Safe Testing

Peter Grünwald[1], Rianne de Heide[2], and Wouter M. Koolen[3]

[1,2,3]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[1,2]Leiden University, Leiden, The Netherlands
[1,2,3] {pdg, heide, wmkoolen}@cwi.nl

June 20, 2019

Abstract

We present a new theory of hypothesis testing. The main concept is the S-value, a notion of evidence which, unlike P-values, allows for effortlessly combining evidence from several tests, even in the common scenario where the decision to perform a new test depends on the previous test outcome: *safe* tests based on S-values generally preserve Type-I error guarantees under such 'optional continuation'. S-values exist for completely general testing problems with composite null and alternatives. Their prime interpretation is in terms of gambling or investing, each S-value corresponding to a particular investment. Surprisingly, optimal "GROW" S-values, 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 S-values also have an interpretation as Bayes factors, with priors given by the JIPr. We illustrate the theory using two classical testing scenarios: the one-sample t-test and the 2×2-contingency table. In the t-test setting, GROW S-values correspond to adopting the right Haar prior on the variance, like in Jeffreys' Bayesian t-test. However, unlike Jeffreys', the *default* safe t-test puts a discrete 2-point prior on the effect size, leading to better behaviour in terms of statistical power. Sharing Fisherian, Neymanian and Jeffreys-Bayesian interpretations, S-values and safe tests may provide a methodology acceptable to adherents of all three schools.

1 Introduction and Overview

We present a new theory of hypothesis testing. 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 S-*test statistics*. These are simply *nonnegative* random variables that satisfy the inequality:

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

Even though they are random variables, we often refer to S-test statistics as S-*values*, emphasizing that they are to be viewed as an alternative to, and in many cases an improvement of, the classical P-value, noting that *large* S-values correspond to evidence against the null: for given S-value S and $0 \leq \alpha \leq 1$, we define $\mathrm{T}_\alpha(S)$, called the *safe test corresponding to S with significance level α*, as the function from $\mathbb{R}_0^+$ to $\{\mathrm{ACCEPT}_0, \mathrm{REJECT}_0\}$ satisfying $\mathrm{T}_\alpha(S) = \mathrm{REJECT}_0$ (i.e. '$\mathcal{H}_0$ is rejected') iff $S \geq 1/\alpha$.

Motivation P-values and standard null hypothesis testing have come under intense scrutiny in recent years (Wasserstein et al., 2016, Benjamin et al., 2018); S-values and safe tests offer several advantages. Most importantly, in contrast to P-values, S-values 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. A second reason is their enhanced *interpretability*, and a third is their flexibility: S-values based on Fisherian, Neyman-Pearsonian and Bayes-Jeffreys' testing philosophies all can be accommodated for. These three types of S-values 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 et al. (2019).

Contribution For simple (singleton) $\mathcal{H}_0$, S-values are closely related to *test martingales*, which have been studied before (see e.g. (Shafer et al., 2011)). Here, we develop the theory of S-values for completely general *composite* $\mathcal{H}_0$, for which hitherto next to nothing was known. We first (Theorem 1, Part 1 and 2) show that nontrivial S-values exist for general $\mathcal{H}_0$ and $\mathcal{H}_1$, as long as $\mathcal{H}_0 \neq \mathcal{H}_1$. Our second contribution is to propose a general design criterion, the *GROW* criterion, for S-values that are in some sense *optimal*. While there are several GROW S-values for the same $\mathcal{H}_0$ and $\mathcal{H}_1$, the most straightforward one is the *default* GROW S-value, which can be defined as long as $\mathcal{H}_1$ shares all parameters with $\mathcal{H}_0$ except a single *parameter of interest*. We next (Theorem 1, Part 3) show that GROW S-values have a surprising representation in terms of KL divergence between two special Bayes marginals which form a so-called *Joint Information Projection* (JIPr) (see also Figure 1). This allows us to compute such optimal S-values numerically by convex optimization. We then, in Section 4, give some examples, showing that for 1-sided tests with exponential families against simple $\mathcal{H}_0$, Johnson's (Johnson, 2013b,a) uniformly most powerful Bayes factors coincide with the default GROW S-values; we also provide 'quick and dirty' (non-GROW) S-values for general multivariate exponential family $\mathcal{H}_0$. Next, we show that Jeffreys' Bayesian t-tests, as well as the many other Bayes factors based on the right Haar prior suggested by Berger and collaborators (Berger et al., 1998, Dass and Berger, 2003, Bayarri et al., 2012) constitute S-values. However, Jeffreys' standard Bayesian t-test, while an S-value, is *not* default GROW, and we present a default GROW version of it that has significantly better properties in terms of statistical power. We also show how to calculate the default GROW S-value for the 2x2 contingency table (which behaves better under optional continuation than either standard frequentist tests (such as Fisher's exact test) or standard Bayes factors (such as Gunel-Dickey (Jamil et al., 2016, Gunel and Dickey, 1974)). Preliminary experiments (Section 5) suggest that with default GROW S-values, 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, although 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. Having thus provided default S-values for what are perhaps the two most commonly encountered testing scenarios, we end — after providing an overview of related work in Section 6 — with a discussion that clarifies how *safe testing* could provide a unification of Fisher's, Neyman's and Jeffreys' ideas on testing.

In the remainder of this introduction, we elaborate our contributions further in the context of the three main interpretations of S-values:

1. First Interpretation: Gambling The first and foremost interpretation of S-values 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 S\$; one may buy several and positive fractional amounts of tickets. (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 $r\mathbf{E}[S] \leq r$\$. Therefore, if the observed value of S 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 P-value is a random variable P such that for all $0 \leq \alpha \leq 1$, all $P \in \mathcal{H}_0$,

$$P(\mathrm{P} \leq \alpha) = \alpha. \tag{2}$$

A *conservative* P-value is a random variable for which (2) holds with '=' replaced by '$\leq$'. There is a close connection between (small) P- and (large) S-values. Indeed:

Proposition 1. *For any given* S*-value* S*, define* $\mathrm{P}_{[\mathrm{S}]} := 1/S$*. Then* $\mathrm{P}_{[\mathrm{S}]}$ *is a conservative* P*-value. As a consequence, for every* S*-value, any* $0 \leq \alpha \leq 1$*, the corresponding safe test* $\mathrm{T}_\alpha(S)$ *has Type-I error guarantee* α*, i.e. for all* $P \in \mathcal{H}_0$*,*

$$P(\mathrm{T}_\alpha(S) = \mathrm{REJECT}_0) \leq \alpha. \tag{3}$$

Proof. **(of Proposition 1)** Markov's inequality gives $P(S \geq \alpha^{-1}) \leq \alpha \mathbf{E}_P[S] \leq \alpha$. The result is now immediate. □

While S-values are thus conservative p-values, standard p-values satisfying (2) are by no means S-values; if S is an S-value and P is a standard p-value, and they are calculated on the same data, then we will usually observe $p \ll 1/S$ so S gives *less* evidence against the null; Example 1 and Section 6 will give some idea of the ratio between $1/S$ and p in various practical settings.

Combining 1. and 2.: Optional Continuation, GROW Proposition 2 below shows that *multiplying* S-values $S_{(1)}, S_{(2)}, \ldots$ for tests based on respective samples $Z_{(1)}$, $Z_{(2)}, \ldots$ (with each $Z_{(j)}$ being the vector of outcomes for the j-th test), gives rise to new S-values, even if the decision whether or not to perform the test resulting in $S_{(j)}$ was based on the value of earlier test outcomes $S_{(j-1)}, S_{(j-2)}, \ldots$. As a result (Prop. 2), *the Type I-Error Guarantee* (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 $S_{(1)}$ and, after observing $S_{(1)}$, reinvest all our new capital $\$S_{(1)}$ into $+S_{(2)}$, then after observing $S_{(2)}$ our new capital will obviously be $\$S_{(1)} \cdot S_{(2)}$, and so on. If, under the null, we do not expect to gain any money for any of the individual gambles $S_{(j)}$, 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 stopping and going home). We do not claim any novelty for Proposition 2 — it is implicit in earlier works such as Shafer et al. (2011). The real novelty is that nontrivial S-values exist for general composite $\mathcal{H}_0$ (Theorem 1, Part 1), that there exists a generically useful means

for constructing them following the *GROW* criterion, and that (Theorem 1, Part 2 and 3) the GROW S-value can be characterized in terms of a *JIPr* (joint information projection), graphically depicted in Figure 1.

In its simplest form, for non-overlapping $\mathcal{H}_0$ and $\mathcal{H}_1$, the GROW *(growth-rate optimal in worst-case)* criterion tells us to pick, among all S-values relative to $\mathcal{H}_0$, the one that maximizes *expected capital growth rate under* $\mathcal{H}_1$ in the worst case, i.e. the S-value S^* that achieves

$$\max_{S:S \text{ is an S-value}} \min_{P \in \mathcal{H}_1} \mathbf{E}_P [\log S] .$$

We give five reasons for using the logarithm in Section 3.1. Briefly when we keep using S-values with additional data batches as explained in Section 2 below, then optimizing for $\log S$ ensures that our capital grows at the fastest rate. Thus: restricting test statistics to S-values means that we do not expect to gain money under $\mathcal{H}_0$; and among all such S-values, the GROW S-value is the one under which our money grows fastest (we get evidence against $\mathcal{H}_0$ fastest) under $\mathcal{H}_1$.

3. Third Interpretation: Bayes Factors For convenience, from now on we write the models $\mathcal{H}_0$ and $\mathcal{H}_1$ as

$$\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\} \quad ; \quad \mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\},$$

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 A for more details). Suppose that Z represents the available data; in all our applications, $Z = (Y_1, \ldots, Y_N)$ is a vector of N outcomes, where 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

$$p_{W_j}(Z) := \int_{\Theta_j} p_\theta(Z) dW_j(\theta). \tag{4}$$

The *Bayes factor* is then given as:

$$\mathsf{BF} := \frac{p_{W_1}(Z)}{p_{W_0}(Z)}. \tag{5}$$

Whenever $\mathcal{H}_0 = \{P_0\}$ is *simple*, i.e., a singleton, then the Bayes factor is also an S-test statistic, 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

$$\mathbf{E}_P[\mathsf{BF}] := \int p_0(z) \cdot \frac{p_{W_1}(z)}{p_0(z)} dz = 1. \tag{6}$$

For such S-values that are really simple-$\mathcal{H}_0$-based Bayes factors, Proposition 1 reduced to the well-known *universal bound* for likelihood ratios that has been rediscovered many times (see Royall (2000) for an overview). If we act as 'strict' Bayesians, we may think of the simple $\mathcal{H}_0$ test really as a test between two simple hypotheses, $\mathcal{H}_1' = \{P_{W_1}\}$ and $\mathcal{H}_0$. In this strict

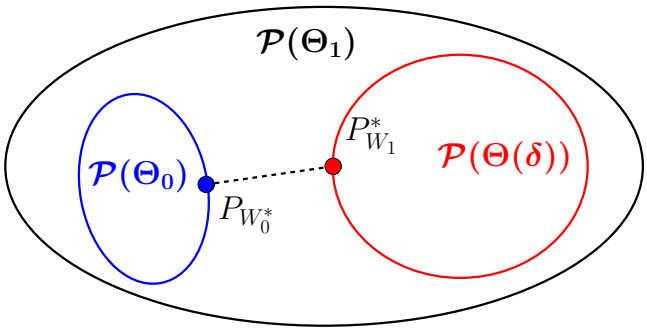

Figure 1: The Joint Information Projection (JIPr), with notation from Section 3. $\Theta_0 \subset \Theta_1$ represent two nested models, $\Theta(\delta)$ is a restricted subset of Θ_1 that does not overlap with Θ_0. $\mathcal{P}(\Theta) = \{P_W : W \in \mathcal{W}(\Theta)\}$, and $\mathcal{W}(\Theta)$ is the set of all priors over Θ, so $\mathcal{P}(\Theta)$ is the set of all Bayes marginals with priors on Θ. Theorem 1 says that the GROW S-value $S^*_{\Theta_1(\delta)}$ between Θ_0 and $\Theta_1(\delta)$ is given by $S^*_{\Theta_1(\delta)} = P_{W_1^*}/P_{W_0^*}$, the Bayes factor between the two Bayes marginals that minimize KL divergence $D(P_{W_1} \| P_{W_0})\}$.

Bayesian view, $\mathcal{H}_0$ is really a singleton, and then the Bayes factor (5) is an S-value — in fact it is then not just any S-value, it is even equal to the GROW S-value for $\mathcal{H}_0$ relative to $\mathcal{H}'_1$. However, almost all priors used in practice are adopted, at least to some extent, for pragmatic reasons, and thus, as pragmaticists, robust Bayesians or frequentists, we may want to adopt the S-value that is GROW relative to some subset Θ'_1 of Θ_1 or more generally, a nonsingleton subset $\mathcal{W}'_1$ of the set of all priors $\mathcal{W}(\Theta_1)$ over Θ_1. In Section 3.1 we describe an appealing default choice for picking Θ'_1. Our main result Theorem 1 expresses that, irrespective of how we define Θ'_1 or $\mathcal{W}'_1$, the resulting GROW S-value is *still* a Bayes factor, but in many (not all) cases it is based on priors quite unlike anything that's used in practice.

When $\mathcal{H}_0$ is itself composite, most Bayes factors $B = p_{W_1}/p_{W_0}$ will *not* be S-values any more, since for B to be an S-value we require (6) to hold for all $P_\theta, \theta \in \Theta_0$, whereas in general it only holds for $P = P_{W_0}$. However, Theorem 1 again expresses that, under regularity conditions, the GROW S-value for this problem is *still* 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 1). Finding the JIPr pair is thus a convex optimization problem, so that it will tend to be computationally feasible.

Again, the priors (W_0^*, W_1^*) are often unlike anything that's used in practice (Section 4.3 gives 2×2 tables as an example), but there does exists a highly important special case in which standard Bayes factors for composite $\mathcal{H}_0$ are S-values after all: the Bayes factors for testing with nuisance parameters satisfying a group invariance as proposed by Berger et al. (1998), Dass and Berger (2003) give S-values. For the special case of the Jeffreys' Bayesian t-test, we formally show in our second main result, Theorem 3, that it is an S-value — though not the default one — and in Theorem 3 we show how to modify the Bayesian t-test so that the resulting Bayes factor is default GROW (and will have better statistical power). Having given an initial overview, we now present the main contributions of this paper: we first formalize optional continuation, then consider the GROW S-value and our characterization of it, then give several examples, and we end by outlining how all this could give rise to a general theory

of *safe(r)* testing — but for concreteness we start with a simple example:

Example 1. [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 $Z_{(1)} = (Y_1, \ldots, Y_n)$. Standard Bayes factor tests in this scenario equip Θ_1 with a prior W that is a scale mixture of normals with mean 0, such as a Cauchy centered at 0, or simply a normal $N(0, \rho^2)$. For simplicity we focus (for now) on the latter case with $\rho = 1$, so that the prior has density $w(\mu) \propto \exp(-\mu^2/2)$. The Bayes factor is given by

$$S_{(1)} := \frac{p_W(Z)}{p_0(Z)} = \frac{\int_{\mu \in \mathbb{R}} p_\mu(Z) w(\mu) d\mu}{p_0(Z)}, \tag{7}$$

where $p_\mu(Z) = p_\mu(Y_1, \ldots, Y_n) \propto \exp(-\sum_{i=1}^n (Y_i - \mu)^2/2)$; by (6) we know that $S_{(1)}$ is an S-value. By straightforward calculation:

$$\log S = -\frac{1}{2}\log(n+1) + \frac{1}{2}(n+1) \cdot \breve{\mu}_n^2,$$

where $\breve{\mu}_n = (\sum_{i=1} Y_i)/(n+1)$ is the Bayes MAP estimator, which only differs from the ML estimator by $O(1/n^2)$: $\breve{\mu}_n - \hat{\mu}_n = \hat{\mu}_n/(n(n+1))$. If we were to reject Θ_0 when $S \geq 20$ (giving, by Proposition 1 a Type-I error guarantee of 0.05), we would thus reject if

$$|\breve{\mu}_n| \geq \sqrt{\frac{5.99 + \log(n+1)}{n+1}}, \text{ i.e. } |\hat{\mu}_n| \succeq \sqrt{(\log n)/n},$$

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 $|\hat{\mu}_n| \geq 1.96/\sqrt{n}$. The *default* GROW S-value for this problem that we describe in Section 4.1 would reach $S^* \geq 20$ if $|\hat{\mu}_n| \geq \tilde{\mu}_n$ with $\tilde{\mu}_n = c_n/\sqrt{n}$ 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 S-value that scores infinitely bad on our GROW optimality criterion (Example 3), the optimal GROW S^* is qualitatively more similar to a standard NP test than a standard Bayes factor approach. For general 1-dimensional exponential families, the default GROW S^* coincides with a 2-sided version of Johnson's (Johnson, 2013b,a) uniformly most powerful Bayes test, which uses a discrete prior W within $\mathcal{H}_1$: for the normal location family, $W(\{\tilde{\mu}_n\}) = W(\{-\tilde{\mu}_n\}) = 1/2$ with $\tilde{\mu}_n$ as above. Since the prior depends on n, we obtain a *local* (in time) Bayes factor by which we mean that for different n, the Bayes marginal P_W represents a different distribution (some statisticians would perhaps not really view this as 'Bayesian').

2 Optional Continuation

Consider a sequence of random variables $Z^{(k_{\max})} \equiv Z_{(1)}, \ldots, Z_{(k_{\max})}$ where each $Z_{(j)}$ is itself a sample, $Z_{(1)} = (Y_1, \ldots, Y_{n_1})$, $Z_{(2)} = (Y_{n_1+1}, \ldots, Y_{n_2}), Z_{(3)} = (Y_{n_2+1}, \ldots, Y_{n_3})$ and $k_{\max}$ is an arbitrarily large number. For example, $Z_{(1)}, Z_{(2)}, \ldots$ may represent a sequence of clinical trials or physical experiments, each $Z_{(j)}$ being the vector of all outcomes in trial j. We observe a first sample, $Z_{(1)}$, and measure our first S-value $S_{(1)}$ based on $Z_{(1)}$, i.e. we can write $S_{(1)} = s_{(1)}(Z_{(1)})$ for some function $s_{(1)} : \mathcal{Y}^{n_1} \to \mathbb{R}_0^+$. Then, if either the value of $S_{(1)}$ or, more generally of the underlying data $Z_{(1)}$, or of *side-information* $V_{(1)}$ is such that we (or

some other research group) would like to perform a second trial, a second data sample $Z_{(2)}$ is generated or made available, and a second test is performed, i.e. an S-value $S_{(2)} = s_{(2)}(Z_{(2)})$ based on data $Z_{(2)}$ is measured. Here the definition of $S_{(2)}$ may be equal to $S_{(1)}$ (i.e. we may have $n_1 = n_2$ and $s_{(1)} = s_{(2)}$) but our 'optional continuation' result still holds if this is not the case. We will require however that $Z_{(2)}$ is independent of $Z_{(1)}$.

After observing $S_{(2)}$, depending again on the value of $S_{(2)}$, $Z_{(2)}$ or $U_{(2)}$, 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 $k_{\max}$ tests have been performed. The decision whether or not to continue after k tests is encoded as the function $B_{(k+1)}$ which takes values in $\{\text{STOP}, \text{CONTINUE}\}$, where $B_{(k+1)} = \text{STOP}$ means that the kth test was the final one to be performed. We allow any deterministic rule or random process for deciding whether to stop or continue *that may depend, in arbitrary, random and unknown ways, on all data and side information observed in the past*, but it *is not allowed to depend on the future.* We can formalize this jointly with our independence assumptions as follows, where we abbreviate $V^{(k)} = (V_{(1)}, \ldots, V_{(k)})$:

Assumption A There exist random vectors $(U_{(1)}, Z_{(1)}), \ldots, (U_{(k_{\max})}, Z_{(k_{\max})})$ on the domain such that the joint distribution P underlying $(U_{(1)}, Z_{(1)}), \ldots, (U_{(k_{\max})}, Z_{(k_{\max})})$ has a marginal on $Z_{(1)}, \ldots, Z_{(k_{\max})}$ that coincides with P_θ for some $\theta \in \Theta_0$ and satisfies, for all $k \in \{1, \ldots, k_{\max}\}$,

$$\begin{aligned}
&(1)\ Z_{(1)}, \ldots, Z_{(k_{\max})} \text{ are independent.} \\
&(2)\ \text{There exist fns } b_1, \ldots, b_{k_{\max}} \text{ with } \ B_{(k)} = b_k(U^{(k)}) \\
&(3)\ \text{for } 1 \le k < k_{\max} : B_{(k)} = \text{STOP} \Rightarrow B_{(k+1)} = \text{STOP} \\
&(4)\ \text{for } 1 \le k < k_{\max} : U_{(k)} \perp Z_{(k)}, \ldots, Z_{(k_{\max})} \mid U^{(k-1)}, Z^{(k-1)}
\end{aligned} \tag{8}$$

The requirement that P is compatible with Θ_0 is needed because we are interested in showing properties of optional continuation *under the null.* Here $U_{(1)}, \ldots, U_{(k)}$ represent all data that is involved in the decision whether or not to continue to a next sample. In standard cases, we have $U_{(1)} \equiv 0$ (no information about the past in the beginning) and $U_{(k)} = (Z_{(k-1)}, V_{(k-1)}, B_{(k-1)})$ 'carries along' the past data, side-information $V_{(k-1)}$ (which may itself be sampled from an unknown distribution) and the previous continuation decision $B_{(k-1)}$. The need for the final requirement in (8) is clear: if it would not hold, we would allow a continuation rule that can peek ahead into the future such as 'continue to the $k+1$st trial if $S_{(k+1)} > 20$, i.e. if this trial will provide a lot of evidence'.

