The E-Posterior

Peter D. Grünwald[*]

January 5, 2023

Abstract

We develop a representation of a decision maker's uncertainty based on e-variables. Like the Bayesian posterior, this *e-posterior* allows for making predictions against arbitrary loss functions that may not be specified ex ante. Unlike the Bayesian posterior, it provides risk bounds that have frequentist validity irrespective of prior adequacy: if the e-collection (which plays a role analogous to the Bayesian prior) is chosen badly, the bounds get loose rather than wrong, making *e-posterior minimax* decision rules safer than Bayesian ones. The resulting *quasi-conditional paradigm* is illustrated by re-interpreting a previous influential partial Bayes-frequentist unification, *Kiefer-Berger-Brown-Wolpert conditional frequentist tests*, in terms of e-posteriors.

1 Introduction and Overview

It is one of the main attractions of Bayesian inference: one may assess the posterior-expected performance of arbitrary decision rules in arbitrary decision tasks, and, as a consequence, perform posterior-optimal decisions for any such task. An important special case arises if we are given a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ equipped with a prior distribution W on Θ, a space of actions $\mathcal{A}$ and a loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$, and we observe data $Y = y$, with $Y \sim P_\theta$ for some $\theta \in \Theta$, taking values in some set $\mathcal{Y}$. We call such a combination of $\mathcal{P}$ and L a *decision problem*. A decision rule δ is then any function mapping data $y \in \mathcal{Y}$ to action $a \in \mathcal{A}$; we write $\delta(y) = a$. We then assess the expected performance of δ given data $Y = y$ as

$$\mathbf{E}_{\theta \sim W|y}[L(\theta, \delta(y))], \tag{1}$$

with $W \mid y$ denoting the Bayes posterior distribution based on prior W and data $Y = y$. In their classic *Bayesian theory* (1994), José Bernardo and Adrian Smith adopt, as they make clear already on page 2, a *wholehearted subjectivist position* on the interpretation of probability. Such a subjective Bayesian stance implies that if *the prior distribution truly describes one's initial degrees of belief*, then the assessment (1) is not just correct for *any* decision rule δ, it also implies optimality of the *Bayes decision rule*, defined to be any rule δ^{BAYES} such that, for all $y \in \mathcal{Y}$,

$$\bar{R}^{\text{BAYES}}(W \mid y) := \mathbf{E}_{\theta \sim W|y}[L(\theta, \delta^{\text{BAYES}}(y))] = \min_{a \in \mathcal{A}} \mathbf{E}_{\theta \sim W|y}[L(\theta, a)]. \tag{2}$$

---

[*]CWI, Amsterdam, and Mathematical Institute, Leiden University. CWI is the National Resarch Institute for Mathematics and Computer Science in the Netherlands.

Unfortunately, in practice, for lack of time, computational power or imagination, one often — in fact, nearly always — uses pragmatic priors in combination with potentially 'misspecified' (wrong) models. Then the priors do not truly represent one's beliefs, and as a consequence (1) may give a highly misleading impression of δ's quality in the real world, and using δ^{BAYES} may lead to disastrous results. Much of my own research in the past has concerned misspecified models (Grünwald, 1999, Grünwald and van Ommen, 2017). In the present paper, I assume the model is correct and concentrate on the case in which "good" priors (i.e. sufficiently sophisticated to justify using (2) in practice) are hard to come by. The pragmatic priors that are then used often lead to *overconfident posteriors*, examples of which abound in the literature: decisions based on $W \mid y$ are then substantially worse in practice than predicted by $W \mid y$ itself via (1) (Oelrich et al., 2020), especially in nonparametrics (Szabó et al., 2015); naturally such assessments can lead to (very) bad decisions (Balch et al., 2019). In this paper, we provide the *e-posterior* as an alternative for such cases: while remaining close to the Bayesian ideal, it has a frequentist justification that avoids overconfidence and other misleading assessments. As an additional advantage, in contrast to Bayesian methods, e-posteriors are readily used in highly nonparametric situations in which a likelihood is not easy to define — we defer discussion of this bonus until the discussion Section 5.

The E-Posterior The e-posterior is an analogue to the Bayes posterior for which frequentist analogues ((5) and (6) below) of Bayesian assessments ((1) and (2) above) hold simultaneously under all $\theta \in \Theta$. It is based on e-values, a far-reaching generalization of likelihood ratios. Crucially, just as with the Bayesian posterior, the same e-posterior can be combined with arbitrary decision problems and loss functions. That is, if a given prior W truly represents beliefs, the Bayesian (1) and (2) are appropriate simultaneously for arbitrary sets of loss functions $\{L_b : b \in \mathcal{B}\}$, substantiating the common Bayesian claim that the Bayesian posterior summarizes all one's uncertainty about the parameter given the data — since all conceivable decisions can be based on it. The same holds for the e-posterior analogues below: they are simultaneously valid for all loss functions, so that, analogously to the Bayesian case, the e-posterior can also be thought of as *a single function summarizing one's uncertainty*. As discussed in Section 2, this is what sets out the e-posterior from previous attempts to provide frequentist guarantees for Bayesian posteriors or frequentist analogues thereof such as confidence distributions. There we also explain why we regard validity under post-hoc choice of loss function as crucial in statistical testing and estimation in the applied sciences.

Both the Bayesian and the e-representation of uncertainty are relative: whereas the Bayesian posterior is defined relative to a prior defined upon a statistical model $\mathcal{P}$, the e-posterior is defined relative to a collection $\mathcal{S}$ of e-variables, which is itself defined relative to $\mathcal{P}$. This 'e-collection' $\mathcal{S}$ in turn may itself be implicitly based on a prior (Example 2) or a set of priors (Section 3.1). Because it gives rise to stochastic upper bounds on risks rather than precise risks, we are able to obtain the result that the e-posterior is *valid* in a frequentist sense irrespective of the collection $\mathcal{S}$ taken; but the bounds may get better or worse depending on $\mathcal{S}$. This makes it considerably more robust than the Bayesian risk assessment (1), which, at least under a subjective Bayes interpretation, one can only trust if one truly fully believes one's prior W and one's model $\mathcal{P}$. It puts the e-posterior approach in the *luckiness*-framework (Grünwald, 2007, Shawe-Taylor and Williamson, 1997, Herbrich and Williamson, 2002, Grünwald and Mehta, 2019) in which Bayesian *belief* (one has to believe one's prior to

trust one's assessment) is replaced by *guarantees* and *hope* that the guarantees are good: if the prior is well-aligned with the data, one is *lucky* and gets tight bounds; if not, the bounds one gets are loose but still valid. It goes substantially further though than standard 'luckiness' approaches (such as PAC-Bayesian methods (Grünwald and Mehta, 2019)), since these invariably incorporate an a-priori given loss function into the likelihood (e.g. via a Gibbs posterior) so that the same (e.g. Gibbs) posterior can only be used for a single loss function. Our e-posterior being combinable with arbitrary loss functions, we chose to call the resulting type of inference *quasi-conditional*, a name we explain further in Section 2. Before we start with the details, we should point out that, of course, not all can be rosy: the specific type of stochastic upper bound that provides our sense of quasi-conditional frequentist validity, as given in Proposition 1 and interpreted underneath in terms of a De Finetti-type of game between a decision-maker and a bookie, while holding universally, may in some instances provide a weaker inference than one would really like to see — we consider it as the main open problem of this work to further investigate this novel type of validity.

