Bayesian learning: Challenges, limitations and pragmatics

Heide, R. de

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 $Z^t \coloneqq Z_1, \ldots, Z_t$ 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 $z^t = z_1, \dots, z_t$, 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

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

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 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 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, $P_{\mathrm{new}}(S) = P_{\mathrm{old}}(S|E)$. 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 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 $\nu_j, j = 1, \ldots, K$ 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 $X^n \coloneqq X_1, \dots, X_n$ 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^τ 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

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

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

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

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

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

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 $\ell'_f(x,y) \coloneqq -\log p_f(y|x)$. 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

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

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 $\bar{\eta}$ in the neighbourhood of f^* under the *$\bar{\eta}$-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 $\mu^\star := \max_{i \in [K]} \mu_i$, 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 recommendation 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.