The following proposition gives the prime motivation for the use of S-values: the fact that the product of S-values remains an S-value, even if the decision to observe additional data and record a new S-value depends on previous outcomes.

Proposition 2. [Optional Continuation] *Suppose that P satisfies Assumption A. Let $S_{(0)} := 1$ and let, for $k = 1, \ldots, k_{\max}$, $S_{(k)} = s_k(Z_{(k)})$ be a function of $Z_{(k)}$ that is an S-value, i.e. $\mathbf{E}_{Z\sim P}[S_{(k)}] \le 1$. Let $S^{(K)} := \prod_{k=0}^{K} S_{(k)}$, and let $K_{\text{STOP}} := K - 1$ where $K \ge 1$ is the smallest number for which $B_{(K)} = \text{STOP}$. Then*

1. *For all $k \ge 1$, $S^{(k)}$ is an S-value.*

2. $S^{(K_{\text{STOP}})}$ *is an* S*-value.*

As a corollary, under all $P_0 \in \mathcal{H}_0$*, for every* $0 \leq \alpha \leq 1$*,*

$$P_0(\mathrm{T}_\alpha(S^{(K_{\text{STOP}})}) = \text{REJECT}_0) \ (= P_0(S^{(K_{\text{STOP}})} \geq \alpha^{-1}) \ ) \ \leq \alpha, \tag{9}$$

i.e. Type I-error guarantees are preserved under optional continuation, even for the most aggressive continuation rule which continues until the first K *is reached such that either* $\prod_{k=1}^{K} S_{(k)} \geq \alpha^{-1}$ *or* $K = k_{\max}$*.*

Proposition 2 verifies the claim we made in the introduction: no matter what optional continuation rule (definition of $B_{(k)}$) we use, as long as the resulting process satisfies (8), our Type-I error guarantee will still hold for the combined (multiplied) test outcome. Informally, we may say that S*-values are safe for Type-I errors under optional continuation.* A formal proof is in Appendix B, but an implicit proof has already been given in the beginning of this paper: the statement is equivalent to 'no matter what your role is for stopping and going home, you cannot expect to win in a real casino'.

As Example 2 below shows, it will be useful to generalize Proposition 2 to a setting in which the definition of the S-value $S_{(k)}$ that is applied to sample $Z_{(k)}$ may itself depend on $U_{(k)}$ and hence on data from the past $Z^{(k-1)}$ or side information from the past $V^{(k-1)}$: by (8) we may have $U^{(k)} = Z^{(k-1)}$ but not, e.g. $U^{(k)} = Z^{(k)}$. The following proposition shows that the result of multiplying until stopping, $S^{(K_{\text{STOP}})}$, is still an S-value as long as, for almost all possible instantiations u of $U^{(k)}$, $S_{(k)}$ is still an S-value conditioned on v:

Proposition 3. *We call a test statistic* $S_{(k)} = s_{(k)}(Z_{(k)}, U^{(k)})$ *that can be written as a function of* $Z_{(k)}$ *and* $U^{(k)}$ *and that satisfies*

$$\mathbf{E}[S_{(k)} \mid U^{(k)}] \leq 1 \tag{10}$$

an S-value for $Z_{(k)}$ conditional on $U^{(k)}$*. We have: Proposition 2 still holds if we replace '*$S_{(k)}$ *is an* S*-value' by '*$S_{(k)}$ *is an* S*-value conditional on* $U^{(k)}$*'.*

Various further extensions are possible. For example, in practice, when the decision whether to perform a new experiment or not is made, the value of $k_{\max}$ may of course be unknown or even undefined. While this is of no great concern, since the result above is valid for arbitrarily large $k_{\max}$, we can still generalize the result to unbounded $k_{\max}$, as long as $k_{\max} < \infty$ with probability 1 by recasting the setting in a measure-theoretic framework. Technically, $S_{(1)}, S_{(1)} \cdot S_{(2)}, \prod_{k=1}^{3} S_{(k)}, \ldots$ then becomes a *nonnegative supermartingale* and the optional continuation result follows trivially from Doob's celebrated *optional stopping theorem.* Once we take this stance, various further generalizations of Proposition 3 are possible; for example, the size $N_j - N_{j-1}$ of sample $Z_{(j)}$ may itself be dependent on past outcomes and side information as summarized in $U^{(j)}$; we omit further details.

Example 2. (Ex. 1, Cont.) Now let us take the standard Bayesian S-value (7) based on a normal prior and suppose we have observed data $Z_{(1)} = Y_1, \ldots, Y_n$, leading to $S_{(1)} = 18$ — promising enough for us to ask our boss for more money to provide some further experiments. Happily our boss grants the extra funding and we perform a new trial leading to data $Z_{(2)} =$

$(Y_{n+1}, \ldots, Y_{n_2})$. If we want to stick to the Bayesian paradigm, we can now use the following conditional S-value:

$$S_{(2)} := \frac{p_{W(\cdot|Z_{(1)})}(Z_{(2)})}{p_0(Z_{(2)})},$$

where $W(\cdot \mid Z_{(1)})$ is the Bayes posterior for μ based on data $Z_{(1)}$. To see that $S_{(2)}$ is a a conditional S-value given $V_{(1)} = Z_{(1)}$, note that $\mathbf{E}_{Z_{(2)} \sim P_0}[S_{(2)} \mid Z_{(1)}] = 1$, independently of the value of $Z_{(1)}$, by a calculation analogous to (6). Yet a simple calculation using Bayes' theorem shows that multiplying $S_{(1)}$ and $S_{(2)}$ (which gives a new S-value by Proposition 2), satisfies

$$S_{(1)} \cdot S_{(2)} = \frac{p_W(Z_{(1)}) \cdot p_{W(\cdot|Z_{(1)})}(Z_{(2)})}{p_0(Z_{(2)})} = \frac{p_W(Y_1, \ldots, Y_{n_2})}{p_0(Y_1, \ldots, Y_{n_2})},$$

which is exactly what one would get by Bayesian updating, showing that, for simple $\mathcal{H}_0$, combining S-values by multiplication can be done consistently with Bayesian updating.

However, it might also be the case that it is not us who get the additional funding but 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 $S'_{(2)}$ which is again a Bayes factor, but now with the original prior W on μ re-used rather than replaced by $W(\cdot \mid Z_{(1)})$. Even though this would not be standard Bayesian, $S_{(1)} \cdot S'_{(2)}$ would still be a valid S-value, 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.

Optional Stopping vs. Optional Continuation In this paper, our claim that S-values are safe under optional *continuation* refers to the fact (Proposition 2 and 3) that under Assumption A, products of S-values calculated on subsequent batches $Z_{(1)}, Z_{(2)}, \ldots$ remain S-values, can still be interpreted in monetary terms (as the capital obtained so far in sequential gambling), and still satisfy Type-I error guarantees. We now contrast this with the behaviour of S-values under optional *stopping*. Suppose that $Y_1, Y_2, \ldots$ are i.i.d. according to all $\theta \in \Theta_0 \cup \Theta_1$. Define an S-*process* to be a sequence of S-values $S_{[1]}, S_{[2]}, S_{[3]}, \ldots$ where for each i, $S_{[i]} : \mathbb{R} \to \mathbb{R}_0^+$ is an S-value that is a function of the first i outcomes (note the difference in notation: $S_{[i]}$, the S-value we use for a sample of i outcomes, vs $S_{(k)}$, an S-value to be used on the k-th sample). Then, optional stopping (in its common interpretation) would refer to stopping at N set to the first i at which we are sufficiently happy with the result, and reporting the S-value $S_{[N]}$. For example, we may set N to be the smallest N such that either $S_{[N]} > 20$ or we run out of money to perform new experiments. In general such an 'S-process' does *not* have a clear monetary interpretation, and consequently it does not lead to preservation of Type-I error probabilities under optional stopping. For example, using the type of S-values of Example 3, we can easily construct a sequence of S-values that satisfies, for $P \in \mathcal{H}_0$, $P(\sup_{i>0} S_{[i]} > 20) = 1$. Thus, in general the claim 'S-processes can handle optional stopping' does not hold. However, there do exist important special cases of S-processes which fare better with optional stopping. In such S-processes subsequent S-values $S_{[i]}$ and $S_{[i+1]}$ have to be interrelated in a particular manner. These S-processes are the so-called *test martingales*. To define these, we note first that we can group the same data $Y_1, Y_2, \ldots$ into batches in various ways; we now assume a grouping $Z'_{(1)} = Y_1, Z'_{(2)} = Y_2, \ldots$ with accompanying information $U'_{(1)}, U'_{(2)}, \ldots$. The following

definition of test martingale generalizes that of Shafer et al. (2011), who considered the case with $U'_{(i)} = Y'_{i-1}$; to make the definition concrete, assume that $U'_{(i)}$ is defined in this way for now.

An S-process $S_{[1]}, S_{[2]}, S_{[3]}, \ldots$ is called a *test martingale* if, with the grouping above, there exists a sequence of conditional S-values $S'_{(1)}, S'_{(2)}, \ldots$ (i.e. each $S'_{(i)}$ satisfies (10) with U replaced by U') such that we can write, for each $n > 0$, $S_{[n]} = S'^{(n)} = \prod_{i=1}^{n} S'_{(i)}$.

In terms of our gambling interpretation, general (conditional) S-values $S_{(1)}, S_{(2)}, \ldots$ applied to samples $Z_{(1)}, +Z_{(2)}, Z_{(3)}, \ldots$ of sizes $n_1, n_2 - n_1, n_3 - n_2, \ldots$ can be understood as pay-offs of bets that are not profitable under the null. $S^{(K)}$ then represents the accumulated capital after K gambles with reinvestment, starting with 1\$. Test martingales $S_{[1]}, S_{[2]}, \ldots$ can be interpreted in the setting where there is one bet per outcome $Z'_{(1)} = Y_1, Z'_{(2)} = Y_2, \ldots$ (rather than one bet per batch $Z_{(1)} = (Y_1, \ldots, Y_{n_1}), Z_{(2)} = (Y_{n_1+1}, \ldots, Y_{n_1+n_2}), \ldots$), each bet is not profitable under the null, and $S_{[i]} = S'^{(i)}$ is the capital after i gambles with reinvestment, starting with 1\$. Thus, if a statistician, at each time i, measures evidence by an S-value $S_{[i]}$ that is part of a test martingale $(S_{[1]}, S_{[2]}, \ldots)$, then Type-I error guarantees are preserved under optional stopping after all:

Corollary 1. [of Proposition 3] *Suppose $S_{[1]}, S_{[2]}, \ldots$ constitute a test martingale. Suppose that Assumption A holds for $Z'_{(i)} = Y_i$ and random variables $U'_{(i)}$ and $B'_{(i)}$. Then (9) holds with $S'^{(K_{\text{STOP}})} = S_{[K_{\text{STOP}}]}$, so that Type-I error guarantees are preserved under optional stopping based on $B'_{(1)}, B'_{(2)}, \ldots$.*

Technically, under Assumption A, products of (conditional) S-values define nonnegative supermartingales; Proposition 2 and 3 are just versions of Doob's optional stopping theorem, which implies that the stopped process $S^{(K_{\text{STOP}})}$ is itself an S-value and satisfies a Type-I error guarantee. This 'optional stopping at the level of batches $Z^{(j)}$' is what we call optional continuation in this paper. When a sequence of S-values $(S_{[1]}, S_{[2]}, \ldots)$ itself can be understood as a sequence of products of (conditional) S-values for batch size one, then it forms a nonnegative supermartingale that we call a *test martingale*; we can then stop at any time $i = 1, 2, \ldots$ and the stopped process $S'^{(i)} = S_{[i]}$ is itself an S-value and satisfies a Type-I error guarantee — this optional stopping at the level of size-1 batches is what in this paper we simply call 'optional stopping' for short.

An example of S-processes that are test martingales (and hence Type-I error guarantees are preserved under optional stopping) is given by the case with $\mathcal{H}_0$ simple, W_1 an arbitrary prior on Θ_1, and for all k, $S_{[k]} = p_W(Y^k)/p_0(Y^k)$ given by the Bayes factor (5); Example 2 describes the special case with Θ representing normal distributions. For these S-values, Assumption A and (10) are satisfied, and by Proposition 3 we can do optional stopping if B'_k can be written as a function of Y^{k-1}. Thus (as is in fact well-known), for simple $\mathcal{H}_0$, Bayes factors with fixed priors that do not depend on n behave well under optional stopping. On the other hand, if we take an S-process $(S_{[1]}, S_{[2]}, \ldots)$ where $S_{[i]}$ is a Bayes factor with a prior dependent on i, then Type-I error guarantees are not preserved under optional stopping. In Section 5.2 we will consider the case where we fix once and for all a prior that is based on optimizing capital growth for some given n, but we then use that prior in an S-process $(S_{[1]}, S_{[2]}, \ldots)$; we can then do optional stopping at time i, and Type I error guarantees will be preserved, even if $i \neq n$. The situation is much more complicated for composite $\mathcal{H}_0$: even if we set $S_{[n]} = p_{W_1}(Y^n)/p_{W_0}(Y^n)$ for fixed priors W_1 and W_0, independent of n, the resulting S-process is not a test martingale

and does not preserve Type-I error guarantees under optional stopping; for example, the S-values we encounter for composite $\mathcal{H}_0$ for 2×2 tables in Section 5.1 do *not* satisfy (10) with the grouping $Z'_{(1)} = Y_1, Z'_{(2)} = Y_2, \ldots$. In the t-test setting though, we can get S-processes that are test martingales, even though there $\mathcal{H}_0$ is composite — but for this test martingale we cannot take $U'_{(i)} = Y_{(i-1)}$, and we can only deal with slightly restricted forms of optional stopping. The upshot of all this is that for composite $\mathcal{H}_0$, it is substantially harder to construct an S-process that handles *optional stopping* (which can be in the middle of an experiment) than an S-value that handles *optional continuation* (which is in between experiments).

3 Main Result

From here onwards 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.

3.1 What is a good S-Value? The GROW Criterion

We start with an example that tells us how *not* to design S-values.

Example 3. [Strict Neyman-Pearson S-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_0 rather than the P-value itself. This can be seen as a safe test based on a special S-value S_{NP}: when P is a P-value determined by data Z, we define $S_{\text{NP}} = 0$ if $P > \alpha$ and $S_{\text{NP}} = 1/\alpha$ otherwise. For any $P_0 \in \mathcal{H}_0$ we then have $\mathbf{E}_{Z \sim P_0}[S_{\text{NP}}] = P_0(P \leq \alpha)\alpha^{-1} \leq 1$, so that S_{NP} is an S-value, and the safe test $\text{T}_\alpha(S_{\text{NP}})$ obviously rejects iff $P \leq \alpha$. However, with this S-value, there is a positive probability α of losing all one's capital. The S-value S_{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 S-value property (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 $\Theta_0 \cap \Theta'_1 = \emptyset$ of 'relevant' or 'sufficiently different from Θ_0' hypotheses (for example, in the t-test Θ'_1 might index all distributions with effect size δ larger than some 'minimum clinically relevant effect size' $\underline{\delta}$; see Section 4.2). If one wants to perform the most 'sensitive' test, one takes the largest Θ'_1 for which at the given sample size a specific power can be obtained; we will develop analogous versions of all these options 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 S-value with disastrous behaviour under optional continuation. However, we now show how to develop a notion of 'good' S-value 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 S-value with *large* $\mathbf{E}_{Z\sim P_\theta}[f(S)]$ 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 will sometimes lead to adoption of an S-value such that $P_\theta(S = 0) > 0$ 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 S-values $S_1, S_2, \ldots$ based on i.i.d. $Z_{(1)}, Z_{(2)}, \ldots$ with, for all j, $\mathbf{E}_{Z_{(i)}\sim P}[\log S_j] \geq L$, will a.s. satisfy $S^{\langle m \rangle} := \prod_{j=1}^m S_j = \exp(mL + o(m))$, i.e. S 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 S-values an interpretation within the MDL framework (Section 7.3); a fourth is that such growth-rate optimal S can be linked to power calculations after all, with an especially strong link in the one-dimensional case (Section 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 S-values S^* that achieve, for some $\Theta_1' \subset \Theta_1 \setminus \Theta_0$:

$$\inf_{\theta\in\Theta_1'} \mathbf{E}_{Z\sim P_\theta}[\log S^*] = \sup_{S\in\mathcal{S}(\Theta_0)} \inf_{\theta\in\Theta_1'} \mathbf{E}_{Z\sim P_\theta}[\log S] =: \mathrm{GR}(\Theta_1'), \tag{11}$$

where $\mathcal{S}(\Theta_0)$ is the set of all S-values that can be defined on Z for Θ_0. We call this special S^*, if it exists and is essentially unique, the GROW (*Growth-Rate-Optimal-in-Worst-case*) S-value relative to Θ_1', and denote it by $S^*_{\Theta_1'}$ (see Appendix 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, and being interested in the GROW S-value that achieves

$$\inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S^*] = \sup_{S\in\mathcal{S}(\Theta_0)} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S]. \tag{12}$$

Again, if an S-value achieving (12) exists and is essentially unique, then we denote it by $S^*_{\mathcal{W}_1'}$. If $\mathcal{W}_1' = \mathcal{W}(\{\theta_1\})$ is a single prior that puts all mass on a singleton θ_1, we write $S^*_{\theta_1}$. Linearity of expectation further implies that (12) and (11) coincide if $\mathcal{W}_1' = \mathcal{W}(\Theta_1')$; thus (12)) generalizes (11.

All S-values in the examples below, except for the 'quick and dirty' ones of Section 4.4, are of this 'maximin' form. They will be defined relative to sets $\mathcal{W}_1'$ with in one case (Section 4.2) $\mathcal{W}'$ representing a set of prior distributions on Θ_1, and in other cases (Section 4.1–4.3) $\mathcal{W}_1' = \mathcal{W}(\Theta_1')$ for a 'default' choice of a subset of Θ_1.

3.2 The JIPr is GROW

We now present our main result, illustrated in Figure 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 S-value *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 S-values 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 S-values take the form of a likelihood ratio/Bayes factor. The third says that GROW S-values for a whole set of distributions Θ_1' can be found, surprisingly, by a KL minimization problem — a characterization that is both intuitively pleasing and practically useful, since, by joint convexity of KL divergence (Van Erven and Harremoës, 2014), it means that the GROW S-value can be found by convex optimization.

Theorem 1. *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 S-value $S^*_{W_1}$ exists, is essentially unique, and satisfies*

$$\mathbf{E}_{Z \sim P_{W_1}}[\log S^*_{W_1}] \quad = \quad \sup_{S \in \mathcal{S}(\Theta_0)} \mathbf{E}_{Z \sim P_{W_1}}[\log S] \quad = \quad \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1} \| P_{W_0})$$

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 S-value takes a simple form: $S^*_{W_1} = p_{W_1}(Z)/p_{W_0^\circ}(Z)$.

3. Now let $\Theta_1' \subset \Theta_1$ and let $\mathcal{W}_1'$ be a convex subset of $\mathcal{W}(\Theta_1')$ such that for all $\theta \in \Theta_0$, all $W_1 \in \mathcal{W}_1'$, P_θ is absolutely continuous relative to P_{W_1}. Suppose that $\min_{W_1 \in \mathcal{W}_1'} \min_{W_0 \in \mathcal{W}_0} D(P_{W_1} \| P_{W_0}) = D(P^*_{W_1} \| P^*_{W_0}) < \infty$ is achieved by some (W_1^*, W_0^*) such that $D(P_{W_1} \| P_{W_0^*}) < \infty$ for all $W_1 \in \mathcal{W}_1'$. Then the minimum is achieved uniquely by (W_1^*, W_0^*), and the GROW S-value $S^*_{\mathcal{W}_1'}$ relative to $\mathcal{W}_1'$ exists, is essentially unique, and is given by

$$S^*_{\mathcal{W}_1'} = \frac{p_{W_1^*}(Z)}{p_{W_0^*}(Z)}, \tag{13}$$

and it satisfies

$$\inf_{W \in \mathcal{W}_1'} \mathbf{E}_{Z \sim P_W}[\log S^*_{\mathcal{W}_1'}] \quad = \quad \sup_{S \in \mathcal{S}(\Theta_0)} \inf_{W \in \mathcal{W}_1'} \mathbf{E}_{Z \sim P_W}[\log S] \quad = \quad D(P_{W_1^*} \| P_{W_0^*}). \tag{14}$$

If $\mathcal{W}_1' = \mathcal{W}(\Theta_1')$, then by linearity of expectation we further have $S^*_{\mathcal{W}_1'} = S^*_{\Theta_1'}$.

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 4.4.

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 (14), that the minima are both achieved, is sufficient but not always necessary: in the examples of Section 4.1 (1-dimensional) and 4.3 (2×2 tables), it holds, but in those of Section 4.2 (t-test) it does not, yet still in Theorem 3 we show a close analogue of (14) for this case.