Formal Definitions Formally, we associate $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with some corresponding collection $\mathcal{S}$ of e-variables $\mathcal{S} = \{S_\theta : \theta \in \Theta\}$. We focus on what we call the *simple setting*, in which all parameters in parameter vector $\theta \in \Theta \subseteq \mathbb{R}^k$ are of interest. An e-variable S_θ relative to any $\theta \in \Theta$ is then simply any *nonnegative* statistic (i.e. a random variable that can be written as a function of the data, $S_\theta = S_\theta(Y)$) that satisfies the inequality:

$$\mathbf{E}_{Y \sim P_\theta}[S_\theta] \leq 1. \tag{3}$$

We defer the case of nonparametrics and nuisance parameters, requiring a slightly extended definition, to Section 5. Over the last three years, interest in e-variables has exploded (Grünwald et al., 2019, Shafer, 2021, Vovk and Wang, 2021, Henzi and Ziegel, 2021, Ramdas et al., 2022), as tools to extend traditional Neyman-Pearson tests to situations with optional stopping and continuation while keeping Type-I error guarantees — see Section 5 for details. Here we use them in a novel way: we define the *e-posterior* based on $\mathcal{S}$ and Y to be the reciprocal of S_θ:

$$\bar{P}(\theta \mid y) := \frac{1}{S_\theta(y)}, \tag{4}$$

where by convention we set $1/0$ to be ∞; and we define the *e-posterior risk assessment* of decision rule δ based on $\mathcal{S}$ by

e-posterior risk bound: $\bar{R}(\delta) := \bar{R}(Y, \delta(Y))$ with $\bar{R}(y, a) := \sup_{\theta \in \Theta} \bar{P}(\theta \mid y) \cdot L(\theta, a),$ (5)

where we use the convention that $\infty \cdot a := 0$ if $a = 0$ and ∞ if $a > 0$. $\bar{R}(\delta)$ is the random variable that maps Y to the corresponding risk bound (5), which is to be interpreted as a specific type of *stochastic upper bound* on the risk that holds under all $\theta \in \Theta$ simultaneously — Proposition 1 below makes this precise. It holds for all loss functions $L : \Theta \times \mathcal{A} \to \mathbb{R}$ satisfying the condition that $\sup_{\theta \in \Theta} L(\theta, a) \geq 0$ for each a. We call this the *no-sure-gain condition*. It is satisfied if $L \geq 0$, but in general it also allows negative losses (as will be useful in Example 6). In contrast to the upper bound-nature of (5), the Bayesian (1) does not just bound but precisely gives the conditional risk. Next, in analogy to (2), we suggest, once $\mathcal{S}$

has been fixed, the $\mathcal{S}$-based *e-posterior-minimax decision rule* as any rule δ^{MM} satisfying, for all $y \in \mathcal{Y}$, any loss function L,

$$\textbf{e-posterior minimax rule:} \ \sup_{\theta \in \Theta} \ \bar{P}(\theta \mid y) \cdot L(\theta, \delta^{\mathrm{MM}}(y)) = \min_{a \in \mathcal{A}} \sup_{\theta \in \Theta} \ \bar{P}(\theta \mid y) \cdot L(\theta, a). \quad (6)$$

Contents of this Paper   Section 2 provides Proposition 1 which establishes the frequentist sense in which $\bar{R}(\delta)$ as defined by (5) is valid. We discuss this *quasi-conditional* validity at length and provide extended examples. The paper continues in Section 3 by listing various types of e-collections, and investigating the risk assessment and the corresponding e-posterior minimax rules that ensue from them when instantiated to simple yet important situations. In particular, Section 3.1 considers the Gaussian location family and a weighted squared error loss, with data-dependent weights. In this situation (extended to 1-dimensional exponential families in the Supplementary Material), Bayesian risk assessment based on standard objective Bayes posteriors can fail spectacularly if the weights strongly vary with the data, whereas our type of risk assessment via (5) is always valid; moreover, if the weights are constant, the assessment for e-posterior minimax rules can match minimax optimal frequentist and Bayesian rates up to a small constant factor, thereby alleviating concerns that the sup in (5) may make assessments overly pessimistic. In Section 3.2, we consider a Bayesian-frequentist unification for hypothesis testing with a simple null that was introduced in an influential paper by Berger, Brown and Wolpert (1994) (BBW from now on), building on earlier work by Kiefer (1976, 1977), Brownie and Kiefer (1977). BBW's approach was later extended with various collaborators (Berger et al., 1994, Wolpert, 1996, Berger and Guglielmi, 2001, Dass and Berger, 2003), culminating in Berger's (2003) IMS Lecture paper *Could Fisher, Neyman and Jeffreys have agreed on testing?* We show how to re-interpret the BBW conditional error probabilities, as well as Kiefer's earlier *conditional confidence estimators* (1976, 1977) and Vovk's (1993) extension of it, in terms of (quite special) e-posteriors. Our re-interpretation suggests that in practice, we will usually want to use other e-variable collections than the BBW one — which is not at all to criticize their pioneering and highly original work, that, together with Vovk's (1993), has served as the main inspiration for this paper. Then, in Section 4, we describe a variation of the e-posterior risk assessment (5) in which the e-posterior is replaced by its *capped* version (see Figure 1) and we investigate when this arguably more intuitive form of the e-posterior gives rise to the same assessment. This turns out to be the case under *Condition Zero*, a fundamental condition on the decision rules employed, which may or may not be enforceable in any given decision problem. Finally, in Section 5, we give more background on the traditional use and interpretation of e-variables and generalize to testing with composite nulls, estimating with nuisance parameters and nonparametrics, and we provide some concluding discussion. But first, we introduce several running examples of e-posteriors, highlighting via (9) the connection with the Bayes posterior.

