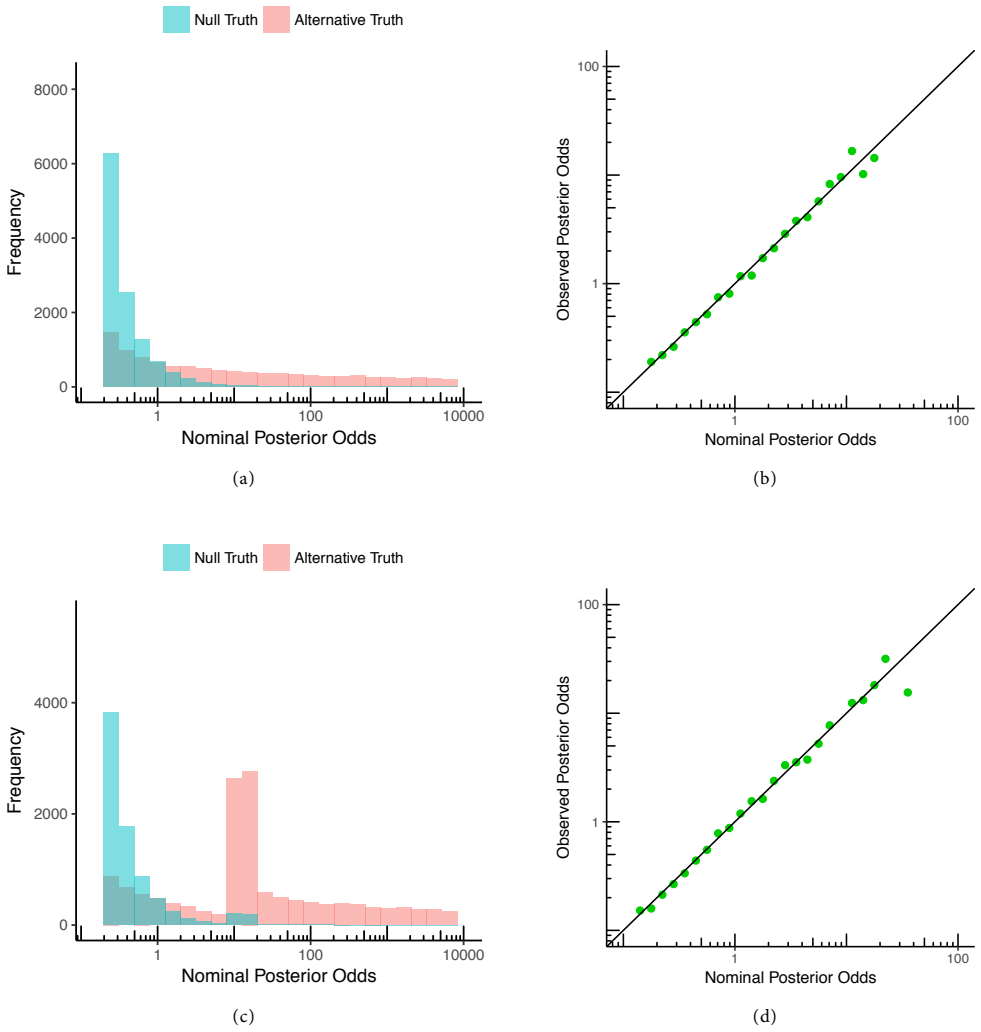


Figure 3.3: Calibration in the t-test experiment, Section 3.4.1, from 20,000 replicate experiments. (a) The distribution of posterior odds as a histogram under $\mathcal{H}_0$ and $\mathcal{H}_1$ in one figure. (b) The observed posterior odds as a function of the nominal posterior odds. (c) Distribution of the posterior odds with optional stopping. (d) The observed posterior odds as a function of the nominal posterior odds with optional stopping.

whether we have strong calibration rather than prior-calibration not just for σ^2, but also for the mean μ. We now first describe such an experiment, and will explain its importance further below.

We generate 10 observations under both models. The mean length of the wheat is again set to be 1 meter with the old fertilizer, and now we pick a particular value for the mean length of the wheat with the new fertilizer: 130 centimeters. For the variance, we again pick $\sigma^2 = 1$. We continue to follow Rouder's experiment and set our prior odds on $\mathcal{H}_0$ and $\mathcal{H}_1$, before observing the data, to 1-to-1. We sample 20,000 replicate experiments with 10 + 10 observations each, 10 from one of the hypotheses (normal with mean 1 for $\mathcal{H}_0$) and 10 from the other (normal with mean $\mu = 1.3$ for $\mathcal{H}_1$), and we calculate the Bayes factors. In Figure 3.4a we see that calibration is, to some extent, violated: the points follow a line that is still approximately, but now not precisely, a straight line. Now what happens in this experiment under optional stopping? We repeated the experiment with the difference that we sampled more data points until the posterior odds were at least 10-to-1, or the maximum number of 25 data points was reached. In Figure 3.4b we see the results: calibration is now violated significantly — when we stop early the nominal posterior odds (on which our stopping rule was based) are on average significantly higher than the actual, observed posterior odds. We repeated the experiment with various choices of μ's within $\mathcal{H}_1$, invariably getting similar results.[5] In mathematical terms, this illustrates that when the stopping time τ is determined by optional stopping, then, for many a and μ',

$$\texttt{post-odds}|\mu = \mu', \text{“}\texttt{post-odds}|x_1, \ldots, x_\tau = a\text{” is very different from } a, \tag{3.9}$$

We conclude that strong calibration for the parameter of interest μ is violated somewhat for fixed sample sizes, but much more strongly under optional stopping. We did similar experiments for a different model with discrete data (see Appendix 3.A), once again getting the same result. We also did experiments in which the means of $\mathcal{H}_1$ were sampled from a different prior than the Cauchy: this also yielded plots which showed violation of calibration. Our experiments are all based on a one-sample t-test; experiments with a two-sample t-test and ANOVA (also with the same overall mean for both $\mathcal{H}_0$ and $\mathcal{H}_1$) yielded severe violation of strong calibration under optional stopping as well.

The Issue Why is this important? When checking Rouder's prior-based calibration, we sampled the effect size from a Cauchy distribution, and then we sampled data from the realized effect size. We repeated this procedure many times to approximate the distribution on posterior odds by a histogram analogous to that in Figure 3.1a. But do we really believe that such a histogram, based on the Cauchy prior, accurately reflects our beliefs about the data? The Cauchy prior was advocated by Jeffreys for the effect size corresponding to a location parameter μ because it has some desirable properties in hypothesis testing, i.e. when comparing two models (Ly, Verhagen and Wagenmakers, 2016). For estimating a one-dimensional location parameter directly, Jeffreys (like most objective Bayesians) would advocate an improper uniform prior on μ. Thus, objective Bayesians may *change their prior depending on the inference task of interest*,

[5] Invariably, strong calibration is violated both with and without optional stopping. In the experiments without optional stopping, the points still lie on an increasing and (approximately) straight line; the extent to which strong calibration is violated — the slope of the straight line — depends on the effect size. In the experiments with optional stopping, strong calibration is violated more strongly in the sense that the points do not follow a straight line anymore.

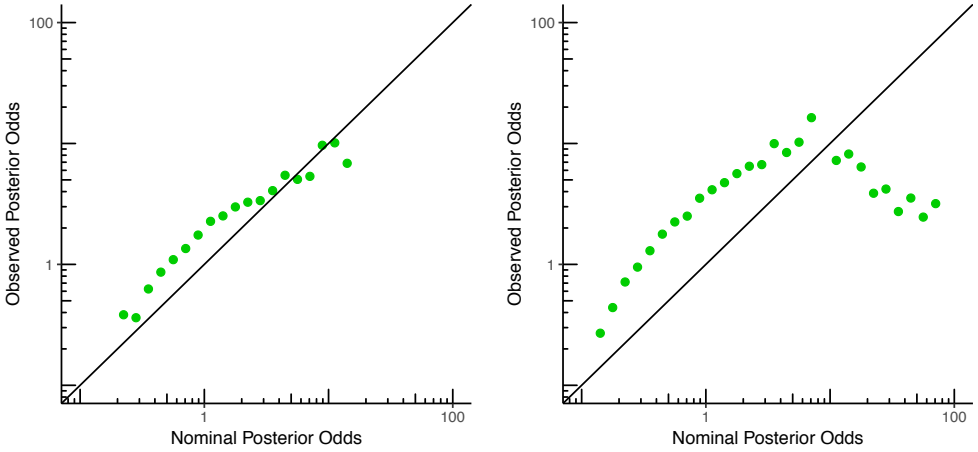


Figure 3.4: Calibration in the t-test experiment with fixed values for the means of $\mathcal{H}_0$ and $\mathcal{H}_1$ (Section 3.4.1, from 40,000 replicate experiments). (a) The observed posterior odds as a function of the nominal posterior odds. (b) The observed posterior odds as a function of the nominal posterior odds with optional stopping.

even when they are dealing with data representing the same underlying phenomenon. It does then not seem realistic to study what happens if data are sampled from the prior; *the prior is used as a* tool *in inferring likely parameters or hypotheses, and not to be thought of as something that prescribes how actual data will arise or tend to look like*. This is the first reason why it is interesting to study not just prior calibration, but also strong calibration for the parameter of interest. One might object that the sampling from the prior done by Rouder, and us, was only done to illustrate the mathematical expression (3.6); perhaps sampling from the prior is not realistic but (3.6) is still meaningful? We think that, because of the mathematical equivalence, it does show that the relevance of (3.6) is questionable as soon as we use default priors.

Prior calibration in terms of (3.6) — which indeed still holds[6] — *would* be meaningful if a Cauchy prior really described our prior beliefs about the data in the subjective Bayesian sense (explained in Section 3.5). But in this particular setup, the Cauchy distribution is highly unrealistic: it is a heavy tailed distribution, which means that the probability of getting very large values is not negligible, and it is very much higher than with, say, a Gaussian distribution. To make the intuition behind this concrete, say that we are interested in measuring the height of a type of corn that with the old fertilizer reaches on average 2 meters. The probability that a new fertilizer would have a mean effect of 6 meters or more under a standard Cauchy distribution would be somewhat larger than one in twenty. For comparison: under a standard Gaussian, this is as small as $9.87 \cdot 10^{-10}$. Do we really believe that it is quite probable (more than one in twenty) that the fertilizer will enable the corn to grow to 8 meters *on average*? Of course we could use a Cauchy with a different spread, but which one? Default Bayesians have emphasized that such choices should be made subjectively (i.e. based on informed prior guesses), but whatever value

[6] Note though that strong calibration still fails.

one choices, the chosen functional form of the prior (a Cauchy has, e.g., no variance) severely restricts the options, making any actual choice to some extent arbitrary. While growing crops (although a standard example in this context) may be particularly ill-suited to be modeled by heavy-tailed distributions, the same issue will arise with many other possible applications for the default Bayesian t-test: one will be practically sure that the effect size will not exceed certain values (not too large, not too small, certainly not negative), but it may be very hard to specify exactly which values. As a purely objective Bayesian, this need not be such a big problem - one resorts to the default prior and uses it anyway; but one has to be aware that in that case, sampling from the prior — as done by Rouder — is not meaningful anymore, since the data one may get may be quite atypical for the underlying process one is modeling.

In practice, most Bayesians are pragmatic, striking a balance between 'flat', 'uninformative' priors, prior knowledge and ease of computation. In the present example, they might put a Gaussian prior with mean μ on the effect size instead, truncated at 0 to avoid negative means. But then there is the question what variance this Gaussian should have — as a pragmatic Bayesian, one has to acknowledge that there will always be arbitrary or 'convenience' aspects about one's priors. This is the second reason why it is interesting to study not just prior calibration, but also strong calibration for the parameter of interest.

Thus, both from a purely objective and from a pragmatic Bayesian point of view, strong calibration is important. Except for nuisance parameters with Type 0 priors, we cannot expect it to hold precisely (see Gu, Hoijtink and Mulder, 2016 for a related point) — but this is fine; like with any sensitivity or robustness test, we acknowledge that our prior is imperfect and we merely ask that our procedure remains reasonable, not perfect. And we see that by and large this is the case if we use a fixed sample size, but not if we perform optional stopping. In our view this indicates that for pragmatic Bayesians using default priors, there is a real problem with optional stopping after all. However, within the taxonomy defined above, we implicitly used Type I priors (Cauchy) here. Default priors are often of Type II, and then, as we will see, the problems get significantly worse.

As a final note, we note that in our strong calibration experiment, we chose parameter values here which we deemed 'reasonable', by this we mean values which reside in a region of large prior density — i.e. we sampled from μ that are not too far from μ_0. Sampling from μ in the tails of the prior would be akin to 'really disbelieving our own prior', and would be asking for trouble. We repeated the experiment for many other values of μ not too far from μ_0 and always obtained similar results. Whether our choices of μ are truly reasonable is of course up to debate, but we feel that the burden of proof that our values are 'unreasonable' lies with those who want to show that Bayesian methods can deal with optional stopping even with default priors.

3.4.2 Example 3: Bayesian linear regression and Type II Priors

We further extend the previous example to a setting of linear regression with fixed design. We employ the default Bayes factor for regression from the `R` package `Bayesfactor` (Morey and Rouder, 2015), based on Liang et al. (2008) and Zellner and Siow (1980), see also Rouder and Morey (2012). This function uses as default prior Jeffreys' prior for the intercept μ and the variance ($\mathbb{P}_J(\mu, \sigma) \sim 1/\sigma$), and a mixture of a normal and an inverse-gamma distribution for

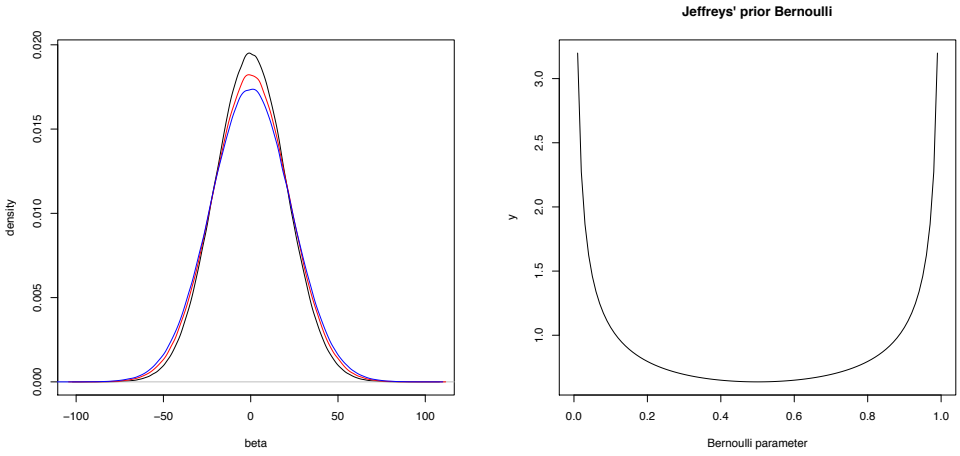
the regression coefficients, henceforth *g-prior*:

$$\begin{aligned} y &\sim \mathrm{N}\left(\mu + X\beta, \sigma^2\right), \\ \beta &\sim \mathrm{N}\left(0, g\sigma^2 n(X'X)^{-1}\right), \\ g &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{\sqrt{2}}{8}\right). \end{aligned} \tag{3.10}$$

Since the publication of Liang et al. (2008), this prior has become very popular as a default prior in Bayesian linear regression. Again we provide an example concerning the growth of wheat. Suppose a researcher wants to investigate the relationship between the level of a fertilizer, and the growth of the crop. We can model this experiment by linear regression with fixed design. We add different levels of the fertilizer to pots with seeds: the first pot gets a dose of 0.1, the second 0.2, ans so on up to the level 2. These are the x-values (covariates) of our simulation experiment. If we would like to repeat the examples of the previous sections and construct the calibration plots, we can generate the y-values — the increase or decrease in length of the wheat from the intercept μ — according to the proposed priors in Eq. (3.10). First we draw a g from an inverse gamma distribution, then we draw a β from the normal prior that we construct with the knowledge of the x-values, and we compute each y_i as the product of β and x_i plus Gaussian noise.

As we can see in Equation 3.10, the prior on β contains a scaling factor that depends on the experimental set-up — while it does not directly depend on the observations (y-values), it does depend on the design/covariates (x-values). If there is no optional stopping, then for a pragmatic Bayesian, the dependency on the x-values of the data is convenient to achieve appropriate scaling; it poses no real problems, since the whole model is conditional on X: the levels of fertilizer we administered to the plants. But under optional stopping, the dependency on X does become problematic, *for it is unclear which prior she should use!* If initially a design with 40 pots was planned (after each dose from 0.1 up to 2, another row of pots, one for each dose is added), but after adding three pots to the original twenty (so now we have two pots with the doses 0.1, 0.2 and 0.3, and one with each other dose), the researcher decides to check whether the results already are interesting enough to stop, should she base her decision on the posterior reached with prior based the initially planned design with 40 pots, or the design at the moment of optional stopping with 23 pots? This is not clear, and it does make a difference, since the g-prior changes as more x-values become available. In Figure 3.5a we see three g-priors on the regression coefficient β for the same fixed value of g, the same x-values as described in the fertilizer experiment above, but increasing sample size. First, each dose is administered to one plant, yielding the black prior distribution for β. Next, 3 plants are added to the experiment, with doses 0.1, 0.2 and 0.3, yielding the red distribution: wider and less peeked, and lastly, another 11 plants are added to the experiment, yielding the blue distribution which puts even less prior mass close to zero.

This problem may perhaps be pragmatically 'solved' in practice in two ways: either one could, as a rule, base the decision to stop at sample size n always using the prior for the given design at sample size n; or one could, as a rule, always use the design for the maximum sample size available. It is very unclear though whether there is any sense in which any of these two (or other) solutions 'handle optional stopping' convincingly. In the first case, the notion of



(a) (b)

Figure 3.5: Default priors that depend on aspects of the experimental setup: (a) G-priors for the regression example of Section 3.4.2 with different sample sizes: $n = 20$ (black), $n = 23$ (red) and $n = 34$ (blue). (b) Jeffreys' prior for the Bernoulli model for the specific case that n is fixed in advance (no optional stopping): a Beta $(1/2, 1/2)$ distribution.

prior calibration is ill-defined, since $\texttt{post-odds}|x_1, \ldots, x_\tau$ in (3.6) is ill-defined (if one tried to illustrate (3.6) by sampling, the procedure would be undefined since one would not know what prior to sample from until after one has stopped); in the second, one can perform it (by sampling β from the prior based on the design at the maximum sample size), but it seems rather meaningless, for if, for some reason or other, even more data were to become available later on, this would imply that the earlier sampled data were somehow 'wrong' and would have to be replaced.

What, then, about strong calibration? Fixing particular, 'reasonable' values of β does seem meaningful in this regression example. However (figures omitted), when we pick reasonable values for β instead of sampling β from the prior, we obtain again the conclusion that strong calibration is, on one hand, violated significantly under optional stopping (where the prior used in the decision to stop can be defined in either of the two ways defined above); but on the other hand, only violated mildly for fixed sample size settings. Using the taxonomy above, we conclude that optional stopping is a significant problem for Bayesians with Type-II priors.

3.4.3 Discrete Data and Type-II Priors

Now let us turn to discrete data: we test whether a coin is fair or not. The data D consist of a sequence of n_1 ones and n_0 zeros. Under $\mathcal{H}_0$, the data are i.i.d. Bernoulli$(1/2)$; under $\mathcal{H}_1$ they can be Bernoulli(θ) for any $0 \leq \theta \leq 1$ except $1/2$, θ representing the bias of the coin. One standard objective and default Bayes method (in this case coinciding with an *MDL (Minimum Description Length) method*, (Grünwald, 2007)) is to use Jeffreys' prior for the Bernoulli model

within $\mathcal{H}_1$. For fixed sample sizes, this prior is proper, and is given by

$$\mathbb{P}_J(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}} \cdot \frac{1}{\pi}, \tag{3.11}$$

where the factor $1/\pi$ is for normalization; see Figure 3.5b. If we repeat Rouder's experiment, and sample from this prior, then the probability that we would pick an extreme θ, within 0.01 of either 1 or 0, would be about 10 times as large as the probability that we would pick a θ within the equally wide interval $[0.49, 0.51]$. But, lacking real prior knowledge, do we really believe that such extreme values are much more probable than values around the middle? Most people would say we do not: under the subjective interpretation, i.e. if one really believes one's prior in the common interpretation of 'belief' given in Section 3.5, then such a prior would imply a willingness to bet at certain stakes. Jeffreys' prior is chosen in this case because it has desirable properties such as invariance under reparameterization and good frequentist properties, but not because it expresses any 'real' prior belief about some parameter values being more likely than others. This is reflected in the fact that in general, it depends on the stopping rule. Using the general definition of Jeffreys' prior (see e.g. Berger (1985)), we see, for example, that in the Bernoulli model, if the sample size is not fixed in advance but depends on the data (for example, we stop sampling as soon as three consecutive 1s are observed), then, as a simple calculation shows, Jeffreys' prior changes and even becomes improper (Jordan, 2010).

In Appendix 3.A we give another example of a common discrete setting, namely the 2×2 contingency table. Here the null hypothesis is a Bernoulli model and its parameter θ is intuitively a nuisance parameter, and thus strong calibration relative to this parameter would be especially desirable. However, the Bernoulli model does not admit a group structure, and hence neither Jeffreys' nor any other prior we know of can serve as a Type 0 prior, and strong calibration can presumably not be attained — the experiments show that it is certainly not attained if the default Gunel and Dickey Bayes factors (Jamil et al., 2016) are used (these are Type-II priors, so we need to be careful about what prior to use in the strong calibration experiment; see Appendix 3.A for details).

3.5 Other Conceptualizations of Optional Stopping

We have seen several problems with optional stopping under default and pragmatic priors. Yet it is known from the literature that, in some senses, optional stopping is indeed no problem for Bayesians (Lindley, 1957; Savage, 1954; Edwards, Lindman and Savage, 1963; Good, 1991). What then, is shown in those papers? Interestingly, different authors show different things; we consider them in turn.

3.5.1 Subjective Bayes optional stopping

The Bayesian pioneers Lindley (1957) and Savage (1954) consider a purely subjective Bayesian setting, appropriate if one truly believes one's prior (and at first sight completely disconnected from strong calibration — but see the two quotations further below). But what does this mean? According to De Finetti, one of the two main founding fathers of modern, subjective Bayesian

statistics, this implies a willingness to bet at small stakes, at the odds given by the prior.[7] For example, a subjective Bayesian who would adopt Jeffreys' prior $\mathbb{P}_J$ for the Bernoulli model as given by (3.11) would be willing to accept a gamble that pays off when the actual parameter lies close to the boundary, since the corresponding region has substantially higher probability, cf. the discussion underneath Eq. (3.11). For example, a gamble where one wins 11 cents if the actual Bernoulli parameter is in the set $[0, 0.01] \cup [0.99, 1]$ and pays 100 cents if it is in the set $[0.49, 0.51]$ and neither pays nor gains otherwise would be considered acceptable[8] because this gamble has positive expected gain under $\mathbb{P}_J$. We asked several Bayesians who are willing to use Jeffreys' prior for testing whether they would also be willing to accept such a gamble; most said no, indicating that they do not interpret Jeffreys prior the way a subjective Bayesian would.[9]

Now, if one adopts priors one really believes in in the above gambling sense, then it is easy to show that Bayesian updating from prior to posterior is not affected by the employed stopping rule; one ends up with the same posterior if one had decided the sample size n in advance or if it had been determined, for example, because one was satisfied with the results at this n. In this sense a subjective Bayesian procedure does not depend on the stopping rule (as we have seen, this is certainly not the case in general for default Bayes procedures). This is the main point concerning optional stopping of Lindley (1957), also made by e.g. Savage (1954) and Bernardo and Smith (1994), among many others. A second point made by Lindley (1957, p. 192) is that the decisions a Bayesian makes will "not, *on average*, be in error, when ignoring the stopping rule". Here the "average" is really an expectation obtained by integrating θ over the prior, and then the data D over the distribution $\mathbb{P}(D \mid \theta)$, making this claim very similar to prior calibration (3.6) — once again, the claim is correct, but works only if one believes that sampling (or taking averages over) the prior gives rise to data of the type one would really expect; and if one would not be willing to bet based on the prior in the above sense, it indicates that perhaps one doesn't really expect that data after all.

We cannot resist to add here that, while for a subjective Bayesian, prior-based calibration is sensible, even the founding fathers of subjective Bayes gave a warning against taking such a prior too seriously:[10]

> " Subjectivists should feel obligated to recognize that any opinion (so much more the initial one) is only vaguely acceptable... So it is important not only to know the exact answer for an exactly specified initial problem, *but what happens changing in a reasonable neighborhood the assumed initial opinion*" De Finetti, as quoted by Dempster (1975). — note that when we checked for strong calibration, we took

[7] Savage, the other father, employs a slightly different conceptualization in terms of preference orderings over outcomes, but that need not concern us here.

[8] One might object that actual Bernoulli parameters are never revealed and arguably do not exist; but one could replace the gamble by the following essentially equivalent gamble: a possibly biased coin is tossed 10,000 times, but rather than the full data only the average number of 1s will be revealed. If it is in the set $[0, 0.01] \cup [0.99, 1]$ one gains 11 cents and if it is in the set $[0.49, 0.51]$ one pays 100 cents. If one really believes Jeffreys' prior, this gamble would be considered acceptable.

[9] Another example is the Cauchy prior with scale one on the standardized effect size (Rouder et al., 2012), as most would agree that this is not realistic in psychological research. Thanks to an anonymous reviewer for pointing this out.

[10] Many thanks to Chris Holmes for bringing these quotations to our attention.

> parameter values μ which were not too unlikely under the prior, which one may perhaps view as 'a reasonable neighborhood of the initial opinion'.
>
> " ...in practice the theory of personal probability is supposed to be an idealization of one's own standard of behavior; the idealization is often imperfect in such a way that an aura of vagueness is attached to many judgments of personal probability..." (Savage, 1954).

Hence, one would expect that even a subjectivist would be interested in seeing what happens under a sensitivity analysis, for example checking for strong rather than prior-based calibration of the posterior. And even a subjectivist cannot escape the conclusion from our experiments that optional stopping leads to more brittle (more sensitive to the prior choice) inference than stopping at a fixed n.

3.5.2 Frequentist optional stopping under $\mathcal{H}_0$

Interestingly, some other well-known Bayesian arguments claiming that 'optional stopping is no problem for Bayesians' really show that some Bayesian procedures can deal, in some cases, with optional stopping in a different, frequentist sense. These include Edwards, Lindman and Savage (1963) and Good (1991) and many others (the difference between this justification and the above one by Lindley (1957) roughly corresponds to Example 1 vs. Example 2 in the appendix to (Wagenmakers, 2007)). We now explain this frequentist notion of optional stopping, emphasizing that some (but — contrary to what is claimed — by no means all!) tests advocated by Bayesians *do* handle optional stopping in this frequentist sense.

The (or at least, 'a common') frequentist interpretation of handling optional stopping is about controlling the Type I error of an experiment. A Type I error occurs when we reject the null hypothesis when it is true, also called a *false positive*. The probability of a Type I error for a certain test is called the *significance level*, usually denoted by α, and in psychology the value of α is usually set to 0.05. A typical classical hypothesis test computes a test statistic from the data and uses it to calculate a p-value. It rejects the null hypothesis if the p-value is below the desired Type I error level α. For other types of hypothesis tests, it is also a crucial property to control the Type I error, by which we mean that we can make sure that the probability of making a Type I error remains below our chosen significance level α. The frequentist interpretation of handling optional stopping is that the Type I error guarantee holds if we do not determine the sampling plan — and thus the stopping rule — in advance, but we may stop when we see a significant result. As we know, see e.g. Wagenmakers (2007), maintaining this guarantee under optional stopping is not possible with most classical p-value based hypothesis tests.

At first sight none of this seems applicable to Bayesian tests, which output posterior odds rather than a p-value. However, in the case that $\mathcal{H}_0$ is *simple* (containing just one hypothesis, as in Example 0), there is a well-known intriguing connection between Bayes factors and Type I error probabilities: — if we reject H_0 iff the posterior odds in favor of $\mathcal{H}_0$ are smaller than some fixed α, then we are guaranteed a Type I error of at most α. And interestingly, this holds not just for fixed sample sizes but even under optional stopping. Thus, if one adopts the rejection rule above (reject iff the posterior odds are smaller than a fixed α), *for* simple $\mathcal{H}_0$, *frequentist optional stopping is no problem for Bayesians*. This is what was noted by Edwards, Lindman

and Savage (1963) (using a different terminology) and Good (1991), based on what Sanborn and Hills (2014) call the *universal bound*, and what in probability theory is known as *Doob's maximal inequality* (Doob, 1971); see also Vovk et al. (2011) and Van der Pas and Grünwald (2018).

But what happens if $\mathcal{H}_0$ is composite? As was only shown very recently (Hendriksen, De Heide and Grünwald, 2020), the Bayes factor still handles optional stopping in the frequentist sense if *all* free parameters in $\mathcal{H}_0$ are nuisance parameters observing a group structure and equipped with the corresponding Type 0 prior and are shared with $\mathcal{H}_1$, an example being Jeffreys' Bayesian t-test of Section 3.4.1. As explained by Hendriksen, De Heide and Grünwald (2020), for general priors and composite $\mathcal{H}_0$ though, this is typically not the case; for example, the Gunel-Dickey default Bayes factors for2×2 tables (another composite $\mathcal{H}_0$) cannot handle optional stopping in the frequentist sense.

An Empirical Frequentist Study of Bayesian Optional Stopping Schönbrodt et al. (2017) performed a thorough simulation study to analyze frequentist performance of optional stopping with Bayes factors both under $\mathcal{H}_0$ and under $\mathcal{H}_1$. They confined their analysis to the Bayesian t-test, i.e. our Example 2, and found excellent results for the Bayesian optional stopping procedure *under a certain frequentist interpretation* of the Bayes factors (posterior odds). As to optional stopping under $\mathcal{H}_0$ (concerning Type I error), this should not surprise us: in the Bayesian t-test, all free parameters in $\mathcal{H}_0$ are equipped with Type 0 priors, which, as we just stated, can handle optional stopping. We thus feel that one should be careful in extrapolating their results to other models such as those for contingency tables, which do not admit such priors. As to optional stopping under $\mathcal{H}_1$, the authors provide a table showing how, for any given effect size δ and desired level of Type II error β, a threshold B can be determined such that the standard Bayesian t-test with (essentially) the following optional stopping and decision rule, has Type II error β:

> Take at least 20 data points. After that stop as soon as posterior odds are larger than B or smaller than $1/B$; accept $\mathcal{H}_0$ if they are smaller than $1/B$, and reject $\mathcal{H}_0$ if larger than B.

For example, if $\delta \geq 0.3$ and one takes $B = 7$ then the Type II error will be smaller than 4% (see their Table 1). They also determined the average sample size needed before this procedure stops, and noted that this is considerably smaller than with the standard t-test optimized for the given desired levels of Type I and Type II error and a priori expected effect size. Thus, if one determines the optional stopping threshold B in the Bayesian t-test based on their table, one can use this Bayesian procedure as a frequentist testing method that significantly improves on the standard t-test in terms of sample size. Under *this* frequentist interpretation (which relies on the specifics of a table), optional stopping with the t-test is indeed unproblematic. Note that this does not contradict our findings in any way: our simulations show that if, when sampling, we fix an effect size in $\mathcal{H}_1$, then the posterior is biased under optional stopping, which means that we cannot interpret the posterior in a *Bayesian* way.

3.6 Discussion and Conclusion

When a researcher using Bayes factors for hypothesis testing truly believes in her prior, she can deal with optional stopping in the Bayesian senses just explained. However, these senses become problematic for every test that makes use of default priors, including all default Bayes factor tests advocated within the Bayesian Psychology community. Such 'default' or 'objective' priors cannot be interpreted in terms of willingness to bet, and sometimes (Type II priors) depend on aspects of the problem at hand such as the stopping rule or the inference task of interest. To make sense of such priors generally, it thus seems necessary to *restrict* their use to their appropriate domain of reference — for example, Jeffreys' prior for the Bernoulli model as given by (3.11) is okay for Bayes factor hypothesis testing with fixed sample size, but not for more complicated stopping rules. This idea, which is unfortunately almost totally lacking from the modern Bayesian literature, is the basis of a novel theory of the very concept of probability called *Safe Probability* which is being developed by one of us (Grünwald, 2013; Grünwald, 2018). That (mis)use of optional stopping is a serious problem in practice, is shown by, among others, John, Loewenstein and Prelec (2012b); however, that paper is (implicitly) mostly about frequentist methods. It would be interesting to investigate to what extent optional stopping when combined with default Bayesian methods is actually a problem not just in theory but also in practice. This would, however, require substantial further study and simulation.

Rouder (2014) argues in response to Sanborn and Hills (2014) that the latter 'evaluate and interpret Bayesian statistics as if they were frequentist statistics', and that 'the more germane question is whether Bayesian statistics are interpretable as *Bayesian statistics*'. Given the betting interpretation above, the essence here is that we need to make a distinction between the purely subjective and the pragmatic approach: we can certainly not evaluate and interpret *all* Bayesian statistics as *purely subjective* Bayesian statistics, what Rouder (2014) seems to imply. He advises Bayesians to use optional stopping — without any remark or restriction to purely subjective Bayesians, and for a readership of experimental psychologists who are in general not familiar with the different flavors of Bayesianism — as he writes further on: 'Bayesians should consider optional stopping in practice. [...] Such an approach strikes me as justifiable and reasonable, perhaps with the caveat that such protocols be made explicit before data collection'. The crucial point here is that this can indeed be done when one works with a purely subjective Bayesian method, but not with the *default Bayes factors* developed for practical use in social science: both strong calibration and the frequentist Type I-error guarantees will typically be violated, and for Bayes factors involving Type II-priors, both prior and strong calibration are even undefined. In Table 3.1 we provide researchers with a simplified overview of four common default Bayes factors indicating which forms of optional stopping they can handle.

While some find the purely subjective Bayesian framework unsuitable for scientific research (see e.g. Berger (2006)), others deem it the only coherent approach to learning from data per se. We do not want to enter this discussion, and we do not have to, since in practice, nowadays most Bayesian statisticians tend to use priors which have both 'default' and 'subjective' aspects. Basically, one uses mathematically convenient priors (which one does not really believe, so they are not purely subjective — and hence, prior calibration is of limited relevance), but they are also chosen to be not overly unrealistic or to match, to some extent, prior knowledge one might have about a problem. This position is almost inevitable in Bayesian practice (especially

	Prior Cal.	Strong Calibration	Freq. OS
Default Bayes Factors			
T-test (Rouder et al., 2009)	✓ **but...** (I)	✓ for σ (o) ✗ for δ (effect size) (I)	✓
ANOVA (Rouder et al., 2012)	✓ **but...** (I)	✓ for μ, σ (o) ✗ for δ (effect size) (I)	✓
Regression (Rouder and Morey, 2012)	✗ (II)	✓ for μ, σ (o) ✗ for β (effects) (II)	✓
Contingency Tables (Jamil et al., 2016)	✗ (II)	✗	✗
Bayes Factors with proper, fully subjective priors (Rouder, 2014)	✓	N/A	N/A

Table 3.1: Overview of several common default Bayes Factors (from the R-package `BayesFactor` (Morey and Rouder, 2015)), and their robustness against different kinds of optional stopping (proofs can be found in Hendriksen, De Heide and Grünwald, 2020). 'Prior Cal.' means 'prior calibration' and 'Freq. OS' means 'frequentist optional stopping'. Between parentheses is the type of prior used, in the taxonomy introduced in this paper. The **but..** indicates that, formally, prior calibration works for the priors, yet, because we are in the default setting, the Bayes factor is not fully subjective, so prior calibration is not too meaningful — which is just the main point of this paper.

since we would not like to burden practitioners with all the subtleties regarding objective and subjective Bayes), and we have no objections to it — but it does imply that, just like frequentists, Bayesians should be careful with optional stopping. For researchers who like to engage in optional stopping but care about frequentist concepts such as Type I error and power, we recommend the *safe tests* of Grünwald, De Heide and Koolen, 2019 based on the novel concept of *S-values*: S-values are related to, and sometimes coincide with, default Bayes factors, but tests based on S-values invariably handle a variation of frequentist optional stopping. For example, the three default Bayes factors that handle frequentist optional stopping in Table 3.1 are also S-values, but there exist other S-values for these three settings that also handle optional stopping but achieve higher frequentist power; and there also exists an S-value for contingency tables that, unlike the default Bayes factor, handles frequentist optional stopping.

Open Practices Statement Since all the data involved in this paper was generated by straightforward computer simulations rather than 'real-world' experiments, we did not make the data available. No experiments were done, and hence no experiments were preregistered.

3.A Example 4: An independence test in a 2x2 contingency table

Suppose that a researcher considers two hypotheses: a null hypothesis $\mathcal{H}_0$ that states that there is no difference in voting preference (Democrat or Republican) between men and women, and an alternative hypothesis $\mathcal{H}_1$ stating that men's voting preferences differ from the women's preferences. Both hypotheses are composite — we may think of a Bernoulli model for $\mathcal{H}_0$: the data are i.i.d. with a fixed probability of 1 (voting Democrat). We are however not interested in the percentage of the persons voting for the Democrats. We are, instead, only interested to learn if this percentage is *equal* for men and women or not. Thus our null hypothesis $\mathcal{H}_0$ consists of all Bernoulli distributions (all possible biases of the coin, infinitely many between 0 and 1) where the model for the men is the same as for the women. Our alternative hypothesis is composite as well: all the sets of two Bernoulli distributions — one for the men and one for the women — that are not equal. Thus, the Bernoulli parameter in $\mathcal{H}_0$ is not a parameter of interest; instead, at least intuitively, it is a nuisance parameter similar to the variance in Example 1; however, it does not observe a group structure and a Type 0-prior for this parameter does not exist.

Once again we follow Rouder's experiments closely. We now use the *Default Gunel and Dickey Bayes Factors for Contingency Tables* (Jamil et al., 2016), which employs specific default choices for the priors within $\mathcal{H}_0$ and $\mathcal{H}_1$, depending on four different sampling schemes (see Section 3.A for the details). We immediately run into a problem similar to the problems described with the g-prior and Jeffreys' prior for Bernoulli: which prior we should choose depends on the sampling plan itself. Based on earlier work by Gunel and Dickey, 1974 (GD from now on), Jamil et al. (2016) provide different default priors depending on whether the sample size n and/or some of the four counts (number of men/women voting democratic/republican) are fixed in advance. For the case that none of these are fixed in advance, they provide a prior which assumes that the four counts are all Poisson distributed; see the next section for details. Intuitively, none of these priors seem to be compatible with the very idea of 'optional stopping' and prior-based calibration under optional stopping cannot be tested (since it is not clear what prior to sample from — a Type II-problem in our earlier terminology). Still, to check the claim that 'optional stopping is no problem for Bayesians' we will again check whether *strong* calibration holds with and without optional stopping. We display here the results of an experiment with the prior advocated for the case in which neither n nor any of the counts are assumed to be fixed in advance, since this seems the choice least incompatible with optional stopping. To avoid discussion on this issue though, we also performed the experiments with the priors advocated for other sampling schemes and combinations of different sampling schemes, which led to very similar results.

We will again fix some 'reasonable' parameter values in each model: when sampling from $\mathcal{H}_0$, we really sample from $\theta = 1/2$, i.e. we suppose that 50% of either gender prefers the Democrats. When we sample from $\mathcal{H}_1$, we suppose that 45% of the men prefers the Democrats, but for the women it is as much as 55%. If there are equally many men as women, under both hypotheses

the average percentage is equal. Like Rouder, we set our prior odds to 1-to-1.

We simulate 20.000 replicate experiments of 100 + 100 samples each, from both $\mathcal{H}_0$ and $\mathcal{H}_1$, and we calculate the Bayes Factors. We construct the histograms and the plots with the odds as before. We can check the calibration in Figure 3.6b; we can see that the nominal posterior odds agree roughly with the observed posterior odds. In Figure 3.6d however, we see the same plot where we did the same experiment with optional stopping. We can clearly see that even the rough linear relationship from Figure 3.6b is completely gone. For this example, we can conclude as well that strong calibration is violated.

We now revisit the example, but we change the proportions under both hypotheses and survey only 25 men and 25 women, and we use a joint multinomial sampling scheme (the grand total, n, is fixed). Under $\mathcal{H}_0$, 70% of both men and women vote for the Democrats, and under $\mathcal{H}_1$, 65% of the men and 75% of the women do. We repeat exactly the same experiment (without optional stopping), and we see the resulting plot in Figure 3.7a. We see that the relationship between the observed and nominal posterior odds looks linear, but the slope is off. If we repeat the same experiment with optional stopping, we see in Figure 3.7b that additionally the linear association is missing.

We do note that the objective priors used in the default Bayes Factor test for contingency tables are proper, so we are able to sample from them. In Figure 3.7c we see what happens if we do exactly the same experiment as in Figure 3.7a, but sampled from the prior: we see the observed posterior odds plotted against the nominal posterior odds, and the points lie approximately on the identity line, in contrast with Figure 3.7a. Furthermore, we performed the same experiment as in Figure 3.7b in this subjective Bayesian way, and we see that (in Rouder's terminology) the interpretation of the posterior odds holds with optional stopping in Figure 3.7d. As said, we do not think this kind of sampling is very meaningful in default prior context; we just add the experiment to show that invariably, *if* one can and wants to sample from priors, then Rouder's conclusions do hold.

Subjective vs. Objective Interpretation In their original paper, Gunel and Dickey, 1974 (GD) give a *subjective* interpretation to their priors. These priors depend on the sampling scheme, i.e. on whether the grand total, and/or one or both of the marginals are known or set by the experimenter in advance. At first sight, this seems to be at odds with the fact that, with subjective priors, Bayesian procedures do not depend on the stopping rule used, as we pointed out in Section 3.5. However, closer inspection reveals that if one follows the method under their subjective interpretation, then the posterior indeed would not depend on the sampling scheme. How is this possible? To see this, note that GD do not model their data as coming in sequentially, but rather they consider a fixed, single datum $D = (N_1, \ldots, N_4)$ consisting of the four entries in the contingency table (see e.g. Table 3.2 below). The different versions of their model and prior are then arrived at by calculating, for example, $\mathbb{P}(D \mid \mathcal{H}_0)$ for the case that no information about the design is given, and $\mathbb{P}(D \mid \mathcal{H}_0, n)$ (where $n = N_1 + N_2 + N_3 + N_4$) for the case that the grand total (sample size) n is determined in the experiment design. In every case, the posterior

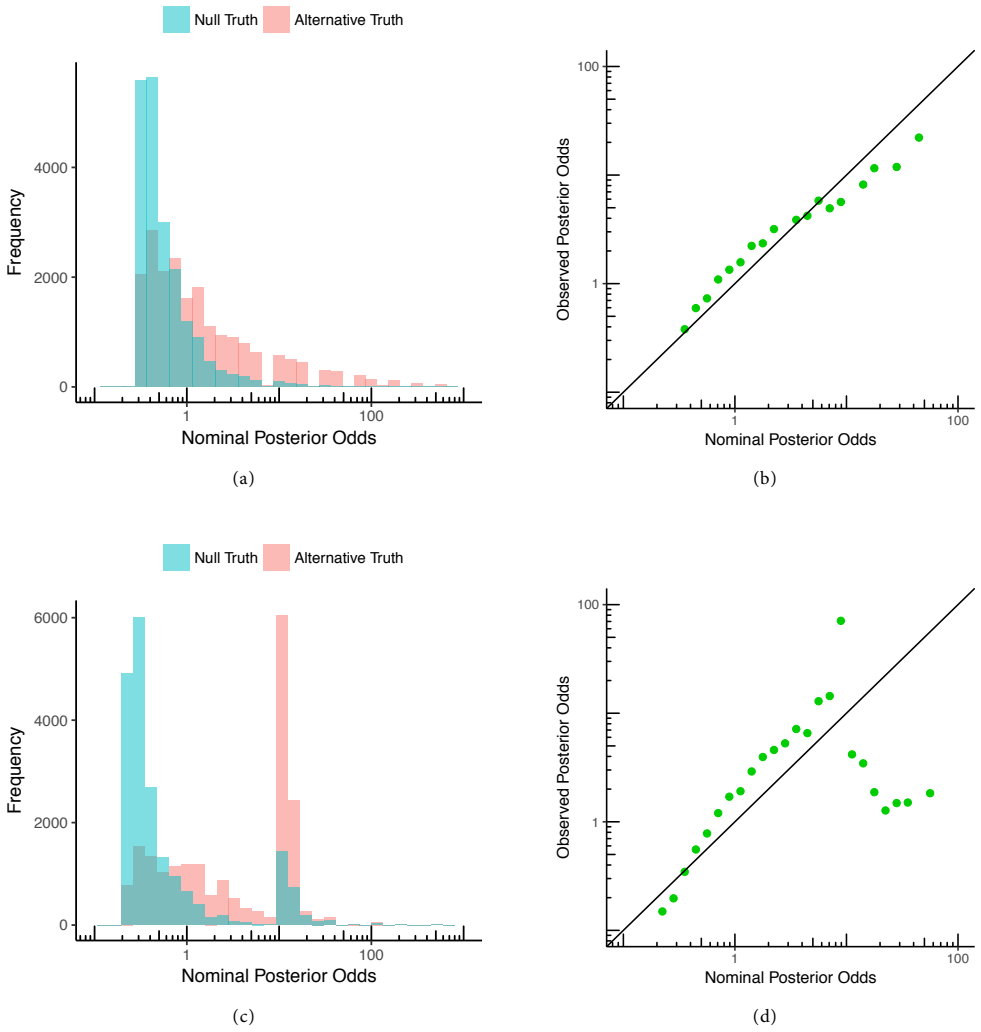


Figure 3.6: Calibration of the contingency table experiment, Section 3.A, from 20.000 replicate experiments. (a) The distribution of posterior odds as a histogram under $\mathcal{H}_0$ and $\mathcal{H}_1$. (b) The observed posterior odds as a function of the nominal posterior odds. (c) Distribution of the posterior odds with optional stopping. (d) The observed posterior odds as a function of the nominal posterior odds with optional stopping.

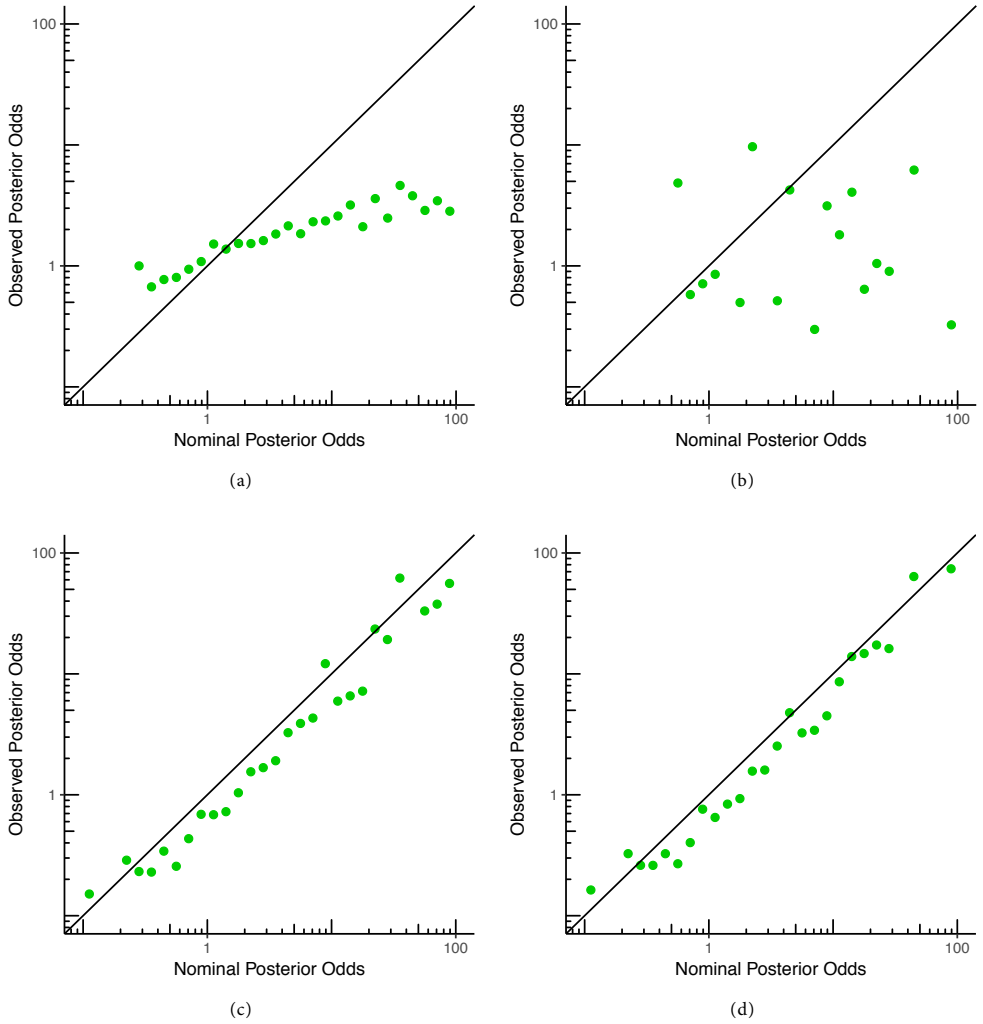


Figure 3.7: The observed posterior odds as a function of the nominal posterior odds, from 20.000 replicate experiments. (a) Contingency table experiment, without optional stopping. (b) Contingency table experiment, with optional stopping. (c) Subjective Bayesian version of the experiment in a. (d) Subjective Bayesian version of the experiment in b.

odds `post-odds`$|D$ will remain the same; for they require the *prior* to be used when n is given, $\mathbb{P}(\mathcal{H}_0 \mid n)$, to be arrived at by conditioning the original prior $\mathbb{P}(\mathcal{H}_0)$ on the grand total n. In particular, this means that a truly subjective Bayesian who follows the GD model would have $\mathbb{P}(\mathcal{H}_0 \mid n) \neq \mathbb{P}(\mathcal{H}_0)$, and could thus not use a $(1/2, 1/2)$ 'uninformative' prior on $(\mathcal{H}_0, \mathcal{H}_1)$ both when the grand total is known in advance and when it is not. In other words, the posterior is *not* affected by the sampling scheme, but the *prior* is.

Details of the experiments For Example 4 above, we used the function `contingencyTableBF`. This function gives the user the option to choose between four different so called sampling schemes, implementing the *Default Gunel and Dickey Bayes Factors for Contingency Tables* of Jamil et al. (2016). Which of the four options to use, depends on which covariates in the contingency table are to be treated as fixed or as random, depending on the design of the experiment.

	0	1	sum
0	$n_1 - k_1$	$n_2 - k_2$	$n - k$
1	k_1	k_2	k
sum	n_1	n_2	n

Table 3.2: 2x2 contingency table; the four entries correspond to the numbers $N_1, N_2, \ldots, N_4$ above.

In the first sampling scheme, none of the cell counts in the contingency table are considered fixed, and the assumption is made that each cell count is Poisson distributed. The default prior for this scheme is a conjugate gamma prior on the Poisson rate parameter, with hyperparameters suggested by Gunel and Dickey. We use this sampling scheme for our first experiment in Section 3.A, but as we noted in our discussion in the same section, the question of 'what is the actual sampling scheme' and hence 'what is the right default prior' for the type of experiment we do — the same experiment with and without optional stopping — is really impossible to answer. Thus, we repeated the experiment with other (combinations of) sampling schemes, in all cases obtaining similar results. Indeed, when we perform the experiment without optional stopping, we sample a fixed number of men and women, whereupon one margin (n_1, n_2) and the grand total (n) is fixed. For our second example (Figure 3.7a and 3.7b) we used the prior advocated for the sampling scheme in which the grand total (n in Table 3.2) is fixed. Under this sampling scheme, the cell counts are assumed to be jointly multinomial distributed, and a Dirichlet conjugate distribution with the suggested parameters (Jamil et al., 2016) is used as prior, which in our case amounts to a uniform prior on the Bernoulli parameter θ; see Jamil et al. (2016) for details. Again, using instead one of the priors advocated for one of the other sampling schemes leads to very similar results.

Chapter 4

Optional stopping with Bayes Factors

Abstract

It is often claimed that Bayesian methods, in particular Bayes factor methods for hypothesis testing, can deal with optional stopping. We first give an overview, using elementary probability theory, of three different mathematical meanings that various authors give to this claim: (1) stopping rule *independence*, (2) posterior *calibration* and (3) (semi-) *frequentist robustness to optional stopping*. We then prove theorems to the effect that these claims do indeed hold in a general measure-theoretic setting. For claims of type (2) and (3), such results are new. By allowing for non-integrable measures based on improper priors, we obtain particularly strong results for the practically important case of models with nuisance parameters satisfying a group invariance (such as location or scale). We also discuss the practical relevance of (1)–(3), and conclude that whether Bayes factor methods actually perform well under optional stopping crucially depends on details of models, priors and the goal of the analysis.

4.1 Introduction

In recent years, a surprising number of scientific results have failed to hold up to continued scrutiny. Part of this 'replicability crisis' may be caused by practices that ignore the assumptions of traditional (frequentist) statistical methods (John, Loewenstein and Prelec, 2012a). One of these assumptions is that the experimental protocol should be completely determined upfront. In practice, researchers often adjust the protocol due to unforeseen circumstances or collect data until a point has been proven. This practice, which is referred to as *optional stopping*, can cause true hypotheses to be wrongly rejected much more often than these statistical methods promise.

Bayes factor hypothesis testing has long been advocated as an alternative to traditional testing

that can resolve several of its problems; in particular, it was claimed early on that Bayesian methods continue to be valid under optional stopping (Lindley, 1957; Raiffa and Schlaifer, 1961; Edwards, Lindman and Savage, 1963). In particular, the latter paper claims that (with Bayesian methods) "it is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience." In light of the replicability crisis, such claims have received much renewed interest (Wagenmakers, 2007; Rouder, 2014; Schönbrodt et al., 2017; Yu et al., 2014; Sanborn and Hills, 2014). But what do they mean mathematically? It turns out that different authors mean quite different things by 'Bayesian methods handle optional stopping'; moreover, such claims are often shown to hold only in an informal sense, or in restricted contexts. Thus, the first goal of the present chapter is to give a systematic overview and formalization of such claims in a simple, expository setting and, still in this simple setting, explain their relevance for practice: can we effectively rely on Bayes factor testing to do a good job under optional stopping or not? As we shall see, the answer is subtle. The second goal is to extend the reach of such claims to more general settings, for which they have never been formally verified and for which verification is not always trivial.

Overview In *Section 4.2*, we give a systematic overview of what we identified to be the three main mathematical senses in which Bayes factor methods can handle optional stopping, which we call *τ-independence, calibration, and (semi-)frequentist*. We first do this in a setting chosen to be as simple as possible — finite sample spaces and strictly positive probabilities — allowing for straightforward statements and proofs of results. In *Section 4.3*, we explain the practical relevance of these three notions. It turns out that whether or not we can say that 'the Bayes factor method can handle optional stopping' in practice is a subtle matter, depending on the specifics of the given situation: what models are used, what priors, and what is the goal of the analysis. We can thus explain the paradox that there have also been claims in the literature that Bayesian methods *cannot* handle optional stopping in certain cases; such claims were made, for example by Yu et al., 2014; Sanborn and Hills, 2014, and also by ourselves (De Heide and Grünwald, 2018). We also briefly discuss *safe tests* (Grünwald, De Heide and Koolen, 2019) which can be interpreted as a novel method for determining priors that behave better under frequentist optional stopping. The chapter has been organized in such a way that these first two sections can be read with only basic knowledge of probability theory and Bayesian statistics. For convenience, we illustrate Section 4.3 with an informally stated example involving group invariances, so that the reader gets a complete overview of what the later, more mathematical sections are about.

Section 4.4 extends the statements and results to a much more general setting allowing for a wide range of sample spaces and measures, including measures based on *improper priors*. These are priors that are not integrable, thus not defining standard probability distributions over parameters, and as such they cause technical complications. Such priors are indispensable within the recently popularized *default Bayes factors* for common hypothesis tests (Rouder et al., 2009; Rouder et al., 2012; Jamil et al., 2016).

In *Section 4.5*, we provide stronger results for the case in which both models satisfy the same group invariance. Several (not all) default Bayes factor settings concern such situations; prominent examples are Jeffreys' (1961) Bayesian one- and two-sample t-tests, in which the models

are location and location-scale families, respectively. Many more examples are given by Berger and various collaborators (Berger, Pericchi and Varshavsky, 1998; Dass and Berger, 2003; Bayarri et al., 2012; Bayarri et al., 2016). These papers provide compelling arguments for using the (typically improper) *right Haar prior* on the nuisance parameters in such situations; for example, in Jeffreys' one-sample t-test, one puts a right Haar prior on the variance. In particular, in our restricted context of Bayes factor hypothesis testing, the right Haar prior does not suffer from the *marginalization paradox* (Dawid, Stone and Zidek, 1973) that often plagues Bayesian inference based on improper priors (we briefly return to this point in the conclusion).

Haar priors and group invariant models were studied extensively by Eaton, 1989; Andersson, 1982; Wijsman, 1990, whose results this chapter depends on considerably. When nuisance parameters (shared by both H_0 and H_1) are of suitable form and the right Haar prior is used, we can strengthen the results of Section 4.4; they now hold uniformly for all possible values of the nuisance parameters, rather than in the marginal, 'on average' sense we consider in Section 4.4. However — and this is an important insight — we *cannot take arbitrary stopping rules* if we want to handle optional stopping in this strong sense: our theorems only hold if the stopping rules satisfy a certain intuitive condition, which will hold in many but not all practical cases: the stopping rule must be "invariant" under some group action. For instance, a rule such as 'stop as soon as the Bayes factor is ≥ 20' is allowed, but a rule (in the Jeffreys' one-sample t-test) such as 'stop as soon as $\sum x_i^2 \geq 20$' is not.

The chapter ends with supplementary material, comprising Section 4.A containing basic background material about groups, and Section 4.B containing all longer mathematical proofs.

Scope and Novelty Our analysis is restricted to Bayesian testing and model selection using the Bayes factor method; we do not make any claims about other types of Bayesian inference. Some of the results we present were already known, at least in simple settings; we refer in each case to the first appearance in the literature that we are aware of. In particular, our results in Section 4.4.1 are implied by earlier results in the seminal work by Berger and Wolpert, 1988 on the likelihood principle; we include them any way since they are a necessary building block for what follows. The real mathematical novelties in the chapter are the results on calibration and (semi-) frequentist optional stopping with general sample spaces and improper priors and the results on the group invariance case (Section 4.4.2–4.5). These results are truly novel, and — although perhaps not very surprising — they do require substantial additional work not covered by Berger and Wolpert, 1988, who are only concerned with τ-independence. In particular, the calibration results require the notion of the 'posterior odds of some particular posterior odds', which need to be defined under arbitrary stopping times. The difficulty here is that, in contrast to the fixed sample sizes where even with continuous-valued data, the Bayes factor and the posterior odds usually have a distribution with full support, with variable stopping times, the support may have 'gaps' at which its density is zero or very near zero. An additional difficulty encountered in the group invariance case is that one has to define filtrations based on maximal invariants, which requires excluding certain measure-zero points from the sample space.

4.2 The Simple Case

Consider a finite set $\mathcal{X}$ and a sample space $\Omega := \mathcal{X}^T$ where T is some very large (but in this section, still finite) integer. One observes a *sample* $x^\tau \equiv x_1, \ldots, x_\tau$, which is an initial segment of $x_1, \ldots, x_T \in \mathcal{X}^T$. In the simplest case, $\tau = n$ is a sample size that is fixed in advance; but, more generally τ is a *stopping time* defined by some stopping rule (which may or may not be known to the data analyst), defined formally below.

We consider a hypothesis testing scenario where we wish to distinguish between a null hypothesis H_0 and an alternative hypothesis H_1. Both H_0 and H_1 are sets of distributions on Ω, and they are each represented by unique probability distributions $\overline{P}_0$ and $\overline{P}_1$ respectively. Usually, these are taken to be Bayesian marginal distributions, defined as follows. First one writes, for both $k \in \{0, 1\}$, $H_k = \{P_{\theta|k} \mid \theta \in \Theta_k\}$ with 'parameter spaces' Θ_k; one then defines or assumes some prior probability distributions π_0 and π_1 on Θ_0 and Θ_1, respectively. The Bayesian marginal probability distributions are then the corresponding marginal distributions, i.e. for any set $A \subset \Omega$ they satisfy:

$$\overline{P}_0(A) = \int_{\Theta_0} P_{\theta|0}(A)\, \mathrm{d}\pi_0(\theta) \;\; ; \;\; \overline{P}_1(A) = \int_{\Theta_1} P_{\theta|1}(A)\, \mathrm{d}\pi_1(\theta). \tag{4.1}$$

For now we also further assume that for every $n \leq T$, every $x^n \in \mathcal{X}^n$, $\overline{P}_0(X^n = x^n) > 0$ and $\overline{P}_1(X^n = x^n) > 0$ (full support), where here, as below, we use random variable notation, $X^n = x^n$ denoting the event $\{x^n\} \subset \Omega$. We note that there exist approaches to testing and model choice such as testing by nonnegative martingales (Shafer et al., 2011; Van der Pas and Grünwald, 2018) and minimum description length (Barron, Rissanen and Yu, 1998; Grünwald, 2007) in which the $\overline{P}_0$ and $\overline{P}_1$ may be defined in different (yet related) ways. Several of the results below extend to general $\overline{P}_0$ and $\overline{P}_1$; we return to this point at the end of the chapter, in Section 4.6. In all cases, we further assume that we have determined an additional probability mass function π on $\{H_0, H_1\}$, indicating the prior probabilities of the hypotheses. The evidence in favor of H_1 relative to H_0 given data x^τ is now measured either by the *Bayes factor* or the *posterior odds*. We now give the standard definition of these quantities for the case that $\tau = n$, i.e., that the sample size is fixed in advance. First, noting that all conditioning below is on events of strictly positive probability, by Bayes' theorem, we can write for any $A \subset \Omega$,

$$\frac{\pi(H_1 \mid A)}{\pi(H_0 \mid A)} = \frac{P(A \mid H_1)}{P(A \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{4.2}$$

where here, as in the remainder of the chapter, we use the symbol π to denote not just prior, but also posterior distributions on $\{H_0, H_1\}$. In the case that we observe x^n for fixed n, the event A is of the form $X^n = x^n$. Plugging this into (4.2), the left-hand side becomes the standard definition of *posterior odds*, and the first factor on the right is called the *Bayes factor*.

4.2.1 First Sense of Handling Optional Stopping: τ-Independence

Now, in reality we do not necessarily observe $X^n = x^n$ for fixed n but rather $X^\tau = x^\tau$ where τ is a stopping time that may itself depend on (past) data (and that in some cases may in fact be unknown to us). This stopping time may be defined in terms of a *stopping rule* $f : \bigcup_{i \geq 0}^{T} \mathcal{X}^i \to$

$\{\texttt{stop}, \texttt{continue}\}$. $\tau \equiv \tau(x^T)$ is then defined as the random variable which, for any sample $x_1, \ldots, x_T$, outputs the smallest n such that $f(x_1, \ldots, x_n) = \texttt{stop}$. For any given stopping time τ, any $1 \leq n \leq T$ and sequence of data $x^n = (x_1, \ldots, x_n)$, we say that x^n *is compatible with* τ if it satisfies $X^n = x^n \Rightarrow \tau = n$. We let $\mathcal{X}^\tau \subset \bigcup_{i=1}^{T} \mathcal{X}^i$ be the set of all sequences compatible with τ.

Observations take the form $X^\tau = x^\tau$, which is equivalent to the event $X^n = x^n; \tau = n$ for some n and some $x^n \in \mathcal{X}^n$ which of necessity must be compatible with τ. We can thus instantiate (4.2) to

$$\frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)} = \frac{P(\tau = n \mid X^n = x^n, H_1) \cdot \pi(H_1 \mid X^n = x^n)}{P(\tau = n \mid X^n = x^n, H_0) \cdot \pi(H_0 \mid X^n = x^n)} = \\ = \frac{\pi(H_1 \mid X^n = x^n)}{\pi(H_0 \mid X^n = x^n)}. \tag{4.3}$$

where in the first equality we used Bayes' theorem (keeping $X^n = x^n$ on the right of the conditioning bar throughout); the second equality stems from the fact that $X^n = x^n$ logically implies $\tau = n$, since x^n is compatible with τ; the probability $P(\tau = n \mid X^n = x^n, H_j)$ must therefore be 1 for $j = 0, 1$. Combining (4.3) with Bayes' theorem we get:

$$\overbrace{\frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)}}^{\gamma(x^n)} = \overbrace{\frac{\overline{P}_1(X^n = x^n)}{\overline{P}_0(X^n = x^n)}}^{\beta(x^n)} \cdot \frac{\pi(H_1)}{\pi(H_0)} \tag{4.4}$$

where we introduce the notation $\gamma(x^n)$ for the posterior odds and $\beta(x^n)$ for the Bayes factor based on sample x^n, calculated as if n were fixed in advance.[1]

We see that the stopping rule plays no role in the expression on the right. Thus, we have shown that, for any two stopping times τ_1 and τ_2 that are both compatible with some observed x^n, the posterior odds one arrives at will be the same irrespective of whether x^n came to be observed because τ_1 was used or if x^n came to be observed because τ_2 was used. We say that the posterior odds do not depend on the stopping rule τ and call this property τ-*independence*. Incidentally, this also justifies that we write the posterior odds as $\gamma(x^n)$, a function of x^n alone, without referring to the stopping time τ.

The fact that the posterior odds given x^n do not depend on the stopping rule is the first (and simplest) sense in which Bayesian methods handle optional stopping. It has its roots in the *stopping rule principle*, the general idea that the conclusions obtained from the data by 'reasonable' statistical methods should not depend on the stopping rule used. This principle was probably first formulated by Barnard (1947; 1949); Barnard, 1949 very implicitly showed that, under some conditions, Bayesian methods satisfy the stopping rule principle (and hence satisfy τ-independence). Other early sources are Lindley (1957) and Edwards, Lindman and Savage (1963). Lindley gave an informal proof in the context of specific parametric models;

[1] A slightly different way to get to (4.4), which some may find even simpler, is to start with $\overline{P}_0(X^n = x^n, \tau = n) = \overline{P}_0(X^n = x^n)$ (since $X^n = x^n$ implies $\tau = n$), whence $\pi(H_j \mid X^n = x^n, \tau = n) \propto \overline{P}_j(X^n = x^n, \tau = n)\pi(H_j) = \overline{P}_j(X^n = x^n)\pi(H_j)$.

in Section 4.4.1 we show that, under some regularity conditions, the result indeed remains true for general σ-finite $\overline{P}_0$ and $\overline{P}_1$. A special case of our result (allowing continuous-valued sample spaces but not general measures) was proven by Raiffa and Schlaifer, 1961, and a more general statement about the connection between the 'likelihood principle' and the 'stopping rule principle' which implies our result in Section 4.4.1 can be found in the seminal work (Berger and Wolpert, 1988), who also provide some historical context. Still, even though not new in itself, we include our result on τ-independence with general sample spaces and measures since it is the basic building block of our later results on calibration and semi-frequentist robustness, which are new.

Finally, we should note that both Raiffa and Schlaifer, 1961 and Berger and Wolpert, 1988 consider more general stopping rules, which can map to a probability of stopping instead of just $\{\texttt{stop}, \texttt{continue}\}$. Also, they allow the stopping rule itself to be parameterized: one deals with a collection of stopping rules $\{f_\xi : \xi \in \Xi\}$ with corresponding stopping times $\{\tau_\xi : \xi \in \Xi\}$, where the parameter ξ is equipped with a prior such that ξ and H_j are required to be a priori independent. Such extensions are straightforward to incorporate into our development as well (very roughly, the second equality in (4.3) now follows because, by conditional independence, we must have that $P(\tau_\xi = n \mid X^n = x^n, H_1) = P(\tau_\xi = n \mid X^n = x^n, H_0)$); we will not go into such extensions any further in this chapter.

4.2.2 Second Sense of Handling Optional Stopping: Calibration

An alternative definition of handling optional stopping was introduced by Rouder, 2014. Rouder calls $\gamma(x^n)$ the *nominal* posterior odds calculated from an obtained sample x^n, and defines the *observed posterior odds* as

$$\frac{\pi(H_1 \mid \gamma(x^n) = c)}{\pi(H_0 \mid \gamma(x^n) = c)}$$

as the posterior odds given the nominal odds. Rouder first notes that, at least if the sample size is fixed in advance to n, one expects these odds to be equal. For instance, if an obtained sample yields nominal posterior odds of 3-to-1 in favor of the alternative hypothesis, then it must be 3 times as likely that the sample was generated by the alternative probability measure. In the terminology of De Heide and Grünwald, 2018, Bayes is *calibrated* for a fixed sample size n. Rouder then goes on to note that, if n is determined by an arbitrary stopping time τ (based for example on optional stopping), then the odds will still be equal — in this sense, Bayesian testing is well-behaved in the calibration sense irrespective of the stopping rule/time. Formally, the requirement that the nominal and observed posterior odds be equal leads us to define the *calibration hypothesis*, which postulates that $c = P(H_1 \mid \gamma = c)/P(H_0 \mid \gamma = c)$ holds for any $c > 0$ that has non-zero probability. For simplicity, for now we only consider the case with equal prior odds for H_0 and H_1 so that $\gamma(x^n) = \beta(x^n)$. Then the calibration hypothesis says that, for arbitrary stopping time τ, for every c such that $\beta(x^\tau) = c$ for some $x^\tau \in \mathcal{X}^\tau$, one has

$$c = \frac{P(\beta(x^\tau) = c \mid H_1)}{P(\beta(x^\tau) = c \mid H_0)}. \tag{4.5}$$

In the present simple setting, this hypothesis is easily shown to hold, because we can write:

$$\frac{P(\beta(X^\tau) = c \mid H_1)}{P(\beta(X^\tau) = c \mid H_0)} = \frac{\sum_{y \in \mathcal{X}^\tau : \beta(y) = c} P(\{y\} \mid H_1)}{\sum_{y \in \mathcal{X}^\tau ; \beta(y) = c} P(\{y\} \mid H_0)} = \frac{\sum_{y \in \mathcal{X}^\tau : \beta(y) = c} cP(\{y\} \mid H_0)}{\sum_{y \in \mathcal{X}^\tau : \beta(y) = c} P(\{y\} \mid H_0)} = c.$$

Rouder noticed that the calibration hypothesis should hold as a mathematical theorem, without giving an explicit proof; he demonstrated it by computer simulation in a simple parametric setting. Deng, Lu and Chen, 2016 gave a proof for a somewhat more extended setting yet still with proper priors. In Section 4.4.2 we show that a version of the calibration hypothesis continues to hold for general measures based on improper priors, and in Section 4.5.4 we extend this further to strong calibration for group invariance settings as discussed below.

We note that this result, too, relies on the priors themselves not depending on the stopping time, an assumption which is violated in several standard default Bayes factor settings. We also note that, if one thinks of one's priors in a default sense — they are practical but not necessarily fully believed — then the practical implications of calibration are limited, as shown experimentally by De Heide and Grünwald, 2018. One would really like a stronger form of calibration in which (4.5) holds under a whole range of distributions in H_0 and H_1, rather than in terms of $\overline{P}_0$ and $\overline{P}_1$ which average over a prior that perhaps does not reflect one's beliefs fully. For the case that H_1 and H_2 share a nuisance parameter g taking values in some set G, one can define this *strong calibration hypothesis* as stating that, for all c with $\beta(x^\tau) = c$ for some $x^\tau \in \mathcal{X}^\tau$, all $g \in G$,

$$c = \frac{P(\beta(x^\tau) = c \mid H_1, g)}{P(\beta(x^\tau) = c \mid H_0, g)}. \tag{4.6}$$

where β is still defined as above; in particular, when calculating β one does not condition on the parameter having the value g, but when assessing its likelihood as in (4.6) one does. De Heide and Grünwald, 2018 show that the strong calibration hypothesis certainly does *not* hold for general parameters, but they also show by simulations that it does hold in the practically important case with group invariance and right Haar priors (Example 4.1 provides an illustration). In Section 4.5.4 we show that in such cases, one can indeed prove that a version of (4.6) holds.

4.2.3 Third Sense of Handling Optional Stopping: (Semi-)Frequentist

In classical, Neyman-Pearson style null hypothesis testing, a main concern is to limit the false positive rate of a hypothesis test. If this false positive rate is bounded above by some $\alpha > 0$, then a null hypothesis significance test (NHST) is said to have *significance level* α, and if the significance level is independent of the stopping rule used, we say that the test is *robust under frequentist optional stopping*.

Definition 4.1. A function $S : \bigcup_{i=m}^{T} \mathcal{X}^i \to \{0,1\}$ is said to be a frequentist sequential test with significance level α and minimal sample size m that is *robust under optional stopping relative to* H_0 if for all $P \in H_0$

$$P\left(\exists n, m < n \leq T : S(X^n) = 1\right) \leq \alpha,$$

i.e. the probability that there is an n at which $S(X^n) = 1$ ('the test rejects H_0 when given sample X^n') is bounded by α.

In our present setting, we can take $m = 0$ (larger m become important in Section 4.4.3), so n runs from 1 to T and it is easy to show that, for any $0 \leq \alpha \leq 1$, we have

$$\overline{P}_0\left(\exists n, 0 < n \leq T : \frac{1}{\beta(x^n)} \leq \alpha\right) \leq \alpha. \tag{4.7}$$

Proof. For any fixed α and any sequence $x^T = x_1, \ldots, x_T$, let $\tau(x^T)$ be the smallest n such that, for the initial segment x^n of x^T, $\beta(x^n) \geq 1/\alpha$ (if no such n exists we set $\tau(x^T) = T$). Then τ is a stopping time, X^τ is a random variable, and the probability in (4.7) is equal to the $\overline{P}_0$-probability that $\beta(X^\tau) \geq 1/\alpha$, which by Markov's inequality is bounded by α. □

It follows that, if H_0 is a singleton, then the sequential test S that rejects H_0 (outputs $S(X^n) = 1$) whenever $\beta(x^n) \geq 1/\alpha$ is a frequentist sequential test with significance level α that is robust under optional stopping.

The fact that Bayes factor testing with singleton H_0 handles optional stopping in this frequentist way was noted by Edwards, Lindman and Savage (1963) and also emphasized by Good, 1991, among many others. If H_0 is not a singleton, then (4.7) still holds, so the Bayes factor still handles optional stopping in a mixed frequentist (Type I-error) and Bayesian (marginalizing over prior within H_0) sense. From a frequentist perspective, one may not consider this to be fully satisfactory, and hence we call it 'semi-frequentist'. In some quite special situations though, it turns out that the Bayes factor satisfies the stronger property of being truly robust to optional stopping in the above frequentist sense, i.e. (4.7) will hold for all $P \in H_0$ and not just 'on average'. This is illustrated in Example 4.1 below and formalized in Section 4.5.5.

4.3 Discussion: why should one care?

Nowadays, even more so than in the past, statistical tests are often performed in an on-line setting, in which data keeps coming in sequentially and one cannot tell in advance at what point the analysis will be stopped and a decision will be made — there may indeed be many such points. Prime examples include group sequential trials (Proschan, Lan and Wittes, 2006) and A/B-testing, to which all internet users who visit the sites of the tech giants are subjected. In such on-line settings, it may or may not be a good idea to use Bayesian tests. But can and should they be used? Together with the companion paper (De Heide and Grünwald, 2018) (DHG from now on — corresponding to Chapter 3 of this dissertation), the present chapter sheds some light on this issue. Let us first highlight a central insight from DHG, which is about the case in which none of the results discussed in the present chapter apply: in many practical situations, many Bayesian statisticians use priors that are *themselves* dependent on parts of the data and/or the sampling plan and stopping time. Examples are Jeffreys prior with the multinomial model and the Gunel-Dickey default priors for 2x2 contingency tables advocated by Jamil et al., 2016. With such priors, final results evidently depend on the stopping rule employed, and even though such methods typically count as 'Bayesian', they do not satisfy τ-independence. The results then become non-interpretable under optional stopping (i.e. stopping using a rule that is not known at the time the prior is decided upon), and as argued by De Heide and Grünwald, 2018,

the notions of calibration and frequentist optional stopping even become undefined in such a case.

In such situations, one cannot rely on Bayesian methods to be valid under optional stopping in any sense at all; in the present chapter we thus focus on the case with priors that are fixed in advance, and that themselves do not depend on the stopping rule or any other aspects of the design. For expository simplicity, we consider the question of whether Bayes factors with such priors are valid under optional stopping in two extreme settings: in the first setting, the goal of the analysis is purely *exploratory* — it should give us some insight in the data and/or suggest novel experiments to gather or novel models to analyze data with. In the second setting we consider the analysis as 'final' and the stakes are much higher — real decisions involving money, health and the like are involved — a typical example would be a Stage 2 clinical trial, which will decide whether a new medication will be put to market or not.

For the first, *exploratory* setting, exact error guarantees might neither be needed at all nor obtainable anyway, so the frequentist sense of handling optional stopping may not be that important. Yet, one would still like to use methods that satisfy some basic *sanity checks* for use under optional stopping. τ-independence is such a check: any method for which it does not hold is simply not suitable for use in a situation in which details of the stopping rule may be unknown. Also calibration can be viewed as such a sanity check: Rouder, 2014 introduced it mainly to show that Bayesian posterior odds remain *meaningful* under optional stopping: they still satisfy some key property that they satisfy for fixed sample sizes.

For the second *high stakes* setting, mere sanity and interpretability checks are not enough: most researchers would want more stringent guarantees, for example on Type-I and/or Type-II error control. At the same time, most researchers would acknowledge that their priors are far from perfect, chosen to some extent for purposes of convenience rather than true belief.[2] Such researchers may thus want the desired Type-I error guarantees to hold for all $P \in H_0$, and not just in average over the prior as in (4.7). Similarly, in the high stakes setting the form of calibration (4.5) that can be guaranteed for the Bayes factor would be considered too weak, and one would hope for a stronger form of calibration as explained at the end of Section 4.2.2.

DHG show empirically that for some often-used models and priors, strong calibration can be severely violated under optional stopping. Similarly, it is possible to show that in general, Type-I error guarantees based on Bayes factors simply do not hold simultaneously for all $P \in H_0$ for such models and priors. Thus, one should be cautious using Bayesian methods in the high stakes setting, despite adhortations such as the quote by Edwards, Lindman and Savage, 1963 in the introduction (or similar quotes by e.g. Rouder et al., 2009): these existing papers invariably use τ-independence, calibration or Type-I error control with simple null hypotheses as a motivation to — essentially — use Bayes factor methods in any situation, including presumably high-stakes situations and situations with composite null hypotheses.[3]

[2] Even De Finetti and Savage, fathers of subjective Bayesianism, acknowledged this: see Section 5 of DHG.

[3] Since the authors of the present chapter are inclined to think frequentist error guarantees are important, we disagree with such claims, as in fact a subset of researchers calling themselves Bayesians would as well. To witness, a large fraction of recent ISBA (Bayesian) meetings is about frequentist properties of Bayesian methods; also the well-known Bayesian authors Good, 1991 and Edwards, Lindman and Savage, 1963 focus on showing that Bayes factor methods achieve a *frequentist Type-I error* guarantee, albeit only for the simple H_0 case.

Still, and this is equally important for practitioners, while frequentist error control and strong calibration are violated in general, in some important special cases they do hold, namely if the models H_0 and H_1 satisfy a group invariance. We proceed to give an informal illustration of this fact, deferring the mathematical details to Section 4.5.5.

Example 4.1. Consider the one-sample t-test as described by Rouder et al., 2009, going back to Jeffreys, 1961. The test considers normally distributed data with unknown standard deviation. The test is meant to answer the question whether the data has mean $\mu = 0$ (the null hypothesis) or some other mean (the alternative hypothesis). Following (Rouder et al., 2009), a Cauchy prior density, denoted by $\pi_\delta(\delta)$, is placed on the effect size $\delta = \mu/\sigma$. The unknown standard deviation is a nuisance parameter and is equipped with the improper prior with density $\pi_\sigma(\sigma) = \frac{1}{\sigma}$ under both hypotheses. This is the so-called *right Haar prior* for the variance. This gives the following densities on n outcomes:

$$p_{0,\sigma}(x^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right) \quad [\, = p_{1,\sigma,0}(x^n)\,] \tag{4.8}$$

$$p_{1,\sigma,\delta}(x^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{n}{2}\left[\left(\frac{\overline{x}}{\sigma} - \delta\right)^2 + \left(\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2}{\sigma^2}\right)\right]\right), \text{ where}$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i,$$

so that the corresponding Bayesian marginal densities are given by

$$\overline{p}_0(x^n) = \int_0^\infty p_{0,\sigma}(x^n)\pi_\sigma(\sigma)\,\mathrm{d}\sigma,$$

$$\overline{p}_1(x^n) = \int_0^\infty \int_{-\infty}^\infty p_{1,\sigma,\delta}(x^n)\pi_\delta(\delta)\pi_\sigma(\sigma)\,\mathrm{d}\delta\,\mathrm{d}\sigma = \int_0^\infty p_{1,\sigma}(x^n)\pi_\sigma(\sigma)\,\mathrm{d}\sigma.$$

Our results in Section 4.5 imply that — under a slight, natural restriction on the stopping rules allowed — the Bayes factor $\overline{p}_1(x^n)/\overline{p}_0(x^n)$ is truly robust to optional stopping in the above frequentist sense. That is, (4.7) will hold for all $P \in H_0$, i.e. all $\sigma > 0$, and not just 'on average'. Thus, we can give Type I error guarantees irrespective of the true value of σ. Similarly, strong calibration in the sense of Section 4.2.2 holds for all $P \in \mathcal{H}_0$. The use of a Cauchy prior is not essential in this construction; the result will continue to hold for any proper prior on δ, including point priors that put all mass on a single value of δ.

As we show in Section 4.5, these results extend to a variety of settings, namely whenever H_0 and H_1 share a common so-called group invariance. In the t-test example, it is a scale invariance — effectively this means that for all δ, all σ, the distributions of

$$X_1, \ldots, X_n \text{ under } p_{1,\sigma,\delta}, \text{ and } \sigma X_1, \ldots, \sigma X_n \text{ under } p_{1,1,\delta}, \text{ coincide.} \tag{4.9}$$

For other models, one could have a translation invariance; for the full normal family, one has both translation and scale invariance; for yet other models, one might have a rotation invariance, and so on. Each such invariance is expressed as a *group* — a set equipped with a binary operation that satisfies certain axioms. The group corresponding to scale invariance is the set of positive

reals, and the operator is scalar multiplication or equivalently division; similarly, the group corresponding to translation invariance is the set of all reals, and the operation is addition.

In the general case, one starts with a group G that satisfies certain further restrictions (detailed in Section 4.5), a model $\{p_{1,g,\theta} : g \in G, \theta \in \Theta\}$ where g represents the invariant parameter (vector) and the parameterization must be such that the analogue of (4.9) holds. In the example above $g = \sigma$ is the variance and θ is set to $\delta := \mu/\sigma$. One then singles out a special value of θ, say θ_0, one sets $H_0 := \{p_{1,g,\theta_0} : g \in G\}$; within H_1 one puts an arbitrary prior on θ. For every group invariance, there exists a corresponding *right Haar prior* on G; one equips both models with this prior on G. Theorem 4.8 and 4.9 imply that in all models constructed this way, we have strong calibration and Type-I error control uniformly for all $g \in G$. While this is hinted at in several papers (e.g. (Bayarri et al., 2016; Dass and Berger, 2003)) and the special case for the Bayesian t-test was implicitly proven in earlier work by Lai, 1976, it seems to never have been proven formally in general before.

Our results thus imply that in some situations (group invariance) with composite null hypotheses, Type-I error control for all $P \in H_0$ under optional stopping is possible with Bayes factors. What about Type-II error control and composite null hypotheses that do *not* satisfy a group structure? This is partially addressed by the *safe testing* approach of Grünwald, De Heide and Koolen, 2019 (see also Howard et al., 2018b for a related approach). They show that for completely arbitrary H_0 and H_1, for any given prior π_1 on H_1, there exists a corresponding prior π_0 on H_0, the *reverse information projection prior*, so that, for all $P \in H_0$, one has Type-I error guarantees under frequentist *optional continuation*, a weakening of the idea of optional stopping. Further, if one wants to get control of Type-II error guarantees under optional stopping/continuation, one can do so by first choosing another special prior π_1^* on H_1 and picking the corresponding π_0^* on H_0. Essentially, like in 'default' or 'objective' Bayes approaches, one chooses special priors in lieu of a subjective choice; but the priors one ends up with are sometimes quite different from the standard default priors, and, unlike these, allow for frequentist error control under optional stopping.

4.4 The General Case

Let $(\Omega, \mathcal{F})$ be a measurable space. Fix some $m \geq 0$ and consider a sequence of functions $X_{m+1}, X_{m+2}, \ldots$ on Ω so that each X_n, $n > m$ takes values in some fixed set ('outcome space') $\mathcal{X}$ with associated σ-algebra Σ. When working with proper priors we invariably take $m = 0$ and then we define $X^n := (X_1, X_2, \ldots, X_n)$ and we let $\Sigma^{(n)}$ be the n-fold product algebra of Σ. When working with improper priors it turns out to be useful (more explanation further below) to take $m > 0$ and define an *initial sample* random variable $\langle X^{(m)} \rangle$ on Ω, taking values in some set $\langle \mathcal{X}^m \rangle \subseteq \mathcal{X}^m$ with associated σ-algebra $\langle \Sigma^{(m)} \rangle$. In that case we set, for $n \geq m$, $\langle \mathcal{X}^n \rangle = \{x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n : x^m = (x_1, \ldots, x_m) \in \langle \mathcal{X}^m \rangle\}$, and $X^n := (\langle X^{(m)} \rangle, X_{m+1}, X_{m+2}, \ldots, X_n)$ and we let $\Sigma^{(n)}$ be $\langle \Sigma^{(m)} \rangle \times \prod_{j=m+1}^{n} \Sigma$. In either case, we let $\mathcal{F}_n$ be the σ-algebra (relative to Ω) generated by $(X^n, \Sigma^{(n)})$. Then $(\mathcal{F}_n)_{n=m,m+1,\ldots}$ is a filtration relative to $\mathcal{F}$ and if we equip $(\Omega, \mathcal{F})$ with a distribution P then $\langle X^{(m)} \rangle, X_{m+1}, X_{m+2}, \ldots$ becomes a random process adapted to $\mathcal{F}$. A *stopping time* is now generalized to be a function $\tau : \Omega \to \{m+1, m+2, \ldots\} \cup \{\infty\}$ such that for each $n > m$, the event $\{\tau = n\}$ is $\mathcal{F}_n$-measurable; note that we only consider

stopping after m initial outcomes. Again, for a given stopping time τ and sequence of data $x^n = (x_1, \ldots, x_n)$, we say that x^n *is compatible with* τ if it satisfies $X^n = x^n \Rightarrow \tau = n$, i.e. $\{\omega \in \Omega \mid X^n(\omega) = x^n\} \subset \{\omega \in \Omega \mid \tau(\omega) = n\}$.

H_0 and H_1 are now sets of probability distributions on $(\Omega, \mathcal{F})$. Again one writes $H_j = \{P_{\theta|j} \mid \theta \in \Theta_j\}$ where now the parameter sets Θ_j (which, however, could itself be infinite-dimensional) are themselves equipped with suitable σ-algebras.

We will still represent both H_0 and H_1 by unique measures $\overline{P}_0$ and $\overline{P}_1$ respectively, which we now allow to be based on (5.5) with improper priors π_0 and π_1 that may be infinite measures. As a result $\overline{P}_0$ and $\overline{P}_1$ are positive real measures that may themselves be infinite. We also allow $\mathcal{X}$ to be a general (in particular uncountable) set. Both non-integrability and uncountability cause complications, but these can be overcome if suitable Radon-Nikodym derivatives exist. To ensure this, we will assume that for all $n \geq \max\{m, 1\}$, for all $k \in \{0, 1\}$ and $\theta \in \Theta_k$, $P^{(n)}_{\theta|k}$, $\overline{P}^{(n)}_0$ and $\overline{P}^{(n)}_1$ are all mutually absolutely continuous and that the measures $\overline{P}^{(n)}_1$ and $\overline{P}^{(n)}_0$ are σ-finite. Then there also exists a measure ρ on $(\Omega, \mathcal{F})$ such that, for all such n, $\overline{P}^{(n)}_1$, $\overline{P}^{(n)}_0$ and $\rho^{(n)}$ are all mutually absolutely continuous: we can simply take $\rho^{(n)} = \overline{P}^{(n)}_0$, but in practice, it is often possible and convenient to take ρ such that $\rho^{(n)}$ is the Lebesgue measure on $\mathbb{R}^n$, which is why we explicitly introduce ρ here.