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. To prove parts 2 and 3 without this condition, we need a nonstandard minimax theorem; and to prove part 1 (which does not rely on minima being achieved and which will be essential for Theorem 3) we need a deep result from Barron and Li (Li, 1999); these extended proofs are in Appendix C.

For Part 2, consider any $W_0' \in \mathcal{W}(\Theta_0)$ with $W_0' \neq W_0^\circ$, 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 $f(\alpha) \coloneqq 1 - \mathbf{E}_{Z\sim P_{W_0'}}[p_{W_1}(Z)/p_{W_0^\circ}(Z)]$. 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 $\mathbf{E}_{Z\sim P_{W_0'}}[p_{W_1}(Z)/p_{W_0^\circ}(Z)] \leq 1$. Since this reasoning holds for all $W_0' \in \mathcal{W}(\Theta_0)$, we get that $p_{W_1}(Z)/p_{W_0^\circ}(Z)$ is an S-value. To see that it is GROW, note that, for every S-value $S = s(Z)$ relative to $\mathcal{S}(\Theta_0)$, we must have, with $q(z) \coloneqq s(z)p_{W_0^\circ}(z)$, that $\int q(z)dz = \mathbf{E}_{Z\sim P_{W_0^\circ}}[S] \leq 1$, so q is a sub-probability density, and by the information inequality of information theory (Cover and Thomas, 1991), we have

$$\mathbf{E}_{P_{W_1}}[\log S] = \mathbf{E}_{P_{W_1}}\left[\log \frac{q(Z)}{p_{W_0^\circ}(Z)}\right] \leq \mathbf{E}_{P_{W_1}}\left[\log \frac{p_{W_1}(Z)}{p_{W_0^\circ}(Z)}\right] = \mathbf{E}_{P_{W_1}}[\log S^*_{W_1}],$$

implying that $S^*_{W_1}$ 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 $\mathbf{E}_{P_{(1-\alpha)W_1^*+\alpha W_1'}}[\log p_{W_1^*}(Z)/p_{W_0^*}(Z)] \leq \mathbf{E}_{P_{W_1^*}}[\log p_{W_1^*}(Z)/p_{W_0^*}(Z)]$. 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 $\inf_{W\in\mathcal{W}_1'} \mathbf{E}_{P_W}[\log S^*_{\mathcal{W}_1'}] = D(P_{W_1^*}\|P_{W_0^*})$, which is already part of (14). Note that we also have

$$\begin{aligned}
\inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S^*_{\mathcal{W}_1'}] &\leq \sup_{S\in\mathcal{S}(\Theta_0)} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S] \\
&\leq \inf_{W\in\mathcal{W}_1'} \sup_{S\in\mathcal{S}(\Theta_0)} \mathbf{E}_{Z\sim P_W}[\log S] = \inf_{W\in\mathcal{W}_1'} \sup_{S\in\mathcal{S}(\mathcal{W}(\Theta_0))} \mathbf{E}_{Z\sim P_W}[\log S] \\
&\leq \inf_{W\in\mathcal{W}_1'} \sup_{S\in\mathcal{S}(\{W_0^*\})} \mathbf{E}_{Z\sim P_W}[\log S] \leq \sup_{S\in\mathcal{S}(\{W_0^*\})} \mathbf{E}_{Z\sim P_{W_1^*}}[\log S].
\end{aligned}$$

where the first two and final inequalities are trivial, the third one follows from definition of S-value and linearity of expectation, and the fourth one follows because, as is immediate from the definition of S-value, for any set $\mathcal{W}_0$ of priors on Θ_0, the set of S-values relative to any set $\mathcal{W}' \subset \mathcal{W}_0$ must be a superset of the set of S-values relative to $\mathcal{W}_0$.

It thus suffices if we can show that $\sup_{S\in\mathcal{S}(\{W_0^*\})} \mathbf{E}_{Z\sim P_{W_1^*}}[\log S] \leq D(P_{W_1^*}\|P_{W_0^*})$. For this, consider S-values $S = s(Z) \in \mathcal{S}(\{W_0^*\})$ defined relative to the singleton hypothesis $\{W_0^*\}$. Since $\mathbf{E}_{Z\sim P_{W_0^*}}[s(Z)] \leq 1$ we can write $s(Z) = q(Z)/p_{W_0^*}(Z)$ for some sub-probability density q, and

$$\begin{aligned}
\sup_{S\in\mathcal{S}(\{P_{W_0^*}\})} \mathbf{E}_{P_{W_1^*}}[\log S] &= \sup_q \mathbf{E}_{Z\sim P_{W_1^*}}\left[\log \frac{q(Z)}{p_{W_0^*}}\right] \\
&= D(P_{W_1^*}\|P_{W_0^*}),
\end{aligned} \tag{15}$$

where the supremum is over all sub-probability densities on Z and the final equality is the information (in)equality again (Cover and Thomas, 1991). The result follows.

3.3 The Default GROW s-Value

To apply Theorem 1 to design s-values with good frequentist properties in the case that $\Theta_0 \subsetneq \Theta_1$, we must choose a subset Θ_1' with $\Theta_1' \cap \Theta_0 = \emptyset$. Usually, we first carve up Θ_1 into nested subsets $\Theta(\epsilon)$. 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

$$\Theta(\epsilon) := \{\theta \in \Theta_1 : d(\theta) \geq \epsilon\}. \tag{16}$$

In many cases (even if $\mathcal{H}_0$ is composite; see Section 4.2 and 4.3), there is just a single *scalar parameter of interest* $\delta \in \Delta \subseteq \mathbb{R}$, and we can (re-)parameterize the model such that $\Theta_1 = \{(\delta, \gamma) : \delta \in \Delta, \gamma \in \Gamma\}$ and $\Theta_0 = \{(0, \gamma) : \gamma \in \Gamma\}$ where parameter space Γ represents all distribution in $\mathcal{H}_0$. In that case, we shall call the family of s-values $\{S^*_{\Theta(\delta)}, \delta > 0\}$ with $d((\delta, \gamma)) = |\delta|$, the *default GROW* s*-values*. For this d, for $\underline{\delta} > 0$, we have

$$\Theta(\underline{\delta}) = \{(\delta, \gamma) : |\delta| \geq \underline{\delta}, \gamma \in \Gamma\},$$

which we call the *default GROW set for* $\underline{\delta}$. Similarly, we call $S^*_{\Theta(\delta)}$ the "default GROW s-value for $\underline{\delta}$" and also, if $\text{GR}(\Theta(\underline{\delta})) = L$, "the default GROW s-value for growth rate L". In general, default s-values are not the only sensible s-values to use (Section 4.3), but, as we now show, there is an important case in which they are.

4 Examples

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 any diffeomorphic (e.g. canonical or mean-value) parameterization, such that $0 \in \Theta$, and take Θ_1 to be some interval (t', t) for some $-\infty \leq t' \leq 0 < t \leq \infty$, 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 $Z = (Y_1, \ldots, Y_n)$ by the i.i.d. assumption. For notational simplicity we set

$$D(\theta \| 0) := D(P_\theta(Z) \| P_0(Z)) = nD(P_\theta(Y_1) \| P_0(Y_1)). \tag{17}$$

We consider the default GROW s-values $S^*_{\Theta(\delta)}$. 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. This gives default GROW sets $\Theta(\underline{\delta}) = \{\theta \in \Theta_1 : |\theta| \geq \underline{\delta}\}$.

One-Sided Test Here we set $t' = 0$ so that $\Theta(\underline{\delta}) = \{\theta \in \Theta_1 : \theta \geq \underline{\delta}\}$. The default GROW s-values take on a particularly simple form here: let W_1 be a prior probability distribution on $\Theta(\underline{\delta})$. As shown in Appendix D, we have for any such $W_1 \in \mathcal{W}(\Theta(\underline{\delta}))$, that

$$D(P_{W_1} \| P_0) \geq D(P_{\underline{\delta}} \| P_0), \tag{18}$$

i.e. the prior achieving the minimum on the left puts all mass on the single point $\underline{\delta}$ achieving the minimum among all points in $\Theta(\underline{\delta})$. It follows that the infimum in Theorem 1, Part 3 is

achieved by $P_{W_1^*} = P_{\underline{\delta}}$ and $P_{W_0^*} = P_0$, and the theorem gives us that

$$\sup_{S \in \mathcal{S}(\{0\})} \inf_{\theta \in \Theta(\underline{\delta})} \mathbf{E}_{Z \sim P_\theta}[\log S] = \inf_{\theta \in \Theta(\underline{\delta})} \mathbf{E}_{Z \sim P_\theta} \mathbf{E}[\log S^*_{\underline{\delta}}]$$
$$= D(P_{\underline{\delta}} \| P_0),$$

i.e. $S^*_{\Theta(\underline{\delta})} = S^*_{\underline{\delta}}$: default GROW S-values can be calculated as a likelihood ratio between two point hypotheses, even though $\Theta(\underline{\delta})$ is composite. Moreover, we now show that, for this 1-sided testing case, $S^*_{\Theta(\underline{\delta})}$ coincides with the *uniformly most powerful Bayes tests* of Johnson (2013b), giving further motivation for their use and an indication of how to choose $\underline{\delta}$. Note first that, since $\Theta_0 = \{0\}$ is a singleton, by Theorem 1, Part 2, we have that $S^*_W = p_W(Z)/p_0(Z)$, i.e. for all $W \in \mathcal{W}(\Theta_1)$, the GROW S-value relative to $\{W\}$ is given by the Bayes factor p_W/p_0. Also recall the definition of the 'safe test' T_α given underneath (1) in Section 1. The following result is a direct consequence of (Johnson, 2013b, Lemma 1); we omit the proof.

Theorem 2 (**Uniformly Most Powerful Local 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 *safe tests based on local Bayes factors, i.e. $\{\mathrm{T}_\alpha(S^*_W) : W \in \mathcal{W}(\Theta_1)\}$, the Type-II error is* uniformly *minimized over Θ_1 by setting W to a degenerate distribution putting all mass on $\underline{\delta}$:*

$$\textit{for all } \theta \in \Theta_1 : \min_{W \in \mathcal{W}(\Theta_1)} P_\theta\left(\mathrm{T}_\alpha(S^*_W) = \textsc{accept}_0\right) = P_\theta\left(\mathrm{T}_\alpha(S^*_{\underline{\delta}}) = \textsc{accept}_0\right),$$

and with the test $\mathrm{T}_\alpha(S^*_{\underline{\delta}}) = \mathrm{T}_\alpha(S^*_{\Theta(\underline{\delta})})$, $\mathcal{H}_0$ will be rejected iff the ML estimator $\hat{\theta}$ satisfies $\hat{\theta} \geq \underline{\delta}$.

Theorem 2 shows that, if the default GROW S-value is to be used in a safe test with given significance level α and one is further interested in maximizing power among all GROW S-values, then one should use $S^*_{\underline{\delta}}$ 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.

Two-Sided Test Let us now consider a two-sided test, with $\Theta_1 = (t', t)$ with $t' < 0$, still focusing on the default GROW S-values based on $\Theta(\underline{\delta}) = \{\theta \in \Theta_1 : |\theta| \geq \underline{\delta}\}$. The normal location family of Example 1 is a special case. While we found no explicit expression for the default GROW S-value $S^*_{\Theta(\underline{\delta})}$, it is easy to come up with an S-value $S^\circ_{\Theta(\underline{\delta})}$ with worst-case growth-rate almost as good as $S^*_{\Theta(\underline{\delta})}$, as follows:

$$S^\circ_{\Theta(\underline{\delta})} := \frac{1}{2} S^*_{\underline{\delta}} + \frac{1}{2} S^*_{-\underline{\delta}} = \frac{\frac{1}{2} p_{-\underline{\delta}}(Z) + \frac{1}{2} p_{\underline{\delta}}(Z)}{p_0(Z)}. \tag{19}$$

We know from the 1-sided case that $S^*_{\underline{\delta}}(Z) = p_{\underline{\delta}}(Z)/p_0(Z)$ is an S-value, and, by symmetry, the same holds for $S^*_{-\underline{\delta}}$. By linearity of expectation, mixtures of S-values are S-values, so $S^\circ_{\Theta(\underline{\delta})}$ must also be an S-value. In Appendix D we show that its worst-case growth rate cannot be substantially smaller than that of the optimal $S^*_{\Theta(\underline{\delta})}$.

Example 4. Consider the normal location setting of Example 1 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 $S^*_{\Theta(\mu)} = p_\mu(Z)/p_0(Z)$ and has $\mathrm{GR}(\Theta(\mu)) = D(\mu \| 0) = (n/2)\|\mu^2\|$. We now see that

the uniformly most powerful default GROW S-value at sample size n is given by the $\tilde{\mu}_n$ with $D(\tilde{\mu}_n \| 0) = -\log \alpha$, so that $\tilde{\mu}_n = \sqrt{2(-\log \alpha)/n}$. Thus (unsurprisingly), this GROW S value is a likelihood ratio test between 0 and $\tilde{\mu}_n$ at distance to 0 of order $1/\sqrt{n}$, and we expect to gain (at least) $-\log \alpha$ in capital growth if data is sampled from $\mu \geq \tilde{\mu}_n$.

In the 2-sided case, with $\Theta_1 = \mathbb{R}$, $S^*_{\Theta(\mu)}$ becomes $((1/2)p_\mu(Z) + (1/2)p_{-\mu}(Z))/p_0(Z)$. Even though we have no more guarantees that it is uniformly most powerful, we can still take the μ such that $\mathrm{GR}(\Theta(\mu)) = -\log \alpha$. This leads to the test we described in Example 1 with threshold $\sqrt{c_n/n} \to 2.72/\sqrt{n}$.

4.2 The Bayesian t-test and the default 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 $Y = (Y_1, \ldots, Y_n)$ given as $\mathcal{H}_0 = \{P_{0,\sigma}(Y) \mid \sigma \in \Gamma\}; \mathcal{H}_1 = \{P_{\delta,\sigma}(Y) \mid (\delta, \sigma) \in \Theta_1\}$, 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[1] $W^{\mathrm{C}}[\delta]$ 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 1 are invariably proper), the resulting Bayes factor is an S-value. We then show that it is in fact even the GROW S-value relative to all distributions in $\mathcal{H}_1$ compatible with $W^{\mathrm{C}}[\delta]$. The reasoning extends to a variety of other priors $W[\delta]$, including standard choices (such as a standard normal) and nonstandard choices (such as the two-point prior we will suggest further below).

Almost Bayesian Case: prior on δ available We fix a (for now, arbitrary) 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:

$$p_{W[\delta],\sigma}(y) = \int_{\delta \in \Delta} p_{\delta,\sigma}(y) \, \mathrm{d}W[\delta]. \tag{20}$$

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 $V := (V_1, \ldots, V_n)$ with $V_i = Y_i/|Y_1|$. We have that Y determines V, and (V, Y_1) determines $Y_1 = (Y_1, Y_2, \ldots, Y_n)$. The distributions in

[1] See Appendix A for the notational conventions used in $W^{\mathrm{C}}[\delta], P[V], \mathcal{S}\langle V \rangle$ and so on.

$\mathcal{H}_0 \cup \mathcal{H}_1$ can thus alternatively be thought of as distributions on the pair (V, Y_1). V is "Y with the scale divided out". It is well-known (and easy to check, see Appendix E) that under all $P \in \mathcal{H}_0$, i.e. all $P_{0,\sigma}$ with $\sigma > 0$, V has the same distribution P_0' with density p_0'. Similarly, one shows that under all $P_{W[\delta],\sigma}$ with $\sigma > 0$, V has the same pdf $p'_{W[\delta]}$ (which therefore does not depend on the prior on σ). We now get that, for all $\sigma > 0$,

$$S^*_{W[\delta]}\langle V\rangle := \frac{p'_{W[\delta]}(V)}{p'_0(V)} \tag{21}$$

satisfies $\mathbf{E}_{V\sim P}[S^*_{W[\delta]}\langle V\rangle] = 1$ for all $P \in \mathcal{H}_0$, hence it is an S-value.

We now restrict to priors $W[\delta]$ that are symmetric around 0; this could, for example, be a normal with mean 0, or a point prior putting mass 1/2 on some $\underline{\delta}$ and 1/2 on $-\underline{\delta}$, or the Cauchy prior mentioned earlier. Remarkably, for such symmetric priors, this 'scale-free' S-value 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 E, we have

$$\frac{\int_\sigma \bar{p}_{W[\delta],\sigma}(Y) w^H(\sigma) d\sigma}{\int_\sigma p_{0,\sigma}(Y) w^H(\sigma) d\sigma} = \frac{p'_{W[\delta]}(V)}{p'_0(V)} = S^*_{W[\delta]}\langle V\rangle. \tag{22}$$

Despite its improperness, w^H induces a valid S-value when used in the Bayes factor. The equivalence of this Bayes factor to $S^*_{W[\delta]}\langle V\rangle$ simply means that it manages to ignore the 'nuisance' part of the model and models the likelihood of the scale-free V instead. The reason this is possible is that w^H coincides with the right-Haar prior for this problem (Eaton, 1989, Berger et al., 1998), about which we will say more below.

Amazingly, it turns out that the S-value (22) is GROW:

Theorem 3. *Fix some prior $W[\delta]$ on δ that is symmetric around 0 and such that the (very weak) tail requirement $\mathbf{E}_{\delta\sim W[\delta]}[\log(1+|\delta|)] < \infty$ holds. Let $\mathcal{W}_1'$ be the set of all probability distributions on $\delta \times \sigma$ whose marginal on δ coincides with $W[\delta]$. Let $n > 1$. For $S^*_{W[\delta]}\langle V\rangle$ as defined by (22)) we have, in very close analogy to (14:*

$$\inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S^*_{W[\delta]}\langle V\rangle] = \sup_{S\in\mathcal{S}(\Theta_0)} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S] \tag{23}$$

$$= \inf_{W\in\mathcal{W}_1'} \inf_{W[\sigma]\in\mathcal{W}(\Gamma)} D(P_W \| P_{0,W[\sigma]}) \tag{24}$$

$$< \infty,$$

where the second infimum is over all priors on $\sigma > 0$. Thus $S^*_{W[\delta]}\langle V\rangle = S^*_{\mathcal{W}_1}$: the Bayes factor based on the right Haar prior, is not just an S-value, but even the GROW S-value relative to the set of all priors on $\delta \times \sigma$ that are compatible with $W[\delta]$.

Default GROW safe t-test: prior on δ not available What if we have no clear idea on how to choose a marginal prior on δ? In that case, we can once again use the *default* GROW S-value for parameter of interest δ, with, for $\underline{\delta} > 0$, GROW sets $\Theta(\underline{\delta}) = \{\delta : |\delta| \geq \underline{\delta}\}$. Let $W_{\underline{\delta}}$ be the prior that puts mass 1/2 on $\underline{\delta}$ and 1/2 on $-\underline{\delta}$. The following theorem, with proof similar to that of Theorem 3, shows that the Bayes factor based on the right Haar prior w^H and this prior is equal to the GROW S-value relative to $\Theta(\underline{\delta})$.

Theorem 4. *Fix $\underline{\delta} > 0$. Fix a convex set of priors $\mathcal{W}_1[\Delta]$ on δ so that for all $W \in \mathcal{W}_1[\Delta]$, $W(|\delta| < \underline{\delta}) = 0$ and such that $\mathcal{W}_1[\Delta]$ contains the prior $W_{\underline{\delta}}$. Let $n > 1$. For $S^*_{W_{\underline{\delta}}}\langle V\rangle$ as defined by (22) (with $W_{\underline{\delta}}$ in the role of $W[\delta]$) we have that*

$$\inf_{W[\sigma]\in\mathcal{W}(\Gamma),W[\delta]\in\mathcal{W}_1[\Delta]} \mathbf{E}_{Z\sim P_{W[\delta],W[\sigma]}}[\log S^*_{W_{\underline{\delta}}}\langle V\rangle] = \sup_{S\in\mathcal{S}(\Theta_0)} \inf_{\substack{W[\delta]\in\mathcal{W}_1[\delta]\\ W[\sigma]\in\mathcal{W}(\Gamma)}} \mathbf{E}_{Z\sim P_{W[\delta],W[\sigma]}}[\log S] \quad (25)$$

$$= \inf_{W[\sigma]\in\mathcal{W}(\Gamma),W[\delta]\in\mathcal{W}_1[\delta]} D(P_{W[\delta],W[\sigma]}\|P_{0,W[\sigma]})$$

$$< \infty,$$

where $P_{W[\delta],W[\sigma]}$ is the Bayes marginal based on the prior W under which δ and σ are independent, with marginals $W[\delta]$ and $W[\sigma]$, respectively.

Extension to General Group Invariant Bayes Factors In a series of papers (Berger et al., 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 et al. (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 θ_0 in both models gives rise to an S-value. 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 suitable analogues of Theorem 3 and Theorem 4 hold. The proof of these theorems seems readily extendable to the general group invariant setting, with the exception of Lemma 2 in Appendix F which uses particular properties of the variance of a normal; generalizing this lemma is a major goal for future work.