1.1   Examples of E-Collections

Example 1 [A Simple E-Posterior: Savage-Dickey] A simple way to design e-variables is to start with a prior distribution W on Θ and setting $p_W(y) := \int p_\theta(y) dW(\theta)$ to be the Bayes marginal based on W. Here, as in the rest of the paper, we assume that for all $\theta \in \Theta$, P_θ has density p_θ relative to some common underlying measure μ. One then defines, for each

$\theta \in \Theta$,

$$S_\theta^{\mathrm{SD}} = \frac{p_W(y)}{p_\theta(y)} \quad ; \quad \bar{P}^{\mathrm{SD}}(\theta \mid y) = \frac{p_\theta(y)}{p_W(y)} = \frac{w(\theta \mid y)}{w(\theta)} \tag{7}$$

where the latter equality holds if W has density w and we denote prior/posterior densities by $w(\theta)$ and $w(\theta \mid y)$, respectively. S_θ^{SD} is an e-variable, since it satisfies $\mathbf{E}_{Y \sim P_\theta}[S_\theta^{\mathrm{SD}}] = \int_{\mathcal{Y}} p_W(y) dy = 1$. Pawel et al. (2022) study this e-variable, focusing on an 'evidential' interpretation; it also figures prominently in (Grünwald et al., 2019) and, under the name 'prior-to-posterior e-variable', in (Neiswanger and Ramdas, 2021). We chose our alternative name because, in the Bayesian literature, the quantity $w(\theta \mid y)/w(\theta)$ is generally known as the *Savage-Dickey density ratio* after the prominent Bayesians L. Savage and J. Dickey. The Savage-Dickey e-variables are by no means the only reasonable ones to base an e-posterior on, but they are educationally insightful because they give a first interpretation of risk assessment (5). Namely, using Bayes' theorem we can trivially but unusually rewrite the *standard* Bayesian uncertainty assessment (1) as follows:

$$\mathbf{E}_{\theta \sim W \mid y}[L(\theta, \delta(y))] = \int \frac{p_\theta(y)}{p_W(y)} \cdot L(\theta, \delta(y)) dW(\theta) = \mathbf{E}_{\theta \sim W}[\bar{P}^{\mathrm{SD}}(\theta \mid y) \cdot L(\theta, \delta(y))], \tag{8}$$

$$\text{whereas with } Y = y, \text{ we have } \bar{R}(\delta) = \max_\theta \bar{P}^{\mathrm{SD}}(\theta \mid y) \cdot L(\theta, \delta(y)). \tag{9}$$

Replacing prior expectation (8) by maximum (9), the Savage-Dickey e-posterior is always more pessimistic than the Bayesian posterior based on the same prior, and might thus not look so outlandish to Bayesians any more.

Example 2 [Savage-Dickey for the Normal Location Family] Let $\{P_\theta : \theta \in \mathbb{R}\}$ represent the normal location family, where P_θ with density p_θ has mean θ and variance 1. We take as prior a normal distribution with mean θ_0 and variance $\rho^2 > 0$, and define the precision $\lambda := \rho^{-2}$. Suppose we observe $Y = x^n = (x_1, \ldots, x_n)$. By standard calculations, the standard Bayesian posterior is given by a normal distribution with mean $\breve{\theta} = (n/(n+\lambda))\hat{\theta} + (\lambda/(n+\lambda))\theta_0$, with $\hat{\theta} = \hat{\theta}(y)$ the MLE $(\sum_{i=1}^n x_i)/n$, and posterior variance $1/(n+\lambda)$, i.e. with density

$$w(\theta \mid x^n) = \sqrt{\frac{n+\lambda}{2\pi}} \cdot e^{-\frac{(n+\lambda)(\theta - \breve{\theta})^2}{2}} \tag{10}$$

so that, by (7),

$$\bar{P}^{\mathrm{SD}}(\theta \mid y) = \sqrt{\frac{n+\lambda}{\lambda}} \cdot e^{-\frac{n+\lambda}{2}(\theta - \breve{\theta})^2 + \frac{\lambda}{2} \cdot (\theta - \theta_0)^2} = \sqrt{\frac{n+\lambda}{\lambda}} \cdot e^{-\frac{n}{2}(\theta - \hat{\theta})^2 + \frac{1}{2} \cdot \frac{n\lambda}{n+\lambda} \cdot (\hat{\theta} - \theta_0)^2} \tag{11}$$

where the latter equality follows by simple calculus when $\theta_0 = 0$ and reducing the general case to this case by considering translated data $x_1 - \theta_0, \ldots, x_n - \theta_0$; see Figure 1.

Example 3 [Two-Point Prior-Based E-Posterior for Normal Location] Fix some $C > 0$ and set n^* equal to the sample size n (in Example 8 we generalize to and explain cases with $n^* \neq n$) and, for each θ, define $\theta^- < \theta$ and $\theta^+ > \theta$ such that

$$\frac{1}{2}(\theta - \theta^+)^2 = \frac{1}{2}(\theta - \theta^-)^2 = \frac{C}{n^*}. \tag{12}$$

Then $S_\theta(y) = \frac{(1/2)p_{\theta^-}(y) + (1/2)p_{\theta^+}(y)}{p_\theta(y)}$ is an e-variable (just plug it into the definition to check). It is not of Savage-Dickey form, since θ^- and θ^+ depend on θ. In fact it coincides with the two-sided version of Johnson's (2013) *uniformly most powerful Bayes factor* at level α if we were to set $C := -\log \alpha$, which might, according to Johnson, motivate its use for testing $\mathcal{H}_0 = \{P_\theta\}$ against $\mathcal{H}_1 = \{P_{\theta'} : \theta' \in \Theta, \theta' \neq \theta\}$. Here we will use it for estimating and not just testing, i.e. for general $\theta \in \Theta$; then its Bayesian interpretation is not so clear, but it will tend to give better risk assessments (5) than the Savage-Dickey e-variable above, as we explain in Section 3. Both Example 2 and 3 are extended to general regular (Barndorff-Nielsen, 1978) 1-dimensional exponential families in the Supplementary Material.

Example 4 [Simple Hypothesis Testing with LR E-Variables] Consider testing a simple $\mathcal{H}_0 = \{P_0\}$ vs. a simple $\mathcal{H}_1 = \{P_1\}$ (we remark on the composite case in Section 5). We can embed this in our setting by taking $\mathcal{P} = \mathcal{H}_0 \cup \mathcal{H}_1$ and as our loss function a *Wald-Neyman-Pearson*-type loss given by $\mathcal{A} = \mathcal{A} = \{0, 1\}$, $L : \{0, 1\}^2 \to \mathbb{R}_0^+$ nonnegative, and $L(0, 1) > 0$, $L(1, 0) > 0$. We further say that the loss function satisfies *Condition Zero*, henceforth abbreviated to **C0**, if $L(0, 0) = L(1, 1) = 0$. The interpretation is that action 0 stands for 'accept the null' leading to loss 0 if the true state of nature is P_0; and 1 stands for 'reject the null', leading to loss 0 if the alternative P_1 is true. The use of such loss functions is implicit in the Neyman-Pearson testing paradigm, as noted in 1939 by Wald (1939) who stated that one may 'of course' set $L(0, 0) = L(1, 1) = 0$, i.e. impose **C0**, and because it simplifies treatment we will do so in most testing examples in this paper — but in Example 6 we will deviate from it, and see that this in a sense makes for the most interesting applications of our risk assessments. We associate both P_0 and P_1 with an e-variable. A natural choice is to take the likelihood ratio, $S_1 = p_0(Y)/p_1(Y)$ and $S_0 = p_1(Y)/p_0(Y)$, seen to be e-variables by noting $\mathbf{E}_{Y \sim P_\theta}(p_{\theta'}(Y)/p_\theta(Y)) = \int p_{\theta'}(y)dy = 1$. In fact, they are the 'prototypical' e-variables for simple-vs.-simple testing (Grünwald et al., 2019). The risk assessment (5) now becomes

$$\bar{R}(\delta) = \max \left\{ \frac{p_1(Y)}{p_0(Y)} L(1, \delta(Y)), \frac{p_0(Y)}{p_1(Y)} L(0, \delta(Y)) \right\} \stackrel{\text{under C0}}{=} \frac{p_1(Y)}{p_0(Y)} L(1, \delta(Y)) + \frac{p_0(Y)}{p_1(Y)} L(0, \delta(Y)).$$
(13)