The absolute continuity conditions guarantee that all required Radon-Nikodym derivatives exist. Finally, we assume that the posteriors $\pi_k(\Theta_k \mid x^m)$ (as defined in the standard manner in (4.12) below; when $m = 0$ these are just the priors) are proper probability measures (i.e. they integrate to 1) for all $x^m \in \langle \mathcal{X}^m \rangle$. This final requirement is the reason why we sometimes need to consider $m > 0$ and nonstandard sample spaces $\langle \mathcal{X}^n \rangle$ in the first place: in practice , one usually starts with the standard setting of a $(\Omega, \mathcal{F})$ where $m = 0$ and all X_i have the same status. In all practical situations with improper priors π_0 and/or π_1 that we know of, there is a smallest finite j and a set $\mathcal{X}^\circ \subset \mathcal{X}^j$ that has measure 0 under all probability distributions in $H_0 \cup H_1$, such that, restricted to the sample space $\mathcal{X}^j \setminus \mathcal{X}^\circ$, the measures $\overline{P}^{(j)}_1$ and $\overline{P}^{(j)}_0$ are σ-finite and mutually absolutely continuous, and the posteriors $\pi_k(\Theta_k \mid x^j)$ are proper probability measures. One then sets m to equal this j, and sets $\langle \mathcal{X}^m \rangle \coloneqq \mathcal{X}^m \setminus \mathcal{X}^\circ$, and the required properness will be guaranteed. Our initial sample $\langle X^{(m)} \rangle$ is a variation of what is called (for example, by Bayarri et al. (2012)) a *minimal sample*. Yet, the sample size of a standard minimal sample is itself a random quantity; by restricting $\mathcal{X}^m$ to $\langle \mathcal{X}^m \rangle$, we can take its sample size m to be constant rather than random, which will greatly simplify the treatment of optional stopping with group invariance; see Example 4.1 and 4.2 below.

We henceforth refer to the setting now defined (with m and initial space $\langle \mathcal{X}^m \rangle$ satisfying the requirements above) as the *general case*.

We need an analogue of (4.4) for this general case. If $\overline{P}_0$ and $\overline{P}_1$ are probability measures, then there is still a standard definition of conditional probability distributions $P(H \mid \mathcal{A})$ in terms of conditional expectation for any given σ-algebra $\mathcal{A}$; based on this, we can derive the required analogue in two steps. First, we consider the case that $\tau \equiv n$ for some $n > m$. We know in advance that we observe X^n for a fixed n: the appropriate $\mathcal{A}$ is then $\mathcal{F}_n$, $\pi(H \mid \mathcal{A})(\omega)$ is determined by $X^n(\omega)$ hence can be written as $\pi(H \mid X^n)$, and a straightforward calculation

gives that

$$\frac{\pi(H_1 \mid X^n = x^n)}{\pi(H_0 \mid X^n = x^n)} = \left(\left(\frac{d\overline{P}_1^{(n)}/d\rho^{(n)}}{d\overline{P}_0^{(n)}/d\rho^{(n)}} \right)(x^n) \right) \cdot \frac{\pi(H_1)}{\pi(H_0)} \tag{4.10}$$

where $(d\overline{P}_1^{(n)}/d\rho^{(n)})$ and $(d\overline{P}_0^{(n)}/d\rho^{(n)})$ are versions of the Radon-Nikodym derivatives defined relative to $\rho^{(n)}$. The second step is now to follow exactly the same steps as in the derivation of (4.4), replacing $\beta(X^n)$ by (4.10) wherever appropriate (we omit the details). This yields, for any n such that $\rho(\tau = n) > 0$, and for $\rho^{(n)}$-almost every x^n that is compatible with τ,

$$\overbrace{\frac{\pi(H_1 \mid x^n)}{\pi(H_0 \mid x^n)}}^{\gamma_n} = \frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)} = \overbrace{\left(\left(\frac{d\overline{P}_1^{(n)}/d\rho^{(n)}}{d\overline{P}_0^{(n)}/d\rho^{(n)}} \right)(x^n) \right)}^{\beta_n} \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{4.11}$$

where here, as below, for $n \geq m$, we abbreviate $\pi(H_k \mid X^n = x^n)$ to $\pi(H_k \mid x^n)$.

The above expression for the posterior is valid if $\overline{P}_0$ and $\overline{P}_1$ are probability measures; we will simply take it as the *definition* of the Bayes factor for the general case. Again this coincides with standard usage for the improper prior case. In particular, let us define the conditional posteriors and Bayes factors given $\langle X^{(m)} \rangle = x^m$ in the standard manner, by the formal application of Bayes' rule, for $k = 0, 1$ and measurable $\Theta_k' \subset \Theta_k$ and $\mathcal{F}$-measurable A,

$$\pi_k(\Theta_k' \mid x^m) := \frac{\int_{\Theta_k'} \frac{dP_{\theta|k}^{(m)}}{d\rho^{(m)}}(x^m) d\pi_k(\theta)}{\int_{\Theta_k} \frac{dP_{\theta|k}^{(m)}}{d\rho^{(m)}}(x^m) d\pi_k(\theta)} \tag{4.12}$$

$$\overline{P}_k(A \mid x^m) := \overline{P}_k(A \mid \langle X^{(m)} \rangle = x^m) := \int_{\Theta_k} P_{\theta|k}(A \mid \langle X^{(m)} \rangle = x^m) \, \mathrm{d}\pi_k(\theta \mid x^m), \tag{4.13}$$

where $P_{\theta|k}(A \mid \langle X^{(m)} \rangle = x^m)$ is defined as the value that (a version of) the conditional probability $P_{\theta|k}(A \mid \mathcal{F}_m)$ takes when $\langle X^{(m)} \rangle = x^m$, and is thus defined up to a set of $\rho^{(m)}$-measure 0.

With these definitions, it is straightforward to derive the following *coherence property*, which automatically holds if the priors are proper, and which in combination with (4.11) expresses that first updating on x^m and then on $x_{m+1}, \ldots, x_n$ (multiplying posterior odds given x^m with the Bayes factor for n outcomes given $X^m = x^m$, which we denote by $\beta_{n|m}$) has the same result as updating based on the full $x_1, \ldots, x_n$ at once (i.e. multiplying the prior odds with the unconditional Bayes factor β_n for n outcomes):

$$\frac{\pi(H_1 \mid X^n = x^n, \tau = n)}{\pi(H_0 \mid X^n = x^n, \tau = n)} = \overbrace{\left(\frac{d\overline{P}_1^{(n)}(\cdot \mid x^m)}{d\overline{P}_0^{(n)}(\cdot \mid x^m)}(x^n) \right)}^{\beta_{n|m}} \cdot \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)}. \tag{4.14}$$

4.4.1 τ-independence, general case

The general version of the claim that the posterior odds do not depend on the specific stopping rule that was used is now immediate, since the expression (4.11) for the Bayes factor does not depend on the stopping time τ.

4.4.2 Calibration, general case

We will now show that the calibration hypothesis continues to hold in our general setting. From here onward, we make the further reasonable assumption that for every $x^m \in \langle \mathcal{X}^m \rangle$, $\overline{P}_0(\tau = \infty \mid x^m) = \overline{P}_1(\tau = \infty \mid x^m) = 0$ (the stopping time is almost surely finite), and we define $\mathcal{T}_\tau := \{n \in \mathbb{N}_{>m} \mid \overline{P}_0(\tau = n) > 0\}$.

To prepare further, let $\{B_j \mid j \in \mathcal{T}_\tau\}$ be any collection of positive random variables such that for each $j \in \mathcal{T}_\tau$, B_j is $\mathcal{F}_j$-measurable. We can define the *stopped* random variable B_τ as

$$B_\tau := \sum_{j=0}^{\infty} \mathbb{1}_{\{\tau=j\}} B_j = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau=j\}} B_j, \tag{4.15}$$

where we note that, under this definition, B_τ is well-defined even if $\mathbf{E}_{\overline{P}_0}[\tau] = \infty$.

We can define the induced measures on the positive real line under the null and alternative hypothesis for any probability measure P on $(\Omega, \mathcal{F})$:

$$P^{[B_\tau]} : \mathcal{B}(\mathbb{R}_{>0}) \to [0,1] : A \mapsto P\left(B_\tau^{-1}(A)\right). \tag{4.16}$$

where $\mathcal{B}(\mathbb{R}_{>0})$ denotes the Borel σ-algebra of $\mathbb{R}_{>0}$. Note that, when we refer to $P^{[B_n]}$, this is identical to $P^{[B_\tau]}$ for the stopping time τ which on all of Ω stops at n. The following lemma is crucial for passing from fixed-sample size to stopping-rule based results.

Lemma 1. *Let $\mathcal{T}_\tau$ and $\{B_n \mid n \in \mathcal{T}_\tau\}$ be as above. Consider two probability measures P_0 and P_1 on $(\Omega, \mathcal{F})$. Suppose that for all $n \in \mathcal{T}_\tau$, the following* fixed-sample size calibration property *holds:*

$$\text{for some fixed } c > 0,\ P_0{}^{[B_n]}\text{-almost all } b : \frac{P_1(\tau = n)}{P_0(\tau = n)} \cdot \frac{\mathrm{d}P_1{}^{[B_n]}(\cdot \mid \tau = n)}{\mathrm{d}P_0{}^{[B_n]}(\cdot \mid \tau = n)}(b) = c \cdot b. \tag{4.17}$$

Then we have

$$\text{for } P_0{}^{[B_\tau]}\text{-almost all } b : \frac{\mathrm{d}P_1{}^{[B_\tau]}}{\mathrm{d}P_0{}^{[B_\tau]}}(b) = c \cdot b. \tag{4.18}$$

The proof is in Section 4.B in the supplementary material.

In this subsection we apply this lemma to the measures $\overline{P}_k(\cdot \mid x^m)$ for arbitrary fixed $x^m \in \langle \mathcal{X}^m \rangle$, with their induced measures $\overline{P}_0^{[\gamma_\tau]}(\cdot \mid x^m), \overline{P}_1^{[\gamma_\tau]}(\cdot \mid x^m)$ for the *stopped posterior odds* γ_τ. Formally, the posterior odds γ_n as defined in (4.11) constitute a random variable for each n, and, under our mutual absolute continuity assumption for $\overline{P}_0$ and $\overline{P}_1$, γ_n can be directly written

as $\frac{\mathrm{d}\overline{P}_1^{(n)}}{\mathrm{d}\overline{P}_0^{(n)}} \cdot \pi(H_1)/\pi(H_0)$. Since, by definition, the measures $\overline{P}_k(\cdot \mid x^m)$ are probability measures, the Radon-Nikodym derivatives in (4.17) and (4.18) are well-defined.

Lemma 2. *We have for all $x^m \in \langle \mathcal{X}^m \rangle$, all $n > m$:*

$$\text{for } \overline{P}_0^{[\gamma_n]}(\cdot \mid x^m)\text{-almost all } b : \frac{\overline{P}_1^{[\gamma_n]}(\tau = n \mid x^m)}{\overline{P}_0^{[\gamma_n]}(\tau = n \mid x^m)} \cdot \frac{\mathrm{d}\overline{P}_1^{[\gamma_n]}(\cdot \mid x^m)}{\mathrm{d}\overline{P}_0^{[\gamma_n]}(\cdot \mid x^m)}(b) = \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)} \cdot b. \tag{4.19}$$

Combining the two lemmas now immediately gives (4.20) below, and combining further with (4.14) and (4.11) gives (4.21):

Corollary 3. *In the setting considered above, we have for all $x^m \in \langle \mathcal{X}^m \rangle$:*

$$\text{for } \overline{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)\text{-almost all } b : \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)} \cdot \frac{\mathrm{d}\overline{P}_1^{[\gamma_\tau]}(\cdot \mid x^m)}{\mathrm{d}\overline{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)}(b) = b, \tag{4.20}$$

and also

$$\text{for } \overline{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)\text{-almost all } b : \frac{\pi(H_1)}{\pi(H_0)} \cdot \frac{\mathrm{d}\overline{P}_1^{[\gamma_\tau]}}{\mathrm{d}\overline{P}_0^{[\gamma_\tau]}}(b) = b, \tag{4.21}$$

In words, the posterior odds remain calibrated under any stopping rule τ which stops almost surely at times $m < \tau < \infty$.

For discrete and strictly positive measures with prior odds $\pi(H_1)/\pi(H_0) = 1$, we always have $m = 0$, and (4.20) is equivalent to (4.5). Note that $\overline{P}_0^{[\gamma_\tau]}(\cdot \mid x^m)$-almost everywhere in (4.20) is equivalent to $\overline{P}_1^{[\gamma_\tau]}(\cdot \mid x^m)$-almost everywhere because the two measures are assumed to be mutually absolutely continuous.

4.4.3 (Semi-)Frequentist Optional Stopping

In this section we consider our general setting as in the beginning of Section 4.4.2, i.e. with the added assumption that the stopping time is a.s. finite, and with $\mathcal{T}_\tau := \{j \in \mathbb{N}_{>m} \mid \overline{P}_0(\tau = j) > 0\}$.

Consider any initial sample $x^m \in \langle \mathcal{X}^m \rangle$ and let $\overline{P}_0 \mid x^m$ and $\overline{P}_1 \mid x^m$ be the conditional Bayes marginal distributions as defined in (4.13). We first note that, by Markov's inequality, for any nonnegative random variable Z on Ω with, for all $x^m \in \langle \mathcal{X}^m \rangle$, $\mathbf{E}_{\overline{P}_0 \mid x^m}[Z] \leq 1$, we must have, for $0 \leq \alpha \leq 1$, $\overline{P}_0(Z^{-1} \leq \alpha \mid x^m) \leq \mathbf{E}_{\overline{P}_0 \mid x^m}[Z]/\alpha^{-1} \leq \alpha$.

Proposition 4. *Let τ be any stopping rule satisfying our requirements. Let $\beta_{\tau|m}$ be the stopped Bayes factor given x^m, i.e., in accordance with (4.15), $\beta_{\tau|m} = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau=j\}} \beta_{j|m}$ with $\beta_{j|m}$ as given by (4.14). Then $\beta_{\tau|m}$ satisfies, for all $x^m \in \langle \mathcal{X}^m \rangle$, $\mathbf{E}_{\overline{P}_0 \mid x^m}[\beta_{\tau|m}] \leq 1$, so that, by the reasoning above, $\overline{P}_0(\frac{1}{\beta_{\tau|m}} \leq \alpha \mid x^m) \leq \alpha$.*

Proof. We have

$$\mathbf{E}_{\overline{P}_0|x^m}[\gamma_\tau] = \int b\overline{P}_0^{[\gamma_\tau]}(\mathrm{d}b \mid x^m) = \int \frac{\mathrm{d}\overline{P}_1^{[\gamma_\tau]}(b \mid x^m)}{\mathrm{d}\overline{P}_0^{[\gamma_\tau]}(b \mid x^m)} \cdot \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)} \overline{P}_0^{[\gamma_\tau]}(\mathrm{d}b \mid x^m) = \frac{\pi(H_1 \mid x^m)}{\pi(H_0 \mid x^m)},$$

where the first equality follows by definition of expectation, the second follows from Corollary 3, and the third follows from the fact that the integral equals 1.

But now note that

$$\beta_{\tau|m} = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau=j\}}\beta_{j|m} = \sum_{j=m+1}^{\infty} \mathbb{1}_{\{\tau=j\}}\gamma_j \cdot \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)} = \gamma_\tau \cdot \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)},$$

where the second equality follows from (4.14) together with the first equality in (4.11). Combining the two equations we get:

$$\mathbf{E}_{\overline{P}_0|x^m}\left[\beta_{\tau|m}\right] = \mathbf{E}_{\overline{P}_0|x^m}\left[\gamma_\tau \cdot \frac{\pi(H_0 \mid x^m)}{\pi(H_1 \mid x^m)}\right] = 1.$$

□

The desired result now follows by plugging in a particular stopping rule: let $S : \bigcup_{i=m+1}^{\infty} \mathcal{X}^i \to \{0,1\}$ be the frequentist sequential test defined by setting, for all $n > m$, $x^n \in \langle \mathcal{X}^n \rangle$: $S(x^n) = 1$ if and only if $\beta_{n|m} \geq 1/\alpha$.

Corollary 5. *Let $t^* \in \{m+1, m+2, \ldots\} \cup \{\infty\}$ be the smallest $t^* > m$ for which $\beta_{t|m}^{-1} \leq \alpha$. Then for arbitrarily large T, when applied to the stopping rule $\tau := \min\{T, t^*\}$, we find that*

$$\overline{P}_0(\exists n, m < n \leq T : S(X^n) = 1 \mid x^m) = \overline{P}_0(\exists n, m < n \leq T : \beta_{n|m}^{-1} \leq \alpha \mid x^m) \leq \alpha.$$

The corollary implies that the test S is robust under optional stopping in the frequentist sense relative to H_0 (Definition 4.1). Note that, just as in the simple case, the setting is really just 'semi-frequentist' whenever H_0 is not a singleton.

4.5 Optional stopping with group invariance

Whenever the null hypothesis is composite, the previous results only hold under the marginal distribution $\overline{P}_0$ or, in the case of improper priors, under $\overline{P}_0(\cdot \mid X^m = x^m)$. When a group structure can be imposed on the outcome space and (a subset of the) parameters that is joint to H_0 and H_1, stronger results can be derived for calibration and frequentist optional stopping. Invariably, such parameters function as *nuisance parameters* and our results are obtained if we equip them with the so-called *right Haar prior* which is usually improper. Below we show how we then obtain results that simultaneously hold for *all* values of the nuisance parameters.

Such cases include many standard testing scenarios such as the (Bayesian variations of the) t-test, as illustrated in the examples below. Note though that our results do not apply to settings with improper priors for which no group structure exists. For example, if $P_{\theta|0}$ expresses that $X_1, X_2, \ldots$ are i.i.d. Poisson(θ), then from an objective Bayes or MDL point of view it makes sense to adopt Jeffreys' prior for the Poisson model; this prior is improper, allows initial sample size $m = 1$, but does not allow for a group structure. For such a prior we can only use the marginal results Corollary 3 and Corollary 5. Group theoretic preliminaries, such as definitions of a (topological) group, the right Haar measure, et cetera can be found in Section 4.A of the supplementary material.

4.5.1 Background for fixed sample sizes

Here we prepare for our results by providing some general background on invariant priors for Bayes factors with fixed sample size n on models with nuisance parameters that admit a group structure, introducing the right Haar measure, the corresponding Bayes marginals, and (maximal) invariants. We use these results in Section 4.5.2 to derive Lemma 7, which gives us a strong version of calibration for fixed n. The setting is extended to variable stopping times in Section 4.5.3, and then Lemma 7 is used in this extended setting to obtain our strong optional stopping results in Section 4.5.4 and 4.5.5.

For now, we assume a sample space $\langle \mathcal{X}^n \rangle$ that is locally compact and Hausdorff, and that is a subset of some product space $\mathcal{X}^n$ where $\mathcal{X}$ is itself locally compact and Hausdorff. This requirement is met, for example, when $\mathcal{X} = \mathbb{R}$ and $\langle \mathcal{X}^n \rangle = \mathcal{X}^n$. In practice, the space $\langle \mathcal{X}^n \rangle$ is invariably a subset of $\mathcal{X}^n$ where some null-set is removed for technical reasons that will become apparent below. We associate $\langle \mathcal{X}^n \rangle$ with its Borel σ-algebra which we denote as $\mathcal{F}_n$. Observations are denoted by the random vector $X^n = (X_1, \ldots, X_n) \in \langle \mathcal{X}^n \rangle$. We thus consider outcomes of fixed sample size, denoting these as $x^n \in \langle \mathcal{X}^n \rangle$, returning to the case with stopping times in Section 4.5.4 and 4.5.5.

From now on we let G be a locally compact group G that acts topologically and properly[4] on the right of $\langle \mathcal{X}^n \rangle$. As hinted to before, this proper action requirement sometimes forces the removal from $\mathcal{X}^n$ of some trivial set with measure zero under all hypotheses involved. This is demonstrated at the end of Example 4.1 below.

Let $P_{0,e}$ and $P_{1,e}$ (notation to become clear below) be two arbitrary probability distributions on $\langle \mathcal{X}^n \rangle$ that are mutually absolutely continuous. We will now generate hypothesis classes H_0 and H_1, both sets of distributions on $\langle \mathcal{X}^n \rangle$ with parameter space G, starting from $P_{0,e}$ and $P_{1,e}$, where $e \in G$ is the group identity element. The group action of G on $\langle \mathcal{X}^n \rangle$ induces a group action on these measures defined by

$$P_{k,g}(A) := (P_{k,e} \cdot g)(A) := P_{k,e}(A \cdot g^{-1}) = \int \mathbb{1}_{\{A\}}(x \cdot g)\, P_{k,e}(\mathrm{d}x) \tag{4.22}$$

for any set $A \in \mathcal{F}_n$, $k = 0, 1$. When applied to $A = \langle \mathcal{X}^n \rangle$, we get $P_{k,g}(A) = 1$, for all $g \in G$,

[4] A group acts properly on a set Y if the mapping $\psi : Y \times G \mapsto Y \times Y$ defined by $\psi(y, g) = (y \cdot g, y)$ is a proper mapping, i.e. the inverse image of ψ of each compact set in $Y \times Y$ is a compact set in $Y \times G$. (Eaton (1989), Definition 5.1)

whence we have created two sets of probability measures parameterized by g, i.e.,

$$H_0 \coloneqq \{P_{0,g} \mid g \in G\} \;\; ; \;\; H_1 \coloneqq \{P_{1,g} \mid g \in G\}. \tag{4.23}$$

In this context, $g \in G$, can typically be viewed as nuisance parameter, i.e. a parameter that is not directly of interest, but needs to be accounted for in the analysis. This is illustrated in Example 4.1 and Example 4.2 below. The examples also illustrate how to extend this setting to cases where there are more parameters than just $g \in G$ in either H_0 or H_1. We extend the whole setup to our general setting with non-fixed n in Section 4.5.4.

We use the right Haar measure for G as a prior to define the Bayes marginals:

$$\overline{P}_k(A) = \int_G \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{A\}} \, \mathrm{d}P_{k,g} \, \nu(\mathrm{d}g) \tag{4.24}$$

for $k = 0, 1$ and $A \in \mathcal{F}_n$. Typically, the right Haar measure is improper so that the Bayes marginals $\overline{P}_k$ are not integrable. Yet, in all cases of interest, they are (a) still σ-finite, and, (b), $\overline{P}_0$, $\overline{P}_1$ and all distributions $P_{k,g}$ with $k = 0, 1$ and $g \in G$ are mutually absolutely continuous; we will henceforth assume that (a) and (b) are the case.

Example 4.1 (continued) Consider the t-test of Example 4.1. For consistency with the earlier Example 4.1, we abbreviate for general measures P on $\langle \mathcal{X}^n \rangle$, $(\mathrm{d}P/\mathrm{d}\lambda)$ (the density of distribution P relative to Lebesgue measure on $\mathbb{R}^n$) to p. Normally, the one-sample t-test is viewed as a test between $H_0 = \{P_{0,\sigma} \mid \sigma \in \mathbb{R}_{>0}\}$ and $H_1' = \{P_{1,\sigma,\delta} \mid \sigma \in \mathbb{R}_{>0}, \delta \in \mathbb{R}\}$, but we can obviously also view it as test between H_0 and $H_1 = \{P_{1,\sigma}\}$ by integrating out the parameter δ to obtain

$$p_{1,\sigma}(x^n) = \int p_{1,\sigma,\delta}(x^n) \pi_\delta(\delta) \, \mathrm{d}\delta. \tag{4.25}$$

The nuisance parameter σ can be identified with the group of scale transformations $G = \{c \mid c \in \mathbb{R}_{>0}\}$. We thus let the sample space be $\langle \mathcal{X}^n \rangle = \mathbb{R}^n \setminus \{0\}^n$, i.e., we remove the measure-zero set $\{0\}^n$, such that the group action is proper on the sample space. The group action is defined by $x^n \cdot c = c\, x^n$ for $x^n \in \langle \mathcal{X}^n \rangle$, $c \in G$. Take $e = 1$ and let, for $k = 0, 1$, $P_{k,e}$ be the distribution with density $p_{k,1}$ as defined in (4.8) and (4.25). The measures $P_{0,g}$ and $P_{1,g}$ defined by (4.22) then turn out to have the densities $p_{0,\sigma}$ and $p_{1,\sigma}$ as defined above, with σ replaced by g. Thus, H_0 and H_1 as defined by (4.8) and (4.25) are indeed in the form (4.23) needed to state our results.

In most standard invariant settings, H_0 and H_1 share the same vector of nuisance parameters, and one can reduce H_0 and H_1 to (4.23) in the same way as above, by integrating out all other parameters; in the example above, the only non-nuisance parameter was δ. The scenario of Example 4.1 can be generalized to a surprisingly wide variety of statistical models. In practice we often start with a model $H_1 = \{P_{1,\gamma,\delta} : \gamma \in \Gamma, \theta \in \Theta\}$ that implicitly already contains a group structure, and we single out a special subset $\{P_1, \gamma, \theta_0 : \gamma \in \Gamma\}$; this is what we informally described in Example 4.1. More generally, we can start with potentially large (or even nonparametric) hypotheses

$$H_k' = \{P_{\theta'|k} : \theta' \in \Theta_k'\} \tag{4.26}$$

which at first are not related to any group invariance, but which we want to equip with an additional nuisance parameter determined by a group G acting on the data. We can turn this into an instance of the present setting by first choosing,for $k = 0, 1$, a proper prior density π_k on Θ'_k, and defining $P_{k,e}$ to equal the corresponding Bayes marginal, i.e.

$$P_{k,e}(A) \coloneqq \int P_{\theta'|k}(A)\ \mathrm{d}\pi_k(\theta'). \tag{4.27}$$

We can then generate $H_k = \{P_{k,g} \mid g \in G\}$ as in (4.22) and (4.23). In the example above, H'_1 would be the set of all Gaussians with a single fixed variance σ_0^2 and $\Theta'_1 = \mathbb{R}$ would be the set of all effect sizes δ, and the group G would be scale transformation; but there are many other possibilities. To give but a few examples, Dass and Berger, 2003 consider testing the Weibull vs. the log-normal model, the exponential vs. the log-normal, correlations in multivariate Gaussians, and Berger, Pericchi and Varshavsky, 1998 consider location-scale families and linear models where H_0 and H_1 differ in their error distribution. Importantly, the group G acting on the data induces groups G_k, $k = 0, 1$, acting on the parameter spaces, which depend on the parameterization. In our example, the G_k were equal to G, but, for example, if H_0 is Weibull and H_1 is log-normal, both given in their standard parameterizations, we get $G_0 = \{g_{0,b,c} \mid g_{0,b,c}(\beta, \gamma) = (b\beta^c, \gamma/c), b > 0, c > 0\}$ and $G_1 = \{g_{1,b,c} \mid g_{1,b,c}(\mu, \sigma) = (c\mu + \log(b), c\sigma), b > 0, c > 0\}$. Several more examples are given by Dass, 1998.

On the other hand, clearly not all hypothesis sets can be generated using the above approach. For instance, the hypothesis $H'_1 = \{P_{\mu,\sigma} \mid \mu = 1, \sigma > 0\}$ with $P_{\mu,\sigma}$ a Gaussian measure with mean μ and standard deviation σ cannot be represented as in (4.23). This is due to the fact that for $\sigma, \sigma' > 0$, $\sigma \neq \sigma'$, no element $g \in \mathbb{R}_{>0}$ exists such that for any measurable set $A \subseteq \langle \mathcal{X}^n \rangle$ the equality

$$P_{1,\sigma}(A) = P_{1,\sigma'}(A \cdot g^{-1})$$

holds. This prevents an equivalent construction of H'_1 in the form of (4.23).

We now turn to the main ingredient that will be needed to obtain results on optional stopping: the quotient σ-algebra.

Definition 4.2 (Eaton (1989), Chapter 2)**.** A group G acting on the right of a set Y induces an equivalence relation: $y_1 \sim y_2$ if and only if there exists $g \in G$ such that $y_1 = y_2 \cdot g$. This equivalence relation partitions the space in *orbits*: $O_y = \{y \cdot g \mid g \in G\}$, the collection of which is called the *quotient space* Y/G. There exists a map, the *natural projection*, from Y to the quotient space which is defined by $\varphi_Y : Y \to Y/G : y \mapsto \{y \cdot g \mid g \in G\}$, and which we use to define the *quotient σ-algebra*

$$\mathcal{G}_n = \{\varphi^{-1}_{\langle \mathcal{X}^n \rangle}(\varphi_{\langle \mathcal{X}^n \rangle}(A)) \mid A \in \mathcal{F}_n\}. \tag{4.28}$$

Definition 4.3 (Eaton (1989), Chapter 2)**.** A random element U_n on $\langle \mathcal{X}^n \rangle$ is *invariant* if for all $g \in G$, $x^n \in \langle \mathcal{X}^n \rangle$, $U_n(x^n) = U_n(x^n \cdot g)$. The random element U_n is *maximal invariant* if U_n is invariant and for all $y^n \in \langle \mathcal{X}^n \rangle$, $U_n(x^n) = U_n(y^n)$ implies $x^n = y^n \cdot g$ for some $g \in G$.

Thus, U_n is maximal invariant if and only if U_n is constant on each orbit, and takes different values on different orbits; $\varphi_{\langle \mathcal{X}^n \rangle}$ is thus an example of a maximal invariant. Note that any

maximal invariant is $\mathcal{G}_n$-measurable. The importance of this quotient σ-algebra $\mathcal{G}_n$ is the following evident fact:

Proposition 6. *For fixed $k \in \{0,1\}$, every invariant U_n has the same distribution under all $P_{k,g}$, $g \in G$.*

Chapter 2 of (Eaton, 1989) provides several methods and examples how to construct a concrete maximal invariant, including the first two given below. Since β_n is invariant under the group action of G (see below), β_n is an example of an invariant, although not necessarily of a maximal invariant.

Example 4.1 (continued) Consider the setting of the one-sample t-test as described above in Example 4.1. A maximal invariant for $x^n \in \langle \mathcal{X}^n \rangle$ is

$$U_n(x^n) = (x_1/|x_1|, x_2/|x_1|, \ldots, x_n/|x_1|).$$

Example 4.2. A second example, with a group invariance structure on two parameters, is the setting of the two-sample t-test with the right Haar prior (which coincides here with Jeffreys' prior) $\pi(\mu, \sigma) = 1/\sigma$ (see Rouder et al. (2009) for details): the group is $G = \{(a, b) \mid a > 0, b \in \mathbb{R}\}$. Let the sample space be $\langle \mathcal{X}^n \rangle = \mathbb{R}^n \setminus \text{span}(e_n)$, where e_n denotes a vector of ones of length n (this is to exclude the measure-zero line for which the $s(x^n)$ is zero), and define the group action by $x^n \cdot (a, b) = ax^n + be_n$ for $x^n \in \langle \mathcal{X}^n \rangle$. Then (Eaton (1989), Example 2.15) a maximal invariant for $x^n \in \langle \mathcal{X}^n \rangle$ is $U_n(x^n) = (x^n - \overline{x}e_n)/s(x^n)$, where $\overline{x}$ is the sample mean and $s(x^n) = \left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^{1/2}$.

However, we can also construct a maximal invariant similar to the one in Example 4.1, which gives a special status to an initial sample:

$$U_n\left(X^n\right) = \left(\frac{X_2 - X_1}{|X_2 - X_1|}, \frac{X_3 - X_1}{|X_2 - X_1|}, \ldots, \frac{X_n - X_1}{|X_2 - X_1|}\right), \quad n \geq 2.$$

4.5.2 Relatively Invariant Measures and Calibration for Fixed n

Let U_n be a maximal invariant, taking values in the measurable space $(\mathcal{U}_n, \mathcal{G}_n)$. Although we have given more concrete examples above, it follows from the results of Andersson, 1982 that, in case we do not know how to construct a U_n, we can always take $U_n = \varphi_{\langle \mathcal{X}^n \rangle}$, the natural projection. Since we assume mutual absolute continuity, the Radon-Nikodym derivative $\mathrm{d}P_{1,g}^{[U_n]} / \mathrm{d}P_{0,g}^{[U_n]}$ must exist and we can apply the following theorem (note it is here that the use of *right* Haar measure is crucial; a different result holds for the left Haar measure):[5]

Theorem Berger, Pericchi and Varshavsky, 1998, Theorem 2.1 Under our previous definitions of and assumptions on G, $P_{k,g}$, $\overline{P}_k$ let $\beta(x^n) := \overline{P}_1(x^n)/\overline{P}_0(x^n)$ be the Bayes factor based on x^n. Let U_n be a maximal invariant as above, with (adopting the notation of (4.16)) marginal

[5] This theorem requires that there exists some relatively invariant measure μ on $\langle \mathcal{X}^n \rangle$ such that for $k = 0, 1$, $g \in G$, the $P_{k,g}$ all have a density relative to μ. Since the Bayes marginal $\overline{P}_0$ based on the right Haar prior is easily seen to be such a relatively invariant measure, the conditions for the theorem apply.

measures $P_{k,g}^{[U_n]}$, for $k = 0, 1$ and $g \in G$. There exists a version of the Radon-Nikodym derivative such that we have for all $g \in G$, all $x^n \in \langle \mathcal{X}^n \rangle$,

$$\frac{\mathrm{d}P_{1,g}^{[U_n]}}{\mathrm{d}P_{0,g}^{[U_n]}}\left(U_n(x^n)\right) = \beta(x^n). \tag{4.29}$$

As a first consequence of the theorem above, we note (as did Berger, Pericchi and Varshavsky (1998)) that the Bayes factor $\beta_n := \beta(X^N)$ is $\mathcal{G}_n$-measurable (it is constant on orbits) , and thus *it has the same distribution under $P_{0,g}$ and $P_{1,g}$ for all $g \in G$*. The theorem also implies the following crucial lemma:

Lemma 7. [Strong Calibration for Fixed n] *Under the assumptions of the theorem above, let U_n be a maximal invariant and let V_n be a $\mathcal{G}_n$-measurable binary random variable with $P_{0,g}(V_n = 1) > 0$, $P_{1,g}(V_n = 1) > 0$. Adopting the notation of (4.16), we can choose the Radon-Nikodym derivative $\mathrm{d}P_{1,g}^{[\beta_n]}(\cdot \mid V_n = 1)/\mathrm{d}P_{0,g}^{[\beta_n]}(\cdot \mid V_n = 1)$ so that we have, for all $x^n \in \langle \mathcal{X}^n \rangle$:*

$$\frac{P_{1,g}(V_n = 1)}{P_{0,g}(V_n = 1)} \cdot \frac{\mathrm{d}P_{1,g}^{[\beta_n]}(\cdot \mid V_n = 1)}{\mathrm{d}P_{0,g}^{[\beta_n]}(\cdot \mid V_n = 1)}\left(\beta_n(x^n)\right) = \beta_n(x^n), \tag{4.30}$$

where for the special case with $P_{k,g}(V_n = 1) = 1$, we get $\dfrac{\mathrm{d}P_{1,g}^{[\beta_n]}}{\mathrm{d}P_{0,g}^{[\beta_n]}}\left(\beta_n(x^n)\right) = \beta_n(x^n)$.

4.5.3 Extending to Our General Setting with Non-Fixed Sample Sizes

We start with the same setting as above: a group G on sample space $\langle \mathcal{X}^n \rangle \subset \mathcal{X}^n$ that acts topologically and properly on the right of $\langle \mathcal{X}^n \rangle$; two distributions $P_{0,e}$ and $P_{1,e}$ on $(\langle \mathcal{X}^n \rangle, \mathcal{F}_n)$ that are used to generate H_0 and H_1, and Bayes marginal measures based on the right Haar measure $\overline{P}_0$ and $\overline{P}_1$, which are both σ-finite. We now denote H_k as $H_k^{(n)}$, $P_{k,e}$ as $P_{k,e}^{(n)}$ and $\overline{P}_k$ as $\overline{P}_k^{(n)}$, all $P \in H_0^{(n)} \cup H_1^{(n)}$ are mutually absolutely continuous.

We now extend this setting to our general random process setting as specified in the beginning of Section 4.4.2 by further assuming that, for the same group G, for some $m > 0$, the above setting is defined for each $n \geq m$. To connect the $H_k^{(n)}$ for all these n, we further assume that there exists a subset $\langle \mathcal{X}^m \rangle \subset \mathcal{X}^m$ that has measure 1 under $P_{k,e}^{(n)}$ (and hence under all $P_{g,e}^{(n)}$) such that for all $n \geq m$:

1. We can write $\langle \mathcal{X}^n \rangle = \{x^n \in \mathcal{X}^n : (x_1, \ldots, x_m) \in \langle \mathcal{X}^m \rangle\}$.
2. For all $x^n \in \langle \mathcal{X}^n \rangle$, the posterior $\nu \mid x^n$ based on the right Haar measure ν is proper.
3. The probability measures $P_{k,e}^{(n)}$ and $P_{k,e}^{(n+1)}$ satisfy Kolmogorov's compatibility condition for a random process.
4. The group action $\cdot$ on the measures $P_{k,e}^{(n)}$ and $P_{k,e}^{(n+1)}$ is compatible, i.e. for every $n > 0$, for every $A \in \mathcal{F}_n$, every $g \in G$, $k \in \{0, 1\}$, we have $P_{k,g}^{(n+1)}(A) = P_{k,g}^{(n)}(A)$.

Requirement 4. simply imposes the condition that the group action considered is the same for all $n \in \mathbb{N}$. As a consequence of 3. and 4., the probability measures $P_{k,g}^{(n)}$ and $P_{k,g}^{(n+1)}$ satisfy Kolmogorov's compatibility condition for all $g \in G$, $k \in \{0,1\}$ which means that there exists a probability measure $P_{k,g}$ on $(\Omega, \mathcal{F})$ (under which $\langle X^{(m)} \rangle, X_{m+1}, X_{m+2}, \ldots$ is a random process), defined as in the beginning of Section 4.4, whose marginals for $n \geq m$ coincide with $P_{k,g}^{(n)}$, and there exist measures $\overline{P}_0$ and $\overline{P}_1$ on $(\Omega, \mathcal{F})$ whose marginals for $n \geq m$ coincide with $\overline{P}_0^{(n)}$ and $\overline{P}_1^{(n)}$. We have thus defined a set H_0 and H_1 of hypotheses on $(\Omega, \mathcal{F})$ and the corresponding Bayes marginals $\overline{P}_0$ and $\overline{P}_1$ and are back in our general setting. It is easily verified that the 1- and 2-sample Bayesian t-tests both satisfy all these assumptions: in Example 4.1, take $m = 1$ and $\langle \mathcal{X}^m \rangle = \mathbb{R} \setminus \{0\}$; in Example 4.5.1, take $m = 2$ and $\langle \mathcal{X}^m \rangle = \mathbb{R}^2 \setminus \{(a,a) : a \in \mathbb{R}\}$. The conditions can also be verified for the variety of examples considered by Berger, Pericchi and Varshavsky (1998) and Bayarri et al., 2012. In fact, our initial sample $x^m \in \langle \mathcal{X}^m \rangle$ is a variation of what they call a *minimal sample*; by excluding 'singular' outcomes from $\mathcal{X}^m$ to ensure that the group acts properly on $\langle \mathcal{X}^m \rangle$, we can guarantee that the initial sample is of fixed size. The size of the minimal sample can be larger, on a set of measure 0 under all $P \in H_0 \cup H_1$, e.g. if, in Example 4.5.1, $X_1 = X_2$. We chose to ensure a fixed size m since it makes the extension to random processes considerably easier.

In Section 4.5.1, underneath Example 4.1 we already outlined how a composite alternative hypothesis can be reduced to a hypothesis with just a free nuisance parameter (or parameter vector) $g \in G$, by putting a proper prior on all other parameters and integrating them out. A similar construction for a single parameter alternative hypothesis in the form of (4.23) can be applied in the non-fixed sample size case.

4.5.4 Strong Calibration

Consider the setting, definitions and assumptions of the previous subsection, with the additional assumptions and definitions made in the beginning of Section 4.4.3, in particular the assumption of a.s. finite stopping time. For simplicity, from now on, we shall also assume equal prior odds, $\pi(H_0) = \pi(H_1) = 1/2$. We will now show a strong calibration theorem for the Bayes factors $\beta_n = (d\overline{P}_0^{(n)})/(d\overline{P}_1^{(n)})(X^n)$ defined in terms of the Bayes marginals $\overline{P}_0$ and $\overline{P}_1$ with the right Haar prior. Thus β_τ is defined as in (4.15) with β in the role of B.

Theorem 4.8 (Strong calibration under optional stopping)**.** *Let τ be a stopping time satisfying our requirements, such that additionally, for each $n > m$, the event $\{\tau = n\}$ is $\mathcal{G}_n$-measurable. Then, adopting the notation of (4.16), for all $g \in G$, for $P_{0,g}^{[\beta_\tau]}$-almost every $b > 0$, we have:*

$$\frac{\mathrm{d}P_{1,g}^{[\beta_\tau]}}{\mathrm{d}P_{0,g}^{[\beta_\tau]}}(b) = b.$$

That means that the posterior odds remain calibrated under every stopping rule τ adapted to the quotient space filtration $\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots$, under all $P_{0,g}$.

Proof. Fix some $g \in G$. We simply first apply Lemma 7 with $V_n = \mathbb{1}_{\{\tau = n\}}$, which gives that the

premise (4.17) of Lemma 1 holds with $c = 1$ and β_n in the role of B_n (it is here that we need that τ_n is $\mathcal{G}_n$-measurable, otherwise we could not apply Lemma 7 with the required definition of V_n). We can now use Lemma 1 with $P_{0,g}$ in the role of P_0 to reach the desired conclusion for the chosen g. Since this works for all $g \in G$, the result follows. □

Example 4.1, Continued: Admissible and Inadmissible Stopping Rules We obtain strong calibration for the one-sample t-test with respect to the nuisance parameter σ (see Example 4.1 above) when the stopping rule is adapted to the quotient filtration $\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots$. Under each $P_{k,g} \in H_k$, the Bayes factors $\beta_m, \beta_{m+1}, \ldots$ define a random process on Ω such that each β_n is $\mathcal{G}_n$-measurable. This means that a stopping time defined in terms of a rule such as 'stop at the smallest t at which $\beta_t > 20$ or $t = 10^6$' is allowed in the result above. Moreover, if the stopping rule is a function of a sequence of maximal invariants, like $x_1/|x_1|, x_2/|x_1|, \ldots$, it is adapted to the filtration $\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots$ and we can likewise apply the result above. On the other hand, this requirement is violated, for example, by a stopping rule that stops when $\sum_{i=1}^{j}(x_i)^2$ exceeds some fixed value, since such a stopping rule explicitly depends on the scale of the sampled data.

4.5.5 Frequentist optional stopping

The special case of the following result for the one-sample Bayesian t-test was proven in the master's thesis (Hendriksen, 2017). Here we extend the result to general group invariances.

Theorem 4.9 (Frequentist optional stopping for composite null hypotheses with group invariance)**.** *Under the same conditions as in Section 4.5.4, let τ be a stopping time such that, for each $n > m$, the event $\{\tau = n\}$ is $\mathcal{G}_n$-measurable. Then, adopting the notation of (4.16), for all $g \in G$, the stopped Bayes factor satisfies $\mathbf{E}_{P_{0,g}}[\beta_\tau] = \int_{\mathbb{R}_{>0}} c \, \mathrm{d}P_{0,g}^{[\beta_\tau]}(c) = 1$, so that, by the reasoning above Proposition 4, we have* for all $g \in G$: $P_{0,g}(\frac{1}{\beta_\tau} \leq \alpha) \leq \alpha$.

Proof. We have

$$\int_{\mathbb{R}_{>0}} c \, dP_{0,g}^{[\beta_\tau]}(c) = \int_{\mathbb{R}_{>0}} \frac{\mathrm{d}P_{1,g}^{[\beta_\tau]}}{\mathrm{d}P_{0,g}^{[\beta_\tau]}}(c) dP_{0,g}^{[\beta_\tau]}(c) = \int_{\mathbb{R}_{>0}} dP_{1,g}^{[\beta_\tau]}(c) = 1.$$

where the first equality follows directly from Theorem 4.8 and the final equality follows because $P_{1,g}$ is a probability measure, integrating to 1. □

Analogously to Corollary 5, the desired result now follows by plugging in a particular stopping rule: let $S : \bigcup_{i=m}^{\infty} \mathcal{X}^i \to \{0, 1\}$ be the frequentist sequential test defined by setting, for all $n > m$, $x^n \in \langle \mathcal{X}^n \rangle$: $S(x^n) = 1$ if and only if $\beta_n \geq 1/\alpha$.

Corollary 10. *Let $t^* \in \{m+1, m+2, \ldots\} \cup \{\infty\}$ be the smallest $t^* > m$ for which $\beta_{t^*}^{-1} \leq \alpha$. Then for arbitrarily large T, when applied to the stopping rule $\tau := \min\{T, t^*\}$, we find that* for all $g \in G$:

$$P_{0,g}(\exists n, m < n \leq T : S(X^n) = 1 \mid x^m) = P_{0,g}(\exists n, m < n \leq T : \beta_n^{-1} \leq \alpha \mid x^m) \leq \alpha.$$

The corollary implies that the test S is robust under optional stopping in the frequentist sense relative to H_0 (Definition 4.1).

Example 4.1 (continued) When we choose a stopping rule that is $(\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots)$-measurable, the hypothesis test is robust under (semi-)frequentist optional stopping. This holds for example, for the one- and two-sample t-test (Rouder et al., 2009), Bayesian ANOVA (Rouder et al., 2012), and Bayesian linear regression (Liang et al., 2008). Again, for stopping rules that are not $(\mathcal{G}_m, \mathcal{G}_{m+1}, \ldots)$-measurable, robustness under frequentist optional stopping cannot be guaranteed and could reasonably be presumed to be violated. The violation of robustness under optional stopping is hard to demonstrate experimentally as frequentist Bayes factor tests are usually quite conservative in approaching the asymptotic significance level α.

4.6 Concluding Remarks

We have identified three types of 'handling optional stopping': τ-independence, calibration and semi-frequentist. We extended the corresponding definitions and results to general sample spaces with potentially improper priors. For the special case of models H_0 and H_1 sharing a nuisance parameter with a group invariance structure, we showed stronger versions of calibration and semi-frequentist robustness to optional stopping. A couple of remarks are in order. First, one of the remarkable properties of the right Haar prior is that, under some additional conditions on $P_{0,g}$ and $P_{1,g}$ in (4.22), $\beta_m = \beta(x^m) = 1$ for all $x^m \in \langle \mathcal{X}^m \rangle$, implying that equal prior odds lead to equal posterior odds after a minimal sample, no matter what the minimal sample is (Berger, Pericchi and Varshavsky, 1998). One might conjecture that our results rely on this property, but this is not the case: in general, one can have $\beta(x^m) \neq 1$, yet our results still hold. For example, in the Bayesian t-test, Example 4.1, $m = 1$ and $\beta(x^1) = 1$ can be guaranteed only if the prior π_δ on δ is symmetric around 0; but our calibration and frequentist robustness results hold irrespective of whether it is symmetric or not.

Secondly, in multiple-parameter problems, the suitable transformation group acting on the parameter space may not be unique, in which case there are multiple possible right Haar priors, see Example 1.2 and 1.3 in (Berger, Bernardo, Sun et al., 2015) and (Berger, Sun et al., 2008). However, in all examples we considered and further know of, this does not lead to ambiguity, because different transformation groups give rise to different sets H_0 of invariant null hypotheses.

As a third remark, it is worth noting that — as is immediate from the proofs — all our group-invariance results continue to hold in the setting with H'_k as in (4.26), and the definition of the Bayes marginal $P_{k,e}$ relative to θ' as in (4.27) replaced by a probability measure on $(\Omega, \mathcal{F})$ that is not necessarily of the Bayes marginal form. The results work for any probability measure; in particular one can take the alternatives for the Bayes marginal with proper prior that are considered in the the minimum description length and sequential prediction literature (Barron, Rissanen and Yu, 1998; Grünwald, 2007) under the name of *universal distribution* relative to $\{P_{\theta'} \mid \theta' \in \Theta'\}$; examples include the prequential or 'switch' distributions considered by Van der Pas and Grünwald, 2018.

As a fourth and final remark, a sizable fraction of Bayesian statisticians is wary of using improper priors at all. An important (though not the only) reason is that their use often leads to some form of the *marginalization paradox* described by Dawid, Stone and Zidek, 1973. It is thus useful to stress that in the context of Bayes factor hypothesis testing, the right Haar prior is immune at least to this particular paradox. In an informal nutshell, the marginalization paradox occurs if the following happens: (a) the Bayes posterior $\pi(\zeta \mid X^n)$ for the quantity of interest ζ based on prior $\pi(\zeta, g)$ with improper marginal on g, only depends on the data X^n through the maximal invariant U_n, i.e. $\pi(\zeta \mid X^n) = f(U_n(X^n))$ for some function f, yet (b) there exists no prior π' on ζ such that the corresponding posterior $\pi'(\zeta \mid U_n(X^n)) = f(U_n(X^n))$. In words, the result of Bayesian updating based on the full data X^n only depends on the maximal invariant U^n; but Bayesian updating directly based on U^n can never give the same result — a paradox indeed. While in general, this can happen even if g is equipped with the right Haar prior [Case 1, page 199](Dawid, Stone and Zidek, 1973), Berger et. al.'s Theorem 2.1 (reproduced in Section 4.5.2 in our chapter) implies that it does not occur in the context of Bayes factor testing, where $\zeta \in \{H_0, H_1\}$, and H_0 and H_1 are null and alternatives satisfying the requirements of Section 4.5. Berger's theorem expresses that for all values of the nuisance parameter $g \in G$, the likelihood ratio $\mathrm{d}P_{1,g}^{[U_n]} / \mathrm{d}P_{0,g}^{[U_n]}(U_n(X^n))$ based on $U_n(X^n)$ is equal to the Bayes factor based on X^n with the right Haar prior on g, so that the paradox cannot occur.

4.A Group theoretic preliminaries

We start with some group-theoretical preliminaries; for more details, see e.g. (Eaton, 1989; Wijsman, 1990; Andersson, 1982).

Definition 4.4 (Topological space)**.** A non-empty set S together with a fixed collection of subsets $\mathcal{T}$ is called a *topological space* $T = (S, \mathcal{T})$ if

1. $S, \varnothing \in \mathcal{T}$,
2. $U \cap V \in \mathcal{T}$ for any two sets $U, V \in \mathcal{T}$, and
3. $S_1 \cup S_2 \in \mathcal{T}$ for any collections of sets $S_1, S_2 \subseteq \mathcal{T}$.

The collection $\mathcal{T}$ is called a *topology* for S, and its members are called the *open sets* of T. A topological space T is called *Hausdorff* if for any two distinct points $x, y \in T$ there exist disjoint open subsets U, V of T containing one point each.

Definition 4.5 ((Local) compactness)**.** A topological space T is *compact* if every *open cover*, that is, every collection $\mathcal{C}$ of open sets of T

$$T = \bigcup_{U \in \mathcal{C}} U,$$

has a *finite subcover*: a finite subcollection $\mathcal{F} \in \mathcal{C}$ such that

$$T = \bigcup_{V \in \mathcal{F}} V.$$

It is *locally compact* if for every $x \in T$ there exist an open set U such that $x \in U$ and the closure of U, denoted by $\mathrm{Cl}(U)$, is compact, that is, the union of U and all its limit points in T is compact. We can also formulate this as each x having a neighborhood U such that $\mathrm{Cl}(U)$ is compact.

Example 4.3 (Locally compact Hausdorff spaces)**.** The reals $\mathbb{R}$ and the Euclidean spaces $\mathbb{R}^n$ together with the Euclidean topology (also called the *usual topology*) are *locally compact Hausdorff spaces*. $\mathbb{R}^n$ (for $n \in \mathbb{N}$) is locally compact because any open ball $B(x, r)$ has a compact closure $\mathrm{Cl}(B(x, r)) = \{y \in \mathbb{R}^n; d(x, y) \leq \varepsilon\}$, where $d(x, y)$ is the Euclidean metric. Any discrete space is locally compact and Hausdorff as well, as any singleton is a neighborhood that equals its closure, and it is compact only if it is finite. Infinite dimensional Banach spaces (function spaces) are for example not locally compact.

Definition 4.6 (Group)**.** A set G together with a binary operation $\circ$, often called the *group law*, is called a *group* when

1. there exists an identity element $e \in G$ for the group law $\circ$,
2. for every three elements $a, b, c \in G$, we have $(a \circ b) \circ c = a \circ (b \circ c)$ (associativity), and
3. for each element $a \in G$, there exists an inverse element, $a^\dagger \in G$, with $a \circ a^\dagger = a^\dagger \circ a = e$.

Transformation groups A group that consists of a set G of transformations on some set S is called a *transformation group*. We also say that *the group G acts on the set S*. A transformation is a mapping from S to itself that preserves certain properties, such as isometries in the Euclidean plane. Transformation groups are usually not commutative, that is $a \circ b \neq b \circ a$ for $a, b \in G$.

Definition 4.7 (Topological group)**.** A topological space G that is also a group is called a *topological group* when the group operation $\circ$ is continuous, that is, for $a, b \in G$, we have that the operations of product

1. $G \times G \to G : (a, b) \mapsto a \circ b$, and taking the inverse
2. $G \to G : a \mapsto a^{\dagger}$,

are continuous, where $G \times G$ has the product topology.

A topological group for which the underlying topology is locally compact and Hausdorff, is called a *locally compact group*.

Definition 4.8 (Eaton (1989), Definition 2.1)**.** Let Y be a set, and let G be a group with identity element e. A function $F : Y \times G \to Y$ satisfying

1. $F(y, e) = y, \quad y \in Y$
2. $F(y, g_1 g_2) = F(F(y, g_1), g_2), \quad g_1, g_2 \in G, y \in Y$

specifies G acting on the right of Y.

In practice, F is omitted: we will write $y \cdot g$ for a group element g acting on the right of $y \in Y$. For a subset $A \subseteq Y$, we write $A \cdot g \coloneqq \{a \cdot g \mid a \in A\}$.

Definition 4.9 (Conway (2013), Example 1.11)**.** Let G be a locally compact topological group. Then the *right invariant Haar* measure (in short: right Haar measure) for G is a Borel measure ν satisfying

1. $\nu(A) > 0$ for every nonempty open set $A \subseteq G$,
2. $\nu(K) < \infty$ for every compact set $K \subseteq G$,
3. $\nu(A \cdot g) = \nu(A)$ for every $g \in G$ and every measurable $A \subseteq G$.

4.B Proofs Omitted from Main Text

Proof. **[of Lemma 1]** Let $A \subset \mathbb{R}_{>0}$ be any Borel measurable set. In the equations below, the sum and integral can be swapped due to the monotone convergence theorem and the fact that B_τ is

a positive function.

$$
\begin{aligned}
\int_A \mathrm{d}P_1^{[B_\tau]} &= \int_\Omega \mathbb{1}_{\{B_\tau \in A\}} \, \mathrm{d}P_1^{[B_\tau]} \\
&= \sum_{n=0}^{\infty} \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{B_\tau \in A\}} \mathbb{1}_{\{\tau = n\}} \, \mathrm{d}P_1^{(n)} \\
&\overset{(3)}{=} \sum_{n=0}^{\infty} \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{B_n \in A\}} \mathbb{1}_{\{\tau = n\}} \, P_1^{(n)}(\tau = n) \cdot \mathrm{d}P_1^{(n)}(\cdot \mid \tau = n) \\
&= \sum_{n=0}^{\infty} \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{B_n \in A\}} \, P_1^{(n)}(\tau = n) \cdot \mathrm{d}P_1^{(n)}(\cdot \mid \tau = n) \\
&\overset{(5)}{=} \sum_{n=0}^{\infty} \int_{r>0} \mathbb{1}_{\{r \in A\}} \, P_1^{(n)}(\tau = n) \cdot \mathrm{d}P_1^{[B_n]}(\cdot \mid \tau = n) \\
&= \sum_{n=0}^{\infty} \int_{r>0} \mathbb{1}_{\{r \in A\}} \frac{\mathrm{d}P_1^{[B_n]}(\cdot \mid \tau = n)}{\mathrm{d}P_0^{[B_n]}(\cdot \mid \tau = n)}(r) \, P_1^{(n)}(\tau = n) \cdot \mathrm{d}P_0^{[B_n]}(\cdot \mid \tau = n) \\
&\overset{(*)}{=} \sum_{n=0}^{\infty} \int_{r>0} \mathbb{1}_{\{r \in A\}} \frac{P_0^{(n)}(\tau = n)}{P_1^{(n)}(\tau = n)} \cdot r \cdot P_1^{(n)}(\tau = n) \cdot \mathrm{d}P_0^{[B_n]}(\cdot \mid \tau = n) \\
&= \sum_{n=0}^{\infty} \int_{r>0} \mathbb{1}_{\{r \in A\}} r \, P_0^{(n)}(\tau = n) \cdot \mathrm{d}P_0^{[B_n]}(\cdot \mid \tau = n) \\
&= \sum_{n=0}^{\infty} \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{B_n \in A\}} \cdot B_n \cdot P_0^{(n)}(\tau = n) \cdot \mathrm{d}P_0^{(n)}(\cdot \mid \tau = n) \\
&= \sum_{n=0}^{\infty} \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{B_n \in A\}} \mathbb{1}_{\{\tau = n\}} \cdot B_n \cdot P_0^{(n)}(\tau = n) \cdot \mathrm{d}P_0^{(n)}(\cdot \mid \tau = n) \\
&= \sum_{n=0}^{\infty} \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{B_\tau \in A\}} \mathbb{1}_{\{\tau = n\}} \cdot B_n \, \mathrm{d}P_0^{(n)} \\
&= \int_\Omega \mathbb{1}_{\{B_\tau \in A\}} \left(\sum_{n=0}^{\infty} \mathbb{1}_{\{\tau = n\}} B_n \right) \mathrm{d}P_0 \\
&= \int_\Omega \mathbb{1}_{\{B_\tau \in A\}} B_\tau \, \mathrm{d}P_0 \\
&\overset{(14)}{=} \int_A t P_0^{[B_\tau]}(dt),
\end{aligned}
$$

where $(*)$ follows because of our fixed n-calibration assumption. Furthermore, (3) follows from the following equality for any $C \in \mathcal{F}$

$$
P_1^{(n)}(C \cap \{\tau = n\}) = P_1^{(n)}(\tau = n) \cdot P_1^{(n)}(C \mid \tau = n), \tag{4.31}
$$

and in (5) we perform a change of variables where we integrate over the possible values of the Bayes Factor instead of over the outcome space, which we repeat in (14).

We have shown that the function g defined by $g(t) = t$ is the Radon-Nikodym derivative $\dfrac{\mathrm{d}P_1^{[B_\tau]}}{\mathrm{d}P_0^{[B_\tau]}}$. □

Proof. **[of Lemma 2]** Let A be any Borel subset of $\mathbb{R}_{>0}$. We have:

$$\begin{aligned}
\int_A \mathrm{d}\overline{P}_1^{[\gamma_n]}(\cdot \mid x^m, \tau = n) &= \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{\gamma_n \in A\}} \, \mathrm{d}\overline{P}_n^{(1)}(\cdot \mid x^m, \tau = n) \\
&= \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{\gamma_n \in A\}} \left(\frac{\mathrm{d}\overline{P}_n^{(1)}(\cdot \mid x^m, \tau = n)}{\mathrm{d}\overline{P}_n^{(0)}(\cdot \mid x^m, \tau = n)} \right) \mathrm{d}\overline{P}_n^{(0)}(\cdot \mid x^m, \tau = n) \\
&\overset{(*)}{=} \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{\gamma_n \in A\}} \gamma_n \cdot \left(\frac{\pi(H_0 \mid x^m, \tau = n)}{\pi(H_1 \mid x^m, \tau = n)} \right) \mathrm{d}\overline{P}_n^{(0)}(\cdot \mid x^m, \tau = n) \\
&= \int_{\langle \mathcal{X}^n \rangle} \mathbb{1}_{\{\gamma_n \in A\}} \gamma_n \left(\frac{\overline{P}_0(\tau = n \mid x^m)\pi(H_0)}{\overline{P}_1(\tau = n \mid x^m)\pi(H_1)} \right) \mathrm{d}\overline{P}_n^{(0)}(\cdot \mid x^m, \tau = n) \\
&= \left(\frac{\overline{P}_0(\tau = n \mid x^m)\pi(H_0 \mid x^m)}{\overline{P}_1(\tau = n \mid x^m)\pi(H_1 \mid x^m)} \right) \cdot \int_A \gamma_n \, \mathrm{d}\overline{P}_0^{[\gamma_n]}(\cdot \mid x^m, \tau = n),
\end{aligned}$$

where, for the case $m = 0$, $(*)$ follows from (4.3), which can be verified to be still valid in our generalized setting. The case $m > 0$ follows in exactly the same way, by shifting the data by m places (so that the new x_1 becomes what was x_{m+1}, and treating, for $k = 0, 1$, $\pi(H_k \mid x^m)$ as the priors for this shifted data problem, and then applying the above with $m = 0$).

We have shown that the Radon-Nikodym derivative $\dfrac{\mathrm{d}\overline{P}_1^{[\gamma_n]}(\cdot \mid x^m)}{\mathrm{d}\overline{P}_0^{[\gamma_n]}(\cdot \mid x^m)}$ at γ_n is given by $\gamma_n \cdot \dfrac{\overline{P}_0(\tau = n \mid x^m)\pi(H_1 \mid x^m)}{\overline{P}_1(\tau = n \mid x^m)\pi(H_0 \mid x^m)}$, which is what we had to show. □

Proof. **[of Lemma 7]** Let A' denote the event $V_n = 1$ and let $A \subset \mathbb{R}_{>0}$ be a Borel measurable subset of the positive real numbers. We have that β_n is a function of the maximal invariant U_n as defined in Definition 4.3, and we write $\beta_n(U_n)$. With this notation, we have:

$$\begin{aligned}
P_{1,g}^{[\beta_n]}(A \mid A') &= \int_{\mathbb{R}_{>0}} \mathbb{1}_{\{A\}} \, \mathrm{d}P_{1,g}^{[\beta_n]}(\cdot \mid A') \\
&\overset{(2)}{=} \int_{\mathcal{U}_n} \mathbb{1}_{\{\beta_n(U_n) \in A\}} \, \mathrm{d}P_{1,g}^{[U_n]}(\cdot \mid A') \\
&= \int_{\mathcal{U}_n} \mathbb{1}_{\{\beta_n(U_n) \in A\}} \frac{\mathrm{d}P_{1,g}^{[U_n]}(\cdot \mid A')}{\mathrm{d}P_{0,g}^{[U_n]}(\cdot \mid A')} \, \mathrm{d}P_{0,g}^{[U_n]}(\cdot \mid A') \\
&\overset{(4)}{=} \int_{\mathcal{U}_n} \mathbb{1}_{\{\beta_n(U^n) \in A\}} \frac{P_{0,g}^{(n)}(A')}{P_{1,g}^{(n)}(A')} \frac{\mathrm{d}P_{1,g}^{[U_n]}}{\mathrm{d}P_{0,g}^{[U_n]}} \, \mathrm{d}P_{0,g}^{[U_n]}(\cdot \mid A') \\
&\overset{(5)}{=} \frac{P_{0,g}^{(n)}(A')}{P_{1,g}^{(n)}(A')} \cdot \int_{\mathcal{U}_n} \mathbb{1}_{\{\beta_n(U^n) \in A\}} \beta_n(U_n) \, \mathrm{d}P_{0,g}^{[U_n]}(\cdot \mid A') \\
&= \frac{P_{0,g}^{(n)}(A')}{P_{1,g}^{(n)}(A')} \cdot \int_{\mathbb{R}_{>0}} \mathbb{1}_{\{A\}} t \, \mathrm{d}P_{0,g}^{[\beta_n]} A'(t),
\end{aligned}$$

where step (2) holds because β_n is $\mathcal{G}_n$-measurable. On the set A' we have

$$\frac{\mathrm{d}P_{1,g}^{[U_n]}(\cdot \mid A')}{\mathrm{d}P_{0,g}^{[U_n]}(\cdot \mid A')} \frac{P_{1,g}^{(n)}(A')}{P_{0,g}^{(n)}(A')} = \frac{\mathrm{d}P_{1,g}^{[U_n]}}{\mathrm{d}P_{0,g}^{[U_n]}},$$

which explains step (4), and step (5) follows from the definition of β_n in Equation (4.29).

We have shown that $\dfrac{P_{0,g}^{(n)}(A')}{P_{1,g}^{(n)}(A')} \cdot t$ is equal to the Radon-Nikodym derivative $\dfrac{\mathrm{d}P_{1,g}^{[\beta_n]}(\cdot \mid V_n = 1)}{\mathrm{d}P_{0,g}^{[\beta_n]}(\cdot \mid V_n = 1)}$, which is what we had to prove. □

Chapter 5

Safe Testing

Abstract

We develop the theory of hypothesis testing based on the E-value, a notion of evidence that, unlike the P-value, allows for effortlessly combining results from several tests. Even in the common scenario of *optional continuation*, where the decision to perform a new test depends on previous test outcomes, 'safe' tests based on E-values generally preserve Type-I error guarantees. Our main result shows that E-values exist for completely general testing problems with composite null and alternatives. Their prime interpretation is in terms of gambling or investing, each E-value corresponding to a particular investment. Surprisingly, optimal "GROW" E-variables, which lead to fastest capital growth, are fully characterized by the *joint information projection* (JIPr) between the set of all Bayes marginal distributions on $\mathcal{H}_0$ and $\mathcal{H}_1$. Thus, optimal E-values also have an interpretation as Bayes factors, with priors given by the JIPr. We illustrate the theory using several 'classic' examples including a one-sample safe t-test and the 2×2 contingency table. Sharing Fisherian, Neymanian and Jeffreys-Bayesian interpretations, E-values and safe tests may provide a methodology acceptable to adherents of all three schools.

5.1 Introduction and Overview

We wish to test the veracity of a *null hypothesis* $\mathcal{H}_0$, often in contrast with some *alternative hypothesis* $\mathcal{H}_1$, where both $\mathcal{H}_0$ and $\mathcal{H}_1$ represent sets of distributions on some given sample space. Our theory is based on *E-test statistics*. These are simply *nonnegative* random variables that satisfy the inequality:

$$\text{for all } P \in \mathcal{H}_0\text{: } \mathbf{E}_P[E] \leq 1. \tag{5.1}$$

We refer to E-test statistics as E-variables, and to the value they take on a given sample as the E-*value*, emphasizing that they are to be viewed as an alternative to, and in many cases an improvement of, the classical P-value. Note that *large* E-values correspond to evidence against the null: for given E-variable E and $0 \leq \alpha \leq 1$, we define the *threshold test corresponding to E with significance level α*, as the test that rejects $\mathcal{H}_0$ iff $E \geq 1/\alpha$. We will see, in a sense to be

defined, that this test is *safe under optional continuation*, which for brevity we will simply call "safe".

Motivation P-values and standard null hypothesis testing have come under intense scrutiny in recent years (Wasserstein, Lazar et al., 2016; Benjamin et al., 2018). E-variables and safe tests offer several advantages. Most importantly, in contrast to P-values, E-variables behave excellently under *optional continuation*, the highly common practice in which the decision to perform additional tests partly depends on the outcome of previous tests; they thus seem particularly promising when used in meta-analysis, avoiding the issue of 'accumulation bias' (Ter Schure and Grünwald, 2019). A second reason is their enhanced *interpretability*, and a third is their flexibility: E-variables based on Fisherian, Neyman-Pearsonian and Bayes-Jeffreys' testing philosophies all can be accommodated for. These three types of E-variables can be freely combined, while preserving Type I error guarantees; at the same time, they keep a clear (monetary) interpretation even if one dismisses 'significance' altogether, as recently advocated by Amrhein, Greenland and McShane, 2019.

Contribution Our aim is to lay out the full theory of testing based on E-variables, both methodologically and mathematically. Methodologically, we explain the advantages that E-variables and safe tests offer over traditional tests, P-values and (some) Bayes factors; we introduce the GROW criterion defining optimal E-variables and provide specific ('simple δ-GROW') E-variables that are well-behaved in terms of GROW and power, and easy to use in practice. Mathematically, we show (Theorem 5.4) that, for arbitrary composite, nonconvex $\mathcal{H}_0$ and $\mathcal{H}_1$, we can construct nontrivial E-variables. In many cases, (Theorem 5.4 and 5.6) we can even construct E-variables that are optimal in the strong GROW sense. E-variables have been invented independently by (at least) Levin (1976) and Zhang, Glancy and Knill (2011) and have been analyzed before by Shafer et al. (2011) and Shafer and Vovk (2019) and Vovk and Wang (2019), who emphasize that they can also be much more easily *merged* than P-values. They are close cousins of test martingales (Shafer et al., 2011) which themselves underlie AV (anytime-valid) P-values (Johari, Pekelis and Walsh, 2015), AV tests and AV confidence sequences (Balsubramani and Ramdas, 2016; Howard et al., 2018b; Howard et al., 2018a). As such, our methodological insights are mostly variations of existing ideas; yet, they have never before been worked out in full. The mathematical results Theorem 5.4 and Theorem 5.6 are new, although a special case of Theorem 5.4 was shown earlier by (Zhang, Glancy and Knill, 2011); see Section 5.6 for more on the novelty and related work.

Contents In this introductory section, we give an overview of the main ideas: Section 5.1.1 provides three interpretations of E-variables and the idea of optional continuation. In Section 5.1.2, we discuss the GROW optimality theorem, and the use of our Theorem 5.4 to find 'good' Bayesian and/or GROW E-variables. Section 5.1.3 gives a first, extended example. The remainder of the paper is structured as follows. Section 5.2 explains how some E-value based tests are not merely safe under optional continuation, but also under the more well-known optional stopping, and explains the close relation between *test martingales* and E-variables. Section 5.3 gives our first main result, Theorem 5.4. Section 5.4 gives several examples, and Section 5.5 reports some preliminary experiments. The paper ends with a section providing

more historical context and an overview of related work in Section 5.6 — including a discussion that clarifies how testing based on E-values could provide a unification of Fisher's, Neyman's and Jeffreys' ideas. All longer proofs are delegated to the appendices, which start with Appendix 5.A providing details about (standard but tacit) assumptions and notations from the main text.

5.1.1 The three main interpretations of E-variables

1. First Interpretation: Gambling The first and foremost interpretation of E-variables is in terms of *money*, or, more precisely, *Kelly (1956) gambling*. Imagine a ticket (contract, gamble, investment) that one can buy for 1$, and that, after realization of the data, pays E $; one may buy several and positive fractional amounts of tickets. (5.1) says that, if the null hypothesis is true, then one expects not to gain any money by buying such tickets: for any $r \in \mathbb{R}^+$, upon buying r tickets one expects to end up with $r\mathbf{E}[E] \leq r$ $. Therefore, if the observed value of E is large, say 20, one would have gained a lot of money after all, indicating that something might be wrong about the null.

2. Second Interpretation: Conservative P-Value, Type I Error Probability Recall that a strict P-value is a random variable P such that for all $0 \leq \alpha \leq 1$, all $P_0 \in \mathcal{H}_0$,

$$P_0(P \leq \alpha) = \alpha. \tag{5.2}$$

A *conservative* P-value is a random variable for which (5.2) holds with '=' replaced by '$\leq$'. There is a close connection between (small) P- and (large) E-values:

Proposition 1. *For any given E-variable E, define $P_{[E]} := 1/E$. Then $P_{[E]}$ is a conservative P-value. As a consequence, for every E-variable E, any $0 \leq \alpha \leq 1$, the corresponding threshold-based test has Type-I error guarantee α, i.e. for all $P \in \mathcal{H}_0$,*

$$P(E \geq 1/\alpha) \leq \alpha. \tag{5.3}$$

Proof. **(of Proposition 1)** Markov's inequality gives $P(E \geq \alpha^{-1}) \leq \alpha \mathbf{E}_P[E] \leq \alpha$. □

While E-variables are thus conservative P-values, standard P-values satisfying (5.2) are by no means E-variables; if E is an E-variable and P is a standard P-value, and they are calculated on the same data, then we will usually observe $P \ll 1/E$ so E gives *less* evidence against the null; Section 5.1.3 and Section 5.6 will give some idea of the ratio between $1/E$ and P in various practical settings.

Combining 1. and 2.: Optional Continuation, GROW Propositions 2, 3 below show that *multiplying* E-variables $E_{(1)}, E_{(2)}, \ldots$ for tests based on respective samples $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots$ (with each $\mathbf{Y}_{(j)}$ being the vector of outcomes for the j-th test), gives rise to new E-variables, even if the decision whether or not to perform the test resulting in $E_{(j)}$ was based on the value of earlier test outcomes $E_{(j-1)}, E_{(j-2)}, \ldots$. As a result (Prop. 2), *the Type I-Error Guarantee (5.3) remains valid even under this 'optional continuation' of testing*. An informal 'proof' is immediate from our gambling interpretation: if we start by investing $1 in $E_{(1)}$ and, after observing $E_{(1)}$, reinvest

all our new capital $\$E_{(1)}$ into $E_{(2)}$, then after observing $E_{(2)}$ our new capital will obviously be $\$E_{(1)} \cdot E_{(2)}$, and so on. If, under the null, we do not expect to gain any money for any of the individual gambles $E_{(j)}$, then, intuitively, we should not expect to gain any money under whichever strategy we employ for deciding whether or not to reinvest (just as you would not expect to gain any money in a casino irrespective of your rule for re-investing and/or stopping and going home).

3. Third Interpretation: Bayes Factors For convenience, from now on we write the models $\mathcal{H}_0$ and $\mathcal{H}_1$ as

$$\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\} \;\; ; \;\; \mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\},$$

where for $\theta \in \Theta_0 \cup \Theta_1$, the P_θ are all probability distributions on the same sample, all have probability densities or mass functions, denoted as p_θ, and we assume the parameterization is 1-to-1 (see Appendix 5.A for more details). $\mathbf{Y} = (Y_1, \ldots, Y_N)$, a vector of N outcomes, represents our data. N may be a fixed sample size n but can also be a random stopping time. In the Bayes factor approach to testing, one associates both $\mathcal{H}_j$ with a *prior* W_j, which is simply a probability distribution on Θ_j, and a *Bayes marginal probability distribution* P_{W_j}, with density (or mass) function given by

$$p_{W_j}(\mathbf{Y}) := \int_{\Theta_j} p_\theta(\mathbf{Y}) \, \mathrm{d}W_j(\theta). \tag{5.4}$$

The *Bayes factor* is then given as:

$$\mathsf{BF} := \frac{p_{W_1}(\mathbf{Y})}{p_{W_0}(\mathbf{Y})}. \tag{5.5}$$

Whenever $\mathcal{H}_0 = \{P_0\}$ is *simple*, i.e., a singleton, then the Bayes factor is also an E-variable, since in that case, we must have that W_0 is degenerate, putting all mass on 0, and $p_{W_0} = p_0$, and then for all $P \in \mathcal{H}_0$, i.e. for P_0, we have

$$\mathbf{E}_P[\mathsf{BF}] := \int p_0(y) \cdot \frac{p_{W_1}(y)}{p_0(y)} \, \mathrm{d}y = 1. \tag{5.6}$$

For such E-variables that are really simple-$\mathcal{H}_0$-based Bayes factors, Proposition 1 reduces to the well-known *universal bound* for likelihood ratios (Royall, 1997). When $\mathcal{H}_0$ is itself composite, most Bayes factors $\mathsf{BF} = p_{W_1}/p_{W_0}$ will *not* be E-variables any more, since for BF to be an E-variable we require (5.6) to hold for *all* $P_\theta, \theta \in \Theta_0$, whereas in general it only holds for $P = P_{W_0}$. Nevertheless, our Theorem 5.4 implies that there always exist many special combinations of W_0 and W_1, for which $\mathsf{BF} = p_{W_1}/p_{W_0}$ is an E-variable after all, and that optimal E-values invariably take on a Bayesian form (though sometimes with unusual priors).

5.1.2 How to find Good E-Values

1. (Semi-) Bayesian Approach Suppose we take a Bayesian stance regarding $\mathcal{H}_1$ and, conditioned on $\mathcal{H}_1$, are prepared to represent our uncertainty by prior distribution W_1 on Θ_1.

Suppose that the set of all probability distributions $\mathcal{W}(\Theta_0)$ that one can define on Θ_0, contains a prior W_0° that minimizes the KL divergence $D(P_{W_1} \| P_{W_0^\circ}) = \min_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1} \| P_{W_0})$ to P_{W_1}. Following Barron and Li, 1999, we call $P_{W_0}^\circ$ the *Reverse Information Projection (RIPr)* of

P_{W_1} on $\mathcal{P}(\Theta_0) = \{P_{W_0} : W_0 \in \mathcal{W}(\Theta_0)\}$. Parts 1 and 2 of our main result Theorem 5.4 essentially state the following:

Corollary of Theorem 5.4 Let W_1 be any prior on Θ_1 and let $P_{W_0^\circ}$ be the RIPr of P_{W_1} on $\mathcal{P}(\Theta_0)$. Then the Bayes factor $\text{E}^*_{W_1} := p_{W_1}(\mathbf{Y})/p_{W_0^\circ}(\mathbf{Y})$ is an E-variable.

The RIPr idea can be extended to the case that the minimum $\min_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1} \| P_{W_0})$ is not achieved, and the theorem provides a W_1-based E-variable for that case as well. We can thus be fully Bayesian about $\mathcal{H}_1$, but any prior W_1 on $\mathcal{H}_1$ that we wish to adopt forces us to adopt a corresponding prior $W_0^\circ \in \mathcal{H}_0$. In general this may feel 'un-Bayesian', but one may perhaps consider it a small price to pay for creating a Bayes factor that should be acceptable to frequentists as well — for the test corresponding to $E^*_{W_1}$ will preserve Type-I error bounds under optional continuation under all $P_0 \in \mathcal{H}_0$, no matter the prior W_1 one chose. Moreover, in the standard case that the models are nested and $\mathcal{H}_0$ is a sub-model of $\mathcal{H}_1$, it is generally recognized that the priors on $\mathcal{H}_0$ and $\mathcal{H}_1$ should somehow be 'matched' with each other (Berger, Pericchi and Varshavsky, 1998); we may view the RIPr construction as providing just such a matching.

2. Frequentist (GROW) Approach We return to the monetary interpretation of E-values. The definition of E-variable ensures that we expect them to stay under 1 (one does not gain money) under any $P \in \mathcal{H}_0$. Analogously, one would like them to be constructed such that they can be expected to grow large as fast as possible (one gets rich, gets evidence against $\mathcal{H}_0$) under all $P \in \mathcal{H}_1$. Informally, E-variables with this property are called GROW. In its simplest form, for $\mathcal{H}_0$ and $\mathcal{H}_1$ that are strictly separated, the GROW *(growth-rate optimal in worst-case)* criterion tells us to pick, among all E-variables relative to $\mathcal{H}_0$, the one that maximizes *expected capital growth rate under* $\mathcal{H}_1$ in the worst case, i.e. the E-variable E^* that achieves

$$\max_{E: E \text{ is an E-variable}} \min_{P \in \mathcal{H}_1} \mathbf{E}_P [\log E] \tag{5.7}$$

We give five reasons for using the logarithm rather than any other increasing function (such as the identity) in Section 5.3.1. Briefly, when we keep using E-variables with additional data batches as explained in Section 5.2 below, then optimizing for $\log E$ ensures that our capital grows at the fastest rate. Optimality in terms of GROW may be viewed as an analogue of the classical frequentist concept of power.

Part 3 of Theorem 5.4 expresses that, under regularity conditions, the GROW E-variable is once again a Bayes factor; remarkably, it is the Bayes factor between the Bayes marginals $(P^*_{W_1}, P^*_{W_0})$ that form the *joint information projection* (JIPr), i.e. that are, among all Bayes marginals indexed by $\mathcal{W}(\Theta_0)$ and $\mathcal{W}'_1$, the *closest* in KL divergence (Figure 5.1). By joint convexity of the KL divergence (Van Erven and Harremoës, 2014), finding the JIPr pair is thus a convex optimization problem, tending to be computationally feasible.

3. δ-GROW E-values In Section 5.3.3 we consider the case that $\mathcal{H}_0$ and $\mathcal{H}_1$ are neither separated nor do we have prior(s) on $\mathcal{H}_1$ available. We can often parameterize the models as $\Theta_0 = \{(0, \gamma) : \gamma \in \Gamma\}$ and $\Theta_1 = \{(\delta, \gamma) : \delta \in \Delta, \gamma \in \Gamma\}$ where δ is a single scalar parameter of

interest. We can then define $\underline{\delta}$-*GROW* E-variables that are GROW relative to some suitable $\mathcal{H}_1' = \{P_{(\delta,\gamma)} : \gamma \in \Gamma, \delta \in \Delta, |\delta| \geq \underline{\delta}\}$. The development is analogous to the classical development of tests that have either maximal power under a minimal relevant effect size, or that have a uniformly most powerful property; and the resulting δ-GROW E-variables will also have reasonable properties in terms of power. δ-GROW E-variables are again Bayes factors. Often the δ-GROW E-variable is *simple* in that it sets W_1^* to be a degenerate prior, putting all its marginal mass on Δ on a single $\underline{\delta}$ (for a one-sided test) or on $\{-\underline{\delta}, \underline{\delta}\}$ (two-sided). If $\mathcal{H}_1$ is a one-dimensional exponential family, then δ-GROW E-values can be connected to the uniformly most powerful Bayes factors of Johnson, 2013b.

We work out simple δ-GROW E-variables for several standard settings: 1-dimensional exponential families, nonparametric tests such as Mann-Whitney, 2×2 contingency tables and the setting of the 1-sample t-test, each time applying Theorem 5.4 to show that the resulting E-variable is GROW. We also provide 'quick and dirty' (non-GROW) E-variables for general multivariate exponential family $\mathcal{H}_0$. Bayesian t-tests with a standard (nondegenerate) prior $W[\delta]$ on δ, while providing a GROW E-variable, are not δ-based in our sense. We present a δ-GROW version of the Bayesian t-test that has significantly better properties in terms of statistical power than the standard versions. We provide a preliminary experiment suggesting that with δ-GROW E-variables, if data comes from $\mathcal{H}_1$ rather than $\mathcal{H}_0$, one needs less data to find out than with standard Bayes factor tests, but a bit more data than with standard frequentist tests. However, in the t-test setting the effective amount of data needed is about the same as with the standard frequentist t-test because one is allowed to do optional stopping.

4. Robust Bayesian view of Theorem 5.4 We may think of the previous Bayesian RIPr result as a special case of the JIPr result: if $\mathcal{H}_1$ is composite, we can 'collapse' it into a single distribution by adopting a prior W_1 on Θ_1 of our choice and re-defining $\mathcal{H}_1$ to be the singleton $\mathcal{H}_1' = \{P_{W_1}\}$. We are then in the setting of Figure 5.1 but with $\mathcal{H}_1$ a singleton, and the JIPr becomes the RIPr. The E-variable $E^*_{W_0^\circ} = p_{W_1}/p_{W_0^\circ}$ can thus be thought of as the GROW E-variable relative to $\mathcal{H}_1'$.

More generally, we may only be able to specify a prior distribution on some, but not all of the parameters. For example, in Bayesian testing with nuisance parameters satisfying a group invariance as proposed by Berger, Pericchi and Varshavsky, 1998 one would like to specify a prior $W[\delta]$ on the effect size (non-nuisance) parameter δ but make no assumptions at all about the nuisance parameter vector γ (a special case is the Bayesian t-test, with γ representing variance). This is an instance of a 'robust Bayesian' approach (Grünwald and Dawid, 2004) in which prior knowledge is encoded as a *set* of priors (in this instance, it would be the set of all priors on (δ, γ) whose marginal on δ coincides with $W[\delta]$). Our Theorem 5.4 continues to apply in this setting. Rather than a full model $\mathcal{H}_1$ as under 2. above, or a single prior W_1 as under 1. above, we may replace the minimum over $P \in \mathcal{H}_1$ in (5.7) by a minimum over $W \in \mathcal{W}_1'$ over any convex *set* of priors $\mathcal{W}_1'$ on Θ_1, $\min_{P \in \mathcal{H}_1} \mathbf{E}_P[\ldots]$ becoming $\min_{W \in \mathcal{W}_1'} \mathbf{E}_{P_W}[\ldots]$. For essentially any such $\mathcal{W}_1'$, our Theorem 5.4 still holds. This high level of generality is needed, for example, in our treatment of the 1-sample t-test. For this we formally show (in our second main result, Theorem 5.6, which enables us to use Theorem 5.4) that the Bayes factor based on the improper right Haar prior, advocated by Berger, Pericchi and Varshavsky, 1998, has a

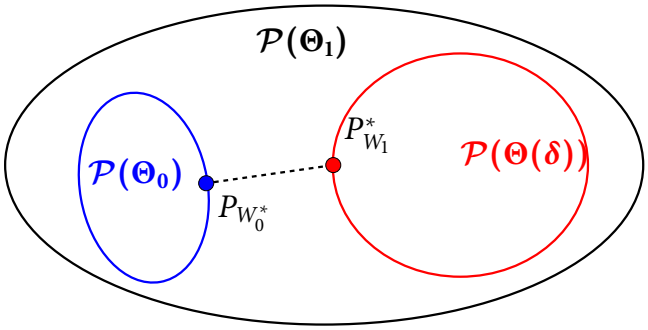


Figure 5.1: The Joint Information Projection (JIPr), with notation from Section 5.3. $\Theta_0 \subset \Theta_1$ represent two nested models, $\Theta(\delta)$ is a restricted subset of Θ_1 that does not overlap with Θ_0. $\mathcal{P}(\Theta) = \{P_W : W \in \mathcal{W}(\Theta)\}$, and $\mathcal{W}(\Theta)$ is the set of all priors over Θ, so $\mathcal{P}(\Theta)$ is the set of all Bayes marginals with priors on Θ. Theorem 5.4 says that the GROW E-variable $E^*_{\Theta_1(\delta)}$ between Θ_0 and $\Theta_1(\delta)$ is given by $E^*_{\Theta_1(\delta)} = P_{W_1^*}/P_{W_0^*}$, the Bayes factor between the two Bayes marginals that minimize KL divergence $D(P_{W_1} \| P_{W_0})\}$.

GROW property.

5. Examples and Experiments We work out simple δ-GROW E-variables for several standard settings: 1-dimensional exponential families, nonparametric tests such as Mann-Whitney, 2×2 contingency tables and the setting of the 1-sample t-test, each time applying Theorem 5.4 to show that the resulting E-variable is GROW. We also provide 'quick and dirty' (non-GROW) E-variables for the case that $\mathcal{H}_0$ is a general multivariate exponential family. Specifically we show that Bayes factors equipped with the right Haar prior on nuisance parameters provide E-variables, despite the prior being improper. The Bayesian t-test with a standard (nondegenerate) prior $W[\delta]$ on δ thus gives an S-variable, but it is not δ-GROW in our sense. We present a δ-GROW version of the Bayesian t-test that has significantly better properties in terms of statistical power than the standard versions. We provide a preliminary experiment suggesting that with δ-GROW E-variables, if data comes from $\mathcal{H}_1$ rather than $\mathcal{H}_0$, one needs less data to find out than with standard Bayes factor tests, but a bit more data than with standard frequentist tests. However, in the t-test setting the effective amount of data needed is about the same as with the standard frequentist t-test because, in this setting, one is allowed to do optional stopping.

5.1.3 A First Example: the Gaussian Location Family

Let $\mathcal{H}_0$ express that the Y_i are i.i.d. $\sim N(0,1)$. According to $\mathcal{H}_1$, the Y_i are i.i.d. $\sim N(\mu,1)$ for some $\mu \in \Theta_1 = \mathbb{R}$. We perform a first test on initial sample $\mathbf{Y} := Y^n := (Y_1, \ldots, Y_n)$. We consider a standard Bayes factor test for this scenario, equiping Θ_1 with a prior W that for simplicity we take to be normal with variance 1, so that W has density $w(\mu) \propto \exp(-\mu^2/2)$. The Bayes factor is given by

$$E_{(1)} := \frac{p_W(\mathbf{Y})}{p_0(\mathbf{Y})} = \frac{\int_{\mu \in \mathbb{R}} p_\mu(\mathbf{Y}) w(\mu) d\mu}{p_0(\mathbf{Y})}, \tag{5.8}$$

where $p_\mu(\mathbf{Y}) = p_\mu(Y_1, \ldots, Y_n) \propto \exp(-\sum_{i=1}^n (Y_i - \mu)^2/2)$; by (5.6) we know that $E_{(1)}$ is an E-value. By straightforward calculation:

$$\log E = -\frac{1}{2}\log(n+1) + \frac{1}{2}(n+1)\cdot \check{\mu}_n^2,$$

where $\check{\mu}_n = (\sum_{i=1} Y_i)/(n+1)$ is the Bayes MAP estimator, which only differs from the ML estimator by $O(1/n^2)$: $\check{\mu}_n - \widehat{\mu}_n = \widehat{\mu}_n/(n(n+1))$. If we were to reject Θ_0 when $E \geq 20$ (giving, by Proposition 1 a Type-I error guarantee of 0.05), we would thus reject if

$$|\check{\mu}_n| \geq \sqrt{\frac{5.99 + \log(n+1)}{n+1}}, \text{ i.e. } |\widehat{\mu}_n| \geq \sqrt{(\log n)/n},$$

where we used $2\log 20 \approx 5.99$. Contrast this with the standard Neyman-Pearson (NP) test, which would reject ($\alpha \leq 0.05$) if $|\widehat{\mu}_n| \geq 1.96/\sqrt{n}$. The δ-GROW E-variables for this problem that we describe in Section 5.4.1 can be chosen so as to guarantee $E^* \geq 20$ if $|\widehat{\mu}_n| \geq \widetilde{\mu}_n$ with $\widetilde{\mu}_n = c_n/\sqrt{n}$ where $c_n > 0$ is increasing and converges exponentially fast to $\sqrt{2\log 40} \approx 2.72$. Thus, while the NP test itself defines an E-variable that scores infinitely bad on our GROW optimality criterion (Example 5.1), we can choose a GROW E^* that is qualitatively more similar to a standard NP test than a standard Bayes factor approach. For general 1-dimensional exponential families, this δ-GROW E^* coincides with a 2-sided version of Johnson's (2013b; 2013a) uniformly most powerful Bayes test, which uses a discrete prior W within $\mathcal{H}_1$: for the normal location family, $W(\{\widetilde{\mu}_n\}) = W(\{-\widetilde{\mu}_n\}) = 1/2$ with $\widetilde{\mu}_n$ as above. Since the prior depends on n, some statisticians would perhaps not really view this as 'Bayesian'; and we also think of such δ-GROW E-variables, despite their formally Bayesian form, as having firstly a frequentist motivation.