4.3 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 $\mathcal{X} \times \mathcal{Y}$, 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, $X = x = (x_1, \ldots, x_n)$, 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 $X = x$, the counts n, $n_a = N_a(x)$ and n_b (see Table 1), the likelihood of an individual sequence $y \mid x$ with statistics $N_{a0}, N_{b0}, N_{b0}, N_{b1}$ becomes:

$$p_{\mu_{1|a},\mu_{1|b}}(y \mid x) = p_{\mu_{1|a},\mu_{1|b}}(y \mid x, n_a, n_b, n) \quad (26)$$

$$= \mu_{1|a}^{N_{a1}}(1-\mu_{1|a})^{N_{a0}} \cdot \mu_{1|b}^{N_{b1}}(1-\mu_{1|b})^{N_{b0}}$$

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 $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ independent,

with $Y_i, \ldots, Y_n$ i.i.d. Ber(μ_1) distributed, $\mu_1 \in \Theta_0 := [0,1]$, i.e. $p_{\mu_1}(y \mid x) = p_{\mu_1}(y) = \mu_1^{N_1}(1-\mu_1)^{N_0}$. To test $\mathcal{H}_0$ against $\mathcal{H}_1$, we numerically calculate the GROW S-value $S^*_{\Theta(\epsilon)}$ where $\Theta(\epsilon)$ is defined

	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 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.

via (16) for two different divergence measures detailed further below. In both cases, $\Theta(\epsilon)$ will be compact, so that by the joint lower-semicontinuity 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 1 to infer that the GROW S-value is given by $S^*_{\mathcal{W}(\Theta(\epsilon))} = S^*_{\Theta(\epsilon)} = p_{W_1^*}(Y \mid X)/p_{W_0^*}(Y)$. Note that the 'priors' W_1^* and W_0^* may depend on the observed 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 F.1 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 $\Theta(\epsilon)$. 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 *default GROW* S-value for parameter of interest δ. To achieve this, we re-parameterize the model in a manner that depends on 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 $p'_{\delta,\mu_1}(y|x)$ (the probability in the new parameterization) to be equal to $p_{mu_{1|a},\mu_{1|b}}(y|x)$ as defined above. As long as 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 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. The second option for defining $\Theta(\epsilon)$ is to take the original parameterization, and have d in (16)) be the KL divergence. This choice is motivated in Appendix G. Then $\Theta(\epsilon)$ is the set of $(\mu_{1|a}, \mu_{1|b}$ with

$$\inf_{\mu'_1 \in [0,1]} \frac{D(P_{\mu_{1|a},\mu_{1|b}} \| P_{\mu'_1})}{n} = \frac{D(P_{\mu_{1|a},\mu_{1|b}} \| P_{\mu_1})}{n} \geq \epsilon.$$

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 $n_{1a}, n_{1,b}$, the set $\Theta(\epsilon)$ does not change if we multiply n_{1a} and n_{1b} by the same fixed positive

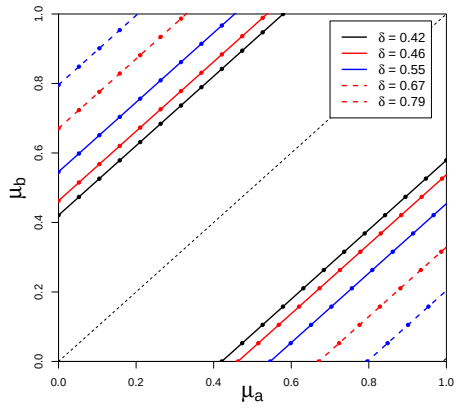

Figure 2: The Beam: Graphical depiction of the default GROW $\Theta(\delta)$.

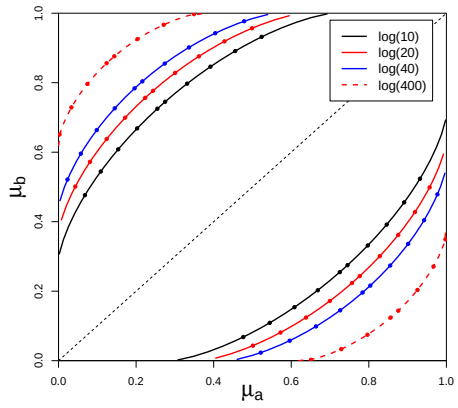

Figure 3: The Lemon: Graphical depiction of the KL-divergence based GROW $\Theta(\epsilon)$.

integer. Note also that the distributions $P_{\mu_{1|a},\mu_{1|b}}$ and P_{μ_1} are again conditional on the given 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 $\Theta(\epsilon)$ for various values of ϵ; this is done in Figure 3, where, for example, the set $\Theta(\epsilon)$ for $\epsilon \in \{\log 10, \log 20, \ldots, \log 400\}$ 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 3.

4.4 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 $\Theta(\epsilon)$ 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 1, Part 3 can be applied with $\mathcal{W}_1' = \mathcal{W}(\Theta(\epsilon))$, the resulting Bayes factor $p_{W_1^*}(Z)/p_{W_0^*}(Z)$ will be a GROW s-value. The main condition for

Part 3 is the requirement that $D(P_{W_1'} \| P_{W_0^*}) < \infty$ for all $W' \in \mathcal{W}(\Theta(\epsilon))$, which automatically holds if $D(P_\theta \| P_{W_0^*}) < \infty$ for all $\theta \in \Theta(\epsilon)$. 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 $\epsilon > 0$, $\Theta(\epsilon)$ is a compact subset of Θ_1 and for each $\theta \in \Theta_1$ that is not on the boundary, there is an $\epsilon > 0$ such that $\theta \in \Theta(\epsilon)$.

For large n though, numerical calculation of GROW S-values 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 $\Theta(\epsilon)$) S-values that take less computational effort. It turns out that these exist: if one is willing to be 'Bayesian' about Θ_1 and specify a prior W_1 on Θ_1, then one can calculate a *conditional* GROW-S-value. We illustrate this for the contingency table setting: conditional on the sufficient statistic relative to $\mathcal{H}_0$, $\hat{\mu}_1(Y) = N_1/n$, all distributions in $\mathcal{H}_0$, assign the same probability mass $p_0(y \mid \hat{\mu}_1(y)) = 1/\binom{n}{N_1}$ to all y with $\hat{\mu}_1(y) = \hat{\mu}_1(Y)$. The conditional S-value is then given by

$$S = \frac{p_{W_1}(Y \mid \hat{\mu}_1(Y), x)}{p_0(Y \mid \hat{\mu}_1(Y))} = \binom{n}{N_1} \cdot \frac{p_{W_1}(Y \mid x)}{p_{W_1}(\hat{\mu}_1(Y) \mid x)}.$$

While this $\mathcal{S}$-value may not be GROW, it is still meaningful if one has reason to adopt W_1. This 'quick and dirty' S-value approach 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 $\hat{\theta}_0(y^n)$ is almost surely well-defined under all $P \in \mathcal{H}_0$, whereas at the same time, $\hat{\theta}_0(Y^n)$ is a sufficient statistic for $\mathcal{H}_0$, i.e. there is a 1-to-1 correspondence between the ML estimator $\hat{\theta}_0(Y^n)$ and the sufficient statistic $\phi(Y^n)$. 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 a reasonable prior W_1 on Θ_1 is available, we can efficiently calculate nontrivial S-values of the form $p_{W_1}(Z \mid \hat{\theta}_1(Z))/p_0(Z \mid \hat{\theta}_0(Z))$ but whether these are sufficiently strong approximations of the GROW S-value 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 power with respect to the full JIPr[2].

5 Testing Our GROW Tests

We perform some initial experiments with our default and nondefault GROW S-values 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.

5.1 Case 1: Fixed n, $\underline{\epsilon}$ unknown

Suppose that n is fixed but we have no idea what the smallest $\underline{\epsilon}$ is such that $\theta_1 \in \Theta((\underline{\epsilon}))$. We may then simply 'give up' on $\theta \in \Theta_1$ that are too close to Θ_0, where 'too close' depends on the

[2]Although it was not connected to S-values, the idea to modify Bayes factors for nested exponential families by conditioning on the smaller model's sufficient statistic is due to T. Seidenfeld (2016).

given sample size. Formally, we fix an $L > 0$ and we determine ϵ_L: the smallest ϵ such that $S^*_{\Theta(\epsilon)}$ achieves rate L, i.e. $\text{GR}(\Theta(\epsilon)) = L$. We then use as our S-value $S^*_{\Theta((\epsilon_L))}$. In other words, we are really testing Θ_0 against $\Theta((\epsilon_L))$ rather than Θ_1. It is of course not clear how exactly we should choose L — but the same can be said for the traditional choice of significance levels and powers. In fact, we can fix an L and use $S^*_{\Theta((\epsilon_L))}$ even if we do not know in advance what α will be used (as long as the choice of α is data-independent): S-values lead to valid Type-I error guarantees for every fixed α, irrespective of the L for which they are defined. We can also use such an S-value if the test will be purely diagnostic and no accept/reject decision will be taken.

However, if we *do* know the significance level α we have used and we have a desired power $1 - \beta$, then we can try to determine the smallest $\underline{\epsilon} := \underline{\epsilon}(\beta)$, i.e. the largest GROW set $\Theta(\underline{\epsilon})$, for which the desired power $1 - \beta$ can be achieved by *some* d-based GROW S-value $S^*_{\Theta(\epsilon^*)}$, uniformly for all $\theta_1 \in \Theta(\underline{\epsilon})$. Thus, We should be very careful here, since we may have $\underline{\epsilon} \neq \epsilon^*$. For example, in case of a point null with a one-sided test as in Section 4.1, we should take ϵ^* such that $\text{GR}(\Theta(\epsilon')) = -\log \alpha$, since by Theorem 2 this will give the *uniformly* most powerful safe test (and ϵ^* does not depend on β). Yet the power of this test will depend on the P_θ with $\theta \in \Theta_1$ from which data are sampled, and will be only larger than $1 - \beta$ for $\theta \in \Theta^{(\underline{\epsilon})}$ for some potentially different $\underline{\epsilon}$.

Mini-Simulation-Study 1: The 2x2 Table Here we investigate these ideas within the contingency table setting.

We first consider the default GROW S-values $S^*_{\Theta(\delta)}$ relative to parameter of interest $\delta = |\mu_{1|a} - \mu_{1|b}|$, the first option considered in Section 4.3. For a grid of $\underline{\delta}$'s in the range $[0.4, 0.9]$ we looked at the best power that can be achieved by a default GROW S-value $S^*_{\Theta(\delta^*)}$, i.e. we looked for the δ^* (again taken from a grid in the range $[0.4, 0.9]$) such that

$$1 - \underline{\beta}(\underline{\delta}, \delta^*) := \inf_{\theta \in \Theta((\underline{\delta}))} P_\theta \left(\log S^*_{\Theta((\delta^*))} \geq -\log \alpha \right) \tag{27}$$

is maximized. We summarized the results in Table 2. We see that, although we know of no analogue to Johnson's Theorem 2 here, something like a "uniformly most powerful default GROW safe test" does seem to exist — it is given by $S^*_{\Theta(\delta^*)}$ with $\delta^* = 0.50$; and we can achieve power 0.8 for all $\theta \in \Theta(\underline{\delta})$ with $\delta \gtrapprox 0.5$. The same exercise is repeated with the GROW S-values defined relative to the KL divergence in Table 3, again indicating that there is something like a uniformly most powerful default 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' default GROW S-value $S^*_{\Theta(\delta*)}$ with $\delta^* = 0.50$ (see Table 2) which we call GROW($\Theta(\delta)$) and the 'uniformly most powerful' KL-based GROW S-value $S^*_{\Theta(\epsilon^*)}$ with $\epsilon^* = \log 16$ (see Table 3) which we call ($\Theta(\epsilon)$). The 0.8-iso-power lines are depicted in Figure 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 default 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 default GROW and default

$\underline{\delta}$	$\text{GR}(\Theta(\underline{\delta})) = D(P_{W_1^*} \| P_{W_0^*})$	δ^*	power $1 - \bar{\beta}$
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 2: Relating $\underline{\delta}, \delta^*$, power and capital growth $\text{GR}(\Theta(\underline{\delta}))$ for $n_a = n_b = 10$ for the default GROW S-values. For example, the row with 0.42 in the first column corresponds to the two black lines in Figure 2 which represent all $\theta_1 = (\mu_{1|a}, \mu_{1|b})$ with $\delta = 0.42$.

$\log n\underline{\epsilon}$	$\text{GR}(\Theta(\underline{\epsilon})) = D(P_{W_1^*} \| P_{W_0^*})$	$\log n\epsilon^*$	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 3: Relating $\underline{\epsilon}, \epsilon^*$, power and capital growth $\text{GR}(\Theta(\underline{\epsilon}))$ for $n_a = n_b = 10$ for the KL-based GROW S-values. For example, the row with 20 in the first column corresponds to the two curved red lines in Figure 3 which represent all $\theta_1 = (\mu_{1|a}, \mu_{1|b})$ with $\inf_{\mu \in [0,1]} D(P_{\theta_1} \| P_\mu) = \log 20$.

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).

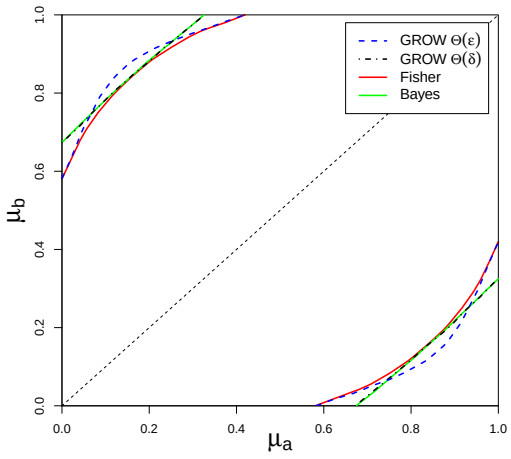

Figure 4: 0.8-iso-powerlines for the four different tests.

5.2 Case 2: n to be determined, $\underline{\delta}$ known

Consider default GROW S-values 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 S-value $S^*_{\Theta(\delta^*)}$ 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 safe test at level α by rejecting when $\mathsf{BF} \geq 20 = 1/\alpha$) and our safe test based on the default GROW S-value $S^*_{\Theta(\delta^*)}$ that maximizes power while having the GROW property. For the standard t-test we can just compute the required (batch) sample size. This is plotted (black line) in Figure 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 default GROW t-test (larger by a factor of around $1.4 - -1.6$).

However, these three lines do not paint the whole picture: for any symmetric prior $W[\delta]$ on δ, the safe test based on $S^*_{W[\delta]}$ given by (22), which included both the Bayesian t-test and our

default GROW t-test preserves Type-I error guarantees not just under optional continuation, but also under a slightly restricted form of *optional stopping*, as was shown by (Hendriksen et al., 2018) (and anticipated in various papers by Berger and collaborators, e.g. Bayarri et al. (2016)). In terms of Proposition 3 and Corollary 1, we observe a sequence of data $Y_1, Y_2, \ldots$; we then define $U_{(j)} = V_{j-1}$ with $V_0 \equiv 1$ and for $j \geq 1$, $V_j = Y_j/Y_1$ as in Section 4.2; $Z'_{(j)} = Y_j$, and $S_{[j]} = S^*_{W[\delta]}\langle V_1, \ldots, V_j\rangle$ the Bayes factor based on the right Haar prior as in (22). Then Hendriksen et al. (2018)) show that$(S_{[1]}, S_{[2]}, \ldots)$ constitute a test martingale. Hence, by Corollary 1, $S_{[K_{\text{STOP}}]} = S_{W[\delta]}\langle V_{K_{\text{STOP}}}\rangle$, i.e. the Bayes factor based on the right Haar prior stopped at K_{STOP} is an S-value as long as the decision $B'_{(t)}$ whether to stop or not after t outcomes is determined by a function of V^t. As can be seen from (22), for each symmetric prior $W[\delta]$ on δ, be it Cauchy or our two-point-prior, the S-value $S^*_{W[\delta]}(Y_1, \ldots, Y_t)$ can be written as a function of $V_1, \ldots, V_t$, and thus, by optional stopping at the smallest t such that $S^*_{W[\delta]}(Y_1, \ldots, Y_t) \geq 1/\alpha$, $B_{(t)}$ can be written as a function of V^t. The corollary thus implies that Type I error guarantees are preserved under this aggressive stopping rule.

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 default GROW S-value $S^*_{\Theta(\delta^*)}$ 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 $S^*_{\Theta(\delta^*)}(Y_1, \ldots, Y_t)$ until either this S-value 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 we see that this effective sample size is about *equal* to the fixed sample size we need with the standard t-test to obtain the required power (it seems even slightly better for small δ, but the difference is on the order of just 1 example and may be due to numerical imprecisions). Thus, quite unlike the classical t-test, our default GROW t-test S-value preserves Type I error probabilities under optional stopping; it needs more data than the classical t-test in the worst-case, *but not more on average under* $\mathcal{H}_1$. For a Neyman-Pearsonian hypothesis tester, this should be a very good reason to adopt it!

6 Earlier, Related and Future Work

Test Martingales, Sequential Tests, Conditional Frequentist Tests As seen in Section 2, S-values constitute a natural weakening of the concept of *test martingale*, a notion that in essence goes back to Ville (1939), the paper that introduced the modern notion of a martingale. S-values themselves have probably been originally introduced by Levin (of P vs NP fame) Levin (1976) (see also (Gács, 2005)) under the name *test of randomness*, but Levin's abstract context is quite different from ours. The first detailed study of S-values in a context more similar to ours is, presumably, Shafer et al. (2011), who call S-values 'Bayes factors', a terminology which can be explained since, like Levin and Gács, the authors (almost) exclusively focus on simple $\mathcal{H}_0$. The same can also be said for *sequential testing* (Lai, 2009) as pioneered by Wald and developed further by Robbins, Lai and others : the methods are related, but the focus in sequential testing is again almost exclusively on point null hypotheses. Very recent related work that builds on sequential testing ideas are the *anytime* P*-values* of Johari et al. (2015), who provide corrections to P-values that allow one to preserve Type I error

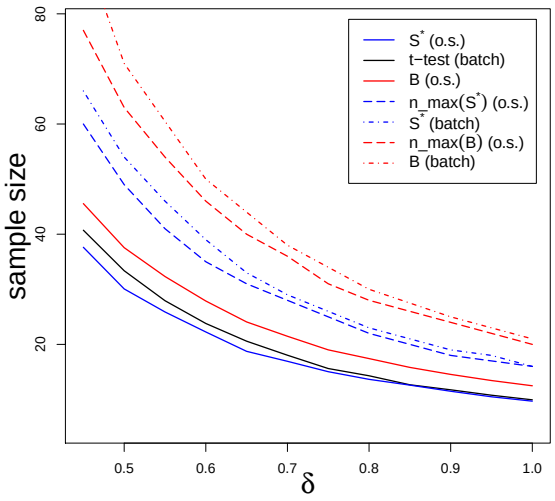

Figure 5: Effective sample size for the classical t-test (black), Bayesian t-test (S-test with Cauchy prior on δ) (red), and the default GROW S-test S^* with a two-point prior on δ (blue). The lines denoted *batch* denote the smallest fixed sample size at which power $\beta = 0.8$ can be obtained under $\mathcal{H}_1$ 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 S^*, 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 (S^* with o.s.), 1.1 (Bayes t-test with o.s.), 1.5 (S^* 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.

guarantees under optional stopping. While the goal of this work is thus very similar to ours, there are no obvious connections to either a Bayesian approach or a monetary interpretation or an information projection; it would be quite interesting to compare the results in terms of amount of data needed before a reject at a given level takes place.

Finally, in a series of papers starting with the landmark (Berger et al., 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 approach there is no direct link to optional continuation.

Related Work on Relating P- and S-values 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, $S := 1/f(p)$ is an S-value. 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 et al., 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, rejection under the safe test with significance level $\alpha = 0.05$, so that $S \geq 20$, would then correspond to reject only if $p \leq f_{\mathrm{vs}}^{-1}(0.05) \approx 0.0032$, requiring a substantial amount of additional data for rejection under a given alternative. Note that the S-values we developed for *given* models in previous sections are more sensitive than such generic calibrators though. For example, in Example 1 the threshold $2.72/\sqrt{n}$ corresponding to $\alpha = 0.05$ corresponds roughly to $p = 0.007$ (for composite H_0 as in the safe t-test there does not seem to be such a generic 'factor' that is independent of n). Another issue with calibrating p-values is that the resulting S-value are generally not capable of handling optional stopping, whereas, as we have seen for e.g. the t-test, in some (not all) settings, GROW S-values allow for optional stopping after all.

Related Work: Testing based on Data-Compression and MDL

Example 5. 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 $y^n = y_1, \ldots, y_n$' as

$$n - L_{\mathtt{zip}}(y^n),$$

where $L_{\mathtt{zip}}(y^n)$ is the number of bits needed to code y^n using (say) `zip`. Now, define $\bar{p}_1(y^n) = 2^{-L_{\mathtt{zip}}(y^n)}$. Via Kraft's inequality (Cover and Thomas, 1991) one can infer that $\sum_{y^n \in \{0,1\}^n} \bar{p}_1(y^n) \leq 1$ (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, $y^n \in \{0,1\}$, $p_0(y^n) = 2^{-n}$. Defining $S := \bar{p}_1(Y^n)/p_0(Y^n)$ we thus find

$$\mathbf{E}_{Y^n \sim P_0}[S] = \sum_{y^n \in \{0,1\}^n} \bar{p}_1(y^n) \leq 1 \quad ; \quad \log S = n - L_{\mathtt{zip}}(Y^n).$$

Thus, the Ryabko-Monarov codelength difference is the logarithm of an S-value. 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' $\bar{p}_1$ is true (if one insists on there being an alternative, one could view $\bar{p}_1$ as a representative of a nonparametric $\mathcal{H}_1$ consisting of *all* distributions P_1 with $\mathbf{E}_{Y^n \sim P_1}[\log S] > 0$, a truly huge and not so useful set).