Example 5 [Simple Testing with BBW E-Variables] Several other e-variables may be considered for simple hypothesis testing as well, including the Savage-Dickey e-variables; we treat these in Section 3. There we also show that BBW's conditional error probabilities (Berger et al., 1994, Wolpert, 1996, Berger, 2003) can be interpreted as e-posteriors. Formally, we obtain a parameterized set of e-posteriors $\{\bar{P}_w^{\text{BBW}}(\theta \mid y)\}$ for $0 < w < 1$. We consider the special case $w = 1/2$ here, deferring the general case and definitions to Section 3.2. As we will see there, this e-posterior is quite special: for the e-posterior minimax decision rule δ^{MM} we get, with $\bar{R}$ defined relative to the BBW e-posterior, under **C0**,

$$\bar{R}(\delta^{\text{MM}}) = \frac{p_1(Y)}{p_0(Y) + p_1(Y)} L(1, \delta^{\text{MM}}(Y)) + \frac{p_0(Y)}{p_0(Y) + p_1(Y)} L(0, \delta^{\text{MM}}(Y)).$$
(14)

Thus, it can again be rewritten as a sum rather than a supremum and we see that it is formally identical to the Bayes assessment (1) when P_0 and P_1 are equipped with the uniform prior. We thus have a *frequentist* risk assessment, not requiring a prior assumption on $\mathcal{P}$, that is as strong as a Bayesian one, which is optimal under a specific prior assumption. This seems almost too good to be true, especially since it is slightly, but uniformly, better than the
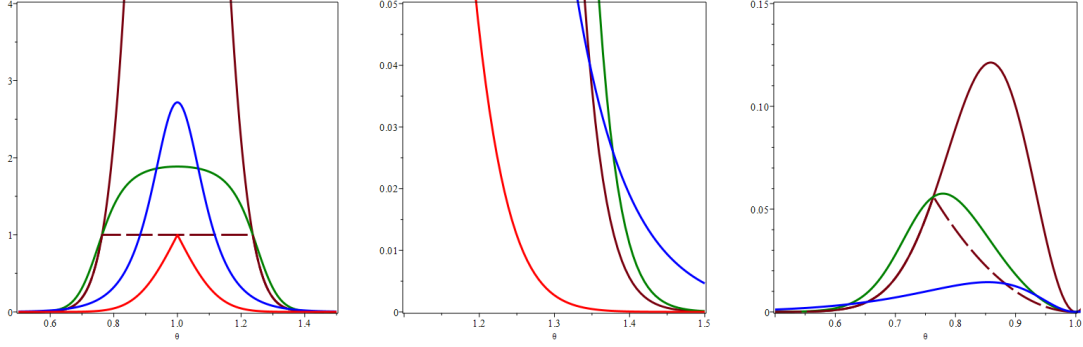
Figure 1: *(left and middle)* The brown solid curve is the pure Savage-Dickey e-posterior $\bar{P}(\theta \mid y)$ for the normal location family based on sample $y = (x_1, \ldots, x_{100})$ with $\hat{\theta}(y) = 1$ and prior W with $\lambda = 1$ and $\theta_0 = 0$ as in (11) (it reaches a maximum of about 17). The green curve represents the $1/2$-dampened version (see (35)) of this e-posterior, by design bounded by 2. The dashed brown curve is the capped version $\bar{P}^{\mathrm{CAP}}(\theta \mid y)$ (Definition 1) of $\bar{P}(\theta \mid y)$, equal to it whenever the latter is ≤ 1. The blue curve represents the generalized Savage-Dickey two-point prior-based e-posterior optimized for $n^* = n = 100$ as in Example 3 and 8. The red curve is twice the tail area of the objective Bayes posterior $w^\circ(\theta \mid y)$ of Example 7, i.e. given by $f(\theta) = 2W^\circ(\bar{\theta} \geq \theta \mid y) = 2 \int_\theta^\infty w^\circ(\theta \mid y) d\theta$. Viz. the remark underneath Definition 1, the set $\{\theta : \bar{P}(\theta \mid y) \geq \alpha\}$ is a conservative 2-sided $(1 - \alpha)$-confidence interval, whereas, viz. Example 7, the set $\{\theta : f(\theta) \geq \alpha\}$ is an exact nonconservative 2-sided $(1 - \alpha)$-confidence interval, so that plotting $f(\theta)$ provides a better comparison to $\bar{P}(\theta \mid y)$ than directly plotting $w^\circ(\theta \mid y)$. The middle display zooms in on the 0.05 right tail.
(right) The value of $(\theta - \hat{\theta})^2 \bar{P}(\theta \mid y)$ is depicted for the same four e-posteriors, using the same colors. The two-point prior-based e-posterior is optimized to give good assessments relative to squared error, its maximum is considerably smaller than for the others; for the pure Savage-Dickey prior, the maximum is excessively high.

alternative (13) — and as we will see in Section 3.2, there is indeed no free lunch: (14) only holds for δ^{MM}, which for the BBW e-variable can sometimes be a rather bad decision rule, whereas (13) holds for arbitrary rules; there are some other prices to pay as well when using BBW's e-variable. In fact, a major motivation for the author's involvement in e-variables was my realization, some fifteen years ago, that when deriving frequentist Type-I error bounds for decisions based on Bayes factors/likelihood ratios in simple testing, one ends up with the expression p_1/p_0 as in (13) if one follows the reasoning of classic Bayesian texts such as (Edwards et al., 1963, Good, 1991) or Royall's *universal bound* (Royall, 1997); whereas using BBW's ideas, one obtains $p_1/(p_1 + p_0)$ as in (14): the small but essential discrepancy between the two assessments tripped me up for about 15 years until I finally developed the necessary tool to understand it — the much more general e-posterior of which both approaches are simply a special case, each with its own (dis)advantages, as discussed in Section 3.2.

2 Quasi-Conditional Risk Assessment

We want risk bounds that are valid, in some precise sense, simultaneously for all loss functions of the form $L : \Theta \times \mathcal{A} \to \mathbb{R}$ that satisfy the no-sure-gain condition (see underneath (5)).

Proposition 1 below shows that, for any choice of e-variable collection $\mathcal{S} = \{S_\theta : \theta \in \Theta\}$ and any given decision rule δ, we have that $\bar{R}(\delta)$ as in (5) satisfies this purpose.

