Bayesian learning: Challenges, limitations and pragmatics

Heide, R. de

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

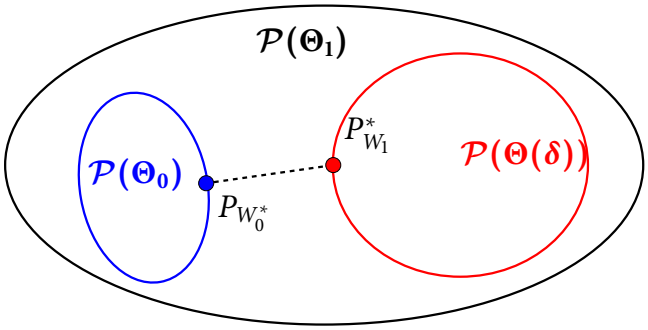

Figure 5.1: The Joint Information Projection (JIPr), with notation from Section 5.3. $\Theta_0 \subset \Theta_1$ represent two nested models, $\Theta(\delta)$ is a restricted subset of Θ_1 that does not overlap with Θ_0. $\mathcal{P}(\Theta) = \{P_W : W \in \mathcal{W}(\Theta)\}$, and $\mathcal{W}(\Theta)$ 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

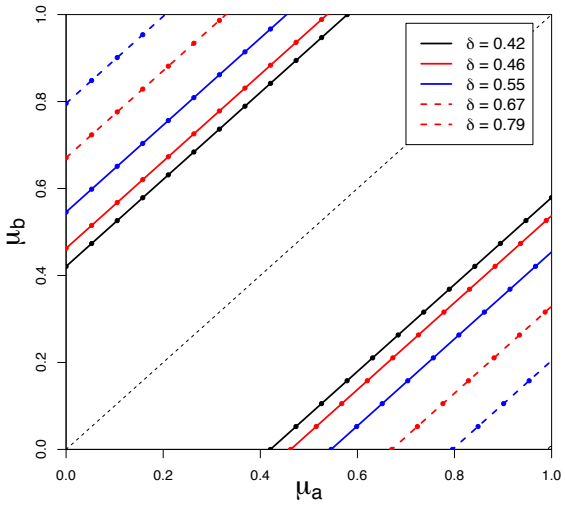

Figure 5.2: The Beam: Graphical depiction of the GROW $\Theta(\delta)$.

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

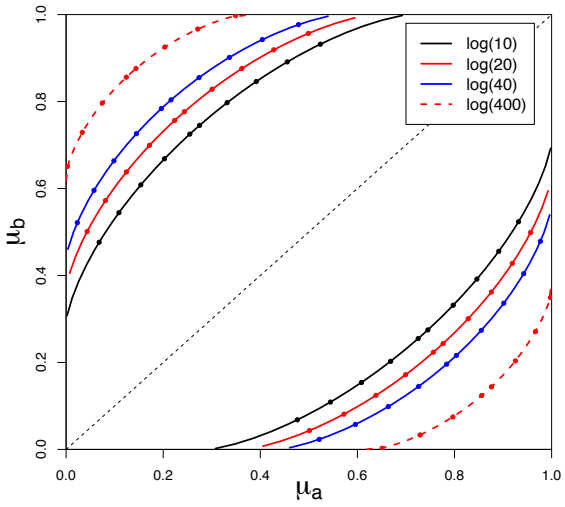

Figure 5.3: The Lemon: Graphical depiction of the KL-divergence based GROW $\Theta(\varepsilon)$.

$n_{1a}, n_{1,b}$, 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 $\inf_{\mu \in [0,1]} D(P_{\theta_1} \| P_\mu) = \log 20$.

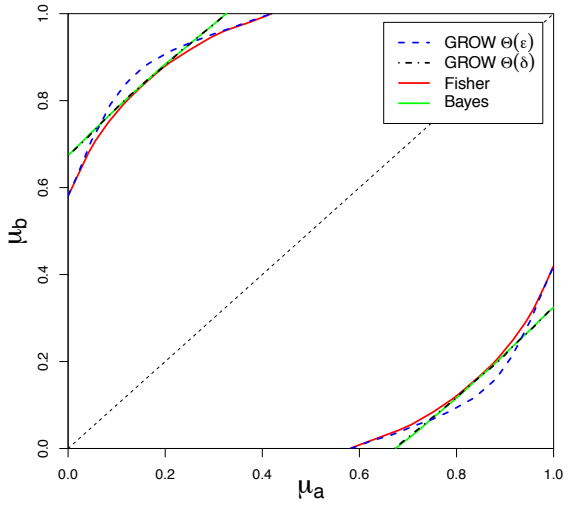

Figure 5.4: 0.8-iso-powerlines for the four different tests.

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!

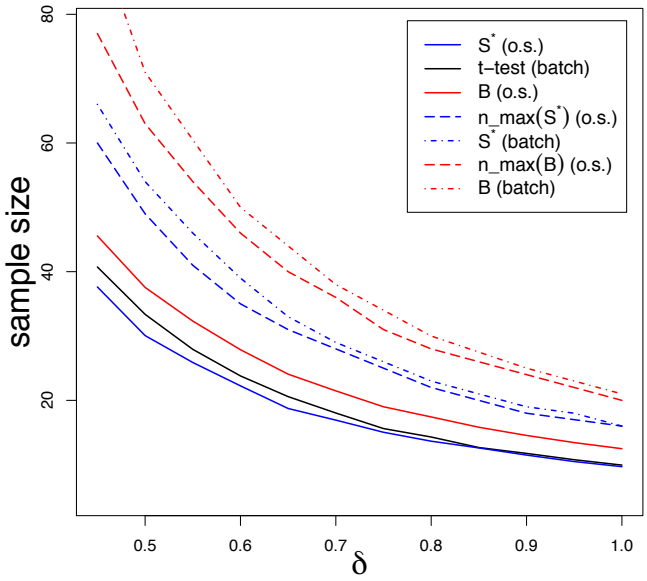

Figure 5.5: Effective sample size for the classical t-test (black), Bayesian t-test (E-test with Cauchy prior on δ) (red), and the δ-GROW E-test E^* with a two-point prior on δ (blue). The lines denoted *batch* denote the smallest fixed sample size at which power $\beta = 0.8$ can be obtained under $\mathcal{H}_1$ 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 $D(\theta_1 \| \Theta_0) = \mathrm{GR}(\{\theta_1\})$.