More generally, by the same reasoning, for singleton $\mathcal{H}_0 = \{P_0\}$, any test statistic of the form $\bar{p}_1(Y^n)/p_0(Y^n)$, with p_0 the density of P_0 and $\bar{p}_1$ a density or sub-density (integrating to less than 1) is an S-value. Such S-values have been considered extensively within the *Minimum Description Length (MDL)* and *prequential* approaches to model selection (Rissanen, 1989, Dawid, 1997, Barron et al., 1998, Grünwald, 2007). In these approaches there usually is a clearly defined alternative $\mathcal{H}_1$, so that a Bayesian would choose $\bar{p}_1 := p_{W_1}$ to be a Bayes marginal density. In contrast, the MDL and prequential approach allow more freedom in the choice of $\bar{p}_1$. MDL merely requires $\bar{p}_1$ 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 S-values; but with composite $\mathcal{H}_0$, just as with Bayes factors, the standard MDL approach may fail to deliver S-values.

Future Work, Open Questions: Practical From a practical perspective, it is now important to come up with software for easy calculation of S-values; packages for the T-test and the 2×2-table are already under way. Also, with some of our GROW S-values, e.g. those for the 2×2 tables, we can only do optional continuation, not optional stopping. For obvious reasons (such as needing less data, see Section 5.2, and applications such as A/B-testing) we would like to have S-values that preserve Type-I error under optional stopping. Perhaps, even for 2×2 tables, this is possible by deviating slightly from our GROW optimality criterion? The following example suggests that optional stopping can sometimes be achieved under very mild conditions, and at the same time points towards the possibility of S-values for semiparametric tests:

Example 6. [Allard's Test Martingale] In his master's thesis, A. Hendriksen (2017) proposes a variation of the t-test when normality cannot be assumed: $\mathcal{H}_1$ represents the hypothesis that Y_i are i.i.d. with some mean μ, $\mathcal{H}_0$ represents the special case that $\mu = 0$; no further distributional assumptions are made. One can then coarsen the data by setting $f(Y_i) = 1$ iff $Y_i \geq 0$ and $f(Y_i) = -1$ otherwise, and then create an S-value on Y^n by setting $S(Y^n) = \frac{p_{W_1}(f(Y_1),\ldots,f(Y_n))}{p_0(f(Y_1,\ldots,Y_n))}$, where according to p_0, the data are i.i.d. Bernoulli(1/2), p_θ is the density of the data according to Bernoulli(θ), and p_{W_1} is the Bayes marginal obtained by putting a prior W_1 (say, uniform or Jeffreys') on θ. Then S is an S-value; in fact it even defines a test martingale, so Type-I error probability guarantees still hold under optional stopping, as long as the data are i.i.d. under *some* distribution with mean 0. A strict Bayesian would not be allowed to do such a coarsening of the data since it loses information (although the great Bayesian I.J. Good acknowledged that such an operation was often very useful, and that Bayesians really needed a *statistician's stooge* to prepare the data for them (Good (1983), in the Chapter titled *46656 Varieties of Bayesians*)). For the safe testing theory developed here, such a coarsening is not an issue at all. If one is unsure of normality, once may replace the default GROW t-test by Allard's martingale. The price one pays is, of course, that, for any

given $\mu > 0$, more data will tend to be needed to get substantial evidence against $\mu = 0$, i.e. a large S-value.

Similarly to this semiparametric case, it would be of major practical interest if we could use S-values even if Assumption A does not hold, as elaborated beneath Example 7; or if we could extend our ideas to confidence intervals, as elaborated at the end of the paper.

Future Work, Open Problems, Theoretical We showed here that Bayes factors based on the right Haar prior in the 1-sample t-test setting constitute GROW S-values. While Hendriksen et al. (2018) implies that using the right Haar prior in general group invariant situations always leads to S-values, the result that these are even GROW S-values is currently restricted to the t-test setting, with the Haar prior on the variance (scalar multiplication group). The only part of the proof that does not extend to the general group invariant setting is Lemma 2 in Appendix F; extending this is a major goal for future work. This would lead to an extension of Theorem 3 and 4 for general group invariant settings. But we want to go even further: Theorem 3 and 4 closely resemble Theorem 1, Part 3, but without the KL infimum being achieved. More generally, a major aim for future work is thus to provide a generalized (even beyond group invariant cases) version of Theorem 1 in which the KL infimum in the first argument is not necessarily achieved.

Finally, from a more conceptual perspective, it would be of major interest to establish the precise link between the present S-value theory and the BBW testing procedures referred to above. 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, 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.

7 A Theory of Hypothesis Testing

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 et al., 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 S-values and the corresponding safe tests can be fit within the three paradigms:

Concerning the *Neyman-Pearson approach*: S-values 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 S-values designed to do exactly this, as we illustrated in Section 5. The one difference to the NP approach is that we optimize power under the constraint that the S-value 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 3, the standard NP tests lead to useless S-values under the GROW criterion.

Concerning the *Fisherian approach*: we have seen that S-values can be reinterpreted as (quite) conservative P-values. But much more importantly within this discussion, S-values 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. Thus, in spirit of Fisher's philosophy, we can use S-values to determine whether there is substantial evidence against $\mathcal{H}_0$, without predetermining any significance level: we simply postulate that the larger S, the more evidence against $\mathcal{H}_0$ without having specific frequentist error guarantees. The major difference though is that these S-values 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* S-values that we encountered can also be written as likelihood ratios, although (e.g. in Example 5 or Section 4.4) 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) S-values we encountered are in fact equivalent to Bayes factors, and Theorem 1 Part 3 strongly suggests that this is a very general phenomenon. While the point priors arising in the default GROW S-values may be quite different from priors commonly adopted in the Bayesian literature, one can also obtain S-values by using priors on $\mathcal{H}_1$ that do reflect prior knowledge or beliefs — we elaborate on this under *Hope vs. Belief* below (note that the prior W_1 on $\mathcal{H}_1$ can be chosen completely freely, but the prior on $\mathcal{H}_0$ cannot: the S-value $S^*_{W_1}$ based on this prior is determined by the RIPr prior W^*_0 determined by W_1; see Theorem 1, Part 2).

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 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, for example, misspecification testing is an essential part of data analysis. Some insist that significance testing should be abolished altogether (Amrhein et al., 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 differing 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 S-values can provide: consider three tests with the same null hypothesis $\mathcal{H}_0$, based on samples $Z_{(1)}$, $Z_{(2)}$ and $Z_{(3)}$ respectively. The results of a default S-value test aimed to optimize power on sample $Z_{(1)}$, an S-value test for sample $Z_{(2)}$ based on a Bayesian prior W_1 on $\mathcal{H}_1$ and a Fisherian S-value 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 the prior on W_1 has $\theta \in \Theta_0 \cup \Theta_1$ independent of the stopping time N. In that case, a celebrated result going back to Barnard (1947) (see Hendriksen et al. (2018) 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 Bayes factors that correspond to default GROW S-values: these generally are 'local', the prior W_1(and potentially also W_0) depending on the sample size n (for example, for the 1-sided test with the normal location family, Example 4, we set all prior mass on $\tilde{\mu}_n = \sqrt{2(-\log \alpha)/n}$; a similar dependence holds for the prior on δ in the default GROW t-test). Thus, while from a purely Bayesian perspective such S-values/Bayes factors are not suitable for optional stopping, in Section 4, both the default GROW S-value 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 et al., 2018), as we did in the experiment reported in Figure 5: the pre-determined n (called there $n_{\max}$) on which the prior on δ is based is determined 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 $s(Y^T) := p_{W_1}(Y^T)/p_{W_0}(Y^T)$ remains an S-value even if $T = n' \neq n_{\max}$ or even if one stops at the first $T \leq n_{\max}$ such that $S(Y^T) \geq 20$. 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 S-value approach. The purely Bayesian approach relies on the *beliefs* being, in some sense, adequate. In the S-value based approach, one *can* use priors that represent subjective a priori assessments; for example, in the Bayesian t-test, instead of a Cauchy or a 2-point prior, one can use any prior W_1 on δ that is symmetric around 0 one likes, with any dependency on sample size n one likes, and still the resulting Bayes factor with the right Haar prior on σ will be a GROW S-value (Theorem 3). *If $\mathcal{H}_1$ is the case, and the data behave as one would expect according to the prior W_1*, then the S-value 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, $S(Y^n)$ will tend to be smaller — but it is still an S-value, hence leads to valid Type-I error guarantees and can be interpreted when multiplied across experiments. Thus, from the S-value 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 S-value 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 either $\mathcal{H}_0$ 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.

7.2 A Safer Form of Testing

We have already seen that some, but not all S-values allow not just for optional continuation but also for optional stopping, as long as Assumption A holds. One might also ask if we can still make valid inferences using (perhaps suitably modified) S-values if Assumption A is violated. Also, while the Type-I error guarantees based on safe tests are always valid under Assumption A, any guarantees for, e.g. expected growth rate or power under $\mathcal{H}_1$ rely on, e.g., the prior assumptions for $\mathcal{H}_1$ or the restriction to $\Theta_1(\delta)$ being correct. Thus, S-values and corresponding safe tests suggest a *range* of inferences, some of which are *safe* (i..e., valid) under quite weak additional assumptions, and some of which only become safe under much stronger additional assumptions. *One can envision a methodology of hypothesis testing in which researchers who base their test on a test statistic S are always required to state what inferences they consider 'safe' based on the test based on test statistic S, i.e. which of these additional assumptions hold.*

If S is actually an S-value, and if it is used in a context in which Assumption A holds, then it will automatically follow that the safe test on S achieves valid Type-I error probabilities bounds under optional continuation. Thus, what we previously called a 'safe test' is really a test that is *safe for Type-I Error Probability under optional continuation when evidence is combined by multiplication*, which we require as a minimal safety guarantee. But there are many other types of safety: for example a researcher may claim that the test $\mathrm{T}_\alpha(S)$ based on S is also *safe for Type-II Error probability.* By this we mean that she believes that the actual power of the test at the given sample size is at least $P_{W_1}(\mathrm{T}_\alpha = \textsc{reject}_0) = P_{W_1}(T \geq 1/\alpha)$. This could be the case if she is a Bayesian who is very convinced that, given Θ_1, W_1 is the appropriate prior (then she believes the power is in fact equal to $P_{W_1}(\mathrm{T}_\alpha = \textsc{reject}_0)$).

Another researcher might read about her results and be convinced that, in her testing scenario, Assumption A held (so that her results are indeed safe for Type-I error), but may have doubts about the prior W_1 — such a researcher would be happy to base Type-I error inferences based on the value of S but not Type-II error probability inferences. Note that, if $\mathcal{H}_1$ is a singleton, then any researcher who thinks the models $\mathcal{H}_0$ and $\mathcal{H}_1$ are well-specified and uses an S-value would be safe, under Assumption A, for both Type-I and Type-II error probability inferences. However, in this case ($\mathcal{H}_1$ singleton), yet another researcher might think that $\mathcal{H}_1$ is not even well-specified (as in Example 5, if $\mathcal{H}_1$ is identified with p_1 as defined there). If such a researcher still thinks that Assumption A holds, then he would consider inferences based on the proposed S-value still safe for Type-I error probabilities but not for the power, i.e. 'probability that $\mathcal{H}_1$ is correctly accepted' — the real probability that a decision to accept $\mathcal{H}_1$ is correct, is 0 according to such a researcher.

We can even go a step further and consider what happens in the case that Assumption A is violated, as happens all too often in the real world:

Example 7. Optional Starting and Double Use of Data In a 2004 (in)famous court case, Dutch nurse Lucia de Berk was convicted of killing several patients, based partly on statistical evidence. Deaths in her ward always occurred while she was on duty; colleagues become

suspicious and the public prosecutor had a statistician look into a matter. The statistician modeled the data as a 2×2 table and performed a null hypothesis test. $\mathcal{H}_0$ represented the hypothesis that it was all a coincidence, i.e. the probability of a patient dying at the jth shift ($Y_j = 1$) would be the same if Lucia were present ($X_j = a$) than if she were not present ($X_j = b$). Fisher's exact test gave a p-value of 1 in 342 million. Of course, a cardinal sin of traditional hypothesis testing was committed here: the data that suggested a hypothesis was used to test the hypothesis itself, leading to wild overestimation of effects.

How is this related to our work? Until now, we only thought of Assumption A in terms of optional *continuation*: we can only use S-values in a context with optional continuation if the decision to continue to do a test on a new sample does not depend on that new sample itself. But Assumption A also rules out *optional starting*, the extreme case of optional stopping at time 1, which arises at the start of the *first* test: Assumption A requires that we only start using S-values *at all* if the decision to calculate the first S-value is independent on the data on which this S-value is based. Thus, if, as a general rule, data analysts would check whether the context in which they do their test is such that Assumption A holds, then the error above is automatically avoided. Assumption A is just a natural extension of the rule 'don't use the data that suggested a hypothesis to test that hypothesis'.

Even though in the example above, Assumption A does not hold, one would still like to be able to say *something* about the strength of evidence of the nurse data. Under strict Bayesian assumptions, one can: one assigns prior $\pi(\mathcal{H}_0) = 1 - \pi(\mathcal{H}_1)$ to the hypotheses, and equips Θ_0 and Θ_1 with priors W_0 and W_1 respectively. As Bayesians, we might reject $\mathcal{H}_0$ if the posterior probability of $\mathcal{H}_0$ given data Z is no larger than some α, i.e.

$$\pi(\mathcal{H}_0 \mid Z) \propto p_{W_0}(Z)\pi(\mathcal{H}_0) \leq \alpha, \tag{28}$$

while, if we only decide to do a test if Z takes values within a certain set $\mathcal{E}$ (e.g. $\mathcal{E}$ contains just those Z for which $\pi(cH_0 \mid Z)$ is smaller than some ϵ), we should really be rejecting if

$$\pi(\mathcal{H}_0 \mid Z, Z \in \mathcal{E}) \leq \alpha.$$

However, for all $Z \in \mathcal{E}$, the two probabilities on the left are in fact equal — since given such a Z, we already know that it is in $\mathcal{E}$. Since a Bayesian conditions on all the data, the fact that $Z \in \mathcal{E}$ is irrelevant for the posterior given Z, and one may say that, assuming the priors are trustworthy, the posterior is trustworthy as well, so that (28) gives the correct conditional Type-I error probability. (in the nurse case, a logical choice for $\mathcal{H}_1$ would be 'nurse murders patients', the prior on which, although one cannot be sure of the precise numerical value, can safely be taken to be very small, making the evidence far weaker — indeed it is now commonly accepted that the nurse was innocent and the deaths were just accidents).

Thus, we may say that *assuming the priors on $\mathcal{H}_0$ and $\mathcal{H}_1$ and the priors W_1 and W_2 are correct, Bayesian inference is Safe for Type-I error probability under optional starting even if Assumption A does not hold.* Similarly, again assuming that all priors are correctly specified, Bayesian inference is safe for Type-I and Type-II error probability under optional stopping and optional continuation. The idea of S-values is to get safety guarantees that still hold without assuming that (all) of the specified priors are correct. An interesting and intriguing question for future work is the case where we may be able to specify (bounds on) priors $\pi(\mathcal{H}_0)$ and $\pi(\mathcal{H}_1)$, but are hard pressed to subjectively yet thrustworthily assess the in-model prior

densities W_0 and/or W_1. For certain types of S-values, it may be possible to replace the Bayes factors in Bayes' theorem by these S-values and still get valid Type-I error guarantees, even though we were not able to specify 'subjective' W_0 and W_1.

This discussion of various forms of 'safety' is admittedly somewhat vague. A first stab at making it precise, in which the focus is on posterior probabilities rather then test decisions, is made in the paper *safe probability* written by one of us (Grünwald, 2018), based on earlier ideas from Grünwald (2000).

7.3 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 S-value based approach isn't 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 S-values, the S-values are easier to explain, via the direct analogy to gambling. Also, we suggested 'default' S-values 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 et al. (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, S-values 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 S-values, even though our comparison to Johnson's uniformly most powerful Bayes tests in Section 4 and the experiments in Section 5 might perhaps suggest this. Rather, our goal is to advocate using GROW S-values 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 (S-values 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 behaviour 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*. To illustrate, consider the following scenario: suppose that one estimates a parameter θ based on initial sample $Z_{(1)}$ and the result is promising but inconclusive – for example, the minimum interesting effect size is 2, the estimate $\hat{\theta}$ was substantially greater than 2 but the left end of the confidence interval was below 2. Thus, because the result is promising, one gathers a second batch of data $Z_{(2)}$. Now if one re-calculates $\hat{\theta}$ based on the joint data $(Z_{(1)}, Z_{(2)})$, the corresponding confidence interval, when calculated in the standard manner, will not be valid any more. These confidence intervals somehow have to be adjusted, or calculated differently. Thus, rather than criticizing confidence interval-based approaches, we would rather like to argue that they do have their uses, but they, too, should be transformed to a novel type of confidence bounds that are *safe under optional continuation*. Developing such safer confidence methods is a major goal for future research.

Acknowledgments Thanks to Andrew Barron, Jim Berger, Peter Harremoes, Alexander Ly, Ronald Meester, Judith ter Schure, Teddy Seidenfeld, Glenn Shafer, Rosanne Turner, Volodya Vovk and Eric-Jan Wagenmakers for many helpful conversations.

References

Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance, 2019.

George A Barnard. Review of *sequential analysis* by Abraham Wald. *Journal of the American Statistical Association*, 42(240), 1947.

O.E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, UK, 1978.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998. Special Commemorative Issue: Information Theory: 1948-1998.

Maria J Bayarri, James O Berger, Anabel Forte, G García-Donato, et al. Criteria for Bayesian model choice with application to variable selection. *The Annals of statistics*, 40(3):1550–1577, 2012.

MJ Bayarri, Daniel J Benjamin, James O Berger, and Thomas M Sellke. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72:90–103, 2016.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6, 2018.

J. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–12, 2003.

James O Berger, Luis R Pericchi, and Julia A Varshavsky. Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 307–321, 1998.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, revised and expanded 2nd edition, 1985.

J.O. Berger, L.D. Brown, and R.L. Wolpert. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Annals of Statistics*, 22(4):1787–1807, 1994.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

G. Cumming. *Understanding the New Statistics: Effect Sizes, Confidence and Meta-Analysis*. Routledge, 2012.

Sarat C. Dass and James O. Berger. Unified conditional frequentist and Bayesian testing of composite hypotheses. *Scandinavian Journal of Statistics*, 30(1):193,Äì210, Mar 2003. ISSN 1467-9469. doi: 10.1111/1467-9469.00326. URL `http://dx.doi.org/10.1111/1467-9469.00326`.

A.P Dawid. Prequential analysis. In S. Kotz, C.B. Read, and D. Banks, editors, *Encyclopedia of Statistical Sciences*, volume 1 (Update), pages 464–470. Wiley-Interscience, New York, 1997.

M.L. Eaton. *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics and American Statistical Association, 1989.

Peter Gács. Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science*, 341(1-3):91–137, 2005.

I.J. Good. *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press, 1983.

P. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

P. D. Grünwald. Maximum entropy and the glasses you are looking through. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pages 238–246, San Francisco, 2000. Morgan Kaufmann.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.

P.D. Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 2018.

Peter Grünwald and Nishant Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *Journal of Machine Learning Research*, 2019. to appear.

E. Gunel and J. Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61(3): 545–557, 1974.

A. Hendriksen, R. de Heide, and P.D. Grünwald. Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *arXiv preprint arXiv:1807.09077*, 2018.

Allard A Hendriksen. Betting as an alternative to p-values. Master's thesis, Leiden University, Dept. of Mathematics, 2017.

T. Jamil, A. Ly, R.D. Morey, J. Love, M. Marsman, and E-J. Wagenmakers. Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods*, pages 1–15, 2016.

H. Jeffreys. *Theory of Probability*. Oxford University Press, London, 3rd edition, 1961.

Ramesh Johari, Leo Pekelis, and David J Walsh. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

Valen E Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013a.

Valen E Johnson. Uniformly most powerful bayesian tests. *Annals of statistics*, 41(4):1716, 2013b.

J.L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, pages 917–926, 1956.

J. Kiefer. Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 72(360):789–808, 1977.

T.L. Lai. Martingales in sequential analysis and time series, 1945–1985. *Electronic Journal for History of Probability and Statistics*, 5(1), 2009.

Leonid A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17(2):337–340, 1976.

J.Q. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, New Haven, CT, 1999.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 2008.

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh ACM Conference on Computational Learning Theory (COLT' 98)*, pages 230–234. ACM Press, 1998.

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245, 2019.

E Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Hackensack, NJ, 1989.

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2): 225,Äì237, Apr 2009. ISSN 1531-5320. doi: 10.3758/pbr.16.2.225. URL `http://dx.doi.org/10.3758/pbr.16.2.225`.