Proposition 1 [Validity of E-Risk Assessments] *Consider a decision problem as defined in the introduction relative to any model $\{P_\theta : \theta \in \Theta\}$ for random variable $Y \in \mathcal{Y}$ and loss function L as above that satisfies the no-sure-gain condition. Let $\bar{R}(\delta)$ be as in (5) relative to some e-variable collection $\mathcal{S} = \{S_\theta : \theta \in \Theta\}$. For all $\theta \in \Theta$ we have:*

$$\mathbf{E}_{Y \sim P_\theta}\left[\frac{L(\theta, \delta(Y))}{\bar{R}(\delta)}\right] \leq 1, \tag{15}$$

where we adopt the convention $0/0 \coloneqq 0$. In particular, this implies, for every function $b : \mathcal{Y} \times \mathcal{U} \times \Theta \times \mathcal{A} \to \mathbb{R}^+$, and for any random variable U taking values in some set $\mathcal{U}$ and any distribution P'_θ for (Y, U) whose marginal on Y coincides with P_θ, that

$$\mathbf{E}_{(Y,U) \sim P'_\theta}\left[b(Y, U, \theta, \delta(Y)) \cdot L(\theta, \delta(Y))\right] \leq \sup_{(y,u,\theta) \in \mathcal{Y} \times \mathcal{U} \times \Theta} b(y, u, \theta, \delta(y)) \cdot \bar{R}(y, \delta(y)). \tag{16}$$

Moreover, for any e-collection $\mathcal{S}$, the risk bound (5) with any minimax decision rule δ^{MM} as given by (6) dominates (5) when applied with any other decision rule δ: for all $y \in \mathcal{Y}$: $\bar{R}(y, \delta^{\mathrm{MM}}(y)) \leq \bar{R}(y, \delta(y))$.

Proof: Fix any $\theta \in \Theta$. By definition of $\bar{R}(y, a)$, we have for all $y \in \mathcal{Y}$ that $\bar{R}(y, \delta(y)) \geq \bar{P}(\theta \mid y)L(\theta, \delta(y))$ so that

$$\mathbf{E}_{Y \sim P_\theta}\left[\frac{L(\theta, \delta(Y))}{\bar{R}(\delta)}\right] \leq \mathbf{E}_{Y \sim P_\theta}\left[\frac{L(\theta, \delta(Y))}{S_\theta^{-1}(Y)L(\theta, \delta(Y))}\right] = \mathbf{E}_{Y \sim P_\theta}\left[S_\theta\right] \leq 1, \tag{17}$$

noting that the derivation is valid as long as the denominator $\bar{R}(\delta)$ inside the first expectation cannot evaluate to 0. In case it can, we split up the expectand into two terms, $\mathbf{1}_{\bar{R}(\delta)=0} \cdot L(\theta, \delta(Y))/\bar{R}(\delta) + \mathbf{1}_{\bar{R}(\delta)>0} \cdot L(\theta, \delta(Y))/\bar{R}(\delta)$, and use that the first fraction is $0/0$ which by our convention evaluates to 0; we proceed with the second term as above. This proves (15). (16) and domination of δ^{MM} is then immediate. $\square$

The bounds (15) and (16) (we mostly use the latter) are unusual and for them to be practically interesting, they should satisfy several desiderata. In particular:

1. They should have a clear interpretation. We show that this is the case directly below.

2. They should exceed standard risk bounds obtainable for standard estimators (actions) in standard decision problems by only a small factor, at least if we choose a reasonable e-collection $\mathcal{S}$. This is indeed the case in our running examples (see Example 8, for the Gaussian location family with standard squared error loss, Example 10 for an extension to exponential families, and Example 9, underneath (28) for simple-vs.-simple testing).

3. There should exist settings in which, just like Bayesian risk assessments, they allow one to make more risky decisions conditional on more extreme data. This happens in Example 6 below.

4. There should exist situations in which they give useful bounds whereas a standard Bayesian assessment, based on a standard convenience prior, is really highly misleading. We give such a case in Example 7.

Bookie Interpretation The following interpretation of (16) has a *De Finettian* flavour to it. Suppose a bookie offers decision-maker (DM) to play the following game, really a meta-decision problem: DM first specifies a decision-rule δ for the original decision-problem. DM then gets paid some amount $\ell > 0$; then θ and Y are revealed to both DM and bookie. Then bookie presents a collection of numbers $\{B_{\theta,a} : \theta \in \Theta, a \in \mathcal{A}\}$, all assumed > 0. Bookie may choose $B_{\theta,a}$ as functions of the data Y and external data, here encoded as U; hence we have $B_{\theta,a} = b(Y, U, \theta, a)$ for some function b. DM then has to pay back $B_{\theta,\delta(Y)} \cdot L(\theta, \delta(Y))$. *If the rules are such that (i.e. bookie a priori guarantees that), no matter what Y, U are observed, bookie will provide a B such that $B_{\theta,\delta(Y)} \leq \ell/\bar{R}(\delta)$ then DM should accept to play this game, for her expected gain is ≥ 0; if this cannot be guaranteed, then DM has no guarantee that the expected gain is nonnegative.* This directly follows from (16); note there is no need for DM to know the details of how $B = b(Y, U, \theta, \delta(Y))$ is defined, as long as the promise $B_{\theta,\delta(Y)} \leq \bar{R}^{-1}(\delta)$ always holds.

In all examples to follow, we consider a *simple* bookie, by which we mean that the $B = b(y, u)$ presented is a single number that can be written as a function of Y and U only, and does not vary with θ or a. The effect of this is that, both for a Bayesian DM and for an e-posterior minimax DM, knowledge of B does not give any incentive to change the action taken: for any given prior W, a decision rule δ is Bayes relative to loss function L iff it is Bayes relative to loss function $L'(\theta, a) := b(y, u) \cdot L(\theta, a)$; and the same holds for e-posterior minimaxity. Bookies who provide a $B_{\theta,a}$ depending on θ and a are certainly worth looking at as well, but the analysis is substantially more difficult and will be left for future work.