Optional Continuation: Compatibility with Bayesian Updating For arbitrary prior W on Θ_1, define $e_{n,W} = p_W(Y^n)/p_0(Y^n)$ to be the Bayes factor with prior W for Θ_1 applied to data Y^n. The Bayesian E-variable (5.8) can then be written as $E_{(1)} = e_{N_{(1)},W_{(1)}}(\mathbf{Y}_{(1)})$, with $N_{(1)} = n$, $\mathbf{Y}_{(1)} = \mathbf{Y} = Y^n$. Suppose we have adopted some inital prior $W_{(1)}$ (say a normal with variance 1), and initial observed data $\mathbf{Y}_{(1)} = Y^n$, leading to a first E-value $E_{(1)} = 18$ — promising enough for us to invest our resources into a subsequent trial. We decide to gather $N_{(2)}$ data points leading to data $\mathbf{Y}_{(2)} = (Y_{N_{(1)}+1}, \ldots, Y_{N_{(2)}})$. We decide to use the following E-variable for this second data batch:

$$E_{(2)} := e_{N_{(2)},W_{(2)}}\left(\mathbf{Y}^{(2)}\right) := \frac{p_{W_{(2)}}\left(\mathbf{Y}_{(2)}\right)}{p_0\left(\mathbf{Y}_{(2)}\right)},$$

for a new prior $W_{(2)}$. Crucially, we are allowed to choose both $N_{(2)}$ and $W_{(2)}$ as a function of past data $\mathbf{Y}^{(1)}$. To see that $E_{(2)}$ gives an E-variable, note that, no matter how we choose $W_{(2)}$, $\mathbf{E}_{\mathbf{Y}^{(2)}\sim P_0}[E_{(2)}] = 1$, by a calculation analogous to (5.6). If we want to stick to the Bayesian paradigm, we can choose $W_{(2)} := W_{(1)}(\cdot \mid \mathbf{Y}_{(1)})$, i.e. $W_{(2)}$ is the Bayes posterior for μ based on data $\mathbf{Y}_{(1)}$ and prior $W_{(1)}$. A simple calculation using Bayes' theorem shows that multiplying $E^{(2)} := E_{(1)} \cdot E_{(2)}$ (which gives a new E-variable by Proposition 2), satisfies

$$E^{(2)} = E_{(1)} \cdot E_{(2)} = \frac{p_{W_{(1)}}(\mathbf{Y}_{(1)}) \cdot p_{W_{(1)}(\cdot|\mathbf{Y}_{(1)})}(\mathbf{Y}_{(2)})}{p_0(\mathbf{Y}_{(2)})} = \frac{p_{W_{(1)}}(Y_1, \ldots, Y_{N_{(2)}})}{p_0(Y_1, \ldots, Y_{N_{(2)}})}, \tag{5.9}$$

which is exactly what one would get by Bayesian updating. This illustrates that, for simple $\mathcal{H}_0$, combining E-variables by multiplication can be done consistently with Bayesian updating if the E-variables are based on Bayes factors with prior on $\mathcal{H}_1$ given by the posterior based on past data. To be precise, if, in Proposition 2 below, one takes as function $g(\mathbf{Y}) := W_{(1)} \mid \mathbf{Y}$, then the resulting products $E^{(k)} = \prod_{j=1}^{k} E_{(j)}$, $k = 1, 2, \ldots$ precisely correspond to the Bayes factors based on prior $W_{(1)}$ after observing data $\mathbf{Y}_1, \ldots, \mathbf{Y}_{(k)}$.

Optional Continuation: Beyond Bayesian Updating However, it might also be the case that it is not us who get the additional funding to obtain extra data, but rather some research group at a different location. If the question is, say, whether a medication works, the null hypothesis would still be that $\mu = 0$ but, if it works, its effectiveness might be slightly different due to slight differences in population. In that case, the research group might decide to use a different test statistic $E'_{(2)}$ which is again a Bayes factor, but now with an alternative prior W on μ (for example, the original prior $W_{(1)}$ might be re-used rather than replaced by $W_{(1)}(\cdot \mid \mathbf{Y}_{(1)})$. Even though this would not be standard Bayesian, $E_{(1)} \cdot E'_{(2)}$ would still be a valid E-variable, and Type-I error guarantees would still be preserved — and the same would hold even if the new research group would use an entirely different prior on Θ_1. It is also conceivable that the group performing the first trial was happy to adopt a Bayesian stance, adopting the normal prior $W_{(1)}$, whereas the second group was frequentist, adopting a δ-GROW E-variable satisfying $E^*_{(2)} \geq 20$ if $|\widehat{\mu}(\mathbf{Y}_{(2)}| \gtrapprox 2.72/\sqrt{n}$, with $\widehat{\mu}(\mathbf{Y}_{(2)}$ the MLE based on the second sample. Still, basing decisions on the product $E^*_{(1)} \cdot E^*_{(2)}$ preserves Type-I error probability bounds. And, after the second batch of data $\mathbf{Y}^{(2)}$, one might consider obtaining a third sample, or even more samples, each time using a different $W_{(k)}$, that is always allowed to depend on the past. In the next section we show how multiplying E-variables against such an arbitrarily long sequence of trials always preserves Type-I error bounds.

Beyond the Normal Location Family Full compatibility of our approach with Bayesian updating remains possible for all testing problems with simple $\mathcal{H}_0$. If $\mathcal{H}_0$ becomes composite, it cannot always be ensured: while we may still choose prior $W_{(2)}$ on Θ_1 to be the Bayes posterior based on $\mathbf{Y}_{(1)}$, the corresponding prior on Θ_0 to be used in the second batch of data may in general not be equal to the posterior on Θ_0 based on $\mathbf{Y}_{(1)}$.

5.2 Optional Continuation

Suppose we have available a collection $\mathcal{E} = \bigcup_{n \geq 1} \mathcal{E}_n$, with $\mathcal{E}_n = \{e_{n,W} : W \in \mathcal{W}\}$, where for each n and $W \in \mathcal{W}_n$, $e_{n,W}$ defines a nonnegative test statistic for data $Y^n = (Y_1, \ldots, Y_n)$ of length n: it is a function from $\mathcal{Y}^n$ to $\mathbb{R}_0^+$. We are mostly interested in the case that $\mathcal{E}$ really represents a collection of E-variables, so that for all n, $W \in \mathcal{W}$, $E := e_{n,W}(Y^n)$ is an E-variable. For example, we could take $e_{n,w}$ to be the E-variable in the example of Section 5.1.3, which depends on the prior W, each different prior leading to a different valid definition of $E = e_{n,W}(\mathbf{Y})$. More generally though, the $e_{n,W}$ may not always have a direct Bayesian interpretation.

We observe a first sample (e.g., data of a first clinical trial), $\mathbf{Y}_{(1)} = Y^{N_{(1)}} = (Y_1, \ldots, Y_{N_{(1)}})$,

and measure our first test statistic $E_{(1)}$ based on $\mathbf{Y}_{(1)}$. That is, $E_{(1)} = E_{N_{(1)},W_{(1)}}(\mathbf{Y}_{(1)})$ for some function $E_{N_{(1)},W_{(1)}} \in \mathcal{E}_{N_{(1)}}$. Then, if either the value of $E_{(1)}$ or, more generally of the underlying data $\mathbf{Y}_{(1)}$ is such that we (or some other research group) would like to continue testing, a second data sample $\mathbf{Y}_{(2)} = (Y_{N_{(1)}+1}, \ldots, Y_{\tau_{(2)}})$ is obtained (e.g. a second clinical trial is done), and a test statistic $E_{(2)}$ based on data $\mathbf{Y}_{(2)}$ is measured. Here $\tau_{(2)} := N_{(1)} + N_{(2)}$, where $N_{(2)}$ is the size of the second sample. We may choose $E_{(2)}$ to be any member from the set $\mathcal{E}$, and $N_{(2)}$ to be any sample size. As illustrated by the example in Section 5.1.3, the particular choice we make may *itself* depend on $\mathbf{Y}_{(1)}$. This means that $N_{(2)}$ and $E_{(2)}$ are determined via two functions $g : \bigcup_{n\geq 0} \mathcal{Y}^n \to \mathcal{W} \cup \{\text{STOP}\}$ and $h : \bigcup_{n\geq 0} \mathcal{Y}^n \to \mathbb{N}$ where, for any data $\mathbf{Y}_{(1)}$, g determines $W_{(2)}$, and h determines $N_{(2)}$, so that together they determine the next E-variable to be used. After observing $\mathbf{Y}_{(2)}$, depending again on the value of $\mathbf{Y}_{(2)}$, a decision is made either to continue to a third test, or to stop testing for the phenomenon under consideration. In this way we go on until either we decide to stop or until some maximum number $k_{\max}$ tests have been performed.

The decision whether to stop after k tests or to continue, and if so, what test statistic to use at the $k+1$-st test, is conveniently encoded into g. Thus, $g(\mathbf{Y}^{(k)}) = \text{STOP}$ means that the k-th test was the final one to be performed. $N_{(k)}$, the size of the k-th batch of data, and $\tau_{(k)} := \sum_{j=1}^{k} N_{(j)}$, the total sample size after k batches are determined as follows: we set $N_{(k)} := h(\mathbf{Y}^{(k-1)})$, where $\mathbf{Y}^{(k)} := (\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(k)})$, and $\mathbf{Y}_{(k)} := (Y_{\tau_{(k-1)}+1}, \ldots, Y_{\tau_{(k)}})$, where we set $\tau_{(0)} := 0$. With this notation, $Y^0 = \mathbf{Y}^{(0)}$ is an 'empty sample' and $N_{(1)} := h(Y^0)$ is a data-independent sample size for the first data batch; for convenience we also set $E_{(0)} := 1$. $E_{(k)}$, the k-th test statistic to be used is similarly determined via $W_{(k)} := g(\mathbf{Y}^{(k-1)})$ and then $E_{(k)} := e_{N_{(k)},W_{(k)}}(\mathbf{Y}_{(k)})$. With $Y_1, Y_2, \ldots$ arriving sequentially, we can recursively use g to first determine $N_{(1)}$ and $E_{(1)}$; we can then use $g(\mathbf{Y}^{(1)})$ to determine $N_{(2)}$, $\tau_{(2)}$ and $E_{(2)}$; we then use $g(\mathbf{Y}^{(2)})$ to determine $N_{(3)}$, $\tau_{(3)}$ and $E_{(3)}$, and so on, until $g(\mathbf{Y}^{(k)}) = \text{STOP}$.

Before presenting definitions and results, we generalize the setting to allow us to deal with optional continuation rules that may be restricted (as needed for e.g. the Bayesian t-test (Section 5.4.3) and with data $Y_1, Y_2, \ldots$ that are not i.i.d. according to all P_θ. For simple i.i.d testing problems, one may simply set $V_n = Y_n$ everywhere for all n below, and skip directly to Definition 5.1 and Proposition 2, ignoring the word 'conditional' in all that follows.

For the general case, we fix a sequence of random variables $V_1, V_2, \ldots$ such that for each n, V_n takes values in a set $\mathcal{V}_n$, and there is a function v_n such that $V_n = v_n(Y^n)$. We call each V_n a *coarsening* of Y^n and, borrowing terminology from measure theory, we call the process $V_1, V_2, \ldots$ a *filtration* of $Y^1, Y^2, \ldots$. We now let $\mathcal{E}\langle(V_i)\rangle = \bigcup_{n>0,m\geq 0} \mathcal{E}_{n|m}$ with $\mathcal{E}_{n|m} = \{e_{n|m,W}\}$ where $e_{n|m,W}$ are functions of V^{n+m}, parameterized not just by the sample size n of samples to which they are to be applied but also by the sample size m of the past sample, after which they are applied. We call such a conditional test statistic $E := e_{n|m,W}(V^{n+m})$ an *E-variable* conditional on V^m relative to filtration $(V_i)_{i\in\mathbb{N}}$ if

$$\text{for all } P \in \mathcal{H}_0\text{: } \mathbf{E}_P[E \mid V^m] \leq 1. \tag{5.10}$$

We change the definition of the function g above by replacing all occurrences of the letter Y with the corresponding instance of the letter V, and with now $E_{(k)} := e_{N_{(k)}|\tau_{(k-1)},W_{(k)}}(\mathbf{Y}_{(k)})$.

Definition 5.1. Let $K_{\text{STOP}} \geq 0$ to be the smallest k for which $g\left(\mathbf{V}^{(k)}\right) = \text{STOP}$, and $K_{\text{STOP}} = k_{\max}$ if no such k exists. Let $\mathcal{E}\langle(V_i)\rangle$ be a collection of nonnegative conditional test statistics as above, defined relative to some filtration $(V_i)_{i\in\mathbb{N}}$ of $(Y_i)_{i\in\mathbb{N}}$. We say that the *threshold test based on $\mathcal{S}$ is safe under optional continuation (for Type-I error probability, under multiplication) for continuation rules based on* (V_i), if for every g as above, with $E^{(k)} := \prod_{j=1}^{k} E_{(j)}$, for all $P_0 \in \mathcal{H}_0$, for every $0 \leq \alpha \leq 1$,

$$P_0\left(E^{(K_{\text{STOP}})} \geq \alpha^{-1}\right) \leq \alpha, \tag{5.11}$$

i.e. the α-Type-I error probability bound is preserved under any optional continuation rule.

Henceforth we simply omit 'for Type-I error, under multiplication' from our descriptions. If for all n, $V_n = Y_n$, then we simply write 'safe under optional continuation'.

A threshold test being safe under optional continuation implies that (5.11) even holds for the most aggressive continuation rule h which continues until the first K is reached such that either $\prod_{k=1}^{K} E_{(k)} \geq \alpha^{-1}$ or $K = k_{\max}$. Thus, safety under optional continuation implies that under all $P_0 \in \mathcal{H}_0$, the probability that there is *any* $k \leq k_{\max}$ such that $E^{(k)} \geq 1/\alpha$ is bounded by α. We can now present our optional continuation result in its most basic form:

Proposition 2. *Take any $(V_i)_{i\in\mathbb{N}}$ as above. If all elements of $\mathcal{E}$ are conditional E-variables as in (5.10), then $E^{(K_{\text{STOP}})}$ is an E-variable, so that by Proposition 1, the threshold test based on $E^{(K_{\text{STOP}})}$ is safe under optional continuation for all continuation rules based on (V_i).*

The proposition gives the prime motivation for the use of E-variables and verifies the claim made in the introduction: the product of E-variables remains an E-variable, even if the decision to observe additional data and record a new E-variable depends on previous outcomes. As a consequence, Type-I error guarantees still hold for the combined (multiplied) test outcome. The definition of safety requires Type-I error probabilities to be preserved under arbitrary functions g, yet a threshold test based on $E^{(K_{\text{STOP}})}$ can be applied without knowing the "off-sample" details of the actual function g that was used: we only need to know, for each k, *once we are at the end of the k-th trial*, the value of $g(Y^{(k)})$. Thus, crucially, we can apply such tests, and have Type-I error guarantees without knowing any other detail of the functions that have actually been (implicitly, or unconsciously) used. For example, suppose that we continued to a second sample $\mathbf{Y}_{(2)}$ because the data looked promising, say we observed a P-value based on $\mathbf{Y}_{(1)}$ equal to 0.02. We may not really know whether we would also have continued to gather a second sample if we had observed $P = 0.04$ — but it does not matter, because irrespective of whether a function g was used that continues if $P(\mathbf{Y}_{(1)}) \in [0.01, 0.03]$ or a function that continues if $P(\mathbf{Y}_{(1)}) \in [0.005, 0.04]$, or any other g (e.g. based on $E_{(1)}$ instead of a P-value), safety under optional continuation guarantees that our Type-I error guarantee is preserved — even without us knowing such details concerning g.

A heuristic proof of Proposition 2 has already been given in the beginning of this paper: the statement is essentially equivalent to 'no matter what your role is for stopping and going home, you cannot expect to win in a real casino'. We give an explicit elementary proof in Appendix 5.B. There we also generalize Proposition 2 in various ways: we include the conditional case where each P_θ defines a conditional distribution for Y^n given covariate information X^n and we allow

the sample size of the j-th sample $Y_{(j)}$ to be not fixed in advance but itself determined by some stopping rule. Finally, we also allow the decision whether or not to perform a new test to depend on (nonstochastic) side-information such as 'there is sufficient money to perform an additional trial with 50 subjects'.

5.2.1 E-values vs. Test Martingales; Optional Continuation vs. Stopping

The purpose of this section is two-fold: this paper is about 'safe testing' — not just under optional continuation, but also under optional stopping, which we therefore must discuss. Second, the prime tools for testing under optional stopping are test martingales, and these can be used to 'generate' useful E-variables, hence are important for us as well.

Optional Stopping We just formalized the idea of continuing from one trial (batch of data) to the next, and potentially stopping at the end of each trial. Now we consider the closely related 'dual' question: we are sequentially observing data within a single trial, but we want to be able to stop in the midst of it, without specifying at the beginning of the trial under what conditions we should stop. For example, we originally planned for a sample size of n but our boss might have peeked at interim results at $n' < n$ and concluded that these were so promising (or futile) that she insists on stopping the experiment, without us having anticipated this in advance. We cannot formalize this directly with E-values, because these are themselves defined for batches of data $\mathbf{Y} = Y^n$ of length n which may in fact come in without any particular order. Even if data does come in a particular order, the number n (or a data-dependent, a priori specified stopping time N as in Appendix 5.B) has to be specified in advance to make an E-value well-defined, so it will not always be clear what evidential value we should assign to the data if we want to stop at $n' < n$. To deal with optional stopping, we should thus not work with test statistics but rather with test *processes*, each process S_W defining an evidential value for each sample size.

Formally, a nonnegative test process $S = (S_i)_{i\in\mathbb{N}}$ relative to a filtration $(V_i)_{i\in\mathbb{N}}$, is defined as a sequence of nonnegative random variables $S_1, S_2, \ldots$ such that each $S_i = s_i(V^i)$ can be written as a function of V^i for some function s_i. We define a *stopping rule* g relative to (V_i) to be any function $g : \bigcup_{n\geq 0} \mathcal{V}^n \to \{\text{STOP}, \text{CONTINUE}\}$ so that there exists an (arbitrarily large but finite) $n_{\max}$ such that $g(v^n) = \text{STOP}$ for all $n \geq n_{\max}$, all $v^n \in \mathcal{V}^n$. We let $\mathcal{G}_{\text{ALL}}$ be the set of all such functions g.

Definition 5.2. Let $(S_i)_{i\in\mathbb{N}}$ be a nonnegative test process and let $\mathcal{G} \subset \mathcal{G}_{\text{ALL}}$ be a set of stopping rules. We say that the *threshold test based on* (S_i) *is safe under all stopping rules in* $\mathcal{G}$ if for every $g \in \mathcal{G}$ as defined above, all $P_0 \in \mathcal{H}_0$, for every $0 \leq \alpha \leq 1$:

$$P_0\left(S_{N_{\text{STOP}}} \geq \alpha^{-1}\right) \leq \alpha, \tag{5.12}$$

where the *stopping time* N_{STOP} is the smallest n at which $g(v^n) = \text{STOP}$.

As is well-known, *test martingales* lead to Type I error guarantees that are preserved under optional stopping. Formally, a *test martingale* relative to filtration (V_i) is a test statistic process $S_1, S_2, \ldots$ where each $S_n := \prod_{i=1}^n S_i$ for another process $S_{1|0}, S_{1|1}, S_{1|2}, \ldots$ such that $S_{1|i}$ is a function of V^i and satisfies, for all $P_0 \in \mathcal{H}_0$, $i \geq 1$,

$$\mathbf{E}_{P_0}\left[S_{1|i-1} \mid V^{i-1}\right] \leq 1. \tag{5.13}$$

We call $(S_{1|i-1})_{i\in\mathbb{N}}$ a *test martingale building block process*. In the proposition below, for $P \in \mathcal{H}_0 \cup \mathcal{H}_1$, $P[V^n]$ denotes the marginal distribution of V^n under P, and we denote its density by $p'(V^n)$. The following results are well-known:

Proposition 3. *Take any filtration (V_i) as above.*

1. *Suppose that $\mathcal{H}_0$ is a simple null for data coarsened to (V_i), i.e. for all $P, Q \in \mathcal{H}_0$, all n, $P[V^n] = Q[V^n]$. Then for every prior W on $\mathcal{H}_1$, the Bayes factor p'_W/p'_0 defines a test martingale, i.e. $(p'_W(V^i)/p'_0(V^i))_{i\in\mathbb{N}}$ is a test martingale relative to $(V_i)_{i\in\mathbb{N}}$.*
2. *Now, take any test martingale $(S_i)_{i\in\mathbb{N}}$ relative to filtration $(V_i)_{i\in\mathbb{N}}$. Then for all $g \in \mathcal{G}_{\text{ALL}}$, $S_{N_{\text{STOP}}}$ is an E-variable, so that by Proposition 1, the threshold test based on $S_{N_{\text{STOP}}}$ is safe under optional stopping for all stopping rules that can be defined relative to (V_i).*

Proof. The first part follows by applying the cancellation trick as in (5.6) to the conditional likelihood ratio $p'_W(V_i \mid V^{i-1})/p'_0(V_i \mid V^{i-1})$; the second part is immediate by Doob's optional stopping theorem. □

Test Martingales vs. E-Variables Part 2 of Proposition 3 shows that test martingales lead to tests that are safe under optional stopping. Just as important for us, it shows that we can use any given martingale and any stopping rule g to define an E-variable. In recent work, A. Ramdas and collaborators (Howard et al., 2018b; Howard et al., 2018a) have developed a large number of practically most useful test martingales (some of these can be thought of as Bayes factors, and some cannot; see Section 7.3 for many more references and history). All these test martingales can thus be used to 'generate' useful E-variables (and in fact Part 2 of Proposition 3 can easily be extended to also generate E-variables conditional on V^m for any desired m).

Conversely, we may ask ourselves whether E-variables can also be used to define test martingales (and hence to allow for tests that are safe under optional stopping). The answer is subtle, as we now illustrate. For simplicity, we only consider unconditional E-variables to be used with data that are i.i.d. under all $P \in \mathcal{H}_0$. In the sections to come, we provide constructions of E-variables for many $\mathcal{H}_0$; all of these can be applied to data of arbitrary fixed sample sizes n. For any given $\mathcal{H}_0$, they thus 'automatically' provide a test statistic process $(E_i)_{i\in\mathbb{N}}$ with $E_i = e_i(V^i)$.

1. A first idea is, for any given $\mathcal{H}_0$ and corresponding E-variables $(e_i(V^i))$, to define the process $(S_i)_{i\in\mathbb{N}}$ where $S_{1|i-1} = e_1(V_i)$, using only the 'first' E-variable. From (5.13) we immediately see that $(S_{1|i-1})_{i\in\mathbb{N}}$ is now a martingale building block process and (S_i) with $S_i = \prod_{i=1}^n e_1(V_i)$ is a test martingale. Since in this way, we can convert all E-variables into martingales, allowing us to do optional stopping, it may seem we have made the concept of E-variable superfluous. But this is not the case: for many of the $\mathcal{H}_0$ we consider below, this method leads to the useless test martingale with $S_i = S_{1|i-1} \equiv 1$, for all i, independent of the data. For example, this is the case for the 2×2-contingency tables (Section 5.4.4), for multivariate exponential families (Section 5.4.5) and for the nonparametric test of Example 5.3 — so that the above construction would lead to useless martingales that almost surely remain 1 forever.
2. In some cases, the test statistic process $(E_i)_{i\in\mathbb{N}}$ does turn out to give a test martingale.

Examples are GROW E-variables for the case that $\mathcal{H}_0$ is simple (as in the one-parameter exponential family case, Section 5.4.1), or for the case that the GROW E-variable for $\mathcal{H}_0$ can be written as a function of (V_i) such that $\mathcal{H}_0$ is simple when data are coarsened to (V_i) (as in the Bayesian t-test, Section 5.4.3). This can be used to modify, if so desired, E_i to another E-variable $E_{N_{\text{STOP}}}$ based on some stopping rule g; see Section 5.5.2 where this idea is used to improve statistical power of E_i.

3. Yet in other cases, $\mathcal{H}_0$ is composite, and there is no natural coarsening/filtration (V_i) under which it becomes simple. Then, at least in general, the process $(e_i(V^i))$ is not a test martingale. Counterexamples again include the E-values for the 2×2-contingency tables, multivariate exponential families and for the nonparametric test of Example 5.3. We do not see an easy way to obtain test martingales, and hence tests that are safe under 'full' optional stopping, for these settings. Still, sometimes tests based on the non-martingale process $(E_i)_{i \in \mathbb{N}}$ do allow for optional stopping under some non-trivial subset $\mathcal{G} \subset \mathcal{G}_{\text{ALL}}$. For example, it is easy to show that the E-values for multivariate exponential families that we consider in Section 5.4.5 satisfy $\mathbf{E}_{P_0}[e(Y^{N_{\text{STOP}}}) \mid x^{N_{\text{STOP}}}] \leq 1$ for all $P_0 \in \mathcal{H}_0$ as long as, for each n, the stopping rule $g(Y^n)$ can be written as a fixed function of the sufficient statistic $\widehat{\theta}_0(Y^n)$ for $\mathcal{H}_0$; the tests based on these E-values are thus safe under optional stopping relative to $(V_i)_{i \in \mathbb{N}} := (Y_i)_{i \in \mathbb{N}}$ under all such g.

5.3 Main Result

From here onward we let $\mathcal{W}(\Theta)$ be the set of all probability distributions (i.e., 'proper priors') on Θ, for any $\Theta \subset \Theta_0 \cup \Theta_1$. Notably, this includes, for each $\theta \in \Theta$, the degenerate distribution W which puts all mass on θ.

5.3.1 What is a good E-Value? The GROW Criterion

The (semi-) Bayesian approach to finding E-variables has already been treated in some detail in Section 5.1.2. Thus, we focus on a frequentist perspective here, getting back to the Bayesian approach later. We start with an example that tells us how *not* to design E-variables.

Example 5.1. [Strict Neyman-Pearson E-Values: valid but useless] In *strict* Neyman-Pearson testing (Berger, 2003), one rejects the null hypothesis if the P-value P satisfies $P \leq \alpha$ for the a priori chosen significance level α, but then one only reports "reject" rather than the P-value itself. This can be seen as a safe test based on a special E-variable E_{NP}: when P is a P-value determined by data $\mathbf{Y}$, we define $E_{\text{NP}} = 0$ if $P > \alpha$ and $E_{\text{NP}} = 1/\alpha$ otherwise. For any $P_0 \in \mathcal{H}_0$ we then have $\mathbf{E}_{\mathbf{Y} \sim P_0}[E_{\text{NP}}] = P_0(P \leq \alpha)\alpha^{-1} \leq 1$, so that E_{NP} is an E-variable, and the 'safe' test that rejects if $E_{\text{NP}} \geq 1/\alpha$ obviously is identical to the test that rejects if $P \leq \alpha$. However, with this E-variable, there is a positive probability α of losing all one's capital. The E-variable E_{NP} leading to the Neyman-Pearson test, i.e. the maximum power test, *now* thus corresponds to an irresponsible gamble that has a positive probability of losing all one's power for *future* experiments. This also illustrates that the E-variable property (5.1) is a *minimal requirement* for being useful under optional continuation; in practice, one also wants guarantees that one cannot completely lose one's capital.

In the Neyman-Pearson paradigm, one measures the quality of a test at a given significance level α by its power in the worst-case over all $P_\theta, \theta \in \Theta_1$. If Θ_0 is nested in Θ_1, one first restricts Θ_1 to a subset $\Theta_1' \subset \Theta_1$ with $\Theta_0 \cap \Theta_1' = \varnothing$ of 'relevant' or 'sufficiently different from Θ_0' hypotheses. For example, one takes the largest Θ_1' for which at the given sample size a specific power can be obtained. We develop analogous versions of this idea below; for now let us assume that we have identified such a Θ_1' that is separated from Θ_0. The standard NP test would now pick, for a given level α, the test which maximizes power over Θ_1'. The example above shows that this corresponds to an E-variable with disastrous behavior under optional continuation. However, we now show how to develop a notion of 'good' E-variable analogous to Neyman-Pearson optimality by replacing 'power' (probability of correct decision under Θ_1') with *expected capital growth rate* under Θ_1', which then can be linked to Bayesian approaches as well.

Taking, like NP, a worst-case approach, we aim for an E-variable with *large* $\mathbf{E}_{\mathbf{Y}\sim P_\theta}[f(E)]$ under any $\theta \in \Theta_1'$. Here $f : \mathbb{R}^+ \to \mathbb{R}$ is some increasing function. At first sight it may seem best to pick f the identity, but this can lead to adoption of an E-variable such that $P_\theta(E = 0) > 0$ for some $\theta \in \Theta_1'$; we have seen in the example above that that is a very bad idea. A similar objection applies to any polynomial f, but it does not apply to the logarithm, which is the single natural choice for f: by the law of large numbers, a sequence of E-variables $E_1, E_2, \ldots$ based on i.i.d. $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots$ with, for all j, $\mathbf{E}_{\mathbf{Y}_{(j)}\sim P}[\log E_j] \geq L$, will a.s. satisfy $E^{\langle m \rangle} := \prod_{j=1}^m E_j = \exp(mL + o(m))$, i.e. E will grow exponentially, and $L(\log_2 e)$ lower bounds the *doubling rate* (Cover and Thomas, 1991). Such exponential growth rates can only be given for the logarithm, which is a second reason for choosing it. A third reason is that it automatically gives E-variables an interpretation within the MDL framework (Section 5.7.2); a fourth is that such growth-rate optimal E can be linked to power calculations after all, with an especially strong link in the one-dimensional case (Section 5.4.1), and a fifth reason is that some existing Bayesian procedures can also be reinterpreted in terms of growth rate.

We thus seek to find E-variables E^* that achieve, for some $\Theta_1' \subset \Theta_1 \setminus \Theta_0$:

$$\inf_{\theta\in\Theta_1'} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log E^*] = \sup_{E\in\mathcal{E}(\Theta_0)} \inf_{\theta\in\Theta_1'} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log E] =: \mathrm{GR}(\Theta_1'), \tag{5.14}$$

where $\mathcal{E}(\Theta_0)$ is the set of all E-variables that can be defined on $\mathbf{Y}$ for Θ_0. We call this special E^*, if it exists and is essentially unique, the GROW (*Growth-Rate-Optimal-in-Worst-case*) E-variable relative to Θ_1', and denote it by $E^*_{\Theta_1'}$ (see Appendix 5.C for the meaning of 'essentially unique').

If we feel Bayesian about $\mathcal{H}_1$, we may be willing to adopt a prior W_1 on Θ_1, and instead of restricting to Θ_1', we may instead want to consider the growth rate under the prior W_1. More generally, as *robust Bayesians* or *imprecise probabilists* (Berger, 1985; Grünwald and Dawid, 2004; Walley, 1991) we may consider a whole 'credal set' of priors $\mathcal{W}_1' \subset \mathcal{W}(\Theta_1)$ and again consider what happens in the worst-case over this set. We are then interested in the GROW E-variable E^* that achieves

$$\inf_{W\in\mathcal{W}_1'} \mathbf{E}_{\mathbf{Y}\sim P_W}[\log E^*] = \sup_{E\in\mathcal{E}(\Theta_0)} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{\mathbf{Y}\sim P_W}[\log E]. \tag{5.15}$$

Again, if an E-variable achieving (5.15) exists and is essentially unique, then we denote it by $E^*_{\mathcal{W}'_1}$. If $\mathcal{W}'_1 = \{W_1\}$ is a single prior, we denote the E-variable by $E^*_{W_1}$. (5.15) then reduces to

$$\mathbf{E}_{\mathbf{Y}\sim P_{W_1}}[\log E^*_{W_1}] = \sup_{E\in\mathcal{E}(\Theta_0)} \mathbf{E}_{\mathbf{Y}\sim P_{W_1}}[\log E],$$

and Theorem 5.4, Part 2 below implies that, under regularity conditions, in this case $E^*_{W_1} = p_{W_1}(\mathbf{Y})/p_{W_0^\circ}(\mathbf{Y})$ for some prior W° on Θ_0: the GROW E^*-variable relative to P_{W_1} is always a Bayes factor with P_{W_1} in the denominator.

If $\mathcal{W}'_1 = \mathcal{W}(\{\theta_1\})$ is a single prior that puts all mass on a singleton θ_1, then we write $E^*_{\mathcal{W}'_1}$ as $E^*_{\theta_1}$. Linearity of expectation further implies that (5.15) and (5.14) coincide if $\mathcal{W}'_1 = \mathcal{W}(\Theta'_1)$; thus (5.15) generalizes (5.14).

All E-variables in the examples below, except for the 'quick and dirty' ones of Section 5.4.5, are of this 'maximin' form. They will be defined relative to sets $\mathcal{W}'_1$ with in one case (Section 5.4.3) $\mathcal{W}'$ representing a set of prior distributions on Θ_1, and in other cases (Section 5.4.1–5.4.4) $\mathcal{W}'_1 = \mathcal{W}(\Theta'_1)$ for a 'default' choice of a subset of Θ_1.

5.3.2 The JIPr is GROW

We now present our main result, illustrated in Figure 5.1. We use $D(P\|Q)$ to denote the *relative entropy* or *Kullback-Leibler (KL) Divergence* between distributions P and Q (Cover and Thomas, 1991). We call an E-variable *trivial* if it is always ≤ 1, irrespective of the data, i.e. no evidence against $\mathcal{H}_0$ can be obtained. The first part of the theorem below implies that nontrivial E-variables essentially always exist as long as $\Theta_0 \neq \Theta_1$. The second part — really implied by the third but stated separately for convenience — characterizes when such E-variables take the form of a likelihood ratio/Bayes factor. The third says that GROW E-variables for a whole set of distributions Θ'_1 can be found by a joint KL minimization problem.

Part 3 of the theorem refers to a *coarsening* of $\mathbf{Y}$. This is any random variable $\mathbf{V}$ that can be written as a function of $\mathbf{Y}$, i.e. $\mathbf{V} = f(\mathbf{Y})$ for some function f; in particular, the result holds with f the identity and $\mathbf{V} = \mathbf{Y}$. For general coarsenings $\mathbf{V}$, the distributions P_θ for $\mathbf{Y}$ induce marginal distributions for $\mathbf{V}$, which we denote by $P_\theta^{[\mathbf{V}]}$.

Theorem 5.4. *1. Let $W_1 \in \mathcal{W}(\Theta_1)$ such that $\inf_{W_0\in\mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) < \infty$ and such that for all $\theta \in \Theta_0$, P_θ is absolutely continuous relative to P_{W_1}. Then the GROW E-variable $E^*_{W_1}$ exists, is essentially unique, and satisfies*

$$\mathbf{E}_{\mathbf{Y}\sim P_{W_1}}[\log E^*_{W_1}] = \sup_{E\in\mathcal{E}(\Theta_0)} \mathbf{E}_{\mathbf{Y}\sim P_{W_1}}[\log E] = \inf_{W_0\in\mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0})$$

2. Let W_1 be as above and suppose further that the inf/min is achieved by some W_0°, i.e. $\inf_{W_0\in\mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) = D(P_{W_1}\|P_{W_0^\circ})$. Then the minimum is achieved uniquely by this W_0° and the GROW E-variable takes a simple form: $E^*_{W_1} = p_{W_1}(\mathbf{Y})/p_{W_0^\circ}(\mathbf{Y})$.

3. Now let $\Theta'_1 \subset \Theta_1$ and let $\mathcal{W}'_1$ be a subset of $\mathcal{W}(\Theta'_1)$ such that for some coarsening $\mathbf{V}$ of $\mathbf{Y}$ (we may have $\mathbf{Y} = \mathbf{V}$) the following holds: for all $\theta \in \Theta_0$, all $W_1 \in \mathcal{W}'_1$, $P_\theta^{[\mathbf{V}]}$ is absolutely con-

tinuous relative to $P_{W_1}^{[\mathbf{V}]}$*, and the set* $\{P_{W_1}^{[\mathbf{V}]} : W_1 \in \mathcal{W}_1'\}$ *is convex (this holds automatically if* $\mathcal{W}_1'$ *is convex). Suppose that*

$$\inf_{W_1 \in \mathcal{W}_1'} \inf_{W_0 \in \mathcal{W}_0} D(P_{W_1} \| P_{W_0}) = \min_{W_1 \in \mathcal{W}_1'} \min_{W_0 \in \mathcal{W}_0} D(P_{W_1}^{[\mathbf{V}]} \| P_{W_0}^{[\mathbf{V}]}) = D(P_{W_1^*}^{[\mathbf{V}]} \| P_{W_0^*}^{[\mathbf{V}]}) < \infty, \quad (5.16)$$

the minimum being achieved by some (W_1^*, W_0^*) *such that* $D(P_{W_1} \| P_{W_0^*}) < \infty$ *for all* $W_1 \in \mathcal{W}_1'$*. If the minimum is achieved uniquely by* (W_1^*, W_0^*)*, then the GROW e-variable* $E_{\mathcal{W}_1'}^*$ *relative to* $\mathcal{W}_1'$ *exists, is essentially unique, and is given by*

$$E_{\mathcal{W}_1'}^* = \frac{p'_{W_1^*}(\mathbf{V})}{p'_{W_0^*}(\mathbf{V})}, \quad (5.17)$$

where p'_W *is the density on* $\mathbf{V}$ *corresponding to* $P_W^{[\mathbf{V}]}$*. Also,* $E_{\mathcal{W}_1'}^*$ *satisfies*

$$\inf_{W \in \mathcal{W}_1'} \mathbf{E}_{\mathbf{Y} \sim P_W}[\log E_{\mathcal{W}_1'}^*] = \sup_{E \in \mathcal{E}(\Theta_0)} \inf_{W \in \mathcal{W}_1'} \mathbf{E}_{\mathbf{Y} \sim P_W}[\log E] = D(P_{W_1^*}^{[\mathbf{V}]} \| P_{W_0^*}^{[\mathbf{V}]}). \quad (5.18)$$

If $\mathcal{W}_1' = \mathcal{W}(\Theta_1')$*, then by linearity of expectation we further have* $E_{\mathcal{W}_1'}^* = E_{\Theta_1'}^*$.

The requirements that, for $\theta \in \Theta_0$, the P_θ are absolutely continuous relative to the P_{W_1}, and, in Part 3, that $D(P_{W_1} \| P_{W_0^*}) < \infty$ for all $W_1 \in \mathcal{W}_1'$ are quite mild — in any case they hold in all specific examples considered below, specifically if $\Theta_0 \subset \Theta_1$ represent general multivariate exponential families, see Section 5.4.5.

Since the KL divergence is strictly convex in both arguments if the other argument is held fixed, and non-strictly jointly convex, we have that if (5.16) holds, then for each (W_1', W_0') achieving the minimum, either $W_1' = W_1^*$, $W_0' = W_0^*$ or both $W_1' \neq W*_1$ and $W_0' \neq W_0^*$. In the latter case, all mixtures $(1-\alpha)(W_1', W_0') + \alpha(W_1, W_0)$ also achieve the minimum.

Following Li, 1999, we call P_{W° as in Part 2 of the theorem, the *Reverse Information Projection (RIPr)* of P_{W_1} on $\{P_W : W \in \mathcal{W}(\Theta_0)\}$. Extending this terminology we call $(P_{W_1^*}, P_{W_0^*})$ the *joint information projection (JIPr)* of $\{P_W : W \in \mathcal{W}_1'\}$ and $\{P_W : W \in \mathcal{W}(\Theta_0)\}$ onto each other.

The requirement for the full JIPr characterization (5.18), that the minima are both achieved is strong in general, but it holds in the examples of Section 5.4.1 (1-dimensional) and 5.4.4 (2×2 tables) with $\mathbf{V} = \mathbf{Y}$. By allowing $\mathbf{V}$ to be a coarsening of $\mathbf{Y}$, we make the condition considerably weaker: it then also holds in the t-test example of Section 5.4.3 — that example will also illustrate that $\{P_{W_1}^{[\mathbf{V}]} : W_1 \in \mathcal{W}_1'\}$ may be convex even if $\mathcal{W}_1'$ is not, and that in cases where the minimum in (5.16) over P_{W_1} on $\mathbf{Y}$ does not exist, still its infimum over P_{W_1} on $\mathbf{Y}$ may be equal to the minimum over P_{W_1} defined on $\mathbf{V}$, which does exist.

Proof Sketch of Parts 2 and 3 We give short proofs of parts 2 and 3 under the (weak) additional condition that we can exchange expectation and differentiation and the (strong) condition that $\mathbf{V}$ is taken equal to $\mathbf{Y}$. To prove parts 2 and 3 without these conditions, we need a nonstandard minimax theorem; and to prove part 1 (which does not rely on minima being

achieved) we need a deep result from Barron and Li (Li, 1999); these extended proofs are in Appendix 5.C.

For Part 2, consider any $W_0' \in \mathcal{W}(\Theta_0)$ with $W_0' \neq W_0^\circ$, with W_0° as in the theorem statement. Straightforward differentiation shows that the derivative $(d/d\alpha)D(P_{W_1}\|P_{(1-\alpha)W_0^\circ+\alpha W_0'})$ at $\alpha = 0$ is given by $f(\alpha) := 1 - \mathbf{E}_{\mathbf{Y}\sim P_{W_0'}}[p_{W_1}(\mathbf{Y})/p_{W_0^\circ}(\mathbf{Y})]$. Since $(1-\alpha)W_0^\circ + \alpha W_0' \in \mathcal{W}(\Theta_0)$ for all $0 \leq \alpha \leq 1$, the fact that W_0° achieves the minimum over $\mathcal{W}(\Theta_0)$ implies that $f(0) \geq 0$, but this implies that $\mathbf{E}_{\mathbf{Y}\sim P_{W_0'}}[p_{W_1}(\mathbf{Y})/p_{W_0^\circ}(\mathbf{Y})] \leq 1$. Since this reasoning holds for all $W_0' \in \mathcal{W}(\Theta_0)$, we get that $p_{W_1}(\mathbf{Y})/p_{W_0^\circ}(\mathbf{Y})$ is an E-variable. To see that it is GROW, note that, for every E-variable $E = e(\mathbf{Y})$ relative to $\mathcal{E}(\Theta_0)$, we must have, with $q(y) := e(y)p_{W_0^\circ}(y)$, that $\int q(y)\,\mathrm{d}y = \mathbf{E}_{\mathbf{Y}\sim P_{W_0^\circ}}[E] \leq 1$, so q is a sub-probability density, and by the information inequality of information theory (Cover and Thomas, 1991), we have

$$\mathbf{E}_{P_{W_1}}[\log E] = \mathbf{E}_{P_{W_1}}\left[\log \frac{q(\mathbf{Y})}{p_{W_0^\circ}(\mathbf{Y})}\right] \leq \mathbf{E}_{P_{W_1}}\left[\log \frac{p_{W_1}(\mathbf{Y})}{p_{W_0^\circ}(\mathbf{Y})}\right] = \mathbf{E}_{P_{W_1}}\left[\log E^*_{W_1}\right],$$

implying that $E^*_{W_1}$ is GROW.

For Part 3, consider any $W_1' \in \mathcal{W}_1'$ with $W_1' \neq W_1^*$, W_1^*, W_0^* as in the theorem statement. Straightforward differentiation and reasoning analogously to Part 2 above shows that the derivative $(d/d\alpha)D(P_{(1-\alpha)W_1^*+\alpha W_1'}\|P_{W_0^*})$ at $\alpha = 0$ is nonnegative iff there is no $\alpha > 0$ such that $\mathbf{E}_{P_{(1-\alpha)W_1^*+\alpha W_1'}}[\log p_{W_1^*}(\mathbf{Y})/p_{W_0^*}(\mathbf{Y})] \leq \mathbf{E}_{P_{W_1^*}}[\log p_{W_1^*}(\mathbf{Y})/p_{W_0^*}(\mathbf{Y})]$. Since this holds for all $W_1' \in \mathcal{W}_1'$, and since $D(P_{W_1^*}\|P_{W_0^*}) = \inf_{W\in\mathcal{W}_1'} D(P_W\|P_{W_0^*})$, it follows that $\inf_{W\in\mathcal{W}_1'} \mathbf{E}_{P_W}[\log E^*_{\mathcal{W}_1'}] = D(P_{W_1^*}\|P_{W_0^*})$, which is already part of (5.18). Note that we also have

$$\begin{aligned}
\inf_{W\in\mathcal{W}_1'} \mathbf{E}_{\mathbf{Y}\sim P_W}[\log E^*_{\mathcal{W}_1'}] &\leq \sup_{E\in\mathcal{E}(\Theta_0)} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{\mathbf{Y}\sim P_W}[\log E] \\
&\leq \inf_{W\in\mathcal{W}_1'} \sup_{E\in\mathcal{E}(\Theta_0)} \mathbf{E}_{\mathbf{Y}\sim P_W}[\log E] \\
&= \inf_{W\in\mathcal{W}_1'} \sup_{E\in\mathcal{E}(\mathcal{W}(\Theta_0))} \mathbf{E}_{\mathbf{Y}\sim P_W}[\log E] \\
&\leq \inf_{W\in\mathcal{W}_1'} \sup_{E\in\mathcal{E}(\{W_0^*\})} \mathbf{E}_{\mathbf{Y}\sim P_W}[\log E] \\
&\leq \sup_{E\in\mathcal{E}(\{W_0^*\})} \mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}[\log E].
\end{aligned}$$

where the first two and final inequalities are trivial, the third one follows from definition of E-variable and linearity of expectation, and the fourth one follows because, as is immediate from the definition of E-variable, for any set $\mathcal{W}_0$ of priors on Θ_0, the set of E-variables relative to any set $\mathcal{W}' \subset \mathcal{W}_0$ must be a superset of the set of E-variables relative to $\mathcal{W}_0$.

It thus suffices if we can show that $\sup_{E\in\mathcal{E}(\{W_0^*\})} \mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}[\log E] \leq D(P_{W_1^*}\|P_{W_0^*})$. For this, consider E-variables $E = e(\mathbf{Y}) \in \mathcal{E}(\{W_0^*\})$ defined relative to the singleton hypothesis $\{W_0^*\}$. Since $\mathbf{E}_{\mathbf{Y}\sim P_{W_0^*}}[e(\mathbf{Y})] \leq 1$ we can write $e(\mathbf{Y}) = q(\mathbf{Y})/p_{W_0^*}(\mathbf{Y})$ for some sub-probability density

q, and

$$\sup_{E \in \mathcal{E}(\{P_{W_0^*}\})} \mathbf{E}_{P_{W_1^*}}[\log E] = \sup_q \mathbf{E}_{\mathbf{Y} \sim P_{W_1^*}}\left[\log \frac{q(\mathbf{Y})}{p_{W_0^*}}\right] \tag{5.19}$$
$$= D(P_{W_1^*} \| P_{W_0^*}),$$

where the supremum is over all sub-probability densities on $\mathbf{Y}$ and the final equality is the information (in)equality again (Cover and Thomas, 1991). The result follows.

5.3.3 δ-GROW and simple δ-GROW E-Values

To apply Theorem 5.4 to design E-variables with good frequentist properties in the case that $\Theta_0 \subsetneq \Theta_1$, we must choose a subset Θ_1' with $\Theta_1' \cap \Theta_0 = \varnothing$. Usually, we first carve up Θ_1 into nested subsets $\Theta(\varepsilon)$. A convenient manner to do this is to pick a divergence measure $d : \Theta_1 \times \Theta_0 \to \mathbb{R}_0^+$ with $d(\theta_1 \| \theta_0) = 0 \Leftrightarrow \theta_1 = \theta_0$, and, defining $d(\theta) := \inf_{\theta_0 \in \Theta_0} d(\theta, \theta_0)$ (examples below) so that

$$\Theta(\varepsilon) := \{\theta \in \Theta_1 : d(\theta) \geq \varepsilon\}. \tag{5.20}$$

In the examples below we are interested in GROW E-variables $E^*_{\Theta(\varepsilon)}$ for a given measure d for some particular value of ε. This is in full analogy to classical frequentist testing, where we look for tests with worst-case optimal power with alternatives restricted to sets $\Theta(\varepsilon)$; we merely replace 'power' by 'growth rate'.

In some cases such E-variables $E^*_{\Theta(\varepsilon)}$ take on a particularly simple form, as Bayes factors with all mass in Θ_1 concentrated on the boundary $\mathrm{BD}(\Theta(\varepsilon)) = \{\theta \in \Theta_1 : d(\theta) = \varepsilon\}$.

To develop these ideas further, for simplicity we restrict attention to the common case with just a single *scalar parameter of interest* $\delta \in \Delta \subseteq \mathbb{R}$ so that $\mathcal{H}_0, \mathcal{H}_1$ can be parameterized as $\Theta_1 = \{(\delta, \gamma) : \delta \in \Delta, \gamma \in \Gamma\}$ and $\Theta_0 = \{(0, \gamma) : \gamma \in \Gamma\}$, with Γ representing all distributions in $\mathcal{H}_0$. We can then simply take $d((\delta, \gamma)) = |\delta|$ so that $\Theta(\underline{\delta}) = \{(\delta, \gamma) : \delta \in \Delta, |\delta| \geq \underline{\delta}, \gamma \in \Gamma\}$. Then the E-variable $E^*_{\Theta(\underline{\delta})}$ with $\underline{\delta} > 0$ will be referred to as the *$\underline{\delta}$-GROW E-variable* for short.

Further defining $E^*_{\underline{\delta}} := E^*_{\{(\delta,\gamma):|\delta|=\underline{\delta},\gamma\in\Gamma\}}$, we call $E^*_{\Theta(\underline{\delta})}$ *simple* if

$$E^*_{\Theta(\underline{\delta})} = E^*_{\underline{\delta}} \tag{5.21}$$

In all examples below, the $\underline{\delta}$-GROW E is also simple, making it particularly easy to deal with.

To illustrate, consider first the one-sided case with $\Delta \subseteq \mathbb{R}_0^+$. Then, applying Theorem 5.4, Part 3 with $\Theta = \{(\underline{\delta}, \gamma) : \gamma \in \Gamma\}$ and assuming the KL-infimum is achieved, we must have $E^*_{\underline{\delta}} = p_{\underline{\delta}, W_1^*[\gamma]}(\mathbf{Y}) / p_{0, W_0^*[\Gamma]}(\mathbf{Y})$ for some priors $W_1^*[\gamma]$, $W_0^*[\gamma]$ on γ. We see that (5.21) holds iff

$$\sup_{E \in \mathcal{E}(\{0\})} \inf_{\theta \in \Theta(\underline{\delta})} \mathbf{E}_{\mathbf{Y} \sim P_\theta}[\log E] = \inf_{\theta \in \Theta(\underline{\delta})} \mathbf{E}_{\mathbf{Y} \sim P_\theta} \mathbf{E}[\log E^*_{\underline{\delta}}] \tag{5.22}$$
$$= D(P_{\underline{\delta}, W_1^*[\gamma]} \| P_{0, W_0^*[\gamma]}). \tag{5.23}$$

In Appendix 5.D, Proposition 9 we provide some sufficient conditions for (5.22) to hold.

Now consider the two-sided case with scalar parameter space Δ' an interval containing 0 in its interior. Since, by linearity of expectation, mixtures of E-variables are obviously E-variables,

$$E^{\circ}_{\underline{\delta}} := \frac{1}{2}E^{*}_{\underline{\delta}} + \frac{1}{2}E^{*}_{-\underline{\delta}} \tag{5.24}$$

is a simple E-variable. While $E^{\circ}_{\underline{\delta}}$ will be seen to be $\underline{\delta}$-GROW in the two-sided Gaussian location and t-test setting, in general, we have no guarantee that it is $\underline{\delta}$-GROW. Still, in Appendix 5.D we show that if its constituents are one-sided GROW, i.e. (5.21) holds for the 1-sided case with Δ set to Δ^+ and with Δ set to $-\Delta^-$, then the worst-case growth rate achieved by $E^{\circ}_{\underline{\delta}}$ is guaranteed to be close (within $\log 2$) of the two-sided δ-based GROW E-variable $E^{*}_{\Theta(\underline{\delta})}$. In such cases we may think of $E^{\circ}_{\underline{\delta}}$ as a *simple $\underline{\delta}$-almost-GROW E-variable*. $E^{\circ}_{\underline{\delta}}$ may be much easier to compute than the actual two-sided GROW E-variable $E^{*}_{\Theta(\underline{\delta})}$.

5.4 Examples

5.4.1 Point null vs. one-parameter exponential family

Let $\{P_\theta \mid \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}$ represent a 1-parameter exponential family for sample space $\mathcal{Y}$, given in its mean-value parameterization, such that $0 \in \Theta$, and take Θ_1 to be some interval (t', t) for some $-\infty \le t' \le 0 < t \le \infty$, such that $t', 0$ and t are contained in the interior of Θ. Let $\Theta_0 = \{0\}$. Both $\mathcal{H}_0 = \{P_0\}$ and $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$ are extended to outcomes in $\mathbf{Y} = (Y_1, \ldots, Y_n)$ by the i.i.d. assumption. For notational simplicity we set

$$D(\theta \| 0) := D(P_\theta(\mathbf{Y}) \| P_0(\mathbf{Y})) = nD(P_\theta(Y_1) \| P_0(Y_1)). \tag{5.25}$$

We consider the δ-GROW E-variables $E^{*}_{\Theta(\delta)}$ relative to sets $\Theta(\delta)$ as in (5.20). Since $\mathcal{H}_0$ is simple, we can simply take θ to be the parameter of interest, hence $\Delta = \Theta_1$ and Γ plays no role, so that $\Theta(\underline{\delta}) = \{\theta \in \Theta_1 : |\theta| \ge \underline{\delta}\}$.

One-Sided Test: simple GROW E-Variable Here we set $t' = 0$ so that $\Theta(\underline{\delta}) = \{\theta \in \Theta_1 : \theta \ge \underline{\delta}\}$. We show in Appendix 5.D that this is a case in which (5.21) holds: the $\underline{\delta}$-GROW E-variable is simple, and can be calculated as a likelihood ratio $E^{*}_{\Theta(\underline{\delta})} = p_{\underline{\delta}}(\mathbf{Y})/p_0(\mathbf{Y})$ between two point hypotheses, even though $\Theta(\underline{\delta})$ is composite.

GROW E-Variables and UMP Bayes tests We now show that, for this 1-sided testing case, for a specific value of $\underline{\delta}$, $E^{*}_{\Theta(\underline{\delta})}$ coincides with the *uniformly most powerful Bayes tests* of Johnson, 2013b, giving further motivation for their use and an indication of how to choose $\underline{\delta}$ if no a priori knowledge is available. Note first that, since $\Theta_0 = \{0\}$ is a singleton, by Theorem 5.4, Part 2, we have that $E^{*}_{W} = p_W(\mathbf{Y})/p_0(\mathbf{Y})$, i.e. for all $W \in \mathcal{W}(\Theta_1)$, the GROW E-variable relative to $\{W\}$ is given by the Bayes factor p_W/p_0. The following result is a direct consequence of Johnson, 2013b, Lemma 1.

Theorem 5.5 (Uniformly Most Powerful Bayes Test (Johnson, 2013b)). *Consider the setting above. Fix any* $0 < \alpha < 1$ *and assume that there is* $\underline{\delta} \in \Theta_1$ *with* $D(\underline{\delta} \| 0) = -\log \alpha$. *Then among the class of* all *threshold-based tests based on local Bayes factors, i.e. all tests of the form "reject iff* $p_W(\mathbf{Y})/p_0(\mathbf{Y}) \geq 1/\alpha$*" for some* $W \in \mathcal{W}(\Theta_1)$*, the Type-II error is* uniformly *minimized over* Θ_1 *by setting* W *to a degenerate distribution putting all mass on* $\underline{\delta}$*:*

$$\text{for all } \theta \in \Theta_1 : \min_{W \in \mathcal{W}(\Theta_1)} P_\theta\left(\frac{p_W(\mathbf{Y})}{p_0(\mathbf{Y})} \geq \frac{1}{\alpha}\right) = P_\theta\left(\frac{p_{\underline{\delta}}(\mathbf{Y})}{p_0(\mathbf{Y})} \geq \frac{1}{\alpha}\right),$$

and with the test that rejects iff $p_{\underline{\delta}}(\mathbf{Y})/p_0(\mathbf{Y}) \geq 1/\alpha$*,* $\mathcal{H}_0$ *will be rejected iff the ML estimator* $\widehat{\theta}$ *satisfies* $\widehat{\theta} \geq \underline{\delta}$.

Theorem 5.5 shows that, in the context of 1-sided testing with 1-parameter exponential families, if a GROW E-variable is to be used in a safe test with given significance level α and one is further interested in maximizing power among all GROW E-variables (i.e. with respect to any set $\mathcal{W}_1'$ of priors on Θ_1), then one should use the simple E-variable $E^*_{\underline{\delta}}$ with $D(P_{\underline{\delta}}(Y_1) \| P_0(Y_1)) = (-\log \alpha)/n$ since this will lead to the uniformly most powerful GROW test.

Example 5.2. [Normal Location, 1- and 2-sided] Consider the normal location setting of Section 5.1.3 with $\Theta_0 = \{0\}$ as before, and $\mu \in \Theta_1$, the mean, the parameter of interest. First take $\Theta_1 = \mathbb{R}^+$, i.e. a one-sided test. Then $E^*_{\Theta(\mu)} = p_\mu(\mathbf{Y})/p_0(\mathbf{Y})$ and has $\mathrm{GR}(\Theta(\mu)) = D(\mu \| 0) = (n/2)\|\mu^2\|$. We now see that the uniformly most powerful δ-GROW E-variable at sample size n is given by the $\widetilde{\mu}_n$ with $D(\widetilde{\mu}_n \| 0) = -\log \alpha$, so that $\widetilde{\mu}_n = \sqrt{2(-\log \alpha)/n}$. Thus (unsurprisingly), this GROW E-variable is a likelihood ratio test between 0 and $\widetilde{\mu}_n$ at distance to 0 of order $1/\sqrt{n}$, and we expect to gain (at least) $-\log \alpha$ in capital growth if data are sampled from $\mu \geq \widetilde{\mu}_n$.

In the two-sided case, with $\Theta_1 = \mathbb{R}$, we can pick the almost-δ-GROW simple E-value (5.24), i.e. $E^\circ_\mu = \big((1/2)p_\mu(\mathbf{Y}) + (1/2)p_{-\mu}(\mathbf{Y})\big)/p_0(\mathbf{Y})$. Using the distributions' symmetry around 0, we can show (Appendix 5.D) that in this case, $E^\circ_\mu = E^*_\mu$, i.e. $E^\circ_{\underline{\mu}}$ is in fact GROW for $\Theta(\underline{\mu}) = \{P_\mu : |\mu| \geq \underline{\mu}\}$. Even though in this 2-sided case we have no proof that it results in a uniformly most powerful δ-GROW E-variable, we can still, when aiming for a high-power test, take our cue from the 1-sided cases and pick $E^\circ_{\widetilde{\mu}_n}$ for the $\widetilde{\mu}_n$ such that $\mathrm{GR}(\Theta(\widetilde{\mu}_n)) = -\log \alpha$. This leads to the test we described in Section 5.1.3 with threshold $\sqrt{c_n/n} \to 2.72/\sqrt{n}$.

5.4.2 Nonparametric E-Variables

Some of the most well-known classical nonparametric tests are based on identifying a statistic $\mathbf{U} = f(\mathbf{Y})$ that has the same distribution $P_0[\mathbf{U}]$ under all $\theta \in \Theta_0$. This $\mathbf{U}$ is then the test statistic on which a P-value is based. At the same time, it is common to report an *(empirical) effect size* $\widehat{\delta}(\mathbf{U})$ for such a test, giving an indication of the found deviation from the null; the precise definition of $\widehat{\delta}$ varies from case to case. For any distribution P for $\mathbf{Y}$ and any given definition of $\widehat{\delta}$ we will write $\delta(P) := \mathbf{E}_{\mathbf{U} \sim P}[\widehat{\delta}(\mathbf{U})]$ for the population effect size. For simplicity we restrict ourselves to cases in which $\widehat{\delta}$ is a monotonically increasing function of U and $\delta(P_0) = 0$. Assuming we have chosen a test statistic U and a definition for $\widehat{\delta}$, we can extend the previous definitions to δ-GROW E-variables based on U or equivalently, $\widehat{\delta}$. The idea is

that $\mathcal{H}_0$ and $\mathcal{H}_1$ are so large that a GROW (or uniformly-most-powerful) E-variable among *all* E-variables for $\mathcal{H}_0$ and $\mathcal{H}_1$ does not exist or is too hard to find; instead we make life easier by searching for the E-variable that is GROW among all E-variables that can be written as a function of $\mathbf{U}$, which is a strict subset of those that can be written as a function of $\mathcal{Y}$. This is easier since $\mathbf{U}$ has the same distribution $P_0[\mathbf{U}]$ under all $P_0 \in \mathcal{H}_0$. To this end, assume $P_0[\mathbf{U}]$ has density p_0 against some background measure μ. We define P_λ as the distribution with density $p_\lambda(u) \propto \exp(\lambda\widehat{\delta}(u))p_0(u)$. Let Λ be the set of λ for which P_λ is well-defined, i.e. for which $\int p_0(u)\exp(\lambda\widehat{\delta}(u))\,\mathrm{d}\mu(u) < \infty$. Then $\mathcal{P} := \{P_\lambda : \lambda \in \Lambda\}$ is an exponential family given in its natural parameterization, and by a standard property of exponential families, $\mathbf{E}_{P_\lambda}[\widehat{\delta}(\mathbf{U})]$ is monotonically increasing in λ. Rephrasing in the mean-value parameterization we can thus write $P_{[\delta]} := P_{\lambda_\delta}$ where λ_δ is the λ such that $\mathbf{E}_{P_\lambda}[\widehat{\delta}(\mathbf{U})] = \delta$.

Consider a one-sided test with $\mathcal{H}_1$ representing $\delta(P) > 0$. Since we have reduced the problem to the 1-sided 1-dimensional exponential family case of Section 5.4.1, we can once again conclude (5.21). That is, for $\underline{\delta} > 0$ such that $P_{[\underline{\delta}]}[\mathbf{U}]$ is well-defined, we have that $E^* = p_{[\underline{\delta}]}(\mathbf{U})/p_{[0]}(\mathbf{U})$ is a simple E-variable that is GROW relative to the set $\{P \in \mathcal{H}_1 : \delta(P) \geq \underline{\delta}\}$, for data coarsened to $\mathbf{U}$. We can then define a simple two-sided E-variable analogously to Example 5.2. Also, Theorem 5.5 for 1-dimensional exponential families above tells us that, for $\underline{\delta}$ chosen so that

$$D\left(P_{[\underline{\delta}]}[\mathbf{U}] \| P_{[0]}[\mathbf{U}]\right) = -\log \alpha, \tag{5.26}$$

the uniformly-most-powerful GROW safe test is the test that rejects iff $E^* \geq 1/\alpha$, *under the assumption that* $\mathbf{U} \sim P_\delta$ *for* $\delta \neq 0$. While by construction we can assume that $\mathbf{U} \sim P_0$ under the null, we cannot assume that $\mathbf{U} \sim P_\delta$ for some δ under the alternative; our constructed model may be misspecified. Whether E^* still has a UMP property is thus an interesting question for future research.

Example 5.3. In the *Mann-Whitney U test*, we are given $n = n_a + n_b$ outcomes, with n_a outcomes in group a and n_b in group b. This can be represented as n pairs (X_i, Y_i) with $X_i \in \{a, b\}$, $Y_i \in \mathbb{R}$, X_i indicating the group of the ith outcome, and $n_j = \sum_{i=1}^n \mathbf{1}_{X_i = j}$, for $j \in \{a, b\}$. Under $\mathcal{H}_1$, all outcomes in group a are i.i.d., all outcomes in group b are i.i.d., but the two distributions are not the same; under $\mathcal{H}_0$, all outcomes are i.i.d. with the same distribution.

The Mann-Whitney U test is based on the Mann-Whitney U statistic (see any text book for a definition). For every fixed n_a and n_b, under all $P \in \mathcal{H}_0$, i.e all distributions such that $\mathbf{Y} = (Y_1, \ldots, Y_{n_a+n_b})$ is i.i.d. with $Y_i \perp X_i$, $\mathbf{U}$ has the same discrete distribution $P_{[0]}[\mathbf{U}]$ with mass function $p_{[0]}(u)$ with some finite support $\mathcal{U}$. $\mathbf{U}$ is normally used to calculate a P-value. Instead, we use it to calculate an E-value in the manner indicated above: a standard effect size for the Mann-Whitney test is $\mathbf{U}/(n_a n_b)$. Instead for convenience we take $\widehat{\delta} = \mathbf{U}/(n_a n_b) - 1/2$, so that $\mathbf{E}_{P_0}[\widehat{\delta}] = 0$. Define

$$p_\lambda(\mathbf{u}) := \frac{p_0(\mathbf{u}) \cdot e^{\lambda\widehat{\delta}(\mathbf{u})}}{\sum_{\mathbf{u}' \in \mathcal{U}} p_0(\mathbf{u}') e^{\lambda\widehat{\delta}(\mathbf{u}')}}$$

Since $\mathbf{U}$ has a finite range, p_λ is well-defined for $\lambda \in \mathbb{R}$ and it is the probability mass function of the P_λ defined earlier. Then $P_{[\delta]}(\mathbf{U}) = P_\lambda(\mathbf{U})$ for the λ with $\mathbf{E}_{P_\lambda}[\mathbf{U}] = \delta$, and the GROW E-variables relative to $\{P \in \mathcal{H}_1 : \delta(P) \geq \underline{\delta}\}$ are simple: they are likelihood ratios for coarsened data $\mathbf{U}$ of the form $p_{[\underline{\delta}]}(\mathbf{U})/p_{[0]}(\mathbf{U})$.

5.4.3 The Bayesian t-test and the simple δ-GROW t-test

Jeffreys, 1961 proposed a Bayesian version of the t-test; see also (Rouder et al., 2009). We start with the models $\mathcal{H}_0$ and $\mathcal{H}_1$ for data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ given as $\mathcal{H}_0 = \{P_{0,\sigma}(\mathbf{Y}) \mid \sigma \in \Gamma\}$; $\mathcal{H}_1 = \{P_{\delta,\sigma}(\mathbf{Y}) \mid (\delta, \sigma) \in \Theta_1\}$, where $\Delta = \mathbb{R}, \Gamma = \mathbb{R}^+$, $\Theta_1 := \Delta \times \Gamma$ and $\Theta_0 = \{(0, \sigma) : \sigma \in \Gamma\}$, and $P_{\delta,\sigma}$ has density

$$p_{\delta,\sigma}(y) = \frac{\exp\left(-\frac{n}{2}\left[\left(\frac{\overline{y}}{\sigma} - \delta\right)^2 + \left(\frac{\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2}{\sigma^2}\right)\right]\right)}{(2\pi\sigma^2)^{n/2}},$$

with $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

Jeffreys proposed to equip $\mathcal{H}_1$ with a Cauchy prior $W^c[\delta]$ on the *effect size* δ, and both $\mathcal{H}_0$ and $\mathcal{H}_0$ with the scale-invariant prior measure with density $w^H(\sigma) \propto 1/\sigma$ on the variance. Below we first show that, even though this prior is improper (whereas the priors appearing in Theorem 5.4 are invariably proper), the resulting Bayes factor is an E-variable. We then show that, for priors $W[\delta]$ with more than 2 moments, it is in fact even the GROW E-variable relative to all distributions in $\mathcal{H}_1$ compatible with $W[\delta]$. Thus, GROW optimality holds for most priors $W[\delta]$ one might want to use, including standard choices (such as a standard normal) and nonstandard choices (such as the two-point prior we will suggest further below) but ironically not to the moment-less Cauchy proposed by Jeffreys.

Almost Bayesian Case: prior on δ available For any proper prior distribution $W[\delta]$ on δ and any proper prior distribution $W[\sigma]$ on σ, we define

$$p_{W[\delta],W[\sigma]}(y) = \int_{\delta\in\Delta}\int_{\sigma\in\Gamma} p_{\delta,\sigma}(y)\,\mathrm{d}W[\delta]\,\mathrm{d}W[\sigma],$$

as the Bayes marginal density under the product prior $W[\delta] \times W[\sigma]$. In case that $W[\sigma]$ puts all its mass on a single σ, this reduces to:

$$p_{W[\delta],\sigma}(y) = \int_{\delta\in\Delta} p_{\delta,\sigma}(y)\,\mathrm{d}W[\delta]. \tag{5.27}$$

For convenience later on we set the sample space to be $\mathcal{Y}^n = (\mathbb{R} \setminus \{0\}) \times \mathbb{R}^{n-1}$, assuming beforehand that the first outcome will not be 0 — an outcome that has measure 0 under all distributions in $\mathcal{H}_0$ and $\mathcal{H}_1$ anyway. Now we define $\mathbf{V} := (V_1, \ldots, V_n)$ with $V_i = Y_i/|Y_1|$. We have that $\mathbf{Y}$ determines $\mathbf{V}$, and $(\mathbf{V}, Y_1)$ determines $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. The distributions in $\mathcal{H}_0 \cup \mathcal{H}_1$ can thus alternatively be thought of as distributions on the pair $(\mathbf{V}, Y_1)$. $\mathbf{V}$ is "$\mathbf{Y}$ with the scale divided out": it is well-known (and easy to check, see Appendix 5.E) that under all $P \in \mathcal{H}_0$, i.e. all $P_{0,\sigma}$ with $\sigma > 0$, $\mathbf{V}$ has the same distribution $P_0[\mathbf{V}]$ with density p'_0. Similarly, one shows that under all $P_{W[\delta],\sigma}$ with $\sigma > 0$, $\mathbf{V}$ has the same pdf $p'_{W[\delta]}$ (which therefore does not depend on the prior on σ). We now get that, for all $\sigma > 0$,

$$E^*_{W[\delta]}\langle \mathbf{V}\rangle := \frac{p'_{W[\delta]}(\mathbf{V})}{p'_0(\mathbf{V})} \tag{5.28}$$

satisfies $\mathbf{E}_{\mathbf{V}\sim P}[E^*_{W[\delta]}\langle\mathbf{V}\rangle] = 1$ for all $P \in \mathcal{H}_0$, hence it is an E-variable. Here we introduced the notation $E^*_{W[\delta]}\langle\mathbf{V}\rangle$ for E-variables that are GROW relative to W for data 'at level' $\mathbf{V}$, i.e among all E-variables that can be written as functions of $\mathbf{V}$ (see Appendix 5.A for further explanation). Remarkably, this 'scale-free' E-variable coincides with the Bayes factor one gets if one uses, for σ, the prior $w^H(\sigma) = 1/\sigma$ suggested by Jeffreys, and treats σ and δ as independent. That is, as shown in Appendix 5.E, we have

$$\frac{\int_\sigma \overline{p}_{W[\delta],\sigma}(\mathbf{Y}) w^H(\sigma)\,\mathrm{d}\sigma}{\int_\sigma p_{0,\sigma}(\mathbf{Y}) w^H(\sigma)\,\mathrm{d}\sigma} = \frac{p'_{W[\delta]}(\mathbf{V})}{p'_0(\mathbf{V})} = E^*_{W[\delta]}\langle\mathbf{V}\rangle. \tag{5.29}$$

Despite its improperness, w^H induces a valid E-variable when used in the Bayes factor. The equivalence of this Bayes factor to $E^*_{W[\delta]}\langle\mathbf{V}\rangle$ simply means that it manages to ignore the 'nuisance' part of the model and models the likelihood of the scale-free $\mathbf{V}$ instead. The reason this is possible is that w^H coincides with the right-Haar prior for this problem (Eaton, 1989; Berger, Pericchi and Varshavsky, 1998), about which we will say more below. Amazingly, it turns out that the E-variable (5.29) is GROW (among all E-variables for data $\mathbf{Y}$, not just the coarsened $\mathbf{V}$!) under the weak condition that the prior $W[\delta]$ has a $(2+\varepsilon)$th moment. This follows from Part 1 of our second main result, Theorem 5.6 below. Its proof is by no means straightforward (at least, we did not find a simple proof). Let, for priors $W[\delta]$, $W[\sigma]$, $P^{[\mathbf{V}]}_{W[\delta],W[\sigma]}$ be the marginal distribution on $\mathbf{V}$, i.e. the distribution with density $p'_{W[\delta],W[\sigma]}$.

Theorem 5.6. *Let $W[\delta]$ be a distribution on δ such that for some $\varepsilon > 0$, $\mathbf{E}_{\delta\sim W[\delta]}[|\delta|^{2+\varepsilon}] < \infty$ for some $\varepsilon > 0$ (in particular this includes all degenerate priors with mass 1 on a single $\overline{\delta}$). Let $\mathcal{W}[\Gamma]$ be the set of all distributions $W[\sigma]$ on the variance σ. We have:*

$$\begin{aligned}\inf_{W'[\sigma],W[\sigma]\in\mathcal{W}(\Gamma)} D(P_{W[\delta],W'[\sigma]}\|P_{0,W[\sigma]}) &= \inf_{W[\sigma]\in\mathcal{W}(\Gamma)} D(P_{W[\delta],W[\sigma]}\|P_{0,W[\sigma]}) \\ &= D(P^{[\mathbf{V}]}_{W[\delta]}\|P^{[\mathbf{V}]}_0).\end{aligned} \tag{5.30}$$

More generally, fix a convex set of distributions $\mathcal{W}[\delta]$ on δ such that, for some $\varepsilon > 0$, each $W[\delta] \in \mathcal{W}[\delta]$ satisfies $\mathbf{E}_{\delta\sim W[\delta]}[|\delta|^{2+\varepsilon}] < \infty$. Let $\mathcal{W}'_1$ be a set of probability distributions on $\delta\times\sigma$ such that, for each $W[\delta] \in \mathcal{W}[\delta]$ and each distribution $W[\sigma] \in \mathcal{W}(\Gamma)$ on σ, $\mathcal{W}'$ contains a distribution whose marginal on δ coincides with $W[\delta]$ and whose marginal on σ coincides with $W[\sigma]$. We then have:

$$\begin{aligned}\inf_{W\in\mathcal{W}'_1}\inf_{W[\sigma]\in\mathcal{W}[\Gamma]} D(P_W\|P_{0,W[\sigma]}) &= \inf_{W[\delta]\in\mathcal{W}[\delta]}\inf_{W[\sigma]\in\mathcal{W}[\Gamma]} D(P_{W[\delta],W[\sigma]}\|P_{0,W[\sigma]}) \\ &= \inf_{W[\delta]\in\mathcal{W}[\delta]} D(P^{[\mathbf{V}]}_{W[\delta]}\|P^{[\mathbf{V}]}_0).\end{aligned} \tag{5.31}$$

Part 1 of this theorem allows us to use Part 3 of Theorem 5.4 to conclude that $E^*_{W[\delta]}\langle\mathbf{V}\rangle = E^*_{\mathcal{W}_1}$: the Bayes factor based on the right Haar prior, is not just an E-variable, but even the GROW E-variable relative to the set of all priors on $\delta\times\sigma$ that are compatible with $W[\delta]$.

Simple GROW safe t-test: prior on δ not available What if we have no clear idea on how to choose a marginal prior on $\underline{\delta}$? In that case, we can once again use the $\underline{\delta}$-GROW E-variable for $\underline{\delta}$. First, consider 1-sided tests. In Appendix 5.D we show that (5.21) holds in this case, i.e. $\min_{W \in \mathcal{W}(\Theta(\underline{\delta}))} D(P_W^{[\mathbf{Y}]} \| P_0^{[\mathbf{Y}]})$ is achieved for the degenerate prior that puts mass 1 on $\underline{\delta}$, i.e. the $\underline{\delta}$-GROW E-variable is simple. We can then use Theorem 5.6 above to infer that the Bayes factor based on the right Haar prior w^H on σ and this point prior on $\underline{\delta}$, i.e. $E^*_{\underline{\delta}} = p'_{\underline{\delta}}(\mathbf{V})/p'_0(\mathbf{V})$ is equal to the GROW E-variable relative to $\Theta(\underline{\delta})$. Mutatis mutandis, the same holds for the 2-sided test: as shown in Appendix 5.D, with the GROW set $\Theta(\underline{\delta}) = \{\delta : |\delta| \geq \underline{\delta}\}$ we get that the $\underline{\delta}$-GROW E-variable is simple, and given by the Bayes factor with, for $\mathcal{H}_1$, the prior on δ that puts mass 1/2 on $\underline{\delta}$ and 1/2 on $-\underline{\delta}$.

Optional Stopping For any prior $W[\delta]$, $E^*_{W[\delta]}$ defines a test statistic process $(E^*_{W[\delta]}\langle V^i \rangle)_{i \in \mathbb{N}}$ with $E*_{W[\delta]}\langle V^i \rangle = p'_{W[\delta]}(V^i)/p'_0(V^i)$. Notably, tests based on this process are safe for optional stopping under Definition 5.2; by Proposition 3, this process defines a test martingale and hence, by the same proposition, the threshold test based on $(E^*_{W[\delta]}\langle V^i \rangle)_{i \in \mathbb{N}}$ preserves Type I error guarantees also under optional stopping. As indicated by (Hendriksen, De Heide and Grünwald, 2020), this test does not necessarily preserve Type-I error guarantees under optional stopping with stopping rules that can only be written as function of $Y_1, Y_2, \ldots$ and not of $V_1, V_2, \ldots$. But, since $E^*_{W[\delta]}\langle V^i \rangle$ is a function of the V_i, it does allow for the prototypical instance of optional stopping, where we stop at the smallest t at which $E^*_{W[\delta]}\langle V^t \rangle > 20 = 1/\alpha$. The insight that $E^*_{W[\delta]}$ provides a test martingale is not new: as we learned from A. Ramdas, it was already considered by Robbins, 1970.

Extension to General Group Invariant Bayes Factors In a series of papers (Berger, Pericchi and Varshavsky, 1998; Dass and Berger, 2003; Bayarri et al., 2012), Berger and collaborators developed a theory of Bayes factors for $\mathcal{H}_0 = \{P_{0,\gamma} : \gamma \in \Gamma\}$ and $\mathcal{H}_1 = \{P_{\delta,\gamma} : \delta \in \Delta, \gamma \in \Gamma\}$ with a nuisance parameter (vector) γ that appears in both models and that satisfies a group invariance; the Bayesian t-test is the special case with $\gamma = \sigma, \Gamma = \mathbb{R}^+$ and with the scalar multiplication group and δ an 'effect size'. Other examples include regression based on mixtures of g-priors (Liang et al., 2008) and the many examples given by e.g. Berger, Pericchi and Varshavsky, 1998; Dass and Berger, 2003, such as testing a Weibull vs. the log-normal or an exponential vs. the log-normal. The reasoning of the first part of this section straightforwardly generalizes to all such cases: under some conditions on the prior on δ, the Bayes factor based on using the right Haar measure on γ in both models gives rise to an E-variable. We furthermore *conjecture* that in all such testing problems, the resulting Bayes factor is even GROW relative to a suitably defined set $\mathcal{W}_1$; i.e. that a suitable analogue of Theorem 5.6 holds. The proof of this theorem seems extendable to the general group invariant setting, with the possible exception of Lemma 12 in Appendix 5.E which uses particular properties of the variance of a normal; generalizing this lemma (which also requires us to handle models with a nonunique right Haar prior (Sun and Berger, 2007), for which it is not immediately clear how a generalization would look like) is a major goal for future work.

5.4.4 Contingency Tables

Let $\mathcal{Y}^n = \{0,1\}^n$ and let $\mathcal{X} = \{a, b\}$ represent two categories. We start with a multinomial model $\mathcal{G}_1$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, extended to n outcomes by independence. We want to test whether the Y_i are dependent on the X_i. To this end, we condition every distribution in $\mathcal{G}_1$ on a fixed, given, $\mathbf{X} = \mathbf{x} = (x_1, \ldots, x_n)$, and we let $\mathcal{H}_1$ be the set of (conditional) distributions on $\mathcal{Z}$ that thus result.

We thus assume the design of $\mathcal{X}^n$ to be set in advance, but N_1, the number of ones, to be random; alternative choices are possible and would lead to a different analysis. Conditioned on $\mathbf{X} = \mathbf{x}$, the counts n, $n_a = N_a(\mathbf{x})$ and n_b (see Table 5.1), the likelihood of an individual sequence $\mathbf{y} \mid \mathbf{x}$ with statistics $N_{a0}, N_{b0}, N_{b0}, N_{b1}$ becomes:

$$\begin{aligned} p_{\mu_{1|a},\mu_{1|b}}(\mathbf{y} \mid \mathbf{x}) &= p_{\mu_{1|a},\mu_{1|b}}(y \mid \mathbf{x}, n_a, n_b, n) \\ &= \mu_{1|a}^{N_{a1}}(1-\mu_{1|a})^{N_{a0}} \cdot \mu_{1|b}^{N_{b1}}(1-\mu_{1|b})^{N_{b0}} \end{aligned} \tag{5.32}$$

These densities define the alternative model $\mathcal{H}_1 = \{P_{\mu_{1|a},\mu_{1|b}} : (\mu_{1|a}, \mu_{1|b}) \in \Theta_1\}$ with $\Theta_1 = [0,1]^2$. $\mathcal{H}_0$, the null model, simply has $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ independent, with $Y_i, \ldots, Y_n$ i.i.d. Ber(μ_1) distributed, $\mu_1 \in \Theta_0 := [0,1]$, i.e.

$$p_{\mu_1}(\mathbf{y} \mid \mathbf{x}) = p_{\mu_1}(\mathbf{y}) = \mu_1^{N_1}(1-\mu_1)^{N_0}.$$

To test $\mathcal{H}_0$ against $\mathcal{H}_1$, we numerically calculate the GROW E-variable $E^*_{\Theta(\varepsilon)}$ where $\Theta(\varepsilon)$

	0	1	sum
a	μ_{a0}	μ_{a1}	μ_a
b	μ_{b0}	μ_{b1}	μ_b
sum	μ_0	μ_1	1

	0	1	sum
a	N_{a0}	N_{a1}	n_a
b	N_{b0}	N_{b1}	n_b
sum	N_0	N_1	n

Table 5.1: 2x2 contingency table: parameters and counts. μ_{ij} is the (unconditional) probability of observing category i and outcome j, and N_{ij} is the corresponding count in the observed sample.

is defined via (5.20) for two different divergence measures detailed further below. In both cases, $\Theta(\varepsilon)$ will be compact, so that by the joint lower-semi-continuity of the KL divergence (Posner, 1975), $\min D(P_{W_1} \| P_{W_0})$ is achieved by some unique (W_1^*, W_2^*), and we can use Part 3 of Theorem 5.4 to infer that the GROW E-variable is given by $E^*_{\mathcal{W}(\Theta(\varepsilon))} = E^*_{\Theta(\varepsilon)} = p_{W_1^*}(\mathbf{Y} \mid \mathbf{X}) / p_{W_0^*}(\mathbf{Y})$. Note that the 'priors' W_1^* and W_0^* may depend on the observed $\mathbf{x} = x^n$, in particular on n_a and n_b, since we take these as given throughout. We can further employ Carathéodory's theorem (see Appendix 5.E.2 for details) to give us that W_1^* and W_0^* must have finite support, which allows us to find them reasonably efficiently by numerical optimization; we give an illustration in the next section.

We now consider two definitions of $\Theta(\varepsilon)$. The first option is to think of μ_1 as a 'nuisance' parameter: we want to test for independence, and are not interested in the precise value of μ_1, but rather in the 'effect size' $\delta := |\mu_{1|a} - \mu_{1|b}|$. We can then, once again, use the δ-GROW E-variable for parameter of interest δ. To achieve this, we re-parameterize the model in a manner that depends

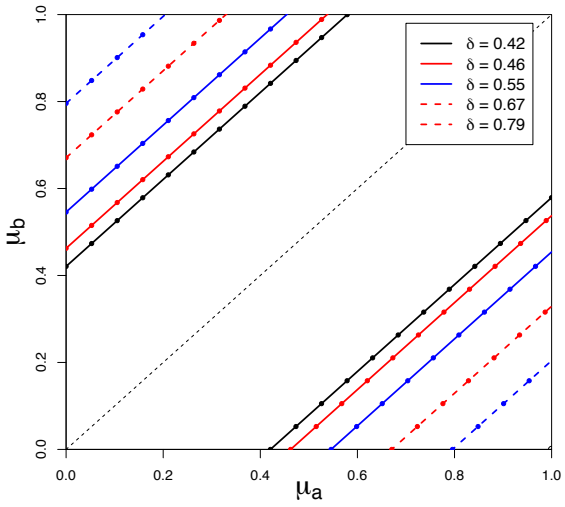


Figure 5.2: The Beam: Graphical depiction of the GROW $\Theta(\delta)$.

on $\mathbf{x}$ via n_a and n_b. For given $\mu_{1|a}$ and $\mu_{1|b}$, we set $\mu_1 = (n_a\mu_{1|a} + n_b\mu_{1|b})/n$, and δ as above, and we define $p'_{\delta,\mu_1}(\mathbf{y}|\mathbf{x})$ (the probability in the new parameterization) to be equal to $p_{\mu_{1|a},\mu_{1|b}}(\mathbf{y}|\mathbf{x})$ as defined above. As long as $\mathbf{x}$ (and hence n_a and n_b) remain fixed, this re-parameterization is 1-to-1, and all distributions in the null model $\mathcal{H}_0$ correspond to a p'_{δ,μ_1} with $\delta = 0$. In Figure 5.2 we show, for the case $n_a = n_b = 10$, the sets $\Theta(\delta)$ for $\delta = \{0.42, 0.46, 0.55, 0.67, 0.79\}$. For example, for $\delta = 0.42$, $\Theta(\delta)$ is given by the region on the boundary, and outside of, the 'beam' defined by the two depicted lines closest to the diagonal. We numerically determined the JIPr, i.e., the prior $(P_{W_0^*}, P_{W_1^*})$ for each choice of δ. This prior has finite support, the support points are depicted by the dots; in line with intuition, we find that the support points for priors on the set $\Theta(\delta)$ are always on the line(s) of points closest to the null model, i.e. the δ-GROW E-variable is simple. Variations of this definition of $\Theta(\delta)$ and corresponding GROW E-values have been considered by Turner, 2019, who showed that for one-sided testing, one can calculate the above JIPr analytically; moreoever, if data comes in as pairs of each group, so that all X_i are give by (a, b) and $Y_i = (y_{ia}, y_{ib}) \in \{0, 1\}^2$, then on this rougher filtration, (where $n_a = n_b$ at all sample points), the JIPR for each n defines a test martingale and, along the lines of Proposition 3, we can use it for testing that is safe under optional stopping. The second option for defining $\Theta(\varepsilon)$ is to take the original parameterization, and have d in (5.20)) be the KL divergence. This choice is motivated in Appendix 5.F. Then $\Theta(\varepsilon)$ is the set of $(\mu_{1|a}, \mu_{1|b})$ with

$$\inf_{\mu'_1 \in [0,1]} \frac{D(P_{\mu_{1|a},\mu_{1|b}} \| P_{\mu'_1})}{n} = \frac{D(P_{\mu_{1|a},\mu_{1|b}} \| P_{\mu_1})}{n} \geq \varepsilon.$$

Note that the scaling by $1/n$ is just for convenience — since $P_{\mu|\cdot}$ are defined as distributions of samples of length n, the KL grows with n and our scaling ensures that, for given $\mu_{1|a}, \mu_{1|b}$ and

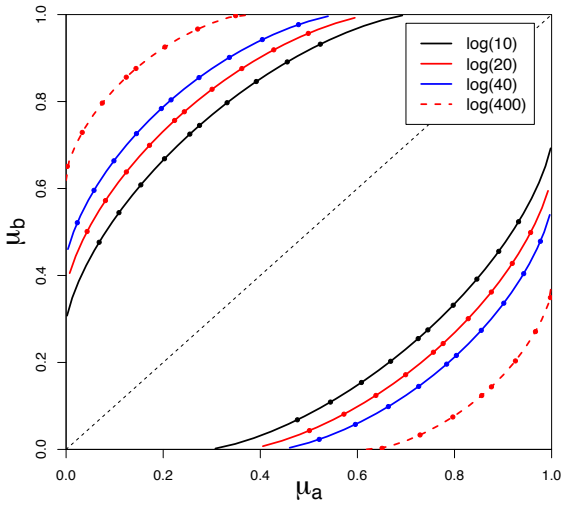


Figure 5.3: The Lemon: Graphical depiction of the KL-divergence based GROW $\Theta(\varepsilon)$.

$n_{1a}, n_{1,b}$, the set $\Theta(\varepsilon)$ does not change if we multiply n_{1a} and n_{1b} by the same fixed positive integer. Note also that the distributions $P_{\mu_{1|a},\mu_{1|b}}$ and P_{μ_1} are again conditional on the given $\mathbf{x}$ (and hence n_a and n_b), and $\mu_1 = (n_a\mu_{1|a} + n_b\mu_{1|b})/n$ as before. We can now numerically determine $\Theta(\varepsilon)$ for various values of ε; this is done in Figure 5.3, where, for example, the set $\Theta(\varepsilon)$ for $\varepsilon \in \{\log 10, \log 20, \ldots, \log 400\}$ is given by all points on and outside of the innermostly depicted 'lemon'. Again, we can calculate the corresponding JIPr; the support points of the corresponding priors are also shown in Figure 5.3.