R. Royall. On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 2000.

B Ya Ryabko and VA Monarev. Using information theory approach to randomness testing. *Journal of Statistical Planning and Inference*, 133(1):95–110, 2005.

T. Seidenfeld. Personal communication, 2016.

Thomas Sellke, MJ Bayarri, and James O Berger. Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

G. Shafer. The language of betting as a strategy for statistical and scientific communication, 2019. working paper, available at `http://probabilityandfinance.com/articles/index.html`.

G. Shafer and V. Vovk. *Game-Theoretic Probability: Theory and Applications to Prediction, Science and Finance*. Wiley, 2019.

Glenn Shafer and Vladimir Vovk. *Probability and Finance – It's Only a Game!* Wiley, New York, 2001.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, pages 84–101, 2011.

J. Shawe-Taylor and R.C. Williamson. A PAC analysis of a Bayesian classifier. In *Proceedings of the Tenth ACM Conference on Computational Learning Theory (COLT' 97)*, pages 2–9, Nashville, Tennessee, 1997.

Tim Van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

J. Ville. Étude critique de la notion de collectif. *Monographies des Probabilités*, 3, 1939.

V.G. Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society, series B*, 55:317–351, 1993. (with discussion).

P. Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1991.

Ronald L Wasserstein, Nicole A Lazar, et al. The ASAs statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

R.L. Wolpert. Testing simple hypotheses. In H.H. Bock and W. Polasek, editors, *Data Analysis and Information Systems: Statistical and Conceptual Approaches*, pages 289–297. Springer, Berlin, 1996.

Appendix A Proof Preliminaries

In the next sections we will prove our theorems. To make all statements in the main text mathematically rigorous and to ensure that all notations in the main text are mutually compatible, we need to provide a few additional definitions and some notation first.

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 measure space $\Omega = \mathcal{Y}^m \times \mathcal{R}^m$ for some large but finite m, where $\mathcal{Y}$, $\mathcal{R}$ are finite-dimensional vector spaces. Whenever we speak about a distribution on Ω or $\mathcal{Y}^m$, we implicitly take its associated σ-algebra to be the Borel σ-algebra for Ω or $\mathcal{Y}^m$, respectively. For each $\theta \in \Theta \coloneqq \Theta_0 \cup \Theta_1$, P_θ is a distribution on $\mathcal{Y}^m$. Thus, unless $\mathcal{R} = \{0\}$, P_θ does not necessarily define a distribution on the full space Ω. We let Y_i be the random vector defined by $Y_i(\omega) = y_i$ when $\omega = (y_1, \ldots, y_m, r_1, \ldots, r_m)$ and similarly $R_i(\omega) = r_i$.

In Section 2 we group outcomes in 'batches' or 'samples', $Z_{(1)} = (Y_1, \ldots, Y_{n_1})$, $Z_{(2)} = (Y_{n_1+1}, \ldots, Y_{n_2})$ and so on, where $n_1 \leq n_2 \leq n_{k_{\max}} \leq m$ are fixed in advance; similarly for $V_{(1)} = (R_1, \ldots, R_{n_1})$, $V_{(2)} = (R_{n_1+1}, \ldots, R_{n_2})$, ... In other sections, we focus on a single batch $Z = (Y_1, \ldots, Y_n)$. We can make this compatible with the grouping in Section 2 by setting $Z = Z_{(1)}$ and $n = n_1$. Whenever we refer to a random variable such as Y without giving an index, it stands for $Y^n = (Y_1, \ldots, Y_n)$; similarly for all other time-indexed random variables.

In some applications though (all marked in the text) we set $Z_{(j)} = Y_j$, thus each outcome is 'its own group'. We could extend this setting, using measure theory, to settings in which the $n_1, n_2, \ldots$ are themselves random stopping times, but for simplicity will not do so here.

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.

Densities Throughout the text, we assume that the Y_i are independent and that for each $\theta \in \Theta$, 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. Recall that invariably Z stands for n outcomes, i.e. $Z = (Y_1, \ldots, Y_n)$ for some $n \leq m$ with m defined as above. Thus, $p_\theta(Z) = \prod_{i=1}^{n} p'_{\theta,i}(Y_i)$ 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 4.3 and the conditional exponential family setting that we briefly mentioned in Section 4.4, we assume that the Y_i are not just independent but also identically distributed, hence $p'_{\theta,i} = p'_{\theta,1}$ for all i. To bring the contingency table and general conditional exponential family setting into our realm, we assume that there is an additional set $\mathcal{X}$ and a fixed vector $(x_1, \ldots, x_m)$, with each $x_i \in \mathcal{X}$. Then $p'_{\theta,i}(y) := p_\theta(y \mid x)$, where, for each $\theta \in \Theta$ and each $x \in \mathcal{X}$, $p_\theta(\cdot \mid x)$ is a density on $\mathcal{Y}$ relative to the same underlying measure λ_1, so that, for each θ, the density of distribution P_θ for $Z = (Y_1, \ldots, Y_n)$ is given by $p_\theta(Y_1, \ldots, Y_n) := p_\theta(Y_1, \ldots Y_n \mid x_1, \ldots, x_n) := \prod_{i=1}^{n} p_\theta(Y_i \mid x_i)$. Since we regard $x_1, \ldots, x_n$ as fixed in advance, we can write, if so inclined, P_θ without conditioning on $x_1, \ldots, x_n$.

Notational Conventions When we mention a distribution P without further qualification, we mean that it is the distribution of the random variable Z defined on Ω. When P is defined on a different random variable U, we write $P[U]$ instead and (unless explicitly stated otherwise) we denote its density by $p[U]$. Similarly, when we mention a distribution W_j without further qualification, we mean that it is a ("prior") distribution on the parameter set Θ_j. In case $\Theta = \Delta \times \Gamma$, $W = W[\Theta]$ denotes a distribution on Θ, and distributions on Δ and Γ are denoted as $W[\delta]$ and $W[\gamma]$ respectively.

A *test statistic* S is by definition a random variable that can itself be written as a function of the data Z. If S can be written as a function of another random variable V (that is itself determined by Z) we write it as $S\langle V \rangle$. Since an S-value is just a test statistic that satisfies some additional constraints, each S-value can also be written as $S\langle Z \rangle$, and the S-value appearing in (21) is written as $S\langle V \rangle$ because it can be written as $S = s(V)$ for some function s. All these definitions are extended to test statistics $S_{(k)}$; those can, by definition, be written as functions of the k-th data sample $Z_{(k)}$.

Appendix B Proofs for Section 2

We will prove Proposition 3, which generalizes Proposition 2.

Proof of Proposition 3 We will only show the result for $S^{(K_{\text{STOP}})}$, the result for $S^{(k)}$ with fixed k being easier. All expectations below are under arbitrary P_θ with $\theta \in \Theta_0$. For $k = 1, \ldots, k_{\max}$, we define $S'_{(k)} := s'_{(k)}(Z_{(k)}, U^{(k)}, B_{(k)})$ by $s'_{(k)}(Z_{(k)}, U^{(k)}, B_{(k)}) = s_{(k)}(Z_{(k)}, U^{(k)})$ if $B_{(k)} = \text{CONTINUE}$ and $s'_{(k)}(Z_{(k)}, U^{(k)}, B_{(k)}) = 1$ if $B_{(k)} = \text{STOP}$. Since we assume that $S_{(k)}$ is an S-value conditional on $U^{(k)}$, i.e. $\mathbf{E}[S_{(k)} \mid U^{(k)}] \leq 1$, and since $B_{(k)} = b_{(k)}(U^{(k)})$ is determined by $U^{(k)}$, we must also have that

$$\mathbf{E}[S'_{(k)} \mid U^{(k)}, B_{(k)}] = \mathbf{E}[s'_{(k)}(Z_{(k)}, U^{(k)}, B_{(k)}) \mid U^{(k)}, B_{(k)}] \leq 1, \tag{29}$$

i.e. $S'_{(k)}$ is an S-value conditional on $(U^{(k)}, B_{(k)})$.

We thus have, under Assumption A, letting $\bar{k} = k_{\max}$,

$$\begin{aligned}
& \mathbf{E}\left[S^{(K_{\text{STOP}})}\right] \\
= \; & \mathbf{E}_{Z^{(\bar{k})}}\left[\mathbf{E}_{U^{(\bar{k})}|Z^{(\bar{k})}}\left[\mathbf{E}_{B^{(\bar{k})}|Z^{(\bar{k})},U^{(\bar{k})}}\left[S^{(K_{\text{STOP}})}\right]\right]\right] \\
= \; & \mathbf{E}_{Z^{(\bar{k})}}\Big[\mathbf{E}_{U_{(1)}|Z^{(\bar{k})}}\mathbf{E}_{U_{(2)}|U_{(1)},Z^{(\bar{k})}} \ldots \mathbf{E}_{U_{(\bar{k})}|U^{(\bar{k}-1)},Z^{(\bar{k})}}\big[\\
& \quad \mathbf{E}_{B_{(1)}|Z^{(\bar{k})},U^{(\bar{k})}}\mathbf{E}_{B_{(2)}|Z^{(\bar{k})},U^{(\bar{k})},B_{(1)}} \ldots \mathbf{E}_{B_{(\bar{k})}|Z^{(\bar{k})},U^{(\bar{k})},B^{(\bar{k}-1)}}\left[S^{(K_{\text{STOP}})}\right]\big]\Big] \\
= \; & \mathbf{E}_{Z_{(1)}} \ldots \mathbf{E}_{Z_{(\bar{k})}}\Big[\mathbf{E}_{U_{(1)}}\mathbf{E}_{U_{(2)}|U_{(1)},Z_{(1)}} \ldots \mathbf{E}_{U_{(\bar{k})}|U^{(\bar{k}-1)},Z^{(\bar{k}-1)}}\Big[\\
& \quad \mathbf{E}_{B_{(1)}}\mathbf{E}_{B_{(2)}|Z_{(1)},U_{(1)},B_{(1)}} \ldots \mathbf{E}_{B_{(\bar{k})}|Z^{(\bar{k}-1)},U^{(\bar{k}-1)},B^{(\bar{k}-1)}}\left[S^{(K_{\text{STOP}})}\right]\Big]\Big] \\
= \; & \mathbf{E}_{U_{(1)}}\mathbf{E}_{B_{(1)}}\mathbf{E}_{Z_{(1)}}\Big[\mathbf{E}_{U_{(2)}|U_{(1)},Z_{(1)}}\mathbf{E}_{B_{(2)}|Z_{(1)},U_{(1)},B_{(1)}}\mathbf{E}_{Z_{(2)}}\Big[\ldots \\
& \quad \ldots \mathbf{E}_{U_{(\bar{k})}|U^{(\bar{k}-1)},Z^{(\bar{k}-1)}}\mathbf{E}_{B_{(\bar{k})}|Z^{(\bar{k}-1)},U^{(\bar{k}-1)},B^{(\bar{k}-1)}}\mathbf{E}_{Z_{(\bar{k})}}\left[\prod_{k=1}^{(\bar{k})} s'_{(k)}(Z_{(k)}, U^{(k)}, B_{(k)})\right] \ldots\Big]\Big] \\
= \; & \mathbf{E}_{U_{(1)}}\mathbf{E}_{B_{(1)}}\left[\mathbf{E}_{Z_{(1)}}\left[s'_{(1)}(Z_{(1)})\right] \cdot \mathbf{E}_{U_{(2)}|U_{(1)},Z_{(1)}}\mathbf{E}_{B_{(2)}|Z_{(1)},U_{(1)},B_{(1)}}\left[\mathbf{E}_{Z_{(2)}}\left[s'_2(Z_{(2)}, U^2, B_{(2)})\right] \cdot \ldots\right.\right. \\
& \quad \left.\left.\mathbf{E}_{U_{(\bar{k})}|U^{(\bar{k}-1)},Z^{(\bar{k}-1)}}\mathbf{E}_{B_{(\bar{k})}|Z^{(\bar{k}-1)},U^{(\bar{k}-1)},B^{(\bar{k}-1)}}\left[\mathbf{E}_{Z_{(\bar{k})}}\left[s'_{(\bar{k})}(Z_{(\bar{k})}, U^{(\bar{k})}, B_{(\bar{k})})\right]\right] \ldots\right]\right] \\
\leq \; & \mathbf{E}_{U_{(1)}}\mathbf{E}_{B_{(1)}}\left[\mathbf{E}_{Z_{(1)}}[1] \cdot \mathbf{E}_{U_{(2)}|U_{(1)},Z_{(1)}}\mathbf{E}_{B_{(2)}|Z_{(1)},U_{(1)},B_{(1)}}\left[\mathbf{E}_{Z_{(2)}}[1] \cdot \ldots\right.\right. \\
& \quad \left.\left.\mathbf{E}_{U_{(\bar{k})}|U^{(\bar{k}-1)},Z^{(\bar{k}-1)}}\mathbf{E}_{B_{(\bar{k})}|Z^{(\bar{k}-1)},U^{(\bar{k}-1)},B^{(\bar{k}-1)}}\left[\mathbf{E}_{Z_{(\bar{k})}}[1]\right] \ldots\right]\right] = 1.
\end{aligned}$$

where we rearranged expectations using our conditional independence assumptions and Tonnelli's theorem. We then used that, by definition of S', we must have $S^{(K_{\text{STOP}})} = \prod_{k=1}^{k_{\max}} S'_{(k)}$ with $S'_{(k)} = s'_{(k)}(Z_{(k)}, U^{(k)}, B_{(k)})$, and, finally, we used (29), i.e. that for each fixed k, $S'_{(k)}$ is an S-value conditional on $U^{(k)}, B_{(k)}$.

Appendix C Elaborations and Proofs for Section 3

Meaning of "S^* as defined by achieving (11) is essentially unique" Consider $\Theta'_1 \subset \Theta_1$ and Θ_0, as in the main text in Section 3. Suppose that there exists an S-value S^* achieving

the infimum in (11). We say that S^* is *essentially unique* if for any other S-value S° achieving the infimum in (11), we have $P_\theta(S^* = S^\circ) = 1$, for all $\theta \in \Theta_1' \cup \Theta_0$. Thus, if the GROW S-value exists and is essentially unique, any two GROW S-values will take on the same value with probability 1 under all hypotheses considered, and then we can simply take one of these GROW S-values and consider it the 'unique' one.

C.1 Proof of Theorem 1

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 $0 < Q(\mathcal{Y}^m) \leq 1$. 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 1. *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*

$$D(P\|Q^\circ) = \inf_{Q\in\mathcal{Q}} D(P\|Q), \tag{30}$$

   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*

$$\mathbf{E}_{Z\sim Q}\left[\frac{p(Z)}{q^\circ(Z)}\right] \leq 1. \tag{31}$$

   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 (30) is achieved by Q_0 in $\mathcal{Q} \Leftrightarrow Q^\circ = Q_0 \Leftrightarrow$ (31) 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

$$\text{for all } q \text{ that are densities of some } Q \in \mathcal{Q}\text{: } \mathbf{E}_{Z\sim P}\left[\frac{q(Z)}{q^\circ(Z)}\right] \leq 1. \tag{32}$$

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 (32) as:

$$\text{for all } Q \in \mathcal{Q}\text{: } \mathbf{E}_{Z\sim Q}\left[\frac{p(Z)}{q^\circ(Z)}\right] \leq 1.$$

Li's Theorem 4.3 still allows for the possibility that $\int q^\circ(z)d\lambda(z) > 1$. 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 1 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 1, 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 $S^*_{W_1} := p_{w_1}(Z)/q^\circ(Z)$ is an S-value, and that it satisfies

$$\mathbf{E}_{P_{W_1}}[\log S^*_{W_1}] = \mathbf{E}_{P_{W_1}}\left[\log \frac{p_{W_1}(Z)}{q^\circ(Z)}\right] = D(P_{W_1}\|Q^\circ) = \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}),$$

where the second equality is immediate and the third is from (30). It only remains to show that (a)

$$\sup_{S \in \mathcal{S}(\Theta_0)} \mathbf{E}_{Z \sim P_{W_1}}[\log S] \leq \mathbf{E}_{P_{W_1}}[\log S^*_{W_1}]$$

and (b) that $S^*_{W_1}$ is essentially unique. To show (a), fix any S-value $S = s(Z)$ in $\mathcal{S}(\Theta_0)$. Now further fix $\epsilon > 0$ and fix a $W_{(\epsilon)} \in \mathcal{W}(\Theta_0)$ with $D(P_{W_1}\|P_{W_{(\epsilon)}}) \leq \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) + \epsilon$. We must have, with $q(z) := s(z)p_{W_{(\epsilon)}}(z)$, that $\int q(z)d\lambda = \mathbf{E}_{Z \sim P_{W_{(\epsilon)}}}[S] \leq 1$, so q is a sub-probability density, and by the information inequality of information theory (Cover and Thomas, 1991), it follows:

$$\mathbf{E}_{P_{W_1}}[\log S] = \mathbf{E}_{P_{W_1}}\left[\log \frac{q(Z)}{p_{W_{(\epsilon)}}(Z)}\right] \leq$$
$$\mathbf{E}_{P_{W_1}}\left[\log \frac{p_{W_1}(Z)}{p_{W_{(\epsilon)}}(Z)}\right] = D(P_{W_1}\|P_{W_{(\epsilon)}}) \leq \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) + \epsilon.$$

Since we can take ϵ to be arbitrarily close to 0, it follows that

$$\mathbf{E}_{P_{W_1}}[\log S] \leq \inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1}\|P_{W_0}) = \mathbf{E}_{P_{W_1}}[\log S^*_{W_1}],$$

where the latter equality was shown earlier. This shows (a).

To show essential uniqueness, let S be any S-value with $\mathbf{E}_{P_{W_1}}[\log S] = \mathbf{E}_{P_{W_1}}[\log S^*_{W_1}]$. By linearity of expectation, $S' = (1/2)S^*_{W_1} + (1/2)S$ is then also an S-value, and by Jensen's inequality applied to the logarithm we must have $\mathbf{E}_{P_{W_1}}[\log S'] > \mathbf{E}_{P_{W_1}}[\log S^*_{W_1}]$ unless $P_{W_1}(S = S^*_{W_1}) = 1$. Since we have already shown that for any S-value S', $\mathbf{E}_{P_{W_1}}[\log S'] \leq \mathbf{E}_{P_{W_1}}[\log S^*_{W_1}]$, it follows that $P_{W_1}(S \neq S^*_{W_1}) = 0$. But then, by our assumption of absolute continuity, we also have $P_{\theta_0}(S \neq S^*_{W_1}) = 0$ so $S^*_{W_1}$ is essentially unique.

Proof of Theorem 1, 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 1, Part 3 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 $L(z,p) := -\log p(z)/p_{W_0^*}(z)$, 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), $H_{W_0^*}(P) := \mathbf{E}_{Z\sim P}[L(Z,p)] = -D(P\|P_{W_0^*})$, the standard definition of *entropy relative to* $P_{W_0^*}$, is well-defined and 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 $W_1 \in \mathcal{W}_1'$, $H_{W_0^*}(P_{W_1}) = -D(P_{W_1}\|P_{W_0^*}) \le 0$.

Theorem 6.3 of Grünwald and Dawid (2004) together with Lemma 4.1 of that same paper then gives

$$H_{W_0^*}(P_{W_1^*}) = \sup_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}\left[-\log \frac{p_W(Z)}{p_{W_0^*}(Z)}\right] = \sup_{W\in\mathcal{W}_1'} \inf_{q\in\mathcal{Q}_{\text{DENS}}} \mathbf{E}_{Z\sim P_W}\left[-\log \frac{q(Z)}{p_{W_0^*}(Z)}\right]$$
$$= \inf_{q\in\mathcal{Q}_{\text{DENS}}} \sup_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}\left[-\log \frac{q(Z)}{p_{W_0^*}(Z)}\right] = \sup_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}\left[-\log \frac{p_{W_1^*}(Z)}{p_{W_0^*}(Z)}\right], \quad (33)$$

where, to be more precise, the first equality is immediate from the fact that $-H_{W_0^*}(P_{W_1^*}) = D(P_{W_1^*}\|P_{W_0^*}) = \inf_{W_1\in\mathcal{W}_1'} D(P_{W_1}\|P_{W_0^*})$; 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 (33) gives, with $S^* = \frac{p_{W_1^*}(Z)}{p_{W_0^*}(Z)}$, that

$$D(P_{W_1^*}\|P_{W_0^*}) = \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S^*] = \sup_{q\in\mathcal{Q}_{\text{DENS}}} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}\left[\log \frac{q(Z)}{p_{W_0^*}(Z)}\right] =$$
$$\sup_{S\in\mathcal{S}(\{W_0^*\})} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S], \quad (34)$$

where the first equality follows because the first and last terms in (33) must be equal, using again that $H_{W_0^*} = -D(\cdot\|P_{W_0^*})$, the second equality follows from the final equality in (33),

and the third equality follows by noting, first, that without loss of generality we can restrict the supremum over $S \in \mathcal{S}(\{W_0^*\}$ with $\mathbf{E}_{P_{W_0^*}}[S] = 1$; and second, that for every such S-value $S = s(Z)$ relative to $\{W_0^*\}$ we can define $q(Z) \coloneqq s(Z)p_{W_0^*}(z)$ and then $\int q(z)d\lambda(z) = \mathbf{E}_{Z\sim P_{W_0^*}}[S] = 11$, so there is a probability density $q \in \mathcal{Q}_{\text{DENS}}$ such that $S = q/p_{W_0^*}$; conversely, for every $q \in \mathcal{S}_{\text{DENS}}$, $S \coloneqq q/p_{W_0^*}$ is an S-value in $\mathcal{S}(\{W_0^*\})$.