Example 6 [Policy Makers vs. the Decision Maker Interpretation] While the bookie interpretation may seem rather abstract, here we show that it can translate into practically relevant settings in which risk bound (16) has actual decision-theoretic consequences, allowing one to reliably make more extreme decisions if the data are more extreme. In standard hypothesis testing problems, this cannot happen, since there are only two actions, 'reject' and 'accept', and then 'more extreme actions' do not exist. But in practical testing problems there is often an additional factor, measuring the *importance* of the decision in light of the observed data. The following real-world example (in essence taken from Grünwald (2022)) illustrates: a study about vaccine efficacy (VE) in a pandemic has been set up as a test between null hypothesis VE $\leq 30\%$ and alternative VE $\geq 50\%$. The original plan was to vaccinate all people above 60 years of age if the null is rejected. But suppose the null actually gets rejected with unexpectedly strong evidence for the alternative — there is substantially more evidence than policy makers had hoped for — and at the same time the virus' reproduction rate may be much higher than anticipated. Based on the observed data and the changed circumstances, the policy makers might now contemplate an alternative action: vaccinate every adult, instead of everyone above 60. This would change the losses of all decisions. We can model this (in an admittedly highly stylized way) by assuming that the loss incurred is really $B \cdot L(\theta, a)$ where B is some positive number depending on how the decision '$\delta(y) = a$' really gets translated into practical measures. For example, $\delta(y) = 1$ (vaccine effective) may translate into 'vaccinate everyone over the age of V'. If $L(1, 1)$ and $L(0, 0)$ are both negative (as we allow), then the lower V, the higher the value of B will be. As indicated above, we simplify things by assuming that the $B = b(y, u)$ ultimately chosen does not depend on θ or the action taken; all losses are multiplied by the same amount. In practice, the employed weight B might arise from a government committee (say the 'policy makers') that analyzes the vaccine study results independently from DM, potentially in a very different way (e.g.

fully Bayesian, or just informally by eyeballing graphs). The more convincing they determine the evidence to be, the higher the B they may propose; similarly, B may depend on external U — if the question is more pressing then one might want to take more drastic measures so the B's go up. The role of the DM (who should now perhaps be called the 'final risk checker') is simply to formulate an absolute upper bound ℓ for the risk and tell the policy makers: as long as you guarantee policies so that $B \cdot \bar{R}(Y, \delta(Y)) \le \ell$, I can guarantee you a risk of no more than ℓ (here ℓ should be determined before the data become available; in contrast, the details of how B arises need not be known and can be arrived at by 'vague' means, as long as B is conditionally (given the data) independence of θ and a).

The Quasi-Conditional Stance   By stressing the importance of (16), we adopt a philosophy closely related to Neyman's (1950) influential *inductive behaviour* view of statistics — we cannot say anything about our performance for an individual study, but we can make claims about our expected performance (or, by the law of large numbers and concentration inequalities, with high probability, about our average performance in several runs). One can say less than under strict Bayesian assumptions, i.e the assumption that one fully believes one's prior: in that case, the assessments one makes will be conditionally correct, and not just in the average sense of (16). But there is still a 'conditional' aspect to (16), in the informal sense that more extreme data (small $\bar{R}(\delta)$) leads to a decision rule being acceptable at higher 'stakes' (value of B). More formally, upon observing $\bar{R}(\delta) = r$, a Bayesian who insists on risk bounded by ℓ would condition on this event and hence accept importance weight B iff $B \cdot r \le \ell$ — a simple bookie can present any B he likes and the Bayesian can decide based on r whether to accept or not. Our DM can safely accept B if the subtly stronger property holds that it is known *a priori* that the bookie must present a B with $B \cdot \bar{R}(\delta) \le \ell$, no matter the value $\bar{R}(\delta)$ takes, or equivalently, DM can tell, after observing $\bar{R}(\delta) = r$, bookie that he must present a $B \le \ell/r$ and bookie is guaranteed to obey. Thus, our DM can safely accept higher stakes if better bounds are observed and is thus more flexible than a DM in the unconditional frequentist Neyman-Pearson setting; but only a subset of the games that are acceptable to the conditioning Bayesian is acceptable to the DM — in this sense, she is 'quasi-conditioning'. Such substantial consequences ensuing from subtle differences in the protocol between bookie and DM are also found, in a different context, in (Grünwald and Halpern, 2011), and are a subject matter of the Vovk-Shafer theory of *game-theoretic probability* (Shafer and Vovk, 2019). They motivate the term *quasi-conditional*, emphasizing that the DM may behave in some but not all senses as if conditioning on the data; and that her final performance is evaluated in expectation over all possible data, and not conditionally on any given data.

   To our knowledge, this is a new paradigm.   It is certainly different from the main existing attempts to unify Bayesian ideas (represent posterior uncertainty by a data-dependent distribution over parameters) and frequentist guarantees — here we think of (a) objective Bayes methods with *matching* priors chosen to get frequentist coverage (Berger et al., 2022); (b) confidence and the related *fiducial* distributions (Schweder and Hjort, 2016); and (c) conditionalist frequentist methods such as BBW's. The example below illustrates this for (a) and (b), focusing on the normal location family, for which matching objective Bayes priors, standard objective Bayes priors, confidence distribution and fiducial distribution all agree. It shows that — despite the matching! — if one were to use them as if they were standard posteriors in decision problems with an incentive to play more extreme actions with more extreme data, they fail to give proper frequentist error bounds. (c) is more subtle: in Section 3.2

we show that BBW and other conditionalist frequentist approaches can be re-interpreted in terms of e-posteriors — so they can be used against arbitrary loss functions even though there were not designed as such. However, these e-posteriors only give meaningful ($< \infty$) bounds for a small subset of all decision rules (in the BBW case, effectively only for a single one (!)), making them highly restricted in practice — whereas other e-collections such as the Savage-Dickey and LR ones give meaningful bounds for essentially any decision rule.

Example 7 [Weighted Squared Error] Take the Gaussian location family with variance 1 as in Example 2 and 3, where we allow data-dependent stopping times such as 'τ is the smallest n at which $x_n > 5$'), defined in the standard manner (Williams, 1991, Hendriksen et al., 2021). Hence, data Y takes the form $Y = (\tau, X^\tau)$ with τ a stopping time and X^τ a vector of τ outcomes. Suppose we aim to find the estimator with smallest mean square error, where the importance of the problem at hand can depend on the data via some function $B = b(Y)$ (we will ignore external circumstances U); it is known to DM that B is some deterministic function of Y. We thus set

$$L(\theta, a) = (\theta - a)^2, L_y(\theta, a) := b(y)L(\theta, a) \tag{18}$$

where L_y is the loss function we are interested in in the end, a squared error loss with importance weighted by some function $b(y)$. In general, $B = b(Y)$ may again be determined by policy makers who may reconsider, using potentially vague ways, the importance of the problem in light of the observed data as in Example 6 above. We can think of any estimator $\breve{\theta}$ for θ as a decision rule, that, upon observing $Y = (n, X^n)$, outputs $\breve{\theta} := \breve{\theta}(n, X^n)$.