5.4.5 General Exponential Families

The contingency table setting is an instance of a test between two nested (conditional) exponential families. We can extend the approach of defining GROW sets $\Theta(\varepsilon)$ relative to distance measures d and numerically calculating corresponding JIPrs $(P_{W_1^*}, P_{W_0^*})$ straightforwardly to this far more general setting. As long as Theorem 5.4, Part 3 can be applied with $\mathcal{W}_1' = \mathcal{W}(\Theta(\varepsilon))$, the resulting Bayes factor $p_{W_1^*}(\mathbf{Y})/p_{W_0^*}(\mathbf{Y})$ will be a GROW E-variable. The main condition for Part 3 is the requirement that $D(P_{W_1'} \| P_{W_0^*}) < \infty$ for all $W' \in \mathcal{W}(\Theta(\varepsilon))$, which automatically holds if $D(P_\theta \| P_{W_0^*}) < \infty$ for all $\theta \in \Theta(\varepsilon)$. Since, for exponential families, $D(P_\theta \| P_{\theta'}) < \infty$ for all θ, θ' in the interior of the parameter space $\Theta = \Theta_1$, this condition can often be enforced to hold though, if we take a divergence measure d such that for each $\varepsilon > 0$, $\Theta(\varepsilon)$ is a compact subset of Θ_1 and for each $\theta \in \Theta_1$ that is not on the boundary, there is an $\varepsilon > 0$ such that $\theta \in \Theta(\varepsilon)$.

For large n though, numerical calculation of GROW E-variables may be time consuming, and

one may wonder whether there exists other nontrivial (but perhaps not GROW, or at least not GROW relative to any intuitive sets $\Theta(\varepsilon)$) E-variables that take less computational effort. It turns out that these exist: one can calculate a *conditional* GROW-E-variable. We illustrate this for the contingency table setting. Fix an arbitrary function g mapping $\mathbf{x}$ to $\mathcal{W}(\Theta_1)$, the set of priors on Θ_1. Conditional on the sufficient statistic relative to $\mathcal{H}_0$, $\widehat{\mu}_1(\mathbf{Y}) = N_1/n$, all distributions in $\mathcal{H}_0$ assign the same probability mass $p_0(\mathbf{y} \mid \widehat{\mu}_1(\mathbf{y})) = 1/\binom{n}{N_1}$ to all $\mathbf{y}$ with $\widehat{\mu}_1(\mathbf{y}) = \widehat{\mu}_1(\mathbf{Y})$. The conditional E-variable based on g is then given by

$$E = \frac{p_{g(\mathbf{x})}(\mathbf{Y} \mid \widehat{\mu}_1(\mathbf{Y}), \mathbf{x})}{p_0(\mathbf{Y} \mid \widehat{\mu}_1(\mathbf{Y}))} = \binom{n}{N_1} \cdot \frac{p_{g(\mathbf{x})}(\mathbf{Y} \mid \mathbf{x})}{p_{g(\mathbf{x})}(\widehat{\mu}_1(\mathbf{Y}) \mid \mathbf{x})}. \tag{5.33}$$

This gives a conditional (and hence also unconditional) E-variable for every choice of function $g(\mathbf{x})$. In fact it coincides with what has been called a method for obtaining 'clean' evidence for the 2×2 table setting by eliminating the nuisance parameter $\widehat{\mu}_1$ (Royall, 1997). In settings with optional stopping based on the value of $\widehat{\mu}_1$, it has a GROW-like optimality property for certain choices of g which we will further explore in future work. In settings with fixed n, it is not GROW and may perhaps be seen as a 'quick and dirty' approach to design an E-variable. It clearly can be extended to any combination of $\mathcal{H}_1$ (not necessarily an exponential family) and any exponential family $\mathcal{H}_0$ such that the ML estimator $\widehat{\theta}_0(\mathbf{y})$ is almost surely well-defined under all $P \in \mathcal{H}_0$, whereas at the same time, $\widehat{\theta}_0(\mathbf{Y})$ is a sufficient statistic for $\mathcal{H}_0$, i.e. there is a 1-to-1 correspondence between the ML estimator $\widehat{\theta}_0(\mathbf{Y})$ and the sufficient statistic $\phi(\mathbf{Y})$. This will hold for most exponential families encountered in practice (to be precise, $\mathcal{H}_0$ has to be a regular or 'aggregate' Barndorff-Nielsen, 1978, page 154-158 exponential family). In such cases, if, for example, a reasonable prior W_1 on Θ_1 is available, we can efficiently calculate nontrivial E-variables based on taking $g(\mathbf{x}) = W_1$, but whether these are sufficiently strong approximations of the GROW E-variable will have to be determined on a case-by-case, i.e. model-by-model basis; we did some experiments for the contingency table, with W_1 a Beta prior, and there we found them to be noncompetitive in terms of GROW and power with respect to the full JIPr[1].

5.5 Testing Our GROW Tests

We perform some initial experiments with GROW E-variables for composite $\mathcal{H}_0$ nested within $\mathcal{H}_1$. We consider two common settings: in one setting, we want to perform the most sensitive test possible for a given sample size n; we illustrate this with the contingency table test. In the second setting, we are given a *minimum clinically relevant effect size* $\underline{\delta}$ and we want to find the smallest sample size n for which we can expect good statistical (power) properties.

5.5.1 Case 1: Fixed n, $\underline{\varepsilon}$ unknown

Mini-Simulation-Study 1: The 2x2 Table We first consider the GROW E-variables $E^*_{\Theta(\delta)}$ relative to parameter of interest $\delta = |\mu_{1|a} - \mu_{1|b}|$, the first option considered in Section 5.4.4. For

[1] Although it was not connected to E-variables, the idea to modify Bayes factors for nested exponential families by conditioning on the smaller model's sufficient statistic was communicated to us by T. Seidenfeld, 2016.

a grid of $\underline{\delta}$'s in the range $[0.4, 0.9]$ we looked at the best power that can be achieved by GROW E-variable $E^*_{\Theta(\delta^*)}$, i.e. we looked for the δ^* (again taken from a grid in the range $[0.4, 0.9]$) such that

$$1 - \underline{\beta}(\underline{\delta}, \delta^*) := \inf_{\theta \in \Theta((\underline{\delta}))} P_\theta \left(\log E^*_{\Theta((\delta^*))} \geq -\log \alpha \right) \tag{5.34}$$

is maximized. We summarized the results in Table 5.2. We see that, although we know of no

$\underline{\delta}$	$\mathrm{GR}(\Theta(\underline{\delta})) = D(P_{W_1^*} \| P_{W_0^*})$	δ^*	power $1 - \overline{\beta}$
0.42	1.20194	0.50	0.20
0.46	1.57280	0.50	0.29
0.50	1.99682	0.50	0.39
0.55	2.47408	0.50	0.49
0.59	3.00539	0.50	0.60
0.63	3.59327	0.50	0.69
0.67	4.23919	0.50	0.77
0.71	4.94988	0.50	0.85
0.75	5.73236	0.50	0.91

Table 5.2: Relating $\underline{\delta}, \delta^*$, power and capital growth $\mathrm{GR}(\Theta(\underline{\delta}))$ for $n_a = n_b = 10$ for the GROW E-variables. For example, the row with 0.42 in the first column corresponds to the two black lines in Figure 5.2 which represent all $\theta_1 = (\mu_{1|a}, \mu_{1|b})$ with $\delta = 0.42$.

analogue to Johnson's Theorem 5.5 here, something like a "uniformly most powerful δ-GROW safe test" does seem to exist — it is given by $E^*_{\Theta(\delta^*)}$ with $\delta^* = 0.50$; and we can achieve power 0.8 for all $\theta \in \Theta(\underline{\delta})$ with $\underline{\delta} \gtrapprox 0.5$. The same exercise is repeated with the GROW E-variables defined relative to the KL divergence in Table 5.3, again indicating that there is something like a uniformly most powerful δ-GROW safe test. We now compare four hypothesis tests for contingency tables for the $n_a = n_b = 10$ design: Fisher's exact test (with significance level $\alpha = 0.05$), the *default Bayes Factor* for contingency tables (Gunel and Dickey, 1974; Jamil et al., 2016) (which is turned into a test by rejecting if the Bayes factor $\geq 20 = -\log \alpha$), the 'uniformly most powerful' GROW E-variable $E^*_{\Theta(\delta*)}$ with $\delta^* = 0.50$ (see Table 5.2) which we call GROW$(\Theta(\delta))$ and the 'uniformly most powerful' KL-GROW E-variable $E^*_{\Theta(\varepsilon^*)}$ with $\varepsilon^* = \log 16$ (see Table 5.3) which we call $(\Theta(\varepsilon))$. The 0.8-iso-power lines are depicted in Figure 5.4; for example, if $\theta_1 = (\mu_{1|a}, \mu_{1|b})$ is on or outside the two curved red lines, then Fisher's exact test achieves power 0.8 or higher. The difference between the four tests is in the shape: Bayes and the δ-based JIPr yield almost straight power lines, the KL-based JIPr and Fisher curved. Fisher gives a power ≥ 0.8 in a region larger than the KL-based JIPr, which makes sense because the corresponding test is *not* safe; the δ-GROW and default Bayes factor behave very similarly, but they are not the same: in larger-scale experiments we do find differences. We see similar figures if we compare the rejection regions rather than the iso-power lines of the four tests (figures omitted).

$\log n\underline{\varepsilon}$	$\text{GR}(\Theta(\underline{\varepsilon})) = D(P_{W_1^*} \| P_{W_0^*})$	$\log n\varepsilon^*$	power
2	0.21884	16	0.06
5	0.98684	16	0.18
10	1.61794	16	0.29
15	1.99988	16	0.35
20	2.27332	16	0.40
25	2.48597	16	0.44
30	2.65997	16	0.47
40	2.93317	16	0.52
50	3.14447	16	0.55
100	3.78479	16	0.65
200	4.48606	16	0.74
300	4.86195	16	0.79
400	5.12058	16	0.82

Table 5.3: Relating $\varepsilon, \varepsilon^*$, power and capital growth $\text{GR}(\Theta(\underline{\varepsilon}))$ for $n_a = n_b = 10$ for the KL-GROW E-variables. For example, the row with 20 in the first column corresponds to the two curved red lines in Figure 5.3 which represent all $\theta_1 = (\mu_{1|a}, \mu_{1|b})$ with $\inf_{\mu \in [0,1]} D(P_{\theta_1} \| P_\mu) = \log 20$.

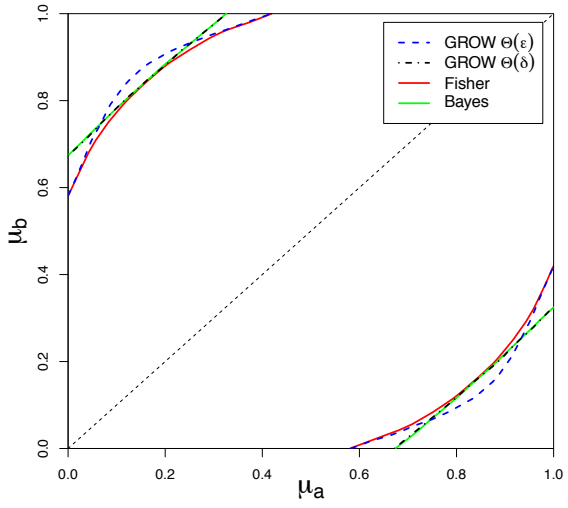


Figure 5.4: 0.8-iso-powerlines for the four different tests.

5.5.2 Case 2: n to be determined, $\underline{\delta}$ known

Consider δ-GROW E-variables for some scalar parameter of interest δ. Whereas in Case 1, the goal was implicitly to detect the 'smallest detectable deviation' from $\mathcal{H}_0$, in Case 2 we know beforehand that we are only really *interested* in rejecting $\mathcal{H}_0$ if $\delta \geq \underline{\delta}$. Here $\underline{\delta} > 0$ is the minimum value at which the statement '$|\delta| \geq \underline{\delta}$' has any practical repercussions. This is common in medical testing in which one talks about the *minimum clinically relevant effect size* $\underline{\delta}$.

Assuming that generating data costs money, we would like to find the smallest possible n at which we have a reasonable chance of detecting that $|\delta| \geq \underline{\delta}$. Proceeding analogously to Case 1, we may determine, for given significance level α and desired power $1 - \underline{\beta}$, the smallest n at which there exist δ^* such that the safe test based on E-variable $E^*_{\Theta(\delta^*)}$ has power at least $1 - \underline{\beta}$ for all $\theta \in \Theta(\underline{\delta})$. Again, both n and δ^* may have to be determined numerically (note that δ^* is not necessarily equal to $\underline{\delta}$).

Mini-Simulation-Study 2: 1-Sample t-test In this simulation study, we test whether the mean of a normal distribution is different from zero, when the variance is unknown. We determine, for a number of tests, the minimum n needed as a function of minimal effect size $\underline{\delta}$ to achieve power at least 0.8 when rejecting at significance level $\alpha = 0.05$. We compare the classical t-test, the Bayesian t-test (with Cauchy prior on δ, turned into a test that is safe under optional continuation by rejecting when $\mathsf{BF} \geq 20 = 1/\alpha$) and our safe test based on the GROW E-variable $E^*_{\Theta(\delta^*)}\langle V^n \rangle = E^*_{\delta^*}\langle V^n \rangle$ that maximizes power while having a GROW property. For the standard t-test we can just compute the required (batch) sample size. This is plotted (black line) in Figure 5.5 as a function of $\underline{\delta}$, where we also plot the corresponding required sample sizes for the Bayesian t-test (larger by a factor of around 1.9 – 2.1) and our maximum power δ^*-GROW t-test (larger by a factor of around 1.4 – 1.6).

However, these three lines do not paint the whole picture: we have already indicated in Section 5.4.3 that for any prior $W[\delta]$, the threshold test based on $(E^*_{W[\delta]}\langle V^i \rangle)_{i \in \mathbb{N}}$ is safe also under optional stopping. Since both the Bayesian t-test and our δ-GROW t-test are an instance of $E^*_{W[\delta]}$ as given by (5.29), we preserve Type-I error guarantees if we stop at the smallest t at which $E^*_{W[\delta]}\langle V^t \rangle > 20 = 1/\alpha$. We can now compute an *effective sample size* under optional stopping in two steps, for given $\underline{\delta}$. First, we determine the smallest n at which the δ^*-GROW E-variable $E^*_{\Theta(\delta^*)}$ which optimizes power achieves a power of at least $0.8 = 1 - \beta$; we call this $n_{\max}$. We then draw data sequentially and record the $E^*_{W[\delta]}\langle V^t \rangle$ until either this E-variable exceeds $1/\alpha$ or $t = n_{\max}$. This new procedure still has Type I error at most α, and it must have power ≥ 0.8. The 'effective sample size' is now the sample size we *expect* if data are drawn from a distribution with effect size at $\underline{\delta}$ and we do optional stopping in the above manner ('stopping' includes both the occasions on which $\mathcal{H}_0$ is accepted and $t = n_{\max}$, and the occasions when $\mathcal{H}_0$ is rejected and $t \leq n_{\max}$). In Figure 5.5 we see that this effective sample size is almost *equal* to the fixed sample size we need with the standard t-test to obtain the required power. Thus, quite unlike the classical t-test, our δ-GROW t-test E-variable preserves Type I error probabilities under optional stopping; it needs more data than the classical t-test in the worst-case, *but hardly more on average under* $\mathcal{H}_1$. For a Neyman-Pearsonian hypothesis tester, this should be a very good reason to adopt it!

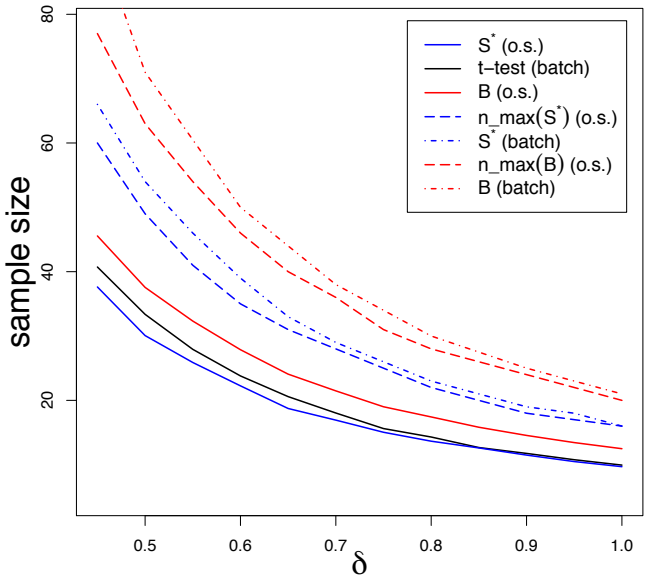


Figure 5.5: Effective sample size for the classical t-test (black), Bayesian t-test (E-test with Cauchy prior on δ) (red), and the δ-GROW E-test E^* with a two-point prior on δ (blue). The lines denoted *batch* denote the smallest fixed sample size at which power $\beta = 0.8$ can be obtained under $\mathcal{H}_1$ as a function of the 'true' effect size δ. The continuous lines, denoted 'o.s.' denote the sample size needed if optional stopping (see main text) is done (and for E^*, the prior is optimized for the batch sizes that were plotted as well. The ratios between the curves at $\delta = 0.5$ and the batch sample size needed for the t-test is 0.9 (E^* with o.s.), 1.1 (Bayes t-test with o.s.), 1.5 (E^* with fixed sample size) and 1.9 (Bayes t-test with fixed sample size). At $\delta = 1$ they are 0.98, 1.26, 1.61 and 2.01 respectively: the amount of data needed compared with the tradition t-test thus increases in δ within the given range. The two lines indicated as '$n_{\max}$ (o.s.)' are explained in the main text.

5.6 Earlier, Related and Future Work

E-Variables, Test Martingales, General Novelty As seen in Section 5.2, E-variables are close cousins of *test martingales*, which go back to Ville, 1939, the paper that introduced the modern notion of a martingale. E-variables themselves have probably been originally introduced by Levin (of P vs NP fame) (1976) (see also (Gács, 2005)) under the name *test of randomness*, but Levin's abstract context is quite different from ours. Independently discovered by Zhang, Glancy and Knill, 2011, they were later analyzed by Shafer et al., 2011; Shafer and Vovk, 2019; Vovk and Wang, 2019; all these authors used different names for the concept. While we originally called them 'S-value', the paper (Vovk and Wang, 2019), which appeared after the first version of the present paper, called them E-variables, a name which we decided to adopt for its better motivation (E can stand both for expectation, just like the P in P-value stands for probability; but also for 'evidence').

Test martingales themselves have been thoroughly investigated by Shafer et al., 2011; Shafer and Vovk, 2019. They themselves underlie AV (anytime-valid) P-values (Johari, Pekelis and Walsh, 2015), AV tests (which we call 'tests that are safe for optional stopping') and AV confidence sequences. The latter were recently developed in great generality by A. Ramdas and collaborators; see e.g. (Balsubramani and Ramdas, 2016; Howard et al., 2018b; Howard et al., 2018a). Both AV tests and confidence sequences have first been developed by H. Robbins and his students (Darling and Robbins, 1967; Lai, 1976; Robbins, 1970). Like we do for E-variables, Ramdas et al. (and also e.g. Pace and Salvan, 2019) stress the promise of the AV notions for a safer kind of statistics that is significantly more robust than standard testing and confidence interval methodology.

Just like regular tests can be turned into confidence intervals by varying the null and 'inverting' the resulting tests, AV confidence intervals can be created by starting with a collection of test martingales, one for each null, and then varying the null and inverting the AV test based on the test martingale for each null. We can do (and plan to investigate in future work) the same thing with E-variables. More generally, the work on AV tests and confidence sequences is very similar in spirit to ours, with our work stressing analysis at the level of batches of data rather than individual data points. Thus, we do not claim any real novelty for the 'safe' or 'always valid' setting. The real novelty is in Theorem 5.4 and 5.6. However, as we discovered after posting the first version of the present paper, a special case of Theorem 5.4 was already formulated and proved[2] by Zhang, Glancy and Knill, 2011 (see also (Zhang, 2013)) who show that GROW E-variables can be constructed for discrete outcome spaces, simple (singleton) $\mathcal{H}_1$ and convex $\mathcal{H}_0$. Theorem 5.4 extends this to its full generality, showing that nontrivial E-variables always exist and that optimal ones can often be constructed, for nonconvex $\mathcal{H}_0$ and $\mathcal{H}_1$ that are both composite — that insight is the main novelty of this paper.

Relation to Sequential Testing *Sequential testing* (Lai, 2009), pioneered by Wald and Barnard and developed much further by H. Robbins and his students, is mathematically similar to testing based on test martingales and (therefore) E-variables. Sequential tests are based on

[2] Zhang, Glancy and Knill, 2011 was in turn inspired by Van Dam, Gill and Grunwald, 2005, co-authored by one of us, which identifies the importance of the KL divergence in test design but falls short of defining E-values.

random processes $(S_i)_{i\in\mathbb{N}}$ that are a *likelihood ratio of (potentially coarsened) data* under all P in both $\mathcal{H}_0$ and $\mathcal{H}_1$. By this we mean that there is a coarsening $\{V_i\}$ of the $\{Y_i\}$ so that both the null and the alternative are simple for data coarsened to $\{V_i\}$, as in Proposition 3, so that for each n, all distributions in $P_0 \in \mathcal{H}_0$ induce the same distribution $Q_0[V_n]$ on V^n with density q'_0, and all distributions $P_1 \in \mathcal{H}_1$ induce the same distribution $Q_1[V^n]$ on V^n with density q'_1, and $S_n = q'_1(V^n)/q'_0(V^n)$. The setting can be extended to the case where $\mathcal{H}_0$ contains additional distributions in $\mathcal{H}_0$ and $\mathcal{H}_1$, as long as for all $P_0 \in \mathcal{H}_0$, $Q_0[S_n]$, the marginal distribution of S_n under $Q_0[V_n]$, stochastically dominates $P_0[V_n]$, and under all $P_1 \in \mathcal{H}_1$, $Q_1[S_n^{-1}]$, the marginal distribution of $1/S_n$ under $Q_1[V_n]$, stochastically dominates $P_1[V_n]$.

For such likelihood ratio processes, $S_1, S_2, \ldots$ has the property of being a test martingale under both $\mathcal{H}_0$ and (after inversion) under $\mathcal{H}_1$. The sequential test based on $S_1, S_2, \ldots$ with prespecified parameters α, β proceeds by calculating $S_1, S_2, \ldots$ and stopping at τ^*, the smallest τ at which either $S_\tau \geq (1-\beta)/\alpha$ ('accept') or $S_\tau \leq (1-\alpha)/\beta$ ('reject'). Wald showed that this test has Type I error probability bounded by α and Type II error bounded by β. The reason one can stop at a smaller threshold ($(1-\beta)/\alpha$ rather than $1/\alpha$) is that one *has* to stop at τ^*, Thus, the method does not allow for optional stopping in our sense: the probability that there is *some* n with $S_n \geq (1-\beta)/\alpha$ is strictly larger than α.

Still, since $S_1, S_2, \ldots$ forms a test martingale under $\mathcal{H}_0$, it can be used to generate useful E-values as explained in Section 5.2.1. Thus, much of the work in sequential testing can be re-cycled to obtain test martingales and E-values. Of course, as discussed in that section, not all useful (δ-GROW) E-variables derive from martingales, let alone from 'two-sided' martingales.

Conditional Frequentist Tests In a series of papers starting with the landmark (Berger, Brown and Wolpert, 1994), Berger, Brown, Wolpert (BBW) and collaborators, extending initial ideas by Kiefer, 1977 develop a theory of frequentist conditional testing that "in spirit" is very similar to ours (see also Wolpert, 1996; Berger, 2003) — one can view the present paper as a radicalization of the BBW stance. Yet in practice there are important differences. For example, our link between posteriors and Type I error is slightly different (Bayes factors, i.e. posterior *ratios* vs. posterior *probabilities*), in our approach there are no 'no-decision regions', in the BBW papers there is no direct link to optional continuation.

Related Work on Relating P-values and E-variables Shafer and Vovk, 2019 give a general formula for *calibrators* f. These are decreasing functions $f : [0,1] \to [0,\infty]$ so that for any P-value P, $E := 1/f(P)$ is an E-variable. Let $f_{\mathrm{vs}}(P) := -eP\log P$, a quantity sometimes called the *Vovk-Sellke bound* (Bayarri et al., 2016)), having roots in earlier work by by Vovk, 1993 and Sellke et al. (Sellke, Bayarri and Berger, 2001). All calibrators satisfy $\lim_{P\downarrow 0} f(P)/f_{\mathrm{vs}}(P) = \infty$, and calibrators f advocated in practice additionally satisfy, for all $P \leq 1/e$, $f(P) \geq f_{\mathrm{vs}}(P)$. For example, for any calibrator f suggested for practice, rejection under the safe test with significance level $\alpha = 0.05$, so that $E \geq 20$, would then correspond to reject only if $P \leq f^{-1}(0.05) > f_{\mathrm{vs}}^{-1}(0.05) \approx 0.0032$, requiring a substantial amount of additional data for rejection under a given alternative. Note that the E-variables we developed for *given* models in previous sections are more sensitive than such generic calibrators though. For example, in Section 5.1.3 the threshold $2.72/\sqrt{n}$ corresponding to $\alpha = 0.05$ corresponds roughly to $p = 0.007$, a factor

2 larger. Experiments in the master's study (Hu, 2020) indicate a similar phenomenon for nonparametric tests: GROW E-values designed specifically for a given $\mathcal{H}_0$ and $\mathcal{H}_1$ achieve higher growth rate and higher power than calibration E-values based on P-values for these $\mathcal{H}_0$ and $\mathcal{H}_1$.

Related Work: Testing based on Data-Compression and MDL

Example 5.4. Ryabko and Monarev, 2005 show that bit strings produced by standard random number generators can be substantially compressed by standard lossless data compression algorithms such as `zip`, which is a clear indication that the bits are not so random after all. Thus, the null hypothesis states that data are 'random' (independent fair coin flips). They measure 'amount of evidence against $\mathcal{H}_0$ provided by data $\mathbf{y} = y_1, \ldots, y_n$' as

$$n - L_{\texttt{zip}}(\mathbf{y}),$$

where $L_{\texttt{zip}}(\mathbf{y})$ is the number of bits needed to code $\mathbf{y}$ using (say) `zip`. Now, define $\overline{p}_1(\mathbf{y}) = 2^{-L_{\texttt{zip}}(\mathbf{y})}$. Via Kraft's inequality (Cover and Thomas, 1991) one can infer that $\sum_{\mathbf{y}\in\{0,1\}^n} \overline{p}_1(\mathbf{y}) \leq 1$ (for this particular case, see the extended discussion by Grünwald, 2007, Chapter 17). At the same time, for the null we have $\mathcal{H}_0 = \{P_0\}$, where P_0 has mass function p_0 with for each n, $\mathbf{y} \in \{0,1\}$, $p_0(\mathbf{y}) = 2^{-n}$. Defining $E := \overline{p}_1(\mathbf{Y})/p_0(\mathbf{Y})$ we thus find

$$\mathbf{E}_{\mathbf{Y}\sim P_0}[E] = \sum_{\mathbf{y}\in\{0,1\}^n} \overline{p}_1(\mathbf{y}) \leq 1 \;\; ; \;\; \log E = n - L_{\texttt{zip}}(\mathbf{Y}).$$

Thus, the Ryabko-Monarov codelength difference is the logarithm of an E-variable. Note that in this example, there is no clearly defined alternative; being able to compress by `zip` simply means that the null hypothesis is false; it certainly does not mean that the 'sub-distribution' $\overline{p}_1$ is true (if one insists on there being an alternative, one could view $\overline{p}_1$ as a representative of a nonparametric $\mathcal{H}_1$ consisting of *all* distributions P_1 with $\mathbf{E}_{\mathbf{Y}\sim P_1}[\log E] > 0$, a truly huge and not so intuitive set).

More generally, by the same reasoning, for singleton $\mathcal{H}_0 = \{P_0\}$, any test statistic of the form $\overline{p}_1(\mathbf{Y})/p_0(\mathbf{Y})$, with p_0 the density of P_0 and $\overline{p}_1$ a density or sub-density (integrating to less than 1) is an E-variable. Such E-variables have been considered extensively within the *Minimum Description Length (MDL)* and *prequential* approaches to model selection (Rissanen, 1989; Dawid, 1997; Barron, Rissanen and Yu, 1998; Grünwald and Roos, 2020). In these approaches there usually is a clearly defined alternative $\mathcal{H}_1$, so that a Bayesian would choose $\overline{p}_1 := p_{W_1}$ to be a Bayes marginal density. In contrast, the MDL and prequential approach allow more freedom in the choice of $\overline{p}_1$. MDL merely requires $\overline{p}_1$ to be a 'universal distribution' such as a Bayes marginal, a normalized maximum likelihood, prequential plug-in or a 'switch' distribution (Grünwald, 2007). With simple $\mathcal{H}_0$, all such 'MDL factors' also constitute E-variables; but with composite $\mathcal{H}_0$, just as with Bayes factors, the standard MDL approach may fail to deliver E-variables.

Future Work, Open Problems In Section 5.3.3 we indicated that standard δ-GROW E-variables often turn out to be 'simple' (and therefore easy to implement): they are defined to be GROW relative to a large set, but they end up as Bayes factors $p_{W_1^*}/p_{W_0^*}$ in which W_1^* puts all mass

on the boundary of Θ_1. We aim to investigate the generality of this phenomenon in future work.

We already indicated that it may be possible to extend Theorem 5.6 to show that the Bayes factor based on the right Haar prior can be GROW in more general group invariant settings; showing or disproving this is a major goal for future work. Also, just as we propose to fully base testing on a method that has a sequential gambling/investment interpretation, Shafer and Vovk have suggested, even more ambitiously, to base the whole edifice of probability theory on sequential-gambling based game theory rather than measure theory (Shafer and Vovk, 2001; Shafer and Vovk, 2019); see also (Shafer, 2019) who emphasizes the ease of the betting interpretation. Obviously our work is related, and it would be of interest to understand the connections more precisely.

5.7 A Theory of Hypothesis Testing

5.7.1 A Common Currency for Testers adhering Jeffreys', Neyman's and Fisher's Testing Philosophies

The three main approaches towards null hypothesis testing are Jeffreys' Bayes factor methods, Fisher's P-value-based testing and the Neyman-Pearson method. Berger, 2003, based on earlier work, e.g. (Berger, Brown and Wolpert, 1994), was the first to note that, while these three methodologies seem superficially highly contradictory, there exist methods that have a place within all three. Our proposal is in the same spirit, yet more radical; it also differs in many technical respects from Berger's. Let us briefly summarize how E-variables and the corresponding safe tests can be fit within the three paradigms:

Concerning the *Neyman-Pearson approach*: E-variables lead to tests with Type-I error guarantees at any fixed significance level α, which is the first requirement of a Neyman-Pearson test. The second requirement is to use the test that maximizes power. But we can use GROW E-variables designed to do exactly this, as we illustrated in Section 5.5. The one difference to the NP approach is that we optimize power under the constraint that the E-variable is GROW — which is *essential* to make the results of various tests of the same null easily combinable, and preserve Type I error probabilities under optional stopping. Note though that this constraint is major: as shown in Example 5.1, the standard NP tests lead to useless E-variables under the GROW criterion.

Concerning the *Fisherian approach*: we have seen that E-variables can be reinterpreted as (quite) conservative P-values. But much more importantly within this discussion, E-variables can be defined, and have a meaningful (monetary) interpretation, *even if no clear (or only a highly nonparametric/nonstationary) alternative can be defined*. This was illustrated in the data compression setting of Example 5.4. Thus, in spirit of Fisher's philosophy, we can use E-variables to determine whether there is substantial evidence against $\mathcal{H}_0$, without predetermining any alternative: we simply postulate that the larger E, the more evidence against $\mathcal{H}_0$ without having specific frequentist error guarantees. The major difference though is that these E-variables continue to have a clear (monetary) interpretation even if we multiply them over different tests,

and even if the decision whether or not to perform a test (gather additional data) depends on the past.

Concerning the *Bayesian approach*: despite their monetary interpretation, *all* E-variables that we encountered can also be written as likelihood ratios, although (e.g. in Example 5.4 or Section 5.4.5) either $\mathcal{H}_0$ or $\mathcal{H}_1$ may be represented by a distribution that is different from a Bayes marginal distribution. Still, all GROW (optimal) E-variables we encountered are in fact equivalent to Bayes factors, and Theorem 5.4 Part 3 strongly suggests that this is a very general phenomenon. While the point priors arising in the δ-GROW E-variables may be quite different from priors commonly adopted in the Bayesian literature, one can also obtain E-variables by using priors on $\mathcal{H}_1$ that do reflect prior knowledge or beliefs — we elaborate on this under *Hope vs. Belief* below.

The Dream With the massive criticisms of P-values in recent years, there seems a consensus that P-values should be used not at all or, at best, with utter care (Wasserstein, Lazar et al., 2016; Benjamin et al., 2018), but otherwise, the disputes among adherents of the three schools continue — intuitions among great scientists still vary dramatically. For example, some highly accomplished statisticians reject the idea of testing without a clear alternative outright; others say that such goodness-of-fit tests are an essential part of data analysis. Some insist that significance testing should be abolished altogether (Amrhein, Greenland and McShane, 2019), others (perhaps slightly cynically) acknowledge that significance may be silly in principle, yet insist that journals and conferences will always require a significance-style 'bar' in practice and thus such bars should be made as meaningful as possible. Finally, within the Bayesian community, the Bayes factor is sometimes presented as a panacea for most testing ills, while others warn against its use, pointing out for example that with different default priors that have been proposed, one can get quite different answers.

Wouldn't it be nice if all these accomplished but disagreeing people could continue to go their way, yet would have a common language or 'currency' to express amounts of evidence, and would be able to combine their results in a meaningful way? This is what E-variables can provide: consider three tests with the same null hypothesis $\mathcal{H}_0$, based on samples $\mathbf{Y}_{(1)}$, $\mathbf{Y}_{(2)}$ and $\mathbf{Y}_{(3)}$ respectively. The results of a δ-based E-variable test aimed to optimize power on sample $\mathbf{Y}_{(1)}$, an E-variable test for sample $\mathbf{Y}_{(2)}$ based on a Bayesian prior W_1 on $\mathcal{H}_1$ and a Fisherian E-variable test in which the alternative $\mathcal{H}_1$ is not explicitly formulated, can all be multiplied — and the result will be meaningful.

Hope vs. Belief In a purely Bayesian set-up, optional stopping is justified if θ viewed as a random variable is independent of the stopping time N under the prior W. In that case, a celebrated result going back to Barnard, 1947 (see Hendriksen, De Heide and Grünwald, 2020 for an overview) says that the posterior does not depend on the stopping rule used; hence it does not matter *how* N was determined (as long as it does not depend on future data). If Bayes factors are 'local', based on priors that depend on the design and thus on the sample size n, then, from a purely Bayesian perspective, optional (early) stopping is not allowed: since the prior depends on n, when stopping at the first $T < n$ at which $p_{W_1}(y^T)/p_{W_0}(y^T) > 20$, neither the original prior based on the fixed n nor the prior based on the observed T (which treats the

random T as fixed in advance) is correct any more. This happens, for example, for the default (Gunel and Dickey, 1974) Bayes factors for 2×2 contingency tables advocated by Jamil et al., 2016 — from a Bayesian perspective, these do not allow for optional stopping.

The same holds for the UMP Bayes factors that we considered in Section 5.4.1. These generally are 'local', the prior W_1 (and, presuming the idea can be extended to composite $\mathcal{H}_0$, potentially also W_0) depending on the sample size n. For example, for the 1-sided test with the normal location family, Example 5.2, we set all prior mass on $\widetilde{\mu}_n = \sqrt{2(-\log \alpha)/n}$; a similar dependence holds for the prior on δ^* in the δ^*-based GROW t-test if we choose δ^* to maximize power. Thus, while from a purely Bayesian perspective such E-variables/Bayes factors are not suitable for optional stopping, in Section 5.4, both the δ-based GROW E-variable for the normal location family and for the t-test setting do allow for optional stopping under *our* definition: one may also stop and report the Bayes factor at any time one likes *during* the experiment, and still Type I error probabilities are preserved (Hendriksen, De Heide and Grünwald, 2020). This is what we did in the experiment of Figure 5.5: the pre-determined n (called there $n_{\max}$) on which the prior W_1 on δ (that puts mass 1/2 on δ^*, and 1/2 on $-\delta^*$) is based is determined there such that, if we stop at any fixed $T = n'$, the statistical power of the test is *optimal* if $n' = n_{\max}$; but the likelihood ratio $e(Y^T) := p_{W_1}(Y^T)/p_{W_0}(Y^T)$ remains an E-variable even if $T = n' \neq n_{\max}$ or even if one stops at the first $T \leq n_{\max}$ such that $E(Y^T) \geq 20$. Thus, we should make a distinction between prior *beliefs* as they arise in Bayesian approaches, and what one may call 'prior *hope*' as it arises in the E-variable approach. The purely Bayesian approach relies on the *beliefs* being, in some sense, adequate. In the E-variable based approach, one *can* use priors that represent subjective a priori assessments; for example, in the Bayesian t-test, one can use any prior W_1 on δ one likes as long as it has more than two moments, and still the resulting Bayes factor with the right Haar prior on σ will be a GROW E-variable (Theorem 5.6). *If $\mathcal{H}_1$ is the case, and the data behave as one would expect according to the prior W_1*, then the E-variable will tend to be large – it GROWs fast. But if the data come from a distribution in $\mathcal{H}_1$ in a region that is very unlikely under W_1, $E(\mathbf{X})$ will tend to be smaller — but it is still an E-variable, hence leads to valid Type-I error guarantees and can be interpreted when multiplied across experiments. Thus, from the E-variable perspective, the prior on W_1 represents something more like 'hope' than 'belief' — if one is *lucky* and data behave like W_1 suggests, one gets better results; but one still gets *valid* and *safe* results even if W_1 is chosen badly (corresponds to false beliefs).

This makes the E-variable approach part of what is perhaps among the most under-recognized paradigms in statistics and machine learning: methods supplying results that have frequentist validity *under a broad range of conditions (in our case: as long as $\mathcal{H}_0$ or $\mathcal{H}_1$ is correct), but that can give much* stronger *results if one is 'lucky' on the data at hand (e.g. the data matches the prior)*. It is, for example, the basis of the so-called *PAC-Bayesian approach* to classification in machine learning (McAllester, 1998; Grünwald and Mehta, 2019), which itself, via Shawe-Taylor and Williamson, 1997, can be traced back to be inspired by the conditional testing approach of Kiefer, 1977 that also inspired the BBW approach to testing. It also connects to the general idea of 'safe' inference (Grünwald, 2000; Grünwald, 2018).

5.7.2 Possible Objections

By the nature of the subject, the relevance of this work is bound to be criticized. We would like to end this paper by briefly anticipating three potential criticisms.

Where does all this leave the poor practitioner? A natural question is, whether the E-variable based approach is not much too difficult and mathematical. Although the present, initial paper is quite technical, we feel the approach in general is in fact easier to understand than any approach based on P-values. The difficulty is that one has to explain it to researchers who have grown up with P-values — we are confident that, to researchers who neither know P-values nor E-variables, the E-variables are easier to explain, via the direct analogy to gambling. Also, we suggested δ-based 'default' E-variables that (unlike some default Bayes factors) can be used in absence of strong prior knowledge about the problem yet still have a valid monetary interpretation and valid Type I Error guarantees. Finally, if, as suggested above, practitioners really were to be forced, when starting an analysis, to think about optional stopping, optional continuation and misspecification — this would make life difficult, but would make practice all the better.

No Binary Decisions, Part I: Removing Significance There is a growing number of influential researchers who hold that the whole concept of 'significance', and ensuing binary 'reject' or 'accept' decisions, should be abandoned altogether (see e.g. the 800 co-signatories of the recent Amrhein, Greenland and McShane, 2019, or the call to abandon significance by McShane et al., 2019). This paper is not the place to take sides in this debate, but we should stress that, although we strongly emphasized Type-I and Type-II error probability bounds here, E-variables still have a meaningful interpretation, as amount of evidence measured in monetary terms, even if one never uses them to make binary decisions; and we stress that, again, this monetary interpretation remains valid under optional continuation, also in the absence of binary decisions. We should also stress here that we do not necessarily want to adopt 'uniformly most powerful E-variables, even though our comparison to Johnson's uniformly most powerful Bayes tests in Section 5.4 and the experiments in Section 5.5 might perhaps suggest this. Rather, our goal is to advocate using GROW E-variables relative to some prior W on Θ_1 or a subset of $\Theta(\delta)$ of Θ_1 — the GROW criterion leaves open some details, and our point in these experiments is merely to compare our approach to classical, power-optimizing Neyman-Pearson approaches — to obtain the sharpest comparison, we decided to fill in the details (the prior W on $\Theta(\delta)$) for which the two approaches (E-variables vs. classical testing) behave most similarly.

No Binary Decisions, Part II: Towards Safe Confidence Intervals Another group of researchers (e.g. Cumming, 2012) has been advocating for generally replacing testing by estimation accompanied by confidence intervals; or, more generally (McShane et al., 2019), that researchers should always provide an analysis of the behavior of and uncertainty inherent in one or more estimators for the given data. While we sympathize with the latter point of view, we stress that standard confidence intervals (as well as other measures of uncertainty of estimators such as standard errors) suffer from a similar problem as P-values: *they are not safe under optional continuation*. The aforementined anytime-valid confidence sequences developed by Lai and later Ramdas and collaborators (Lai, 1976; Howard et al., 2018b; Howard et al., 2018a)

do allow for optional stopping and hence, if subsequent experimenters keep using the same underlying test martingales, optional continuation. We strongly feel that if one really wants to replace testing by confidence approaches, one should adopt anytime-valid rather than standard confidence intervals, even though the former ones are invariably a bit broader. In future work we hope to study whether it is useful to consider 'safe confidence intervals', merely allowing for optional continuation rather than optional stopping (at each data point).

5.A Proof Preliminaries

In the next sections we prove our theorems. To make all statements in the main text mathematically rigorous and their notations mutually compatible, we first provide a few additional definitions and notation.

Sample Spaces and σ-Algebras In all mathematical results and examples in the main text, we tacitly make the following assumptions: all random elements mentioned in the main text are defined on some measurable space $(\Omega, \mathcal{A})$. We assume that $\{Y_i\}_{i\in I}$ and $\{R_i\}_{i\in I}$ are two collections of measurable functions from Ω to measurable spaces $(\mathcal{Y}, \mathcal{A}')$ and $(\mathcal{R}, \mathcal{A}'')$ respectively, where either $I = \{1, 2, \ldots, n_{\max}\}$ for some finite $n_{\max}$ or $I = \mathbb{N}$. We additionally assume that each Y_i takes values in $\mathcal{Y} \subseteq \mathbb{R}^m$ for some finite m, and we equip $(\Omega, \mathcal{A})$ with the filtration $(\mathcal{F}_i)_{i\in I}$ where $\mathcal{F}_i$ is the σ-algebra generated by (Y^i, R^i).

For each $\theta \in \Theta := \Theta_0 \cup \Theta_1$, in the unconditional case, P_θ is a distribution for the random process $(Y_i)_{i\in I}$. In the conditional case, we assume finite I and existence of a fixed function ϕ and another collection of functions $\{X_i\}_{i\in I}$ such that for all $i \in I$, $X_i = \phi(R_i)$, with X_i taking values in some set $\mathcal{X}$. For each $x^n \in \mathcal{X}^n$, $P_\theta(\cdot \mid X^n = x^n)$ is then a distribution on $(Y_1, \ldots, Y_{n_{\max}})$. We assume throughout that $P_\theta(Y^n \mid X^n = x^n) = P_\theta(Y^n \mid X^m = x^m)$ for every $n, m > n, x^m \in \mathcal{X}^m$: present data is independent of future covariates given present covariates. Whenever we refer to a random variable such as $\mathbf{Y}$ without giving an index, it stands for $Y^n = (Y_1, \ldots, Y_n)$; similarly for all other time-indexed random variables.

We stated in the main text that we assume that the parameterization is 1-to-1. By this we mean that for each $\theta, \theta' \in \Theta$ with $\theta \neq \theta'$, the associated distributions are also different, so that $P_\theta \neq P_{\theta'}$. We also assume that Θ_0 and Θ_1 are themselves associated with appropriate σ-algebras. In general, Θ_j need not be finite-dimensional, so we allow non-parametric settings.

(In)-Dependence and Densities In Section 5.2 on optional continuation we make no further assumptions about P_θ. Specifically, the Y_i need not be independent. In all other sections, unless we explicitly state otherwise, we assume independence. Specifically, when the P_θ represent unconditional distributions, then we assume that the random variables $Y_1, Y_2, \ldots$ are independent under each P_θ with $\theta \in \Theta$, and that for all i, the marginal distribution $P_\theta(Y_i)$ has a density relative to some underlying measure λ_1. That is, we for each j we can write $p_\theta(Y^j) = p_\theta(Y_1, \ldots, Y_j) = \prod_{i=1}^j p'_{\theta,i}(Y_i)$ as a product density where $p'_{\theta,i}$ is a density relative to λ_1. In all our examples, λ_1 is either a probability mass function on $\mathcal{Y}$ or a density on $\mathcal{Y}$ relative to Lebesgue measure, but the theorems work for general λ_1. Then $p_\theta(\mathbf{Y}) = \prod_{i=1}^n p'_{\theta,i}(Y_i)$ is a density relative to $\lambda := \lambda_n$, defined as the n-fold product measure of λ_1.

With the exception of the contingency table setting of Section 5.4.4 and the conditional exponential family setting that we briefly mentioned in Section 5.4.5 (the only sections in which the $+P_\theta$ are conditional (on $\mathbf{x}$) distributions), we assume that the Y_i are not just independent but also identically distributed, hence $p'_{\theta,i} = p'_{\theta,1}$ for all i.

Notational Conventions When we mention a distribution P_θ without further qualification, we mean that it is the distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n) = Y^n$ defined on Ω; and we use p_θ for its

density as defined above. We sometimes refer to the marginal distribution of a random variable $\mathbf{U}$ under P_θ, where $\mathbf{U}$ is a function (coarsening) of $\mathbf{Y}$. We denote this distribution as $P_\theta[\mathbf{U}]$, and its density by $p'_\theta(u_1, \ldots, u_n)$, avoiding the cumbersome $p_\theta[\mathbf{U}](u_1, \ldots, u_n)$.

We generically use $E^*_{\ldots}$ to denote E-variables that are GROW relative to some prior, set, or set of priors, e.g. $E^*_{W_1}, E^*_{\Theta(\Theta)}, E^*_{\mathcal{W}_1}$, and so on. If we consider E-variables that can be written as a function of a coarsened random variable $\mathbf{V} = f(\mathbf{Y})$, and that are also GROW on the 'coarsened' level of distributions on $\mathbf{V}$ rather than $\mathbf{Y}$, then we write $E^*_{\ldots}\langle\mathbf{V}\rangle$. Thus, standard GROW E-variables could equivalently be written as $E^*_{\ldots}\langle\mathbf{Y}\rangle$.

5.B Optional Continuation with Side-Information

Proof of Proposition 2 Although Proposition 2 is easily proved using Doob's optional stopping theorem, it may be useful to give a direct proof:

Proof. **(sketch)** We first consider the case with $K_{\text{STOP}} = k_{\max}$. Under all P_θ, we have

$$\begin{aligned}
\mathbf{E}\left[E^{(k)}\right] &= \mathbf{E}\left[e_{h(V^0)|\tau_{(0)},g(V^0)}\left(\mathbf{V}^{(1)}\right)\cdot\ldots\cdot e_{h(\mathbf{V}^{(k-1)})|\tau_{(k-1)},g(\mathbf{V}^{(k-1)})}\left(\mathbf{V}^{(k)}\right)\right] \\
&= \mathbf{E}_{\mathbf{V}_{(1)}\sim P_\theta}\mathbf{E}_{\mathbf{V}_{(2)}\sim P_\theta|\mathbf{V}^{(1)}}\ldots\mathbf{E}_{\mathbf{V}_{(k)}\sim P_\theta|\mathbf{V}^{(k-1)}}\left[e_{h(V^0)|\tau_{(0)},g(V^0)}\left(\mathbf{V}^{(1)}\right)\cdot\right. \\
&\quad \left. e_{h(\mathbf{V}^{(1)})|\tau_{(1)},g(\mathbf{V}^{(1)})}\left(\mathbf{V}^{(2)}\right)\cdot\ldots\cdot e_{h(\mathbf{V}^{(k-1)})|\tau_{(k-1)},g(\mathbf{V}^{(k-1)})}\left(\mathbf{V}^{(k)}\right)\right] \\
&= \mathbf{E}_{\mathbf{V}_{(1)}\sim P_\theta}\left[e_{h(V^0)|\tau_{(0)},g(V^0)}\left(\mathbf{V}^{(1)}\right)\cdot\mathbf{E}_{\mathbf{V}_{(2)}\sim P_\theta|\mathbf{V}^{(1)}}\left[e_{h(\mathbf{V}^{(1)})|\tau_{(1)},g(\mathbf{V}^{(1)})}\left(\mathbf{V}^{(2)}\right)\cdot\right.\right. \\
&\quad \left.\left.\ldots\cdot\mathbf{E}_{\mathbf{V}_{(k)}\sim P_\theta|\mathbf{V}^{(k-1)}}\left[e_{h(\mathbf{V}^{(k-1)})|\tau_{(k-1)},g(\mathbf{V}^{(k-1)})}\left(\mathbf{V}^{(k)}\right)\right]\ldots\right]\right].
\end{aligned}$$

By definition of E-variables, all factors in the product are bounded by 1, and the result follows. For general $K_{\text{STOP}} \leq k_{\max}$, note that without loss of generality we may assume that $\mathcal{W}$ contains the parameter $\mathbf{1}$, where for all n, m, $e_{n|m,\mathbf{1}}$ is the *trivial* E-variable $e_{n|m,\mathbf{1}}(v^{n+m}) \equiv 1$ for all $v^{n+m} \in \mathcal{V}^{n+m}$. For any sequence $v_1, v_2 \ldots$ we modify g, h to g', h' recursively as follows: we let $h'(\mathbf{v}^{(1)}) := h(\mathbf{v}^{(1)})$, $h'(\mathbf{v}^{(2)}) = h(\mathbf{v}^{(2)}), \ldots$, similarly for g' and g, until we reach the smallest k such that $g(\mathbf{v}^{(k)}) = \text{STOP}$. Then we set $g'(v^n) = g'(v_1, \ldots, v_n) = \mathbf{1}$ and $h'(y^n) = 1$ for all $n \geq \tau_{(k)}$ and all v^n that are extensions of $v^{\tau_{(k)}}$. The E' based on the new g', h' will have $E'^{(k_{\max})} = E^{(K)}$. It follows from (a) that $E'^{(k_{\max})}$ is an E-variable, so the result follows. □

Extending Proposition 2 We want to extend the proposition to allow for two possibilities, First, the sample size for the j-th batch of data may be determined by a *stopping time* $N_{(j)}$, which generalizes the $N_{(j)}$ used in the main text to the case that the sample size of the j-th sample $\mathbf{Y}_{(j)}$ is not fixed in advance. For example, in the 2×2 table (Example 5.4.4) we might continue sampling until we have obtained 10 new examples of category a. Second, we want to model the idea of 'side information'. For this, we assume we make additional observations $Z_{(0)}, Z_{(1)}, Z_{(2)}, \ldots$. The idea is that at the end of analyzing the k-th data batch $\mathbf{Y}_{(k)}$, we also get some side information $Z_{(k)}$ which may influence our decision whether or not to take into account a new data batch $\mathbf{Y}_{(k+1)}$. We want to make as few assumptions as possible about this side-information; specifically, we will not assume that is itself of stochastic nature (i.e. will

assume no distribution on it), and the $Z_{(k)}$ may take values in an unspecified countable set $\mathcal{Z}_{(k)}$. Thus, whereas the data $\mathbf{Y}_{(k)}$ can always be viewed as a vector $(Y_{\tau_{(k-1)}+1}, \ldots, Y_{\tau_{(k)}})$, we do not assume that $Z_{(k)}$ has such (or any other) sub-structure. To make this compatible with the measure-theoretic setting of the previous section, we assume that all $Z_{(j)}$ are random variables on $(\Omega, \mathcal{A})$. Whereas before, the filtration $(\mathcal{F}_i)_{i\in I}$ was defined by setting $\mathcal{F}_i$ to be the σ-algebra generated by (Y^i, R^i), we now set $\mathcal{F}_i$ to be the σ-algebra generated by $(Y^i, R^i, Z_{(J_i)})$ where J_i is the largest $J \geq 0$ such that $\tau_{(J)} \leq i$, where $\tau_{(J)}$ is defined as below. Since $\tau_{(0)} = 0$, J_i is a measurable function. It represents 'which batch sample size i is part of'. For example, if the first batch has sample size $N_{(1)} = 5$ and the second $N_{(2)} = 10$, then, for $1 \leq i \leq 5$, before observing Y_i, the available information is $Y^{i-1}, R^{i-1}, Z_{(0)}$. Then, for $6 \leq i \leq 10$, we are 'in the second batch', and the available information is $Y^{i-1}, R^{i-1}, Z^{(1)}$. Afterwards, $Z_{(2)}$ becomes available, and so on
.

As formalized in (5.35) below, we will assume that past outcomes may influence the value of $Z_{(k)}$, but $Z_{(k)}$ should be independent of any future $\mathbf{Y}_{(k+j)}$. Our optional continuation result continues to hold *irrespective* of the actual definition of $Z_{(k)}$ and $\mathcal{Z}_{(k)}$, as long as these independences hold. Thus, we may think of $Z_{(k)}$ as encoding information that is difficult to think of stochastically, such as 'more money to perform future tests is available'. Still, the confinements of classical probability theory (or rather the measure theory on which it is based) force us to assume the existence of sets of possible outcomes $\mathcal{Z}_{(k)}$, even if we do not need to specify them. It seems that even this can be avoided by re-expressing the optional continuation result in terms of the *open* protocols enabled by the Game-Theoretic Theory of Probability due to Shafer and Vovk, 2019; but that would really go beyond the scope of this paper.

Batch Stopping Times To further incorporate $Z_{(k)}$ into our framework together with sample sizes $N_{(j)}$ that are not fixed in advance, we need a slight generalization of the idea of stopping time and stopping rule. In our context, a *stopping rule for the k-th batch with start time t* is a collection of functions $f_{(k),t,i}$, $i \in \mathbb{N}$, where $f_{(k),t,i}$ maps $(Z_{(k-1)}, X^{t+i}, V^{t+i})$ to $\{\text{STOP}, \text{CONTINUE}\}$ such that for every $z \in \mathcal{Z}_{(k-1)}$, every sequence $(x_1, v_1), (x_2, v_2), \ldots$, there is an $i > t$ such that

$$f_{(k),t,i}(z, ((x_1, v_1), \ldots, (x_{t+i}, v_{t+i})) = \text{STOP}.$$

Thus, we require stopping times that are finite on all sample paths rather than the more usual 'almost surely finite' stopping times because the X_i and $Z_{(k)}$ do not have a distribution associated with them.

We now define $\tau_{(k)}$ as the *stopping time for the k-th batch* in terms of stopping rules $f_{(k)}$ defined above. We set $\tau_{(1)} := N_{(1)}$ to be the smallest i such that $f_{(1),0,i}(Z(0), X^i, V^i) = \text{STOP}$, and more generally, we set $\tau_{(k)}$ to be $\tau_{(k-1)} + N_{(k)}$, where $N_{(k)}$ is the smallest i such that

$$f_{(k),\tau_{(k-1)},i}(Z^{(k-1)}, X^{\tau_{(k-1)}+i}, V^{\tau_{(k-1)}+i}) = \text{STOP}.$$

To make all required probabilities and expectations well-defined we set, for all $i \geq 1$,

$$P_\theta(Y_{\tau_{(j)}+1}, \ldots, Y_{\tau_{(j)}+i} \mid Z^{(j)}, \mathbf{Y}^{(j)}, X^{\tau_{(j)}+i}) := P_\theta(Y_{\tau_{(j)}+1}, \ldots, Y_{\tau_{(j)}+i} \mid \mathbf{Y}^{(j)}, X^{\tau_{(j)}+i}). \quad (5.35)$$

That is, according to all distributions P_θ under consideration, the 'side-information' $Z^{(j)}$ available after the j-th data batch cannot influence future outcomes $Y_{\tau_{(j)}+i}$; on the other hand,

the formulation allows that all data obtained up to and including $\mathbf{Y}^{(j)}$ may influence the side-information $Z_{(j)}$.

The definition below evidently generalizes (5.10), and the proposition evidently generalizes Proposition 2:

Definition 5.3 (Conditional E-Variables)**.** Let X_i, Y_i, V_i and $\tau_{(1)}, \ldots, \tau_{(k)}$ with $1 \le k \le k_{\max}$ be as above. Let $E_{(k)}$ be a nonnegative random variable that can be written as a function of $(X^{(k)}, V^{(k)})$. We call $E_{(k)}$ *an E-variable for* $V_{(k)}$ *conditional on* $\mathbf{X}^{(k)}, \mathbf{V}^{(k-1)}$ if it satisfies, for all $P \in \mathcal{H}_0$,

$$\mathbf{E}_P[E_{(k)} \mid \mathbf{X}^{(k)}, \mathbf{V}^{(k-1)}] \le 1. \tag{5.36}$$

Proposition 7. [Optional Continuation with Side-Information] *Let $\tau_{(1)}, \ldots, \tau_{(k)}$ with $k \le k_{\max}$ and τ^* be generalized stopping times as above such that on all sample paths, τ^* coincides with $\tau_{(j)}$ for some $j = 1..k$. Let $E_{(1)}, E_{(2)}, \ldots, E_{(k)}$ be a sequence of random variables such that for each $j = 1..k$, $E_{(j)}$ is an E-variable for $\mathbf{V}_{(j)}$ conditional on $\mathbf{X}^{(j)}, \mathbf{V}^{(j-1)}$. Let the random variable K_{STOP} be such that $\tau^* = \tau_{(K_{\text{STOP}})}$. Then $E^{(K_{\text{STOP}})}$ is an E-variable, so that under all $P_0 \in \mathcal{H}_0$, for every $0 \le \alpha \le 1$, (5.11) of Proposition 2 and all its consequences hold.*

Proof. **(sketch)** By (5.35), $E_{(j)}$ being an E-variable conditional on $\mathbf{X}_{(j)}, \mathbf{V}^{(j-1)}$ implies that $E_{(j)}$ is also an E-variable conditional on $\mathbf{X}_{(j)}, \mathbf{V}^{(j-1)}, Z^{(j-1)}$. Then, since $E^{(j-1)}$ can be written as a function of $\mathbf{X}^{(j-1)}, \mathbf{V}^{(j-1)}, Z^{(j-1)}$, we have, under all $P \in \mathcal{H}_0$, for $j \ge 1$,

$$\begin{aligned}\mathbf{E}_P[E^{(j)} \mid \mathbf{X}^{(j)}, \mathbf{V}^{(j-1)}, Z^{(j-1)}] &= \mathbf{E}_P[E_{(j)} \cdot E^{(j-1)} \mid \mathbf{X}^{(j)}, \mathbf{V}^{(j-1)}, Z^{(j-1)}] \\ &= \mathbf{E}_P[E_{(j)} \mid \mathbf{X}^{(j)}, \mathbf{V}^{(j-1)}, Z^{(j-1)}] \cdot E_{(j-1)} \le E_{(j-1)},\end{aligned}$$

where the final step is just the definition of conditional E-variable. This shows that the process $E^{(1)}, E^{(2)}, \ldots$ constitutes a nonnegative supermartingale relative to the process $\mathbf{X}^{(1)}, \mathbf{V}^{(0)}, Z^{(0)}, \mathbf{X}^{(2)}, \mathbf{V}^{(1)}, Z^{(1)}, \ldots$. The result now follows by Doob's optional stopping theorem. □

5.C Elaborations and Proofs for Section 5.3

Meaning of "E^* as defined by achieving (5.14) is essentially unique" Consider $\Theta_1' \subset \Theta_1$ and Θ_0, as in the main text in Section 5.3. Suppose that there exists an E-variable E^* achieving the infimum in (5.14). We say that E^* is *essentially unique* if for any other E-variable E° achieving the infimum in (5.14), we have $P_\theta(E^* = E^\circ) = 1$, for all $\theta \in \Theta_1' \cup \Theta_0$. Thus, if the GROW E-variable exists and is essentially unique, any two GROW E-variables will take on the same value with probability 1 under all hypotheses considered, and then we can simply take one of these GROW E-variables and consider it the 'unique' one.

5.C.1 Proof of Theorem 5.4

For Part 1 of the result, we first need the following lemma. We call a measure Q on $\mathcal{Y}^m$ a *sub-probability distribution* if $0 < Q(\mathcal{Y}^m) \le 1$. Note that the KL divergence $D(P\|Q)$ remains

well-defined even if the measure Q is not a probability measure (e.g. Q could be a sub-probability distribution or might not be integrable), as long as P and Q both have a density relative to a common underlying measure (the definition of KL divergence does require the first argument P to be a probability measure though).

Lemma 8. *Let $\{Q_W : W \in \mathcal{W}_0\}$ be a set of probability measures where each Q_W has a density q_w relative to some fixed underlying measure λ. Let $\mathcal{Q}$ be any* convex *subset of these pdfs. Fix any pdf p (defined relative to measure λ) with corresponding probability measure P so that $\inf_{Q\in\mathcal{Q}} D(P\|Q) < \infty$ and so that all $Q \in \mathcal{Q}$ are absolutely continuous relative to P. Then:*

1. *There exists a unique sub-distribution Q° with density q° such that*

$$D(P\|Q^\circ) = \inf_{Q\in\mathcal{Q}} D(P\|Q), \tag{5.37}$$

   i.e. Q° is the Reverse Information Projection *of P on $\mathcal{Q}$.*

2. *For q° as above, for all $Q \in \mathcal{Q}$, we have*

$$\mathbf{E}_{\mathbf{Y}\sim Q}\left[\frac{p(\mathbf{Y})}{q^\circ(\mathbf{Y})}\right] \le 1. \tag{5.38}$$

   We note that we may have $Q^\circ \notin \mathcal{Q}$.

3. *Let Q_0 be a probability measure in $\mathcal{Q}$ with density q_0. Then: the infimum in (5.37) is achieved by $Q_0 \Leftrightarrow Q^\circ = Q_0 \Leftrightarrow$ (5.38) holds for $q^\circ = q_0$.*

Proof. The existence and uniqueness of a measure Q° (not necessarily a probability measure) with density q° that satisfies $D(P\|Q^\circ) = \inf_{Q\in\mathcal{Q}} D(P\|Q)$, and furthermore has the property

$$\text{for all } q \text{ that are densities of some } Q \in \mathcal{Q}\text{: } \mathbf{E}_{\mathbf{Y}\sim P}\left[\frac{q(\mathbf{Y})}{q^\circ(\mathbf{Y})}\right] \le 1. \tag{5.39}$$

follows directly from Li, 1999, Theorem 4.3. But by writing out the integral in the expectation explicitly we immediately see that we can rewrite (5.39) as:

$$\text{for all } Q \in \mathcal{Q}\text{: } \mathbf{E}_{\mathbf{Y}\sim Q}\left[\frac{p(\mathbf{Y})}{q^\circ(\mathbf{Y})}\right] \le 1.$$

Li's Theorem 4.3 still allows for the possibility that $\int q^\circ(y)\,\mathrm{d}\lambda(y) > 1$. To see that in fact this is impossible, i.e. q° defines a (sub-) probability density, use Lemma 4.5 of Li, 1999. This shows Part 1 and 2 of the lemma. The third part of the result follows directly from Lemma 4.1 of Li, 1999). (additional proofs of (extensions of) Li's results can be found in the refereed paper Grünwald and Mehta, 2019). □

We shall now prove Theorem 5.4 itself. Throughout the proof, λ stands for the n-fold product measure as defined in the introduction of this appendix, so that all distributions P_W with $W \in \mathcal{W}_1' \cup \mathcal{W}(\Theta_0)$ have a density p_W relative to λ, and whenever we speak of a 'density' we mean 'a density relative to λ'.

Proof of Theorem 5.4, Part 1 Let $\mathcal{W}_0 := \mathcal{W}(\Theta_0)$ and let $\mathcal{Q} = \{P_W : W \in \mathcal{W}(\Theta_0)\}$ and $P := P_{W_1}$. We see that $\mathcal{Q}$ is convex so we can apply Part 1 and 2 of the lemma above to P and $\mathcal{Q}$ and we find that $E^*_{W_1} := p_{W_1}(\mathbf{Y})/q^\circ(\mathbf{Y})$ is an E-variable, and that it satisfies

$$\mathbf{E}_{P_{W_1}}\left[\log E^*_{W_1}\right] = \mathbf{E}_{P_{W_1}}\left[\log \frac{p_{W_1}(\mathbf{Y})}{q^\circ(\mathbf{Y})}\right] = D\left(P_{W_1}\|Q^\circ\right) = \inf_{W_0 \in \mathcal{W}(\Theta_0)} D\left(P_{W_1}\|P_{W_0}\right),$$

where the second equality is immediate and the third is from (5.37). It only remains to show that (a)

$$\sup_{E \in \mathcal{E}(\Theta_0)} \mathbf{E}_{\mathbf{Y}\sim P_{W_1}}\left[\log E\right] \le \mathbf{E}_{P_{W_1}}\left[\log E^*_{W_1}\right]$$

and (b) that $E^*_{W_1}$ is essentially unique. To show (a), fix any E-variable $E = e(\mathbf{Y})$ in $\mathcal{E}(\Theta_0)$. Now further fix $\varepsilon > 0$ and fix a $W_{(\varepsilon)} \in \mathcal{W}(\Theta_0)$ with $D(P_{W_1}\|P_{W_{(\varepsilon)}}) \le \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) + \varepsilon$. We must have, with $q(y) := e(y)p_{W_{(\varepsilon)}}(y)$, that $\int q(y)\,\mathrm{d}\lambda = \mathbf{E}_{\mathbf{Y}\sim P_{W_{(\varepsilon)}}}[E] \le 1$, so q is a sub-probability density, and by the information inequality of information theory (Cover and Thomas, 1991), it follows:

$$\begin{aligned}
\mathbf{E}_{P_{W_1}}[\log E] &= \mathbf{E}_{P_{W_1}}\left[\log \frac{q(\mathbf{Y})}{p_{W_{(\varepsilon)}}(\mathbf{Y})}\right] \\
&\le \mathbf{E}_{P_{W_1}}\left[\log \frac{p_{W_1}(\mathbf{Y})}{p_{W_{(\varepsilon)}}(\mathbf{Y})}\right] \\
&= D(P_{W_1}\|P_{W_{(\varepsilon)}}) \\
&\le \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) + \varepsilon.
\end{aligned}$$

Since we can take ε to be arbitrarily close to 0, it follows that

$$\mathbf{E}_{P_{W_1}}[\log E] \le \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) = \mathbf{E}_{P_{W_1}}[\log E^*_{W_1}],$$

where the latter equality was shown earlier. This shows (a).

To show essential uniqueness, let E be any E-variable with $\mathbf{E}_{P_{W_1}}[\log E] = \mathbf{E}_{P_{W_1}}[\log E^*_{W_1}]$. By linearity of expectation, $E' = (1/2)E^*_{W_1} + (1/2)E$ is then also an E-variable, and by Jensen's inequality applied to the logarithm we must have $\mathbf{E}_{P_{W_1}}[\log E'] > \mathbf{E}_{P_{W_1}}[\log E^*_{W_1}]$ unless $P_{W_1}(E = E^*_{W_1}) = 1$. Since we have already shown that for any E-variable E', $\mathbf{E}_{P_{W_1}}[\log E'] \le \mathbf{E}_{P_{W_1}}[\log E^*_{W_1}]$, it follows that $P_{W_1}(E \ne E^*_{W_1}) = 0$. But then, by our assumption of absolute continuity, we also have $P_{\theta_0}(E \ne E^*_{W_1}) = 0$ so $E^*_{W_1}$ is essentially unique.

Proof of Theorem 5.4, Part 2 The general result of Part 2 (without the differentiability condition imposed in the proof in the main text) is now a direct extension of Part 1 which we just proved above: by Part 3 of the lemma above, we must have that $Q^\circ = P_{W_0^*}$ and everything follows.

Proof of Theorem 5.4, Part 3 The proof consists of two sub-parts, Part 3(a) relying on Part 1 above (and the RIPr-construction, which works for the case that $\mathcal{W}_1'$ is a singleton), Part 3(b) relying on a minimax theorem from Grünwald and Dawid, 2004 (relying heavily on an earlier result from Topsøe, 1979) that itself works for the case that Θ_0 is a singleton.

Part 3(a). We show the following inequalities:

$$D(P^{[\mathbf{V}]}_{W_1^*} \| P^{[\mathbf{V}]}_{W_0^*}) = \inf_{W_1 \in \mathcal{W}_1'} \inf_{W_0 \in \mathcal{W}_0} D(P_{W_1} \| P_{W_0}) \geq \sup_{E \in \mathcal{E}(\Theta_0)} \inf_{W \in \mathcal{W}_1'} \mathbf{E}_{P_W}[\log E] \geq \inf_{W \in \mathcal{W}_1'} \mathbf{E}_{P_W}[\log E^*_{\mathcal{W}_1'}]. \tag{5.40}$$

The first equality follows by assumption of the Theorem. For the first inequality, note that by Theorem 5.4, Part 1, we have for each fixed $W_1 \in \mathcal{W}_1'$ that

$$\inf_{W_0 \in \mathcal{W}_0} D(P_{W_1} \| P_{W_0}) = \sup_{E \in \mathcal{E}(\Theta_0)} \mathbf{E}_{P_{W_1}}[\log E]$$

and this directly implies the inequality by a standard "inf sup $\geq$ sup inf" argument (the trivial side of the minimax theorem). The second inequality is then immediate since $E^*_{\mathcal{W}_1'} \in \mathcal{E}(\Theta_0)$.

Part (3(b). From (5.40) we see that it now suffices to show that

$$D(P^{[\mathbf{V}]}_{W_1^*} \| P^{[\mathbf{V}]}_{W_0^*}) \leq \inf_{W \in \mathcal{W}_1'} \mathbf{E}_{P_W}[\log E^*_{\mathcal{W}_1'}], \tag{5.41}$$

where by the assumptions of the theorem we may assume that $\min_{W_1 \in \mathcal{W}_1'} D(P^{[\mathbf{V}]}_{W_1} \| P^{[\mathbf{V}]}_{W_0^*}) = D(P^{[\mathbf{V}]}_{W_1^*} \| P^{[\mathbf{V}]}_{W_0^*})$. Since all distributions occurring in (5.41) are marginals on $\mathbf{V}$, and E^* can be written as a function of $\mathbf{V}$, we will from now on simply refer to the marginal densities on $\mathbf{V}$ corresponding to P_W as p_W (rather than p'_W as in the main text), and we will omit the superscripts $[\mathbf{V}]$ from P; thus we take as our basic outcome now $\mathbf{V}$ rather than $\mathbf{Y}$.

We will show the stronger statement that (5.41) holds with equality. For this, let W_0^* and W_1^* be as in the statement of the theorem. Let P be a probability measure that is absolutely continuous with respect to $P^*_{W_0}$. Such P must have a density p and the logarithmic score of p relative to measure $P_{W_0^*}$ is defined, in the standard manner, as $L(z, p) := -\log p(v)/p_{W_0^*}(v)$, which is P-almost surely finite, so that, following standard conventions for expectations of random variables that are unbounded both from above and from below (see Grünwald and Dawid, 2004, Section 3.1), $H_{W_0^*}(P) := \mathbf{E}_{\mathbf{V} \sim P}[L(\mathbf{V}, p)] = -D(P \| P_{W_0^*})$, the standard definition of *entropy relative to* $P_{W_0^*}$, is well-defined and nonpositive.

We will apply the minimax Theorem 6.3 of (Grünwald and Dawid, 2004) with L as defined above. For this, we need to verify Conditions 6.2–6.4 of that paper, where Γ in Condition 6.3 and 6.4 is set to be our $\mathcal{W}_1'$, and the set $\mathcal{Q}$ mentioned in Condition 6.2 must be a superset of Γ. We will take $\mathcal{Q}$ to be the set of all probability distributions absolutely continuous relative to $P_{W_0^*}$; note that each $Q \in \mathcal{Q}$ then has a density q; we let $\mathcal{Q}_{\text{DENS}}$ be the set of all densities corresponding to $\mathcal{Q}$. By our requirement that $D(P_{W_1} \| P_{W_0^*}) < \infty$ for all $W_1 \in \mathcal{W}_1'$, we then have that $\mathcal{W}_1' = \Gamma \subset \mathcal{Q}$ as required. By our definition of $\mathcal{Q}$, Condition 6.2 then follows from Proposition A.1. from the same paper (Grünwald and Dawid, 2004) (with μ in the role of $P_{W_0^*}$), and it remains to

verify Condition 6.3 and 6.4, which, taken together, in our notation together amount to the requirements (a) $\mathcal{W}_1'$ is convex, (b1) for every $W_1 \in \mathcal{W}_1'$, P_{W_1} has a Bayes act relative to L and (b2) $H_{W_0^*}(P_{W_1}) > -\infty$, and (c) there exists W_1^* with $H_{W_0^*}(P_{W_1^*}) = \sup_{W_1 \in \mathcal{W}_1'} H_{W_0^*}(P_{W_1}) < \infty$. Now, (a) holds by definition; (b1) holds because L is a proper scoring rule so the density p of any P is an L-Bayes act for P (see Grünwald and Dawid, 2004 for details); (b2) holds by our assumption that $-H_{W_0^*}(P_{W_1}) = D(P_{W_1} \| P_{W_0^*}) < \infty$ and (c) holds because for all $W_1 \in \mathcal{W}_1'$, $H_{W_0^*}(P_{W_1}) = -D(P_{W_1} \| P_{W_0^*}) \leq 0$.

Theorem 6.3 of Grünwald and Dawid, 2004 together with Lemma 4.1 of that same paper then gives

$$H_{W_0^*}(P_{W_1^*}) = \sup_{W \in \mathcal{W}_1'} \mathbf{E}_{\mathbf{V} \sim P_W}\left[-\log \frac{p_W(\mathbf{V})}{p_{W_0^*}(\mathbf{V})}\right] = \sup_{W \in \mathcal{W}_1'} \inf_{q \in \mathcal{Q}_{\text{DENS}}} \mathbf{E}_{\mathbf{Y} \sim P_W}\left[-\log \frac{q(\mathbf{V})}{p_{W_0^*}(\mathbf{V})}\right]$$
$$= \inf_{q \in \mathcal{Q}_{\text{DENS}}} \sup_{W \in \mathcal{W}_1'} \mathbf{E}_{\mathbf{V} \sim P_W}\left[-\log \frac{q(\mathbf{V})}{p_{W_0^*}(\mathbf{V})}\right] = \sup_{W \in \mathcal{W}_1'} \mathbf{E}_{\mathbf{V} \sim P_W}\left[-\log \frac{p_{W_1^*}(\mathbf{V})}{p_{W_0^*}(\mathbf{V})}\right], \quad (5.42)$$

where, to be more precise, the first equality is immediate from the fact that $-H_{W_0^*}(P_{W_1^*}) = D(P_{W_1^*} \| P_{W_0^*}) = \inf_{W_1 \in \mathcal{W}_1'} D(P_{W_1} \| P_{W_0^*})$ (which we may assume as stated underneath (5.41)). The second follows because the W_0^*-logarithmic score is a proper scoring rule, the third is Theorem 6.3 of Grünwald and Dawid, 2004; this Theorem also gives that the infimum must be achieved by some $W_1' \in \mathcal{W}_1'$, and Lemma 4.1 of that paper then gives that it must be equal to W_1^*, which gives the fourth equality.

But, because the first and last terms in (5.42) must be equal, and using again that $H_{W_0^*} = -D(\cdot \| P_{W_0^*})$, (5.42) implies (5.41), which is what we had to prove.

5.D Proofs that δ-GROW E-variables claimed to be simple really are simple

All our results will rely on the following proposition, which we state and prove first:

Proposition 9. [stochastic dominance and simple E-variables] *Let $\Theta_0 = \{0\}$, let, for $\delta > 0$, $\Theta(\delta)$ be defined as in (5.20) and let $\text{BD}(\Theta(\delta))$ be the boundary $\text{BD}(\Theta(\delta)) = \{\theta \in \Theta_1 : d(\theta \| \Theta_0) = \delta\}$. Suppose that $\min_{W \in \mathcal{W}(\text{BD}(\Theta(\delta)))} D(P_W \| P_0)$ is achieved by some W_1^* (note that this will automatically be the case if $\text{BD}(\Theta(\delta))$ is a finite set), so that by Theorem 5.4, Part 3, $E^*_{\text{BD}(\Theta(\delta))} = p_{W_1^*}(\mathbf{Y})/p_0(\mathbf{Y})$. Then the following statements are equivalent:*

1. $$\inf_{\theta \in \Theta(\delta)} \mathbf{E}_{\mathbf{Y} \sim P_\theta}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right] = \inf_{\theta \in \text{BD}(\Theta(\delta))} \mathbf{E}_{\mathbf{Y} \sim P_\theta}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right]. \quad (5.43)$$
2. *For all $W_1 \in \mathcal{W}(\Theta(\delta))$, we have $D(P_{W_1} \| P_0) \geq D(P_{W_1^*} \| P_0)$.*
3. *We have $E^*_{\Theta(\delta)} = E^*_{\text{BD}(\Theta(\delta))}$ which, if Θ_0 and Θ_1 are as above (5.21), is equivalent to (5.21).*

Furthermore, suppose that there exist a function t, a random variable $T = t(\mathbf{Y})$ (whose density under θ we denote by p'_θ), a $\theta^* \in \text{BD}(\Theta(\delta))$ and a strictly increasing function f such that $\log p_{W_1^*}(\mathbf{Y})/p_0(\mathbf{Y}) = \log p'_{\theta^*}(t(\mathbf{Y}))/p'_0(t(\mathbf{Y})) = f(t(\mathbf{Y}))$ and such that for all $\theta \in \Theta(\delta) \setminus \text{BD}(\Theta(\delta))$, $P_\theta[T]$, the distribution of T under P_θ, first-order stochastically dominates *$P_{\theta^*}[T]$ (i.e. for all t, $F_\theta(t) \leq F_{\theta^*}(t)$ where F_θ is the distribution function of $P_\theta[T]$). Then (5.43) holds.*

Proof. (1) $\Rightarrow$ (2) We first note that the conditions of the proposition imply that for all $\theta \in \text{BD}(\Theta(\delta))$,

$$\mathbf{E}_{\mathbf{Y}\sim P_\theta}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right] \geq \mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right] = D(P_{W_1^*}\|P_0), \tag{5.44}$$

as is immediate from Theorem 5.4, Part 3, which gives that $P_{W_1^*}$ is the information projection on the set $\mathcal{W}'_1 = \mathcal{W}(\text{BD}(\Theta(\delta)))$. Now, fix any $W_1 \in \mathcal{W}(\Theta(\delta))$ and consider the function $f(\alpha) = D((1-\alpha)P_{W_1^*} + \alpha P_{W_1}\|P_0)$ on $\alpha \in [0,1]$. Straightforward differentiation gives the following: the second derivative of f is nonnegative, so f is convex on $[0,1]$. The first derivative of $f(\alpha)$ at $\alpha = 0$ is given by

$$\mathbf{E}_{\mathbf{Y}\sim P_{W_1}}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right] - \mathbf{E}_{\mathbf{Y}\sim P_{W_1^*}}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right] \geq \mathbf{E}_{\mathbf{Y}\sim P_{W_1}}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right] - \inf_{\theta\in\text{BD}(\Theta(\delta))} \mathbf{E}_{\mathbf{Y}\sim P_\theta}\left[\log \frac{p_{W_1^*}(\mathbf{Y})}{p_0(\mathbf{Y})}\right], \tag{5.45}$$

where the first expression is just differentiation and the inequality follows from (5.44). So, if we can show that, no matter how W_1 was chosen, the right-hand side of (5.45) is nonnegative, we must have $f(1) \geq f(0)$ and the desired result follows. But nonnegativity of (5.45) follows by the premise (5.43) and linearity of expectation.

(2) $\Rightarrow$ (3) Since $\inf_{W_1\in\mathcal{W}(\Theta(\delta)),W_0\in\mathcal{W}(\Theta_0)} D(P_{W_1}\|P_0) = D(P_{W_1^*}\|P_0)$ we can apply Theorem 5.4, Part 3, which gives the required result.

(3) $\Rightarrow$ (1) is immediate using the definitions of $E^*_{\Theta(\underline{\delta})}$ and $E^*_{\text{BD}(\Theta(\underline{\delta}))}$

For the second part, note that, by a general property of stochastic dominance (Pomatto, Strack and Tamuz, 2020) we have for arbitrary distributions $P[T]$: if $P[T]$ stochastically dominates $P_{\theta^*}[T]$, then we must also have $\mathbf{E}_{P[T]}[f(T)] \geq \mathbf{E}_{P_{\theta^*}}[f(T)]$. This immediately implies the result. □

Proofs that δ-GROW E-variables claimed to be simple are simple We need to show this for four cases mentioned in the main text. In all these cases we show this by establishing the existence of a statistic T as needed to apply the second part of Proposition 9.

1. One-Sided Exponential Families (Section 5.4.1) In this case $\text{BD}(\Theta(\underline{\delta}))$ is a singleton, so W_1^* is the degenerate distribution putting all mass on $\underline{\delta}$. We take $T = t(\mathbf{Y})$ to be the sufficient statistic for the family at the given sample size. That is, we re-represent our exponential family in the canonical parameterization, and let β_δ be the canonical parameter corresponding to

$\delta > 0$; we can choose the parameterization such that $\beta_0 = 0$. With $T = t(\mathbf{Y})$ the sufficient statistic, we then have $\log p_{\underline{\delta}}(\mathbf{Y})/p_0(\mathbf{Y}) = \beta_{\underline{\delta}} t(\mathbf{Y}) + \log(Z(0)/Z(\beta_{\underline{\delta}})) = f(t(\mathbf{Y}))$; here $Z(\cdot)$ is the normalization function. Since β_δ is strictly increasing with δ (another general property of exponential families) and $\beta_0 = 0$, we have that $f(T)$ is increasing in T. It thus remains to show that $P_\delta^{[T]}$ stochastically dominates $P_{\underline{\delta}}^{[T]}$ for $\delta > \underline{\delta}$. But this is immediate by basic rewriting, giving $F_\beta(t) = \int_{-\infty}^{t} \exp(\beta t) dP_0^{[T]}(t) / \int_{-\infty}^{\infty} \exp(\beta t) dP_0^{[T]}$, and then taking derivatives.

2. Two-Sided Normal Location Family (Section 5.4.1) We take $T = \widehat{\mu}^2$, the square of the empirical mean. The result then follows by reasoning similarly to 4. below but is easier, hence we omit details.

3. One-Sided normal with unknown variance (Section 5.4.3) Note first that $E^*_{\underline{\delta}} = p'_{\underline{\delta}}(\mathbf{V})/p'_0(\mathbf{V})$. Thus, by expressing E-variables in terms of $\mathbf{V}$ we can re-represent the problem as having a simple $\mathcal{H}_0$ so that we can use Proposition 9. We take $T = t_s(\mathbf{Y})$ to be the Student's T-statistic. Straightforward rewriting gives that, for $\delta > 0$, for all σ, $p_{\underline{\delta}}(\mathbf{V})/p_0(\mathbf{V}) = f(T)$ for some increasing function f of T. We thus need to show that the distribution of T under $P_{\underline{\delta}}^{[T]}$ is stochastically dominated by its distribution under $P_{\delta'}^{[T]}$, for $\delta' > \underline{\delta}$. But these are just two noncentral t-distributions with $\nu := n - 1$ degrees of freedom and noncentrality parameter $\mu = \sqrt{n}\underline{\delta}$ vs. $\mu = \sqrt{n}\delta'$ respectively. Since a noncentral t distribution with parameters (ν, μ) can be viewed as the distribution of $(Z + \mu)/\sqrt{V/\nu}$ where Z is standard normal and V is an independent χ^2 random variable, stochastic dominance is immediate from the fact that $\underline{\delta} > 0$.

4. Two-sided normal with unknown variance (Section 5.4.3) This case is similar to the previous one but now we take $T = (t_s(\mathbf{Y}))^2$ to be the absolute value of Student's t-statistic $t_s(\mathbf{Y})$. Symmetry considerations dictate that $E^*_{\underline{\delta}} = ((1/2)p'_{-\underline{\delta}}(\mathbf{V}) + (1/2)p'_{\underline{\delta}}(\mathbf{V}))/p'_0(\mathbf{V})$. It is easy to verify that this quantity only depends on T and is strictly increasing in T. Again by symmetry, the distribution of T under $P_\delta[T]$ is the same as its distribution under $P_{-\delta}[T]$ and then also the same as its distribution under $P_{(1/2)\delta-(1/2)\delta}[T]$. It thus suffices to show that $P_{\underline{\delta}}[T]$ is stochastically dominated by $P_{\delta'}[T]$ for $\delta' > \underline{\delta} > 0$. But the distribution of T under P_δ is now the ratio of two independent χ^2 distributions, a noncentral χ^2 with one degree of freedom and noncentrality δ and a central χ^2 with $n-1$ degrees of freedom. By independence, it is sufficient to prove that noncentral χ^2's with one degree of freedom and noncentrality $\delta' > \delta$ dominates a noncentral χ^2 with one degree of freedom and noncentrality δ. But this is straightforward by differentiating the cumulative distribution functions.

Relating $E^\circ_{\Theta(\underline{\delta})}$ and $E^*_{\Theta(\underline{\delta})}$ in the two-sided case We have, on all samples,

$$\log E^\circ_{\Theta(\underline{\delta})} \geq \max\{\log(1/2)E^*_{\underline{\delta}}, \log(1/2)E^*_{-\underline{\delta}}\},$$

so that

$$\begin{aligned}
\inf_{\theta:|\theta|\geq\underline{\delta}} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log E^\circ_{\Theta(\underline{\delta})}] &\geq \inf_{\theta:|\theta|\geq\underline{\delta}} \max\left\{\mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log \frac{1}{2}E^*_{\underline{\delta}}], \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log \frac{1}{2}E^*_{-\underline{\delta}}]\right\} \qquad (5.46)\\
&\geq \max\left\{\inf_{\theta:|\theta|\geq\underline{\delta}} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log \frac{1}{2}E^*_{\underline{\delta}}], \inf_{\theta:|\theta|\geq\underline{\delta}} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log \frac{1}{2}E^*_{-\underline{\delta}}]\right\}\\
&\geq \max\left\{\inf_{\theta:\theta\geq\underline{\delta}} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log \frac{1}{2}E^*_{\underline{\delta}}], \inf_{\theta:\theta\leq-\underline{\delta}} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log \frac{1}{2}E^*_{-\underline{\delta}}]\right\}\\
&= \max\left\{\mathbf{E}_{\mathbf{Y}\sim P_{\underline{\delta}}}[\log \frac{1}{2}E^*_{\underline{\delta}}], \mathbf{E}_{\mathbf{Y}\sim P_{-\underline{\delta}}}[\log \frac{1}{2}E^*_{-\underline{\delta}}]\right\},
\end{aligned}$$

where the final equality is just condition (5.43) of the proposition above again for the one-sided case, which above we already showed to hold for 1-dimensional exponential families. On the other hand, letting $W_{\underline{\delta}}$ be the prior that puts mass $1/2$ on $\underline{\delta}$ and $1/2$ on $-\underline{\delta}$, we have:

$$\begin{aligned}
\inf_{\theta:|\theta|\geq\underline{\delta}} \mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log E^*_{\Theta(\underline{\delta})}] &\leq \mathbf{E}_{\theta\sim W_{\underline{\delta}}}\mathbf{E}_{\mathbf{Y}\sim P_\theta}[\log E^*_{\Theta(\underline{\delta})}] \qquad (5.47)\\
&\leq \mathbf{E}_{\theta\sim W_{\underline{\delta}}}\mathbf{E}_{\mathbf{Y}\sim P_\theta}\left[\log \frac{P_{W_{\underline{\delta}}}(\mathbf{Y})}{P_0(\mathbf{Y})}\right]\\
&= \mathbf{E}_{\theta\sim W_{\underline{\delta}}}\mathbf{E}_{\mathbf{Y}\sim P_\theta}\left[\log E^\circ_{\Theta(\underline{\delta})}\right]\\
&= \frac{1}{2}\mathbf{E}_{\underline{\delta}}[\log \frac{1}{2}E^*_{\underline{\delta}}] + \frac{1}{2}\mathbf{E}_{-\underline{\delta}}[\log \frac{1}{2}E^*_{-\underline{\delta}}] + \varepsilon_n\\
&\leq \max\left\{\mathbf{E}_{\mathbf{Y}\sim P_{\underline{\delta}}}[\log \frac{1}{2}E^*_{\underline{\delta}}], \mathbf{E}_{\mathbf{Y}\sim P_{-\underline{\delta}}}[\log \frac{1}{2}E^*_{-\underline{\delta}}]\right\} + \varepsilon_n,
\end{aligned}$$

where the first inequality is linearity of expectation and the second inequality follows because, since $E^*_{\Theta(\underline{\delta})}$ is an E-variable relative to $\{P_0\}$, we can set $q := E^*_{\Theta(\underline{\delta})} \cdot p_0$; then $\int q(\mathbf{Y})\,\mathrm{d}\lambda \leq 1$ and $E^*_{\Theta(\underline{\delta})} = q(\mathbf{Y})/p_0(\mathbf{Y})$, and the inequality follows by the information inequality of information theory. ε_n above is defined as:

$$\begin{aligned}
\varepsilon_n &= \frac{1}{2}\cdot\left(\mathbf{E}_{\underline{\delta}}[\log E^\circ_{\Theta(\underline{\delta})} - \log \frac{1}{2}E^*_{\underline{\delta}}] + \mathbf{E}_{-\underline{\delta}}[\log E^\circ_{\Theta(\underline{\delta})} - \log \frac{1}{2}E^*_{-\underline{\delta}}]\right)\\
&= \log 2 + \frac{1}{2}\cdot\left(\mathbf{E}_{\underline{\delta}}[\log E^\circ_{\Theta(\underline{\delta})}/E^*_{\underline{\delta}}] + \mathbf{E}_{-\underline{\delta}}[\log E^\circ_{\Theta(\underline{\delta})}/E^*_{-\underline{\delta}}]\right)\\
&= \log 2 - \frac{1}{2}\left(D(P_{\underline{\delta}}(\mathbf{Y})\|P_{W_{\underline{\delta}}}(\mathbf{Y})) + D(P_{-\underline{\delta}}(\mathbf{Y})\|P_{W_{\underline{\delta}}}(\mathbf{Y}))\right).
\end{aligned}$$

Together, (5.46) and (5.47) show that $E^\circ_{\Theta(\underline{\delta})}$ is an E-variable whose worst-case growth rate is always within $\varepsilon_n \leq \log 2$ ('1 bit') of that of the minimax optimal $E^*_{\Theta(\underline{\delta})}$; moreover, for fixed $\underline{\delta}$, ε_n quickly converges to 0, since, for $\theta \in \{\underline{\delta}, -\underline{\delta}\}$, if $\mathbf{Y} \sim P_\theta$, then with high probability, $P_{-\theta}/P_\theta$ will be exponentially small in n, so that $D(P_\theta(\mathbf{Y})\|P_{W_{\underline{\delta}}}(\mathbf{Y})) \approx -\log(1/2) = \log 2$.