If we could replace $\mathcal{S}(\{W_0^*\})$ in (34) by $\mathcal{S}(\Theta_0)$, then (14) would follow and we would be done. But we can achieve this by noting that

$$\sup_{S\in\mathcal{S}(\{W_0^*\})} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S] \geq \sup_{S\in\mathcal{S}(\mathcal{W}(\Theta_0))} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S] =$$

$$\sup_{S\in\mathcal{S}(\Theta_0)} \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S] \geq \inf_{W\in\mathcal{W}_1'} \mathbf{E}_{Z\sim P_W}[\log S^*] = D(P_{W_1^*}\|P_{W_0^*}), \tag{35}$$

where the first inequality follows because, as is immediate from the definition of S-value, the set of S-values relative to any set $\mathcal{W}' \subset \mathcal{W}$ must be a superset of the set of S-values relative to $\mathcal{W}$. The equality follows by linearity of expectation and the definition of S-value. The second inequality follows because $S^* \in \mathcal{S}(\Theta_0)$, as is implied by Theorem 1, Part 2, applied with $W_1 \coloneqq W_1^*$, and the final equality is the first equality of (34) again. Together (35) and (34) imply the required result (14).

Appendix D Proofs for Section 4.1

(18) is a consequence of the following proposition:

Proposition Let $\Theta_0 = \{0\}$, let $\Theta(\delta)$ be defined as in (16) 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). We then have for all $\theta \in \text{BD}(\Theta(\delta))$,

$$\mathbf{E}_{Z\sim P_\theta}\left[\log \frac{p_{W_1^*}(Z)}{p_0(Z)}\right] \geq \mathbf{E}_{Z\sim P_{W_1^*}}\left[\log \frac{p_{W_1^*}(Z)}{p_0(Z)}\right] = D(P_{W_1^*}\|P_0). \tag{36}$$

Now, suppose further that

$$\inf_{\theta\in\Theta(\delta)} \mathbf{E}_{Z\sim P_\theta}\left[\log \frac{p_{W_1^*}(Z)}{p_0(Z)}\right] = \inf_{\theta\in\text{BD}(\Theta(\delta))} \mathbf{E}_{Z\sim P_\theta}\left[\log \frac{p_{W_1^*}(Z)}{p_0(Z)}\right]. \tag{37}$$

Then for all $W_1 \in \mathcal{W}(\Theta(\delta))$, we also have $D(P_{W_1}\|P_0) \geq D(P_{W_1^*}\|P_0)$.

Proof. (36) is immediate from Theorem 1, 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

$$\mathbf{E}_{Z\sim P_{W_1}}\left[\log\frac{p_{W_1^*}(Z)}{p_0(Z)}\right] - \mathbf{E}_{Z\sim P_{W_1^*}}\left[\log\frac{p_{W_1^*}(Z)}{p_0(Z)}\right] \geq \mathbf{E}_{Z\sim P_{W_1}}\left[\log\frac{p_{W_1^*}(Z)}{p_0(Z)}\right] - \inf_{\theta\in \text{BD}(\Theta(\delta))}\mathbf{E}_{Z\sim P_\theta}\left[\log\frac{p_{W_1^*}(Z)}{p_0(Z)}\right], \quad (38)$$

where the first expression is just differentiation and the inequality follows from (36). so if we can show that, no matter how W_1 was chosen, the right-hand side of (38) is nonnegative, we must have $f(1) \geq f(0)$ and the desired result follows. But nonnegativity of (38) follows by the premise (37) and linearity of expectation. □

We need to prove (18) from the main text. (18) immediately follows from the proposition above if we can prove that (37) holds for $\delta = \underline{\delta}$, with the 1-dimensional $\Theta(\delta) = \Theta(\underline{\delta})$ under consideration. But this is straightforward: in this case $\text{BD}(\Theta(\underline{\delta}))$ is a singleton, so W_1^* is the degenerate distribution putting all mass on $\underline{\delta}$, and the right-hand side of (37) is just $D(P_{\underline{\delta}}\|P_0)$ whereas the θ on the left-hand side must satisfy $\theta \geq \underline{\delta}$. Without loss of generality, we may assume that $\{P_\theta : \theta \in \Theta\}$ is given in the canonical parameterization, so that $p_\theta(z) = \exp(\theta\phi(z))p_0(z)Z^{-1}(\theta)$, for some function ϕ. We can then write

$$\mathbf{E}_{Z\sim P_\theta}\left[\log\frac{p_{W_1^*}(Z)}{p_0(Z)}\right] = \underline{\delta}\cdot\mathbf{E}_{Z\sim P_\theta}\left[\phi(Z)\right] - \log Z(\underline{\delta}).$$

Since for general exponential families, $\mathbf{E}_{Z\sim P_\theta}\left[\phi(Z)\right]$ is an increasing function in θ, (37) and thus also (18) follows.

Relating $S^\circ_{\Theta(\underline{\delta})}$ and $S^*_{\Theta(\underline{\delta})}$ in the two-sided case We have, on all samples, $\log S^\circ_{\Theta(\underline{\delta})} \geq \max\{\log(1/2)S^*_{\underline{\delta}}, \log(1/2)S^*_{-\underline{\delta}}\}$, so that

$$\begin{aligned}
&\inf_{\theta:|\theta|\geq\underline{\delta}}\mathbf{E}_{Z\sim P_\theta}[\log S^\circ_{\Theta(\underline{\delta})}] \geq \inf_{\theta:|\theta|\geq\underline{\delta}}\max\{\mathbf{E}_{Z\sim P_\theta}[\log\frac{1}{2}S^*_{\underline{\delta}}], \mathbf{E}_{Z\sim P_\theta}[\log\frac{1}{2}S^*_{-\underline{\delta}}]\} \geq \\
&\max\{\inf_{\theta:|\theta|\geq\underline{\delta}}\mathbf{E}_{Z\sim P_\theta}[\log\frac{1}{2}S^*_{\underline{\delta}}], \inf_{\theta:|\theta|\geq\underline{\delta}}\mathbf{E}_{Z\sim P_\theta}[\log\frac{1}{2}S^*_{-\underline{\delta}}]\} \geq \\
&\max\{\inf_{\theta:\theta\geq\underline{\delta}}\mathbf{E}_{Z\sim P_\theta}[\log\frac{1}{2}S^*_{\underline{\delta}}], \inf_{\theta:\theta\leq-\underline{\delta}}\mathbf{E}_{Z\sim P_\theta}[\log\frac{1}{2}S^*_{-\underline{\delta}}]\} = \\
&\max\{\mathbf{E}_{Z\sim P_{\underline{\delta}}}[\log\frac{1}{2}S^*_{\underline{\delta}}], \mathbf{E}_{Z\sim P_{-\underline{\delta}}}[\log\frac{1}{2}S^*_{-\underline{\delta}}]\}, \qquad (39)
\end{aligned}$$

where the final equality is just condition (37) 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:

$$\begin{aligned}
&\inf_{\theta:|\theta|\geq\underline{\delta}}\mathbf{E}_{Z\sim P_\theta}[\log S^*_{\Theta(\underline{\delta})}] \leq \mathbf{E}_{\theta\sim W_{\underline{\delta}}}\mathbf{E}_{Z\sim P_\theta}[\log S^*_{\Theta(\underline{\delta})}] \leq \\
&\mathbf{E}_{\theta\sim W_{\underline{\delta}}}\mathbf{E}_{Z\sim P_\theta}\left[\log\frac{P_{W_{\underline{\delta}}}(Z)}{P_0(Z)}\right] = \mathbf{E}_{\theta\sim W_{\underline{\delta}}}\mathbf{E}_{Z\sim P_\theta}\left[\log S^\circ_{\Theta(\underline{\delta})}\right] = \\
&\frac{1}{2}\mathbf{E}_{\underline{\delta}}[\log\frac{1}{2}S^*_{\underline{\delta}}] + \frac{1}{2}\mathbf{E}_{\underline{-\delta}}[\log\frac{1}{2}S^*_{-\underline{\delta}}] + \epsilon_n \leq \\
&\max\{\mathbf{E}_{Z\sim P_{\underline{\delta}}}[\log\frac{1}{2}S^*_{\underline{\delta}}], \mathbf{E}_{Z\sim P_{-\underline{\delta}}}[\log\frac{1}{2}S^*_{-\underline{\delta}}] + \epsilon_n, \qquad (40)
\end{aligned}$$

where the first inequality is linearity of expectation and the second inequality follows because, since $S^*_{\Theta(\underline{\delta})}$ is an S-value relative to $\{P_0\}$, we can set $q \coloneqq S^*_{\Theta(\underline{\delta})} \cdot p_0$; then $\int q(Z)d\lambda \leq 1$ and $S^*_{\Theta(\underline{\delta})} = q(Z)/p_0(Z)$, and the inequality follows by the information inequality of information theory. ϵ_n above is defined as:

$$\begin{aligned}\epsilon_n &= \frac{1}{2} \cdot \left(\mathbf{E}_{\underline{\delta}}[\log S^\circ_{\Theta(\underline{\delta})} - \log \frac{1}{2} S^*_{\underline{\delta}}] + \mathbf{E}_{\underline{-\delta}}[\log S^\circ_{\Theta(\underline{\delta})} - \log \frac{1}{2} S^*_{\underline{-\delta}}] \right) \\ &= \log 2 + \frac{1}{2} \cdot \left(\mathbf{E}_{\underline{\delta}}[\log S^\circ_{\Theta(\underline{\delta})}/S^*_{\underline{\delta}}] + \mathbf{E}_{\underline{-\delta}}[\log S^\circ_{\Theta(\underline{\delta})}/S^*_{\underline{-\delta}}] \right) \\ &= \log 2 - \frac{1}{2} \left(D(P_{\underline{\delta}}(Y^n) \| P_{W_{\underline{\delta}}}(Y^n)) + D(P_{\underline{-\delta}}(Y^n) \| P_{W_{\underline{\delta}}}(Y^n)) \right).\end{aligned}$$

Together, (39) and (40) show that $S^\circ_{\Theta(\underline{\delta})}$ is an S-value whose worst-case growth rate is always within $\epsilon_n \leq \log 2$ ('1 bit') of that of the minimax optimal $S^*_{\Theta(\underline{\delta})}$; moreover, for fixed $\underline{\delta}$, ϵ_n quickly converges to 0, since, for $\theta \in \{\underline{\delta}, \underline{-\delta}\}$, if $Y_n \sim P_\theta$, then with high probability, $P_{-\theta}/P_\theta$ will be exponentially small in n, so that $D(P_\theta(Y^n) \| P_{W_{\underline{\delta}}}(Y^n)) \approx -\log(1/2) = \log 2$.

Appendix E Proofs and Details for Section 4.2

We first walk through the claims made in Section 4.2. The first claim is that under all $P_{0,\sigma}$ with $\sigma > 0$, V has the same distribution, say P_0, and under all $P_{W[\delta],\sigma}$ with $\sigma > 0$, V has the same distribution, say $P_{W[\delta]}(V)$. 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 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 V as a function of $(Y'_1, \ldots, Y'_n)$, hence the distribution of V does not depend on σ either (note that at this stage, symmetry of the prior is not yet required).

To show (22), we first need to show how to re-express the Bayes factor in terms of densities on V and Y_1. For every prior $W[\delta]$, the corresponding Bayes marginal distribution $P_{W[\delta],\sigma}$, given by (20), viewed as a distribution on (V, Y_1), has density $p'_{W[\delta],\sigma} \coloneqq p_{W[\delta],\sigma}[V, Y_1]$ on $\mathbb{R}^n \times \mathbb{R}+$ that factorizes as

$$p'_{W[\delta],\sigma}(V) \cdot p'_{W[\delta],\sigma}(Y_1 \mid V_1) = p'_{W[\delta]}(V) \cdot p'_{W[\delta],\sigma}(Y_1 \mid V_1),$$

the second equation following because, as we already showed, the density of V does not depend on σ under either $\mathcal{H}_1$ or $\mathcal{H}_0$, so that we can write $p'_{W[\delta]}(V) = p'_{W[\delta],\sigma}(V)$ for all $\sigma > 0$. Now, under every distribution in $\mathcal{H}_1 \cup \mathcal{H}_0$, $V_1 \in \{-1, 1\}$ a.s. Thus, iff, as we assume, the prior $W[\delta]$ is *symmetric* around 0, then V_1 has a Bernoulli(1/2)-distribution under all $P \in \mathcal{H}_0 \cup \mathcal{H}_1$ and we also have

$$p'_{W[\delta],\sigma}(Y_1 \mid V_1) = 2p'_{W[\delta],\sigma}(Y_1) \text{ and } p'_{0,\sigma}(Y_1 \mid V_1) = 2p'_{0,\sigma}(Y_1). \tag{41}$$

We can thus rewrite the Bayes factor with the right Haar prior, (22), as

$$\begin{aligned}\frac{\int_\sigma p_{W[\delta],\sigma}(Y) w^H(\sigma) d\sigma}{\int_\sigma p_{0,\sigma}(Y) w^H(\sigma) d\sigma} &= \frac{\int_\sigma p'_{W[\delta],\sigma}(V) \cdot p'_{W[\delta],\sigma}(Y_1) w^H(\sigma) d\sigma}{\int_\sigma p'_{0,\sigma}(V) \cdot p'_{0,\sigma}(Y_1) w^H(\sigma) d\sigma} \\ &= \frac{p'_{W[\delta]}(V) \cdot \int_\sigma p_{W[\delta],\sigma}(Y_1) w^H(\sigma) d\sigma}{p'_0(V) \cdot \int_\sigma p'_{0,\sigma}(Y_1) w^H(\sigma) d\sigma} = \frac{p'_{W[\delta]}(V)}{p'_0(V)} \cdot g(V_1)\end{aligned} \tag{42}$$

where g is some function of V_1. This final equation follows from Theorem 2.1. of Berger et al. (1998) (this is best seen from its statement in the notation used by (Hendriksen et al., 2018, Section 4.2)). Since V_1 can only take on two values, -1 and 1, it suffices to show that $g(1) = g(-1)$. But this must be the case, since the ratio of integrals over σ has the same value for any value of Y_1 and $-Y_1$, by our assumption that $W[\delta]$ is a prior that is symmetric around 0; (22) is thus proved.

Proof of Theorem 3 We actually prove a more general statement: let $\mathcal{W}_1$ be a set of probability distributions on $\delta \times \sigma$ such that (a) for all $W \in \mathcal{W}_1$, the marginal of W on δ coincides with the given $W[\delta]$ and (b) for all distributions $W[\sigma] \in \mathcal{W}(\Gamma)$ (the set of all probability distributions on $\sigma \in \Gamma = \mathbb{R}^+$), $\mathcal{W}_1$ contains a distribution W whose marginal on σ coincides with $W[\sigma]$ and under which σ and δ are independent. Clearly in general we have $\mathcal{W}_1 \subseteq \mathcal{W}_1'$ where $\mathcal{W}_1'$ is defined as in Theorem 3 in the main text, and we can also choose $\mathcal{W}_1$ to be equal to $\mathcal{W}_1'$. We now show:

Theorem 3, Strengthened *Theorem 3 holds not just with $\mathcal{W}_1'$ but with any $\mathcal{W}_1$ of the form just given.*

To prove this, fix $W[\delta]$ as in the theorem statement, and any corresponding $\mathcal{W}_1$ as above.

We need to show (a),

$$\sup_{S \in \mathcal{S}(\Theta_0)} \inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W}[\log S] = \inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W}[\log S^*_{W[\delta]}\langle V \rangle] = \inf_{W \in \mathcal{W}_1} \inf_{W[\sigma] \in \mathcal{W}(\Gamma)} D(P_W \| P_{0,W[\sigma]}), \tag{43}$$

and (b), that the expression in (43) is $< \infty$. We first turn to (a). Clearly the first expression on the left is no smaller than the second:

$$\sup_{S \in \mathcal{S}(\Theta_0)} \inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W}[\log S] \geq \inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W}[\log S^*_{W[\delta]}\langle V \rangle]$$

Thus, if we can also show that the second is no smaller than the third,

$$\inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W}[\log S^*_{W[\delta]}\langle V \rangle] \geq \inf_{W \in \mathcal{W}_1} \inf_{W[\sigma] \in \mathcal{W}(\Gamma)} D(P_W \| P_{0,W[\sigma]}) \tag{44}$$

and that the third is no smaller than the first,

$$\inf_{W_1 \in \mathcal{W}_1} \inf_{W[\sigma] \in \mathcal{W}(\Gamma)} D(P_{W_1} \| P_{0,W[\sigma]}) \geq \sup_{S \in \mathcal{S}(\Theta_0)} \inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W}[\log S], \tag{45}$$

then we're done. We first show the latter equation. By Theorem 1, Part 1, we have for each fixed $W_1 \in \mathcal{W}_1$ that

$$\inf_{W[\sigma] \in \mathcal{W}(\Gamma)} D(P_{W_1} \| P_{0,W[\sigma]}) = \sup_{S \in \mathcal{S}(\Theta_0)} \mathbf{E}_{P_{W_1}}[\log S]$$

and this directly implies (45) by a standard "inf sup $\geq$ sup inf" argument (the trivial side of the minimax theorem).

It thus remains to show (44). Since $S^*_{W[\delta]}\langle V\rangle$ can be written as a function of V, and, as we already showed (see main text above (21) and proof in beginning of Appendix E above), the distribution of V under $P_{\delta,\sigma}$ does not depend on σ and hence is completely specified by $W[\delta]$, we have for any $W \in \mathcal{W}_1$ that $\mathbf{E}_{P_W}[\log S^*_{W[\delta]}\langle V\rangle] = \mathbf{E}_{P_{W[\delta]}}[\log S^*_{W[\delta]}\langle V\rangle]$ and hence we are done if we can show the following: there exist a set of priors $\{W^{(t)}[\sigma] : t > 0\}$ on $\sigma \in \Gamma$ such that, for some function $u : \mathbb{R}^+ \to \mathbb{R}^+$,

$$\mathbf{E}_{P_{W[\delta]}}[\log S^*_{W[\delta]}\langle V\rangle] \geq \liminf_{t\downarrow 0} D(P_{W[\delta],W^{(t)}[\sigma]} \| P_{0,W^{(u(t))}[\sigma]}), \tag{46}$$

where $P_{W[\delta],W^{(t)}[\sigma]}$ is the marginal over the product prior on $\delta \times \sigma$ with marginals $W[\delta]$ and $W^{(t)}[\sigma]$ which is contained in $\mathcal{W}_1$, and $W^{(t)}[\sigma] \in \mathcal{W}(\Gamma)$.

We proceed to show that a family $\{W^{(t)}[\sigma] : t > 0\}$ satisfying (46) exists. First, let $w^H(\sigma) = 1/\sigma$ be the density of the right Haar prior. For each $t > 0$, let $v_t : \mathbb{R}^+_0$ be such that for all $\sigma > 0$, $v_t(\sigma) \leq w^H(\sigma)$ is an integrable function (to be specified explicitly further below) such that $w_t(\sigma) = \frac{v_t(\sigma)}{\int_0^\infty v_t(\sigma)d\sigma}$ is a probability density. We define $W^{(t)}[\sigma]$ to be the distribution with density w_t. Let $\bar{p}^{(t)}_0$ be the marginal density for the data based on marginalizing $p_{0,\sigma}$ over v_t on σ (which in general does not integrate to 1, so $\bar{p}^{(t)}_0$ may not be a probability density). Let $\bar{q}^{(t)}_0 = p_{W^{(t)}[\sigma]}$ be the Bayes marginal probability density based on the prior density w_t (with $\bar{Q}^{(t)}_0$ the corresponding distribution) and let $\bar{p}^H_0$ be the marginal you get with w_t replaced by the right Haar measure $w^H(\sigma) = 1/\sigma$. Let $\bar{p}^{(t)}_1$ and let $\bar{q}^{(t)}_1$ and $\bar{p}^H_1$ be the corresponding marginal densities based on marginalizing over the product measure on $\delta \times \sigma$ with marginals $W[\delta]$ and, respectively, densities v_t, w_t and w^H. Summarizing:

$$\begin{aligned}
\bar{p}^{(t)}_0(y) &= \int p_{0,\sigma}(y)v_t(\sigma)d\sigma \\
\bar{p}^{(t)}_1(y) &= \int p_{\delta,\sigma}(y)v_t(\sigma)ddW[\delta]d\sigma \\
\bar{q}^{(t)}_0(y) &= \int p_{0,\sigma}(y)w_t(\sigma)d\sigma \\
\bar{q}^{(t)}_1(y) &= \int p_{\delta,\sigma}(y)w_t(\sigma)ddW[\delta]d\sigma \\
\bar{p}^H_0(y) &= \int p_{0,\sigma}(y)w^H(\sigma)d\sigma \\
\bar{p}^H_1(y) &= \int p_{\delta,\sigma}(y)w^H(\sigma)ddW[\delta]d\sigma.
\end{aligned}$$

The idea is that, as $t \downarrow 0$, the $\bar{Q}^{(t)}_j$ become successively better approximations of the P^H_j.