We first consider how one would go about this problem in an 'objective Bayes' manner. The standard (uniform, improper) 'objective Bayes' prior for this family and data x^n corresponds to the limit for $\lambda \downarrow 0$ in (10). The ensuing posterior $W^\circ \mid x^n$ has a normal density $w^\circ(\theta \mid x^n)$ with mean and median equal to the maximum likelihood estimator (MLE) $\hat{\theta}(x^n) = n^{-1} \sum_{i=1}^n x_i$ and variance $1/n$, and it is 'matching' (Berger et al., 2022) in the sense that $1 - \alpha$ credible posterior intervals (taken symmetrically around the posterior mean) coincide exactly, for each fixed n, with $1 - \alpha$ confidence intervals. In this case the objective Bayes posterior also coincides with the *fiducial* and the *confidence* distribution (Schweder and Hjort, 2016) based on x^n. The name 'confidence distribution' already warns that this posterior should perhaps only be used for 'confidence' (statements about confidence intervals) but in the fiducial and objective Bayes world, it seems to have been advocated more generally. That this is problematic was already exhibited in different settings by (Grünwald, 2018, Example 7) and Balch et al. (2019). As is well-known, Bayesian posteriors do not depend on the definition of the stopping time (Berger and Wolpert, 1988, Hendriksen et al., 2021); also, they condition on $Y = (n, X^n)$ so further conditioning on $B = b(Y)$ also does not affect them, since it is completely determined by Y. Thus we have $w^\circ(\theta \mid Y = (n, X^n), B = b_0) = w^\circ(\theta \mid x^n)$ (Hendriksen et al., 2021) so that the Bayes optimal action based on (18) and $W^\circ(\theta \mid Y = (n, x^n), B = b_0)$ is simply the MLE, $\hat{\theta}(x^n)$ irrespective of the definition of τ, the definition of b and the value of $b(y)$. Now, based on this posterior, the loss we *think* we make upon observing $Y = (n, x^n)$ is given, as in (2), by

$$\bar{R}^{\text{BAYES}}(W^\circ \mid n, x^n) = \mathbf{E}_{\bar{\theta} \sim W^\circ \mid x^n}[L_y(\bar{\theta}, \hat{\theta})] = \frac{b(n, x^n)}{n}, \tag{19}$$

where we used that $W^\circ \mid Y$ is a normal with mean $\hat{\theta}$ and variance $1/n$, independently of the definition of the stopping time random variable τ. Therefore, if data are in fact sampled from

θ^*, then the average of the losses we expect to make, in several studies, is given by

$$\mathbf{E}_{(\tau,X^\tau)\sim P_{\theta^*}}\left[\mathbf{E}_{\bar{\theta}\sim W^\circ|X^\tau}[L_Y(\bar{\theta},\hat{\theta}(X^\tau))]\right] = \mathbf{E}_{(\tau,X^\tau)\sim P_{\theta^*}}\left[\frac{b(\tau,X^\tau)}{\tau}\right], \tag{20}$$

whereas the loss we *actually* make on average is

$$\bar{R}^* := \mathbf{E}_{(\tau,X^\tau)\sim P_{\theta^*}}[L_Y(\theta^*,\hat{\theta}(X^\tau))] = \mathbf{E}_{(\tau,X^\tau)\sim P_{\theta^*}}[b(\tau,X^\tau)\cdot(\theta^*-\hat{\theta}(X^\tau))^2] \tag{21}$$

We now show that the difference between believed risk (20) and actual (21) can be unbounded, i.e. we can get it to be at least k^*, for arbitrarily large $k^* > 0$. To this end, simply set $b(\tau,X^\tau) := \tau$ and set the stopping time τ to be the smallest n such that $(\hat{\theta}(x^n)-\theta^*)^2/n \geq k^*$. By the law of the iterated logarithm (Williams, 1991) we know that τ is finite, P_{θ^*}-almost surely, under any $\theta^* \in \Theta$. With this definition of b, the believed risk (20) is 1 yet the actual risk (21) is at least k^*. By Proposition 1 such a discrepancy is impossible if we base our risk assessment on e-posteriors via (5) instead. This will be made concrete in Example 8 and 11 below.

3 A Small Taxonomy of E-Specifications

Existing papers on e-variables such as Turner et al. (2021), Pérez-Ortiz et al. (2022), Wasserman et al. (2020) consider P_θ that define a random process $X_1, X_2, \ldots$, and then they provide constructions for an e-variable for sample size n, for each n. Thus, they really provide e-*specifications*: a sequence of e-variables $S_\theta^{[1]}, S_\theta^{[2]}, \ldots$, for each $\theta \in \Theta$, where $S_\theta^{[n]}$ is an e-variable for data $Y = (X_1, \ldots, X_n)$.

Straightforwardly, upon defining $q(X^n) = p_\theta(X^n)S_\theta^{[n]}(X^n)$, we have $\int q(x^n)d\mu(x^n) \leq 1$ where μ is the underlying measure, so q can be thought of as a probability density (if $\int q d\mu < 1$, we may think of the missing mass being put on an outcome that, under P_θ, has probability 0). $S_\theta^{[n]}$ can then be written as a likelihood ratio $q(X^n)/p_\theta(X^n)$. Thus, in the *simple* (no-nuisance) parameter case, *we can rewrite every e-variable as a likelihood ratio*. Importantly though, q is allowed to depend on n and θ; the specification $S_\theta^{[1]}, S_\theta^{[2]}, \ldots$ may deliver, for different n and θ, entirely different q's. To emphasize this possibility, we will from now on make the potential dependence on n and θ explicit, and henceforth write

$$S_\theta^{[n]} = \frac{q_{\rightsquigarrow\theta}^{[n]}(X^n)}{p_\theta(X^n)}. \tag{22}$$

For an important subclass of e-variable specifications, the corresponding $q_{\rightsquigarrow\theta}^{[n]} = q_{\rightsquigarrow\theta}$ does not depend on n after all (it may still depend on θ). In that case $q_{\rightsquigarrow\theta}$ defines a random process for $X_1, X_2, \ldots$ and the specification $(S_\theta^{[n]})_n$ is called an *e-process* (Ramdas et al., 2022). E-processes have a major advantage: for every e-process, for every stopping time τ (e.g. 'stop at the smallest n at which $S^{[n]} \geq 1/\alpha$, and set τ to this n'), we have (Ramdas et al., 2022):

$$\mathbf{E}_{X^\tau\sim P_\theta}\left[S^{[\tau]}(X^\tau)\right] = \mathbf{E}_{X^\tau\sim P_\theta}\left[\frac{q_{\rightsquigarrow\theta}(X^\tau)}{p_\theta(X^\tau)}\right] \leq 1.$$

As a result, with e-processes we can engage in *optional stopping* — under any stopping time τ — even the aggressive stopping rule mentioned above or some externally imposed rule,

the details of the definition of which we do not need to know — $S_\theta^{[\tau]}$ is still an e-variable, and decision making based on (5) as above can proceed. With e-specifications that do not provide e-processes, one cannot engage in such optional stopping (see the extended discussion by (Grünwald et al., 2019, Section 5)). Indeed, in Example 7, which involved a non-constant stopping time τ, one only gets valid assessments with e-processes rather than mere e-variables, and we shall use these in Example 8 and 11 to provide risk assessments for that setting.

We will now reconsider the e-collections of all our running examples — as we will see, some provide e-processes, others do not.