5.E Proofs and Details for Section 5.4.3

We first walk through the claims made in Section 5.4.3. The first claim is that under all $P_{0,\sigma}$ with $\sigma > 0$, $\mathbf{V}$ has the same distribution, say P_0, and under all $P_{W[\delta],\sigma}$ with $\sigma > 0$, $\mathbf{V}$ has the same

distribution, say $P_{W[\delta]}(\mathbf{V})$. To show this, it is sufficient to prove that for all σ, all $\delta \in \mathbb{R}$, under all $P_{\delta,\sigma}$, the distribution of $\mathbf{V}$ only depends on δ but not on σ. But this follows easily: for $i \in 1..n$, we define $Y_i' = Y_i/\sigma$. Then Y_i' is $\sim N(\delta, 1)$. But we can write $\mathbf{V}$ as a function of $(Y_1', \ldots, Y_n')$, hence the distribution of $\mathbf{V}$ does not depend on σ either (note that at this stage, symmetry of the prior is not yet required).

(5.29) (we only need to show the first equality) is straightforward to show: one first notes that, for every $c > 0$,

$$\frac{\int_\sigma \overline{p}_{W[\delta],\sigma}(\mathbf{Y}/c) w^H(\sigma)\,\mathrm{d}\sigma}{\int_\sigma p_{0,\sigma}(\mathbf{Y}/c) w^H(\sigma)\,\mathrm{d}\sigma} = \frac{\int_\sigma \overline{p}_{W[\delta],\sigma}(\mathbf{Y}) w^H(\sigma)\,\mathrm{d}\sigma}{\int_\sigma p_{0,\sigma}(\mathbf{Y}) w^H(\sigma)\,\mathrm{d}\sigma},$$

which follows easily by changing the domain of integration in the leftmost expression in both numerator and denominator from σ to $c\sigma$ and noting that this incurs the same factor c^n in both numerator and denominator, which therefore cancels. Since we assume $Y_1 \neq 0$, the first equality in (5.29) now follows by setting $c \coloneqq Y_1$.

Proof of Theorem 5.6 *Part 1.* For $0 < a < b < \infty$, denote by $W_{[a,b]}$ the *restricted Haar prior*, i.e. the probability distribution on σ with density

$$w_{[a,b]}(\sigma) \coloneqq \begin{cases} \frac{1}{\sigma} \cdot \frac{1}{\log b/a} & \text{if } \sigma \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$$

For notational convenience we abbreviate the joint distribution of σ and $\mathbf{Y}$ for effect size prior $W[\delta]$ and restricted Haar prior $W_{[a,b]}$ on σ to $P_{W[\delta],[a,b]} \coloneqq P_{W[\delta],W_{[a,b]}[\sigma]}$. The Bayes factor for effect size prior $W[\delta]$ vs. effect size 0 at sample size n based on using the restricted Haar prior $W_{[a,b]}$ in both $\mathcal{H}_1$ and $\mathcal{H}_0$ and data $\mathbf{Y}$ will be denoted as

$$B_{[a,b]}(\mathbf{Y}) = \frac{\int_{\sigma\in[a,b]} \overline{p}_{W[\delta],\sigma}(\mathbf{Y}) w_{[a,b]}(\sigma)\,\mathrm{d}\sigma}{\int_{\sigma\in[a,b]} p_{0,\sigma}(\mathbf{Y}) w_{[a,b]}(\sigma)\,\mathrm{d}\sigma}.$$

The Bayes factor based on the right Haar prior can then be written as $B_{[0,\infty]}(\mathbf{Y})$. From (5.29), we have for all $\sigma > 0$ that

$$D\left(P^{[\mathbf{V}]}_{W[\delta]} \| P^{[\mathbf{V}]}_0\right) = \mathbf{E}_{\mathbf{V}\sim P_{W[\delta]}}\left[\frac{p'_{W[\delta]}(\mathbf{V})}{p'_0(\mathbf{V})}\right] = \mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],\sigma}}\left[\log B_{[0,\infty]}(\mathbf{Y})\right]. \tag{5.48}$$

Since $\mathbf{V}$ is a coarsening of $\mathbf{Y}$, by the information inequality (Cover and Thomas, 1991), we must also have, for all priors $W[\sigma]$, $W[\sigma']$:

$$D\left(P_{W[\delta],W'[\sigma]} \| P_{0,W[\sigma]}\right) \geq D\left(P^{[\mathbf{V}]}_{W[\delta],W'[\sigma]} \| P^{[\mathbf{V}]}_{0,W[\sigma]}\right) = D\left(P^{[\mathbf{V}]}_{W[\delta]} \| P^{[\mathbf{V}]}_0\right), \tag{5.49}$$

where we also used that the marginal distributions on $\mathbf{V}$ do not depend on σ. Combining (5.48) and (5.49), we find that it suffices to prove the following lemma, which is done further below.

Lemma 10. *For all $W[\delta]$ satisfying the condition of Theorem 5.6, for all $\sigma > 0$, we have:*

$$\lim_{i\to\infty} D\left(P_{W[\delta],[1/i,i]} \| P_{0,[1/i,i]}\right) = \mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],\sigma}}\left[\log B_{[0,\infty]}(\mathbf{Y})\right]. \tag{5.50}$$

Part 2. Fix $\mathcal{W}[\delta]$ as in the theorem statement, and any corresponding $\mathcal{W}_1'$ as above. We have:

$$\begin{aligned}\inf_{W[\delta]\in\mathcal{W}[\delta]} D\left(P^{[\mathbf{V}]}_{W[\delta]} \| P^{[\mathbf{V}]}_0\right) &\le \inf_{W\in\mathcal{W}_1'} \inf_{W[\sigma]\in\mathcal{W}[\Gamma]} D(P_W \| P_{0,W[\sigma]}) \\ &\le \inf_{W[\delta]\in\mathcal{W}[\delta]} \inf_{W[\sigma]\in\mathcal{W}[\Gamma]} D(P_{W[\delta],W[\sigma]} \| P_{0,W[\sigma]}) = \inf_{W[\delta]\in\mathcal{W}[\delta]} D\left(P^{[\mathbf{V}]}_{W[\delta]} \| P^{[\mathbf{V}]}_0\right).\end{aligned} \tag{5.51}$$

Here the first inequality is based on (5.49), the second is immediate and the third follows by noting that, by Part 1, for any fixed $W[\delta] \in \mathcal{W}[\delta]$, we have

$$\inf_{W[\sigma]\in\mathcal{W}[\Gamma]} D(P_{W[\delta],W[\sigma]} \| P_{0,W[\sigma]}) = D\left(P^{[\mathbf{V}]}_{W[\delta]} \| P^{[\mathbf{V}]}_0\right).$$

But (5.51) is equivalent to the desired result.

5.E.1 Proof of Lemma 10

Define random variables $\overline{U} := \sqrt{n^{-1}\sum Y_i^2}$, $\overline{\mathbf{Y}} := n^{-1}\sum Y_i$ and $T := \overline{\mathbf{Y}}/\overline{U} \in [-1,1]$ is an invariant, i.e. a function of $\mathbf{V}$. We will sometimes express $\overline{U}$ and T as functions of $\mathbf{Y}$ and freely write $\overline{U}(\mathbf{Y})$, $T(\mathbf{Y})$ when this notation is more convenient.

The Bayes factor $B_{[a,b]}(\mathbf{Y})$ depends on $\mathbf{Y}$ only through the functions $\overline{U}(\mathbf{Y})$ and $T(\mathbf{Y})$. We will therefore also write it, whenever convenient, as a function of these random variables, and denote it as $B_{[a,b]}(\overline{U}, T)$.

The proof will combine the following two (sub-) lemmas, whose proof is deferred to further below. The first lemma allows us to conclude that, *when restricted to events of small (marginal) probability, the expectation of the log Bayes factor is also small.*

The second lemma allows us to conclude that, as $i \to \infty$, the expected log Bayes factor uniformly converges on $\mathbf{y} \in A_i$, where A_i is a set that itself grows towards $\mathbb{R}^n$. Thus, while uniform convergence for all $\mathbf{y} \in \mathbb{R}^n$ is too much to ask for, remarkably we do get uniform convergence on a 'noncompact' sequence of sets: the sets A_i are not included in any compact set.

Lemma 11. [Uniform Integrability-Flavored Lemma] *Let A be a measurable subset of $\mathbb{R}^n$. We have for all $0 < a < b < \infty$, $W[\delta]$ as in the theorem statement, that:*

$$\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a,b]}}\left[\mathbb{1}_{\{\mathbf{Y}\in A\}} \cdot \left(-\log B_{[0,\infty]}(\mathbf{Y})\right)\right] \le P_{W[\delta],[a,b]}(\mathbf{Y}\in A) \log \frac{1}{P_{W[\delta],[a,b]}(\mathbf{Y}\in A)} \tag{5.52}$$

Suppose further that $\mathbf{E}_{\delta\sim W[\delta]}[|\delta|^{2+\varepsilon}] < \infty$ for some $\varepsilon > 0$. Then

$$\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a,b]}}\left[\mathbb{1}_{\{\mathbf{Y}\in A\}} \cdot \log B_{[a,b]}(\mathbf{Y})\right] \le P_{W[\delta],[a,b]}(\mathbf{Y}\in A)^{\varepsilon/(1-\varepsilon)} \cdot C \tag{5.53}$$

were C is a constant depending on $W[\delta]$, n (but not on a, b).

Lemma 12. [Uniform Convergence Beyond Compactness] *Let $(a_i, b_i, \underline{c}_i, \overline{c}_i)_{i\in\mathbb{N}}$ be a sequence of numbers in $\mathbb{R}^+$ such that for all i, $\underline{c}_i > 1$ and $\overline{c}_i < 1$, $\underline{c}_i a_i < \overline{c}_i b_i$ (hence also $a_i < b_i$), and $\lim_{i\to\infty} a_i = 0, \lim_{i\to\infty} b_i = \infty, \lim_{i\to\infty} \underline{c}_i = \infty, \lim_{i\to\infty} \overline{c}_i = 0$, $\lim(\overline{c}_i b_i - \underline{c}_i a_i) = \infty$ (For example, take $a_i = 1/i$, $b_i = i$, $\underline{c}_i = \log(i+1)$, $\overline{c}_i = 1/\log(i+1)$). Then:*

$$\limsup_{i\to\infty} \sup_{t\in[-1,1],\overline{u}\in[a_i\underline{c}_i, b_i\overline{c}_i]} \left(\log B_{[a_i,b_i]}(\overline{u}, t) - \log B_{[0,\infty]}(\overline{u}, t)\right) = 0.$$

The proof of Lemma 12 is itself based on another key observation, which is an immediate consequence of the fact that $W_{[a,b]}$ is proportional to the Haar measure on $[a, b]$:

Proposition 13. [Change-of-Variables] *We have for all $\overline{u} > 0$, all $t \in [-1, 1]$, $B_{[a,b]}(\overline{u}, t) = B_{[a/\overline{u}, b/\overline{u}]}(1, t)$.*

We now first show how the two lemmas imply the main result. Take any sequence $(a_i, b_i, \underline{c}_i, \overline{c}_i)$ satisfying the requirements of Lemma 12. Let

$$A_i = \{\mathbf{Y} \in \mathbb{R}^n : \underline{c}_i a_i \le \overline{U}(\mathbf{Y}) \le \overline{c}_i b_i\}.$$

and let $\overline{A}_i \subset \mathbb{R}^n$ be its complement. We have

$$\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a_i,b_i]}}\left[\log B_{[a_i,b_i]}(\mathbf{Y}) - \log B_{[0,\infty]}(\mathbf{Y})\right] = f(i) + g(i),$$

where

$$f(i) = \mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a_i,b_i]}}\left[\mathbb{1}_{\{\mathbf{Y}\in A_i\}} \cdot \log \frac{B_{[a_i,b_i]}(\mathbf{Y})}{B_{[0,\infty]}(\mathbf{Y})}\right],$$
$$g(i) = \mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a_i,b_i]}}\left[\mathbb{1}_{\{\mathbf{Y}\in \overline{A}_i\}} \cdot \log \frac{B_{[a_i,b_i]}(\mathbf{Y})}{B_{[0,\infty]}(\mathbf{Y})}\right].$$

Now, take $a_i = 1/i, b_i = i, \underline{c}_i = \log(i+1), \overline{c}_i = 1/\log(i+1)$. We already indicated in Lemma 12 that this choice allows us to apply Lemma 12 to $f(i)$, which will therefore converge to 0 as $i \to \infty$. It thus remains to show that $g(i) \to 0$. By Lemma 11 we have $g(i) = o(P_{W[\delta],[a_i,b_i]}(\mathbf{Y} \in \overline{A}_i))$. It thus suffices to show that $P_{W[\delta],[a_i,b_i]}(\mathbf{Y} \in \overline{A}_i) \to 0$. For this, note that we can write:

$$\begin{aligned}
&P_{W[\delta],[a_i,b_i]}(\mathbf{Y} \in \overline{A}_i) = \mathbf{E}_{\sigma\sim W_{[a_i,b_i]}}\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],1}}\left[\mathbb{1}_{\{(\sigma Y_1,\ldots,\sigma Y_n)\in\overline{A}_i\}}\right] \\
&= \mathbf{E}_{\sigma\sim W_{[a_i,b_i]}}\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],1}}\left[\mathbb{1}_{\{\sigma\overline{U}(\mathbf{Y})<\underline{c}_i a_i \vee \sigma\overline{U}(\mathbf{Y})>\overline{c}_i b_i\}}\right] \\
&\le W_{[a_i,b_i]}(\sigma < \underline{c}_i a_i \vee \sigma > \overline{c}_i b_i) + \mathbf{E}_{\sigma\sim W_{[a_i,b_i]}}\left[\mathbb{1}_{\{\underline{c}_i a_i<\sigma<\overline{c}_i b_i\}} \cdot \mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],1}}\left[\mathbb{1}_{\{\sigma\overline{U}(\mathbf{Y})<\underline{c}_i a_i\vee\sigma\overline{U}(\mathbf{Y})>\overline{c}_i b_i\}}\right]\right] \\
&= W_{[a_i,b_i]}(\sigma < \underline{c}_i a_i) + W_{[a_i,b_i]}(\sigma > \overline{c}_i b_i) + P_{W[\delta],1}(\overline{U} < \underline{c}_i a_i) + P_{W[\delta],1}(\overline{U} > \overline{c}_i b_i),
\end{aligned}$$

where we used the union bound. Now, by our choice of $(a_i, b_i, \underline{c}_i, \overline{c}_i)$, the first two probabilities go to 0 as $i \to \infty$. And, since $a_i\underline{c}_i \to 0$ and $\overline{c}_i b_i \to \infty$ and $\overline{U}$ has a fixed distribution which has no mass at $\overline{U} \le 0$ (to be precise, $n\overline{U}^2$ has a noncentral χ^2 distribution), the third and fourth term go to 0 as well. The result is proved.

Remaining Proofs underlying Lemma 10

Proof. **(of Proposition 13)** Changing the integration variable from σ to $\rho := \sigma/u$, we have:

$$B_{[a,b]}(u,t) = \frac{\int_\delta \int_{\sigma=a}^{\sigma=b} \frac{1}{\sigma} e^{n\cdot(-\frac{1}{2}\delta^2+\delta u t/\sigma - \frac{1}{2}u^2/\sigma^2)} \,\mathrm{d}\sigma \,\mathrm{d}W[\delta]}{\int_a^b \frac{1}{\sigma} e^{-(n/2)\cdot u^2/\sigma^2}\,\mathrm{d}\sigma}$$
$$= \frac{\int_\delta \int_{\rho=a/u}^{\rho=b/u} \frac{1}{u\rho} e^{n\cdot(-\frac{1}{2}\delta^2+\delta u t/(u\rho) - \frac{1}{2}u^2/(u^2\rho^2))} \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\rho}\right) \mathrm{d}\rho \,\mathrm{d}W[\delta]}{\int_{\rho=a/u}^{\rho=b/u} \frac{1}{u\rho} e^{-(n/2)\cdot u^2/(u^2\rho^2)} \left(\frac{\mathrm{d}\sigma}{\mathrm{d}\rho}\right) \mathrm{d}\rho},$$

and the result follows by rewriting. □

Proof. **(of Lemma 11)** *Part 2.* Let $W_{[a,b]} \mid \mathbf{y}$ be the posterior distribution on (δ,σ) based on prior $W[\delta] \times W_{[a,b]}$. By straightforward rewriting we can re-express $1/B_{[a,b]}(\mathbf{y})$ as an expectation over the posterior $W_{[a,b]} \mid \mathbf{y}$. We do this in the second step below, and then continue using Jensen's inequality:

$$\log B_{[a,b]}(\mathbf{y}) = -\log \frac{\int_\delta \int_{\sigma\in[a,b]} e^{-n(\overline{\mathbf{y}}^2/2\sigma^2+\delta^2/2-\delta\cdot\overline{\mathbf{y}}/\sigma)+n(\delta^2/2-\delta\cdot\overline{\mathbf{y}}/\sigma)}\,\mathrm{d}\sigma\,\mathrm{d}W[\delta]}{e^{-n(\overline{\mathbf{y}}^2/2\sigma^2+\delta^2/2-\delta\cdot\overline{\mathbf{y}}/\sigma)}\,\mathrm{d}\sigma\,\mathrm{d}W[\delta]}$$
$$= -\log \mathbf{E}_{(\delta,\sigma)\sim W_{[a,b]}|\mathbf{y}}\left[e^{n\cdot(\frac{1}{2}\delta^2-\delta\overline{\mathbf{y}}/\sigma)}\right]$$
$$\le -\frac{1}{2}\cdot n\delta^2 + \frac{1}{2}n\cdot\mathbf{E}_{(\delta,\sigma)\sim W_{[a,b]}|\mathbf{y}}\left[\overline{\mathbf{y}}\cdot\delta/\sigma\right] \le \frac{1}{2}n\cdot\mathbf{E}_{(\delta,\sigma)\sim W_{[a,b]}|\mathbf{y}}\left[|\overline{\mathbf{y}}|\cdot|\delta|/\sigma\right].$$

We thus have, by Hölder's inequality, for $q, r > 1$ with $1/r + 1/q = 1$:

$$\mathbf{E}_{\mathbf{Y}}\left[\mathbb{1}_{\{\mathbf{Y}\in A\}}\cdot\log B_{[a,b]}(\mathbf{Y}, W[\delta])\right] \le \left(\mathbf{E}_{\mathbf{Y}}\left[\mathbb{1}^q_{\{\mathbf{Y}\in A\}}\right]\right)^{1/q}\cdot\left(\mathbf{E}_{\mathbf{Y}}\left(\mathbf{E}_{(\delta,\sigma)\sim W_{[a,b]}|\mathbf{Y}}\left[(n/2)|\overline{\mathbf{Y}}||\delta|/\sigma\right]\right)^r\right)^{1/r}$$
$$\le P(\mathbf{Y}\in A)^{1/q}\cdot(n/2)\cdot\left(\mathbf{E}_{\mathbf{Y}}\mathbf{E}_{(\delta,\sigma)\sim W_{[a,b]}|\mathbf{Y}}\left(|\overline{\mathbf{Y}}||\delta|/\sigma\right)^r\right)^{1/r},$$

where in the final line we once again used Jensen. The expectation can be rewritten as:

$$\mathbf{E}_{\mathbf{Y}}\mathbf{E}_{(\delta,\sigma)\sim W_{a,b}|\mathbf{Y}}\left(|\overline{\mathbf{Y}}||\delta|/\sigma\right)^r = \mathbf{E}_{\delta\sim W[\delta],\sigma\sim W_{[a,b]}}\mathbf{E}_{Y_1,\ldots,Y_n\ \text{i.i.d.}\sim P_{\delta,\sigma}}\left(|\overline{\mathbf{Y}}||\delta|/\sigma\right)^r$$
$$= \mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{\sigma\sim W_{[a,b]}}\mathbf{E}_{\mathbf{Y}'\sim N(\delta/n,1/n)}\left(|\mathbf{Y}'||\delta|\right)^r$$
$$= n^{-r}\mathbf{E}_{\delta\sim W[\delta]}|\delta|^r\mathbf{E}_{\mathbf{Y}'\sim N(\delta,1)}|\mathbf{Y}'|^r$$
$$\le 2^r n^{-r}\mathbf{E}_{\delta\sim W[\delta]}|\delta|^r\mathbf{E}_{\mathbf{Y}'\sim N(1,\delta)}\left[(|\mathbf{Y}'-\delta|+|\delta|)^r\right]$$
$$\le 2^r n^{-r}\mathbf{E}_{\delta\sim W[\delta]}\left[|\delta|^{2r}+|\delta|^r C_r\right],$$

where we used that $|a+b|^r \le (2\max\{|a|,|b|\})^r \le 2^r(|a|^r+|b|^r)$ and that, if $\mathbf{Y}\sim N_{0,1}$, then $\mathbf{E}[|\mathbf{Y}|^r] \le C_r$ for a constant C_r that does not depend on δ. The result follows.

Part 1. Recall that $\mathbf{V}$ denotes the maximal invariant. Its marginal distribution does not depend on σ, so for any $0 < a' < b'$ we can write:

$$\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a,b]}}\left[\mathbb{1}_{\{\mathbf{Y}\in A\}}\cdot\left(-\log B_{[0,\infty]}(\mathbf{Y})\right)\right] =$$
$$\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a,b]}}\left[\mathbb{1}_{\{\mathbf{Y}\in A\}}\cdot\left(\log\frac{p_{[a,b],0}(\mathbf{V}(\mathbf{Y}))}{p_{W[\delta],[a,b]}(\mathbf{V}(\mathbf{Y}))}\right)\right] =$$
$$P_{W[\delta],[a,b]}(\mathbf{Y}\in A)\cdot\mathbf{E}_{\mathbf{Y}\sim P_{W[\delta],[a,b]}|\mathbf{Y}\in A}\left[\log\frac{p_{[a,b],0}(\mathbf{V}(\mathbf{Y})\mid\mathbf{Y}\in A)}{p_{W[\delta],[a,b]}(\mathbf{V}(\mathbf{Y})\mid\mathbf{Y}\in A)}+\log\frac{P_{[a,b],0}(\mathbf{Y}\in A)}{P_{W[\delta],[a,b]}(\mathbf{Y}\in A)}\right]\le$$
$$P_{W[\delta],[a,b]}(\mathbf{Y}\in A)\cdot\left(\log P_{[a,b],0}(\mathbf{Y}\in A)-\log P_{W[\delta],[a,b]}(\mathbf{Y}\in A)\right)\le$$
$$-P_{W[\delta],[a,b]}(\mathbf{Y}\in A)\log P_{W[\delta],[a,b]}(\mathbf{Y}\in A)$$

where we used Jensen's inequality. □

Proof. **(of Lemma 12)** Using Proposition 13 and its consequence that $B_{[0,\infty]}$ depends on the invariant only, i.e. for all $\overline{u} > 0$, $B_{[0,\infty]}(\overline{u}, t) = B_{[0,\infty]}(1, t)$, we can rewrite the supremum as

$$\sup_{t\in[-1,1],\,\overline{u}\in[a_i\underline{c}_i,b_i\overline{c}_i]}\left(\log B_{[a_i/\overline{u},b_i/\overline{u}]}(1,t)-\log B_{[0,\infty]}(1,t)\right)\le$$
$$\sup_{t\in[-1,1],\,0<c<1/\underline{c}_i,\,c'>1/\overline{c}_i}\left(\log B_{[c,c']}(1,t)-\log B_{[0,\infty]}(1,t)\right)\le$$
$$\sup_{0<c<1/\underline{c}_i,\,c'>1/\overline{c}_i}\left(\log\int_0^\infty\frac{1}{\sigma}e^{-(n/2)\sigma^{-2}}\,\mathrm{d}\sigma-\log\int_c^{c'}\frac{1}{\sigma}e^{-(n/2)\sigma^{-2}}\,\mathrm{d}\sigma\right)\le$$
$$\left(\log\int_0^\infty\frac{1}{\sigma}e^{-(n/2)\sigma^{-2}}\,\mathrm{d}\sigma-\log\int_{1/\underline{c}_i}^{1/\overline{c}_i}\frac{1}{\sigma}e^{-(n/2)\sigma^{-2}}\,\mathrm{d}\sigma\right)=f(\underline{c}_i,\overline{c}_i)$$

for some function $f(\underline{c},\overline{c})$ with $\lim_{\underline{c}\to\infty,\overline{c}\downarrow 0}f(\underline{c},\overline{c}) = 0$ (note that the dependence on t has disappeared); the result follows. Here we used that, for general $u, t, 0 < a < b$,

$$\log B_{[a,b]}(u,t)-\log B_{[0,\infty]}(u,t) =$$
$$\log\frac{\int_\delta\int_{\sigma=a}^b\frac{1}{\sigma}e^{n\cdot(-\frac{1}{2}\delta^2+\delta ut/\sigma-\frac{1}{2}u^2/\sigma^2)}\,\mathrm{d}\sigma\,\mathrm{d}W[\delta]}{\int_a^b\frac{1}{\sigma}e^{-(n/2)\cdot u^2/\sigma^2}\,\mathrm{d}\sigma}-\log\frac{\int_\delta\int_{\sigma=0}^\infty\frac{1}{\sigma}e^{n\cdot(-\frac{1}{2}\delta^2+\delta ut/\sigma-\frac{1}{2}u^2/\sigma^2)}\,\mathrm{d}\sigma\,\mathrm{d}W[\sigma]}{\int_0^\infty\frac{1}{\sigma}e^{-(n/2)\cdot u^2/\sigma^2}\,\mathrm{d}\sigma}\le$$
$$\log\int_0^\infty\frac{1}{\sigma}e^{-(n/2)\cdot u^2/\sigma^2}\,\mathrm{d}\sigma-\log\int_a^b\frac{1}{\sigma}e^{-(n/2)\cdot u^2/\sigma^2}\,\mathrm{d}\sigma.$$

□

5.E.2 Why W_1^* and W_0^* are achieved and have finite support in Section 5.4.5

The minima are achieved because of the joint lower-semi-continuity of KL divergence (Posner, 1975). To see that the supports are finite, note the following: for given sample size n, the probability distribution P_W is completely determined by the probabilities assigned to the sufficient statistics $N_{1|a}, N_{1|b}$. This means that for each prior $W \in \mathcal{W}(\Theta_1)$, the Bayes marginal

P_W can be identified with a vector of $M_n := (n_a+1)\cdot(n_b+1)$ real-valued components. Every such P_W can also be written as a mixture of P_θ's for $\theta = (\mu_{a|1}, \mu_{b|1}) \in \Theta_1$, a convex set. By Carathéodory's theorem we need at most M_n components to describe an arbitrary P_W.

5.F Motivation for use of KL to define GROW sets

If there is more than a single parameter of interest, then a natural (but certainly not the only reasonable!) divergence measure to use in (5.20) is to set d equal to the KL divergence $D(\theta_1\|\Theta_0) := \inf_{\theta_0\in\Theta_0} D(\theta_1\|\theta_0)$.

To see why, note that ε indicates the easiness of testing $\Theta(\varepsilon)$ vs. Θ_0: the larger ε, the 'further' $\Theta(\varepsilon)$ from Θ_0 and the larger the value of $\mathrm{GR}(\varepsilon)$. The KL divergence is the *only divergence measure* in which 'easiness' of testing $\Theta(\varepsilon)$ is consistent with easiness of testing individual elements of Θ_1. By this we mean the following: suppose there exist $\theta_1, \theta_1' \in \Theta_1$ with $\theta_1 \neq \theta_1'$ achieving equal growth rates $\mathrm{GR}(\{\theta_1'\}) = \mathrm{GR}(\{\theta_1\})$ in the tests of the individual point hypotheses $\{\theta_1\}$ vs Θ_0 and $\{\theta_1'\}$ vs. Θ_0 Then if d is *not* the KL it can happen that, for some $\varepsilon > 0$, $\theta_1 \in \Theta(\varepsilon)$ yet $\theta_2' \notin \Theta(\varepsilon)$. With d equal to KL this is impossible. This follows immediately from Theorem 5.4, Part 1, which tells us $D(\theta_1\|\Theta_0) = \mathrm{GR}(\{\theta_1\})$.

Chapter 6

Safe-Bayesian generalized linear regression

Abstract

We study generalized Bayesian inference under misspecification, i.e. when the model is 'wrong but useful'. Generalized Bayes equips the likelihood with a learning rate η. We show that for generalized linear models (GLMs), η-generalized Bayes concentrates around the best approximation of the truth within the model for specific $\eta \neq 1$, even under severely misspecified noise, as long as the tails of the true distribution are exponential. We derive MCMC samplers for generalized Bayesian lasso and logistic regression and give examples of both simulated and real-world data in which generalized Bayes substantially outperforms standard Bayes.

6.1 Introduction

Over the last ten years it has become abundantly clear that Bayesian inference can behave quite badly under misspecification, i.e., if the model $\mathcal{F}$ under consideration is 'wrong but useful' (Grünwald and Langford, 2007; Erven, Grünwald and Rooij, 2007; Müller, 2013; Syring and Martin, 2017; Yao et al., 2018; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017). For example, Grünwald and Langford (2007) exhibit a simple nonparametric classification setting in which, even though the prior puts positive mass on the unique distribution in $\mathcal{F}$ that is closest in KL divergence to the data generating distribution P, the posterior never concentrates around this distribution. Grünwald and Van Ommen (2017) give a simple misspecified setting in which standard Bayesian ridge regression, model selection and model averaging severely overfit small-sample data.

Grünwald and Van Ommen (2017) also propose a remedy for this problem: equip the likelihood with an exponent or *learning rate* η (see (6.1) below). Such a *generalized Bayesian* (also known as *fractional* or *tempered* Bayesian) approach was considered earlier by e.g. Barron and Cover,

1991; Walker and Hjort, 2002; Zhang, 2006b. In practice, η will usually (but not always — see Section 6.5.1 below) be chosen smaller than one, making the prior have a stronger regularizing influence. Grünwald and Van Ommen (2017) show that for Bayesian ridge regression and model selection/averaging, this results in excellent performance, being competitive with standard Bayes if the model is correct and very significantly outperforming standard Bayes if it is not. Extending Zhang's (2006a; 2006b) earlier work, Grünwald and Mehta (2019) (GM from now on) show that, under what was earlier called the $\overline{\eta}$-*central condition* (Definition 6.1 below), generalized Bayes with a specific finite learning rate $\overline{\eta}$ (usually $\neq 1$) will indeed concentrate in the neighborhood of the 'best' $f \in \mathcal{F}$ with high probability. Here, the 'best' f means the one closest in KL divergence to P.

Yet, three important parts of the story are missing in this existing work: (1) Can Grünwald-Van Ommen-type examples, showing failure of standard Bayes ($\eta = 1$) and empirical success of generalized Bayes with the right η, be given more generally, for different priors π (say of lasso-type ($\pi(f) \propto \exp(-\lambda \|f\|_1)$) rather than ridge-type $\pi(f) \propto \exp(-\lambda \|f\|_2^2)$), and for different models, say for *generalized* linear models (GLMs)? (2) Can we find examples of generalized Bayes outperforming standard Bayes with real-world data rather than with toy problems such as those considered by Grünwald and Van Ommen? (3) Does the central condition — which allows for good theoretical behavior of generalized Bayes — hold for GLMs, under reasonable further conditions?

We answer all three questions in the affirmative: in Section 6.2.1 below, we give (a) a toy example on which the Bayesian lasso and the Horseshoe estimator fail; later in the chapter, in Section 6.5 we also (b) give a toy example on which standard Bayes logistic regression fails, and (c) two real-world data sets on which Bayesian lasso and Horseshoe regression fail; in all cases, (d) generalized Bayes with the right η shows much better performance. In Section 6.3, we show (e) that for GLMs, even if the noise is severely misspecified, as long as the distribution of the predictor variable Y has exponentially small tails (which is automatically the case in classification, where the domain of Y is finite), the central condition holds for some $\eta > 0$. In combination with (e), GM's existing theoretical results suggest that generalized Bayes with this η should lead to good results — this is corroborated by our experimental results in Section 6.5. These findings are not obvious: one might for example think that the sparsity-inducing prior used by Bayesian lasso regression circumvents the need for the additional regularization induced by taking an $\eta < 1$, especially since in the original setting of Grünwald and Van Ommen, the standard Bayesian lasso ($\eta = 1$) succeeds. Yet, Example 6.1 below shows that under a modification of their example, it fails after all. In order to demonstrate the failure of standard Bayes and the success of generalized Bayes, we devise (in Section 6.4) MCMC algorithms (f) for generalized Bayes posterior sampling for Bayesian lasso and logistic regression. (a)-(f) are all novel contributions.

In Section 6.2 we first define our setting more precisely. Section 6.2.1) gives a first example of bad standard-Bayesian behavior and Section 6.2.2) recalls a theorem from GM indicating that under the $\overline{\eta}$-central condition, generalized Bayes for $\eta < \overline{\eta}$ should perform well. We

present our new theoretical results in Section 6.3. We next (Section 6.4), present our algorithms for generalized Bayesian posterior sampling, and we continue (Section 6.5) to empirically demonstrate how generalized Bayes outperforms standard Bayes under misspecification. All proofs are in Appendix 6.A.

6.2 The setting

A *learning problem* can be characterized by a tuple $(P, \ell, \mathcal{F})$, where $\mathcal{F}$ is a set of predictors, also referred to as a *model*, P is a distribution on sample space $\mathcal{Z}$, and $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ is a loss function. We denote by $\ell_f(z) \coloneqq \ell(f, z)$ the loss of predictor $f \in \mathcal{F}$ under outcome $z \in \mathcal{Z}$. If $Z \sim P$, we abbreviate $\ell_f(Z)$ to ℓ_f. In all our examples, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We obtain e.g. standard (random-design) regression with squared loss by taking $\mathcal{Y} = \mathbb{R}$ and $\mathcal{F}$ to be some subset of the class of all functions $f : \mathcal{X} \to \mathbb{R}$ and, for $z = (x, y)$, $\ell_f(x, y) = (y - f(x))^2$; logistic regression is obtained by taking $\mathcal{F}$ as before, $\mathcal{Y} = \{-1, 1\}$ and $\ell_f(x, y) = \log(1 + \exp(-y f(x))$. We get conditional density estimation by taking $\{p_f(Y \mid X) : f \in \mathcal{F}\}$ to be a family of conditional probability mass or density functions (defined relative to some measure μ), extended to n outcomes by the i.i.d. assumption, and taking conditional log-loss $\ell_f(x, y) \coloneqq -\log p_f(y \mid x)$.

We are given an i.i.d. sample $Z^n \coloneqq Z_1, Z_2, \ldots, Z_n \sim P$ where each Z_i takes values in $\mathcal{Z}$, and we consider, as our learning algorithm, the *generalized Bayesian posterior*, also known as the *Gibbs posterior*, Π_n on $\mathcal{F}$, defined by its density

$$\pi_n(f) \coloneqq \frac{\exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f)}{\int_{\mathcal{F}} \exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f) \mathrm{d}\rho(f)}, \tag{6.1}$$

where $\eta > 0$ is the *learning rate*, and π_0 is the density of some prior distribution Π_0 on $\mathcal{F}$ relative to an underlying measure ρ. Note that, in the conditional log-loss setting, we get that

$$\pi_n(f) \propto \prod_{i=1}^n (p_f(y_i \mid x_i))^\eta \pi_0(f), \tag{6.2}$$

which, if $\eta = 1$, reduces to standard Bayesian inference. While GM's result (quoted as Theorem 6.1 below) works for arbitrary loss functions, Theorem 6.2 and our empirical simulations (this chapter's new results) revolve around (generalized) linear models. For these models, (6.1) can be equivalently interpreted either in terms of the original loss functions ℓ_f or in terms of the conditional likelihood p_f. For example, consider regression with $\ell_f(x, y) = (y - f(x))^2$ and fixed η. Then (6.1) induces the same posterior distribution $\pi_n(f)$ over $\mathcal{F}$ as does (6.2) with the conditional distributions $p_f(y|x) \propto \exp(-(y - f(x))^2$, which is again the same as (6.1) with ℓ_f replaced by the conditional log-loss $\ell'_f(x, y) \coloneqq -\log p_f(y|x)$, giving a likelihood corresponding to Gaussian errors with a particular fixed variance; an analogous statement holds for logistic regression. Thus, all our examples can be interpreted in terms of (6.2) for a model that is misspecified, i.e., the density of $P(Y|X)$ is not equal to p_f for any $f \in \mathcal{F}$. As is customary (see e.g. Bartlett, Bousquet and Mendelson (2005)), we assume throughout that there exists an optimal $f^* \in \mathcal{F}$ that achieves the smallest *risk* (expected loss) $\mathbb{E}[\ell_{f^*}(Z)] = \inf_{f \in \mathcal{F}} \mathbb{E}[\ell_f(Z)]$. If $\mathcal{F}$ is a GLM, the risk minimizer again has additional interpretations: first, f^* minimizes,

among all $f \in \mathcal{F}$, the conditional KL divergence $\mathbf{E}_{(X,Y)\sim P}[\log\big(p(Y|X)/p_f(Y|X)\big)]$ to the true distribution P. Second, if there is an $f \in \mathcal{F}$ with $\mathbf{E}_{X,Y\sim P}[Y \mid X] = f(X)$ (i.e. $\mathcal{F}$ contains the *true regression function*, or equivalently, *true conditional mean*), then the risk minimizer satisfies $f^* = f$.

6.2.1 Bad Behavior of Standard Bayes

Example 6.1. We consider a Bayesian lasso regression setting (Park and Casella, 2008) with random design, with a Fourier basis. We sample data $Z_i = (X_i, Y_i)$ i.i.d. $\sim P$, where P is defined as follows: we first sample *preliminary* (X_i', Y_i') with $X_i' \overset{i.i.d.}{\sim}$ Uniform$([-1,1])$; the dependent variable Y_i' is set to $Y_i' = 0 + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for some fixed value of σ, independently of X_i'. In other words: the true distribution for (X_i', Y_i') is 'zero with Gaussian noise'. Now we toss a fair coin for each i. If the coin lands heads, we set the actual $(X_i, Y_i) := (X_i', Y_i')$, i.e. we keep the (X_i', Y_i') as they are, and if the coin lands tails, we put the pair to zero: $(X_i, Y_i) := (0, 0)$.

We model the relationship between X and Y with a p^{th} order Fourier basis. Thus, $\mathcal{F} = \{f_\beta : \beta \in \mathbb{R}^{2p+1}\}$, with $f_\beta(x)$ given by

$$\left\langle \beta, \frac{1}{\pi} \cdot \left(2^{-1/2}, \cos(x), \sin(x), \cos(2x), \ldots, \sin(px)\right)\right\rangle,$$

and the η-posterior is defined by (6.1) with $\ell_{f_\beta}(x, y) = (y - f_\beta(x))^2$; the prior is the Bayesian lasso prior whose definition we recall in Section 6.4.1. Since our 'true' regression function $\mathbf{E}[Y_i \mid X_i]$ is 0, in an actual sample around 50% of points will be noiseless, *easy* points, lying on the true regression function. Since the actual sample of (X_i, Y_i) has less noise then the original sample (X_i', Y_i'), we would expect Bayesian lasso regression to learn the correct regression function, but as we see in the blue line in Figure 6.1, it overfits and learns the noise instead (later on (Figure 6.3 in Section 6.5.1) we shall see that, not surprisingly, this results in terrible predictive behavior). By removing the noise in half the data points, we misspecified the model: we made the noise heteroscedastic, whereas the model assumes homoscedastic noise. Thus, in this experiment the *model is wrong*. Still, the distribution in $\mathcal{F}$ closest to the true P, both in KL divergence and in terms of minimizing the squared error risk, is given by the conditional distribution corresponding to $Y_i = 0 + \varepsilon_i$, where ε_i is i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. While this element of $\mathcal{F}$ is in fact *favored* by the prior (the lasso prior prefers β with small $\|\beta\|_1$), nevertheless, for small samples, the standard Bayesian posterior puts most if its mass at f with many nonzero coefficients. In contrast, the generalized posterior (6.1) with $\eta = 0.25$ gives excellent results here. To learn this η from the data, we can use the Safe-Bayesian algorithm of Grünwald (2012). The result is depicted as the red line in Figure 6.1. Implementation details are in Section 6.4.1 and Appendix 6.D; the details of the figure are in Appendix 6.E.

The example is similar to that of Grünwald and Van Ommen (2017), who use multidimensional X and a ridge (normal) prior on $\|\beta\|$; in their example, standard Bayes succeeds when equipped with a lasso prior; by using a trigonometric basis we can make it 'fail' after all. Grünwald and Van Ommen (2017) relate the potential for the overfitting-type of behavior of standard Bayes, as well as the potential for full inconsistency (i.e. even holding as $n \to \infty$) as noted by Grünwald

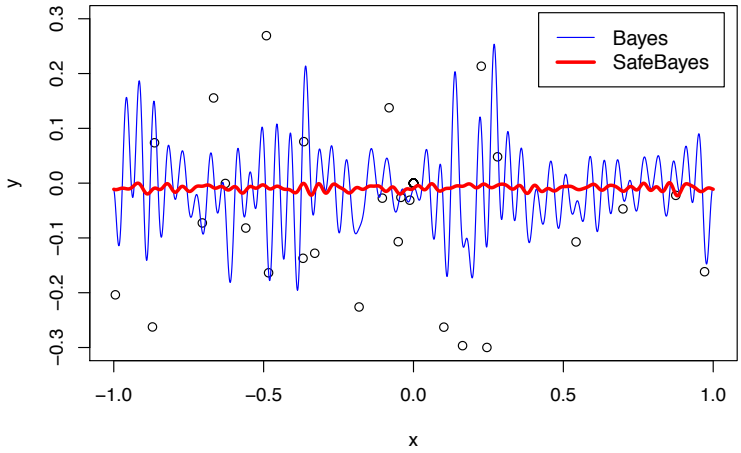


Figure 6.1: *Predictions of standard Bayes (blue) and SafeBayes (red),* $n = 50$, $p = 101$.

and Langford (2007) to properties of the Bayesian predictive distribution

$$\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n) := \int_{\mathcal{F}} p_f(Y_{n+1} \mid X_{n+1}) \pi_n(f \mid Z^n) \mathrm{d}\rho(f).$$

Being a mixture of $f \in \mathcal{F}$, $\overline{p}(Y_{n+1} \mid X_{n+1})$, is a member of the convex hull of densities $\mathcal{F}$ but not necessarily of $\mathcal{F}$ itself. As explained by Grünwald and Van Ommen, severe overfitting may take place if $\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ is 'far' from any of the distributions in $\mathcal{F}$. It turns out that this is exactly what happens in the lasso example above, as we see from Figure 6.2 (details in Appendix 6.E). This figure plots the data points as $(X_i, 0)$ to indicate their location; we see that the predictive variance of standard Bayes fluctuates, being small around the data points and large elsewhere. However, it is obvious that for every density p_f in our model $\mathcal{F}$, the variance is fixed independently of X, and thus $\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ is indeed very far from any particular p_f with $f \in \mathcal{F}$. In contrast, for the generalized Bayesian lasso with $\eta = 0.25$, the corresponding predictive variance is almost constant; thus, at the level $\eta = 0.25$ the predictive distribution is almost 'in-model' (in machine learning terminology, we may say that $\overline{p}$ is 'proper' (Shalev-Shwartz and Ben-David, 2014), and the overfitting behavior then does not occur anymore.

6.2.2 When Generalized Bayes Concentrates

Having just seen bad behavior for $\eta = 1$, we now recall some results from GM. Under some conditions, GM show that generalized Bayes, for appropriately chosen η, does concentrate at fast rates even under misspecification. We first recall (a very special case of) the asymptotic behavior under misspecification theorem of GM. GM bound (a) the *misspecification metric* $d_{\overline{\eta}}$

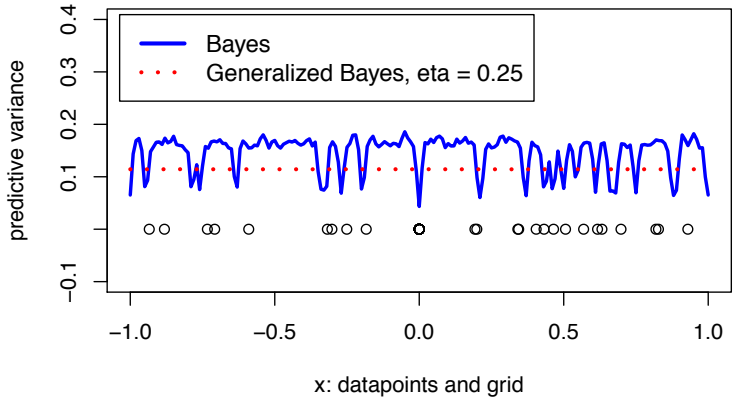


Figure 6.2: *Variance of Predictive Distribution* $\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ *for a single run with* $n = 50$.

in terms of (b) the *information complexity*. The bound (c) holds under a simple condition on the learning problem that was termed the *central condition* by Van Erven et al. (2015). Before presenting the theorem we explain (a)–(c). As to (a), we define the *misspecification metric* $d_{\overline{\eta}}$ in terms of its square by

$$d^2_{\overline{\eta}}(f, f') := \frac{2}{\overline{\eta}} \left(1 - \int \sqrt{p_{f,\overline{\eta}}(z) p_{f',\overline{\eta}}(z)} \mathrm{d}\mu(z)\right)$$

which is the $(2/\overline{\eta}$-scaled) squared Hellinger distance between $p_{f,\overline{\eta}}$ and $p_{f',\overline{\eta}}$. Here, a density $p_{f,\overline{\eta}}$ is defined as

$$p_{f,\overline{\eta}}(z) := p(z) \frac{\exp(-\overline{\eta} L_f(z))}{\mathbb{E}[\exp(-\overline{\eta} L_f(Z))]},$$

where $L_f = \ell_f - \ell_{f^*}$ is the *excess loss* of f. GM show that $d_{\overline{\eta}}$ defines a metric for all $\overline{\eta} > 0$. If $\overline{\eta} = 1$, ℓ is log-loss, and the model is well-specified, then it is straightforward to verify that $p_{f,\overline{\eta}} = p_f$, and so $(1/2) \cdot d_{\overline{\eta}}$ becomes the standard squared Hellinger distance.

As to (b), we denote by $\mathrm{IC}_{n,\eta}(\Pi_0)$ the information complexity, defined as:

$$\mathrm{IC}_{n,\eta}(\Pi_0) := \mathbf{E}_{\underline{f} \sim \Pi_n} \left[\frac{1}{n} \sum_{i=1}^n L_{\underline{f}}(Z_i)\right] + \frac{KL(\Pi_n \,\|\, \Pi_0)}{\eta \cdot n} =$$
$$-\frac{1}{\eta n} \log \int_{\mathcal{F}} \pi_0(f) e^{-\eta \sum_{i=1}^n \ell_f(Z_i)} \mathrm{d}\rho(f) - \sum_{i=1}^n \ell_{f^*}(Z_i), \tag{6.3}$$

where $\underline{f}$ denotes the predictor sampled from the posterior Π_n and KL denotes KL divergence; we suppress dependency of IC on f^* in the notation. The fact that both lines above are equal

(noticed by, among others, Zhang (2006b); GM give an explicit proof) allows us to write the information complexity in terms of a generalized Bayesian predictive density which is also known as *extended stochastic complexity* (Yamanishi, 1998). It also plays a central role in the field of prediction with expert advice as the *mix-loss* (Van Erven et al., 2015; Cesa-Bianchi and Lugosi, 2006) and coincides with the minus log of the standard Bayesian predictive density if $\eta = 1$ and ℓ is log-loss. It can be thought of as a complexity measure analogous to VC dimension and Rademacher complexity.

As to (c), GM's result holds under the *central condition* ((Li, 1999); name due to Van Erven et al., 2015) which expresses that, for some fixed $\overline{\eta} > 0$, for all fixed f, the probability that the loss of f exceeds that of the optimal f^* by $a/\overline{\eta}$ is exponentially small in a:

Definition 6.1 (Central Condition, Def. 7 of GM)**.** Let $\overline{\eta} > 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the *$\overline{\eta}$-strong central condition* if, for all $f \in \mathcal{F}$: $\mathbb{E}\left[e^{-\overline{\eta} L_f}\right] \leq 1$.

As straightforward rewriting shows, this condition holds *automatically*, for any $\overline{\eta} \leq 1$ in the density estimation setting, if the model is correct; Van Erven et al. (2015) provide some other cases in which it holds, and show that many other conditions on ℓ and P that allow fast rate convergence that have been considered before in the statistical and on-line learning literature, such as *exp-concavity* (Cesa-Bianchi and Lugosi, 2006), the *Tsybakov* and *Bernstein* conditions (Bartlett, Bousquet and Mendelson, 2005; Tsybakov, 2004) and several others, can be viewed as special cases of the central condition; yet they don't discuss GLMs. Here is GM's result:

Theorem 6.1 (Theorem 10 from GM)**.** *Suppose that the $\overline{\eta}$-strong central condition holds. Then for any $0 < \eta < \overline{\eta}$, the metric $d_{\overline{\eta}}$ satisfies*

$$\mathbb{E}_{Z^n \sim P} \mathbb{E}_{\underline{f} \sim \Pi_n}\left[d^2_{\overline{\eta}}(f^*, \underline{f})\right] \leq C_\eta \cdot \mathbb{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}(\Pi_0)\right]$$

with $C_\eta = \eta/(\overline{\eta} - \eta)$. In particular, $C_\eta < \infty$ for $0 < \eta < \overline{\eta}$, and $C_\eta = 1$ for $\eta = \overline{\eta}/2$.

Thus, we expect the posterior to concentrate at a rate dictated by $\mathbf{E}[\mathrm{IC}_{n,\eta}]$ in neighborhoods of the best (risk-minimizing, KL optimal, or even true regression function) f^*. The misspecification metric $d^2_{\overline{\eta}}$ on the left hand side is a weak metric, however, in Appendix 6.B we show that we can replace it by stronger notions such as KL-divergence, squared error or logistic loss. Theorem 6.1 generalizes previous results (e.g. Zhang (2006a) and Zhang (2006b)) to the misspecified setting. In the well-specified case, Zhang, as well as several other authors (Walker and Hjort, 2002; Martin, Mess and Walker, 2017), state a result that holds for any $\eta < 1$ but not $\eta = 1$. This suggests that there is an advantage to taking η slightly smaller than one even when the model is well-specified (for more details see Zhang (2006a)).

To make the theorem work for GLMs under misspecification, we must verify (a) that the central condition still holds (which is in general not guaranteed) and that (b) the information complexity is sufficiently small. As to (a), in the following section we show that the central condition holds (with $\overline{\eta}$ usually $\neq 1$) for 1-dimensional exponential families and high-dimensional generalized linear models (GLMs) if the noise is misspecified, as long as P has exponentially small tails; in particular, we relate $\overline{\eta}$ to the variance of P. As to (b), if the model is correct (the conditional distribution $P(Y \mid X)$ has density f equal to p_f with $f \in \mathcal{F}$), where $\mathcal{F}$ represents

a d-dimensional GLM, then it is known (see e.g. Zhang (2006b)) that, for any prior Π_0 with continuous, strictly positive density on $\mathcal{F}$, the information complexity satisfies

$$\mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}(\Pi_0)\right] = O\left(\frac{d}{n} \cdot \log n\right), \tag{6.4}$$

which leads to bounds within a log-factor of the minimax optimal rate (among all possible estimators, Bayesian or not), which is $O(d/n)$. While such results were only known for the well-specified case, in Proposition 3 below we show that, for GLMs, they continue to hold for the misspecified case.

6.3 Generalized GLM Bayes

Below we first show that the central condition holds for natural univariate exponential families; we then extend this result to the GLM case, and establish bounds in information complexity of GLMs. Let the class $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ be a univariate natural exponential family of distributions on $\mathcal{Z} = \mathcal{Y}$, represented by their densities, indexed by natural parameter $\theta \in \Theta \subset \mathbb{R}$ (Barndorff-Nielsen, 1978). The elements of this restricted family have probability density functions

$$p_\theta(y) := \exp(\theta y - F(\theta) + r(y)), \tag{6.5}$$

for log-normalizer F and carrier measure r. We denote the corresponding distribution as P_θ. In the first part of the theorem below we assume that Θ is restricted to an arbitrary closed interval $[\underline{\theta}, \overline{\theta}]$ with $\underline{\theta} < \overline{\theta}$ that resides in the interior of the natural parameter space $\overline{\Theta} = \{\theta : F(\theta) < \infty\}$. Such Θ allow for a simplified analysis because within Θ the log-normalizer F as well as all its derivatives are uniformly bounded from above and below; see (6.7) in Appendix 4.B. As is well-known (see e.g. Barndorff-Nielsen (1978)), exponential families can equivalently be parameterized in terms of the mean-value parameterization: there exists a 1-to-1 strictly increasing function $\mu : \overline{\Theta} \to \mathbb{R}$ such that $\mathbb{E}_{Y\sim P_\theta}[Y] = \mu(\theta)$. As is also well-known, the density $p_{f^*} \equiv p_{\theta^*}$ within $\mathcal{F}$ minimizing KL divergence to the true distribution P satisfies $\mu(\theta^*) = \mathbf{E}_{Y\sim P}[Y]$, whenever the latter quantity is contained in $\mu(\Theta)$ (Grünwald, 2007). In words, the best approximation to P in $\mathcal{F}$ in terms of KL divergence has the same mean of Y as P.

Theorem 6.2. *Consider a learning problem $(P, \ell, \mathcal{F})$ with $\ell_\theta(y) = -\log p_\theta(y)$ the log loss and $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ a univariate exponential family as above.*
(1). Suppose that $\Theta = [\underline{\theta}, \overline{\theta}]$ is compact as above and that $\theta^* = \arg\min_{\theta\in\overline{\Theta}} D(P\|P_\theta)$ lies in Θ. Let $\sigma^2 > 0$ be the true variance $\mathbb{E}_{Y\sim P}(Y - \mathbb{E}[Y])^2$ and let $(\sigma^*)^2$ be the variance $\mathbb{E}_{Y\sim P_{\theta^*}}(Y - \mathbb{E}[Y])^2$ according to θ^*. Then

- *(i) for all $\overline{\eta} > (\sigma^*)^2/\sigma^2$, the $\overline{\eta}$-central condition does* not *hold.*
- *(ii) Suppose there exists $\eta^\circ > 0$ such that $\overline{C} := \mathbb{E}_P[\exp(\eta^\circ|Y|)] < \infty$. Then there exists $\overline{\eta} > 0$, depending only on η°, $\overline{C}$, $\underline{\theta}$ and $\overline{\theta}$ such that the $\overline{\eta}$-central condition holds. Moreover,*
- *(iii), for all $\delta > 0$, there is an $\varepsilon > 0$ such that, for all $\overline{\eta} \le (\sigma^*)^2/\sigma^2 - \delta$, the $\overline{\eta}$-central condition holds relative to the restricted model $\mathcal{F}_\varepsilon = \{p_\theta : \theta \in [\theta^* - \varepsilon, \theta^* + \varepsilon]\}$.*

(2). Suppose that P is Gaussian with variance $\sigma^2 > 0$ and that $\mathcal{F}$ indexes a full Gaussian location family. Then the $\overline{\eta}$-central condition holds iff $\overline{\eta} \leq (\sigma^*)^2/\sigma^2$.

We provide (iii) just to give insight — 'locally', i.e. in restricted models that are small neighborhoods around the best-approximating θ^*, the smallest $\overline{\eta}$ for which the central condition holds is determined by a ratio of variances. The final part shows that for the Gaussian family, the same holds not just locally but globally (note that we do not make the compactness assumption on Θ there); we warn the reader though that the standard posterior ($\eta = 1$) based on a model with fixed variance σ^* is quite different from the generalized posterior with $\eta = (\sigma^*)^2/\sigma^2$ and a model with variance σ^2 (Grünwald and Van Ommen, 2017). Finally, while in practical cases we often find $\overline{\eta} < 1$ (suggesting that Bayes may only succeed if we learn 'slower' than with the standard $\eta = 1$, i.e. the prior becomes more important), the result shows that we can also very well have $\overline{\eta} > 1$; we give a practical example at the end of Section 6.5. Theorem 6.2 is new and supplements Van Erven et al.'s (2015) various examples of $\mathcal{F}$ which satisfy the central condition. In the theorem we require that both tails of Y have exponentially small probability.

Central Condition: GLMs Let $\mathcal{F}$ be the generalized linear model (McCullagh and Nelder, 1989) (GLM) indexed by parameter $\beta \in \mathcal{B} \subset \mathbb{R}^d$ with link function $g : \mathbb{R} \to \mathbb{R}$. By definition this means that there exists a set $\mathcal{X} \subset \mathbb{R}^d$ and a univariate exponential family $\mathcal{Q} = \{p_\theta : \theta \in \overline{\Theta}\}$ on $\mathcal{Y}$ of the form (6.5) such that the conditional distribution of Y given $X = x$ is, for all possible values of $x \in \mathcal{X}$, a member of the family $\mathcal{Q}$, with mean-value parameter $g^{-1}(\langle \beta, x \rangle)$. Then the class $\mathcal{F}$ can be written as $\mathcal{F} = \{p_\beta : \beta \in \mathcal{B}\}$, a set of conditional probability density functions such that

$$p_\beta(y \mid x) := \exp\big(\theta_x(\beta) y - F(\theta_x(\beta)) + r(y)\big), \tag{6.6}$$

where $\theta_x(\beta) := \mu^{-1}(g^{-1}(\langle \beta, x \rangle))$, and μ^{-1}, the inverse of μ defined above, sends mean parameters to natural parameters. We then have $\mathbb{E}_{P_\beta}[Y \mid X] = g^{-1}(\langle \beta, X \rangle)$, as required.

Proposition 3. *Under the following three assumptions, the learning problem $(P, \ell, \mathcal{F})$ with $\mathcal{F}$ as above satisfies the $\overline{\eta}$-central condition for some $\overline{\eta} > 0$ depending only on the parameters of the problem:*

1. *(Conditions on g): the inverse link function g^{-1} has bounded derivative on the domain $\mathcal{B} \times \mathcal{X}$, and the image of the inverse link on the same domain is a bounded interval in the interior of the mean-value parameter space $\{\mu \in \mathbb{R} : \mu = \mathbb{E}_{Y \sim q}[Y] \ : \ q \in \mathcal{Q}\}$ (for all standard link functions, this can be enforced by restricting $\mathcal{B}$ and $\mathcal{X}$ to an (arbitrarily large but still) compact domain).*
2. *(Condition on 'true' P): for some $\eta > 0$ we have $\sup_{x \in \mathcal{X}} \mathbb{E}_{Y \sim P}[\exp(\eta |Y|) \mid X = x] < \infty$.*
3. *(Well-specification of conditional mean): there exists $\beta^\circ \in \mathcal{B}$ such that $\mathbb{E}[Y \mid X] = g^{-1}(\langle \beta^\circ, X \rangle)$.*

A simple argument (differentiation with respect to β) shows that under the third condition, it must be the case that $\beta^\circ = \beta^*$, where $\beta^* \in \mathcal{B}$ is the index corresponding to the density $p_{f^*} \equiv p_{\beta^*}$ within $\mathcal{F}$ that minimizes KL divergence to the true distribution P. Thus, our conditions imply that $\mathcal{F}$ contains a β^* which correctly captures the conditional mean (and this will then be the

risk minimizer); thus, as is indeed the case in Example 6.1, the regression function must be well-specified but the noise can be severely misspecified.

We stress that the three conditions have very different statuses. The first is mathematically convenient; it can be enforced by truncating parameters and data, which is awkward but may not lead to substantial deterioration in practice. Whether it is even really needed or not is not clear (and may in fact depend on the chosen exponential family). The second condition is really necessary — as can immediately be seen from Definition 6.1, the strong central condition cannot hold if Y has polynomial tails and for some f and x, $\ell_f(x, Y)$ increases polynomially in Y (in Section 6 of their paper, GM consider weakenings of the central condition that still work in such situations). For the third condition, however, we suspect that there are many cases in which it does not hold yet still the strong central condition holds; so then the GM convergence result would still be applicable under 'full misspecification'; investigating this will be the subject of future work.

GLM Information Complexity To apply Theorem 6.1 to get convergence bounds for exponential families and GLMs, we need to verify that the central condition holds (which we just did) and we need to bound the information complexity, which we proceed to do now. It turns out that the bound on $\mathrm{IC}_{n,\eta}$ of $O((d/n)\log n)$ of (6.4) continues to hold unchanged under misspecification, as is an immediate corollary of applying the following proposition to the definition of $\mathrm{IC}_{n,\eta}$ given above (6.3):

Proposition 4. *Let $(P, \ell, \mathcal{F})$ be a learning problem with $\mathcal{F}$ a GLM satisfying Conditions 1–3 above. Then for all $f \in \mathcal{F}$, $\mathbb{E}_{X,Y\sim P}[L_f] = \mathbb{E}_{X,Y\sim P_{f^*}}[L_f]$.*

This result follows almost immediately from the 'robustness property of exponential families' (Chapter 19 of Grünwald (2007)); for convenience we provide a proof in Appendix 4.B. The result implies that any bound in $\mathrm{IC}_{n,\eta}(\Pi_0)$ for a particular prior in the well-specified GLM case, in particular (6.4), immediately transfers to the same bound for the misspecified case, as long as our regularity conditions hold, allowing us to apply Theorem 6.1 to obtain the parametric rate for GLMs under misspecification.

6.4 MCMC Sampling

Below we devise MCMC algorithms for obtaining samples from the η-generalized posterior distribution for two problems: regression and classification. In the regression context we consider one of the most commonly used sparse parameter estimation techniques, the lasso. For classification we use the logistic regression model. In our experiments in Section 6.5, we compare the performance of generalized Bayesian lasso with Horseshoe regression (Carvalho, Polson and Scott, 2010). The derivations of samplers are given in Appendix 6.D.

6.4.1 Bayesian lasso regression

Consider the regression model $Y = X\beta + \varepsilon$, where $\beta \in \mathbb{R}^p$ is the vector of parameters of interest, $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is a noise vector. The Least Absolute Shrinkage

and Selection Operator (LASSO) of Tibshirani (1996) is a regularization method used in regression problems for shrinkage and selection of features. The lasso estimator is defined as $\widehat{\beta}_{\text{lasso}} \coloneqq \arg\min_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$, where $\|\cdot\|_1, \|\cdot\|_2$ are l_1 and l_2 norms correspondingly. It can be interpreted as a Bayesian posterior mode (MAP) estimate when the priors on β are given by independent Laplace distributions. As discovered by Park and Casella (2008), the same posterior on β is also obtained by the following Gibbs sampling scheme: set $\eta = 1$ and denote $D_\tau \coloneqq \operatorname{diag}(\tau_1, \ldots, \tau_n)$. Also, let $a \coloneqq \frac{\eta}{2}(n-1) + \frac{p}{2} + \alpha$ and $b_\tau \coloneqq \frac{\eta}{2}(Y - X\beta)^T(Y - X\beta) + \frac{1}{2}\beta^T {D_\tau}^{-1}\beta + \gamma$, where $\alpha, \gamma > 0$ are hyperparameters. Then the Gibbs sampler is constructed as follows.

$$\begin{aligned}
\beta &\sim \mathcal{N}\left(\eta M_\tau X^T Y, \sigma^2 M_\tau\right), \\
\sigma^2 &\sim \text{Inv-Gamma}\left(a, b_\tau\right), \\
\tau_j^{-2} &\sim \text{IG}\left(\sqrt{\lambda^2\sigma^2/\beta_j^2}, \lambda^2\right),
\end{aligned}$$

where IG is the inverse Gaussian distribution and $M_\tau \coloneqq (\eta X^T X + {D_\tau}^{-1})^{-1}$. Following Park and Casella (2008), we put a Gamma prior on the shrinkage parameter λ. Now, in their paper Park and Casella only give the scheme for $\eta = 1$, but, as is straightforward to derive from their paper, the scheme above actually gives the *η-generalized* posterior corresponding to the lasso prior for general η (more details in Appendix 6.D). We will use the Safe-Bayesian algorithm for choosing the optimal η developed by Grünwald and Van Ommen (2017) (see Appendix 6.D.3). The code for Generalized- and Safe-Bayesian lasso regression can be found in the CRAN R-package 'SafeBayes' (De Heide, 2016).

Horseshoe estimator The Horseshoe prior is the state-of-the-art global-local shrinkage prior for tackling high-dimensional regularization, introduced by Carvalho, Polson and Scott (2010). Unlike the Bayesian lasso, it has flat Cauchy-like tails, which allow strong signals to remain unshrunk a posteriori. For completeness we include the horseshoe in our regression comparison, using the implementation of Van der Pas et al. (2016).

6.4.2 Bayesian logistic regression

Consider the standard logistic regression model $\{f_\beta : \beta \in \mathbb{R}^p\}$, the data $Y_1, \ldots, Y_n \in \{0, 1\}$ are independent binary random variables observed at the points $X \coloneqq (X_1, \ldots, X_n) \in \mathbb{R}^{n \times p}$ with

$$P_{f_\beta}(Y_i = 1 \mid X_i) \coloneqq p_{f_\beta}(1 \mid X_i) \coloneqq \frac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}}.$$

The standard Bayesian approach involves putting a Gaussian prior on the parameter $\beta \sim \mathcal{N}(b, B)$ with mean $b \in \mathbb{R}^p$ and the covariance matrix $B \in \mathbb{R}^{p \times p}$. To sample from the η-generalized posterior we modify a Pólya–Gamma latent variable scheme described in Polson, Scott and Windle (2013). We first introduce latent variables $\omega_1, \ldots, \omega_n \in \mathbb{R}$, which will be sampled from Pòlya-Gamma distribution (constructed to yield a simple Gibbs sampler for

Bayesian logistic regression, for more details see Polson, Scott and Windle (2013)). Let

$$\begin{aligned}\Omega &\coloneqq \operatorname{diag}\{\omega_1, \dots, \omega_n\},\\ \kappa &\coloneqq (Y_1 - 1/2, \dots, Y_n - 1/2)^T,\\ V_\omega &\coloneqq (X^T \Omega X + B^{-1})^{-1}, \text{ and}\\ m_\omega &\coloneqq V_\omega(\eta X^T \kappa + B^{-1} b).\end{aligned}$$

Then the Gibbs sampler for η-generalized posterior is given by

$$\omega_i \sim \mathrm{PG}(\eta, X_i^T \beta), \quad \beta \sim \mathcal{N}(m_\omega, V_\omega),$$

where PG is the Pòlya-Gamma distribution.

6.5 Experiments

Below we present the results of experiments that compare the performance of the derived Gibbs samplers with their standard counterparts. More details/experiments are in Appendix 6.E.

6.5.1 Simulated data

Regression In our experiments we focus on prediction, and we run simulations to determine the *square-risk* (expected squared error loss) of our estimate relative to the underlying distribution P: $\mathbb{E}_{(X,Y)\sim P}(Y - X\beta)^2$, where $X\beta$ would be the conditional expectation, and thus the square-risk minimizer, if β would be the true parameter (vector).

Consider the data generated as described in Example 6.1. We study the performance of the η-generalized Bayesian lasso with η chosen by the Safe-Bayesian algorithm (we call it the Safe-Bayesian lasso) in comparison with two popular estimation procedures for this context: the Bayesian lasso (which corresponds to η=1), and the Horseshoe method. In Figure 6.3 the simulated square-risk is plotted as a function of the sample size for all three methods. We average over enough samples so that the graph appears to be smooth (25 iterations for SafeBayes, 1000 for the two standard Bayesian methods). It shows that both the standard Bayesian lasso and the Horseshoe perform significantly worse than the Safe-Bayesian lasso. Moreover we see that the risks for the standard methods initially grows with the sample size (additional experiments not reported here suggest that Bayes will 'recover' at very large n).

Classification We focus on finding coefficients β for prediction, and our error measure is the expected logarithmic loss, which we call *log-risk*: $\mathbb{E}_{(X,Y)\sim P}\left[-\log \mathrm{Li}_\beta(Y \mid X)\right]$, where $\mathrm{Li}_\beta(Y \mid X) \coloneqq e^{YX^T\beta}/(1 + e^{X^T\beta})$. We start with an example that is very similar to the previous one. We generate a $n \times p$ matrix of independent standard normal random variables with $p = 25$. For every feature vector X_i we sample a corresponding $Z_i \sim \mathcal{N}(0, \sigma^2)$, as before, and we misspecify the model by putting approximately half of the Z_i and the corresponding $X_{i,1}$ to zero. Next, we sample the labels $Y_i \sim \mathrm{Binom}(\exp(Z_i)/(1 + \exp(Z_i))$. We compare standard Bayesian logistic regression ($\eta = 1$) to a generalized version ($\eta = 0.125$). In Figure 6.4 we plot

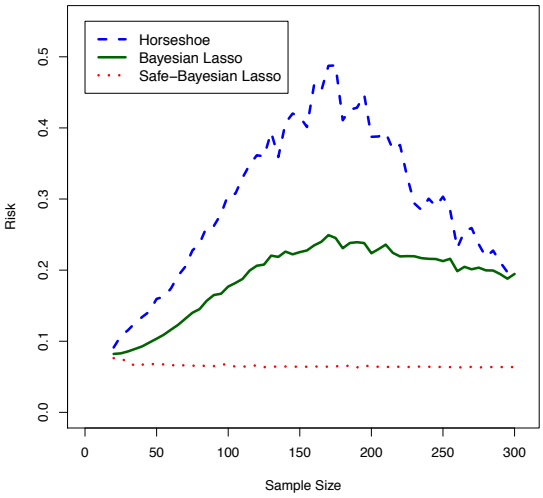


Figure 6.3: *Simulated squared error risk (test error) with respect to P as function of sample size for the* wrong-model *experiments of Section 6.5.1 using the posterior predictive distribution of the standard Bayesian lasso (green, solid), the Safe-Bayesian lasso (red, dotted), both with standard improper priors, and the Horseshoe (blue, dashed); and* 201 *Fourier basis functions.*

the log-risk as a function of the sample size. As in the regression case, the risk for standard Bayesian logistic regression ($\eta = 1$) is substantially worse than the one for generalized Bayes ($\eta = 0.125$). Even for generalized Bayes, the risk initially goes up a little bit, the reason being that the prior is *too good*: it is strongly concentrated around the risk-optimal $\beta^* = 0$. Thus, the first prediction made by the Bayesian predictive distribution coincides with the optimal ($\beta = 0$) prediction, and in the beginning, due to noise in the data, predictions will first get slightly worse. This is a phenomenon that also applies to standard Bayes with well-specified models; see for example Grünwald and Halpern, 2004, Example 3.1.

Even for the well-specified case it can be beneficial to use $\eta \neq 1$. It is easy to see that the maximum *a posteriori* estimate for generalized logistic regression corresponds to the ridge logistic regression method (which penalizes large $\|\beta\|_2$) with the shrinkage parameter $\lambda = \eta^{-1}$. However, when the the prior mean is zero but the risk minimizer β^* is far from zero, penalizing large norms of β is inefficient, and we find that the best performance is achieved with $\eta > 1$.

6.5.2 Real World Data

We present two examples with real world data to demonstrate that bad behavior under misspecification also occurs in practice. For these data sets, we compare the performance of Safe-Bayesian lasso and standard Bayesian lasso. As the first example we consider the data of the daily maximum temperatures at Seattle Airport as a function of the time and date (source: `R`-package `weatherData`, also available at `www.wunderground.com`). A second example is

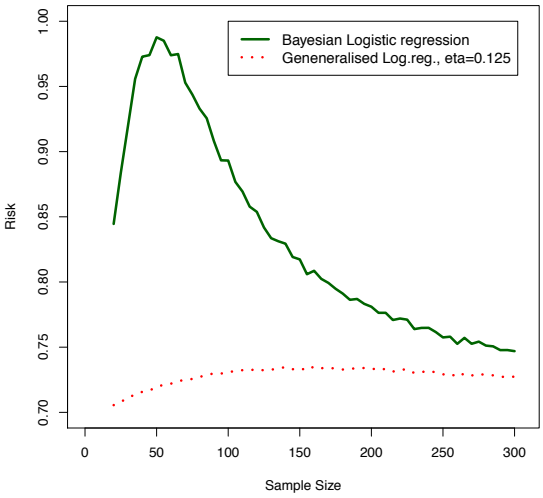


Figure 6.4: *Simulated logistic risk as function of sample size for* wrong-model *experiments of Section 6.5.1 using posterior predictive distribution of standard Bayesian logistic regression (green, solid), and generalized Bayes ($\eta = 0.125$, red, dotted) with 25 noise dimensions.*

	Horse-shoe	Bayesian lasso	SafeBayes lasso
MSE $((°C)^2)$	6.53	6.16	6.04
MSE $((ppm)^2)$	1169	1201	1142

Table 6.1: Mean square errors for predictions on the Seattle and London data sets of Section 6.5.2

London air pollution data (source: `R`-package `Openair`, for more details see Carslaw and Ropkins (2012) and Carslaw (2015)). Here the quantity of interest is the concentration of nitrogen dioxide (NO_2), again as a function of time and date. In both settings we divide the data into a training set and a test set and focus on the prediction error. In both examples, SafeBayes picks an $\widehat{\eta}$ strictly smaller than one. Also, for both data sets the Safe-Bayesian lasso clearly outperforms the standard Bayesian lasso and the Horseshoe in terms of mean square prediction error, as seen from Table 6.1 (details in Appendix 6.E).

6.6 Future work

We provided both theoretical and empirical evidence that η-generalized Bayes can significantly outperform standard Bayes for GLMs. However, the empirical examples are only given for Bayesian lasso linear regression and logistic regression. In future work we would like to devise generalized posterior samplers for other GLMs and speed up the sampler for generalized Bayesian logistic regression, since our current implementation is slow and (unlike our linear

regression implementation) cannot deal with high-dimensional (and thus, real-world) data yet. Furthermore, the Safe-Bayesian algorithm of Grünwald, 2012, used to learn η, enjoys good theoretical performance but is computationally very slow. Since learning η for which the central condition holds (preferably the largest possible value, since small values of η mean slower learning) is essential for using generalized Bayes in practice, there is a necessity for speeding up SafeBayes or finding an alternative. A potential solution might be using cross-validation to learn η, but its theoretical properties (e.g. satisfying the central condition) are yet to be established.

6.A Proofs

6.A.1 Proof of Theorem 6.2

The second part of the theorem about the Gaussian location family is a straightforward calculation, which we omit. As to the first part (Part (i)—(iii)), we will repeatedly use the following fact: for every Θ that is a nonempty compact subset of the interior of $\overline{\Theta}$, in particular for $\Theta = [\underline{\theta}, \overline{\theta}]$ with $\underline{\theta} < \overline{\theta}$ both in the interior of $\overline{\Theta}$, we have:

$$\begin{aligned} -\infty < \inf_{\theta\in\Theta} F(\theta) < \sup_{\theta\in\Theta} F(\theta) < \infty \\ -\infty < \inf_{\theta\in\Theta} F'(\theta) < \sup_{\theta\in\Theta} F'(\theta) < \infty \\ 0 < \inf_{\theta\in\Theta} F''(\theta) < \sup_{\theta\in\Theta} F''(\theta) < \infty. \end{aligned} \tag{6.7}$$

Now, let $\theta, \theta^* \in \Theta$. We can write

$$\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] = \mathbb{E}_{Y\sim P}\left[\left(\frac{p_\theta(Y)}{p_{\theta^*}(Y)}\right)^\eta\right] = \exp\left(-G(\eta(\theta-\theta^*)) + \eta F(\theta^*) - \eta F(\theta)\right). \tag{6.8}$$

where $G(\lambda) = -\log \mathbb{E}_{Y\sim P}\left[\exp(\lambda Y)\right]$. If this quantity is $-\infty$ for all $\eta > 0$, then (i) holds trivially. If not, then (i) is implied by the following statement:

$$\limsup_{\varepsilon\to 0}\left\{\eta : \text{for all } \theta \in [\theta^* - \varepsilon, \theta^* + \varepsilon],\ \mathbb{E}[\exp(\eta L_{p_\theta})] \le 1\right\} = \frac{(\sigma^*)^2}{\sigma^2}. \tag{6.9}$$

Clearly, this statement also implies (iii). To prove (i), (ii) and (iii), it is thus sufficient to prove (ii) and (6.9). We prove both by a second-order Taylor expansion (around θ^*) of the right-hand side of (6.8).

Preliminary Facts. By our assumption there is a $\eta^\circ > 0$ such that $\mathbb{E}[\exp(\eta^\circ|Y|)] = \overline{C} < \infty$. Since $\theta^* \in \Theta = [\underline{\theta}, \overline{\theta}]$ we must have for every $0 < \eta < \eta^\circ/(2|\overline{\theta} - \underline{\theta}|)$, every $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}[\exp(2\eta(\theta-\theta^*)\cdot Y)] &\le \mathbb{E}[\exp(2\eta|\theta-\theta^*|\cdot|Y|)] \\ &\le \mathbb{E}[\exp(\eta^\circ(|\theta-\theta^*|/|\overline{\theta}-\underline{\theta}|)\cdot|Y|)] \\ &\le \overline{C} \\ &< \infty. \end{aligned} \tag{6.10}$$

The first derivative of the right of (6.8) is:

$$\eta\mathbb{E}\left[(Y - F'(\theta))\exp\Big(\eta\big((\theta-\theta^*)Y + F(\theta^*) - F(\theta)\big)\Big)\right]. \tag{6.11}$$

The second derivative is:

$$\mathbb{E}\left[\left(-\eta F''(\theta) + \eta^2(Y - F'(\theta))^2\right)\cdot\exp\Big(\eta\big((\theta-\theta^*)Y + F(\theta^*) - F(\theta)\big)\Big)\right]. \tag{6.12}$$

We will also use the standard result (Grünwald, 2007; Barndorff-Nielsen, 1978) that, since we assume $\theta^* \in \Theta$,

$$\mathbb{E}[Y] = \mathbb{E}_{Y\sim P_{\theta^*}}[Y] = \mu(\theta^*); \quad \text{for all } \theta \in \overline{\Theta}\colon F'(\theta) = \mu(\theta); \quad F''(\theta) = \mathbb{E}_{Y\sim P_\theta}(Y - E(Y))^2, \tag{6.13}$$

the latter two following because F is the cumulant generating function.

Part (ii). We use an exact second-order Taylor expansion via the Lagrange form of the remainder. We already showed there exist $\eta' > 0$ such that, for all $0 < \eta \le \eta'$, all $\theta \in \Theta$, $\mathbb{E}[\exp(2\eta(\theta - \theta^*)Y)] < \infty$. Fix any such η. For some $\theta' \in \{(1-\alpha)\theta + \alpha\theta^* \colon \alpha \in [0,1]\}$, the (exact) expansion is:

$$\begin{aligned}\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] = 1 + \eta(\theta - \theta^*)\mathbb{E}\left[Y - F'(\theta^*)\right] - \frac{\eta}{2}(\theta - \theta^*)^2 F''(\theta') \ldots \\ \ldots \cdot \mathbb{E}\left[\exp\Big(\eta\big((\theta' - \theta^*)Y + F(\theta^*) - F(\theta')\big)\Big)\right] \ldots \\ \ldots + \frac{\eta^2}{2}(\theta - \theta^*)^2 \mathbb{E}\left[(Y - F'(\theta'))^2 \cdot \exp\Big(\eta\big((\theta' - \theta^*)Y + F(\theta^*) - F(\theta')\big)\Big)\right].\end{aligned}$$

Defining $\Delta = \theta' - \theta$, and since $F'(\theta^*) = \mathbb{E}[Y]$ (see (6.13)), we see that the central condition is equivalent to the inequality:

$$\eta\mathbb{E}\left[(Y - F'(\theta'))^2 e^{\eta\Delta Y}\right] \le F''(\theta')\mathbb{E}\left[e^{\eta\Delta Y}\right].$$

From Cauchy-Schwarz, to show that the η-central condition holds it is sufficient to show that

$$\eta \left\|(Y - F'(\theta'))^2\right\|_{L_2(P)} \left\|e^{\eta\Delta Y}\right\|_{L_2(P)} \le F''(\theta')\mathbb{E}\left[e^{\eta\Delta Y}\right],$$

which is equivalent to

$$\eta \le \frac{F''(\theta')\mathbb{E}\left[e^{\eta\Delta Y}\right]}{\sqrt{\mathbb{E}\left[(Y - F'(\theta'))^4\right]\mathbb{E}\left[e^{2\eta\Delta Y}\right]}}. \tag{6.14}$$

We proceed to lower bound the RHS by lower bounding each of the terms in the numerator and upper bounding each of the terms in the denominator. We begin with the numerator. $F'(\theta)$ is bounded by (6.7). Next, by Jensen's inequality,

$$\mathbb{E}\left[\exp(\eta\Delta Y)\right] \ge \exp(\mathbb{E}[\eta\Delta \cdot Y]) \ge \exp(-\eta^\circ|\overline{\theta} - \underline{\theta}||\mu(\theta^*)|)$$

is lower bounded by a positive constant. It remains to upper bound the denominator. Note that the second factor is upper bounded by the constant $\overline{C}$ in (6.10). The first factor is bounded by a fixed multiple of $\mathbb{E}|Y|^4 + \mathbb{E}[F'(\theta)^4]$. The second term is bounded by (6.7), so it remains to bound the first term. By assumption $\mathbb{E}[\exp(\eta^\circ|Y|)] \le \overline{C}$ and this implies that $\mathbb{E}|Y^4| \le a^4 + \overline{C}$ for any $a \ge e$ such that $a^4 \le \exp(\eta^\circ a)$; such an a clearly exists and only depends on η°.

We have thus shown that the RHS of (6.14) is upper bounded by a quantity that only depends on $\overline{C}, \eta^\circ$ and the values of the extrema in (6.7), which is what we had to show.

Proof of (iii). We now use the asymptotic form of Taylor's theorem. Fix any $\eta > 0$, and pick any θ close enough to θ^* so that (6.8) is finite for all θ' in between θ and θ^*; such a $\theta \ne \theta^*$ must

exist since for any $\delta > 0$, if $|\theta - \theta^*| \le \delta$, then by assumption (6.8) must be finite for all $\eta \le \eta^\circ/\delta$. Evaluating the first and second derivative (6.11) and (6.12) at $\theta = \theta^*$ gives:

$$\begin{aligned}\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] &= 1 + \eta(\theta - \theta^*)\mathbb{E}\left[Y - F'(\theta^*)\right] \dots \\ \dots &- \left(\frac{\eta}{2}(\theta - \theta^*)^2 F''(\theta^*) - \frac{\eta^2}{2}(\theta - \theta^*)^2 \cdot \mathbb{E}\left[(Y - F'(\theta^*))^2\right]\right) + h(\theta)(\theta - \theta^*)^2 \\ &= 1 - \frac{\eta}{2}(\theta - \theta^*)^2 F''(\theta^*) + \frac{\eta^2}{2}(\theta - \theta^*)^2 \mathbb{E}\left[(Y - F'(\theta^*))^2\right] + h(\theta)(\theta - \theta^*)^2,\end{aligned}$$

where $h(\theta)$ is a function satisfying $\lim_{\theta \to \theta^*} h(\theta) = 0$, where we again used (6.13), i.e. that $F'(\theta^*) = \mathbb{E}[Y]$. Using further that $\sigma^2 = \mathbb{E}\left[(Y - F'(\theta^*))^2\right]$ and $F''(\theta^*) = (\sigma^*)^2$, we find that $\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \le 1$ iff

$$-\frac{\eta}{2}(\theta - \theta^*)^2(\sigma^*)^2 + \frac{\eta^2}{2}(\theta - \theta^*)^2\sigma^2 + h(\theta)(\theta - \theta^*)^2 \le 0.$$

It follows that for all $\delta > 0$, there is an $\varepsilon > 0$ such that for all $\theta \in [\theta^* - \varepsilon, \theta^* + \varepsilon]$, all $\eta > 0$,

$$\frac{\eta^2}{2}\sigma^2 \le \frac{\eta}{2}(\sigma^*)^2 - \delta \Rightarrow \mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \le 1 \tag{6.15}$$

$$\frac{\eta^2}{2}\sigma^2 \ge \frac{\eta}{2}(\sigma^*)^2 + \delta \Rightarrow \mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \ge 1 \tag{6.16}$$

The condition in (6.15) is implied if:

$$0 < \eta \le \frac{(\sigma^*)^2}{\sigma^2} - \frac{2\delta}{\eta\sigma^2}.$$

Setting $C = 4\sigma^2/(\sigma^*)^4$ and $\eta_\delta = (1 - C\delta)(\sigma^*)^2/\sigma^2$ we find that for any $\delta < (\sigma^*)^4/(8\sigma^2)$, we have $1 - C\delta \ge 1/2$ and thus $\eta_\delta > 0$ so that in particular the premise in (6.15) is satisfied for η_δ. Thus, for all small enough δ, both the premise and the conclusion in (6.15) hold for $\eta_\delta > 0$; since $\lim_{\delta \downarrow 0} \eta_\delta = (\sigma^*)^2/\sigma^2$, it follows that there is an increasing sequence $\eta_{(1)}, \eta_{(2)}, \dots$ converging to $(\sigma^*)^2/\sigma^2$ such that for each $\eta_{(j)}$, there is $\varepsilon_{(j)} > 0$ such that for all $\theta \in [\theta^* - \varepsilon_{(j)}, \theta^* + \varepsilon_{(j)}]$, $\mathbb{E}\left[e^{-\eta_{(j)}(\ell_\theta - \ell_{\theta^*})}\right] \le 1$. It follows that the lim sup in (6.9) is at least $(\sigma^*)^2/\sigma^2$. A similar argument (details omitted) using (6.16) shows that the lim sup is at most this value; the result follows.

6.A.2 Proof of Proposition 4

For arbitrary conditional densities $p'(y \mid x)$ with corresponding distribution $P' \mid X$ for which

$$\mathbb{E}_{P'}[Y|X] = g^{-1}(\langle \beta, X \rangle), \tag{6.17}$$

and densities $p_{f^*} = p_{\beta^*}$ and p_β with $\beta^*, \beta \in \mathcal{B}$, we can write:

$$\begin{aligned}\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim P'|X}\left[\log \frac{p_{\beta^*}(Y \mid X)}{p_\beta(Y|X)}\right] &= \mathbb{E}\mathbb{E}\left[(\theta_X(\beta^*) - \theta_X(\beta))Y - \log\frac{F(\theta_X(\beta^*))}{F(\theta_X(\beta))} \mid X\right] \\ &= \mathbb{E}_{X\sim P}\left[(\theta_X(\beta^*) - \theta_X(\beta))g^{-1}(\langle \beta, X]_d\rangle \dots\right. \\ &\left. \dots - \log F(\theta_X(\beta^*)) + \log F(\theta_X(\beta)) \mid X\right],\end{aligned}$$

where the latter equation follows by (6.17). The result now follows because (6.17) both holds for the 'true' P and for P_{f^*}.

6.A.3 Proof of Proposition 3

The fact that under the three imposed conditions the $\overline{\eta}$-central condition holds for some $\overline{\eta} > 0$ is a simple consequence of Theorem 6.2: Condition 1 implies that there is some compact Θ such that for all $x \in \mathcal{X}$, $\beta \in \mathcal{B}$, $\theta_x(\beta) \in \Theta$. Condition 3 then ensures that $\theta_x(\beta)$ lies in the interior of this Θ. And Condition 2 implies that $\overline{\eta}$ in Theorem 6.2 can be chosen uniformly for all $x \in \mathcal{X}$.