For any function $u(t) : \mathbb{R}^+ \to \mathbb{R}^+$, we have, for all $t > 0$:

$$\begin{aligned}
\mathbf{E}_{V\sim P_{W[\delta]}}[\log S^*_{W[\delta]}\langle V\rangle] &= \mathbf{E}_{Y\sim\bar{Q}_1^{(t)}}[\log \bar{p}_1^H(Y)/\bar{p}_0^H(Y)] \\
&\geq \mathbf{E}_{Y\sim\bar{Q}_1^{(t)}}[\log \bar{p}_1^{(t)}(Y)/\bar{p}_0^H(Y)] \\
&= \mathbf{E}_{Y\sim\bar{Q}_1^{(t)}}[\log \bar{p}_1^{(t)}(Y)/\bar{p}_0^{(u(t))}(Y)] + \log \bar{p}_0^{(u(t))}(Y)/\bar{p}_0^H(Y)] \\
&= D(\bar{Q}_1^{(t)}\|\bar{Q}_0^{(u(t))}) + \mathbf{E}_{Y\sim\bar{Q}_1^{(t)}}[\log \bar{p}_0^{(u(t))}(Y)/\bar{p}_0^H(Y)] \\
&\geq \liminf_{t\downarrow 0}\{D(P_{W[\delta],W^{(t)}[\sigma]}\|P_{0,W^{(u(t))}[\sigma]}) + \mathbf{E}_{Y\sim\bar{Q}_1^{(t)}}[\log \bar{p}_0^{(u(t))}(Y)/\bar{p}_0^H(Y)]\},
\end{aligned}$$

where we note that, under our assumptions, the first expectation on the left is well-defined, nonnegative and bounded by Lemma 2, part 1.

Here the first equation above is just the property of the right-Haar prior expressed by (21) in the main text, proven in the beginning of Appendix E; the property holds as long as the prior $W[\delta]$ on δ is symmetric around 0, as we require. The second is by definition of $\bar{p}_1^{(t)}$, the final equation is evident and the fourth follows because, as we will show below, the right-most expression $\mathbf{E}_{Y\sim\bar{Q}_1^{(t)}}[\log \bar{p}_0^{(u(t))}(Y)/\bar{p}_0^H(Y)]$ is finite; hence the sum in the fourth line is finite as well, and the fourth and third line must be equal.

To show (46), it is thus sufficient to show that we can choose the function $u(t)$ and the densities $\{v^{(t)} : t > 0\}$ satisfying the requirements above so that we have

$$\liminf_{t\downarrow 0} \mathbf{E}_{Y\sim\bar{Q}_1^{(t)}}[\log \bar{p}_0^{(u(t))}(Y)/\bar{p}_0^H(Y)] \geq 0,$$

which is equivalent to:

$$\liminf_{t\downarrow 0} \mathbf{E}_{\sigma\sim W^{(t)}[\sigma],\delta\sim W[\delta]}\mathbf{E}_{Y\sim P_{\delta,\sigma}}[\log \bar{p}_0^{(u(t))}(Y)/\bar{p}_0^H(Y)] \geq 0, \tag{47}$$

But Lemma 2 in the next section shows that we can indeed choose the v_t and $u(t)$ such that (47)) holds. We also need to show that the left-hand side of (46 is bounded, which is also implied by Lemma 2; the theorem is proved.

Proof of Theorem 4 Let $\mathcal{W}_1$ be the set of all product priors on $\delta \times \sigma$ so that the marginal on δ is contained in $\mathcal{W}_1[\delta]$ as defined in the statement of the theorem and the marginal on σ is contained in $\mathcal{W}[\Gamma]$, the set of all priors on σ. Clearly we can reformulate Theorem 4 as follows:

Theorem 4, rephrased *Theorem 3 (the strengthened version stated and proved above) still holds with $\mathcal{W}_1$ as defined above.*

We will now prove this rephrased statement. We see that all steps in the proof of Theorem 3 continue to hold with the new definition of $\mathcal{W}_1$, as long as we can prove the following analogue of (44):

$$\inf_{W\in\mathcal{W}_1} \mathbf{E}_{P_W}[\log S^*_{W_{\underline{\delta}}}\langle V\rangle] \geq \inf_{W\in\mathcal{W}_1}\inf_{W[\sigma]\in\mathcal{W}(\Gamma)} D(P_W\|P_{0,W[\sigma]}). \tag{48}$$

Since the product prior on (δ, σ) with marginal $W_{\underline{\delta}}$ on δ and any prior $W[\sigma]$ on σ is in $\mathcal{W}_1$, we have

$$\inf_{W[\sigma] \in \mathcal{W}(\Theta_0)} D(P_{W_{\underline{\delta}}, W[\sigma]} \| P_{0, W[\sigma]}) \geq \inf_{W \in \mathcal{W}_1} \inf_{W_{[\sigma]} \in \mathcal{W}(\Gamma)} D(P_W \| P_{0, W_{[\sigma]}}),$$

and hence, to prove (48) it is sufficient if we can prove that the following strengthening of (46) holds:

$$\inf_{W \in \mathcal{W}_1} \mathbf{E}_{P_W}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle] \geq \liminf_{t \downarrow 0} D(P_{W_{\underline{\delta}}, W^{(t)}[\sigma]} \| P_{0, W^{(u(t))}[\sigma]}).$$

Since (46) itself still holds with the prior $W[\delta]$ set to $W_{\underline{\delta}}$, using the same arguments as in the proof of Theorem 3, it is sufficient if we can show that

$$\inf_{W[\delta] \in \mathcal{W}_1[\delta]} \mathbf{E}_{P_{W[\delta]}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle] = \mathbf{E}_{P_{W_{\underline{\delta}}}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle]. \tag{49}$$

Now, let $\mathcal{W}_1^\circ[\delta] = \{W^\circ : W \in \mathcal{W}_1[\delta]\}$ be the 'symmetrized' version of $\mathcal{W}_1[\delta]$: for given prior W on δ, W° is defined as $(1/2)W + (1/2)W^-$, where W^- is the mirror prior of W. That is, letting F be the cdf of δ under prior W, F is also the cdf of $-\delta$ under prior W^-. We have that $W_{\underline{\delta}} \in \mathcal{W}_1^\circ[\delta]$ and, by symmetry and linearity of expectation,

$$\inf_{W[\delta] \in \mathcal{W}_1[\delta]} \mathbf{E}_{P_{W[\delta]}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle] = \inf_{W[\delta]: W[\delta]^{-1} \in \mathcal{W}_1} \mathbf{E}_{P_{W[\delta]}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle] = \inf_{W[\delta] \in \mathcal{W}_1^\circ[\delta]} \mathbf{E}_{P_{W[\delta]}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle],$$

so it suffices to show (49) with $\mathcal{W}_1$ replaced by $\mathcal{W}_1^\circ$. Let W_δ be the prior that puts mass $1/2$ on δ and $1/2$ on $-\delta$. By linearity of expectation, it thus suffices to show that

$$\inf_{\delta \geq \underline{\delta}} \mathbf{E}_{P_{W_\delta}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle] = \mathbf{E}_{P_{W_{\underline{\delta}}}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle],$$

which will hold iff $\mathbf{E}_{P_{W_\delta}}[\log S^*_{W_{\underline{\delta}}}\langle V \rangle]$ is increasing in $\delta \geq \underline{\delta}$. But showing the latter is straightforward.

Appendix F Approximation Lemma for Right Haar Measure on σ

Whereas all aspects of the proofs of Theorem 3 and Theorem 4 appear to generalize to arbitrary group invariant settings with right Haar measures, the proof of the following lemma is highly specific to the case where the nuisance parameter of interest is the variance σ^2; thus, if one wants to generalize the results to arbitrary group invariant nuisance parameters, it is the following lemma that needs to be generalized (that's why we organized it into a separate section).

Lemma 2. *Let $W[\delta]$, $W^{(t)}[\sigma]$ with density w_t, $\bar{Q}_{0,t}$, $\bar{Q}_{1,t}$, etc. be defined as in and satisfy the requirements of Theorem 3. Suppose that $\mathbf{E}_{\delta \sim W[\delta]}[\log(1 + |\delta|)] < \infty$ and that $n > 1$. We have, (a), that $\mathbf{E}_{Y \sim \bar{Q}_1^{(t)}}[\log \bar{p}_1^H(Y)/\bar{p}_0^H(Y)] < \infty$. Furthermore, (b), if we set $v_t := v_{t,t}$ where for $t, a > 0$:*

$$v_{t,a}(\sigma) = \frac{1}{\sigma^{1+a}} \exp(-t/(2\sigma^2))$$

and we set $u(t) = t^2$, then (47) holds, i.e. we have:

$$\liminf_{t \downarrow 0} \mathbf{E}_{\sigma \sim W^{(t)}[\sigma], \delta \sim W[\delta]} \mathbf{E}_{Y \sim P_{\delta,\sigma}}[\log \bar{p}_0^{(t^2)}(Y)/\bar{p}_0^H(Y)] \geq 0. \tag{50}$$

Proof of Lemma 2 To show (a), note that it is equal to the left-hand side of (51), and as established earlier (see (21)), we have, for all distributions $W[\sigma]$ on σ,

$$\mathbf{E}_{V\sim P_{W[\delta]}}[\log S^*_{W[\delta]}\langle V\rangle] = \mathbf{E}_{\sigma\sim W[\sigma]}\mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{Y\sim P_{\delta,\sigma}}[\log \bar{p}_1^H(Y)/\bar{p}_0^H(Y)] \tag{51}$$

We recognize the term on the left as the KL divergence between two marginal distributions on V, $P_{W[\delta]}[V]$ and $P_0[V]$, hence it is well-defined and ≥ 0; we need to show that it is $< \infty$. The right-hand side takes on the same value no matter the definition of $W[\sigma]$, hence without loss of generality we can take $\sigma = 1$, and the right-hand side then becomes

$$\mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{Y\sim N(\delta,1)}[-\log \bar{p}_0^H(Y) + \log \bar{p}_0^H(Y)] = \\ \mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{Y\sim N(\delta,1)}[-\log \bar{p}_0^H(Y)] + \mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{Y\sim N(\delta,1)}[\log \bar{p}_1^H(Y)] \tag{52}$$

provided that the expectations on the right-hand side are both well-defined and not equal to $-\infty, \infty$ or $\infty, -\infty$. We show below that the first expectation on the right is finite, so that the splitting of expectations in (52) is justified.

The Haar integral $\bar{p}_0^H$ on the right can be evaluated, which gives

$$\begin{aligned}\mathbf{E}_{Y\sim N(\delta,1)}[-\log \bar{p}_0^H(Y)] &= f(n) + \mathbf{E}_{Y\sim N(\delta,1)}\left[\frac{n}{2}\log U^2\right] \\ &\leq f(n) + \log \mathbf{E}_{Y\sim N(\delta,1)}\left[\frac{n}{2}U^2\right] \\ &= f(n) + \frac{n}{2}\log\left((k(1+\delta^2))^2 + 2k(1+2\delta^2)\right), \end{aligned} \tag{53}$$

where $f(n)$ is a fixed function from $\mathbb{N}$ to $\mathbb{R}$ and $U^2 = \sum_{i=1}^n Y_i^2$. Here the first inequality is Jensen's and the second follows because U^2 has a noncentral χ^2 distribution with parameters k and $k\delta^2$. Under our condition that $\log(1+|\delta|)$ has finite expectation under $W[\delta]$, the expectation of (53) under $W[\delta]$ is bounded, hence the first term on the right in (52) is bounded from above; since the left-hand side of (52) is ≥ 0, it follows that the whole expression is bounded above as well.

For the second part of the result, part (b), note first that $\bar{p}_0^{(t^2)}(Y)$ and $\bar{p}_0^H(Y)$ can both be written as functions of $U^2 = \sum_{i=1}^n Y_i^2$. From now on we denote by $\bar{p}_0^{(t^2)}$ and $\bar{p}_0^H$ the respective densities for U^2 rather than Y. By definition of the distributions $P_{\delta,\sigma}$, the inner expectation in (50) can be rewritten as

$$\mathbf{E}_{U^2\sim P_{\delta,\sigma}}[\log \bar{p}_0^{(t^2)}(U^2)/\bar{p}_0^H(U^2)] = \mathbf{E}_{U^2\sim P_{\delta,1}}[\log \bar{p}_0^{(t^2)}(\sigma^2U^2)/\bar{p}_0^H(\sigma U^2)]$$

and we see that (50) is equivalent to (note the sign change)

$$\limsup_{t\downarrow 0} h(t) \leq 0$$

$$\text{where } h(t) = \mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{U^2\sim P_{\delta,1}}\left[\mathbf{E}_{\sigma\sim W_{(t)}}\left[\log \frac{\bar{p}_0^H(\sigma^2U^2)]}{\bar{p}_0^{(t^2)}(\sigma^2U^2)}\right]\right].$$

For convenience we define $\bar{p}_0^{(t,a)}$ as the marginal density of Y, integrated over $v_{t,a}$ (so $\bar{p}_0^{(t^2)} = \bar{p}_0^{(t^2,t^2)}$). We can explicitly evaluate the integrals in $\bar{p}_0^H$ and $\bar{p}_0^{(t,a)}$, which gives:

$$\log \frac{\bar{p}_0^H(u^2)}{\bar{p}_0^{(t,a)}(u^2)} = \frac{n+a}{2}\log\frac{u^2+t}{u^2} + a\cdot\frac{1}{2}\log(u^2) - a\cdot\frac{1}{2}\log 2 + \log\frac{\Gamma(n/2)}{\Gamma(a/2+n/2)}$$

Substituting (t, a) by (t^2, t^2), and noting that the latter two terms do not depend on u and vanish as $t \downarrow 0, a \downarrow 0$, it is thus sufficient if we can show that

$$\limsup_{t \downarrow 0} h^\circ(t) \leq 0 \tag{54}$$

where

$$\begin{aligned} h^\circ(t) =& \mathbf{E}_{\delta \sim W[\delta]} \mathbf{E}_{U^2 \sim P_{1,\delta}} \mathbf{E}_{\sigma \sim W_{(t)}} \left[\frac{n+t^2}{2} \cdot \log \frac{t^2 + \sigma^2 U^2}{\sigma^2 U^2} + t^2 \cdot \frac{1}{2} \log(\sigma^2 U^2) \right] \leq \\ & \mathbf{E}_{\delta \sim W[\delta]} \mathbf{E}_{U^2 \sim P_{1,\delta}} \left[\frac{n+t^2}{2} \cdot \log \left(1 + \mathbf{E}_{\sigma \sim W_{(t)}} \left[\frac{t^2}{\sigma^2 U^2} \right] \right) + \mathbf{E}_{\sigma \sim W_{(t)}} \left[t^2 \cdot \frac{1}{2} \log(\sigma^2 U^2) \right] \right] = \\ & \mathbf{E}_{\delta \sim W[\delta]} \mathbf{E}_{U^2 \sim P_{\delta,1}} \left[\frac{n+t^2}{2} \cdot \log \left(1 + \frac{t^2}{U^2} \right) + t^2 \cdot \frac{1}{2} \left(\log U^2 + \log(t/2) - \Psi(t/2) \right) \right] \end{aligned} \tag{55}$$

where the inequality is Jensen's and the last line follows by evaluating the expectations over $1/\sigma^2$ and $\log \sigma^2$ if $\sigma \sim W_{(t)} = W_{(t,t)}$, which can be done analytically: the first is equal to 1 and the second is equal to $\log(t/2) - \Psi(t/2)$, where Ψ is the digamma function.

Before evaluating (55) further, we first define:

$$\begin{aligned} g_1(t) &= \frac{t^2}{2} \left(\log(t/2) - \Psi(t/2) \right) \\ g_2(t) &= g_1(t) + \frac{t^2}{2} \cdot \mathbf{E}_{\delta \sim W[\delta]} [\log(n(1+\delta^2))] \\ g_3(t) &= g_1(t) + g_2(t) + \frac{n+t^2}{2} \log(1+t^2) \\ g_4(t) &= \frac{n+t^2}{2} \cdot \int_0^1 (\log(t+u^2) - \log u^2) du. \end{aligned}$$

It is straightforward to establish that $\lim_{t \downarrow 0} g_j(t) = 0$ for $j = 1..4$ (if the leftmost t^2 in $g_1(t)$ is replaced by t then $g_1(t)$ does not converge to 0; this is the reason why we compared $Q_1^{(t)}$ with $Q_0^{(t^2)}$ rather than $Q_0^{(t)}$). For g_2, this follows by our assumption on $W_{[\delta]}$.

The idea is now to bound (55) further in terms of the g_j:

$$\begin{aligned} h^\circ(t) = g_1(t) + \frac{t^2}{2} \cdot \mathbf{E}_{\delta \sim W[\delta]} \mathbf{E}_{U^2 \sim P_{\delta,1}} \left[\log U^2 \right] + \frac{n+t^2}{2} \mathbf{E}_{\delta \sim W[\delta]} \mathbf{E}_{U^2 \sim P_{\delta,1}} \left[\log \left(1 + \frac{t^2}{U^2} \right) \right] \leq \\ g_1(t) + \frac{t^2}{2} \mathbf{E}_{\delta \sim W[\delta]} \left[\log \mathbf{E}_{U^2 \sim P_{\delta,1}} \left[U^2 \right] \right] + \frac{n+t^2}{2} \mathbf{E}_{\delta \sim W[\delta]} \mathbf{E}_{U^2 \sim P_{\delta,1}} \left[\log \frac{t^2+U^2}{U^2} \right] = \\ g_1(t) + \frac{t^2}{2} \mathbf{E}_{\delta \sim W[\delta]} \left[\log(n + n\delta^2) \right] + \frac{n+t^2}{2} \mathbf{E}_{\delta \sim W[\delta]} \mathbf{E}_{U^2 \sim P_{\delta,1}} \left[\log \frac{t^2+U^2}{U^2} \right] = \end{aligned}$$

$$= g_2(t) + \frac{n+t^2}{2}\mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{U^2\sim P_{\delta,1}}\left[\mathbf{1}_{U^2\geq 1}\log\left(1+\frac{t^2}{U^2}\right)+\mathbf{1}_{0\leq U^2<1}\log\frac{t^2+U^2}{U^2}\right] \leq$$

$$g_2(t) + \frac{n+t^2}{2}\mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{U^2\sim P_{\delta,1}}\left[\mathbf{1}_{U^2\geq 1}\log(1+t^2)+\mathbf{1}_{0\leq U^2<1}\left(\log(t^2+U^2)-\log U^2\right)\right] =$$

$$g_3(t) + \frac{n+t^2}{2}\mathbf{E}_{\delta\sim W[\delta]}\mathbf{E}_{U^2\sim P_{\delta,1}}\left[\mathbf{1}_{0\leq U^2<1}\left(\log(t^2+U^2)-\log U^2\right)\right] \leq$$

$$g_3(t) + \frac{n+t^2}{2}\mathbf{E}_{\delta\sim W[\delta]}\left[\int_0^1 f_{\delta,n}(u)(\log(t^2+u^2)-\log u^2)du\right] \leq$$

$$g_3(t) + \frac{n+t^2}{2}\cdot\sup_{\delta\in\mathbb{R}}\max_{v\in[0,1]} f_{\delta,n}(v)\int_0^1(\log(t^2+u^2)-\log u^2)du =$$

$$g_3(t) + g_4(t)\cdot\sup_{\delta\in\mathbb{R}}\max_{v\in[0,1]} f_{\delta,n}(v).$$

Here the second line follows from Jensen's inequality. For the third, we used that U^2 has a noncentral χ^2-distribution with n degrees of freedom and noncentrality parameter $n\delta^2$; $f_{\delta,n}$, appearing later, represents the density of such a distribution. All other (in)equalities are immediate.

Since we already showed that $\lim_{t\downarrow 0} g_j(t) = 0$ for $j = 3, 4$, it suffices if we can show that $\sup_{\delta\in\mathbb{R}}\max_{v\in[0,1]} f_{\delta,n}(v) < \infty$. Since we assume $n > 1$, we have that $\max_{v\in[0,1]} f_{0,n}(v) < \infty$. And since $f_{\delta,n}(v)$ is decreasing in $|\delta|$ for each $v \in [0, 1]$, the result follows.

F.1 Why W_1^* and W_0^* are achieved and have finite support in Section 4.4

The minima are achieved because of the joint lower-semicontinuity 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.

Appendix G 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 (16) 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 $\Theta(\epsilon)$ vs. Θ_0: the larger ϵ, the 'further' $\Theta(\epsilon)$ from Θ_0 and the larger the value of $\text{GR}(\epsilon)$. The KL divergence is the *only divergence measure* in which 'easiness' of testing $\Theta(\epsilon)$ 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 $\text{GR}(\{\theta_1'\}) = \text{GR}(\{\theta_1\})$ 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 $\epsilon > 0$, $\theta_1 \in \Theta(\epsilon)$ yet $\theta_2' \notin \Theta(\epsilon)$. With d equal to KL this is impossible. This follows immediately from Theorem 1, Part 1, which tells us $D(\theta_1\|\Theta_0) = \text{GR}(\{\theta_1\})$.