6.B Excess risk and KL divergence instead of generalized Hellinger distance

The misspecification metric/generalized Hellinger distance $d_{\overline{\eta}}$ appearing in Theorem 6.1 is rather weak (it is 'easy' for two distributions to be close) and lacks a clear interpretation for general, non-logarithmic loss functions. Motivated by these facts, GM study in depth under what additional conditions the (square of this) metric can be replaced by a stronger and more readily interpretable divergence measure. They come up with a new, surprisingly weak condition, the *witness condition*, under which $d_{\overline{\eta}}$ can be replaced by the *excess risk* $\mathbf{E}_P[L_f]$, which is the additional risk incurred by f as compared to the optimal f^*. For example, with the squared error loss, this is the additional mean square error of f compared to f^*; and with (conditional) log-loss, it is the well-known *generalized KL divergence* $\mathbf{E}_{X,Y\sim P}[\log \frac{p_{f^*}(Y|X)}{p_f(Y|X)}]$, coinciding with standard KL divergence if the model is correctly specified. Bounding the excess risk is a standard goal in statistical learning theory; see for example (Bartlett, Bousquet and Mendelson, 2005; Van Erven et al., 2015).

The following definition appears (with substantial explanation including the reason for its name) as Definition 12 in GM:

Definition 6.2 (Empirical Witness of Badness)**.** We say that $(P, \ell, \mathcal{F})$ satisfies the (u, c)-*empirical witness of badness condition* (or *witness condition*) for constants $u > 0$ and $c \in (0, 1]$ if for all $f \in \mathcal{F}$

$$\mathbb{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbb{1}_{\{\cdot\}}\ell_f - \ell_{f^*} \le u\right] \ge c\mathbb{E}[\ell_f - \ell_{f^*}].$$

More generally, for a function $\tau : \mathbb{R}^+ \to [1, \infty)$ and constant $c \in (0, 1)$ we say that $(P, \ell, \mathcal{F})$ satisfies the (τ, c)-*witness condition* if for all $f \in \mathcal{F}$, $\mathbb{E}[\ell_f - \ell_{f^*}] < \infty$ and

$$\mathbb{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbb{1}_{\{\cdot\}}\ell_f - \ell_{f^*} \le \tau(\mathbb{E}[\ell_f - \ell_{f^*}])\right] \ge c\mathbb{E}[\ell_f - \ell_{f^*}].$$

It turns out that the (τ, c)-witness condition holds in many practical situations, including our GLM-under-misspecification setting. Before elaborating on this, let us review (a special case of) Theorem 12 of GM, which is the analogue of Theorem 6.1 but with the misspecification metric replaced by the excess risk.

First, let, for arbitrary $0 < \eta < \overline{\eta}$, $c_u \coloneqq \frac{1}{c}\frac{\eta u+1}{1-\frac{\eta}{\overline{\eta}}}$. Note that for large u, c_u is approximately linear in u/c.

Theorem 6.5. [Specialization of Theorem 12 of GM] *Consider a learning problem $(P, \ell, \mathcal{F})$. Suppose that the $\overline{\eta}$-strong central condition holds. If the (u, c)-witness condition holds, then for any $\eta \in (0, \overline{\eta})$,*

$$\mathbf{E}_{Z^n\sim P}\mathbb{E}_{\underline{f}\sim\Pi_n}\left[\mathbb{E}[L_f]\right] \leq c_u \cdot \mathbf{E}_{Z^n\sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_0\right)\right], \tag{6.18}$$

with c_u as above. If instead the (τ, c)-witness condition holds for some nonincreasing function τ as above, then for any $\lambda > 0$,

$$\mathbf{E}_{Z^n\sim P}\mathbb{E}_{\underline{f}\sim\Pi_n}\left[\mathbb{E}[L_f]\right] \;\leq\; \lambda + c_{\tau(\lambda)} \cdot \mathbf{E}_{Z^n\sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_0\right)\right].$$

The actual theorem given by GM generalizes this to an in-probability statement for general (not just generalized Bayesian) learning methods. If the (u, c)-witness condition holds, then, as is obvious from (6.18) and Theorem 6.1, the same rates can be obtained for the excess risk as for the squared misspecification metric. For the (τ, c)-witness condition things are a bit more complicated; the following lemma (Lemma 16 of GM) says that, under an exponential tail condition, (τ, c)-witness holds for a sufficiently 'nice' function τ, for which we loose at most a logarithmic factor:

Lemma 6. *Define $M_\kappa \coloneqq \sup_{f\in\mathcal{F}} \mathbb{E}\left[e^{\kappa L_f}\right]$ and assume that the excess loss L_f has a uniformly exponential upper tail, i.e. $M_\kappa < \infty$. Then, for the map $\tau : x \mapsto 1 \vee \kappa^{-1}\log\frac{2M_\kappa}{\kappa x} = O(1 \vee \log(1/x))$, the (τ, c)-witness condition holds with $c = 1/2$.*

As an immediate consequence of this lemma, GM's theorem above gives that for any $\eta \in (0, \overline{\eta})$, (using $\lambda = 1/n$), there is $C_\eta < \infty$ such that

$$\mathbf{E}_{Z^n\sim P}\mathbb{E}_{\underline{f}\sim\Pi_n}\left[\mathbb{E}[L_{\underline{f}}]\right] \leq \frac{1}{n} + C_\eta \cdot (\log n) \cdot \mathbf{E}_{Z^n\sim P}\left[\mathrm{IC}_{\eta,n}\left(f^* \parallel \Pi_|\right)\right], \tag{6.19}$$

so our excess risk bound is only a log factor worse than the bound that can be obtained for the squared misspecification metric in Theorem 6.1. We now apply this to the misspecified GLM setting:

Generalized Linear Models and Witness Recall that the central condition holds for generalized linear models under the three assumptions made in Proposition 3. Let $\ell_\beta \coloneqq \ell_\beta(X, Y) = -\log p_\beta(Y \mid X)$ be the loss of action $\beta \in \mathcal{B}$ on random outcome $(X, Y) \sim P$, and let β^* denote the risk minimizer over $\mathcal{B}$. The first two assumptions taken together imply, via (6.7), that there is a $\kappa > 0$ such that

$$\begin{aligned}\sup_{\beta\in B}\mathbb{E}_{X,Y\sim P}\left[e^{\kappa(\ell_\beta-\ell_{\beta^*})}\right] &\leq \sup_{\beta\in\mathcal{B},x\in\mathcal{X}}\mathbb{E}_{Y\sim P|X=x}\left[e^{\kappa(\ell_\beta-\ell_{\beta^*})}\right] \\ &= \sup_{\beta\in\mathcal{B},x\in\mathcal{X}}\left(\frac{F_{\theta_x(\beta)}}{F_{\theta_x(\beta^*)}}\right)^\kappa \cdot \mathbb{E}_{Y\sim P|X=x}\left[e^{\kappa|Y|}\right] < \infty.\end{aligned}$$

The conditions of Lemma 6 are thus satisfied, and so the (τ, c)-witness condition holds for the τ and c in that lemma. From (6.19) we now see that we get an $O((\log n)^2/n)$ bound on the expected excess risk, which is equal to the parametric (minimax) rate up to a $(\log n)^2$ factor. Thus, fast learning rates in terms of excess risks and KL divergence under misspecification with GLMs are possible under the conditions of Proposition 3.

6.C Learning rate > 1 for misspecified models

In what follows we give an example of a misspecified setting, where the best performance is achieved with the learning rate $\eta > 1$. Consider a model $\{P_\beta, \beta \in [0.2, 0.8]\}$, where P_β is a Bernoulli distribution with $\mathbb{P}_\beta(Y = 1) = \beta$. Let the data $Y_1, \dots, Y_n$ be sampled i.i.d. from P_0, i.e. $Y_i = 0$ for all $i = 1, \dots, n$. In this case the log-likelihood function is given by

$$\log p(Y_1, \dots, Y_n \,|\, \beta) = n \log(1 - \beta).$$

Observe that in this setting $\beta^\star = 0.2$. Now assume that the model is correct and data $Y_1', \dots, Y_n'$ is sampled i.i.d. from P_β with $\beta = 0.2$. Then the log-likelihood is

$$\log p(Y_1', \dots, Y_n' \,|\, \beta = 0.2) \approx 0.2n \log 0.2 + 0.8n \log 0.8 \ll n \log 0.8 = \log p(Y_1, \dots, Y_n \,|\, \beta = 0.2).$$

Thus, the data are more informative about the best distribution than they would be if the model were correct. Therefore, we can afford to learn 'faster': let the data be more important and the (regularizing) prior be less important. This is realized by taking $\eta >> 1$

6.D MCMC sampling

6.D.1 The η-generalized Bayesian lasso

Here, following Park and Casella (2008) we consider a slightly more general version of the regression problem:

$$Y = \mu + X\beta + \varepsilon,$$

where $\mu \in \mathbb{R}^n$ is the overall mean, $\beta \in \mathbb{R}^p$ is the vector of parameters of interest, $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\varepsilon \sim N(0, \sigma^2 I_n)$ is a noise vector. For a given shrinkage parameter $\lambda > 0$ the Bayesian lasso of Park and Casella (2008) can be represented as follows.

$$\begin{aligned} Y|\mu, X, \beta, \sigma^2 &\sim N(\mu + X\beta, \sigma^2 I_n)\,, \\ \beta|\tau_1^2, \dots, \tau_p^2, \sigma^2 &\sim N(0, \sigma^2 D_\tau), \quad D_\tau = \operatorname{diag}(\tau_1^2, \dots, \tau_p^2)\,, \\ \tau_1^2, \dots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \quad \tau_1^2, \dots, \tau_p^2 > 0\,, \\ \sigma^2 &\sim \pi(\sigma^2)\, d\sigma^2\,. \end{aligned} \tag{6.20}$$

In this model formulation the μ on which the outcome variables Y depend, is the overall mean, from which $X\beta$ are deviations. The parameter μ can be given a flat prior and subsequently integrated out, as we do in the coming sections.

We will use the typical inverse gamma prior distribution on σ^2, i.e. for $\sigma^2 > 0$

$$\pi(\sigma^2) = \frac{\gamma^\alpha}{\Gamma(\alpha)} \sigma^{-2\alpha-2} e^{-\gamma/\sigma^2},$$

where $\alpha, \gamma > 0$ are hyperparameters. With the hierarchy of (6.20) the joint density for the posterior with the likelihood to the power η becomes

$$\begin{aligned}
&(f(Y|\mu,\beta,\sigma^2))^\eta \, \pi(\sigma^2)\, \pi(\mu) \prod_{j=1}^{p} \pi(\beta_j|\tau_j^2,\sigma^2)\, \pi(\tau_j^2) \\
&\qquad = \left(\frac{1}{(2\pi\sigma^2)^{n/2}} \, e^{\frac{1}{2\sigma^2}(Y-\mu 1_n - X\beta)^T (Y-\mu 1_n - X\beta)} \right)^\eta \cdots \\
&\qquad\qquad \cdots \frac{\gamma^\alpha}{\Gamma(\alpha)} \sigma^{-2\alpha-2} e^{-\frac{\gamma}{\sigma^2}} \prod_{j=1}^{p} \frac{1}{(2\sigma^2\tau_j^2)^{1/2}} \, e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2} \frac{\lambda^2}{2} \, e^{-\lambda^2\tau_j^2/2}. \qquad (6.21)
\end{aligned}$$

Let $\widetilde{Y}$ be $Y - \overline{Y}$. If we integrate out μ, the joint density marginal over μ is proportional to

$$\sigma^{-\eta(n-1)} \, e^{-\frac{\eta}{2\sigma^2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)} \, \sigma^{-2\alpha-2} \, e^{-\frac{\gamma}{\sigma^2}} \prod_{j=1}^{p} \frac{1}{(\sigma^2\tau_j^2)^{1/2}} \, e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2} \, e^{-\lambda^2\tau_j^2/2}. \qquad (6.22)$$

First, observe that the full conditional for β is multivariate normal: the exponent terms involving β in (6.22) are

$$\begin{aligned}
&-\frac{\eta}{2\sigma^2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta) - \frac{1}{2\sigma^2}\beta^T {D_\tau}^{-1}\beta \\
&\qquad = -\frac{1}{2\sigma^2}\left\{ (\beta^T(\eta X^T X + {D_\tau}^{-1})\beta - 2\eta\widetilde{Y}X\beta + \eta\widetilde{Y}^T\widetilde{Y}) \right\}. \qquad (6.23)
\end{aligned}$$

If we now write $M_\tau = (\eta X^T X + {D_\tau}^{-1})^{-1}$ and complete the square, we arrive at

$$-\frac{1}{2\sigma^2}\left\{ (\beta - \eta M_\tau X^T\widetilde{Y})^T M_\tau^{-1} (\beta - \eta M_\tau X^T\widetilde{Y}) + \widetilde{Y}^T(\eta I_n - \eta^2 X^{-1} M_\tau X^T)\widetilde{Y} \right\}.$$

Accordingly we can see that β is conditionally multivariate normal with mean $\eta M_\tau X^T\widetilde{Y}$ and variance $\sigma^2 M_\tau$.

The terms in (6.22) that involve σ^2 are:

$$(\sigma^2)^{\{-\eta(n-1)/2-p/2-\alpha-1\}} \exp\Big\{-\frac{\eta}{2\sigma^2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)-\frac{1}{2\sigma^2}\beta^T {D_\tau}^{-1}\beta-\frac{\gamma}{\sigma^2}\Big\}.$$

We can conclude that σ^2 is conditionally inverse gamma with shape parameter $\eta\,\dfrac{n-1}{2}+\dfrac{p}{2}+\alpha$ and scale parameter $\dfrac{\eta}{2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)+\beta^T {D_\tau}^{-1}\beta/2+\gamma$.

Since τ_j^2 is not involved in the likelihood, we need not modify the implementation of it and follow Park and Casella (2008):

$$\frac{1}{\tau_j^2}\sim \mathrm{IG}\Big(\sqrt{\lambda^2\sigma^2/\beta_j^2},\,\lambda^2\Big).$$

Summarizing, we can implement a Gibbs sampler with the following distributions:

$$\beta\sim \mathrm{N}\left(\eta(\eta X^TX+{D_\tau}^{-1})^{-1}X^T\widetilde{Y},\,\sigma^2(\eta X^TX+{D_\tau}^{-1})^{-1}\right), \tag{6.24}$$

$$\sigma^2\sim \text{Inv-Gamma}\Big(\frac{\eta}{2}(n-1)+p/2+\alpha,\,\frac{\eta}{2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)+\beta^T {D_\tau}^{-1}\beta/2+\gamma\Big), \tag{6.25}$$

$$\frac{1}{\tau_j^2}\sim \mathrm{IG}\Big(\sqrt{\lambda^2\sigma^2/\beta_j^2},\,\lambda^2\Big). \tag{6.26}$$

There are several ways to deal with the shrinkage parameter λ. We follow the hierarchical Bayesian approach and place a hyperprior on the parameter. In our implementation we provide three ways to do so: a point mass (resulting in a fixed λ), a gamma prior on λ^2 following Park and Casella (2008) and a beta prior following De los Campos et al. (2009), details about the implementation of the latter two priors can be found in those papers respectively.

6.D.2 The η-generalized Bayesian logistic regression

We follow the construction of the Pólya–Gamma latent variable scheme for constructing a Bayesian estimator in the logistic regression context described in Polson, Scott and Windle, 2013.

First, for $b>0$ consider the density function of a Pólya-Gamma random variable $PG(b,0)$

$$p(x\,|\,b,0)=\frac{2^{b-1}}{\Gamma(b)}\sum_{n=1}^{\infty}(-1)^n\frac{\Gamma(n+b)}{\Gamma(n+1)}\frac{(2n+b)}{\sqrt{2\pi x^3}}e^{-\frac{(2n+b)^2}{8x}}.$$

The general class $PG(b,c)$ $(b,c>0)$ is defined through an exponential tilting of the $PG(b,0)$ and has the density function

$$p(x\,|\,b,c)=\frac{e^{-\frac{c^2x}{2}}p(x|b,0)}{\mathbb{E}\left[e\right]^{-\frac{c^2\omega}{2}}},$$

where $\omega\sim PG(b,0)$.

To derive our Gibbs sampler we use the following result from Polson, Scott and Windle, 2013.

Theorem 6.D.1. *Let $p_{b,0}(\omega)$ denote the density of $PG(b,0)$. Then for all $a \in \mathbb{R}$*

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p_{b,0}(\omega) d\omega,$$

where $\kappa = a - b/2$.

According to Theorem 6.D.1 the likelihood contribution of the observation i taken to the power η can be written as

$$L_{i,\eta}(\beta) = \left[\frac{(e^{X_i^T\beta})^{y_i}}{1+e^{X_i^T\beta}}\right]^\eta \propto e^{\eta\kappa_i X_i^T\beta} \int_0^\infty e^{-\omega_i \frac{(X_i^T\beta)^2}{2}} p(\omega_i \mid \eta, 0),$$

where $\kappa_i := y_i - 1/2$ and $p(\omega_i \mid \eta, 0)$ is the density function of $PG(\eta, 0)$.

Let

$$X := (X_1, \ldots, X_n)^T, \quad Y := (Y_1, \ldots, Y_n)^T, \quad \kappa := (\kappa_1, \ldots, \kappa_n)^T,$$
$$\omega := (\omega_1, \ldots, \omega_n)^T, \quad \Omega := \operatorname{diag}(\omega_1, \ldots, \omega_n).$$

Also, denote the density of the prior on β by $\pi(\beta)$. Then the conditional posterior of β given ω is

$$p(\beta \mid \omega, Y) \propto \pi(\beta) \prod_{i=1}^n L_{i,\eta}(\beta \mid \omega_i) = \pi(\beta) \prod_{i=1}^n e^{\eta\kappa_i X_i^T\beta - \omega_i \frac{(X_i^T\beta)^2}{2}} \propto \pi(\beta) e^{-\frac{1}{2}(z-X\beta)^T\Omega(z-X\beta)},$$

where $z := \eta(\frac{\kappa_1}{\omega_1}, \ldots, \frac{\kappa_n}{\omega_n})$. Observe that the likelihood part is conditionally Gaussian in β. Since the prior on β is Gaussian, a simple linear-model calculation leads to the following Gibbs sampler. To sample from the the η-generalized posterior one has to iterate these two steps

$$\omega_i \mid \beta \sim PG(\eta, X_i^T\beta), \tag{6.27}$$
$$\beta \mid Y, \omega \sim \mathcal{N}(m_\omega, V_\omega), \tag{6.28}$$

where

$$V_\omega := (X^T\Omega X + B^{-1})^{-1},$$
$$m_\omega := V_\omega(\eta X^T\kappa + B^{-1}b).$$

To sample from the Pólya-Gamma distribution $PG(b,c)$ we adopt a method from (Windle, Polson and Scott, 2014), which is based on the following representation result. According to Polson, Scott and Windle, 2013 a random variable $\omega \sim PG(b,c)$ admits the following representation

$$\omega \overset{\mathrm{d}}{=} \sum_{n=0}^\infty \frac{g_n}{d_n},$$

where $g_n \sim Ga(b,1)$ are independent Gamma distributed random variables, and

$$d_n := 2\pi^2(n+\frac{1}{2})^2 + 2c^2.$$

Therefore, we approximate the PG random variable by a truncated sum of weighted Gamma random variables. (Windle, Polson and Scott, 2014) shows that the approximation method performs well with the truncation level $N = 300$. Furthermore, we performed our own comparison of the sampler with the STAN implementation for Bayesian logistic regression, which showed no difference between the methods (for $\eta = 1$).

6.D.3 The Safe-Bayesian Algorithms

The version of the Safe-Bayesian algorithm we are using for the experiments is called *R-log-SafeBayes*, more details and other versions can be found in Grünwald and Van Ommen (2017). The $\widehat{\eta}$ is chosen from a grid of learning rates η that minimizes the *cumulative Posterior-Expected Posterior-Randomized log-loss*:

$$\sum_{i=1}^{n} \mathbb{E}_{\beta,\sigma^2\sim\Pi|z^{i-1},\eta}\left[-\log f(Y_i|X_i,\beta,\sigma^2)\right].$$

Minimizing this comes down to minimizing

$$\sum_{i=1}^{n-1} \text{AV}\left[\frac{1}{2}\log 2\pi\sigma_{i,\eta}^2 + \frac{1}{2}\frac{(Y_{i+1}-X_{i+1}\beta_{i,\eta})^2}{\sigma_{i,\eta}^2}\right].$$

The loss between the brackets is averaged over many draws of $(\beta_{i,\eta}, \sigma_{i,\eta}^2)$ from the posterior, where $\beta_{i,\eta}$ (or $\sigma_{i,\eta}^2$) denotes one random draw from the conditional η-generalized posterior based on data points z^i. For the sake of completeness we present the algorithm below.

Algorithm 1 The R-Safe-Bayesian algorithm

Input: data $z_1, \ldots z_n$, model $\mathcal{M} = \{f(\cdot|\theta)|\theta \in \Theta\}$, prior Π on Θ, step-size $\mathcal{K}_{\text{STEP}}$, max. exponent $\mathcal{K}_{\text{MAX}}$, loss function $\ell_\theta(z)$
$\mathcal{S}_n := \{1, 2^{-\mathcal{K}_{\text{STEP}}}, 2^{-2\mathcal{K}_{\text{STEP}}}, 2^{-3\mathcal{K}_{\text{STEP}}}, \ldots, 2^{-\mathcal{K}_{\text{MAX}}}, \}$
for all $\eta \in \mathcal{S}_n$ **do**
    $s_\eta := 0$
    for $i = 1 \ldots n$ **do**
        Determine generalized posterior $\Pi(\cdot|z^{i-1}, \eta)$ of Bayes with learning rate η.
        Calculate posterior-expected posterior-randomized loss of predicting actual next outcome:

$$r := \ell_{\Pi|z^{i-1},\eta}(z_i) = \mathbb{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[\ell_\theta(z_i)\right] \tag{6.29}$$

        $s_\eta := s_\eta + r$
    end for
end for
Ouput: Learning rate $\widehat{\eta}$

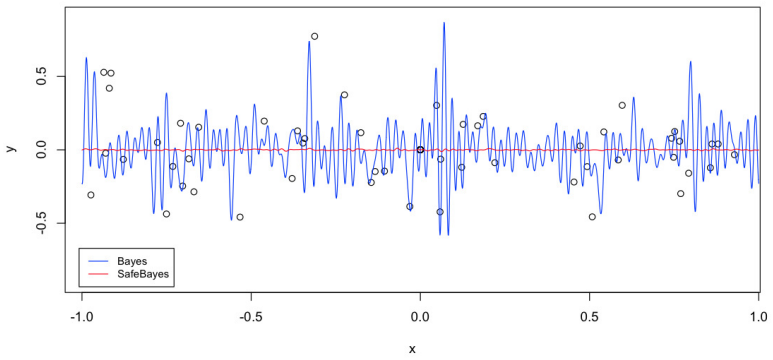


Figure 6.5: *Prediction of standard Bayesian lasso (blue) and Safe-Bayesian lasso (red, $\eta = 0.5$) with $n = 200$, $p = 100$.*

6.E Details for the experiments and figures

Below we present the results of additional simulation experiments for Section 6.5.1 (Appendix 6.E.1) and the description of experiments with real-world data (Appendix 6.E.2). We also give details for Figure 6.2 in Appendix 6.E.3.

6.E.1 Additional Figures for Section 6.5.1

Consider the regression context described in Section 6.5.1. Here, we explore different choices of the number of Fourier basis functions, showing that regardless of the choice Safe-Baysian lasso outperforms its standard counterpart. In Figures 6.5 and 6.6 we see conditional expectations $\mathbb{E}[Y \mid X]$ according to the posteriors of the standard Bayesian lasso (blue) and the Safe-Bayesian lasso (red, $\widehat{\eta} = 0.5$) for the *wrong-model* experiment described in Section 6.5.1, with 100 data points. We take 201 and 25 Fourier basis functions respectively.

Now we consider logistic regression setting and show that even for some well-specified problems it is beneficial to choose $\eta \neq 1$. In Figure 6.7 we see a comparison of the log-risk for $\eta = 1$ and $\eta = 3$ in the well-specified logistic regression case (described in Section 6.5.1). Here $p = 1$ and $\beta = 4$.

6.E.2 Real-world data

Seattle Weather Data The R-package `weatherData` (Narasimhan, 2014) loads weather data available online from `www.wunderground.com`. Besides data from many thousands of personal weather stations and government agencies, the website provides access to data from Automated Surface Observation Systems (ASOS) stations located at airports in the US, owned and maintained by the Federal Aviation Administration. Among them is a weather station at Seattle Tacoma International Airport, Washington (WMO ID 72793). From this station we collected the data for this experiment.

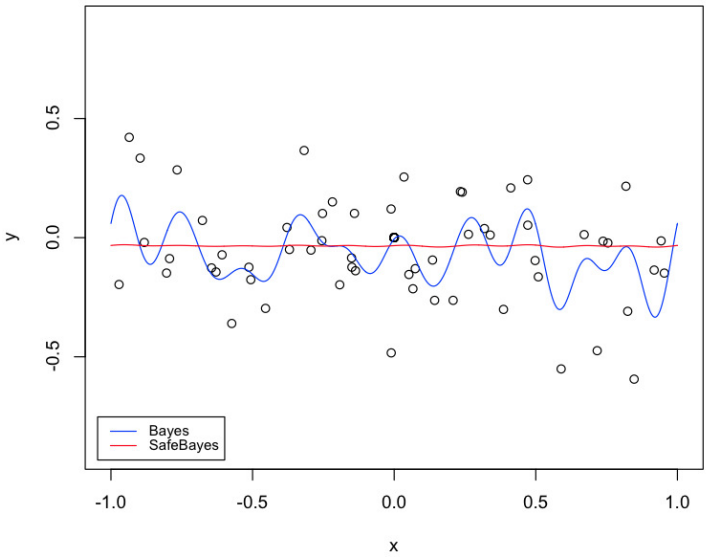


Figure 6.6: *Prediction of standard Bayesian lasso (blue) and Safe-Bayesian lasso (red, $\eta = 0.5$) with $n = 200$, $p = 12$.*

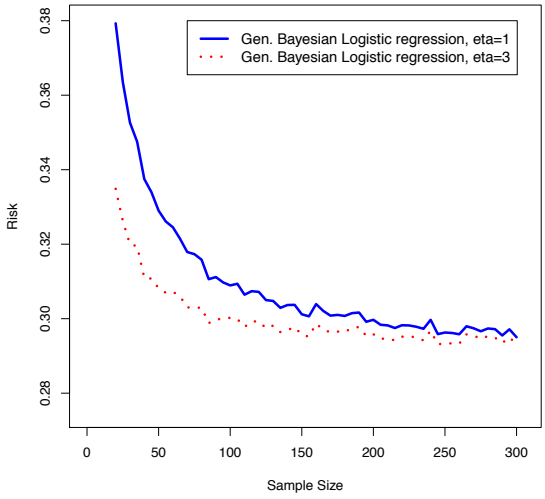


Figure 6.7: *Simulated logistic risk as a function of the sample size for the* correct-model *experiments described in Section 6.5.1 according to the posterior predictive distribution of standard Bayesian logistic regression ($\eta = 1$), and generalized Bayes ($\eta = 3$).*

The training data are the maximum temperatures for each day of the year 2011 at Seattle airport. We divided the data randomly in a training set (300 measurements) and a test set (65 measurements). First, we sampled the posterior of the standard Bayesian lasso with a 201-dimensional Fourier basis and standard improper priors on the training set, and we did the same for the Horseshoe. Next, we sampled the generalized posterior with the learning rate $\widehat{\eta}$ learned by the Safe-Bayesian algorithm, with the same model and priors on the same training set. The grid of η's we used was $1, 0.9, 0.8, 0.7, 0.6, 0.5$. We compare the performance of the standard Bayesian lasso and Horseshoe and the Safe-Bayesian versions of the lasso (SB) in terms of mean square error. In all experiments performed with different partitions, priors and number of iterations, SafeBayes never picked $\widehat{\eta} = 1$. We averaged over 10 runs. Moreover, whichever learning rate was chosen by SafeBayes, it always outperformed standard Bayes (with $\eta = 1$) in an unchanged set-up. Experiments with different priors for λ yielded similar results.

London Air Pollution Data As training set we use the following data. We start with the first four weeks of the year 2013, starting at Monday January 7 at midnight. We have a measurement for (almost) every hour until Sunday February 3rd, 23.00. We also have data for the first four weeks of 2014, starting at Monday January 6 at midnight, until Sunday February 2nd, 23.00. For each hour in the four weeks we randomly pick a data point from either 2013 or 2014. We remove the missing values. We predict for the same time of year in 2015: starting at Monday January 5 at midnight, until Sunday February 1st at 23.00. We do this with a (Safe-)Bayesian lasso and Horseshoe with a 201-dimensional Fourier basis and standard improper priors. The grid of η's we used for the Safe-Bayesian algorithm was again $1, 0.9, 0.8, 0.7, 0.6, 0.5$. We look at the mean square prediction errors, and average the errors over 20 runs of the generalized Bayesian lasso with the η learned by SafeBayes, and the standard Bayesian lasso and Horseshoe. Again we find that SafeBayes clearly performs better than standard Bayes.

6.E.3 Details for Figure 6.2

Here we sampled the posteriors of the standard and generalized Bayesian lasso ($\eta = 0.25$) on 50 model-wrong data points (approximately half easy points) with 101 Fourier basis functions, and estimated the predictive variance on a grid of new data points $X_{\text{new}} = \{-1.00, -0.99, \ldots, 1.00\}$ with the Monte Carlo estimate:

$$\widehat{\text{VAR}}\big(Y_{\text{new}} \mid X_{\text{new}}, Z_{\text{old}}\big) = \mathbb{E}_{\theta|Z_{\text{old}}}\big[\text{VAR}\big(Y_{\text{new}} \mid \theta\big)\big] + \widehat{\text{VAR}}\big[\mathbb{E}\big(Y_{\text{new}} \mid \theta\big)\big], \tag{6.30}$$

where

$$\mathbb{E}_{\theta|Z_{\text{old}}}\big[\text{VAR}\big(Y_{\text{new}} \mid \theta\big)\big] = \frac{1}{m}\sum_{k=1}^{m} \sigma^{2[k]} = \overline{\sigma^2},$$

$$\widehat{\text{VAR}}\big[\mathbb{E}\big(Y_{\text{new}} \mid \theta\big)\big] = \widehat{\text{VAR}}\big[X_{\text{new}}\beta\big] = \frac{1}{m}\sum_{k=1}^{m}\Big(X_{\text{new}}\beta^{[k]}\Big)^2 - \Big(X_{\text{new}}\overline{\beta}\Big)^2.$$

Here $\overline{\beta}$ is the posterior mean of the parameter for the coefficients and $\overline{\sigma^2}$ is the posterior mean of the variance.

Chapter 7

Fixed-confidence guarantees for Bayesian best-arm identification

Abstract

We investigate and provide new insights on the sampling rule called Top-Two Thompson Sampling (TTTS). In particular, we justify its use for *fixed-confidence best-arm identification*. We further propose a variant of TTTS called Top-Two Transportation Cost (T3C), which disposes of the computational burden of TTTS. As our main contribution, we provide the first sample complexity analysis of TTTS and T3C when coupled with a very natural Bayesian stopping rule, for bandits with Gaussian rewards, solving one of the open questions raised by Russo (2016). We also provide new posterior convergence results for TTTS under two models that are commonly used in practice: bandits with Gaussian and Bernoulli rewards and conjugate priors.

7.1 Introduction

In multi-armed bandits, a learner repeatedly chooses an *arm* to play, and receives a reward from the associated unknown probability distribution. When the task is *best-arm* identification (BAI), the learner is not only asked to sample an arm at each stage, but is also asked to output a recommendation (i.e., a guess for the arm with the largest mean reward) after a certain period. Unlike in another well-studied bandit setting, the learner is not interested in maximizing the sum of rewards gathered during the exploration (or minimizing *regret*), but only cares about the quality of her recommendation. As such, BAI is a particular *pure exploration* setting (Bubeck, Munos and Stoltz, 2009).

Formally, we consider a finite-arm bandit model, which is a collection of K probability distributions, called arms $\mathcal{A} \triangleq \{1, \dots, K\}$, parametrized by their means $\mu_1, \dots, \mu_K$. We assume the (unknown) best arm is unique and we denote it by $I^\star \triangleq \arg\max_i \mu_i$. A best-arm identification

strategy (I_n, J_n, τ) consists of three components. The first is a *sampling rule*, which selects an arm I_n at round n. At each round n, a vector of rewards $\mathbf{Y}_n = (Y_{n,1}, \cdots, Y_{n,K})$ is generated for all arms independently from past observations, but only Y_{n,I_n} is revealed to the learner. Let $\mathcal{F}_n$ be the σ-algebra generated by $(U_0, I_1, Y_{1,I_1}, U_1, \cdots, I_n, Y_{n,I_n}, U_n)$, then I_n is $\mathcal{F}_{n-1}$-measurable, i.e., it can only depend on the past $n-1$ observations, and some exogenous randomness, materialized into $U_{n-1} \sim \mathcal{U}([0,1])$. The second component is a $\mathcal{F}_n$-measurable *recommendation rule* J_n, which returns a guess for the best arm, and thirdly, the *stopping rule* τ, a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, decides when the exploration is over.

BAI is studied within several theoretical frameworks. In this chapter we consider the fixed-confidence setting, introduced by Even-dar, Mannor and Mansour, 2003. Given a risk parameter $\delta \in [0,1]$, the goal is to ensure that the probability to stop and recommend a wrong arm, $\mathbb{P}[J_\tau \neq I^\star \wedge \tau < \infty]$, is smaller than δ, while minimizing the expected total number of samples to make this accurate recommendation, $\mathbb{E}[\tau]$. The most studied alternative setting is the fixed-budget setting for which the stopping rule τ is fixed to some (known) maximal budget n, and the goal is to minimize the error probability $\mathbb{P}[J_n \neq I^\star]$ (Audibert and Bubeck, 2010). Note that these two frameworks are very different in general and do not share transferable regret bounds (see Carpentier and Locatelli 2016 for an additional discussion).

Most existing sampling rules for the fixed-confidence setting depend on the risk parameter δ. Some of them rely on confidence intervals such as `LUCB` (Kalyanakrishnan et al., 2012), `UGapE` (Gabillon, Ghavamzadeh and Lazaric, 2012), or `lil'UCB` (Jamieson et al., 2014); others are based on eliminations such as `SuccessiveElimination` (Even-dar, Mannor and Mansour, 2003) and `ExponentialGapElimination` (Karnin, Koren and Somekh, 2013). The first known sampling rule for BAI that does not depend on δ is the *tracking* rule proposed by Garivier and Kaufmann, 2016, which is proved to achieve the minimal sample complexity when combined with the Chernoff stopping rule when δ goes to zero. Such an *anytime* sampling rule (neither depending on a risk δ or a budget n) is very appealing for applications, as advocated by Jun and Nowak, 2016 who introduce the anytime best-arm identification framework. In this chapter, we investigate another anytime sampling rule for BAI: Top-Two Thompson Sampling (`TTTS`), and propose a second anytime sampling rule: Top-Two Transportation Cost (`T3C`).

Thompson Sampling (Thompson, 1933) is a Bayesian algorithm well known for regret minimization, for which it is now seen as a major competitor to `UCB`-typed approaches (Burnetas and Katehakis, 1996; Auer, Cesa-Bianchi and Fischer, 2002; Cappé et al., 2013). However, it is also well known that regret minimizing algorithms cannot yield optimal performance for BAI (Bubeck, Munos and Stoltz, 2011; Kaufmann and Garivier, 2017) and as we opt Thompson Sampling for BAI, then its adaptation is necessary. Such an adaptation, `TTTS`, was given by Russo (2016) along with two other top-two sampling rules `TTPS` and `TTVS`. By choosing between two different candidate arms in each round, these sampling rules enforce the exploration of sub-optimal arms, which would be under-sampled by vanilla Thompson sampling due to its objective of maximizing rewards.

While `TTTS` appears to be a good anytime sampling rule for fixed-confidence BAI when coupled with an appropriate stopping rule, so far there is no theoretical support for this employment. Indeed, the (Bayesian-flavored) asymptotic analysis of Russo, 2016 shows that under `TTTS`, the posterior probability that $I^\star$ is the best arm converges almost surely to 1 at the best possible

rate. However, this property does not by itself translate into sample complexity guarantees. Since the result of Russo, 2016, Qin, Klabjan and Russo (2017) proposed and analyzed TTEI, another Bayesian sampling rule, both in the fixed-confidence setting and in terms of posterior convergence rate. Nonetheless, similar guarantees for TTTS have been left as an open question by Russo, 2016. In the present chapter, we answer the question whether we can obtain fixed-confidence guarantees and optimal posterior convergence rates for TTTS. In addition, we propose T3C, a computationally more favorable variant of TTTS and extend the fixed-confidence guarantees to T3C as well.

Contributions (1) We propose a new Bayesian sampling rule, T3C, which is inspired by TTTS but easier to implement and computationally advantageous (2) We investigate two Bayesian stopping and recommendation rules and establish their δ-correctness for a bandit model with Gaussian rewards.[1] (3) We provide the first sample complexity analysis of TTTS and T3C for a Gaussian model and our proposed stopping rule. (4) Russo's posterior convergence results for TTTS were obtained under restrictive assumptions on the models and priors, which exclude the two mostly used models in practice: Gaussian bandits with Gaussian priors and bandits with Bernoulli rewards[2] with Beta priors. We prove that optimal posterior convergence rates can be obtained for those two as well.

Outline In Section 7.2, we restate TTTS and introduce T3C along with our proposed recommendation and stopping rules. Then, in Section 7.3, we describe in detail two important notions of optimality that are invoked in this chapter. The main fixed-confidence analysis follows in Section 7.4, and further Bayesian optimality results are given in Section 7.5. Numerical illustrations are given in Section 7.6.

7.2 Bayesian BAI Strategies

In this section, we give an overview of the sampling rule TTTS and introduce T3C. We provide details for Bayesian updating for Gaussian and Bernoulli models respectively, and introduce associated Bayesian stopping and recommendation rules.

7.2.1 Sampling rules

Both TTTS and T3C employ a Bayesian machinery and make use of a prior distribution Π_1 over a set of parameters Θ, which is assumed to contain the unknown true parameter vector $\boldsymbol{\mu}$. Upon acquiring observations $(Y_{1,I_1}, \cdots, Y_{n-1,I_{n-1}})$, we update our beliefs according to Bayes' rule and obtain a posterior distribution Π_n which we assume to have density π_n w.r.t. the Lebesgue measure. Russo's analysis is requires strong regularity properties on the models and priors, which exclude two important useful cases we consider in this chapter: (1) the observations of each arm i follow a Gaussian distribution $\mathcal{N}(\mu_i, \sigma^2)$ with common known variance σ^2, with imposed Gaussian prior $\mathcal{N}(\mu_{1,i}, \sigma^2_{1,i})$, (2) all arms receive Bernoulli rewards with unknown means, with a uniform ($\mathcal{B}eta(1,1)$) prior on each arm.

[1] Hereafter *Gaussian bandits* or *Gaussian model.*
[2] Hereafter *Bernoulli bandits.*

Gaussian model For Gaussian bandits with a $\mathcal{N}(0, \kappa^2)$ prior on each mean, the posterior distribution of μ_i at round n is Gaussian with mean and variance that are respectively given by

$$\frac{\sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell}}{T_{n,i} + \sigma^2/\kappa^2} \quad \text{and} \quad \frac{\sigma^2}{T_{n,i} + \sigma^2/\kappa^2},$$

where $T_{n,i} \triangleq \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\}$ is the number of selections of arm i before round n. For the sake of simplicity, we consider improper Gaussian priors with $\mu_{1,i} = 0$ and $\sigma_{1,i} = +\infty$ for all $i \in \mathcal{A}$, for which

$$\mu_{n,i} = \frac{1}{T_{n,i}} \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell} \quad \text{and} \quad \sigma^2_{n,i} = \frac{\sigma^2}{T_{n,i}}.$$

Observe that in this case the posterior mean $\mu_{n,i}$ coincides with the empirical mean.

Beta-Bernoulli model For Bernoulli bandits with a uniform ($\mathcal{B}eta(1,1)$) prior on each mean, the posterior distribution of μ_i at round n is a Beta distribution with shape parameters $\alpha_{n,i} = \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell} + 1$ and $\beta_{n,i} = T_{n,i} - \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell, I_\ell} + 1$.

Now we briefly recall TTTS and introduce T3C. The pseudo-code of TTTS and T3C are shown in Algorithm 2.

Description of TTTS At each time step n, TTTS has two potential actions: (1) with probability β, a parameter vector $\boldsymbol{\theta}$ is sampled from Π_n, and TTTS chooses to play $I_n^{(1)} \triangleq \arg\max_{i \in \mathcal{A}} \theta_i$, (2) and with probability $1-\beta$, the algorithm continues sampling new $\boldsymbol{\theta}'$ until we obtain a *challenger* $I_n^{(2)} \triangleq \arg\max_{i \in \mathcal{A}} \theta'_i$ that is different from $I_n^{(1)}$, and TTTS chooses to play $I_n^{(2)}$.

Description of T3C One drawback of TTTS is that, in practice, when the posteriors become concentrated, it takes many Thompson samples before the challenger $I_n^{(2)}$ is obtained. We thus propose a variant of TTTS, called T3C, which alleviates this computational burden. Instead of re-sampling from the posterior until a different candidate appears, we define the challenger as the arm that has the lowest *transportation cost* $W_n(I_n^{(1)}, i)$ with respect to the first candidate (with ties broken uniformly at random).

Let $\mu_{n,i}$ be the empirical mean of arm i and $\mu_{n,i,j} \triangleq (T_{n,i}\mu_{n,i} + T_{n,j}\mu_{n,j})/(T_{n,i} + T_{n,j})$, then we define

$$W_n(i,j) \triangleq \begin{cases} 0 & \text{if } \mu_{n,j} \geq \mu_{n,i}, \\ W_{n,i,j} + W_{n,j,i} & \text{otherwise,} \end{cases} \tag{7.1}$$

where $W_{n,i,j} \triangleq T_{n,i} d\left(\mu_{n,i}, \mu_{n,i,j}\right)$ for any i, j and $d(\mu; \mu')$ denotes the Kullback-Leibler between the distribution with mean μ and that of mean μ'. In the Gaussian case, $d(\mu; \mu') = (\mu - \mu')^2/(2\sigma^2)$ while in the Bernoulli case $d(\mu; \mu') = \mu \ln(\mu/\mu') + (1-\mu)\ln(1-\mu)/(1-\mu')$. In particular, for Gaussian bandits

$$W_n(i,j) = \frac{(\mu_{n,i} - \mu_{n,j})^2}{2\sigma^2(1/T_{n,i} + 1/T_{n,j})} \mathbb{1}\{\mu_{n,j} < \mu_{n,i}\}.$$

Note that under the Gaussian model with improper priors, one should pull each arm once at the beginning for the sake of obtaining proper posteriors.

Algorithm 2 Sampling rule (TTTS/T3C)

Input: β
for $n \leftarrow 1, 2, \cdots$ **do**
  sample $\boldsymbol{\theta} \sim \Pi_n$
  $I^{(1)} \leftarrow \arg\max_{i \in \mathcal{A}} \theta_i$
  sample $b \sim \mathcal{B}ern(\beta)$
  if $b = 1$ **then**
    evaluate arm $I^{(1)}$
  else
    repeat sample $\boldsymbol{\theta}' \sim \Pi_n$
    $I^{(2)} \leftarrow \arg\max_{i \in \mathcal{A}} \theta'_i$ } TTTS
    until $I^{(2)} \neq I^{(1)}$
    $I^{(2)} \leftarrow \arg\min_{i \neq I^{(1)}} W_n(I^{(1)}, i)$, cf. (7.1) T3C
    evaluate arm $I^{(2)}$
  end if
  update mean and variance
  $t = t + 1$
end for

7.2.2 Rationale for T3C

In order to explain how T3C can be seen as an approximation of the re-sampling performed by TTTS, we first need to define the *optimal action probabilities*.

Optimal action probability The optimal action probability $a_{n,i}$ is defined as the posterior probability that arm i is optimal. Formally, letting Θ_i be the subset of Θ such that arm i is the optimal arm,

$$\Theta_i \triangleq \left\{ \boldsymbol{\theta} \in \Theta \;\middle|\; \theta_i > \max_{j \neq i} \theta_j \right\},$$

then we define

$$a_{n,i} \triangleq \Pi_n(\Theta_i) = \int_{\Theta_i} \pi_n(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{7.2}$$

With this notation, one can show that under TTTS,

$$\Pi_n\left(I_n^{(2)} = j | I_n^{(1)} = i\right) = \frac{a_{n,j}}{\sum_{k \neq i} a_{n,k}}. \tag{7.3}$$

Furthermore, when i coincides with the empirical best mean (and this will often be the case for $I_n^{(1)}$ when n is large due to posterior convergence) one can write

$$a_{n,j} \simeq \Pi_n\left(\theta_j \geq \theta_i\right) \simeq \exp\left(-W_n(i,j)\right),$$

where the last step is justified in Lemma 6 in the Gaussian case (and Lemma 32 in Appendix 7.I.3 in the Bernoulli case). Hence, T3C replaces sampling from the distribution (7.3) by an approx-

imation of its mode which is *easy to compute*. Note that directly computing the mode would require to compute $a_{n,j}$, which is much more costly than the computation of $W_n(i,j)$[3].

7.2.3 Stopping and recommendation rules

In order to use TTTS or T3C as the sampling rule for fixed-confidence BAI, we need to additionally define stopping and recommendation rules. While Qin, Klabjan and Russo, 2017 suggest to couple TTEI with the "frequentist" Chernoff stopping rule (Garivier and Kaufmann, 2016), we propose in this section natural Bayesian stopping and recommendation rules. They both rely on the optimal action probabilities defined in (7.2).

Bayesian recommendation rule At time step n, a natural candidate for the best arm is the arm with largest optimal action probability, hence we define

$$J_n \triangleq \arg\max_{i \in \mathcal{A}} a_{n,i} \,.$$

Bayesian stopping rule In view of the recommendation rule, it is natural to stop when the posterior probability that the recommended action is optimal is large, and exceeds some threshold $c_{n,\delta}$ which gets close to 1. Hence our Bayesian stopping rule is

$$\tau_\delta \triangleq \inf\left\{ n \in \colon \max_{i \in \mathcal{A}} a_{n,i} \geq c_{n,\delta} \right\} . \tag{7.4}$$

Links with frequentist counterparts Using the transportation cost $W_n(i,j)$ defined in (7.1), the Chernoff stopping rule of Garivier and Kaufmann, 2016 can actually be rewritten as

$$\tau_\delta^{\text{Ch.}} \triangleq \inf\left\{ n \in \mathbb{N} : \max_{i \in \mathcal{A}} \min_{j \in \mathcal{A} \setminus \{i\}} W_n(i,j) > d_{n,\delta} \right\} . \tag{7.5}$$

This stopping rule is coupled with the recommendation rule $J_n = \arg\max_i \mu_{n,i}$.

As explained in that paper, $W_n(i,j)$ can be interpreted as a (log) Generalized Likelihood Ratio statistic for rejecting the hypothesis $\mathcal{H}_0 : (\mu_i < \mu_j)$. Through our Bayesian lens, we rather have in mind the approximation $\Pi_n(\theta_j > \theta_i) \simeq \exp\{-W_n(i,j)\}$, valid when $\mu_{n,i} > \mu_{n,j}$, which permits to analyze the two stopping rules using similar tools, as will be seen in the proof of Theorem 7.3.

As shown later in Sec. 7.4, τ_δ and $\tau_\delta^{\text{Ch.}}$ prove to be fairly similar for some corresponding choices of the thresholds $c_{n,\delta}$ and $d_{n,\delta}$. This similarity endorses the use of the Chernoff stopping rule in practice, which does not require the (heavy) computation of optimal action probabilities. Still, our sample complexity analysis applies to the two stopping rules, and we believe that a frequentist sample complexity analysis of a fully Bayesian-flavored BAI strategy is a nice theoretical contribution.

[3] TTPS (Russo, 2016) also requires the computation of $a_{n,i}$, thus we do not report simulations for it in Sec. 7.6.

Useful notation We follow the notation of Russo (2016) and define the following measures of effort allocated to arm i up to time n,

$$\psi_{n,i} \triangleq \mathbb{P}\left[I_n = i | \mathcal{F}_{n-1}\right] \quad \text{and} \quad \Psi_{n,i} \triangleq \sum_{l=1}^{n} \psi_{l,i}.$$

In particular, for TTTS we have

$$\psi_{n,i} = \beta a_{n,i} + (1-\beta) a_{n,i} \sum_{j \neq i} \frac{a_{n,j}}{1 - a_{n,j}},$$

while for T3C

$$\psi_{n,i} = \beta a_{n,i} + (1-\beta) \sum_{j \neq i} a_{n,j} \frac{\mathbb{1}\{W_n(j,i) = \min_{k \neq j} W_n(j,k)\}}{\# \left| \arg\min_{k \neq j} W_n(j,k) \right|}.$$

7.3 Two Related Optimality Notions

In the fixed-confidence setting, we aim for building δ-correct strategies, i.e. strategies that identify the best arm with high confidence on any problem instance.

Definition 7.1. A strategy (I_n, J_n, τ) is δ-correct if for all bandit models $\boldsymbol{\mu}$ with a unique optimal arm, it holds that $\mathbb{P}_{\boldsymbol{\mu}}\left[J_\tau \neq I^\star \wedge \tau < \infty\right] \leq \delta$.

Among δ-correct strategies, we seek the one with the smallest sample complexity $\mathbb{E}\left[\tau_\delta\right]$. So far, TTTS has not been analyzed in terms of sample complexity; Russo (2016) focuses on posterior consistency and optimal convergence rates. Interestingly, both the smallest possible sample complexity and the fastest rate of posterior convergence can be expressed in terms of the following quantities.

Definition 7.2. Let $\Sigma_K = \{\boldsymbol{\omega} : \sum_{k=1}^K \omega_k = 1, \omega_k \geq 0\}$ and define for all $i \neq I^\star$

$$C_i(\omega, \omega') \triangleq \min_{x \in \mathcal{I}} \ \omega d(\mu_{I^\star}; x) + \omega' d(\mu_i; x),$$

where $d(\mu, \mu')$ is the KL-divergence defined above and $\mathcal{I} = \mathbb{R}$ in the Gaussian case and $\mathcal{I} = [0,1]$ in the Bernoulli case. We define

$$\begin{aligned} \Gamma^\star &\triangleq \max_{\boldsymbol{\omega} \in \Sigma_K} \min_{i \neq I^\star} C_i(\omega_{I^\star}, \omega_i), \\ \Gamma^\star_\beta &\triangleq \max_{\substack{\boldsymbol{\omega} \in \Sigma_K \\ \omega_{I^\star} = \beta}} \min_{i \neq I^\star} C_i(\omega_{I^\star}, \omega_i). \end{aligned} \tag{7.6}$$

The quantity $C_i(\omega_{I^\star}, \omega_i)$ can be interpreted as a "transportation cost"[4] from the original bandit instance $\boldsymbol{\mu}$ to an alternative instance in which the mean of arm i is larger than that of $I^\star$, when the proportion of samples allocated to each arm is given by the vector $\boldsymbol{\omega} \in \Sigma_K$. As shown by Russo, 2016, the $\boldsymbol{\omega}$ that maximizes (7.6) is unique, which allows us to define the *β-optimal allocation* $\boldsymbol{\omega}^\beta$ in the following proposition.

[4] for which $W_n(I^\star, i)$ is an empirical counterpart

Proposition 1. *There is a unique solution* $\boldsymbol{\omega}^\beta$ *to the optimization problem* (7.6) *satisfying* $\omega^\beta_{I^\star} = \beta$*, and for all* $i, j \neq I^\star$*,* $C_i(\beta, \omega^\beta_i) = C_j(\beta, \omega^\beta_j)$.

For models with more than two arms, there is no closed form expression for $\Gamma^\star_\beta$ or $\Gamma^\star$, even for Gaussian bandits with variance σ^2 for which we have

$$\Gamma^\star_\beta = \max_{\boldsymbol{\omega}: \omega_{I^\star} = \beta} \min_{i \neq I^\star} \frac{(\mu_{I^\star} - \mu_i)^2}{2\sigma^2(1/\omega_i + 1/\beta)}.$$

Bayesian β-optimality Russo (2016) proves that any sampling rule allocating a fraction β to the optimal arm ($\Psi_{n,I^\star}/n \to \beta$) satisfies $1 - a_{n,I^\star} \geq e^{-n(\Gamma^\star_\beta + o(1))}$ (a.s.).We define a *Bayesian β-optimal* sampling rule as a sampling rule matching this lower bound, i.e. satisfying $\Psi_{n,I^\star}/n \to \beta$ and $1 - a_{n,I^\star} \leq e^{-n(\Gamma^\star_\beta + o(1))}$.

Russo (2016) proves that TTTS with parameter β is Bayesian β-optimal. However, the result is valid only under strong regularity assumptions, excluding the two practically important cases of Gaussian and Bernoulli bandits. In this chapter, we complete the picture by establishing Bayesian β-optimality for those models in Sec. 7.5. For the Gaussian bandit, Bayesian β-optimality was established for TTEI by Qin, Klabjan and Russo, 2017 with Gaussian priors, but this remained an open problem for TTTS.

A fundamental ingredient of these proofs is to establish the convergence of the allocation of measurement effort to the β-optimal allocation: $\Psi_{n,i}/n \to \omega^\beta_i$ for all i, which is equivalent to $T_{n,i}/n \to \omega^\beta_i$ (cf. Lemma 8).

β-optimality in the fixed-confidence setting In the fixed confidence setting, the performance of an algorithm is evaluated in terms of sample complexity. A lower bound given by Garivier and Kaufmann, 2016 states that any δ-correct strategy satisfies $\mathbb{E}[\tau_\delta] \geq (\Gamma^\star)^{-1} \ln(1/(3\delta))$.

Observe that $\Gamma^\star = \max_{\beta \in [0,1]} \Gamma^\star_\beta$. Using the same lower bound techniques, one can also prove that under any δ-correct strategy satisfying $T_{n,I^\star}/n \to \beta$,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\ln(1/\delta)} \geq \frac{1}{\Gamma^\star_\beta}.$$

This motivates the relaxed optimality notion that we introduce in this chapter: A BAI strategy is called *asymptotically β-optimal* if it satisfies

$$\frac{T_{n,I^\star}}{n} \to \beta \quad \text{and} \quad \limsup_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\ln(1/\delta)} \leq \frac{1}{\Gamma^\star_\beta}.$$

In this chapter, we provide the first sample complexity analysis of a BAI algorithm based on TTTS (with the stopping and recommendation rules described in Sec. 7.2), establishing its asymptotic β-optimality.

As already observed by Qin, Klabjan and Russo, 2017, any sampling rule converging to the β-optimal allocation (i.e. satisfying $T_{n,i}/n \to w_i^\beta$ for all i) can be shown to satisfy

$$\limsup_{\delta \to 0} \frac{\tau_\delta}{\ln(1/\delta)} \leq (\Gamma_\beta^\star)^{-1}$$

almost surely, when coupled with the Chernoff stopping rule. The fixed confidence optimality that we define above is stronger as it provides guarantees on $\mathbb{E}[\tau_\delta]$.

7.4 Fixed-Confidence Analysis

In this section, we consider Gaussian bandits and the Bayesian rules using an improper prior on the means. We state our main result below, showing that TTTS and T3C are asymptotically β-optimal in the fixed confidence setting, when coupled with appropriate stopping and recommendation rules.

Theorem 7.2. *With $\mathcal{C}^{g_G}$ the function defined in Corollary 10 of Kaufmann and Koolen, 2018, which satisfies $\mathcal{C}^{g_G}(x) \simeq x + \ln(x)$, we introduce the threshold*

$$d_{n,\delta} = 4\ln(4 + \ln(n)) + 2\mathcal{C}^{g_G}\left(\frac{\ln((K-1)/\delta)}{2}\right). \tag{7.7}$$

The TTTS and T3C sampling rules coupled with either

- *the Bayesian stopping rule (7.4) with threshold*

$$c_{n,\delta} = 1 - \frac{1}{\sqrt{2\pi}} e^{-\left(\sqrt{d_{n,\delta}} + \frac{1}{\sqrt{2}}\right)^2}$$

  and recommendation rule $J_t = \arg\max_i a_{n,i}$, or
- *the Chernoff stopping rule (7.5) with threshold $d_{n,\delta}$ and recommendation rule $J_t = \arg\max_i \mu_{n,i}$,*

form a δ-correct BAI strategy. Moreover, if all the arms means are distinct, it satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \frac{1}{\Gamma_\beta^\star}.$$

We now give the proof of Theorem 7.2, which is divided into three parts. The **first step** of the analysis is to prove the δ-correctness of the studied BAI strategies.

Theorem 7.3. *Regardless of the sampling rule, the stopping rule (7.4) with the threshold $c_{n,\delta}$ and the Chernoff stopping rule (7.5) with threshold $d_{n,\delta}$ defined in (7.7) satisfy $\mathbb{P}[\tau_\delta < \infty \wedge J_{\tau_\delta} \neq I^\star] \leq \delta$.*

To prove that TTTS and T3C allow to reach a β-optimal sample complexity, one needs to quantify how fast the measurement effort for each arm is concentrating to its corresponding optimal weight. For this purpose, we introduce the random variable

$$T_\beta^\varepsilon \triangleq \inf\left\{N \in : \max_{i \in \mathcal{A}} |T_{n,i}/n - \omega_i^\beta| \leq \varepsilon, \forall n \geq N\right\}.$$

The **second step** of our analysis is a sufficient condition for β-optimality, stated in Lemma 4. Its proof is given in Appendix 7.F. The same result was proven for the Chernoff stopping rule by Qin, Klabjan and Russo, 2017.

Lemma 4. *Let $\delta, \beta \in (0,1)$. For any sampling rule which satisfies $\mathbb{E}\left[T_\beta^\varepsilon\right] < \infty$ for all $\varepsilon > 0$, we have*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}\left[\tau_\delta\right]}{\log(1/\delta)} \leq \frac{1}{\Gamma_\beta^\star},$$

if the sampling rule is coupled with stopping rule (7.4),

Finally, it remains to show that TTTS and T3C meet the sufficient condition, and therefore the **last step**, which is the core component and the most technical part our analysis, consists of showing the following.

Theorem 7.5. *Under TTTS or T3C, $\mathbb{E}\left[T_\beta^\varepsilon\right] < +\infty$.*

In the rest of this section, we prove Theorem 7.3 and sketch the proof of Theorem 7.5. But we first highlight some important ingredients for these proofs.

7.4.1 Core ingredients

Our analysis hinges on properties of the Gaussian posteriors, in particular on the following tail bounds, which follow from Lemma 1 of Qin, Klabjan and Russo, 2017.

Lemma 6. *For any $i, j \in \mathcal{A}$, if $\mu_{n,i} \leq \mu_{n,j}$*

$$\Pi_n\left[\theta_i \geq \theta_j\right] \leq \frac{1}{2} \exp\left\{-\frac{\left(\mu_{n,j} - \mu_{n,i}\right)^2}{2\sigma_{n,i,j}^2}\right\}, \tag{7.8}$$

$$\Pi_n\left[\theta_i \geq \theta_j\right] \geq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\left(\mu_{n,j} - \mu_{n,i} + \sigma_{n,i,j}\right)^2}{2\sigma_{n,i,j}^2}\right\}, \tag{7.9}$$

where $\sigma_{n,i,j}^2 \triangleq \sigma^2/T_{n,i} + \sigma^2/T_{n,j}$.

This lemma is crucial to control $a_{n,i}$ and $\psi_{n,i}$, the optimal action and selection probabilities.

7.4.2 Proof of Theorem 7.3

We upper bound the desired probability as follows

$$\begin{aligned}
\mathbb{P}\left[\tau_\delta < \infty \wedge J_{\tau_\delta} \neq I^\star\right] &\leq \sum_{i \neq I^\star} \mathbb{P}\left[\exists n \in \colon a_{n,i} > c_{n,\delta}\right] \\
&\leq \sum_{i \neq I^\star} \mathbb{P}\left[\exists n \in \colon \Pi_n(\theta_i \geq \theta_{I_\star}) > c_{n,\delta}, \mu_{n,I^\star} \leq \mu_{n,i}\right] \\
&\leq \sum_{i \neq I^\star} \mathbb{P}\left[\exists n \in \colon 1 - c_{n,\delta} > \Pi_n(\theta_{I^\star} > \theta_i), \mu_{n,I^\star} \leq \mu_{n,i}\right].
\end{aligned}$$

The second step uses the fact that as $c_{n,\delta} \geq 1/2$, a necessary condition for $\Pi_n(\theta_i \geq \theta_{I_\star}) \geq c_{n,\delta}$ is that $\mu_{n,i} \geq \mu_{n,I_\star}$. Now using the lower bound (7.9), if $\mu_{n,I^\star} \leq \mu_{n,i}$, the inequality $1 - c_{n,\delta} > \Pi_n(\theta_{I^\star} > \theta_i)$ implies

$$\frac{(\mu_{n,i} - \mu_{n,I^\star})^2}{2\sigma^2_{n,i,I^\star}} \geq \left(\sqrt{\ln \frac{1}{\sqrt{2\pi}(1 - c_{n,\delta})}} - \frac{1}{\sqrt{2}} \right)^2 = d_{n,\delta},$$

where the equality follows from the expression of $c_{n,\delta}$ as function of $d_{n,\delta}$. Hence to conclude the proof it remains to check that

$$\mathbb{P}\left[\exists n \in \mu_{n,i} \geq \mu_{n,I^\star}, \frac{(\mu_{n,i} - \mu_{n,I^\star})^2}{2\sigma^2_{n,i,I^\star}} \geq d_{n,\delta} \right] \leq \frac{\delta}{K-1}. \tag{7.10}$$

To prove this, we observe that for $\mu_{n,i} \geq \mu_{n,I^\star}$,

$$\begin{aligned} \frac{(\mu_{n,i} - \mu_{n,I^\star})^2}{2\sigma^2_{n,i,I^\star}} &= \inf_{\theta_i < \theta_{I^\star}} T_{n,i} d(\mu_{n,i}; \theta_i) + T_{n,I^\star} d(\mu_{n,I^\star}; \theta_{I^\star}) \\ &\leq T_{n,i} d(\mu_{n,i}; \mu_i) + T_{n,I^\star} d(\mu_{n,I^\star}; \mu_{I^\star}). \end{aligned}$$

Corollary 10 of Kaufmann and Koolen, 2018 then allows us to upper bound the probability

$$\mathbb{P}\left[\exists n \in : T_{n,i} d(\mu_{n,i}; \mu_i) + T_{n,I^\star} d(\mu_{n,I^\star}, \mu_{I^\star}) \geq d_{n,\delta} \right]$$

by $\delta/(K-1)$ for the choice of threshold given in (7.7), which completes the proof that the stopping rule (7.4) is δ-correct. The fact that the Chernoff stopping rule with the above threshold $d_{n,\delta}$ given above is δ-correct straightforwardly follows from (7.10).

7.4.3 Sketch of the proof of Theorem 7.5

We present a unified proof sketch of Theorem 7.5 for TTTS and T3C. While the two analyses follow the same steps, some of the lemmas given below have different proofs for TTTS and T3C, which can be found in Appendix 7.D and 7.E respectively.

We first state two important concentration results, that hold under any sampling rule.

Lemma 7. *[Lemma 5 of Qin, Klabjan and Russo 2017] There exists a random variable W_1, such that for all $i \in \mathcal{A}$,*

$$\forall n \in , \quad |\mu_{n,i} - \mu_i| \leq \sigma W_1 \sqrt{\frac{\log(e + T_{n,i})}{1 + T_{n,i}}} \ a.s.,$$

and $\mathbb{E}\left[e^{\lambda W_1}\right] < \infty$ for all $\lambda > 0$.

Lemma 8. *There exists a random variable W_2, such that for all $i \in \mathcal{A}$,*

$$\forall n \in , |T_{n,i} - \Psi_{n,i}| \leq W_2 \sqrt{(n+1)\log(e^2 + n)} \ a.s.,$$

and $\mathbb{E}\left[e^{\lambda W_2}\right] < \infty$ for any $\lambda > 0$.

Lemma 7 controls the concentration of the posterior means towards the true means and Lemma 8 establishes that $T_{n,i}$ and $\Psi_{n,i}$ are close. Both results rely on uniform deviation inequalities for martingales.

Our analysis uses the same principle as that of TTEI: We establish that T^{ε}_{β} is upper bounded by some random variable N which is a polynomial of the random variables W_1 and W_2 introduced in the above lemmas, denoted by $\mathrm{Poly}(W_1, W_2) \triangleq \mathcal{O}(W_1^{c_1} W_2^{c_2})$, where c_1 and c_2 are two constants (that may depend on the arms' means and the constant hidden in the $\mathcal{O}$). As all exponential moments of W_1 and W_2 are finite, N has a finite expectation as well, concluding the proof.

The first step to exhibit such an upper bound N is to establish that every arm is pulled sufficiently often.

Lemma 9. *Under TTTS or T3C, there exists* $N_1 = \mathrm{Poly}(W_1, W_2)$ *s.t.*

$$\forall n \geq N_1, \forall i,\ T_{n,i} \geq \sqrt{\frac{n}{K}},\ a.s..$$

Due to the randomized nature of TTTS and T3C, the proof of Lemma 9 is significantly more involved than for a deterministic rule like TTEI. Intuitively, the posterior of each arm would be well concentrated once the arm is sufficiently pulled. If the optimal arm is under-sampled, then it would be chosen as the first candidate with large probability. If a sub-optimal arm is under-sampled, then its posterior distribution would possess a relatively wide tail that overlaps with or cover the somehow narrow tails of other overly-sampled arms. The probability of that sub-optimal arm being chosen as the challenger would be large enough then.

Combining Lemma 9 with Lemma 7 straightforwardly leads to the following result.

Lemma 10. *Under TTTS or T3C, fix a constant* $\varepsilon > 0$, *there exists* $N_2 = \mathrm{Poly}(1/\varepsilon, W_1, W_2)$ *s.t.* $\forall n \geq N_2, \forall i \in \mathcal{A}, \quad |\mu_{n,i} - \mu_i| \leq \varepsilon$.

We can then deduce a very nice property about the optimal action probability for sub-optimal arms from the previous two lemmas. Indeed, we can show that

$$\forall i \neq I^\star, \quad a_{n,i} \leq \exp\left\{-\frac{\Delta^2_{\min}}{16\sigma^2}\sqrt{\frac{n}{K}}\right\}$$

for n larger than some $\mathrm{Poly}(W_1, W_2)$, where $\Delta_{\min}$ is the smallest mean difference among all the arms.

Plugging this in the expression of $\psi_{n,i}$, one can easily quantify how fast $\psi_{n,I^\star}$ converges to β, which eventually yields the following result.

Lemma 11. *Under TTTS or T3C, fix* $\varepsilon > 0$, *then there exists* $N_3 = \mathrm{Poly}(1/\varepsilon, W_1, W_2)$ *s.t.* $\forall n \geq N_3$,

$$\left|\frac{T_{n,I^\star}}{n} - \beta\right| \leq \varepsilon.$$

The last, more involved, step is to establish that the fraction of measurement allocation to every sub-optimal arm i is indeed similarly close to its optimal proportion ω_i^β.

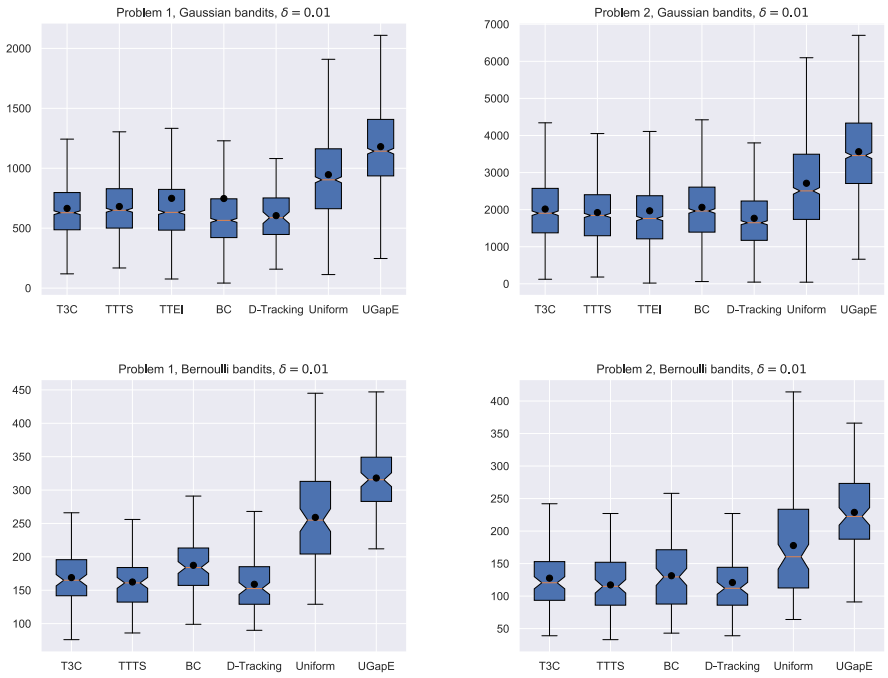


Figure 7.1: Black dots represent means and oranges lines represent medians.

Lemma 12. *Under TTTS or T3C, fix a constant $\varepsilon > 0$, there exists $N_4 = Poly(1/\varepsilon, W_1, W_2)$ s.t. $\forall n \geq N_4$,*

$$\forall i \neq I^\star, \quad \left| \frac{T_{n,i}}{n} - \omega_i^\beta \right| \leq \varepsilon.$$

The major step in the proof of Lemma 12 for each sampling rule, is to establish that if some arm is over-sampled, then its probability to be selected is exponentially small. Formally, we show that for n larger than some $\mathrm{Poly}(1/\varepsilon, W_1, W_2)$,

$$\frac{\Psi_{n,i}}{n} \geq \omega_i^\beta + \xi \quad \Rightarrow \quad \psi_{n,i} \leq \exp\left\{-f(n,\xi)\right\},$$

for some function $f(n,\xi)$ to be specified for each sampling rule, satisfying $f(n) \geq C_\xi \sqrt{n}$ (a.s.). This result leads to the concentration of $\Psi_{n,i}/n$, thus can be easily converted to the concentration of $T_{n,i}/n$ by Lemma 8.

Finally, Lemma 11 and Lemma 12 show that T_β^ε is upper bounded by $N \triangleq \max(N_3, N_4)$, which yields

$$\mathbb{E}[T_\beta^\varepsilon] \leq \max(\mathbb{E}[N_3], \mathbb{E}[N_4]) < \infty.$$

Sampling rule	Execution time (s)
T3C	1.6×10^{-5}
TTTS	2.3×10^{-4}
TTEI	1×10^{-5}
BC	1.4×10^{-5}
D-Tracking	1.3×10^{-3}
Uniform	6×10^{-6}
UGapE	5×10^{-6}

Table 7.1: Average execution time in seconds for different sampling rules.

7.5 Optimal Posterior Convergence

Recall that $a_{n,I^\star}$ denotes the posterior mass assigned to the event that action $I^\star$ (i.e. the true optimal arm) is optimal at time n. As the number of observations tends to infinity, we want the posterior distribution to converge to the truth. In this section we show equivalently that the posterior mass on the complementary event, $1 - a_{n,I^\star}$, the event that arm $I^\star$ is not optimal, converges to zero at an exponential rate, and that it does so at optimal rate $\Gamma^\star_\beta$.

Russo (2016) proves a similar theorem under three confining boundedness assumptions (see Russo 2016, Assumption 1) on the parameter space, the prior density and the (first derivative of the) log-normalizer of the exponential family. Hence, the theorems in Russo, 2016 do not apply to the two bandit models most used in practice, which we consider in this chapter: the Gaussian and Bernoulli model.

In the first case, the parameter space is unbounded, in the latter model, the derivative of the log-normalizer (which is $e^\eta/(1+e^\eta)$) is unbounded. Here we provide a theorem, proving that under TTTS, the optimal, exponential posterior convergence rates are obtained for the Gaussian model with uninformative (improper) Gaussian priors (proof in Appendix 7.H), and the Bernoulli model with $\mathcal{B}eta(1,1)$ priors (proof in Appendix 7.I).

Theorem 7.13. *Under TTTS, for Gaussian bandits with improper Gaussian priors and for Bernoulli bandits with uniform priors, it holds almost surely that*

$$\lim_{n\to\infty} -\frac{1}{n}\log(1 - a_{n,I^\star}) = \Gamma^\star_\beta.$$

7.6 Numerical Illustrations

This section is aimed at illustrating our theoretical results and supporting the practical use of Bayesian sampling rules for fixed-confidence BAI.

We experiment with 3 Bayesian sampling rules: T3C, TTTS and TTEI with $\beta = 1/2$, against the Direct Tracking (D-Tracking) of Garivier and Kaufmann, 2016 (which is adaptive to β), UGapE of Gabillon, Ghavamzadeh and Lazaric, 2012, and a uniform baseline. To make fair

comparisons, we use the stopping rule (7.5) and associated recommendation rule for all of the sampling rules except for UGapE which has its own stopping rule.

We further include a top-two variant of the Best Challenger (BC) heuristic (see Ménard, 2019). BC selects the empirical best arm $\widehat{I}_n$ with probability β and the maximizer of $W_n(\widehat{I}_n, j)$ with probability $1-\beta$, but also performs forced exploration (selecting any arm sampled less than $\sqrt{n}$ times at round n). T3C can thus be viewed as a variant of BC in which no forced exploration is needed to converge to $\boldsymbol{\omega}^\beta$, due to the noise added by replacing $\widehat{I}_n$ with $I_n^{(1)}$. This randomization is crucial as BC without forced exploration can fail: we observed that on bandit instances with two identical sub-optimal arms, BC has some probability to alternate forever between these two arms and never stop.

We consider two simple instances with arm means given by $\boldsymbol{\mu}_1 = [0.5\ 0.9\ 0.4\ 0.45\ 0.44999]$, and $\boldsymbol{\mu}_2 = [1\ 0.8\ 0.75\ 0.7]$. We run simulations for both Gaussian ($\sigma = 1$) and Bernoulli bandits, with a risk parameter $\delta = 0.01$. Fig. 7.1 reports the empirical distribution of τ_δ under the different sampling rules, estimated over 1000 independent runs. We also indicate the values of $N^\star \triangleq \log(1/\delta)/\Gamma^\star$ (resp. $N^\star_{0.5} \triangleq \log(1/\delta)/\Gamma^\star_{0.5}$), the theoretical minimal number of samples needed for any strategy (resp. any 1/2-optimal strategy). In Appendix 7.C, we further illustrate how the empirical stopping time of T3C matches the theoretical one.

These figures provide several insights: (1) T3C is competitive with, and sometimes slightly better than TTTS/TTEI in terms of sample complexity. (2) The UGapE algorithm has a larger sample complexity than the uniform sampling rule, which highlights the importance of the stopping rule in the fixed-confidence setting. (3) The fact that D-Tracking performs best is not surprising, since it converges to $\boldsymbol{\omega}^{\beta^\star}$ and achieves minimal sample complexity. However, in terms of computation time, D-Tracking is much worse than others, as shown in Table 7.1, which reports the average execution time of one step of each sampling rule for $\boldsymbol{\mu}_1$ in the Gaussian case. (4) TTTS also suffers from computational costs, whose origins are explained in Sec. 7.2, unlike T3C or TTEI. Although TTEI is already computationally more attractive than TTTS, its practical benefits are limited to the Gaussian case, since the *Expected Improvement* (EI) does not have a closed form beyond this case and its approximation would be costly. In contrast, T3C can be applied for other distributions.

7.7 Conclusion

We have advocated the use of Bayesian sampling rules for BAI. In particular, we proved that TTTS and a computationally advantageous approach T3C, are both β-optimal in the fixed-confidence setting, for Gaussian bandits. We further extended the Bayesian optimality properties (Russo, 2016) to more practical choices of models and prior distributions. In order to be optimal, these sampling rules would need the oracle tuning $\beta^\star = \arg\max_{\beta\in[0,1]} \Gamma^\star_\beta$, which is not feasible. In future work, we will investigate the efficient online tuning of β to circumvent this issue. We also wish to obtain explicit finite-time sample complexity bound for these Bayesian strategies, and justify the use of these appealing anytime sampling rules in the fixed-budget setting. The latter is often more plausible in application scenarios such as BAI for automated machine learning (Li et al., 2017; Shang, Kaufmann and Valko, 2019).

7.A Outline

The appendix of this chapter is organized as follows:

Appendix 7.C provides some further numerical illustration for better understanding of T3C.
Appendix 7.D provides the complete fixed-confidence analysis of TTTS (Gaussian case).
Appendix 7.E provides the complete fixed-confidence analysis of T3C (Gaussian case).
Appendix 7.F is dedicated to Lemma 4.
Appendix 7.G is dedicated to crucial technical lemmas.
Appendix 7.H is the proof to the posterior convergence Theorem 7.27 (Gaussian case).
Appendix 7.I is the proof to the posterior convergence Theorem 7.34 (Beta-Bernoulli case).

7.B Useful Notation

In this section, we provide a list of useful notation that is applied in appendices (including reminders of previous notation in the main text and some new ones).

- Recall that $d(\mu_1;\mu_2)$ denotes the KL-divergence between two distributions parametrized by their means μ_1 and μ_2. For Gaussian distributions, we know that

$$d(\mu_1;\mu_2) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2}.$$

  When it comes to Bernoulli distributions, we denote this with kl, i.e.

$$kl(\mu_1;\mu_2) = \mu_1 \ln\left(\frac{\mu_1}{\mu_2}\right) + (1-\mu_1)\ln\left(\frac{1-\mu_1}{1-\mu_2}\right).$$

- $\mathcal{B}eta(\cdot,\cdot)$ denotes a Beta distribution.
- $\mathcal{B}ern(\cdot)$ denotes a Bernoulli distribution.
- $\mathcal{B}(\cdot)$ denotes a Binomial distribution.
- $\mathcal{N}(\cdot,\cdot)$ denotes a normal distribution.
- $Y_{n,i}$ is the reward of arm i at time n.
- Y_{n,I_n} is the observation of the sampling rule at time n.
- $\mathcal{F}_n \triangleq \sigma(I_1, Y_{1,I_1}, I_2, Y_{2,I_2}, \cdots, I_n, Y_{n,I_n})$ is the filtration generated by the first n observations.
- $\psi_{n,i} \triangleq \mathbb{P}\left[I_n = i | \mathcal{F}_{n-1}\right]$.
- $\Psi_{n,i} \triangleq \sum_{l=1}^{n} \psi_{l,i}$.
- For the sake of simplicity, we further define $\overline{\psi}_{n,i} \triangleq \frac{\Psi_{n,i}}{n}$.
- $T_{n,i}$ is the number of pulls of arm i before round n.
- $\mathbf{T}_n$ denotes the vector of the number of arm selections.
- $I_n^\star \triangleq \arg\max_{i\in\mathcal{A}} \mu_{n,i}$ denotes the empirical best arm at time n.
- For any $a, b > 0$, define a function $C_{a,b}$ s.t. $\forall y$,

$$C_{a,b}(y) \triangleq (a+b-1)kl(\frac{a-1}{a+b-1}; y).$$

- We define the minimum and the maximum means gap as

$$\Delta_{\min} \triangleq \min_{i \neq j} |\mu_i - \mu_j|; \qquad \Delta_{\max} \triangleq \max_{i \neq j} |\mu_i - \mu_j|.$$

- We introduce two indices

$$J_n^{(1)} \triangleq \arg\max_j a_{n,j}; \qquad J_n^{(2)} \triangleq \arg\max_{j \neq J_n^{(1)}} a_{n,j}.$$

  Note that $J_n^{(1)}$ coincides with the Bayesian recommendation index J_n.
- Two real-valued sequences (a_n) and (b_n) are are said to be logarithmically equivalent if

$$\lim_{n \to \infty} \frac{1}{n} \log\left(\frac{a_n}{b_n}\right) = 0,$$

  and we denote this by $a_n \doteq b_n$.

7.C Empirical vs. theoretical sample complexity

In Fig. 7.2, we plot expected stopping time of T3C for $\delta = 0.01$ as a function of $1/\Gamma_\beta^\star$ on 100 randomly generated problem instances. We see on this plot that the empirical stopping time has the right linear scaling in $1/\Gamma_\beta^\star$ (ignoring a few outliers).

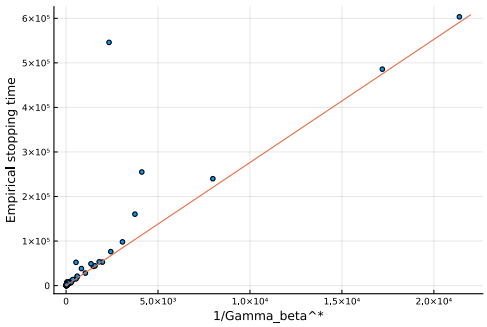


Figure 7.2: dots: empirical sample complexity, solid line: theoretical sample complexity.

7.D Fixed-Confidence Analysis for TTTS

This section is entirely dedicated to TTTS.

7.D.1 Technical novelties and some intuitions

Before we start the analysis, we first highlight some technical novelties and intuitions. The main novelty in our analysis is the proof of Lemma 9, establishing that all arms are sufficiently explored

by our randomized strategies. Although Qin, Klabjan and Russo, 2017 indeed establish a similar result, our proof is much more intricate due to the randomized nature of the two candidate arms $I^{(1)}$ and $I^{(2)}$ for TTTS (resp. $I^{(1)}$ for T3C). In the proof of Lemma 9 (in Appendix 7.D.2 and Appendix 7.E.1 respectively), we need to add a sort of 'extra layer' where we first study the behaviour of $J^{(1)}$ and $J^{(2)}$ for TTTS (resp. $J^{(1)}$ and $\widetilde{J^{(2)}}$ for T3C). We show in Lemma 14 (resp. Lemma 21 for T3C) that if there exists some under-sampled arm, then either $J^{(1)}$ or $J^{(2)}$ is also under-sampled. A link between I and J is then established using the expression of $\psi_{n,i}$, which also allows to upper bound the optimal action probability with a known rate (see Lemma 17).

7.D.2 Sufficient exploration of all arms proof of Lemma 9 under TTTS

To prove this lemma, we introduce the two following sets of indices for a given $L > 0$: $\forall n \in \mathbb{N}$ we define

$$U_n^L \triangleq \{i : T_{n,i} < \sqrt{L}\},$$
$$V_n^L \triangleq \{i : T_{n,i} < L^{3/4}\}.$$

It is seemingly non trivial to manipulate directly TTTS's candidate arms, we thus start by connecting TTTS with TTPS (top two probability sampling). TTPS is another sampling rule presented by Russo, 2016 for which the two candidate samples are defined as in Appendix 7.B, we recall them in the following.

$$J_n^{(1)} \triangleq \arg\max_j a_{n,j}, J_n^{(2)} \triangleq \arg\max_{j \neq J_n^{(1)}} a_{n,j}.$$

Lemma 9 is proved via the following sequence of lemmas.

Lemma 14. *There exists $L_1 = Poly(W_1)$ s.t. if $L > L_1$, for all n, $U_n^L \neq \varnothing$ implies $J_n^{(1)} \in V_n^L$ or $J_n^{(2)} \in V_n^L$.*

Proof. If $J_n^{(1)} \in V_n^L$, then the proof is finished. Now we assume that $J_n^{(1)} \in \overline{V_n^L}$, and we prove that $J_n^{(2)} \in V_n^L$.

Step 1 According to Lemma 7, there exists $L_2 = \text{Poly}(W_1)$ s.t. $\forall L > L_2, \forall i \in \overline{U_n^L}$,

$$\begin{aligned}
|\mu_{n,i} - \mu_i| &\leq \sigma W_1 \sqrt{\frac{\log(e + T_{n,i})}{1 + T_{n,i}}} \\
&\leq \sigma W_1 \sqrt{\frac{\log(e + \sqrt{L})}{1 + \sqrt{L}}} \\
&\leq \sigma W_1 \frac{\Delta_{\min}}{4\sigma W_1} = \frac{\Delta_{\min}}{4}.
\end{aligned}$$

The second inequality holds since $x \mapsto \frac{\log(e+x)}{1+x}$ is a decreasing function. The third inequality holds for a large $L > L_2$ with $L_2 = \ldots$.

Step 2 We now assume that $L > L_2$, and we define

$$\overline{J_n^\star} \triangleq \arg\max_{j \in \overline{U_n^L}} \mu_{n,j} = \arg\max_{j \in \overline{U_n^L}} \mu_j.$$

The last equality holds since $\forall j \in \overline{U_n^L}$, $|\mu_{n,i} - \mu_i| \leq \Delta_{\min}/4$. We show that there exists $L_3 = \mathrm{Poly}(W_1)$ s.t. $\forall L > L_3$,

$$\overline{J_n^\star} = J_n^{(1)}.$$

We proceed by contradiction, and suppose that $\overline{J_n^\star} \neq J_n^{(1)}$, then $\mu_{n,J_n^{(1)}} < \mu_{n,\overline{J_n^\star}}$, since $J_n^{(1)} \in \overline{V_n^L} \subset \overline{U_n^L}$. However, we have

$$\begin{aligned} a_{n,J_n^{(1)}} &= \Pi_n\left[\theta_{J_n^{(1)}} > \max_{j \neq J_n^{(1)}} \theta_j\right] \\ &\leq \Pi_n\left[\theta_{J_n^{(1)}} > \theta_{\overline{J_n^\star}}\right] \\ &\leq \frac{1}{2}\exp\left\{-\frac{(\mu_{n,J_n^{(1)}} - \mu_{n,\overline{J_n^\star}})^2}{2\sigma^2(1/T_{n,J_n^{(1)}} + 1/T_{n,\overline{J_n^\star}})}\right\}. \end{aligned}$$

The last inequality uses the Gaussian tail inequality (7.8) of Lemma 6. On the other hand,

$$\begin{aligned} |\mu_{n,J_n^{(1)}} - \mu_{n,\overline{J_n^\star}}| &= |\mu_{n,J_n^{(1)}} - \mu_{J_n^{(1)}} + \mu_{J_n^{(1)}} - \mu_{\overline{J_n^\star}} + \mu_{\overline{J_n^\star}} - \mu_{n,\overline{J_n^\star}}| \\ &\geq |\mu_{J_n^{(1)}} - \mu_{\overline{J_n^\star}}| - |\mu_{n,J_n^{(1)}} - \mu_{J_n^{(1)}} + \mu_{\overline{J_n^\star}} - \mu_{n,\overline{J_n^\star}}| \\ &\geq \Delta_{\min} - \left(\frac{\Delta_{\min}}{4} + \frac{\Delta_{\min}}{4}\right) \\ &= \frac{\Delta_{\min}}{2}, \end{aligned}$$

and

$$\frac{1}{T_{n,J_n^{(1)}}} + \frac{1}{T_{n,\overline{J_n^\star}}} \leq \frac{2}{\sqrt{L}}.$$

Thus, if we take L_3 s.t.

$$\exp\left\{-\frac{\sqrt{L_3}\Delta_{\min}^2}{16\sigma^2}\right\} \leq \frac{1}{2K},$$

then for any $L > L_3$, we have

$$a_{n,J_n^{(1)}} \leq \frac{1}{2K} < \frac{1}{K},$$

which contradicts the definition of $J_n^{(1)}$. We now assume that $L > L_3$, thus $J_n^{(1)} = \overline{J_n^\star}$.

Step 3 We finally show that for L large enough, $J_n^{(2)} \in V_n^L$. First note that $\forall j \in \overline{V_n^L}$, we have

$$a_{n,j} \leq \Pi_n\left[\theta_j \geq \theta_{\overline{J_n^\star}}\right] \leq \exp\left\{-\frac{L^{3/4}\Delta_{\min}^2}{16\sigma^2}\right\}. \tag{7.11}$$

This last inequality can be proved using the same argument as Step 2. Now we define another index $J_n^\star \triangleq \arg\max_{j \in U_n^L} \mu_{n,j}$ and the quantity $c_n \triangleq \max(\mu_{n,J_n^\star}, \mu_{n,\overline{J_n^\star}})$. We can lower bound $a_{n,J_n^\star}$ as follows:

$$\begin{aligned} a_{n,J_n^\star} &\geq \Pi_n\left[\theta_{J_n^\star} \geq c_n\right] \prod_{j \neq J_n^\star} \Pi_n\left[\theta_j \leq c_n\right] \\ &= \Pi_n\left[\theta_{J_n^\star} \geq c_n\right] \prod_{j \neq J_n^\star; j \in U_n^L} \Pi_n\left[\theta_j \leq c_n\right] \prod_{j \in \overline{U_n^L}} \Pi_n\left[\theta_j \leq c_n\right] \\ &\geq \Pi_n\left[\theta_{J_n^\star} \geq c_n\right] \frac{1}{2^{K-1}}. \end{aligned}$$

Now there are two cases:

- If $\mu_{n,J_n^\star} > \mu_{n,\overline{J_n^\star}}$, then we have

$$\Pi_n\left[\theta_{J_n^\star} \geq c_n\right] = \Pi_n\left[\theta_{J_n^\star} \geq \mu_{n,J_n^\star}\right] \geq \frac{1}{2}.$$

- If $\mu_{n,J_n^\star} < \mu_{n,\overline{J_n^\star}}$, then we can apply the Gaussian tail bound (7.9) of Lemma 6, and we obtain

$$\begin{aligned} \Pi_n\left[\theta_{J_n^\star} \geq c_n\right] &= \Pi_n\left[\theta_{J_n^\star} \geq \mu_{n,\overline{J_n^\star}}\right] = \Pi_n\left[\theta_{J_n^\star} \geq \mu_{n,J_n^\star} + (\mu_{n,\overline{J_n^\star}} - \mu_{n,J_n^\star})\right] \\ &\geq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(1 - \frac{\sqrt{T_{n,J_n^\star}}}{\sigma}(\mu_{n,J_n^\star} - \mu_{n,\overline{J_n^\star}})\right)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(1 + \frac{\sqrt{T_{n,J_n^\star}}}{\sigma}(\mu_{n,\overline{J_n^\star}} - \mu_{n,J_n^\star})\right)^2\right\}. \end{aligned}$$

  On the other hand, by Lemma 7, we know that

$$\begin{aligned} |\mu_{n,J_n^\star} - \mu_{n,\overline{J_n^\star}}| &= |\mu_{n,J_n^\star} - \mu_{J_n^\star} + \mu_{J_n^\star} - \mu_{\overline{J_n^\star}} + \mu_{\overline{J_n^\star}} - \mu_{n,\overline{J_n^\star}}| \\ &\leq |\mu_{J_n^\star} - \mu_{\overline{J_n^\star}}| + \sigma W_1 \sqrt{\frac{\log(e + T_{n,J_n^\star})}{1 + T_{n,J_n^\star}}} + \sigma W_1 \sqrt{\frac{\log(e + T_{n,\overline{J_n^\star}})}{1 + T_{n,\overline{J_n^\star}}}} \\ &\leq |\mu_{J_n^\star} - \mu_{\overline{J_n^\star}}| + 2\sigma W_1 \sqrt{\frac{\log(e + T_{n,J_n^\star})}{1 + T_{n,J_n^\star}}} \\ &\leq \Delta_{\max} + 2\sigma W_1 \sqrt{\frac{\log(e + T_{n,J_n^\star})}{1 + T_{n,J_n^\star}}}. \end{aligned}$$

Therefore,

$$\begin{aligned}\Pi_n\left[\theta_{J_n^\star} \geq c_n\right] &\geq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(1+\frac{\sqrt{T_{n,J_n^\star}}}{\sigma}\left(\Delta_{\max}+2\sigma W_1\sqrt{\frac{\log(e+T_{n,J_n^\star})}{1+T_{n,J_n^\star}}}\right)\right)^2\right\} \\ &\geq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(1+\frac{\sqrt{\sqrt{L}}}{\sigma}\left(\Delta_{\max}+2\sigma W_1\sqrt{\frac{\log(e+\sqrt{L})}{1+\sqrt{L}}}\right)\right)^2\right\} \\ &\geq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(1+\frac{L^{1/4}\Delta_{\max}}{\sigma}+2W_1\sqrt{\log(e+\sqrt{L})}\right)^2\right\}.\end{aligned}$$

Now we have

$$a_{n,J_n^\star} \geq \max\left(\left(\frac{1}{2}\right)^K, \left(\frac{1}{2}\right)^{K-1}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(1+\frac{L^{1/4}\Delta_{\max}}{\sigma}+2W_1\sqrt{\log(e+\sqrt{L})}\right)^2\right\}\right),$$

and we have $\forall j \in \overline{V_n^L}$, $a_{n,j} \leq \exp\left\{-L^{3/4}\Delta_{\min}^2/(16\sigma^2)\right\}$, thus there exists $L_4 = \mathrm{Poly}(W_1)$ s.t. $\forall L > L_4$, $\forall j \in \overline{V_n^L}$,

$$a_{n,j} \leq \frac{a_{n,J_n^\star}}{2},$$

and by consequence, $J_n^{(2)} \in V_n^L$.

Finally, taking $L_1 = \max(L_2, L_3, L_4)$, we have $\forall L > L_1$, either $J_n^{(1)} \in V_n^L$ or $J_n^{(2)} \in V_n^L$. □

Next we show that there exists at least one arm in V_n^L for whom the probability of being pulled is large enough. More precisely, we prove the following lemma.

Lemma 15. *There exists $L_1 = \mathrm{Poly}(W_1)$ s.t. for $L > L_1$ and for all n s.t. $U_n^L \neq \emptyset$, then there exists $J_n \in V_n^L$ s.t.*

$$\psi_{n,J_n} \geq \frac{\min(\beta, 1-\beta)}{K^2} \triangleq \psi_{\min}.$$

Proof. Using Lemma 14, we know that $J_n^{(1)}$ or $J_n^{(2)} \in V_n^L$. On the other hand, we know that

$$\forall i \in \mathcal{A}, \psi_{n,i} = a_{n,i}\left(\beta + (1-\beta)\sum_{j\neq i}\frac{a_{n,j}}{1-a_{n,j}}\right).$$

Therefore we have

$$\psi_{n,J_n^{(1)}} \geq \beta a_{n,J_n^{(1)}} \geq \frac{\beta}{K},$$

since $\sum_{i\in\mathcal{A}} a_{n,i} = 1$, and

$$\begin{aligned}\psi_{n,J_n^{(2)}} &\geq (1-\beta)a_{n,J_n^{(2)}}\frac{a_{n,J_n^{(1)}}}{1-a_{n,J_n^{(1)}}}\\ &= (1-\beta)a_{n,J_n^{(1)}}\frac{a_{n,J_n^{(2)}}}{1-a_{n,J_n^{(1)}}}\\ &\geq \frac{1-\beta}{K^2},\end{aligned}$$

since $a_{n,J_n^{(1)}} \geq 1/K$ and $\sum_{i\neq J_n^{(1)}} a_{n,i}/(1-a_{n,J_n^{(1)}}) = 1$, thus $a_{n,J_n^{(2)}}/(1-a_{n,J_n^{(1)}}) \geq 1/K$. □

The rest of this subsection is quite similar to that of Qin, Klabjan and Russo, 2017. Indeed, with the above lemma, we can show that the set of poorly explored arms U_n^L is empty when n is large enough.

Lemma 16. *Under TTTS, there exists* $L_0 = Poly(W_1, W_2)$ *s.t.* $\forall L > L_0$, $U^L_{\lfloor KL\rfloor} = \varnothing$.

Proof. We proceed by contradiction, and we assume that $U^L_{\lfloor KL\rfloor}$ is not empty. Then for any $1 \leq \ell \leq \lfloor KL\rfloor$, U_ℓ^L and V_ℓ^L are non empty as well.

There exists a deterministic L_5 s.t. $\forall L > L_5$,

$$\lfloor L\rfloor \geq KL^{3/4}.$$

Using the pigeonhole principle, there exists some $i \in \mathcal{A}$ s.t. $T_{\lfloor L\rfloor,i} \geq L^{3/4}$. Thus, we have $|V^L_{\lfloor L\rfloor}| \leq K-1$.

Next, we prove $|V^L_{\lfloor 2L\rfloor}| \leq K-2$. Otherwise, since U_ℓ^L is non-empty for any $\lfloor L\rfloor + 1 \leq \ell \leq \lfloor 2L\rfloor$, thus by Lemma 15, there exists $J_\ell \in V_\ell^L$ s.t. $\psi_{\ell,J_\ell} \geq \psi_{\min}$. Therefore,

$$\sum_{i\in V_\ell^L} \psi_{\ell,i} \geq \psi_{\min},$$

and

$$\sum_{i\in V^L_{\lfloor L\rfloor}} \psi_{\ell,i} \geq \psi_{\min}$$

since $V_\ell^L \subset V^L_{\lfloor L\rfloor}$. Hence, we have

$$\sum_{i\in V^L_{\lfloor L\rfloor}} (\Psi_{\lfloor 2L\rfloor,i} - \Psi_{\lfloor L\rfloor,i}) = \sum_{\ell=\lfloor L\rfloor+1}^{\lfloor 2L\rfloor} \sum_{i\in V^L_{\lfloor L\rfloor}} \psi_{\ell,i} \geq \psi_{\min}\lfloor L\rfloor.$$

Then, using Lemma 8, there exists $L_6 = \mathrm{Poly}(W_2)$ s.t. $\forall L > L_6$, we have

$$\begin{aligned}\sum_{i \in V^L_{\lfloor L \rfloor}} (T_{\lfloor 2L \rfloor, i} - T_{\lfloor L \rfloor, i}) &\geq \sum_{i \in V^L_{\lfloor L \rfloor}} (\Psi_{\lfloor 2L \rfloor, i} - \Psi_{\lfloor L \rfloor, i} - 2W_2 \sqrt{\lfloor 2L \rfloor \log(e^2 + \lfloor 2L \rfloor)}) \\ &\geq \sum_{i \in V^L_{\lfloor L \rfloor}} (\Psi_{\lfloor 2L \rfloor, i} - \Psi_{\lfloor L \rfloor, i}) - 2KW_2 \sqrt{\lfloor 2L \rfloor \log(e^2 + \lfloor 2L \rfloor)} \\ &\geq \psi_{\min} \lfloor L \rfloor - 2KW_2 C_2 \lfloor L \rfloor^{3/4} \\ &\geq KL^{3/4},\end{aligned}$$

where C_2 is some absolute constant. Thus, we have one arm in $V^L_{\lfloor L \rfloor}$ that is pulled at least $L^{3/4}$ times between $\lfloor L \rfloor + 1$ and $\lfloor 2L \rfloor$, thus $|V^L_{\lfloor 2L \rfloor}| \leq K - 2$.

By induction, for any $1 \leq k \leq K$, we have $|V^L_{\lfloor kL \rfloor}| \leq K - k$, and finally if we take $L_0 = \max(L_1, L_5, L_6)$, then $\forall L > L_0$, $U^L_{\lfloor KL \rfloor} = \varnothing$. □

We can finally conclude the proof of Lemma 9 for TTTS.

Proof of Lemma 9 Let $N_1 = KL_0$ where $L_0 = \mathrm{Poly}(W_1, W_2)$ is chosen according to Lemma 16. For all $n > N_1$, we let $L = n/K$, then by Lemma 16, we have $U^L_{\lfloor KL \rfloor} = U^{n/K}_n$ is empty, which concludes the proof.

■

7.D.3 Concentration of the empirical means, proof of Lemma 10 under TTTS

As a corollary of the previous section, we can show the concentration of $\mu_{n,i}$ to μ_i for TTTS[5].

By Lemma 7, we know that $\forall i \in \mathcal{A}$ and $n \in \mathbb{N}$,

$$|\mu_{n,i} - \mu_i| \leq \sigma W_1 \sqrt{\frac{\log(e + T_{n,i})}{T_{n,i} + 1}}.$$

According to the previous section, there exists $N_1 = \mathrm{Poly}(W_1, W_2)$ s.t. $\forall n \geq N_1$ and $\forall i \in \mathcal{A}$, $T_{n,i} \geq \sqrt{n/K}$. Therefore,

$$|\mu_{n,i} - \mu_i| \leq \sqrt{\frac{\log(e + \sqrt{n/K})}{\sqrt{n/K} + 1}},$$

[5] this proof is the same as Proposition 3 of Qin, Klabjan and Russo, 2017

since $x \mapsto \log(e+x)/(x+1)$ is a decreasing function. There exists $N_2' = \mathrm{Poly}(\varepsilon, W_1)$ s.t. $\forall n \geq N_2'$,

$$\sqrt{\frac{\log(e+\sqrt{n/K})}{\sqrt{n/K}+1}} \leq \sqrt{\frac{2(n/K)^{1/4}}{\sqrt{n/K}+1}} \leq \frac{\varepsilon}{\sigma W_1}.$$

Therefore, $\forall n \geq N_2 \triangleq \max\{N_1, N_2'\}$, we have

$$|\mu_{n,i} - \mu_i| \leq \sigma W_1 \frac{\varepsilon}{\sigma W_1}.$$

7.D.4 Measurement effort concentration of the optimal arm, proof of Lemma 11 under TTTS

In this section we show that the empirical arm draws proportion of the true best arm for TTTS concentrates to β when the total number of arm draws is sufficiently large.

The proof is established upon the following lemmas. First, we prove that the empirical best arm coincides with the true best arm when the total number of arm draws goes sufficiently large.

Lemma 17. *Under TTTS, there exists $M_1 = Poly(W_1, W_2)$ s.t. $\forall n > M_1$, we have $I_n^\star = I^\star = J_n^{(1)}$ and $\forall i \neq I^\star$,*

$$a_{n,i} \leq \exp\left\{-\frac{\Delta_{min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\}.$$

Proof. Using Lemma 10 with $\varepsilon = \Delta_{\min}/4$, there exists $N_1' = \mathrm{Poly}(4/\Delta_{\min}, W_1, W_2)$ s.t. $\forall n > N_1'$,

$$\forall i \in \mathcal{A}, |\mu_{n,i} - \mu_i| \leq \frac{\Delta_{\min}}{4},$$

which implies that starting from a known moment, $\mu_{n,I^\star} > \mu_{n,i}$ for all $i \neq I^\star$, hence $I_n^\star = I^\star$. Thus, $\forall i \neq I^\star$,

$$\begin{aligned} a_{n,i} &= \Pi_n\left[\theta_i > \max_{j\neq i}\theta_j\right] \\ &\leq \Pi_n\left[\theta_i > \theta_{I^\star}\right] \\ &\leq \frac{1}{2}\exp\left\{-\frac{(\mu_{n,i}-\mu_{n,I^\star})^2}{2\sigma^2(1/T_{n,i}+1/T_{n,I^\star})}\right\}. \end{aligned}$$

The last inequality uses the Gaussian tail inequality of (7.8) Lemma 6. Furthermore,

$$\begin{aligned} (\mu_{n,i}-\mu_{n,I^\star})^2 &= (|\mu_{n,i}-\mu_{n,I^\star}|)^2 \\ &= (|\mu_{n,i}-\mu_i+\mu_i-\mu_{I^\star}+\mu_{I^\star}-\mu_{n,I^\star}|)^2 \\ &\geq (|\mu_i-\mu_{I^\star}|-|\mu_{n,i}-\mu_i+\mu_{I^\star}-\mu_{n,I^\star}|)^2 \\ &\geq \left(\Delta_{\min}-\left(\frac{\Delta_{\min}}{4}+\frac{\Delta_{\min}}{4}\right)\right)^2 = \frac{\Delta_{\min}^2}{4}, \end{aligned}$$

and according to Lemma 9, we know that there exists $M_2 = \mathrm{Poly}(W_1, W_2)$ s.t. $\forall n > M_2$,

$$\frac{1}{T_{n,i}} + \frac{1}{T_{n,I^\star}} \leq \frac{2}{\sqrt{n/K}}.$$

Thus, $\forall n > \max\{N_1', M_2\}$, we have

$$\forall i \neq I^\star, a_{n,i} \leq \exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\}.$$

Then, we have

$$a_{n,I^\star} = 1 - \sum_{i \neq I^\star} a_{n,i} \geq 1 - (K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\}.$$

There exists M_2' s.t. $\forall n > M_2'$, $a_{n,I^\star} > 1/2$, and by consequence $I^\star = J_n^{(1)}$. Finally taking $M_1 \triangleq \max\{N_1', M_2, M_2'\}$ concludes the proof. □

Before we prove Lemma 11, we first show that $\Psi_{n,I^\star}/n$ concentrates to β.

Lemma 18. *Under TTTS, fix a constant $\varepsilon > 0$, there exists $M_3 = \mathrm{Poly}(\varepsilon, W_1, W_2)$ s.t. $\forall n > M_3$, we have*

$$\left|\frac{\Psi_{n,I^\star}}{n} - \beta\right| \leq \varepsilon.$$

Proof. By Lemma 17, we know that there exists $M_1' = \mathrm{Poly}(W_1, W_2)$ s.t. $\forall n > M_1'$, we have $I_n^\star = I^\star = J_n^{(1)}$ and $\forall i \neq I^\star$,

$$a_{n,i} \leq \exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\}.$$

Note also that $\forall n \in \mathbb{N}$, we have

$$\psi_{n,I^\star} = a_{n,I^\star}\left(\beta + (1-\beta)\sum_{j \neq I^\star}\frac{a_{n,j}}{1-a_{n,j}}\right).$$

We proceed the proof with the following two steps.

Step 1 We first lower bound $\Psi_{n,I^\star}$ for a given ε. Take $M_4 > M_1'$ that we decide later, we have $\forall n > M_4$,

$$\begin{aligned}
\frac{\Psi_{n,I^\star}}{n} &= \frac{1}{n}\sum_{l=1}^{n}\psi_{l,I^\star} = \frac{1}{n}\sum_{l=I^\star}^{M_4}\psi_{l,I^\star} + \frac{1}{n}\sum_{l=M_4+1}^{n}\psi_{l,I^\star} \\
&\geq \frac{1}{n}\sum_{l=M_4+1}^{n}\psi_{l,I^\star} \geq \frac{1}{n}\sum_{l=M_4+1}^{n} a_{l,I^\star}\beta \\
&= \frac{\beta}{n}\sum_{l=M_4+1}^{n}\left(1 - \sum_{j\neq I^\star} a_{l,j}\right) \\
&\geq \frac{\beta}{n}\sum_{l=M_4+1}^{n}\left(1 - (K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}\right) \\
&= \beta - \frac{M_4}{n}\beta - \frac{\beta}{n}\sum_{l=M_4+1}^{n}(K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\} \\
&\geq \beta - \frac{M_4}{n}\beta - \frac{(n-M_4)}{n}\beta(K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{M_4}{K}}\right\} \\
&\geq \beta - \frac{M_4}{n}\beta - \beta(K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{M_4}{K}}\right\}.
\end{aligned}$$

For a given constant $\varepsilon > 0$, there exists M_5 s.t. $\forall n > M_5$,

$$\beta(K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\} < \frac{\varepsilon}{2}.$$

Furthermore, there exists $M_6 = \mathrm{Poly}(\varepsilon/2, M_5)$ s.t. $\forall n > M_6$,

$$\frac{M_5}{n}\beta < \frac{\varepsilon}{2}.$$

Therefore, if we take $M_4 \triangleq \max\{M_1', M_5, M_6\}$, we have $\forall n > M_4$,

$$\frac{\Psi_{n,I^\star}}{n} \geq \beta - \varepsilon.$$

Step 2 On the other hand, we can also upper bound $\Psi_{n,I^\star}$. We have $\forall n > M_3$,

$$\begin{aligned}
\frac{\Psi_{n,I^\star}}{n} &= \frac{1}{n}\sum_{l=1}^{n}\psi_{l,I^\star} \\
&= \frac{1}{n}\sum_{l=1}^{n} a_{l,I^\star}\left(\beta + (1-\beta)\sum_{j\neq I^\star}\frac{a_{l,j}}{1-a_{l,j}}\right) \\
&\leq \frac{1}{n}\sum_{l=1}^{n} a_{l,I^\star}\beta + \frac{1}{n}\sum_{l=1}^{n} a_{l,I^\star}(1-\beta)\sum_{j\neq I^\star}\frac{a_{l,j}}{1-a_{l,j}} \\
&\leq \beta + \frac{1}{n}\sum_{l=1}^{n}(1-\beta)\sum_{j\neq I^\star}\frac{a_{l,j}}{1-a_{l,j}} \\
&\leq \beta + \frac{1}{n}\sum_{l=1}^{n}(1-\beta)\sum_{j\neq I^\star}\frac{\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}{1-\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}.
\end{aligned}$$

Since, for a given $\varepsilon > 0$, there exists M_8 s.t. $\forall n > M_8$,

$$\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\} < \frac{1}{2},$$

and there exists M_9 s.t. $\forall n > M_9$,

$$(1-\beta)(K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\} < \frac{\varepsilon}{4}.$$

Thus, $\forall n > M_{10} \triangleq \max\{M_8, M_9\}$,

$$\begin{aligned}
\frac{\Psi_{n,I^\star}}{n} &\leq \beta + \frac{1-\beta}{n}\left(\sum_{l=1}^{M_{10}}\sum_{j\neq I^\star}\frac{\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}{1-\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}} + \sum_{l=M_{10}+1}^{n}\sum_{j\neq I^\star}\frac{\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}{1-\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}\right) \\
&\leq \beta + \frac{1-\beta}{n}\sum_{l=1}^{M_{10}}\sum_{j\neq I^\star}\frac{\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}{1-\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}} + 2(1-\beta)(K-1)\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{M_{10}}{K}}\right\} \\
&\leq \beta + \frac{1-\beta}{n}\sum_{l=1}^{M_{10}}\sum_{j\neq I^\star}\frac{\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}{1-\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}} + \frac{\varepsilon}{2}.
\end{aligned}$$

There exists $M_{11} = \mathrm{Poly}(\varepsilon/2, M_{10})$ s.t. $\forall n > M_{11}$,

$$\frac{1-\beta}{n}\sum_{l=1}^{M_{10}}\sum_{j\neq I^\star}\frac{\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}}{1-\exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{l}{K}}\right\}} < \frac{\varepsilon}{2}.$$

Therefore, $\forall n > M_7 \triangleq \max\{M_3, M_{11}\}$, we have

$$\frac{\Psi_{n,I^\star}}{n} \leq \beta + \varepsilon.$$

Conclusion Finally, combining the two steps and define $M_3 \triangleq \max\{M_4, M_7\}$, we have $\forall n > M_3$,

$$\left|\frac{\Psi_{n,I^\star}}{n} - \beta\right| \leq \varepsilon.$$

□

With the help of the previous lemma and Lemma 8, we can finally prove Lemma 11.

Proof of Lemma 11 Fix an $\varepsilon > 0$. Using Lemma 8, we have $\forall n \in \mathbb{N}$,

$$\left|\frac{T_{n,I^\star}}{n} - \frac{\Psi_{n,I^\star}}{n}\right| \leq \frac{W_2\sqrt{(n+1)\log(e^2+n)}}{n}.$$

Thus there exists M_{12} s.t. $\forall n > M_{12}$,

$$\left|\frac{T_{n,I^\star}}{n} - \frac{\Psi_{n,I^\star}}{n}\right| \leq \frac{\varepsilon}{2}.$$

And using Lemma 18, there exists $M_3' = \mathrm{Poly}(\varepsilon/2, W_1, W_2)$ s.t. $\forall n > M_3'$,

$$\left|\frac{\Psi_{n,I^\star}}{n} - \beta\right| \leq \frac{\varepsilon}{2}.$$

Again, according to Lemma 15, there exists M_3' s.t. $\forall n > M_3'$,

$$\frac{\Psi_{n,I^\star}}{n} \leq \beta + \frac{\varepsilon}{2}.$$

Thus, if we take $N_3 \triangleq \max\{M_3', M_{12}\}$, then $\forall n > N_3$, we have

$$\left|\frac{T_{n,I^\star}}{n} - \beta\right| \leq \varepsilon.$$

■

7.D.5 Measurement effort concentration of other arms, proof of Lemma 12 under TTTS

In this section, we show that, for TTTS, the empirical measurement effort concentration also holds for other arms than the true best arm. We first show that if some arm is overly sampled at time n, then its probability of being picked is reduced exponentially.

Lemma 19. *Under TTTS, for every $\xi \in (0,1)$, there exists $S_1 = Poly(1/\xi, W_1, W_2)$ such that for all $n > S_1$, for all $i \neq I^\star$,*

$$\frac{\Psi_{n,i}}{n} \geq \omega_i^\beta + \xi \;\;\Rightarrow\;\; \psi_{n,i} \leq \exp\left\{-\varepsilon_0(\xi)n\right\},$$

where ε_0 is defined in (7.12) below.

Proof. First, by Lemma 17, there exists $M_1'' = \mathrm{Poly}(W_1, W_2)$ s.t. $\forall n > M_1''$,

$$I^\star = I_n^\star = J_n^{(1)}.$$

Then, following the similar argument as in Lemma 31, one can show that for all $i \neq I^\star$ and for all $n > M_1''$,

$$\begin{aligned}
\psi_{n,i} &= a_{n,i}\left(\beta + (1-\beta)\sum_{j\neq i}\frac{a_{n,j}}{1-a_{n,j}}\right)\\
&\leq a_{n,i}\beta + a_{n,i}(1-\beta)\frac{\sum_{j\neq i} a_{n,j}}{1-a_{n,J_n^{(1)}}}\\
&= a_{n,i}\beta + a_{n,i}(1-\beta)\frac{\sum_{j\neq i} a_{n,j}}{1-a_{n,I^\star}}\\
&\leq a_{n,i}\beta + a_{n,i}(1-\beta)\frac{1}{1-a_{n,I^\star}}\\
&\leq \frac{a_{n,i}}{1-a_{n,I^\star}}\\
&\leq \frac{\Pi_n\left[\theta_i \geq \theta_{I^\star}\right]}{\Pi_n\left[\cup_{j\neq I^\star}\theta_j \geq \theta_{I^\star}\right]}\\
&\leq \frac{\Pi_n\left[\theta_i \geq \theta_{I^\star}\right]}{\max_{j\neq I^\star}\Pi_n\left[\theta_j \geq \theta_{I^\star}\right]}.
\end{aligned}$$

Using the upper and lower Gaussian tail bounds from Lemma 6, we have

$$\begin{aligned}
\psi_{n,i} &\leq \frac{\exp\left\{-\dfrac{(\mu_{n,I^\star}-\mu_{n,i})^2}{2\sigma^2\left(1/T_{n,I^\star}+1/T_{n,i}\right)}\right\}}{\exp\left\{-\min\limits_{j\neq I^\star}\dfrac{1}{2}\left(\dfrac{(\mu_{n,I^\star}-\mu_{n,j})}{\sigma\sqrt{\left(1/T_{n,I^\star}+1/T_{n,j}\right)}}-1\right)^2\right\}}\\
&= \frac{\exp\left\{-n\dfrac{(\mu_{n,I^\star}-\mu_{n,i})^2}{2\sigma^2\left(n/T_{n,I^\star}+n/T_{n,i}\right)}\right\}}{\exp\left\{-n\left(\min\limits_{j\neq I^\star}\dfrac{(\mu_{n,I^\star}-\mu_{n,j})}{\sqrt{2\sigma^2\left(n/T_{n,I^\star}+n/T_{n,j}\right)}}-\dfrac{1}{\sqrt{2n}}\right)^2\right\}},
\end{aligned}$$

where we assume that $n > S_2 = \text{Poly}(W_1, W_2)$ for which

$$\frac{(\mu_{n,I^\star} - \mu_{n,i})^2}{\sigma^2 \left(1/T_{n,I^\star} + 1/T_{n,i}\right)} \geq 1$$

according to Lemma 9. From there we take a supremum over the possible allocations to lower bound the denominator and write

$$\begin{aligned}
\psi_{n,i} &\leq \frac{\exp\left\{-n \frac{(\mu_{n,I^\star} - \mu_{n,i})^2}{2\sigma^2 \left(n/T_{n,I^\star} + n/T_{n,i}\right)}\right\}}{\exp\left\{-n \left(\sup_{\boldsymbol{\omega}: \omega_{I^\star} = T_{n,I^\star}/n} \min_{j \neq I^\star} \frac{(\mu_{n,I^\star} - \mu_{n,i})}{\sqrt{2\sigma^2 \left(1/\omega_{I^\star} + 1/\omega_j\right)}} - \frac{1}{\sqrt{2n}}\right)^2\right\}} \\
&= \frac{\exp\left\{-n \frac{(\mu_{n,I^\star} - \mu_{n,i})^2}{2\sigma^2 \left(n/T_{n,I^\star} + n/T_{n,i}\right)}\right\}}{\exp\left\{-n \left(\sqrt{\Gamma^\star_{T_{n,I^\star}/n}(\boldsymbol{\mu}_n)} - \frac{1}{\sqrt{2n}}\right)^2\right\}},
\end{aligned}$$

where $\boldsymbol{\mu}_n \triangleq (\mu_{n,1}, \cdots, \mu_{n,K})$, and $(\beta, \boldsymbol{\mu}) \mapsto \Gamma^\star_\beta(\boldsymbol{\mu})$ represents a function that maps β and $\boldsymbol{\mu}$ to the parameterized optimal error decay that any allocation rule can reach given parameter β and a set of arms with means $\boldsymbol{\mu}$. Note that this function is continuous with respect to β and $\boldsymbol{\mu}$ respectively.

Now, assuming $\Psi_{n,i}/n \geq \omega_i^\beta + \xi$ yields that there exists $S_2' \triangleq \text{Poly}(2/\xi, W_2)$ s.t. for all $n > S_2'$, $T_{n,i}/n \geq \omega_i^\beta + \xi/2$, and by consequence,

$$\psi_{n,i} \leq \exp\left\{ \underbrace{-n \left(\frac{(\mu_{n,I^\star} - \mu_{n,i})^2}{2\sigma^2 \left(n/T_{n,I^\star} + 1/(\omega_i^\beta + \xi/2)\right)} - \Gamma^\star_{T_{n,I^\star}/n}(\boldsymbol{\mu}_n) - \frac{1}{2n} + \sqrt{\frac{2\Gamma^\star_{T_{n,I^\star}/n}(\boldsymbol{\mu}_n)}{n}} \right)}_{\varepsilon_n(\xi)} \right\}.$$

Using Lemma 11, we know that for any ε, there exists $S_3 = \text{Poly}(1/\varepsilon, W_1, W_2)$ s.t. $\forall n > S_3$, $|T_{n,I^\star}/n - \beta| \leq \varepsilon$, and $\forall j \in \mathcal{A}, |\mu_{n,j} - \mu_j| \leq \varepsilon$. Furthermore, $(\beta, \boldsymbol{\mu}) \mapsto \Gamma^\star_\beta(\boldsymbol{\mu})$ is continuous with respect to β and $\boldsymbol{\mu}$, thus for a given ε_0, there exists $S_3' = \text{Poly}(1/\varepsilon_0, W_1, W_2)$ s.t. $\forall n > S_3'$, we have

$$\left| \varepsilon_n(\xi) - \left(\frac{(\mu_{I^\star} - \mu_i)^2}{2\sigma^2 \left(1/\beta + 1/(\omega_i^\beta + \xi/2)\right)} - \Gamma^\star_\beta \right) \right| \leq \varepsilon_0.$$

Finally, define $S_1 \triangleq \max\{S_2, S_2', S_3'\}$, we have $\forall n > S_1$,

$$\psi_{n,i} \leq \exp\left\{-\varepsilon_0(\xi) n\right\},$$

where

$$\varepsilon_0(\xi) = \frac{(\mu_{I^\star} - \mu_i)^2}{2\sigma^2\left(1/\beta + 1/(\omega_i^\beta + \xi/2)\right)} - \Gamma_\beta^\star + \varepsilon_0\,. \tag{7.12}$$

□

Next, starting from some known moment, no arm is overly allocated. More precisely, we show the following lemma.

Lemma 20. *Under TTTS, for every ξ, there exists $S_4 = Poly(1/\xi, W_1, W_2)$ s.t. $\forall n > S_4$,*

$$\forall i \in \mathcal{A},\quad \frac{\Psi_{n,i}}{n} \le \omega_i^\beta + \xi.$$

Proof. From Lemma 19, there exists $S_1' = \mathrm{Poly}(2/\xi, W_1, W_2)$ such that for all $n > S_1'$ and for all $i \neq I^\star$,

$$\frac{\Psi_{n,i}}{n} \ge \omega_i^\beta + \frac{\xi}{2} \quad\Rightarrow\quad \psi_{n,i} \le \exp\left\{-\varepsilon_0(\xi/2)n\right\}.$$

Thus, for all $i \neq I^\star$,

$$\begin{aligned}\frac{\Psi_{n,i}}{n} &\le \frac{S_1'}{n} + \frac{\sum_{\ell=S_1'+1}^{n} \psi_{\ell,i}\mathbb{1}\left(\frac{\Psi_{\ell,i}}{n} \ge \omega_i^\beta + \frac{\xi}{2}\right)}{n} + \frac{\sum_{\ell=S_1'+1}^{n} \psi_{\ell,i}\mathbb{1}\left(\frac{\Psi_{\ell,i}}{n} \le \omega_i^\beta + \frac{\xi}{2}\right)}{n} \\ &\le \frac{S_1'}{n} + \frac{\sum_{\ell=1}^{n} \exp\left\{-\varepsilon_0(\xi/2)n\right\}}{n} + \frac{\sum_{\ell=S_1'+1}^{\ell_n(\xi)} \psi_{\ell,i}\mathbb{1}\left(\frac{\Psi_{\ell,i}}{n} \le \omega_i^\beta + \frac{\xi}{2}\right)}{n},\end{aligned}$$

where we let $\ell_n(\xi) = \max\left\{\ell \le n : \Psi_{\ell,i}/n \le \omega_i^\beta + \xi/2\right\}$. Then

$$\begin{aligned}\frac{\Psi_{n,i}}{n} &\le \frac{S_1'}{n} + \frac{\sum_{\ell=1}^{n} \exp\left\{-\varepsilon_0(\xi/2)n\right\}}{n} + \Psi_{\ell_n(\xi),i} \\ &\le \frac{S_1' + (1 - \exp(-\varepsilon_0(\xi/2))^{-1}}{n} + \omega_i^\beta + \frac{\xi}{2}\end{aligned}$$

Then, there exists S_5 such that for all $n \ge S_5$,

$$\frac{S_1' + (1 - \exp(-\varepsilon_0(\xi/2))^{-1}}{n} \le \frac{\xi}{2}.$$

Therefore, for any $n > S_4 \triangleq \max\{S_1', S_5\}$, $\Psi_{n,i} \le \omega_i^\beta + \xi$ holds for all $i \neq I^\star$. For $i = I^\star$, it is already proved for the optimal arm. □

We now prove Lemma 12 under TTTS.

Proof of Lemma 12 From Lemma 20, there exists $S_4' = \mathrm{Poly}((K-1)/\xi, W_1, W_2)$ such that for all $n > S_4'$,

$$\forall i \in \mathcal{A}, \frac{\Psi_{n,i}}{n} \leq \omega_i^\beta + \frac{\xi}{K-1}.$$

Using the fact that $\Psi_{n,i}/n$ and ω_i^β all sum to 1, we have $\forall i \in \mathcal{A}$,

$$\begin{aligned}\frac{\Psi_{n,i}}{n} &= 1 - \sum_{j \neq i} \frac{\Psi_{n,j}}{n} \\ &\geq 1 - \sum_{j \neq i} \left(\omega_j^\beta + \frac{\xi}{K-1} \right) \\ &= \omega_i^\beta - \xi.\end{aligned}$$

Thus, for all $n > S_4'$, we have

$$\forall i \in \mathcal{A}, \left| \frac{\Psi_{n,i}}{n} - \omega_i^\beta \right| \leq \xi.$$

And finally we use the same reasoning as the proof of Lemma 11 to link $T_{n,i}$ and $\Psi_{n,i}$. Fix an $\varepsilon > 0$. Using Lemma 8, we have $\forall n \in \mathbb{N}$,

$$\forall i \in \mathcal{A}, \left| \frac{T_{n,i}}{n} - \frac{\Psi_{n,i}}{n} \right| \leq \frac{W_2 \sqrt{(n+1)\log(e^2+n)}}{n}.$$

Thus there exists S_5 s.t. $\forall n > S_5$,

$$\left| \frac{T_{n,I^\star}}{n} - \frac{\Psi_{n,I^\star}}{n} \right| \leq \frac{\varepsilon}{2}.$$

And using the above result, there exists $S_4'' = \mathrm{Poly}(2/\varepsilon, W_1, W_2)$ s.t. $\forall n > S_4''$,

$$\left| \frac{\Psi_{n,i}}{n} - \omega_i^\beta \right| \leq \frac{\varepsilon}{2}.$$

Thus, if we take $N_4 \triangleq \max\{S_4'', S_5\}$, then $\forall n > N_4$, we have

$$\forall i \in \mathcal{A}, \left| \frac{T_{n,i}}{n} - \omega_i^\beta \right| \leq \varepsilon.$$

■

7.E Fixed-Confidence Analysis for T3C

This section is entirely dedicated to T3C. Note that the analysis to follow share the same proof line with that of TTTS, and some parts even completely coincide with those of TTTS. For the sake of clarity and simplicity, we shall only focus on the parts that differ and skip some redundant proofs.

7.E.1 Sufficient exploration of all arms, proof of Lemma 9 under T3C

To prove this lemma, we still need the two sets of indices for under-sampled arms like in Appendix 7.D.2. We recall that for a given $L > 0$: $\forall n \in \mathbb{N}$ we define

$$U_n^L \triangleq \{i : T_{n,i} < \sqrt{L}\},$$

$$V_n^L \triangleq \{i : T_{n,i} < L^{3/4}\}.$$

For T3C however, we investigate the following two indices,

$$J_n^{(1)} \triangleq \arg\max_j a_{n,j}; \qquad \widetilde{J_n^{(2)}} \triangleq \arg\min_{j \neq J_n^{(1)}} W_n(J_n^{(1)}, j).$$

Lemma 9 is proved via the following sequence of lemmas.

Lemma 21. *There exists $L_1 = Poly(W_1)$ s.t. if $L > L_1$, for all n, $U_n^L \neq \varnothing$ implies $J_n^{(1)} \in V_n^L$ or $\widetilde{J_n^{(2)}} \in V_n^L$.*

Proof. If $J_n^{(1)} \in V_n^L$, then the proof is finished. Now we assume that $J_n^{(1)} \in \overline{V_n^L} \subset \overline{U_n^L}$, and we prove that $J_n^{(2)} \in V_n^L$.

Step 1 Following the same reasoning as Step 1 and Step 2 of the proof of Lemma 14, we know that there exists $L_2 = \mathrm{Poly}(W_1)$ s.t. if $L > L_2$, then

$$\overline{J_n^\star} \triangleq \arg\max_{j \in \overline{U_n^L}} \mu_{n,j} = \arg\max_{j \in \overline{U_n^L}} \mu_j = J_n^{(1)}.$$

Step 2 Now assuming that $L > L_2$, and we show that for L large enough, $\widetilde{J_n^{(2)}} \in V_n^L$. In the same way that we proved (7.11) one can show that for all $\forall j \in \overline{V_n^L}$,

$$W_n(J_n^{(1)}, j) = \frac{(\mu_{n,I^\star} - \mu_{n,j})^2}{2\sigma^2\left(\frac{1}{T_{n,I^\star}} + \frac{1}{T_{n,j}}\right)} \geq \frac{L^{3/4}\Delta_{\min}^2}{16\sigma^2}.$$

Again, denote $J_n^\star \triangleq \arg\max_{j \in U_n^L} \mu_{n,j}$, we obtain

$$W_n(J_n^{(1)}, J_n^\star) = \begin{cases} 0 & \text{if } \mu_{n,J_n^\star} \geq \mu_{n,J_n^{(1)}}, \\ \dfrac{(\mu_{n,J_n^{(1)}} - \mu_{n,J_n^\star})^2}{2\sigma^2\left(\frac{1}{T_{n,J_n^{(1)}}} + \frac{1}{T_{n,J_n^\star}}\right)} & \text{else.} \end{cases}$$

In the second case, as already shown in Step 3 of Lemma 14 we have that

$$|\mu_{n,J_n^\star} - \mu_{n,\overline{J_n^\star}}| \le \Delta_{\max} + 2\sigma W_1 \sqrt{\frac{\log(e + T_{n,J_n^\star})}{1 + T_{n,J_n^\star}}}$$
$$\le \Delta_{\max} + 2\sigma W_1 \sqrt{\frac{\log(e + \sqrt{L})}{1 + \sqrt{L}}},$$

since $J_n^\star \in U_n^L$. We also know that

$$2\sigma^2 \left(\frac{1}{T_{n,J_n^{(1)}}} + \frac{1}{T_{n,J_n^\star}} \right) \ge \frac{2\sigma^2}{T_{n,J_n^\star}} \ge \frac{2\sigma^2}{\sqrt{L}}.$$

Therefore, we get

$$W_n(J_n^{(1)}, J_n^\star) \le \frac{\sqrt{L}}{2\sigma^2} \left(\Delta_{\max} + 2\sigma W_1 \sqrt{\frac{\log(e + \sqrt{L})}{1 + \sqrt{L}}} \right)^2.$$

On the other hand, we know that for all $j \in \overline{V_n^L}$,

$$W_n(J_n^{(1)}, j) \ge \frac{L^{3/4} \Delta_{\min}^2}{16\sigma^2}.$$

Thus, there exists L_3 s.t. if $L > L_3$, then

$$\forall j \in \overline{V_n^L},\ W_n(J_n^{(1)}, j) \ge 2 W_n(J_n^{(1)}, J_n^\star).$$

That means $\widetilde{J_n^{(2)}} \notin \overline{V_n^L}$ and by consequence, $\widetilde{J_n^{(2)}} \in V_n^L$.

Finally, taking $L_1 = \max(L_2, L_3)$, we have $\forall L > L_1$, either $J_n^{(1)} \in V_n^L$ or $\widetilde{J_n^{(2)}} \in V_n^L$. □

Next we show that there exists at least one arm in V_n^L for whom the probability of being pulled is large enough. More precisely, we prove the following lemma.

Lemma 22. *There exists $L_1 = Poly(W_1)$ s.t. for $L > L_1$ and for all n s.t. $U_n^L \neq \varnothing$, then there exists $J_n \in V_n^L$ s.t.*

$$\psi_{n,J_n} \ge \frac{\min(\beta, 1-\beta)}{K^2} \triangleq \psi_{\min}.$$

Proof. Using Lemma 21, we know that $J_n^{(1)}$ or $\widetilde{J_n^{(2)}} \in V_n^L$. We also know that under T3C, for any arm i, $\psi_{n,i}$ can be written as

$$\psi_{n,i} = \beta a_{n,i} + (1-\beta) \sum_{j \neq i} a_{n,j} \frac{\mathbb{1}\{W_n(j,i) = \min_{k \neq j} W_n(j,k)\}}{\left| \arg\min_{k \neq j} W_n(j,k) \right|}.$$

Note that $(\psi_{n,i})_i$ sums to 1,

$$\begin{aligned}\sum_i \psi_{n,i} &= \beta + (1-\beta)\sum_j a_{n,j} \sum_{i\neq j} \frac{\mathbb{1}\{W_n(j,i) = \min_{k\neq j} W_n(j,k)\}}{\left|\arg\min_{k\neq j} W_n(j,k)\right|} \\ &= \beta + (1-\beta)\sum_j a_{n,j} = 1\,.\end{aligned}$$

Therefore, we have

$$\psi_{n,J_n^{(1)}} \geq \beta a_{n,J_n^{(1)}} \geq \frac{\beta}{K}$$

on one hand, since $\sum_{i\in\mathcal{A}} a_{n,i} = 1$. On the other hand, we have

$$\begin{aligned}\psi_{n,\widehat{J_n^{(2)}}} &\geq (1-\beta)\frac{a_{n,J_n^{(1)}}}{K} \\ &\geq \frac{1-\beta}{K^2},\end{aligned}$$

which concludes the proof. □

The rest of this subsection is exactly the same to that of TTTS. Indeed, with the above lemma, we can show that the set of poorly explored arms U_n^L is empty when n is large enough.

Lemma 23. *Under T3C, there exists* $L_0 = Poly(W_1, W_2)$ *s.t.* $\forall L > L_0$, $U^L_{\lfloor KL \rfloor} = \varnothing$.

Proof. See proof of Lemma 16 in Appendix 7.D.2. □

We can finally conclude the proof of Lemma 9 for T3C in the same way as for TTTS in Appendix 7.D.2. ■

7.E.2 Concentration of the empirical means, proof of Lemma 10 under T3C

As a corollary of the previous section, we can show the concentration of $\mu_{n,i}$ to μ_i, and the proof remains the same as that of TTTS in Appendix 7.D.3.

7.E.3 Measurement effort concentration of the optimal arm, proof of Lemma 11 under T3C

Next, we show that the empirical arm draws proportion of the true best arm for T3C concentrates to β when the total number of arm draws is sufficiently large. This proof also remains the same as that of TTTS in Appendix 7.D.4.

7.E.4 Measurement effort concentration of other arms, proof of Lemma 12 under T3C

In this section, we show that, for T3C, the empirical measurement effort concentration also holds for other arms than the true best arm. Note that this part differs from that of TTTS.

We again establish first an over-allocation implies negligible probability result as follow.

Lemma 24. *Under T3C, for every $\xi \le \varepsilon_0$ with ε_0 problem dependent, there exists $S_1 = Poly(1/\xi, W_1, W_2)$ such that for all $n > S_1$, for all $i \neq I^\star$,*

$$\frac{\Psi_{n,i}}{n} \ge \omega_i^\beta + 2\xi \;\Rightarrow\; \psi_{n,i} \le (K-1)\exp\left\{-\frac{\Delta_{min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\}.$$

Proof. Fix $i \neq I^\star$ s.t. $\Psi_{n,i}/n \ge \omega_i^\beta + 2\xi$, then using Lemma 8, there exists $S_2 = \mathrm{Poly}(1/\xi, W_2)$ such that for any $n > S_2$, we have

$$\frac{T_{n,i}}{n} \ge \omega_i^\beta + \xi.$$

Then,

$$\begin{aligned}
\psi_{n,i} &\le \beta a_{n,i} + (1-\beta)\sum_{j\neq i} a_{n,j}\mathbb{1}\{W_n(j,i) = \min_{k\neq j} W_n(j,k)\} \\
&\le \beta a_{n,i} + (1-\beta)\left(\sum_{j\neq i, I^\star} a_{n,j} + a_{n,I^\star}\mathbb{1}\{W_n(I^\star,i) = \min_{k\neq I^\star} W_n(I^\star,k)\}\right) \\
&\le \sum_{j\neq I^\star} a_{n,j} + \mathbb{1}\{W_n(I^\star,i) = \min_{k\neq I^\star} W_n(I^\star,k)\}.
\end{aligned}$$

Next we show that the indicator function term in the previous inequality equals 0.

Using Lemma 7 and Lemma 11 for T3C, there exists $S_3 = \mathrm{Poly}(1/\xi, W_1, W_2)$ such that for any $n > S_3$,

$$\left|\frac{T_{n,I^\star}}{n} - \beta\right| \le \xi^2 \text{ and } \forall j \in \mathcal{A}, |\mu_{n,j} - \mu_j| \le \xi^2.$$

Now if $\forall j \neq I^\star, i$, we have $T_{n,j}/n > \omega_j^\beta$, then

$$\begin{aligned}
\frac{n-1}{n} &= \sum_{j\in\mathcal{A}} \frac{T_{n,j}}{n} \\
&= \frac{T_{n,I^\star}}{n} + \frac{T_{n,i}}{n} + \sum_{j\neq I^\star, i} \frac{T_{n,j}}{n} \\
&> \beta - \varepsilon^2 + \omega_i^\beta + \varepsilon + \sum_{j\neq I^\star, i} \omega_j^\beta \ge 1,
\end{aligned}$$

which is a contradiction.

Thus there exists at least one $j_0 \neq I^\star, i$, such that $T_{n,j_0}/n \leq \omega_j^\beta$. Assuming $n > \max(S_2, S_3)$, we have

$$W_n(I^\star, i) - W_n(I^\star, j_0) = \frac{(\mu_{n,I^\star} - \mu_{n,i})^2}{2\sigma^2\left(\frac{1}{T_{n,I^\star}} + \frac{1}{T_{n,i}}\right)} - \frac{(\mu_{n,I^\star} - \mu_{n,j_0})^2}{2\sigma^2\left(\frac{1}{T_{n,I^\star}} + \frac{1}{T_{n,j_0}}\right)}$$
$$\geq \underbrace{\frac{(\mu_{I^\star} - \mu_i - 2\xi^2)^2}{2\sigma^2\left(\frac{1}{\beta - \xi^2} + \frac{1}{\omega_i^\beta + \xi}\right)} - \frac{(\mu_{I^\star} - \mu_{j_0} + 2\xi^2)^2}{2\sigma^2\left(\frac{1}{\beta + \xi^2} + \frac{1}{\omega_{j_0}^\beta}\right)}}_{W_{i,j_0}^\xi}.$$

According to Proposition 1, W_{i,j_0}^ξ converges to 0 when ξ goes to 0, more precisely we have

$$W_{i,j_0}^\xi = \frac{(\mu_{I^\star} - \mu_i)^2}{2\sigma^2}\left(\frac{\beta}{\beta + \omega_i^\beta}\right)^2 \xi + O(\xi^2),$$

thus there exists a ε_0 such that for all $\xi < \varepsilon_0$ it holds for all $i, j_0 \neq I^\star$, $W_{i,j_0}^\xi > 0$. It follows then

$$W_n(I^\star, i) - \min_{k \neq I^\star} W_n(I^\star, k) \geq W_n(I^\star, i) - W_n(I^\star, j_0) > 0,$$

and $\mathbb{1}\{W_n(I^\star, i) = \min_{k \neq I^\star} W_n(I^\star, k)\} = 0$.

Knowing that Lemma 17 is also valid for T3C, thus there exists $M_1 = \mathrm{Poly}(4/\Delta_{\min}, W_1, W_2)$ such that for all $n > M_1$,

$$\forall j \neq I^\star, a_{n,j} \leq \exp\left\{-\frac{\Delta_{\min}^2}{16\sigma^2}\sqrt{\frac{n}{K}}\right\},$$

which then concludes the proof by taking $S_1 \triangleq \max(M_1, S_2, S_3)$. □

The rest of this subsection almost coincides with that of TTTS. We first show that, starting from some known moment, no arm is overly allocated. More precisely, we show the following lemma.

Lemma 25. *Under T3C, for every ξ, there exists $S_4 = \mathrm{Poly}(1/\xi, W_1, W_2)$ s.t. $\forall n > S_4$,*

$$\forall i \in \mathcal{A}, \quad \frac{\Psi_{n,i}}{n} \leq \omega_i^\beta + 2\xi.$$

Proof. See proof of Lemma 20 in Appendix 7.D.5. Note that the previous step does not match exactly that of TTTS, so the proof would be slightly different. However, the difference is only a matter of constant, we thus still choose to skip this proof. □

It remains to prove Lemma 12 for T3C, which stays the same as that of TTTS.

Proof of Lemma 12 for T3C See proof of Lemma 12 for TTTS in Appendix 7.D.5. ■

7.F Proof of Lemma 4

Finally, it remains to prove Lemma 4 under the Gaussian case before we can conclude for Theorem 7.2 for TTTS or T3C.

Lemma 4. *Let $\delta, \beta \in (0,1)$. For any sampling rule which satisfies $\mathbb{E}\left[T_\beta^\varepsilon\right] < \infty$ for all $\varepsilon > 0$, we have*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}\left[\tau_\delta\right]}{\log(1/\delta)} \leq \frac{1}{\Gamma_\beta^\star},$$

if the sampling rule is coupled with stopping rule (7.4),

For the clarity, we recall the definition of generalized likelihood ratio. For any pair of arms i, j, We first define a weighted average of their empirical means,

$$\widehat{\mu}_{n,i,j} \triangleq \frac{T_{n,i}}{T_{n,i} + T_{n,j}} \widehat{\mu}_{n,i} + \frac{T_{n,j}}{T_{n,i} + T_{n,j}} \widehat{\mu}_{n,j}.$$

And if $\widehat{\mu}_{n,i} \geq \widehat{\mu}_{n,j}$, then the generalized likelihood ratio $Z_{n,i,j}$ for Gaussian noise distributions has the following analytic expression,

$$Z_{n,i,j} \triangleq T_{n,i} d(\widehat{\mu}_{n,i}; \widehat{\mu}_{n,i,j}) + T_{n,j} d(\widehat{\mu}_{n,j}; \widehat{\mu}_{n,i,j}).$$

We further define a statistic Z_n as

$$Z_n \triangleq \max_{i \in \mathcal{A}} \min_{j \in \mathcal{A} \setminus \{i\}} Z_{n,i,j}.$$

The following lemma stated by Qin, Klabjan and Russo (2017) is needed in our proof.

Lemma 26. *For any $\zeta > 0$, there exists ε s.t. $\forall n \geq T_\beta^\varepsilon$, $Z_n \geq (\Gamma_\beta^\star - \zeta) n$.*

To prove Lemma 4, we need the Gaussian tail inequality (7.8) of Lemma 6.

Proof. We know that

$$\begin{aligned}
1 - a_{n,I^\star} &= \sum_{i \neq I^\star} a_{n,i} \\
&\leq \sum_{i \neq I^\star} \Pi_n \left[\theta_i > \theta_{I^\star}\right] \\
&= \sum_{i \neq I^\star} \Pi_n \left[\theta_i - \theta_{I^\star} > 0\right] \\
&\leq (K-1) \max_{i \neq I^\star} \Pi_n \left[\theta_i - \theta_{I^\star} > 0\right].
\end{aligned}$$

We can further rewrite $\Pi_n\left[\theta_i - \theta_{I^\star} > 0\right]$ as

$$\Pi_n\left[\theta_i - \theta_{I^\star} > \mu_{n,i} - \mu_{n,I^\star} + \mu_{n,I^\star} - \mu_{n,i}\right].$$

We choose ε sufficiently small such that the empirical best arm $I_n^\star = I^\star$. Then, for all $n \geq T_\beta^n$ and for any $i \neq I^\star$, $\mu_{n,I^\star} \geq \mu_{n,i}$. Thus, fix any $\zeta \in (0, \Gamma_\beta^\star/2)$ and apply inequality (7.8) of Lemma 6 with $\mu_{n,I^\star}$ and $\mu_{n,i}$, we have for any $n \geq T_\beta^\varepsilon$,

$$\begin{aligned} 1 - a_{n,I^\star} &\leq (K-1) \max_{i \neq I^\star} \frac{1}{2} \exp\left\{-\frac{(\mu_{n,I^\star} - \mu_{n,i})^2}{2\sigma_{n,i,I^\star}^2}\right\} \\ &= \frac{(K-1)\exp\{-Z_n\}}{2} \\ &\leq \frac{(K-1)\exp\left\{-(\Gamma_\beta^\star - \zeta)n\right\}}{2}. \end{aligned}$$

The last inequality is deduced from Lemma 26. By consequence,

$$\forall n \geq T_\beta^\varepsilon, \ln(1 - a_{n,I^\star}) \leq \ln\frac{K-1}{2} - (\Gamma_\beta^\star - \zeta)n.$$

On the other hand, we have for any n,

$$1 - c_{n,\delta} = \frac{\delta}{2n(K-1)\sqrt{2\pi e}\exp\left\{\sqrt{2\ln\frac{2n(K-1)}{\delta}}\right\}}.$$

Thus, there exists a deterministic time N s.t. $\forall n \geq N$,

$$\begin{aligned} \ln(1 - c_{n,\delta}) &= \ln\frac{\delta}{(K-1)\sqrt{8\pi e}} - \ln n - \sqrt{2\ln\frac{2n(K-1)}{\delta}} \\ &\geq \ln\frac{\delta}{2(K-1)\sqrt{2\pi e}} - \zeta n. \end{aligned}$$

Let $C_3 \triangleq (K-1)^2\sqrt{2\pi e}$, we have for any $n \geq N_0 \triangleq T_\beta^\varepsilon + N$,

$$\ln(1 - a_{n,I^\star}) - \ln(1 - c_{n,\delta}) \leq \ln\frac{C_3}{\delta} - (\Gamma_\beta^\star - 2\zeta)n, \tag{7.13}$$

and it is clear that $\mathbb{E}[N_0] < \infty$.

Let us consider the following two cases:

Case 1 There exists $n \in [1, N_0]$ s.t. $a_{n,I^\star} \geq c_{n,\delta}$, then by definition,

$$\tau_\delta \leq n \leq N_1.$$

Case 2 For any $n \in [1, N_0]$, we have $a_{n,I^\star} < c_{n,\delta}$, then $\tau_\delta \geq N_0 + 1$, thus by Equation 7.13,

$$\begin{aligned} 0 &\leq \ln\left(1 - a_{\tau_\delta - 1, I^\star}\right) - \ln\left(1 - c_{\tau_\delta - 1, \delta}\right) \\ &\leq \ln \frac{C_3}{\delta} - (\Gamma^\star_\beta - 2\zeta)(\tau_\delta - 1), \end{aligned}$$

and we obtain

$$\tau_\delta \leq \frac{\ln(C_3/\delta)}{\Gamma^\star_\beta - 2\zeta} + 1.$$

Combining the two cases, and we have for any $\zeta \in (0, \Gamma^\star_\beta/2)$,

$$\begin{aligned} \tau_\delta &\leq \max\left\{N_0, \frac{\ln(C_3/\delta)}{\Gamma^\star_\beta - 2\zeta} + 1\right\} \\ &\leq N_0 + 1 + \frac{\ln(C_3)}{\Gamma^\star_\beta - 2\zeta} + \frac{\ln(1/\delta)}{\Gamma^\star_\beta - 2\zeta}. \end{aligned}$$

Since $\mathbb{E}[N_1] < \infty$, therefore

$$\limsup_\delta \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \frac{1}{\Gamma^\star_\beta - 2\zeta}, \forall \zeta \in (0, \Gamma^\star_\beta/2),$$

which concludes the proof. □

7.G Technical Lemmas

The whole fixed-confidence analysis for the two sampling rules are both substantially based on two lemmas: Lemma 5 of Qin, Klabjan and Russo, 2017 and Lemma 8. We prove Lemma 8 in this section.

Lemma 8. *There exists a random variable W_2, such that for all $i \in \mathcal{A}$,*

$$\forall n \in, |T_{n,i} - \Psi_{n,i}| \leq W_2 \sqrt{(n+1)\log(e^2 + n)} \text{ a.s.},$$

and $\mathbb{E}\left[e^{\lambda W_2}\right] < \infty$ for any $\lambda > 0$.

Proof. The proof shares some similarities with that of Lemma 6 of Qin, Klabjan and Russo, 2017. For any arm $i \in \mathcal{A}$, define $\forall n \in \mathbb{N}$,

$$D_n \triangleq T_{n,i} - \Psi_{n,i},$$

$$d_n \triangleq \mathbb{1}\{I_n = i\} - \psi_{n,i}.$$

It is clear that $D_n = \sum_{l=1}^{n-1} d_l$ and $\mathbb{E}\left[d_n|\mathcal{F}_{n-1}\right] = 0$. Indeed,

$$\begin{aligned}\mathbb{E}\left[d_n|\mathcal{F}_{n-1}\right] &= \mathbb{E}\left[\mathbb{1}\{I_n = i\} - \psi_{n,i}|\mathcal{F}_{n-1}\right] \\ &= \mathbb{P}\left[I_n = i|\mathcal{F}_{n-1}\right] - \mathbb{E}\left[\mathbb{P}\left[I_n = i|\mathcal{F}_{n-1}\right]|\mathcal{F}_{n-1}\right] \\ &= \mathbb{P}\left[I_n = i|\mathcal{F}_{n-1}\right] - \mathbb{P}\left[I_n = i|\mathcal{F}_{n-1}\right] = 0.\end{aligned}$$

The second last equality holds since $\mathbb{P}\left[I_n = i|\mathcal{F}_{n-1}\right]$ is $\mathcal{F}_{n-1}$-measurable. Thus D_n is a martingale, whose increment are 1 sub-Gaussian as $d_n \in [-1, 1]$ for all n.

Applying Corollary 8 of Abbasi-Yadkori, Pál and Szepesvári, 2012[6], it holds that, with probability larger than $1 - \delta$, for all n,

$$|D_n| \leq \sqrt{2\left(1+n\right)\ln\left(\frac{\sqrt{1+n}}{\delta}\right)}$$

which yields the first statement of Lemma 8.

We now introduce the random variable

$$W_2 \triangleq \max_{n\in\mathbb{N}} \max_{i\in\mathcal{A}} \frac{\left|T_{n,i} - \Psi_{n,i}\right|}{\sqrt{(n+1)\ln(e^2+n)}}.$$

Applying the previous inequality with $\delta = e^{-x^2/2}$ yields

$$\begin{aligned}\mathbb{P}\left[\exists n \in \mathbb{N}^\star : |D_n| > \sqrt{\left(1+n\right)\left(\ln\left(1+n\right) + x^2\right)}\right] &\leq e^{-x^2/2}, \\ \mathbb{P}\left[\exists n \in \mathbb{N}^\star : |D_n| > \sqrt{\left(1+n\right)\ln\left(e^2+n\right)x^2}\right] &\leq e^{-x^2/2},\end{aligned}$$

where the last inequality uses that for all $a, b \geq 2$, we have $ab \geq a + b$.

Consequently $\forall x \geq 2$, for all $i \in \mathcal{A}$

$$\mathbb{P}\left[\max_{n\in\mathbb{N}} \frac{\left|T_{n,i} - \Psi_{n,i}\right|}{\sqrt{(n+1)\log\left(e^2+n\right)}} \geq x\right] \leq e^{-x^2/2}.$$

Now taking a union bound over $i \in \mathcal{A}$, we have $\forall x \geq 2$,

$$\begin{aligned}\mathbb{P}\left[W_2 \geq x\right] &\leq \mathbb{P}\left[\max_{i\in\mathcal{A}} \max_{n\in\mathbb{N}} \frac{\left|T_{n,i} - \Psi_{n,i}\right|}{(n+1)\log\left(\sqrt{e^2+n}\right)} \geq x\right] \\ &\leq \mathbb{P}\left[\bigcup_{i\in\mathcal{A}} \max_{n\in\mathbb{N}} \frac{\left|T_{n,i} - \Psi_{n,i}\right|}{(n+1)\log\left(\sqrt{e^2+n}\right)} \geq x\right] \\ &\leq \sum_{i\in\mathcal{A}} \mathbb{P}\left[\max_{n\in\mathbb{N}} \frac{\left|T_{n,i} - \Psi_{n,i}\right|}{(n+1)\log\left(\sqrt{e^2+n}\right)} \geq x\right] \\ &\leq Ke^{-x^2/2}.\end{aligned}$$

[6] but we could actually use several deviation inequalities that hold uniformly over time for martingales with sub-Gaussian increments

The previous inequalities imply that $\forall i \in \mathcal{A}$ and $\forall n \in \mathbb{N}$, we have

$$|T_{n,i} - \Psi_{n,i}| \leq W_2 \sqrt{(n+1)\log(e^2+n)}$$

almost surely. Now it remains to show that $\forall \lambda > 0, \mathbb{E}\left[e^{\lambda W_2}\right] < \infty$. Fix some $\lambda > 0$.

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda W_2}\right] &= \int_{x=1}^{\infty} \mathbb{P}\left[e^{\lambda W_2} \geq x\right] \mathrm{d}x = \int_{y=0}^{\infty} \mathbb{P}\left[e^{\lambda W_2} \geq e^{2\lambda y}\right] 2\lambda e^{2\lambda y}\, \mathrm{d}y \\
&= 2\lambda \int_{y=0}^{2} \mathbb{P}\left[W_2 \geq 2y\right] e^{2\lambda y}\, \mathrm{d}y + 2\lambda \int_{y=2}^{\infty} \mathbb{P}\left[W_2 \geq 2y\right] e^{2\lambda y}\, \mathrm{d}y \\
&\leq \underbrace{2\lambda \int_{y=0}^{2} \mathbb{P}\left[W_2 \geq 2y\right] e^{2\lambda y}\, \mathrm{d}y}_{= e^{4\lambda - 1}} + \underbrace{2\lambda C_1 \int_{y=2}^{\infty} e^{-y^2/2} e^{2\lambda y}\, \mathrm{d}y}_{< \infty} < \infty,
\end{aligned}$$

where C_1 is some constant.

□

7.H Proof of Posterior Convergence for the Gaussian Bandit

7.H.1 Proof of Theorem 7.13, Gaussian case

Theorem 7.27. *Under TTTS, for Gaussian bandits with improper Gaussian priors, it holds almost surely that*

$$\lim_{n\to\infty} -\frac{1}{n}\log(1 - a_{n,I^\star}) = \Gamma_\beta^\star.$$

From Theorem 2 in Qin, Klabjan and Russo, 2017, any allocation rule satisfying $T_{n,i}/n \to \omega_i^\beta$ for each $i \in \mathcal{A}$, satisfies

$$\lim_{n\to\infty} -\frac{1}{n}\log(1 - a_{n,I^\star}) = \Gamma_\beta^\star.$$

Therefore, to prove Theorem 7.27, it is sufficient to prove that under TTTS,

$$\forall i \in \{1, \ldots, K\}, \quad \lim_{n\to\infty} \frac{T_{n,i}}{n} \overset{a.s}{=} \omega_i^\beta. \tag{7.14}$$

Due to the concentration result in Lemma 8 that we restate below (and proved in Appendix 7.D), which will be useful at several places in the proof, observe that

$$\lim_{n\to\infty} \frac{T_{n,i}}{n} \overset{a.s}{=} \omega_i^\beta \quad \Leftrightarrow \quad \lim_{n\to\infty} \frac{\Psi_{n,i}}{n} \overset{a.s}{=} \omega_i^\beta,$$

therefore it suffices to establish the convergence of $\overline{\psi}_{n,i} = \Psi_{n,i}/n$ to ω_i^β, which we do next. For that purpose, we need again the following maximality inequality lemma.

Lemma 8. *There exists a random variable W_2, such that for all $i \in \mathcal{A}$,*

$$\forall n \in, |T_{n,i} - \Psi_{n,i}| \leq W_2 \sqrt{(n+1)\log(e^2+n)}\ a.s.,$$

and $\mathbb{E}\left[e^{\lambda W_2}\right] < \infty$ for any $\lambda > 0$.

Step 1: TTTS draws all arms infinitely often and satisfies $T_{n,I^\star}/n \to \beta$. More precisely, we prove the following lemma.

Lemma 28. *Under TTTS, it holds almost surely that*

1. *for all $i \in \mathcal{A}$, $\lim_{n\to\infty} T_{n,i} = \infty$.*
2. *$a_{n,I^\star} \to 1$.*
3. *$T_{n,I^\star}/n \to \beta$.*

Proof. Our first ingredient is a lemma showing the implications of finite measurement, and consistency when all arms are sampled infinitely often. Its proof follows standard posterior concentration arguments and is given in Appendix 7.H.2.

Lemma 29 (Consistency and implications of finite measurement)**.**
Denote with $\overline{\mathcal{I}}$ the arms that are sampled only a finite amount of times:

$$\overline{\mathcal{I}} = \{i \in \{1, \dots, k\} : \forall n, T_{n,i} < \infty\}.$$

If $\overline{\mathcal{I}}$ is empty, $a_{n,i}$ converges almost surely to 1 when $i = I^\star$ and to 0 when $i \neq I^\star$. If $\overline{\mathcal{I}}$ is non-empty, then for every $i \in \overline{\mathcal{I}}$, we have $\liminf_{n\to\infty} a_{n,i} > 0$ a.s.

First we show that $\sum_{n\in\mathbb{N}} T_{n,j} = \infty$ for each arm j. Suppose otherwise. Let $\overline{\mathcal{I}}$ again be the set of arms to which only finite measurement effort is allocated. Under TTTS, we have

$$\psi_{n,i} = a_{n,i}\left(\beta + (1-\beta)\sum_{j\neq i}\frac{a_{n,j}}{1-a_{n,j}}\right),$$

so $\psi_{n,i} \geq \beta a_{n,i}$. Therefore, by Lemma 29, if $i \in \overline{\mathcal{I}}$, then $\liminf a_{n,i} > 0$ implies that $\sum_n \psi_{n,i} = \infty$. By Lemma 8, we then must have that $\lim_{n\to\infty} T_{n,i} = \infty$ as well: contradiction. Thus, $\lim_{n\to\infty} T_{n,i} = \infty$ for all i, and we conclude that $a_{n,I^\star} \to 1$, by Lemma 29.

For TTTS with parameter β this implies that $\overline{\psi}_{n,I^\star} \to \beta$, and since we have a bound on $|T_{n,i}/n - \overline{\psi}_{n,i}|$ in Lemma 8, we have $T_{n,I^\star}/n \to \beta$ as well. □

Step 2: Controlling the over-allocation of sub-optimal arms. The convergence of $T_{n,I^\star}/n$ to β leads to following interesting consequence, expressed in Lemma 30: if an arm is sampled more often than its optimal proportion, the posterior probability of this arm to be optimal is reduced compared to that of other sub-optimal arms.

Lemma 30 (Over-allocation implies negligible probability)**.** [7] *Fix any $\xi > 0$ and $j \neq I^\star$. With probability 1, under any allocation rule, if $T_{n,I^\star}/n \to \beta$, there exist $\xi' > 0$ and a sequence ε_n with $\varepsilon_n \to 0$ such that for any $n \in \mathbb{N}$,*

$$\frac{T_{n,j}}{n} \geq \omega_j^\beta + \xi \Rightarrow \frac{a_{n,j}}{\max_{i\neq I^\star} a_{n,i}} \leq e^{-n(\xi'+\varepsilon_n)}.$$

[7] analogue of Lemma 13 of Russo, 2016

Proof. We have $\Pi_n(\Theta_{\cup i \neq I^\star}) = \sum_{i \neq I^\star} a_{n,i} = 1 - a_{n,I^\star}$, therefore $\max_{i \neq I^\star} a_{n,i} \leq 1 - a_{n,I^\star}$. By Theorem 2 of Qin, Klabjan and Russo, 2017 we have, as $T_{n,I^\star}/n \to \beta$,

$$\limsup_{n \to \infty} -\frac{1}{n} \log \left(\max_{i \neq I^\star} a_{n,i} \right) \leq \Gamma^\star_\beta.$$

We also have the following from the standard Gaussian tail inequality, for $n \geq \tau$ after which $\mu_{n,I^\star} \geq \mu_{n,i}$, using that $\theta_i - \theta_{I^\star} \sim \mathcal{N}(\mu_{n,i} - \mu_{n,I^\star}, \sigma^2_{n,i} + \sigma^2_{n,I^\star})$ and $\sigma^2_{n,i} + \sigma^2_{n,I^\star} = \sigma^2(1/T_{n,i} + 1/T_{n,I^\star})$,

$$a_{n,i} \leq \Pi_n(\theta_i \geq \theta_{I^\star}) \leq \exp\left(\frac{-(\mu_{n,i} - \mu_{n,I^\star})^2}{2\sigma^2(1/T_{n,I^\star} + 1/T_{n,i})} \right) = \exp\left(-n \frac{(\mu_{n,i} - \mu_{n,1})^2}{2\sigma^2(n/T_{n,I^\star} + n/T_{n,i})} \right).$$

Thus, there exists a sequence $\varepsilon_n \to 0$, for which

$$\begin{aligned} \frac{a_{n,j}}{\max_{i \neq I^\star} a_{n,i}} &\leq \frac{\exp\left\{ -n \left(\frac{(\mu_{n,j} - \mu_{n,I^\star})^2}{2\sigma^2(n/T_{n,I^\star} + n/T_{n,j})} - \varepsilon_n/2 \right) \right\}}{\exp\left\{ -n \left(\Gamma^\star_\beta + \varepsilon_n/2 \right) \right\})} \\ &= \exp\left\{ -n \left(\frac{(\mu_{n,j} - \mu_{n,I^\star})^2}{2\sigma^2(n/T_{n,I^\star} + n/T_{n,j})} - \Gamma^\star_\beta - \varepsilon_n \right) \right\}. \end{aligned}$$

Now we take a look at the two terms in the middle:

$$\frac{(\mu_{n,j} - \mu_{n,I^\star})^2}{2\sigma^2(n/T_{n,I^\star} + n/T_{n,j})} - \Gamma^\star_\beta.$$

Note that the first term is increasing in $T_{n,j}/n$. We have the definition from Qin, Klabjan and Russo, 2017, for any $j \neq I^\star$,

$$\Gamma^\star_\beta = \frac{(\mu_j - \mu_{I^\star})^2}{2\sigma^2 \left(1/\omega^\beta_{I^\star} + 1/\omega^\beta_j \right)},$$

and we have the premise

$$\frac{T_{n,j}}{n} \geq \omega^\beta_j + \xi.$$

Combining these with the convergence of the empirical means to the true means (consistency, see Lemma 29), we can conclude that for all $\varepsilon > 0$, there exists a time n_0 such that for all later times $n \geq n_0$, we have

$$\frac{(\mu_{n,j} - \mu_{n,I^\star})^2}{2\sigma^2(n/T_{n,I^\star} + n/T_{n,j})} \geq \frac{(\mu_j - \mu_{I^\star})^2}{2\sigma^2 \left(1/\beta + n/T_{n,j} \right)} - \varepsilon \geq \frac{(\mu_j - \mu_{I^\star})^2}{2\sigma^2 \left(1/\beta + 1/(\omega^\beta_j + \xi) \right)} - \varepsilon > \Gamma^\star_\beta,$$

where the first inequality follows from consistency, the second from monotonicity in $T_{n,j}/n$. That means that there exist a $\xi' > 0$ such that

$$\frac{(\mu_{n,j} - \mu_{n,I^\star})^2}{2\sigma^2(n/T_{n,I^\star} + n/T_{n,j})} - \Gamma^\star_\beta > \xi',$$

and thus the claim follows that when $\frac{T_{n,j}}{n} \geq \omega_j^\beta + \xi$, we have

$$\frac{a_{n,j}}{\max_{i \neq I^\star} a_{n,i}} \leq \exp\left\{-n\left(\frac{(\mu_{n,j} - \mu_{n,I^\star})^2}{2\sigma^2(n/T_{n,I^\star} + n/T_{n,j})} - \Gamma_\beta^\star - \varepsilon_n\right)\right\} \leq e^{-n(\xi' + \varepsilon_n)}.$$

□

Step 3: $\overline{\psi}_{n,i}$ converges to ω_i^β for all arms. To establish the convergence of the allocation effort of all arms, we rely on the same sufficient condition used in the analysis of Russo, 2016, that we recall below.

Lemma 31 (Sufficient condition for optimality). [8] *Consider any adaptive allocation rule. If we have*

$$\overline{\psi}_{n,I^\star} \to \beta, \quad \text{and} \quad \sum_{n \in \mathbb{N}} \psi_{n,j} \mathbf{1}\left\{\overline{\psi}_{n,j} \geq \omega_j^\beta + \xi\right\} < \infty, \quad \forall j \neq I^\star, \xi > 0, \tag{7.15}$$

then $\overline{\psi}_n \to \psi^\beta$.

First, note that from Lemma 28 we know that $T_{n,I^\star}/n \to \beta$, an by Lemma 8 this implies $\overline{\psi}_{n,I^\star} \to \beta$, hence we can use Lemma 31 to prove convergence to the optimal proportions. Thus, we now show that (7.15) holds under TTTS. Recall that $J_n^{(1)} = \arg\max_j a_{n,j}$ and $J_n^{(2)} = \arg\max_{j \neq J_n^{(1)}} a_{n,j}$. Since $a_{n,I^\star} \to 1$ by Lemma 28, there is some finite time τ after which for all $n > \tau$, $J_n^{(1)} = I^\star$. Under TTTS,

$$\begin{aligned}
\psi_{n,i} &= a_{n,i}\left(\beta + (1-\beta)\sum_{j \neq i} \frac{a_{n,j}}{1 - a_{n,j}}\right) \\
&\leq a_{n,i}\beta + a_{n,i}(1-\beta)\frac{\sum_{j \neq i} a_{n,j}}{1 - a_{n,J_n^{(1)}}} \\
&\leq a_{n,i}\beta + a_{n,i}(1-\beta)\frac{\sum_{j \neq i} a_{n,j}}{a_{n,J_n^{(2)}}} \\
&\leq a_{n,i}\beta + a_{n,i}(1-\beta)\frac{1}{a_{n,J_n^{(2)}}} \\
&\leq \frac{a_{n,i}}{a_{n,J_n^{(2)}}},
\end{aligned}$$

where we use the fact that for $j \neq J_n^{(1)}$, we have $a_{n,J_n^{(1)}} \geq a_{n,j}$ and $a_{n,J_n^{(2)}} \leq 1 - a_{n,J_n^{(1)}}$. For $n \geq \tau$ this means that $\psi_{n,i} \leq a_{n,i}/\max_{j \neq I^\star} a_{n,i}$ for any $i \neq I^\star$.

By Lemma 30, there is a constant $\xi' > 0$ such and a sequence $\varepsilon_n \to 0$ such that

$$T_{n,i}/n \geq w_i^\beta + \xi \Rightarrow \frac{a_{n,i}}{\max_{j \neq I^\star} a_{n,j}} \leq e^{-n(\xi' - \varepsilon_n)}.$$

[8] Lemma 12 of Russo, 2016

Now take a time τ large enough, such that for $n \geq \tau$ we have $|T_{n,j}/n - \overline{\psi}_{n,j}| \leq \xi$ (which can be found by Lemma 8). Then we have

$$\mathbb{1}\left\{\overline{\psi}_{n,j} \geq \psi_j^\beta + \xi\right\} \leq \mathbb{1}\left\{\frac{T_{n,j}}{n} \geq \omega_j^\beta + 2\xi\right\}$$

Therefore, for all $i \neq I^\star$, we have

$$\sum_{n\geq\tau} \psi_{n,i}\mathbb{1}\left\{\overline{\psi}_{n,j} \geq \psi_j^\beta + \xi\right\} \leq \sum_{n\geq\tau} \psi_{n,i}\mathbb{1}\left\{\frac{T_{n,j}}{n} \geq \omega_j^\beta + 2\xi\right\} \leq \sum_{n\geq\tau} e^{-n(\xi'-\varepsilon_n)} < \infty.$$

Thus (7.15) holds and the convergence to the optimal proportions follows by Lemma 31.

7.H.2 Proof of auxiliary lemmas

Proof of Lemma 29 Let $\overline{\mathcal{I}}$ be nonempty. Define

$$\mu_{\infty,n} \triangleq \lim_{n\to\infty} \mu_{n,i}, \text{ and } \sigma^2_{\infty,i} \triangleq \lim_{n\to\infty} \sigma^2_{n,i},$$

and recall that for $i \in \mathcal{A}$ for which $T_{n,i} = 0$, we have $\mu_{n_i} = \mu_{1,i} = 0$ and $\sigma^2_{n,i} = \sigma^2_{1,i} = \infty$, and if $T_{n,i} > 0$, we have

$$\mu_{n,i} = \frac{1}{T_{n,i}} \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell,I_\ell}, \text{ and } \sigma^2_{n,i} = \frac{\sigma^2}{T_{n,i}}.$$

For all arms that are sampled infinitely often, we therefore have $\mu_{\infty,i} = \mu_i$ and $\sigma^2_{\infty,i} = 0$. For all arms that are sampled only a finite number of times, i.e. $i \in \overline{\mathcal{I}}$, we have $\sigma^2_{\infty,i} > 0$, and there exists a time n_0 after which for all $n \geq n_0$ and $i \in \overline{\mathcal{I}}$, we have $T_{n,i} = T_{n_0,i}$. Define

$$\Pi_\infty \triangleq \mathcal{N}(\mu_{\infty,1}, \sigma^2_{\infty,1}) \otimes \mathcal{N}(\mu_{\infty,2}, \sigma^2_{\infty,2}) \otimes \ldots \otimes \mathcal{N}(\mu_{\infty,k}, \sigma^2_{\infty,k}) = \bigotimes_{i\notin\overline{\mathcal{I}}} \delta_{\mu_i} \otimes \bigotimes_{i\in\overline{\mathcal{I}}} \Pi_{n_0}.$$

Then for each $i \in \mathcal{A}$ we define

$$a_{\infty,i} \triangleq \Pi_\infty\left(\theta_i > \max_{j\neq i} \theta_j\right).$$

Then we have for all $i \in \overline{\mathcal{I}}$, $a_{\infty,i} \in (0,1)$, since $\sigma^2_{\infty,i} > 0$, and thus $a_{\infty,I^\star} < 1$.

When $\overline{\mathcal{I}}$ is empty, we have $a_{n,I^\star} = \Pi_n(\theta_{I^\star} > \max_{i\neq I^\star} \theta_i)$, but since $\Pi_\infty = \bigotimes_{i\in\mathcal{A}} \delta_{\mu_i}$, we have $a_{\infty,I^\star} = 1$ and $a_{\infty,i} = 0$ for all $i \neq I^\star$.

■

7.I Proof of Posterior Convergence for the Bernoulli Bandit

7.I.1 Preliminaries

We first introduce a crucial Beta tail bound inequality. Let $F^{\text{Beta}}_{a,b}$ denote the cdf of a Beta distribution with parameters a and b, and $F^{\text{B}}_{c,d}$ the cdf of a Binomial distribution with parameters c and d, then we have the following relationship, often called the 'Beta-Binomial trick',

$$F^{\text{Beta}}_{a,b}(y) = 1 - F^{\text{B}}_{a+b-1,y}(a-1),$$

so that we have

$$\mathbb{P}\left[X \geq x\right] = \mathbb{P}\left[B_{a+b-1,x} \leq a-1\right] = \mathbb{P}\left[B_{a+b-1,1-x} \geq b\right].$$

We can bound Binomial tails with Sanov's inequality:

$$\frac{e^{-nd(k/n,x)}}{n+1} \leq \mathbb{P}\left[B_{n,x} \geq k\right] \leq e^{-nd(k/n,x)},$$

where the last inequalities hold when $k \geq nx$.

Lemma 32. *Let $X \sim \mathcal{B}eta(a,b)$ and $Y \sim \mathcal{B}eta(c,d)$ with $0 < \frac{a-1}{a+b-1} < \frac{c-1}{c+d-1}$. Then we have $\mathbb{P}\left[X > Y\right] \leq De^{-C}$ where*

$$C = \inf_{\frac{a-1}{a+b-1} \leq y \leq \frac{c-1}{c+d-1}} C_{a,b}(y) + C_{c,d}(y),$$

and

$$D = 3 + \min\left(C_{a,b}\left(\frac{c-1}{c+d-1}\right), C_{c,d}\left(\frac{a-1}{a+b-1}\right)\right).$$

Note that this lemma is the Bernoulli version of Lemma 6.

Theorem 7.33. *Consider the Beta-Bernoulli setting. For $\beta \in (0,1)$, under any allocation rule satisfying $T_{n,I^\star}/n \to \omega^\beta_{I^\star}$,*

$$\lim_{n\to\infty} -\frac{1}{n}\log(1 - a_{n,I^\star}) \leq \Gamma^\star_\beta,$$

and under any allocation rule satisfying $T_{n,i}/n \to \omega^\beta_i$ for each $i \in \mathcal{A}$,

$$\lim_{n\to\infty} -\frac{1}{n}\log(1 - a_{n,I^\star}) = \Gamma^\star_\beta.$$

Proof. Denote again with $\overline{\mathcal{I}}$ again the set of arms sampled only finitely many times. For $\overline{\mathcal{I}}$ empty, we thus have $\mu_{\infty,i} \triangleq \lim_{n\to\infty} \mu_{n,i} = \mu_i$. The posterior variance is

$$\sigma^2_{n,i} = \frac{\alpha_{n,i}\beta_{n,i}}{(\alpha_{n,i}+\beta_{n,i})^2(\alpha_{n,i}+\beta_{n,i}+1)} = \frac{(1+\sum_{\ell=1}^{n-1}\mathbb{1}\{I_\ell = i\}Y_{\ell,I_\ell})(1+T_{n,i}-\sum_{\ell=1}^{n-1}\mathbb{1}\{I_\ell = i\}Y_{\ell,I_\ell})}{(2+T_{n,i})^2(2+T_{n,i}+1)}.$$

We see that when $\overline{\mathcal{I}}$ is empty, we have $\sigma^2_{\infty,i} \triangleq \lim_{n\to\infty}\sigma^2_{n,i} = 0$, i.e., the posterior is concentrated.

Step 1: A lower bound when some arms are sampled only finitely often. First, note that when $T_{n,i} = 0$ for some $i \in \mathcal{A}$, the empirical mean for that arm equals the prior mean

$$\mu_{n,i} = \alpha_{0,i}/(\alpha_{0,i} + \beta_{0,i}),$$

and the variance is strictly positive:

$$\sigma^2_{n,i} = (\alpha_{0,i}\beta_{0,i})/\left((\alpha_{0,i} + \beta_{0,i})^2(\alpha_{0,i} + \beta_{0,i} + 1)\right) > 0.$$

When $\overline{\mathcal{I}}$ is not empty, then for every $i \in \overline{\mathcal{I}}$ we have $\sigma^2_{\infty,i} > 0$, and $a_{\infty,i} \in (0,1)$, implying $a_{\infty,I^\star} < 1$, and thus

$$\lim_{n\to\infty} -\frac{1}{n}\log(1 - a_{n,I^\star}) = -\frac{1}{n}\log(1 - a_{\infty,I^\star}) = 0.$$

Step 2: A lower bound when every arm is sampled infinitely often. Suppose now that $\overline{\mathcal{I}}$ is empty, then we have

$$\max_{i\neq I^\star} \Pi_n(\theta_i \geq \theta_{I^\star}) \leq 1 - a_{n,I^\star} \leq \sum_{i\neq I^\star} \Pi_n(\theta_i \geq \theta_{I^\star}) \leq (k-1)\max_{i\neq I^\star} \Pi_n(\theta_i \geq \theta_{I^\star}).$$

Thus, we have $1 - a_{n,I^\star} \leq (k-1)\max_{i\neq I^\star} \Pi_n(\theta_i \geq \theta_{I^\star})$ and also $1 - a_{n,I^\star} \doteq \max_{i\neq I^\star} \Pi_n(\theta_i \geq \theta_{I^\star})$. We have

$$\begin{aligned}
\Gamma^\star &= \max_{w\in W} \min_{i\neq I^\star} C_i(\omega_{I^\star}, \omega_i), \\
\Gamma^\star_\beta &= \max_{w\in W; \omega_{I^\star} = \beta} \min_{i\neq I^\star} C_i(\beta, \omega_i), \text{ with} \\
C_i(\omega_{I^\star}, \omega_i) &= \min_{x\in\mathbb{R}} \omega_{I^\star} d(\theta_{I^\star}; x) + \omega_i d(\theta_i; x) = \omega_{I^\star} d(\theta_{I^\star}; \overline{\theta}) + \omega_i d(\theta_i; \overline{\theta}),
\end{aligned}$$

where $\overline{\theta} \in [\theta_i, \theta_{I^\star}]$ is the solution to

$$A'(\overline{\theta}) = \frac{\omega_{I^\star} A'(\theta_{I^\star}) + \omega_i A'(\theta_i)}{\omega_{I^\star} + \omega_i}.$$

Since every arm is sampled infinitely often, when n is large, we have $\mu_{n,I^\star} > \mu_{n,i}$. Define $S_{n,i} \triangleq \sum_{\ell=1}^{n-1} \mathbb{1}\{I_\ell = i\} Y_{\ell,I_\ell}$. Recall that the posterior is a Beta distribution with parameters $a_{n,i} = S_{n,i} + 1$ and $\beta_{n,i} = T_{n,i} - S_{n,i} + 1$. Let $\tau \in \mathbb{N}$ be such that for every $n \geq \tau$, we have $S_{n,i}/(T_{n,i}+1) < S_{n,I^\star}/(T_{n,I^\star}+1)$. For the sake of simplicity, we define for any $i \in \mathcal{A}$ the interval

$$I_{i,I^\star} \triangleq \left[\frac{S_{n,i}}{T_{n,i}+1}, \frac{S_{n,I^\star}}{T_{n,I^\star}+1}\right].$$

Then using Lemma 32 with $a = S_{n,i}+1, b = T_{n,i} - S_{n,i} + 1, c = S_{n,I^\star} + 1, d = T_{n,I^\star} - S_{n,I^\star} + 1$, we have

$$\Pi_n(\theta_i - \theta_{I^\star} \geq 0) \leq D\exp\left\{-\inf_{y\in I_{i,I^\star}} C_{S_{n,i}+1, T_{n,i}-S_{n,i}+1}(y) + C_{S_{n,I^\star}+1, T_{n,I^\star}-S_{n,I^\star}+1}(y)\right\}.$$

This implies

$$\frac{1}{n}\log\left(\frac{\Pi_n(\theta_i \geq \theta_{I^\star})}{\exp\left\{-\inf_{y\in I_{i,I^\star}} C_{S_{n,i}+1,T_{n,i}-S_{n,i}+1}(y) + C_{S_{n,I^\star}+1,T_{n,I^\star}-S_{n,I^\star}+1}(y)\right\}}\right) \leq \frac{1}{n}\log(D),$$

which goes to zero as n goes to infinity. Indeed replacing a, b, c, d by their values in the definition of D we get

$$\begin{aligned} D &\leq 3 + (T_{n,i}-1)kl\left(\frac{S_{n,i}}{T_{n,i}+1}; \frac{S_{n,I^\star}}{T_{n,I^\star}+1}\right) \\ &\leq 3 + (n+1)kl\left(0; \frac{n}{n+1}\right) \\ &= (n+1)\log(n+1)\,. \end{aligned}$$

Hence,

$$\Pi_n(\theta_i \geq \theta_{I^\star}) \doteq \exp\left\{-\inf_{y\in I_{i,I^\star}} C_{S_{n,i}+1,T_{n,i}-S_{n,i}+1}(y) + C_{S_{n,I^\star}+1,T_{n,I^\star}-S_{n,I^\star}+1}(y)\right\}.$$

We thus have for any i,

$$\begin{aligned} 1-a_{n,i} &\doteq \max_{j\neq I^\star} \Pi_n\left[\theta_j \geq \theta_{I^\star}\right] \\ &\doteq \max_{j\neq I^\star} \exp\left\{-\inf_{y\in I_{j,I^\star}} C_{S_{n,j}+1,T_{n,j}-S_{n,j}+1}(y) + C_{S_{n,I^\star}+1,T_{n,I^\star}-S_{n,I^\star}+1}(y)\right\} \\ &\doteq \exp\left\{-n\min_{j\neq I^\star}\inf_{y\in I_{j,I^\star}} \frac{T_{n,j}+1}{n} kl\left(\frac{S_{n,j}}{T_{n,j}+1}; y\right) + \frac{T_{n,I^\star}+1}{n} kl\left(\frac{S_{n,I^\star}}{T_{n,I^\star}+1}; y\right)\right\} \\ &\geq \exp\left\{-n\max_{\boldsymbol{\omega}}\min_{j\neq I^\star}\inf_{y\in I_{j,I^\star}} \omega_i kl\left(\frac{S_{n,j}}{T_{n,j}+1}; y\right) + \omega_{I^\star} kl\left(\frac{S_{n,I^\star}}{T_{n,j}+1}; y\right)\right\}. \end{aligned}$$

Fix some $\varepsilon > 0$, then there exists some $n_0(\varepsilon)$ such that for all $n \geq n_0(\varepsilon)$, we have for any j,

$$I_{j,I^\star} = \left[\frac{S_{n,j}}{T_{n,j}+1}, \frac{S_{n,I^\star}}{T_{n,I^\star}+1},\right] \subset \left[\mu_j + \varepsilon, \mu_{I^\star} - \varepsilon\right] \triangleq I^\star_{j,\varepsilon},$$

and because KL-divergence is uniformly continuous on the compact interval $I^\star_{j,\varepsilon}$, there exists an n_1 such that for every $n \geq n_1$ we have

$$kl\left(\frac{S_{n,j}}{T_{n,j}+1}; y\right) \geq (1-\varepsilon)kl\left(\mu_j; y\right),$$

for any y and for all $j \in \mathcal{A}$. Therefore, we have

$$\begin{aligned} 1-a_{n,i} &\doteq \exp\left\{-n\max_{\boldsymbol{\omega}}\min_{j\neq I^\star}\inf_{y\in I_{j,I^\star}} \omega_j kl\left(\frac{S_{n,j}}{T_{n,j}+1}; y\right) + \omega_{I^\star} kl\left(\frac{S_{n,I^\star}}{T_{n,I^\star}+1}; y\right)\right\} \\ &\geq \exp\left\{-n\max_{\boldsymbol{\omega}}\min_{i\neq I^\star}\inf_{y\in I^\star_{j,\varepsilon}} \omega_i kl(\mu_j; y) + \omega_{I^\star} kl(\mu_{I^\star}; y)\right\}. \end{aligned}$$

Therefore, we have

$$\limsup_{n\to\infty} -\frac{1}{n}\log(1-a_{n,i}) \le \Gamma^\star.$$

If $T_{n,i}/n \to \omega_i^\star$ for each $i \in \mathcal{A}$, we have

$$\begin{aligned}\lim_{n\to\infty} \inf_{y\in I_{i,I^\star}} \frac{T_{n,i}+1}{n} kl\left(\frac{S_{n,i}}{T_{n,i}+1}; y\right) + \frac{T_{n,I^\star}+1}{n} kl\left(\frac{S_{n,I^\star}}{T_{n,i}+1}; y\right)\\ = \inf_{y\in[\mu_i,\,\mu_{I^\star}]} \omega_i^\star kl(\mu_i; y) + \omega_{I^\star}^\star kl(\mu_{I^\star}; y)\\ = \Gamma^\star,\end{aligned}$$

and thus

$$\begin{aligned}1-a_{n,i} &\doteq \exp\left\{-n \max_{\boldsymbol{\omega}} \min_{j\neq I^\star} \inf_{y\in I_\varepsilon^\star} \omega_i kl(\mu_j; y) + \omega_{I^\star} kl(\mu_{I^\star}; y)\right\}\\ &\doteq \exp\left\{-n\Gamma^\star\right\},\end{aligned}$$

implying

$$\lim_{n\to\infty} -\frac{1}{n}\log\left(1-a_{n,i}\right) = \Gamma^\star.$$

Everything goes similarly when $\omega_{I^\star} = \beta \in (0,1)$, so under any sampling rule satisfying $T_{n,I^\star}/n \to \beta$ we have

$$\limsup_{n\to\infty} -\frac{1}{n}\log(1-a_{n,i}) \le \Gamma_\beta^\star$$

and under any sampling rule satisfying $T_{n,i}/n \to \omega_i^\beta$ for each $i \in \mathcal{A}$, we have

$$\lim_{n\to\infty} -\frac{1}{n}\log(1-a_{n,i}) = \Gamma_\beta^\star.$$

□

7.I.2 Proof of Theorem 7.13, Bernoulli case

Theorem 7.34. *Under TTTS, for Bernoulli bandits and uniform priors, it holds almost surely that*

$$\lim_{n\to\infty} -\frac{1}{n}\log(1-a_{n,I^\star}) = \Gamma_\beta^\star.$$

From Theorem 7.33 we know that under any allocation rule satisfying $T_{n,i}/n \to \omega_i^\beta$ for every $i \in \mathcal{A}$, we have

$$\lim_{n\to\infty} -\frac{1}{n}\log\left(1-a_{n,I^\star}\right) = \Gamma_\beta^\star.$$

Thus, we only need to prove that under TTTS, for all $i \in \mathcal{A}$, we have

$$\lim_{n\to\infty} \frac{T_{n,i}}{n} \overset{a.s}{=} \omega_i^\beta.$$

Just as for the proof of the Gaussian case, we can use Lemma 8 (proof in Appendix 7.H.2), which implies

$$\lim_{n\to\infty} \frac{T_{n,i}}{n} \overset{a.s}{=} \omega_i^\beta \quad \Leftrightarrow \quad \lim_{n\to\infty} \frac{\Psi_{n,i}}{n} \overset{a.s}{=} \omega_i^\beta.$$

Therefore, it suffices to show convergence for $\overline{\psi}_{n,i} = \Psi_{n,i}/n$ to ω_i^β, which we will do next, following the same steps as in the proof for the Gaussian case.

Step 1: TTTS draws all arms infinitely often and satisfies $T_{n,I^\star}/n \to \beta$. We prove the following lemma.

Lemma 35. *Under TTTS, it holds almost surely that*

1. *for all $i \in \mathcal{A}$, $\lim_{n\to\infty} T_{n,i} = \infty$.*
2. *$a_{n,I^\star} \to 1$.*
3. *$\frac{T_{n,I^\star}}{n} \to \beta$.*

Proof. First, we give a lemma showing the implications of finite measurement, and consistency when all arms are sampled infinitely often, which provides a proof for 2. The proof of this lemma follows from the proof of Theorem 7.33, and is given in Appendix 7.I.3.

Lemma 36 (Consistency and implications of finite measurement)**.**
Denote with $\overline{\mathcal{I}}$ the arms that are sampled only a finite amount of times:

$$\overline{\mathcal{I}} = \{i \in \{1, \dots, k\} : \forall n, T_{n,i} < \infty\}.$$

If $\overline{\mathcal{I}}$ is empty, $a_{n,i}$ converges almost surely to 1 when $i = I^\star$ and to 0 when $i \neq I^\star$. If $\overline{\mathcal{I}}$ is non-empty, then for every $i \in \overline{\mathcal{I}}$, we have $\liminf_{n\to\infty} a_{n,i} > 0$ a.s.

Now we can show 1. of Lemma 35: we show that under TTTS, for each $j \in A$, we have $\sum_{n\in\mathbb{N}} T_{n,j} = \infty$. The proof is exactly equal to the proof for Gaussian arms.

Under TTTS, we have

$$\psi_{n,i} = a_{n,i}\left(\beta + (1-\beta)\sum_{j\neq i} \frac{a_{n,j}}{1-a_{n,j}}\right),$$

so $\psi_{n,i} \geq \beta a_{n,i}$, therefore, by Lemma 29, if $i \in \overline{\mathcal{I}}$, then $\liminf a_{n,i} > 0$ implies that $\sum_n \psi_{n,i} = \infty$. By Lemma 8, we then must have that $\lim_{n\to\infty} T_{n,i} = \infty$ as well: contradiction. Thus, $\lim_{n\to\infty} T_{n,i} = \infty$ for all i, and we conclude that $a_{n,I^\star} \to 1$, by Lemma 29.

Lastly we prove point 3. of Lemma 35. For TTTS with parameter β, the above implies that $\overline{\psi}_{n,I^\star} \to \beta$, and since we have a bound on $|T_{n,i}/n - \overline{\psi}_{n,i}|$ in Lemma 8, we have $T_{n,I^\star}/n \to \beta$ as well.

□

Step 2: Controlling the over-allocation of sub-optimal arms. Following the proof for the Gaussian case again, we can establish a consequence of the convergence of $T_{n,I^\star}/n$ to β : if an arm is sampled more often than its optimal proportion, the posterior probability of this arm to be optimal is reduced compared to that of other sub-optimal arms. We can prove this by using ingredients from the proof of the lower bound in Theorem 7.33.

Lemma 37 (Over-allocation implies negligible probability)**.** [9]
Fix any $\xi > 0$ and $j \neq I^\star$. With probability 1, under any allocation rule, if $T_{n,I^\star}/n \to \beta$, there exist $\xi' > 0$ and a sequence ε_n with $\varepsilon_n \to 0$ such that for any $n \in \mathbb{N}$,

$$\frac{T_{n,j}}{n} \geq \omega_j^\beta + \xi \implies \frac{a_{n,j}}{\max_{i \neq I^\star} a_{n,i}} \leq e^{-n(\xi' + \varepsilon_n)}.$$

Proof. By Theorem 7.33, we have, as $T_{n,I^\star}/n \to \beta$,

$$\limsup_{n \to \infty} -\frac{1}{n} \log \left(\max_{i \neq I^\star} a_{n,i} \right) \leq \Gamma_\beta^\star,$$

since $\max_{i \neq I^\star} a_{n,i} \leq 1 - a_{n,I^\star}$. We also have from Lemma 32 a deviation inequality, so that we can establish the following logarithmic equivalence:

$$a_{n,j} \leq \Pi_n(\theta_j \geq \theta_{I^\star}) \doteq \exp\left\{-nC_j\left(w_{n,I^\star}, \omega_{n,j}\right)\right\} \doteq \exp\left\{-nC_j\left(\beta, \omega_{n,j}\right)\right\},$$

where we denote $\omega_{n,j} \triangleq \frac{T_{n,j}}{n}$. We can combine these results, which implies that there exists a non-negative sequence $\varepsilon_n \to 0$ such that

$$\frac{a_{n,j}}{\max_{i \neq I^\star} a_{n,i}} \leq \frac{\exp\left\{-nC_j\left(\beta, \omega_{n,j}\right) - \varepsilon_n/2\right\}}{\exp\left\{-n(\Gamma_\beta^\star + \varepsilon/2)\right\}} = \exp\left\{-n\left(C_j\left(\beta, \omega_{n,j}\right) - \Gamma_\beta^\star\right) - \varepsilon_n\right\}.$$

We know that $C_j\left(\beta, \omega_j^\beta\right)$ is strictly increasing in ω_j^β, and $C_j\left(\beta, \omega_j^\beta\right) = \Gamma_\beta^\star$, thus, there exists some $\xi' > 0$ such that

$$\omega_{n,j} \geq \omega_j^\beta + \xi \implies C_j\left(\beta, \omega_{n,j}\right) - \Gamma_\beta^\star > \xi'.$$

□

[9] analogue of Lemma 13 of Russo, 2016

Step 3: $\overline{\psi}_{n,i}$ converges to ω_i^β for all arms. To establish the convergence of the allocation effort of all arms, we rely on the same sufficient condition used in the analysis of Russo, 2016, restated above in Lemma 31, and we will restate it here again for convenience.

Lemma 38 (Sufficient condition for optimality)**.**
Consider any adaptive allocation rule. If

$$\overline{\psi}_{n,I^\star} \to \beta, \quad \textit{and} \quad \sum_{n\in\mathbb{N}} \psi_{n,j} \mathbf{1}\left\{\overline{\psi}_{n,j} \geq \omega_j^\beta + \xi\right\} < \infty, \ \ \forall\, j \neq I^\star, \xi > 0, \tag{7.16}$$

then $\overline{\psi}_n \to \psi^\beta$.

First, note that from Lemma 35 we know that $\frac{T_{n,I^\star}}{n} \to \beta$, and by Lemma 8 this implies $\overline{\psi}_{n,I^\star} \to \beta$, hence we can use the lemma above to prove convergence to the optimal proportions. This proof is already given in Step 3 of the proof for the Gaussian case, and since it does not depend on the specifics of the Gaussian case, except for invoking Lemma 29 (consistency), which for the Bernoulli case we replace by Lemma 36, it gives a proof for the Bernoulli case as well. We conclude that (7.15) holds, and the convergence to the optimal proportions follows by Lemma 31.

7.I.3 Proof of auxiliary lemmas

Lemma 32. *Let $X \sim \mathcal{B}eta(a,b)$ and $Y \sim \mathcal{B}eta(c,d)$ with $0 < \frac{a-1}{a+b-1} < \frac{c-1}{c+d-1}$. Then we have $\mathbb{P}[X > Y] \leq De^{-C}$ where*

$$C = \inf_{\frac{a-1}{a+b-1} \leq y \leq \frac{c-1}{c+d-1}} C_{a,b}(y) + C_{c,d}(y),$$

and

$$D = 3 + \min\left(C_{a,b}\left(\frac{c-1}{c+d-1}\right), C_{c,d}\left(\frac{a-1}{a+b-1}\right)\right).$$

Proof

$$\begin{aligned}
\mathbb{P}[X > Y] = \mathbb{E}[\mathbb{P}[X > Y|Y]] &\leq \mathbb{E}\left[\mathbb{1}\{Y < \frac{a-1}{a+b-1}\} + \mathbb{1}\{Y \geq \frac{a-1}{a+b-1}\}\mathbb{P}[X > Y|Y]\right] \\
&\leq \exp\left\{-(c+d-1)kl\left(\frac{c-1}{c+d-1}; \frac{a-1}{a+b-1}\right)\right\} \\
&\quad + \underbrace{\mathbb{E}\left[\exp\left\{-(a+b-1)kl\left(\frac{a-1}{a+b-1}; Y\right)\right\}\mathbb{1}\{Y \geq \frac{a-1}{a+b-1}\}\right]}_{A},
\end{aligned}$$

Using the Beta-Binomial trick in the second inequality. Furthermore, we have

$$A \leq \underbrace{\mathbb{E}\left[\mathbb{1}\{\frac{a-1}{a+b-1} \leq Y \leq \frac{c-1}{c+d-1}\}\right] \exp\left\{-(a+b-1)kl\left(\frac{a-1}{a+b-1}; Y\right)\right\}}_{B} + \exp\left\{-(a+b-1)kl\left(\frac{a-1}{a+b-1}; \frac{c-1}{c+d-1}\right)\right\}$$

Denote with f the density of Y, then

$$B = \int_{\frac{a-1}{a+b-1}}^{\frac{c-1}{c+d-1}} \exp\left\{-(a+b-1)kl\left(\frac{a-1}{a+b-1}; y\right)\right\} f(y)\,\mathrm{d}y.$$

Via integration by parts we obtain

$$\begin{aligned} B &= \left[\exp\left\{-(a+b-1)kl\left(\frac{a-1}{a+b-1}; y\right)\right\}\mathbb{P}\left[Y \leq y\right]\right]_{\frac{a-1}{a+b-1}}^{\frac{c-1}{c+d-1}} \\ &\quad + \int_{\frac{a-1}{a+b-1}}^{\frac{c-1}{c+d-1}} (a+b-1)\frac{\mathrm{d}}{\mathrm{d}y}kl\left(\frac{a-1}{a+b-1}; y\right)\exp\left\{-C_{a,b}(y)\right\}P(Y \leq y)\,\mathrm{d}y \\ &\leq \int_{\frac{a-1}{a+b-1}}^{\frac{c-1}{c+d-1}} (a+b-1)\frac{\mathrm{d}}{\mathrm{d}y}kl\left(\frac{a-1}{a+b-1}; y\right)\exp\left\{-(C_{a,b}(y)+C_{c,d}(y))\right\}\mathrm{d}y \\ &\quad + \exp\left\{-(a+b-1)kl\left(\frac{a-1}{a+b-1}; \frac{c-1}{c+d-1}\right)\right\}, \end{aligned}$$

where the first inequality uses the Binomial trick again. Let

$$\begin{aligned} C &= \inf_{\frac{a-1}{a+b-1} \leq y \leq \frac{c-1}{c+d-1}} (a+b-1)kl\left(\frac{a-1}{a+b-1}; y\right) + (c+d-1)kl\left(\frac{c-1}{c+d-1}; y\right) \\ &= \inf_{\frac{a-1}{a+b-1} \leq y \leq \frac{c-1}{c+d-1}} C_{a,b}(y) + C_{c,d}(y), \end{aligned}$$

then note that in particular we have

$$\begin{aligned} C &\leq \min\left((a+b-1)kl\left(\frac{a-1}{a+b-1}; \frac{c-1}{c+d-1}\right), (c+d-1)kl\left(\frac{c-1}{c+d-1}; \frac{a-1}{a+b-1}\right)\right) \\ &= \min\left(C_{a,b}\left(\frac{c-1}{c+d-1}\right), C_{c,d}\left(\frac{a-1}{a+b-1}\right)\right). \end{aligned}$$

Then

$$\begin{aligned} B &\leq e^{-C}\int_{\frac{a-1}{a+b-1}}^{\frac{c-1}{c+d-1}} (a+b-1)\frac{\mathrm{d}}{\mathrm{d}y}kl\left(\frac{a-1}{a+b-1}; y\right)\mathrm{d}y + e^{-C} \\ &= \left[(a+b-1)kl\left(\frac{a-1}{a+b-1}; \frac{c-1}{c+d-1}\right)+1\right]e^{-C}. \end{aligned}$$

Thus we have

$$\mathbb{P}\left[X > Y\right] \leq \left(3 + (a+b-1)kl\left(\frac{a-1}{a+b-1}; \frac{c-1}{c+d-1}\right)\right)e^{-C}.$$

By symmetry, we have

$$\mathbb{P}\left[X > Y\right] \leq \left(3 + \min\left(C_{a,b}\left(\frac{c-1}{c+d-1}\right), C_{c,d}\left(\frac{a-1}{a+b-1}\right)\right)\right)e^{-C},$$

where

$$C = \inf_{\frac{a-1}{a+b-1} \leq y \leq \frac{c-1}{c+d-1}} (a+b-1)kl\left(\frac{a-1}{a+b-1}; y\right) + (c+d-1)kl\left(\frac{c-1}{c+d-1}; y\right).$$

■

Proof of Lemma 36 Let $\overline{\mathcal{I}}$ be empty, then we have $\mu_{\infty,i} \triangleq \lim_{n\to\infty} \mu_{n,i} = \mu_i$. The posterior variance is

$$\begin{aligned}\sigma^2_{n,i} &= \frac{\alpha_{n,i}\beta_{n,i}}{(\alpha_{n,i}+\beta_{n,i})^2(\alpha_{n,i}+\beta_{n,i}+1)} \\ &= \frac{(1+\sum_{\ell=1}^{n-1}\mathbb{1}\{I_\ell = i\}Y_{\ell,I_\ell})(1+T_{n,i}-\sum_{\ell=1}^{n-1}\mathbb{1}\{I_\ell = i\}Y_{\ell,I_\ell})}{(2+T_{n,i})^2(2+T_{n,i}+1)},\end{aligned}$$

We see that when $\overline{\mathcal{I}}$ is empty, we have $\sigma^2_{\infty,i} \triangleq \lim_{n\to\infty} \sigma^2_{n,i} = 0$, i.e., the posterior is concentrated.

When $T_{n,i} = 0$ for some $i \in \mathcal{A}$, the empirical mean for that arm equals the prior mean

$$\mu_{n,i} = \alpha_{1,i}/(\alpha_{1,i}+\beta_{1,i}),$$

and the variance is strictly positive:

$$\sigma^2_{n,i} = (\alpha_{n,i}\beta_{n,i})/\left((\alpha_{1,i}+\beta_{1,i})^2(\alpha_{1,i}+\beta_{1,i}+1)\right) > 0.$$

When $\overline{\mathcal{I}}$ is not empty, then for every $i \in \overline{\mathcal{I}}$ we have $\sigma^2_{\infty,i} > 0$, and $\alpha_{\infty,i} \in (0,1)$, implying $\alpha_{\infty,I^\star} < 1$, hence the posterior is not concentrated.

■

Chapter 8

Discussion and future work

In this chapter I concisely review the previous six chapters of this dissertation, and explore some open challenges and possible directions for future work.

8.1 Forward-looking Bayesians

In Chapter 2 we studied the failure of weak truth-merger of Wenmackers and Romeijn's open-minded Bayesians, and we proposed two versions of forward-looking open-minded Bayesians that do weakly merge with the truth when the truth is added at some point in time. In Chapter 2 we only focus on *how* to incorporate new hypotheses. A direction for future research, possibly for me and my co-author on this chapter, is to formalise *when* new hypotheses should be considered, and to investigate how this interacts with the guarantee of truth-merger.

Chapter 2 inspired the following idea for a future project for myself in the area of continuous-armed best-arm identification in machine learning. This protocol can be viewed as similar to the protocol of the forward-looking Bayesians, if we let arms correspond to hypotheses, however, it is still unclear what the relation is between truth-merger and identification. The algorithms proposed in papers on best-arm identification in continuous-armed bandits (Bubeck, Munos and Stoltz, 2009; Carpentier and Valko, 2015; Aziz et al., 2018) employ two phases: First, a finite subset of arms from a continuous reservoir is selected, and subsequently a finite-armed bandit algorithm is run on this subset to identify the best arm. An interesting idea would be to propose an algorithm that decides during the learning process to add (or remove) arms from the finite set under consideration, which might lead to simple regret bounds scaling better in the confidence parameter δ in the fixed-confidence setting. Another future course would be to propose a Bayesian algorithm for best-arm identification in continuous-armed bandits, which can also be seen as an extension of the algorithms discussed in Chapter 7, see also the upcoming Section 8.4. This is both conceptually interesting because of the link with the forward-looking Bayesians and Bayesian confirmation theory, and also interesting because the Bayesian sampling rules of Chapter 7 do not depend on a confidence parameter or time

horizon. The combination of these two challenges is to propose a Bayesian algorithm for best-arm identification in continuous-armed bandits that adds or removes arms in course of the learning process. This algorithm could also provide some insights for the problem of *when* to add new hypotheses in the framework of the forward-looking Bayesians.

8.2 Hypothesis testing

Chapters 3 and 4 deal with the question whether Bayes factor hypothesis testing is robust under *optional stopping*. The bottom line of these chapters is that the answer to this question depends on one's perspective on Bayesianism (see also Section 1.2) and which definition of optional stopping one employs — we give three distinct mathematical definitions in Chapter 4. It is remarkable how resolutely some authors advocate the use of their favourite method for hypothesis testing, and how firm their reproach sometimes is to other authors who nuance or criticise claims about these methods, see for example (Benjamin et al., 2018) and (McShane et al., 2019); and even before being published, Chapter 3 provoked several responses (Rouder, 2019; Wagenmakers, Gronau and Vandekerckhove, 2019; Rouder and Haaf, n.d.) . In light of this fierce defence of some specific methods for hypothesis testing, an interesting project would be to investigate the role of hypothesis testing in the behavioural sciences. In a paper related to this subject, Gigerenzer and Marewski (2014) argue that "determining significance has become a surrogate for good research". The current discussion on optional stopping with Bayes factors that is the subject of Chapter 3 seems to be an example of that shift in focus from the actual goals of science to the surrogate of "mindless mechanical statistics". Goals of science include gaining knowledge about the world around us, and hypothesis testing is one of the means scientists have at their disposal to achieve that. How clear this distinction between goals and means is in current research in the behavioural sciences, and what the role of hypothesis testing in scientific research should be, are subjects to be addressed, possibly by philosophers of science.

In Chapter 5 we proposed a new theory for hypothesis testing based on E-values. From a practical perspective, it is now important to develop software for calculating E-values for common hypothesis tests, so that practitioners can start working with E-value based hypothesis tests. From a theoretical perspective, there are some open questions arising in particular from the combination of Chapter 4 and 5. The former chapter provides results showing that using the right Haar prior in general group invariant cases leads to E-values, however, in Chapter 5 is only shown that these are GROW E-values for the particular (important) case of the t-test. An objective for future work is thus to extend this to a general group-invariant setting. Further goals for future work on Safe Testing include the construction of confidence intervals by *inverting* a safe test. When this safe test constitutes a *test martingale*, these confidence intervals are *always valid confidence intervals* in the sense of Howard et al.'s 2018b framework of uniform, nonparametric, non-asymptotic confidence sequences (Darling and Robbins, 1967; Lai, 1984). The intuitions behind the construction of safe tests can lead to other constructions of confidence intervals. Further future objectives are to investigate the connections of safe testing to Shafer and Vovk's 2019 game-theoretic probability framework, and to the framework of always-valid P-values (Robbins, 1970; Robbins and Siegmund, 1972; Robbins and Siegmund, 1974; Johari, Pekelis and Walsh, 2015). The group of prof. Grünwald at CWI is working on these practical and theoretical challenges.

8.3 Safe-Bayesian generalised linear regression

Chapter 6 provides theoretical evidence that η-generalised Bayes can outperform standard Bayes for generalised linear models, and provides empirical evidence for Bayesian lasso and logistic regression. We also provided MCMC samplers for the generalised Bayesian lasso and logistic regression. The Gibbs sampler for the latter is based on a Pólya-Gamma latent variable scheme, in which the Pólya-Gamma random variable is approximated by a truncated sum of weighted Gamma random variables. Our current implementation is slow and unable to deal with high-dimensional data, presumably because of the approximation via the truncated sum. There exist another implementation of Bayesian logistic regression, in the programming language STAN (Carpenter et al., 2017), using No-U-Turn-Sampling (Hoffman and Gelman, 2014), which is an extension of Hamiltonian Monte Carlo (HMC) (Duane et al., 1987). An interesting direction for future work, possibly for a master's or PhD student, would be to develop HMC algorithms for η-generalised Bayesian methods. This could also lead to a better and possibly faster implementation of η-generalised Bayesian logistic regression.

An issue with generalised Bayesian methods is the dependency on the learning rate parameter η. Grünwald's 2012 Safe-Bayesian algorithm provably finds the appropriate η for bounded excess loss functions and likelihood ratio's, and experiments of Grünwald and Van Ommen (2017) and Chapter 6 indicate that SafeBayes performs excellently in the unbounded case as well, but theoretical guarantees still need to be established. Furthermore, a drawback of the Safe-Bayesian algorithm is that it is computationally very slow. Another future objective is to propose a faster algorithm for learning η, possibly based on cross-validation, naturally together with theoretical guarantees, e.g. that the data distribution satisfies the central condition at the learning rate η output by the algorithm.

Objectives for future work thus are:

- providing a better MCMC sampler for η-generalised logistic regression, possibly via Hamiltonian Monte Carlo,
- providing MCMC samplers for other η-generalised GLMs,
- providing guarantees on the Safe-Bayesian algorithm for the unbounded case,
- proposing a faster algorithm than SafeBayes for learning the appropriate learning rate η, together with
- providing theoretical guarantees for this algorithm.

8.4 Pure exploration

In Chapter 7 we studied two Bayesian sampling rules, `TTTS` and `T3C`, for best-arm identification (BAI) in the fixed confidence setting. We introduced the notion of asymptotic β-optimality and proved that `TTTS` and `T3C` are asymptotically β-optimal. This optimality notion has two drawbacks. First, in order to be optimal, we would need the unknown true optimal $\beta^\star = \arg\max_{\beta\in[0,1]} \Gamma^\star_\beta$. Secondly, the guarantees are asymptotic, whereas finite-time sample complexity bounds would be more practicable.

Evident objectives for my future work are:

- fixed-confidence guarantees with online tuning of β for TTTS and T3C,
- finite-time sample complexity bounds,
- an extension to continuous-armed bandit models (see Section 8.1 above), and
- fixed-budget guarantees.

Furthermore, Chapter 7 provides a piece of the puzzle of the following two bigger pictures.

Any-time sampling rules BAI has been studied in different frameworks: the fixed-budget setting, the fixed-confidence setting, which has been studied in Chapter 7, and the *any-time* BAI setting, introduced by Jun and Nowak, 2016. In the any-time setting, the sampling rule does not depend on the risk parameter or the budget. The first sampling rule for BAI that does not depend on the risk parameter is the *tracking rule* proposed by Garivier and Kaufmann (2016). The sampling rules studied in Chapter 7, TTTS and T3C, are also examples of any-time sampling rules. This sparks the question: does there exist a sampling rule that is, albeit with modifications depending on the setting and objective, optimal in all settings? Thompson sampling (TS) could be a possible candidate for this: vanilla TS for regret minimization, TTTS for fixed-confidence best-arm identification, and (see below), Murphy sampling for the minimum of means problem.

Pure-exploration objectives Pure exploration problems can have other objectives than finding the best arm. Naturally, different objectives require different sampling rules. However, an interesting avenue for future work is to investigate how the lower bounds and sampling rules for the different objectives and frameworks relate. Here are two pure-exploration problems with objectives different from BAI.

Kaufmann, Koolen and Garivier (2018) study a problem related to BAI: They consider the task of adaptively learning how the minimum mean of a finite set of arms compares to a given threshold. They provide a lower bound on the sample complexity in the fixed-confidence setting, and propose an algorithm inspired by TTTS, called *Murphy Sampling*. Murphy Sampling is, just as TTTS and T3C, an any-time sampling rule. An open problem is to find a fixed-budget lower bound and algorithm for this problem.

Antos, Grover and Szepesvári (2010) and Carpentier et al. (2011) study the problem of estimating the means of a finite number of arms in the fixed-budget setting uniformly well. The objective is to minimise the worst expected squared error loss of the arms, and the performance of the algorithm is measured by comparing its loss to that of the optimal allocation algorithm, that is, *regret*. This notion of regret is however not cumulative, and this problem is therefore more related to the pure-exploration setting than to the standard MAB framework. This is also reflected in the property that good strategies for this problem should play all arms linearly in the number of draws, whereas in the standard stochastic bandit setting suboptimal arms should be played logarithmically in the number of draws. The problem can be extended to learning the transition probabilities of Markov Chains (Talebi and Maillard, 2019). An open problem is to find problem-dependent lower bounds for this problem. Furthermore, the algorithms proposed in both papers depend on the budget and/or the confidence level. An interesting avenue for future work is to find a problem-dependent lower bound and to propose an any-time, possibly Thompson Sampling related sampling rule.

References

Y. Abbasi-Yadkori, D. Pál and C. Szepesvári (2012). "Online-to-confidence-set conversions and application to sparse stochastic bandits". In: *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AIStats)* (p. 245).

V. Amrhein, S. Greenland and B. McShane (2019). *Scientists rise up against statistical significance* (pp. 120, 156, 158).

S. Andersson (1982). "Distributions of Maximal Invariants Using Quotient Measures". In: *The Annals of Statistics* 10.3, pp. 955–961 (pp. 91, 108, 114).

A. Antos, V. Grover and C. Szepesvári (2010). "Active learning in heteroscedastic noise". In: *Theoretical Computer Science* 411.29-30, pp. 2712–2728 (p. 264).

J.B. Asendorpf et al. (2014). "Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation". In: *International Journal of Behavioral Development* 38.5, pp. 453–460 (p. 7).

J-Y. Audibert and S. Bubeck (2010). "Best arm identification in multi-armed bandits". In: *Proceedings of the 23rd Conference on Learning Theory (CoLT)* (pp. 16, 206).

P. Auer, N. Cesa-Bianchi and P. Fischer (2002). "Finite-time analysis of the multi-armed bandit problem". In: *Machine Learning Journal* 47.2-3, pp. 235–256 (p. 206).

M. Aziz et al. (2018). "Pure exploration in infinitely-armed bandit models with fixed-confidence". In: *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*. arXiv: 1803.04665 (p. 261).

A. Balsubramani and A. Ramdas (2016). "Sequential nonparametric testing with the law of the iterated logarithm". In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 42–51 (pp. 120, 152).

G.A. Barnard (1947). "Review of *Sequential Analysis* by Abraham Wald". In: *Journal of the American Statistical Association* 42.240 (pp. 93, 156).

G.A. Barnard (1949). "Statistical inference". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 11.2, pp. 115–149 (p. 93).

G.A. Barnard and D.R. Cox, eds. (1962). *The Foundations of Statistical Inference: A Discussion*. Methuen's Monographs on Applied Probability and Statistics. London: Methuen (p. 24).

O.E. Barndorff-Nielsen (1978). *Information and Exponential Families in Statistical Theory*. Chichester, UK: Wiley (pp. 147, 184, 193).

A. Barron, J. Rissanen and B. Yu (1998). "The Minimum Description Length principle in coding and modeling". In: *IEEE Transactions on Information Theory* 44.6. Special Commemorative Issue: Information Theory: 1948-1998, pp. 2743–2760 (pp. 92, 112, 154).

A.R. Barron and T.M. Cover (1991). "Minimum complexity density estimation". In: *Information Theory, IEEE Transactions on* 37.4, pp. 1034–1054 (p. 177).

P.L. Bartlett, O. Bousquet and S. Mendelson (2005). "Local Rademacher complexities". In: *The Annals of Statistics* 33.4, pp. 1497–1537 (pp. 179, 183, 195).

M.J. Bayarri et al. (2012). “Criteria for Bayesian model choice with application to variable selection”. In: *The Annals of statistics* 40.3, pp. 1550–1577 (pp. 91, 100, 110, 143).

M.J. Bayarri et al. (2016). “Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses”. In: *Journal of Mathematical Psychology* 72, pp. 90–103 (pp. 91, 99, 153).

G. Belot (2013). “Bayesian Orgulity”. In: *Philosophy of Science* 80.4, pp. 483–503 (p. 19).

D.J. Benjamin et al. (2018). “Redefine statistical significance”. In: *Nature Human Behaviour* 2.1, p. 6 (pp. 12, 120, 156, 262).

J.O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. revised and expanded 2nd. Springer Series in Statistics. New York: Springer-Verlag (pp. 63, 78, 133).

J.O. Berger (2003). “Could Fisher, Jeffreys and Neyman Have Agreed on Testing?” In: *Statistical Science* 18.1, pp. 1–12 (pp. 132, 153, 155).

J.O. Berger (2006). “The case for objective Bayesian analysis”. In: *Bayesian Analysis* 1.3, pp. 385–402 (pp. 5, 69, 82).

J.O. Berger, J.M. Bernardo, D. Sun et al. (2015). “Overall objective priors”. In: *Bayesian Analysis* 10.1, pp. 189–221 (p. 112).

J.O. Berger, L.D. Brown and R.L. Wolpert (1994). “A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing”. In: *Annals of Statistics* 22.4, pp. 1787–1807 (pp. 153, 155).

J.O. Berger, L.R. Pericchi and J.A. Varshavsky (1998). “Bayes factors and marginal distributions in invariant situations”. In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 307–321 (pp. 68, 70, 91, 107–110, 112, 123, 124, 142, 143).

J.O. Berger and T. Sellke (1987). “Testing a point null hypothesis: The irreconcilability of p values and evidence”. In: *Journal of the American statistical Association* 82.397, pp. 112–122 (p. 12).

J.O. Berger, D. Sun et al. (2008). “Objective priors for the bivariate normal model”. In: *The Annals of Statistics* 36.2, pp. 963–982 (p. 112).

J.O. Berger and R.L. Wolpert (1988). *The Likelihood Principle*. 2nd. Hayward, C.A.: Institute of Mathematical Statistics (pp. 61, 91, 94).

J.M. Bernardo (1979). “Reference posterior distributions for Bayesian inference”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2, pp. 113–128 (p. 5).

J.M. Bernardo (1996). “The concept of exchangeability and its applications”. In: *Far East Journal of Mathematical Sciences* 4, pp. 111–122 (p. 6).

J.M. Bernardo and A.F.M Smith (1994). *Bayesian Theory*. Chichester: Wiley (pp. 4, 63, 79).

D. Blackwell and L. Dubins (1962). “Merging of Opinion With Increasing Information”. In: *The Annals of Mathematical Statistics* 33, pp. 882–886 (pp. 10, 30).

S. Bubeck, R. Munos and G. Stoltz (2009). “Pure exploration in multi-armed bandits problems”. In: *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*. arXiv: 0802.2655 (pp. 205, 261).

S. Bubeck, R. Munos and G. Stoltz (2011). “Pure exploration in finitely-armed and continuous-armed bandits”. In: *Theoretical Computer Science* 412.19, pp. 1832–1852 (p. 206).

A.N. Burnetas and M.N. Katehakis (1996). “Optimal adaptive policies for sequential allocation problems”. In: *Advances in Applied Mathematics* 17.2, pp. 122–142 (p. 206).

G. de los Campos et al. (2009). “Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree”. In: *Genetics* 182, pp. 375–385 (p. 199).

O. Cappé et al. (2013). “Kullback-Leibler upper confidence bounds for optimal sequential allocation”. In: *Annals of Statistics* 41.3, pp. 1516–1541. arXiv: arXiv:1210.1136v4 (p. 206).

B. Carpenter et al. (2017). “Stan: A probabilistic programming language”. In: *Journal of statistical software* 76.1 (p. 263).

A. Carpentier and A. Locatelli (2016). "Tight (lower) bounds for the fixed budget best arm identification bandit problem". In: *Proceedings of the 29th Conference on Learning Theory (CoLT)*. arXiv: 1605.09004 (p. 206).

A. Carpentier and M. Valko (2015). "Simple regret for infinitely many armed bandits". In: *Proceedings of the 32nd International conference on Machine Learning (ICML)*, pp. 1133–1141. arXiv: 1505.04627 (p. 261).

A. Carpentier et al. (2011). "Upper-confidence-bound algorithms for active learning in multi-armed bandits". In: *International Conference on Algorithmic Learning Theory*. Springer, pp. 189–203 (p. 264).

D.C. Carslaw (2015). *The openair manual - open-source tools for analysing air pollution data. Manual for version 1.1-4*. King's College London (p. 190).

D.C. Carslaw and K. Ropkins (2012). "Openair - an R package for air quality data analysis". In: *Environmental Modelling & Software* 27-18, pp. 52–61 (p. 190).

C.M. Carvalho, N.G. Polson and J.G. Scott (2010). "The horseshoe estimator for sparse signals". In: *Biometrika* 97.2, pp. 465–480 (pp. 186, 187).

N. Cesa-Bianchi and G. Lugosi (2006). *Prediction, Learning and Games*. Cambridge, UK: Cambridge University Press (pp. 46, 183).

A. Chernov and V.G. Vovk (2009). "Prediction with expert evaluators' advice". In: *Proceedings ALT 2009*. Ed. by R. Gavaldà et al. Springer, pp. 8–22 (p. 46).

C.S. Chihara (1987). "Some Problems for Bayesian Confirmation Theory". In: *British Journal for the Philosophy of Science* 38.4, pp. 551–560 (p. 20).

J.B. Conway (2013). *A course in functional analysis*. Vol. 96. Springer Science & Business Media (p. 115).

T.M. Cover and J.A. Thomas (1991). *Elements of Information Theory*. First. New York: Wiley (pp. 133, 134, 136, 137, 154, 165, 171).

G. Cumming (2012). *Understanding the New Statistics: Effect Sizes, Confidence and Meta-Analysis*. Routledge (p. 158).

W. van Dam, R.D. Gill and P.D. Grunwald (2005). "The statistical strength of nonlocality proofs". In: *IEEE transactions on information theory* 51.8, pp. 2812–2835 (p. 152).

D.A. Darling and H. Robbins (1967). "Confidence sequences for mean, variance, and median". In: *Proceedings of the National Academy of Sciences of the United States of America* 58.1, p. 66 (pp. 152, 262).

S.C. Dass (1998). "Unified Bayesian and conditional frequentist testing procedures". PhD thesis. University of Michigan (p. 107).

S.C. Dass and J.O. Berger (2003). "Unified conditional frequentist and Bayesian testing of composite hypotheses". In: *Scandinavian Journal of Statistics* 30.1, pp. 193–210 (pp. 68, 70, 91, 99, 107, 143).

A.P. Dawid (1982). "The Well-Calibrated Bayesian". In: *Journal of the American Statistical Association* 77.379, pp. 605–611 (pp. 20, 272).

A.P. Dawid (1985). "The Impossibility of Inductive Inference. Comment on Oakes (1985)". In: *Journal of the American Statistical Association* 80.390, pp. 340–341 (p. 19).

A.P. Dawid (1997). "Prequential Analysis". In: *Encyclopedia of Statistical Sciences*. Ed. by S. Kotz, C.B. Read and D. Banks. Vol. 1 (Update). New York: Wiley-Interscience, pp. 464–470 (p. 154).

A.P. Dawid, M. Stone and J.V. Zidek (1973). "Marginalization paradoxes in Bayesian and structural inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 35.2, pp. 189–213 (pp. 91, 113).

A.P. Dawid and V.G. Vovk (1999). "Prequential Probability: Principles and Properties". In: *Bernoulli* 5.1, pp. 125–162 (p. 6).

A.P. Dempster (1975). *A Subjective Look at Robustness*. Tech. rep. Research Report S-33. Harvard University (p. 79).

A. Deng, J. Lu and S. Chen (2016). "Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing". In: *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. IEEE, pp. 243–252 (pp. 60, 64, 95).

J.L. Doob (1971). "What is a Martingale?" In: *The American Mathematical Monthly* 78.5, pp. 451–463 (p. 81).
S. Duane et al. (1987). "Hybrid monte carlo". In: *Physics letters B* 195.2, pp. 216–222 (p. 263).
J. Earman (1992). *Bayes or Bust?* Cambridge, MA: MIT Press (pp. 7, 19, 20, 25, 31, 35).
K. Easwaran (2011). "Bayesianism I: Introduction and arguments in favor". In: *Philosophy Compass* 6.5, pp. 312–320 (pp. 4, 6).
M.L. Eaton (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics and American Statistical Association (pp. 68, 91, 105, 107, 108, 114, 115, 142).
W. Edwards, H. Lindman and L.J. Savage (1963). "Bayesian Statistical Inference for Psychological Research". In: *Psychological Review* 70.3, pp. 193–242 (pp. 12, 60, 78, 80, 90, 93, 96, 97).
C. Elliot (2014). "Countable additivity in the philosophical foundations of probability". MA thesis (p. 6).
T. van Erven, P.D. Grünwald and S. de Rooij (2007). "Catching up Faster in Bayesian Model Selection and Model Averaging". In: *Advances in Neural Information Processing Systems*. Vol. 20 (p. 177).
T. van Erven and P. Harremoës (2014). "Rényi Divergence and Kullback-Leibler Divergence". In: *IEEE Transactions on Information Theory* 60.7, pp. 3797–3820 (p. 123).
T. van Erven et al. (2015). "Fast Rates in Statistical and Online Learning". In: *Journal of Machine Learning Research* 16, pp. 1793–1861 (pp. 182, 183, 185, 195).
E. Even-dar, S. Mannor and Y. Mansour (2003). "Action elimination and stopping conditions for reinforcement learning". In: *Proceedings of the 20th International Conference on Machine Learning (ICML)* (pp. 16, 206).
B. de Finetti (1937). "La prévision: ses lois logiques, ses sources subjectives". In: *Annales de l'Institut Henri Poincaré* 7, pp. 1–68 (pp. 5, 6, 68).
B. de Finetti (1974). *Theory of probability: a critical introductory treatment*. John Wiley & Sons (p. 4).
R.A. Fisher (1934). "Statistical methods for research workers". In: (p. 11).
R.A. Fisher (1955). "Statistical methods and scientific induction". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 17.1, pp. 69–78 (p. 11).
R.A. Fisher (1956). "Statistical methods and scientific inference". In: (p. 11).
D.A. Freedman (2003). "Notes on the Dutch Book Argument". In: *Lecture Notes, Department of Statistics, University of Berkley at Berkley, http://www. stat. berkeley. edu/˜ census/dutchdef. pdf* (p. 6).
Y. Freund et al. (1997). "Using and combining predictors that specialize". In: *Proceedings STOC 1997*. New York: ACM Press, pp. 334–343 (p. 46).
V. Gabillon, M. Ghavamzadeh and A. Lazaric (2012). "Best arm identification: A unified approach to fixed budget and fixed confidence". In: *Advances in Neural Information Processing Systems 25 (NIPS)* (pp. 206, 218).
P. Gács (2005). "Uniform test of algorithmic randomness over a general space". In: *Theoretical Computer Science* 341.1-3, pp. 91–137 (p. 152).
H. Gaifman and M. Snir (1982). "Probabilities Over Rich Languages, Testing and Randomness". In: *Journal of Symbolic Logic* 47.3, pp. 495–548 (p. 31).
A. Garivier and E. Kaufmann (2016). "Optimal best arm identification with fixed confidence". In: *Proceedings of the 29th Conference on Learning Theory (CoLT)*. arXiv: 1602.04589 (pp. 16, 206, 210, 212, 218, 264).
A.E. Gelfand and A.F.M. Smith (1990). "Sampling-based approaches to calculating marginal densities". In: *Journal of the American statistical association* 85.410, pp. 398–409 (p. 3).
A. Gelman (2008). "Objections to Bayesian statistics". In: *Bayesian Analysis* 3.3, pp. 445–449 (p. 7).
A. Gelman (2017). *Statistical Modeling, Causal Inference, and Social Science* (p. 69).
A. Gelman and C. Shalizi (2012). "Philosophy and the practice of Bayesian statistics". In: *British Journal of Mathematical and Statistical Psychology* (p. 6).
A. Gelman and C.R. Shalizi (2013). "Philosophy and the Practice of Bayesian Statistics". In: *British Journal of Mathematical and Statistical Psychology* 66, pp. 8–38 (p. 20).
A. Gelman et al. (2003). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press (p. 4).

J.K. Ghosh, M. Delampady and T. Samanta (2007). *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media (pp. 2, 4).

G. Gigerenzer (1993). "The superego, the ego, and the id in statistical reasoning". In: *A handbook for data analysis in the behavioral sciences: Methodological issues*, pp. 311–339 (p. 12).

G. Gigerenzer and J.N. Marewski (Jan. 2014). "Surrogate Science: The Idol of a Universal Method for Scientific Inference". In: *Journal of Management* 41, pp. 421–440 (pp. 11, 12, 59, 262).

G. Gigerenzer et al. (1990). *The empire of chance: How probability changed science and everyday life*. Vol. 12. Cambridge University Press (p. 12).

D.A. Gillies (2001). "Bayesianism and the Fixity of the Theoretical Framework". In: *Foundations of Bayesianism*. Ed. by D. Corfield and J. Williamson. Springer, pp. 363–379 (p. 20).

C. Glymour (1981). "Why I am not a Bayesian". In: *Theory and evidence*, pp. 63–93 (p. 7).

C. Glymour (2016). "Responses to contributions to special issue". In: *Synthese* 193.4, pp. 1251–1285 (p. 27).

I.J. Good (1971). "46656 varieties of Bayesians". In: *American Statistician* 25.5, p. 62 (p. 3).

I.J. Good (1991). "C383. A comment concerning optional stopping". In: *Journal of Statistical Computation and Simulation* 39.3, pp. 191–192 (pp. 60, 78, 80, 81, 96, 97).

P. Grünwald and T. Roos (2020). "Minimum Description Length Revisited". In: *International Journal of Mathematics for Industry* 11.1 (p. 154).

P.D. Grünwald (2000). "Maximum Entropy and the Glasses You are Looking Through". In: *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*. San Francisco: Morgan Kaufmann, pp. 238–246 (p. 157).

P.D. Grünwald (2007). *The Minimum Description Length Principle*. MIT Press (pp. 5, 77, 92, 112, 154, 184, 186, 193).

P.D. Grünwald (2012). "The Safe Bayesian: learning the learning rate via the mixability gap". In: *Proceedings 23rd International Conference on Algorithmic Learning Theory (ALT '12)*. Springer (pp. 180, 191, 263).

P.D. Grünwald (2013). "Safe Probability: restricted conditioning and extended marginalization". In: *Proceedings Twelfth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2013)*. Vol. 7958. Lecture Notes in Computer Science. Springer, pp. 242–252 (p. 82).

P.D. Grünwald (2016). "Toetsen als gokken: een redelijk alternatief voor de p-waarde". In: *Nieuw Archief voor Wiskunde* 5/17.4, pp. 236–244 (p. 12).

P.D. Grünwald (2018). "Safe Probability". In: *Journal of Statistical Planning and Inference* 195. See also arXiv preprint 1604.01785, pp. 47–63 (pp. 61, 82, 157).

P.D. Grünwald and A.P. Dawid (2004). "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory". In: *Annals of Statistics* 32.4, pp. 1367–1433 (pp. 124, 133, 166, 167).

P.D. Grünwald and J.Y. Halpern (July 2004). "When ignorance is bliss". In: *Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2004)*. Banff, Canada (p. 189).

P.D. Grünwald, R. de Heide and W.M. Koolen (2019). *Safe Testing*. arXiv preprint arXiv:1906.07801 (pp. 17, 83, 90, 99).

P.D. Grünwald and J. Langford (2007). "Suboptimal behavior of Bayes and MDL in classification under misspecification". In: *Machine Learning* 66.2-3, pp. 119–149 (pp. 177, 180).

P.D. Grünwald and N.A. Mehta (2019). "Fast Rates for General Unbounded Loss Functions: from ERM to Generalized Bayes". In: *Journal of Machine Learning Research*. Accepted pending minor revision (pp. 15, 157, 164, 178).

P.D. Grünwald and T. van Ommen (2017). "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it". In: *Bayesian Analysis* 12.4, pp. 1069–1103 (pp. 15, 177, 178, 180, 185, 187, 201, 263).

X. Gu, H. Hoijtink and J. Mulder (2016). "Error probability in default Bayesian hypothesis testing". In: *Journal of Mathematical Psychology* (p. 75).

E. Gunel and J. Dickey (1974). "Bayes factors for independence in contingency tables". In: *Biometrika* 61.3, pp. 545–557 (pp. 84, 85, 148, 157).
I. Hacking (2006). *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. 2nd. Cambridge University Press (p. 4).
A. Hájek (2019). "Interpretations of Probability". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University (p. 7).
R. de Heide (2016). *SafeBayes: Generalized and Safe-Bayesian Ridge and Lasso Regression*. R package version 1.1 (p. 187).
R. de Heide and P.D. Grünwald (2018). "Why optional stopping is a problem for Bayesians". In: *arXiv preprint arXiv:1708.08278* (pp. 17, 90, 94–96).
R. de Heide et al. (2020). "Safe-Bayesian Generalized Linear Regression". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2623–2633 (p. 17).
A.A. Hendriksen (2017). "Betting as an alternative to *p*-values". MA thesis. Leiden University, Dept. of Mathematics (pp. 68, 111).
A.A. Hendriksen, R. de Heide and P.D. Grünwald (2020). "Optional Stopping with Bayes Factors: a categorization and extension of folklore results, with an application to invariant situations". In: *Bayesian Analysis* (pp. 17, 60, 64, 68, 70, 81, 83, 143, 156, 157).
M.D. Hoffman and A. Gelman (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623 (p. 263).
C. Holmes and S. Walker (2017). "Assigning a value to a power likelihood in a general Bayesian model". In: *Biometrika* 104.2, pp. 497–503 (p. 177).
S.R. Howard et al. (2018a). "Exponential line-crossing inequalities". In: *arXiv preprint arXiv:1808.03204* (pp. 120, 131, 152, 158).
S.R. Howard et al. (2018b). "Uniform, nonparametric, non-asymptotic confidence sequences". In: *arXiv preprint arXiv:1810.08240* (pp. 99, 120, 131, 152, 158, 262).
C. Howson (1988). "On the Consistency of Jeffreys's Simplicity Postulate, and its Role in Bayesian Inference". In: *The Philosophical Quarterly* 38.150, pp. 68–83 (pp. 20, 27).
C. Howson (2000). *Hume's Problem*. New York: Oxford University Press (p. 19).
C. Howson and P. Urbach (2006). *Scientific reasoning: the Bayesian approach*. Open Court Publishing (p. 4).
Y. Hu (2020). "Safe Wilcoxon test (working title, thesis in preparation)". MA thesis. Leiden University, Mathematical Institute (p. 154).
R. Hubbard (2004). "Alphabet Soup: Blurring the Distinctions Between p's and a's in Psychological Research". In: *Theory & Psychology* 14.3, pp. 295–327 (pp. 11, 12).
C.J. Huberty (1993). "Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks". In: *The Journal of Experimental Education* 61.4, pp. 317–333 (p. 12).
S.M. Huttegger (2015). "Merging of Opinions and Probability Kinematics". In: *The Review of Symbolic Logic* 8.4, pp. 611–648 (pp. 9, 19).
K. Jamieson et al. (2014). "lil'UCB: An optimal exploration algorithm for multi-armed bandits". In: *Proceedings of the 27th Conference on Learning Theory (CoLT)* (p. 206).
T. Jamil et al. (2016). "Default "Gunel and Dickey" Bayes factors for contingency tables". In: *Behavior Research Methods*, pp. 1–15 (pp. 13, 59, 69, 78, 83, 84, 88, 90, 96, 148, 157).
E.T. Jaynes (1957). "Information Theory and Statistical Mechanics". In: *Physical Review* 108.2, pp. 171–190 (p. 5).
R.C. Jeffrey (1965). *The logic of decision*. University of Chicago Press (pp. 5, 6).
R.C. Jeffrey (1992). *Probability and the Art of Judgment*. Cambridge University Press (p. 4).
H. Jeffreys (1939). *Theory of probability*. 1st (p. 5).

H. Jeffreys (1946). "An invariant form for the prior probability in estimation problems". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186.1007, pp. 453–461 (p. 5).
H. Jeffreys (1948). *Theory of probability*. 2nd (p. 5).
H. Jeffreys (1961). *Theory of Probability*. Oxford, England: Oxford (pp. 4, 12, 20, 62, 69, 70, 90, 98, 141).
R. Johari, L. Pekelis and D.J. Walsh (2015). "Always valid inference: Bringing sequential analysis to A/B testing". In: *arXiv preprint arXiv:1512.04922* (pp. 120, 152, 262).
L.K. John, G. Loewenstein and D. Prelec (2012a). "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling". In: *Psychological science* (p. 89).
L.K. John, G. Loewenstein and D. Prelec (2012b). "Measuring the prevalence of questionable research practices with incentives for truth telling". In: *Psychological science* 23.5, pp. 524–532 (p. 82).
V.E. Johnson (2013a). "Revised standards for statistical evidence". In: *Proceedings of the National Academy of Sciences* 110.48, pp. 19313–19317 (p. 126).
V.E. Johnson (2013b). "Uniformly most powerful Bayesian tests". In: *Annals of statistics* 41.4, p. 1716 (pp. 124, 126, 138, 139).
M. Jordan (2010). *Lecture notes for Stat260 class on Bayesian Modeling and Inference*. retrieved from `http://www.cs.berkeley.edu/ jordan/courses/260-spring10/lectures/lecture10.pdf` (p. 78).
K-S. Jun and R. Nowak (2016). "Anytime exploration for multi-armed bandits using confidence information". In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)* (pp. 16, 206, 264).
J.B. Kadane, M.J. Schervish and T. Seidenfeld (1999). "Statistical implications of finitely additive probability". In: *Rethinking the Foundations of Statistics*, p. 211 (p. 6).
E. Kalai and E. Lehrer (1993). "Rational Learning Leads to Nash Equilibrium". In: *Econometrica* 61.5, pp. 1019–1045 (pp. 10, 30).
E. Kalai and E. Lehrer (1994). "Weak and strong merging of opinions". In: *Journal of Mathematical Economics* 23.1, pp. 73–86 (p. 30).
S. Kalyanakrishnan et al. (2012). "PAC subset selection in stochastic multi-armed bandits". In: *Proceedings of the 29th International Conference on Machine Learning (ICML)* (p. 206).
Z. Karnin, T. Koren and O. Somekh (2013). "Almost optimal exploration in multi-armed bandits". In: *Proceedings of the 30th International Conference on Machine Learning (ICML)* (p. 206).
R.E. Kass and A.E. Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795 (pp. 12, 62).
E. Kaufmann and A. Garivier (2017). "Learning the distribution with largest mean: two bandit frameworks". In: *ESAIM: Proceedings and Surveys* 60, pp. 114–131 (p. 206).
E. Kaufmann and W. Koolen (2018). "Mixture martingales revisited with applications to sequential tests and confidence intervals". In: *arXiv preprint arXiv:1811.11419* (pp. 213, 215).
E. Kaufmann, W.M. Koolen and A. Garivier (2018). "Sequential test for the lowest mean: From Thompson to Murphy sampling". In: *Advances in Neural Information Processing Systems*, pp. 6332–6342 (p. 264).
J.L. Kelly (1956). "A New Interpretation of Information Rate". In: *Bell System Technical Journal*, pp. 917–926 (p. 121).
J. Kiefer (1977). "Conditional Confidence Statements and Confidence Estimators". In: *Journal of the American Statistical Association* 72.360, pp. 789–808 (pp. 153, 157).
A.N. Kolmogorov (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer (p. 4).
W.M. Koolen, D. Adamskiy and M.K. Warmuth (2012). "Putting Bayes to Sleep". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS 2012*. Ed. by Fernando Pereira et al., pp. 135–143 (p. 46).
T.L. Lai (1976). "On confidence sequences". In: *The Annals of Statistics* 4.2, pp. 265–280 (pp. 99, 152, 158).

T.L. Lai (1984). "Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach". In: *Communications in Statistics-Theory and Methods* 13.19, pp. 2355–2368 (p. 262).

T.L. Lai (2009). "Martingales in Sequential Analysis and Time Series, 1945–1985". In: *Electronic Journal for History of Probability and Statistics* 5.1 (p. 152).

P.-S. Laplace (1814). "Essai philosophique sur les probabilités. 2e éd". In: *Paris: Mme Ve Courcier* (p. 3).

E. Lehrer and R. Smorodinsky (1996). "Merging and learning". In: *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*. Ed. by Thomas S. Ferguson, Lloyd S. Shapley and James B. MacQueen. Vol. 30. Lecture Notes - Monograph Series. Institute of Mathematical Statistics, pp. 147–168 (pp. 10, 31).

J. Leike (2016). "Nonparametric General Reinforcement Learning". PhD Dissertation. Australian National University (pp. 10, 31).

I. Levi (1980). *THe Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. Cambridge, MA: MIT Press (p. 19).

L.A. Levin (1976). "Uniform tests of randomness". In: *Soviet Mathematics Doklady* 17.2, pp. 337–340 (pp. 120, 152).

J.Q. Li (1999). "Estimation of Mixture Models". PhD thesis. New Haven, CT: Yale University (pp. 122, 135, 136, 164, 183).

L. Li et al. (2017). "Hyperband: Bandit-based configuration evaluation for hyperparameter optimization". In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)* (p. 219).

F. Liang et al. (2008). "Mixtures of g Priors for Bayesian Variable Selection". In: *Journal of the American Statistical Association* 103.481, pp. 410–423 (pp. 70, 75, 76, 112, 143).

D.V. Lindley (1957). "A statistical paradox". In: *Biometrika* 44.1/2, pp. 187–192 (pp. 60, 78–80, 90, 93).

D.V. Lindley (1982). "Comment on Dawid (1982)". In: *Journal of the American Statistical Association* 77.379, pp. 611–612 (p. 27).

D.V. Lindley (2000). "The Philosophy of Statistics". In: *Journal of the Royal Statistical Society* 49.3, pp. 293–337 (pp. 21, 27).

A. Ly, A.J. Verhagen and E-J. Wagenmakers (2016). "Harold Jeffreys' default Bayes factor hypothesis tests: Explanation, extension, and application in psychology". In: *Journal of Mathematical Psychology*, pp. 19–32 (pp. 13, 73).

R. Martin, R. Mess and S.G. Walker (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models". In: *Bernoulli* 23.3, pp. 1822–1847 (p. 183).

D. McAllester (1998). "Some PAC-Bayesian Theorems". In: *Proceedings of the Eleventh ACM Conference on Computational Learning Theory (COLT' 98)*. ACM Press, pp. 230–234 (p. 157).

P. McCullagh and J. Nelder (1989). *Generalized Linear Models*. Second. Boca Raton: Chapman and Hall/CRC (p. 185).

B.B. McShane et al. (2019). "Abandon Statistical Significance". In: *The American Statistician* 73.sup1, pp. 235–245 (pp. 158, 262).

P. Ménard (2019). "Gradient ascent for active exploration in bandit problems". In: *arXiv preprint arXiv:1905.08165*. arXiv: 1905.08165 (p. 219).

R. von Mises (1939). *Probability, statistics and truth*. Macmillan (p. 7).

R.D. Morey and J.N. Rouder (2015). *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.12-2 (pp. 75, 83).

J. Mourtada and O-A. Maillard (2017). "Efficient tracking of a growing number of experts". In: *Proceedings ALT 2017*. Ed. by S. Hanneke and L. Reyzin. Vol. 76. Proceedings of Machine Learning Research. PMLR, pp. 517–539 (p. 46).

U.K. Müller (2013). "Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix". In: *Econometrica* 81.5, pp. 1805–1849 (p. 177).

R. Narasimhan (June 2014). *Package weatherData, get Weather Data from the Web.* URL: http://ram-n.github.io/weatherData/ (p. 202).

J. Neyman and E.S. Pearson (1933). "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706, pp. 289–337 (p. 12).

D. Oakes (1985). "Self-Calibrating Priors Do Not Exist". In: *Journal of the American Statistical Association* 80.390, p. 339 (p. 267).

L. Pace and A. Salvan (2019). "Likelihood, Replicability and Robbins' Confidence Sequences". In: *International Statistical Review* (p. 152).

T. Park and G. Casella (2008). "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103, pp. 369–412 (pp. 180, 187, 197, 199).

S. van der Pas and P.D. Grünwald (2018). "Almost the Best of Three Worlds: Risk, Consistency and Optional Stopping for the Switch Criterion in Nested Model Selection". In: *Statistica Sinica* 1.18. see also arXiv preprint 1408.5724, pp. 229–255 (pp. 81, 92, 112).

S. van der Pas et al. (2016). "horseshoe: Implementation of the Horseshoe Prior". In: *R package version 0.1.0* (p. 187).

N.G. Polson, J.G. Scott and J. Windle (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables". In: *Journal of the American statistical Association* 108.504, pp. 1339–1349 (pp. 187, 188, 199, 200).

L. Pomatto, P. Strack and O. Tamuz (2020). "Stochastic dominance under independent noise". In: *Journal of Political Economy* 128.5 (p. 168).

E. Posner (1975). "Random coding strategies for minimum entropy". In: *IEEE Transactions on Information Theory* 21.4, pp. 388–391 (pp. 144, 175).

M.A. Proschan, K.K.G. Lan and J.T. Wittes (2006). *Statistical monitoring of clinical trials: a unified approach*. Springer Science & Business Media (p. 96).

H. Putnam (1963). "'Degree of Confirmation' and Inductive Logic". In: *The Philosophy of Rudolf Carnap*. Ed. by Paul A. Schilpp. LaSalle, IL: Open Court, pp. 761–783 (pp. 19, 20).

C. Qin, D. Klabjan and D. Russo (2017). "Improving the expected improvement algorithm". In: *Advances in Neural Information Processing Systems 30 (NIPS)*. arXiv: 1705.10033 (pp. 207, 210, 212–215, 222, 226, 227, 242, 244, 246, 248).

Eric Raidl (2020). "Open-Minded Orthodox Bayesianism by Epsilon-Conditionalization". In: *British Journal for the Philosophy of Science* 71.1, pp. 139–176 (p. 21).

H. Raiffa and R. Schlaifer (1961). *Applied Statistical Decision Theory*. Cambridge, MA: Harvard University Press (pp. 90, 94).

F.P. Ramsey (1926). "Truth and probability. Reprinted in". In: *Studies in subjective probability*, pp. 61–92 (pp. 4, 6).

J. Rissanen (1989). *Stochastic Complexity in Statistical Inquiry*. Hackensack, N.J.: World Scientific (p. 154).

H. Robbins (1970). "Statistical methods related to the law of the iterated logarithm". In: *The Annals of Mathematical Statistics* 41.5, pp. 1397–1409 (pp. 143, 152, 262).

H. Robbins and D. Siegmund (1972). "A class of stopping rules for testing parametric hypotheses". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, CA, 1970/1971)*. Vol. 4, pp. 37–41 (p. 262).

H. Robbins and D. Siegmund (1974). "The expected sample size of some tests of power one". In: *The Annals of Statistics*, pp. 415–436 (p. 262).

C.P. Robert (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media (p. 6).

J-W. Romeijn (2004). "Hypotheses and Inductive Predictions". In: *Synthese* 141.3, pp. 333–364 (pp. 19, 46).

J-W. Romeijn (2005a). "Bayesian inductive logic". PhD Dissertation. University of Groningen (p. 4).

J-W. Romeijn (2005b). "Theory Change and Bayesian Statistical Inference". In: *Philosophy of Science* 72.5, pp. 1174–1186 (p. 20).
J-W. Romeijn (2017). "Philosophy of Statistics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University (p. 6).
J.N. Rouder (2014). "Optional stopping: No problem for Bayesians". In: *Psychonomic Bulletin & Review* 21.2, pp. 301–308 (pp. 13, 59–63, 65, 82, 83, 90, 94, 97).
J.N. Rouder (2019). "On The Interpretation of Bayes Factors: A Reply to de Heide and Grunwald". In: (p. 262).
J.N. Rouder and J.M. Haaf (n.d.). "Optional Stopping and the Interpretation of The Bayes Factor". In: () (p. 262).
J.N. Rouder and R.D. Morey (2012). "Default Bayes Factors for Model Selection in Regression". In: *Multivariate Behavioral Research* 47.6, pp. 877–903 (pp. 75, 83).
J.N. Rouder et al. (2009). "Bayesian t tests for accepting and rejecting the null hypothesis". In: *Psychonomic Bulletin & Review* 16.2, pp. 225–237 (pp. 13, 59, 66, 69–71, 83, 90, 97, 98, 108, 112, 141).
J.N. Rouder et al. (2012). "Default Bayes factors for ANOVA designs". In: *Journal of Mathematical Psychology* 56.5, pp. 356–374 (pp. 59, 69, 79, 83, 90, 112).
R. Royall (1997). *Statistical evidence: a likelihood paradigm*. Chapman and Hall (pp. 122, 147).
D. Russo (2016). "Simple Bayesian algorithms for best arm identification". In: *Proceedings of the 29th Conference on Learning Theory (CoLT)*. arXiv: 1602.08448 (pp. 16, 205–207, 210–212, 218, 219, 222, 247, 249, 256, 257).
B.Y. Ryabko and V.A. Monarev (2005). "Using information theory approach to randomness testing". In: *Journal of Statistical Planning and Inference* 133.1, pp. 95–110 (p. 154).
A.N. Sanborn and T.T. Hills (2014). "The frequentist implications of optional stopping on Bayesian hypothesis tests". In: *Psychonomic Bulletin & Review* 21.2, pp. 283–300 (pp. 60, 64, 81, 82, 90).
L.J. Savage (1954). *The Foundations of Statistics*. New York: John Wiley and Sons (pp. 4, 5, 60, 68, 78–80).
F.D. Schönbrodt et al. (2017). "Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences". In: *Psychological Methods* 22.2, pp. 322–339 (pp. 81, 90).
J. ter Schure and P. Grünwald (2019). "Accumulation Bias in meta-analysis: the need to consider time in error control". In: *F1000Research* 8 (p. 120).
T. Seidenfeld (1979). "Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz". In: *Theory and Decision* 11.4, pp. 413–440 (pp. 5, 61).
T. Seidenfeld (2016). *Personal Communication* (p. 147).
T. Seidenfeld and M.J. Schervish (1983). "A conflict between finite additivity and avoiding Dutch book". In: *Philosophy of Science* 50.3, pp. 398–412 (p. 6).
T. Sellke, M.J. Bayarri and J.O. Berger (2001). "Calibration of ρ values for testing precise null hypotheses". In: *The American Statistician* 55.1, pp. 62–71 (p. 153).
G. Shafer (2019). *The Language of Betting as a Strategy for Statistical and Scientific Communication*. working paper, available at `http://probabilityandfinance.com/articles/index.html` (p. 155).
G. Shafer and V. Vovk (2001). *Probability and Finance – It's Only a Game!* New York: Wiley (p. 155).
G. Shafer and V. Vovk (2019). *Game-Theoretic Probability: Theory and Applications to Prediction, Science and Finance*. Wiley (pp. 120, 152, 153, 155, 162, 262).
G. Shafer et al. (Feb. 2011). "Test Martingales, Bayes Factors and p-Values". In: *Statistical Science* 26.1, pp. 84–101 (pp. 92, 120, 152).
S. Shalev-Shwartz and S. Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press (p. 181).
X. Shang, E. Kaufmann and M. Valko (2019). "A simple dynamic bandit algorithm for hyper-parameter tuning". In: *6th Workshop on Automated Machine Learning at International Conference on Machine Learning (ICML-AutoML)* (p. 219).

X. Shang et al. (2020). "Fixed-confidence guarantees for Bayesian best-arm identification". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1823–1832 (p. 17).

J. Shawe-Taylor and R.C. Williamson (1997). "A PAC Analysis of a Bayesian classifier". In: *Proceedings of the Tenth ACM Conference on Computational Learning Theory (COLT' 97)*. Nashville, T., pp. 2–9 (p. 157).

A. Shimony (1969). *Letter to Rudolf Carnap, September 20*. Rudolf Carnap Papers, 1905-1970, ASP.1974.01, Series XII. Notes and Correspondence: Probability, Mathematics, Publishers, UCLA Administrative, and Lecture Notes, 1927-1970, Subseries 1: Probability Authors, Box 84b, Folder 55, Special Collections Department, University of Pittsburgh. (p. 20).

A. Shimony (1970). "Scientific inference". In: *The Nature and Function of Scientific Theories*. Ed. by Robert G. Colodny. University of Pittsburgh Press, pp. 79–172 (pp. 20, 24, 27).

J. Sprenger and S. Hartmann (2019). *Bayesian Philosophy of Science*. Oxford University Press (p. 27).

Jan Sprenger (2020). "Conditional Degree of Belief and Bayesian Inference". In: *Philosophy of Science* 87.2, pp. 319–335 (p. 27).

T.F. Sterkenburg (2019). "Putnam's Diagonal Argument and the Impossibility of a Universal Learning Machine". In: *Erkenntnis* 84.3, pp. 633–656 (p. 19).

T.F. Sterkenburg and R. de Heide (2019). "On the truth-convergence of open-minded Bayesianism". In: *tbk* (p. 17).

D. Sun and J.O. Berger (2007). "Objective Bayesian analysis for the multivariate normal model". In: *Bayesian Statistics* 8, pp. 525–562 (p. 143).

N. Syring and R. Martin (2017). "Calibrating general posterior credible regions". In: *arXiv preprint arXiv:1509.00922* (p. 177).

M.S. Talebi and O-A. Maillard (2019). "Learning Multiple Markov Chains via Adaptive Allocation". In: *Advances in Neural Information Processing Systems*, pp. 13322–13332 (p. 264).

J.N. Tendeiro and H.A.L. Kiers (2019). "A review of issues about null hypothesis Bayesian testing." In: *Psychological methods* (p. 60).

W.R. Thompson (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4, p. 285 (p. 206).

R. Tibshirani (1996). "Regression shrinkage and selection via the lasso". In: *J. Roy. Statist. Soc. Ser. B* 58.1, pp. 267–288 (p. 187).

F. Topsøe (1979). "Information-Theoretical Optimization Techniques". In: *Kybernetika* 15.1, pp. 8–27 (p. 166).

A.B. Tsybakov (2004). "Optimal aggregation of classifiers in statistical learning". In: *Annals of Statistics* 32, pp. 135–166 (p. 183).

J.W. Tukey (1980). "We need both exploratory and confirmatory". In: *The American Statistician* 34.1, pp. 23–25 (p. 2).

R. Turner (2019). "Safe tests for 2×2 contingency tables and the Cochran-Mantel-Haenszel test". MA thesis. Leiden University, Mathematical Institute (p. 145).

Olav Benjamin Vassend (2019). "New Semantics for Bayesian Inference: The Interpretive Problem and Its Solutions". In: *Philosophy of Science* 86.4, pp. 696–718 (p. 27).

J-A. Ville (1939). *Étude critique de la notion de collectif*. Monographies des probabilités. Paris: Gauthier-Villars (p. 152).

V. Vovk and R. Wang (2019). "Combining e-values and p-values". In: *Available at SSRN* (pp. 120, 152).

V.G. Vovk (1993). "A logic of probability, with application to the foundations of statistics". In: *Journal of the Royal Statistical Society, series B* 55. (with discussion), pp. 317–351 (p. 153).

V. Vovk et al. (2011). "Test Martingales, Bayes Factors and p-Values". In: *Statistical Science* 26.1, pp. 84–101 (p. 81).

E-J. Wagenmakers (2007). "A practical solution to the pervasive problems of p-values". In: *Psychonomic Bulletin & Review* 14.5, pp. 779–804 (pp. 12, 61, 62, 80, 90).

E-J. Wagenmakers et al. (2012). "An Agenda for Purely Confirmatory Research". In: *Perspectives on Psychological Science* 7, pp. 627–633 (p. 59).

E.-J. Wagenmakers, Q.F. Gronau and J. Vandekerckhove (2019). "Five Bayesian Intuitions for the Stopping Rule Principle". In: (p. 262).

A. Wald (1947). "An essentially complete class of admissible decision functions". In: *The Annals of Mathematical Statistics*, pp. 549–555 (p. 6).

S. Walker and N.L. Hjort (2002). "On Bayesian consistency". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.4, pp. 811–821 (pp. 178, 183).

P. Walley (1991). *Statistical Reasoning with Imprecise Probabilities*. Vol. 42. Monographs on Statistics and Applied Probability. London: Chapman and Hall (p. 133).

R.L. Wasserstein, N.A. Lazar et al. (2016). "The ASAs statement on p-values: context, process, and purpose". In: *The American Statistician* 70.2, pp. 129–133 (pp. 12, 120, 156).

S. Wenmackers and J-W. Romeijn (2016). "New theory about old evidence: A framework for open-minded Bayesianism". In: *Synthese* 193.4, pp. 1225–1250 (pp. 10, 20, 24, 26–28, 35, 261).

R.A. Wijsman (1990). *Invariant measures on groups and their use in statistics*. Institute of Mathematical Statistics (pp. 91, 114).

J. Williamson (1999). "Countable additivity and subjective probability". In: *The British Journal for the Philosophy of Science* 50.3, pp. 401–416 (p. 6).

J. Windle, N.G. Polson and J.G. Scott (2014). "Sampling Pólya-gamma random variates: alternate and approximate techniques". In: *arXiv preprint arXiv:1405.0506* (pp. 200, 201).

R.L. Wolpert (1996). "Testing Simple Hypotheses". In: *Data Analysis and Information Systems: Statistical and Conceptual Approaches*. Ed. by H.H. Bock and W. Polasek. Berlin: Springer, pp. 289–297 (p. 153).

K. Yamanishi (1998). "A Decision-Theoretic Extension of Stochastic Complexity and its Applications to Learning". In: *IEEE Transactions on Information Theory* 44.4, pp. 1424–1439 (p. 183).

Y. Yao et al. (2018). "Using Stacking to Average Bayesian Predictive Distributions". In: *Bayesian Analysis*. Advance publication; Number and pages to be announced (p. 177).

E.C. Yu et al. (2014). "When decision heuristics and science collide". In: *Psychonomic Bulletin & Review* 21.2, pp. 268–282 (pp. 60, 64, 90).

A. Zellner and A. Siow (Feb. 1980). "Posterior odds ratios for selected regression hypotheses". In: *Trabajos de Estadistica Y de Investigacion Operativa* 31.1, pp. 585–603 (p. 75).

T. Zhang (2006a). "From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation". In: *The Annals of Statistics* 34.5, pp. 2180–2210 (pp. 178, 183).

T. Zhang (2006b). "Information-theoretic upper and lower bounds for statistical estimation". In: *IEEE Transactions on Information Theory* 52.4, pp. 1307–1321 (pp. 178, 183, 184).

Y. Zhang (2013). "Analysis of tests of local realism". PhD thesis. Boulder, CO: University of Colorado (p. 152).

Y. Zhang, S. Glancy and E. Knill (2011). "Asymptotically optimal data analysis for rejecting local realism". In: *Physical Review A* 84.6, p. 062118 (pp. 120, 152).

Alphabetical Index

Samenvatting

Dit proefschrift gaat over het leren van data op een Bayesiaanse wijze. *Statistiek* en *machine learning* gaan over de vraag hoe mensen en computers kunnen leren van data. Bayesiaanse methodes worden in deze vakgebieden veel gebruikt, echter, ze hebben bepaalde beperkingen en interpretatieproblemen die niet altijd worden onderkend. In twee hoofdstukken van dit proefschrift onderzoeken we een dergelijke beperking en omzeilen we deze door een verruiming van het standaardkader van de Bayesiaanse methode. In twee andere hoofdstukken nemen we door hoe verschillende filosofische interpretaties van het Bayesianisme wiskundige definities en stellingen beïnvloeden, en hoe dat zijn uitwerking heeft op de praktische toepassing van Bayesiaanse methodes. In de overige twee hoofdstukken passen we zelf Bayesiaanse methodes toe op een *pragmatische* wijze: enkel als *werktuig* voor een interessant statistisch probleem, een probleem dat ook op een niet-Bayesiaanse manier had kunnen worden aangepakt.

Leren Als een onderzoeker iets wil leren over een onbekend proces, vindt er een interactie plaats tussen haar en de *data* die door het proces zijn voortgebracht. De taak van de onderzoeker is *inductie*: een manier van redeneren waarbij er op grond van waarnemingen tot een algemene regel — een generalisatie — wordt gekomen. De onderzoeker begint met enkele veronderstellingen over het onbekende proces, omdat zonder deze *voorkennis* de datapunten op iedere mogelijke manier zouden kunnen samenhangen en het onmogelijk is tot een generalisatie te komen. Daarnaast bestaat er een verzameling van *hypotheses* die de onderzoeker kan opstellen of onderzoeken: algemene beschrijvingen van het onbekende proces. In de context van dit proefschrift, statistiek en machine learning, beschouwen we hypotheses die kunnen worden uitgedrukt als een waarschijnlijkheidsverdeling over een uitkomstenruimte, en deze noemen we *statistische hypotheses*. Een verzameling statistische hypotheses vormt een (statistisch) *model*. Een model is een wiskundige weergave van de voorkennis.

Bayesianisme Naast een model en de data hebben we een laatste ingrediënt nodig voor inductie: een methode. Het hoofdthema van dit proefschrift is de *Bayesiaanse* methode. In essentie is dit een methode die niet alleen waarschijnlijkheidsverdelingen over de data hanteert, maar ook over de statistische hypotheses. De onderzoeker begint met het specificeren van een *prior*, een waarschijnlijkheidsverdeling die haar onzekerheid over de statistische hypotheses uitdrukt, voordat ze heeft kennis genomen van de data. Na waarneming van de data, wordt met *de stelling van Bayes* een *posterior* berekend: een conditionele waarschijnlijkheidsverdeling

over de statistische hypotheses gegeven de data.

Bayesianisme is een term die verwijst naar een verzameling aan deze methode verwante ideeën in verschillende takken van de wetenschap. Echter, *het* Bayesianisme bestaat niet: er zijn verschillende stromingen, die er bijvoorbeeld verschillende theoriën op na houden over hoe de priors tot stand komen. Twee noemenswaardige, invloedrijke stromingen zijn het *subjectivisme* en het *objectivisme*. In de tweede helft van dit proefschrift staat een derde stroming centraal: het *pragmatisme*: onderzoekers die de Bayesiaanse methode niet uit filosofische overtuigingen bezigen, maar enkel vanwege haar nuttige eigenschappen of andere praktische beweegredenen. Discussies over de grondslagen van het Bayesianisme worden vaak in de filosofie gevoerd; welke stroming men aanhangt heeft nochtans consequenties voor de (statistische) praktijk: welke priors men kiest, welke wiskundige definities men formuleert en welke stellingen men poneert, hangt hier vanaf.

Misspecificatie van het model Zoals hierboven beschreven, begint de onderzoeker met het specificeren van een model en het toekennen van priorwaarschijnlijkheden aan zijn elementen. Als het ware datagenererende proces onderdeel is van het model, en niet uitgesloten wordt door de prior, is *consistentie* gegarandeerd: naar mate we meer en meer data verkrijgen, valt de onderzoekers posterior meer en meer samen met de ware verdeling. Niettemin kan het voorkomen dat het model *gemisspecificeerd* is: het ware datagenererende proces is geen onderdeel van het model (of heeft prior nul toegekend gekregen). Dit kan op verschillende manieren problematisch zijn en in dit proefschrift wordt de Bayesiaanse methode op twee verschillende manieren uitgebreid om twee van deze problemen te boven te komen.

Ten eerste kan het gebeuren dat de onderzoeker tijdens het leerproces een nieuwe hypothese bedenkt en deze wil toevoegen aan het model. In het standaardkader van de Bayesiaanse methode is dit in principe niet mogelijk: de onderzoeker moet de reeds verkregen data weggooien en opnieuw beginnen met het toekennen van priorwaarschijnlijkheden aan de elementen van het nieuwe, grotere model. In hoofdstuk 2 bestuderen we een *ruimdenkende Bayesiaanse logica*, die het dynamisch bijvoegen van nieuwe hypotheses tijdens het leerproces mogelijk maakt.

Ten tweede kan het gebeuren dat we willen dat de Bayesiaanse posterior samenvalt met het *beste* element in het model, in plaats van met de ware verdeling die buiten het model ligt. In hoofdstuk 6 laten we zien hoe dit kan mislukken met de Bayesiaanse standaardmethode. Vervolgens verrichten we een aanpassing aan de stelling van Bayes: de aannemelijkheidsverdeling wordt tot een macht verheven, en we noemen dit de *gegeneraliseerde Bayesiaanse methode*. Indien deze macht gevoeglijk wordt gekozen, lost dit het probleem op, en valt de gegeneraliseerde Bayesiaanse posterior na vergaring van data samen met het beste element in het model, ondanks de modelmisspecificatie.

Optioneel stoppen met de Bayes-factor-hypothesetoets De *Bayes factor* is een Bayesiaanse methodiek voor hypothesetoetsen. In hoofdstuk 3 en 4 bestuderen we *optioneel stoppen*. Informeel betekent dit 'tijdens het leerproces naar de tussenresultaten kijken om te beslissen of er meer datapunten vergaard moeten worden'. Verschillende auteurs beweren dat Bayesiaanse methodes *bestand zijn tegen optioneel stoppen*, maar het blijkt onduidelijk te zijn wat dat precies betekent. In hoofdstuk 4 geven we drie verschillende wiskundige definities van deze uitspraak.

In hoofdstuk 3 en 4 zetten we uiteen hoe het aanhangen van een van de stromingen van het Bayesianisme invloed heeft op welke beweringen men kan doen in de praktijk. In hoofdstuk 3 laten wij bijvoorbeeld zien dat sommige beweringen over optioneel stoppen met Bayes factors alleen betekenis hebben als ze vanuit een puur subjectieve invalshoek worden gedaan, desalniettemin worden deze beweringen vaak gedaan als zouden ze ook gelden voor een pragmatisch Bayesiaanse benadering.

Een nieuwe theorie voor hypothesetoetsen In hoofdstuk 5 presenteren we een nieuwe theorie voor hypothesetoetsen. Deze theorie draait om het concept genaamd 'E-variabele' of 'E-waarde', een stochast die de mate van bewijs tegen de nulhypothese aanduidt en die in de toekomst hopelijk de P-waarde zal vervangen in de toegepaste statistiek. Tevens introduceren we een optimaliteitscriterium voor de constructie van E-variabelen, genaamd *GROW*, wat een acroniem is voor het Engelse *Growth-Optimal in Worst Case*. Het blijkt dat de GROW E-variabele een Bayesiaanse interpretatie kent, zij het met een geheel ander soort priors dan priors die in de huidige Bayesiaanse praktijk worden gebruikt.

Identificatie van de beste waarschijnlijkheidsverdeling In hoofdstuk 7 bestuderen we een Bayesiaanse manier om uit een verzameling waarschijnlijkheidsverdelingen degene met de hoogste verwachtingswaarde te onderscheiden. We kunnen aan de verschillende verdelingen, die ook wel *armen* worden genoemd, een prior toekennen die de waarschijnlijkheid uitdrukt dat deze verdeling de hoogste verwachtingswaarde heeft. Vervolgens stellen we een regel op om op ieder tijdstip een arm te kiezen waarvan we een observatie willen ontvangen. Nadien berekenen we de posteriorwaarschijnlijkheid dat deze arm de hoogste verwachtingswaarde heeft. In hoofdstuk 7 bewijzen we asymptotische frequentistische garanties voor deze Bayesiaanse strategie.

Acknowledgements

The time has come to thank all the people without whom this dissertation would not even exist, and who made the past few years an unforgettable experience. I hope that these words are not the only way they learn how grateful I am to them.

First and foremost, I want to thank Jacqueline for creating the opportunity to pursue a PhD at Leiden University and CWI, for always supporting me, for creating the best working environment for me, and for giving me complete freedom to pursue my own interests. You are the best boss I could possibly wish for.

I am extremely grateful to my two other (co-)promotores and supervisors Peter and Wouter for their guidance, inspiration, and support. I learnt a lot from both of them and they gave me the freedom to follow my various interests with them and with others. I am especially grateful to Peter for forcing me to be absolutely rigorous in my mathematical arguments. I was very lucky to have you both as supervisors.

The papers on which this thesis is based would not have been possible without my co-authors Peter, Wouter, Allard, Tom, Alice, Nishant, Emilie, Michal, Pierre and Xuedong. I hope to have the pleasure of writing papers with you again some time in the future.

Most of this thesis was written in Amsterdam at the Centrum Wiskunde & Informatica, the best working environment one can imagine. Thanks to everyone who made and makes this a great place by having a chat, playing foosball or ping pong, going for runs, and so on. I especially want to mention CWI's fantastic library, which I made much use of. In Leiden I had a lot of fun with colleagues and friends Maarten and Sanne. Muriel and Sanne: thank you for being my paranymphs and guiding me through the formalities of the PhD defense.

I want to thank Tom for introducing me to the discipline of mathematical philosophy. Not only did you teach me a great deal and pointed me in the right directions for further reading material, and did I learn much from writing the paper together, you also invited me to visit the MCMP in Munich, and we went to present our paper at the FEW in Turin, where you introduced me to many more friendly philosophers.

In spring 2019 I had a wonderful time at SequeL in Lille, France, and I'd like to thank, in particular, Emilie, Michal, Sadegh, Odalric, Omar, Pierre and Xuedong for the great atmosphere both during and outside working hours, for inviting me to dinners and parties, for singing concerts together, for teaching me about important French (and Alsacian) culture (apéro, extensive lunch, good food in general) and for making me really feel at home, though I must say my French is still not at a conversational level. I will surely return to Lille often in the future.

I also enjoyed the research visits to Alexandra in Magdeburg and Csaba in Edmonton: thank you for having me, and I look forward to collaborating in the future. The time in trains and planes was also well spent on this dissertation.

Of course there was plenty of time for distractions during the past years as well: board games with Wouter, Sirée, Muriel, Peter P., Rosanne; sports with Ruud, Muriel, Wouter, Sanne, Maarten (running); Julia, Peter-Ben, Loek (ice-skating); Ruud (plenty of chess); Yvonne, Sem, Martijn tV., Alice (tennis, squash); great cycling adventures with Wouter; and many hiking adventures with Peter P. There are friends, across the world and close by, always in for a cup of tea or online chat: Jolien, Linda, Julia, Sanja, Metta and notably: Lineke.

But the most important distraction (and I still can't imagine having to leave you when I will move abroad for a postdoc): singing with the VU-Kamerkoor. Besides performing good music with one of the best conductors on this planet, Krista, you are all amazing people, and some of you are my best friends. I'm really going to miss having dinner every week with Roel, Jasper, Stefan and varying others, travelling with *club Leiden*: Ruud, Stefan, Jasper, Ellen, (and former Leienaar Saskia); my low-alto friends and neighbours Willemien and Chantal — who also made the beautiful cover of this dissertation; I will miss our good conversations Peter-Ben, Martijn S. and Eveline; the wonderful bassi profundi behind me (Ricus, Roel, Martijn, André); Gerben's cheerfulness... and finally yet importantly: Liza and Yvonne. Your friendship is priceless, and this dissertation would not have been completed without you. Thanks a lot.

And *tak skal du have*, Peter P., for keeping up with me, with my ever-changing plans, my wild adventures, my enthusiastically telling you about everything I'm interested in at any given moment of the day, but also for being fine with me moving abroad for months or years. It will be *lekker rustig* without me, but I'll make up for it twice over when we visit each other!

Rianne de Heide
Leiden, August 2020.

Curriculum Vitae

The author of this dissertation was born in 1989 in Rotterdam, The Netherlands. After completing VWO at Piter Jelles Gymnasium in Leeuwarden (2007), she obtained a bachelor's degree in classical music (horn) from the Prins Claus Conservatoire in Groningen (2012), a bachelor's degree in mathematics from the University of Groningen (2013), a master's degree in classical music (horn) from the Royal Conservatoire in The Hague (2014), and a master's degree in mathematics (Statistical Science, *cum laude*) from Leiden University (2016). The research presented in this dissertation was performed while working as a PhD candidate for 0.8 fte from July 2016 until January 2019, and and full-time from January 2019 until April 2020. The author was working as a member of the coordinating committee of the M.Sc. programme Statistical Science for 0.2 fte from July 2016 until January 2019. The author was employed by prof.dr. Jacqueline Meulman at the mathematical institute of Leiden University, and the research was performed at the Centrum Wiskunde & Informatica in Amsterdam, under the supervision of prof. dr. Peter Grünwald and dr. Wouter Koolen, except for Chapter 7, for which the research was performed during the author's stay at Inria Lille (France) from April 23 until August 3, 2019.