3.1 Generalized Savage-Dickey E-Processes

We have already seen two instances of e-processes: the (specification, for each n, of) Savage-Dickey e-variables S^{SD} and the likelihood ratio e-variables S^{LR}. Both are instances of what we may call *generalized Savage-Dickey e-processes* $(S^{[n]})_\theta$, with

$$S_\theta^{[n]} = \frac{p_{W \leftsquigarrow \theta}(X^n)}{p_\theta(X^n)} \tag{23}$$

where $p_W(X^n) = \int p_\theta(X^n) dW(\theta)$ is a Bayes marginal distribution on n. Thus, the e-variable numerator is a Bayes marginal, but the prior W one uses is allowed to depend on the parameter θ one is comparing to.

Example 8 [Gaussian Location, Example 3 Continued] Fix some *anticipated* sample size n^*, representing our best a priori guess of the actual sample size n that we will observe. Our aim is to obtain an e-process that is valid for all n, including $n \neq n^*$, but is optimized to give the best possible risk bound if we happen to observe $y = x^n$ with n equal or close to n^*. To this end, for each $\theta \in \mathbb{R}$ we define $\theta^- < \theta$ and $\theta^+ > \theta$ to satisfy (12) with $C = 1$. The construction below gives a valid e-process for every $C > 0$, but $C = 1$ gives the best bounds for the squared error, at the cost of potentially somewhat worse yet of course still valid, bounds for other loss functions. Now define, for each n, the e-variable $S_\theta^{[n]}(y) = \frac{(1/2)p_{\theta^-}(y) + (1/2)p_{\theta^+}(y)}{p_\theta(y)}$. Clearly the $S_\theta^{[n]}$ form a generalized Savage-Dickey e-process, with $W \leftsquigarrow \theta$ the uniform distribution on $\{\theta^-, \theta^+\}$. In the Supplementary Material (Section B, Proposition 3) we show that for the corresponding e-posterior $\bar{P}(\theta \mid y)$, we have that

$$\min_a \max_{\theta \in \mathbb{R}} (\theta - a)^2 \bar{P}(\theta \mid y)$$

is achieved by setting $a = \hat{\theta}$ (independently of n and n^*), so that $\delta^{\mathrm{MM}}(y) = \hat{\theta}$, and we also show that (Section A) (5) evaluates to

$$\max_{\theta \in \mathbb{R}} (\theta - \hat{\theta})^2 \bar{P}(\theta \mid y) = \frac{1}{n} \cdot g(n, n^*) \cdot 1.45 \ldots \overset{\text{if } n^* = n}{=} \frac{1}{n} \cdot 1.45 \ldots \tag{24}$$

where $g(n, n^*) = \left(\frac{n^*}{n}\right) \cdot e^{n/n^* - 1}$. Note that $g(\cdot, n^*)$ has a minimum at $n = n^*$ with $g(n, n^*) = 1$ and increases very slowly for n/n^* close to 1. The maximum over θ is, for $n = n^*$, achieved approximately at $\theta = \hat{\theta} \pm 1.46/\sqrt{n}$ at which $\bar{P}(\theta \mid y) \approx 0.68$, independently of the value of $n = n^*$. $S_\theta^{[n]}$ gives an e-process for any n^* we choose in (12). In the Supplementary Material, Section A, we show that (24) still holds up to lower order factors, for general 1-dimensional families, as long as they are restricted to a compact subinterval of the parameter space.

13

Incidentally, this shows that for *standard* decision problems (e.g. n fixed in advance, uniform weight function), the e-posterior based risk bounds (5) are not much worse than the Bayesian risk assessment and minimax risk bounds for the same model, as long as sufficiently 'clever' e-collections are used: the factor is a modest 1.45 if n can be assumed known in advance, and a $\log n$-factor otherwise.

Example 9 [Hypothesis Testing with Generalized Savage-Dickey and LR] Consider a simple-vs.-simple test as in Example 4. From (23) we see that the generalized Savage-Dickey e-variable at fixed sample size n, with $Y = X^n$, and based on priors $W \leftsquigarrow \theta$ with mass function w_θ, for $\theta \in \Theta = \{0,1\}$, is given by

$$S_\theta(y) = \frac{w_\theta(0)p_0(y) + w_\theta(1)p_1(y)}{p_\theta(y)} \text{ with } w_\theta(1) = 1 - w_\theta(0). \tag{25}$$

For degenerate priors with $w_0(0) = w_1(1) = 0$ this reduces to the LR e-process of (13), and for 'pure' (i.e. nongeneralized) Savage-Dickey e-posteriors, we have $w_0 = w_1$. Consider Wald-Neyman-Pearson loss functions under Condition Zero (**C0**) as in Example 4. Generalizing (13), the e-posterior minimax decision rule δ^{MM} upon observing $Y = y$ is then given by $\delta^{\mathrm{MM}}(y) = a$, for a achieving

$$\min_{a \in \{0,1\}} \max \left\{ L(0,a) \cdot \frac{p_0(y)}{w_0(0)p_0(y) + w_0(1)p_1(y)}, L(1,a) \cdot \frac{p_1(y)}{w_1(0)p_0(y) + w_1(1)p_1(y)} \right\}$$

$$\overset{\mathbf{C0}}{=} \min_{a \in \{0,1\}} \left(L(0,a) \cdot \frac{p_0(y)}{w_0(0)p_0(y) + w_0(1)p_1(y)} + L(1,a) \cdot \frac{p_1(y)}{w_1(0)p_0(y) + w_1(1)p_1(y)} \right). \tag{26}$$

It follows that under $\mathbf{C_0}$, δ^{MM} selects hypothesis 0 ($a = 0$) iff $f\left(\frac{p_0(y)}{p_1(y)}\right) \geq \frac{L(1,0)}{L(0,1)}$ where $f(r) = r \cdot \frac{(1-w_1)r+w_1}{(1-w_0)r+w_0}$ with $w_j = w_j(1)$. Straightforward calculus gives that, irrespective of the choice for w_0 and w_1, $f(r)$ is strictly increasing in r, so that there is a unique point r^* depending on w_1 and w_0, such that for all $r \geq r^*$, $\delta^{\mathrm{MM}}(y) = 0$ and for all $r < r^*$, $\delta^{\mathrm{MM}}(y) = 1$. Further, for any pure Savage-Dickey e-collection with $w_1 = w_0$, we have $f(r) = r$, i.e. we select action 1 if

$$p_1(y) \cdot (L(1,0))^\gamma \geq p_0(y) \cdot (L(0,1))^\gamma \tag{27}$$

for $\gamma = 1$. This is equivalent to taking the Bayes decision under a uniform prior. On the other hand, for the LR e-collection, $\delta^{\mathrm{MM}}(y)$ selects action 1 if the above holds for $\gamma = 1/2$. Finally, if the loss is symmetric, $L(0,1) = L(1,0)$, then in fact we select action 1 iff $p_1(y) > p_0(y)$ even for general $w_1 \neq w_0$.

How about our basic risk assessment (5) for general decision rules, including these minimax ones? While (5) does not give rise to easily interpretable formulae that work for all generalized Savage-Dickey e-variables, under **C0** we do get intuitive expressions for pure Savage-Dickey e-variables with $w_0 = w_1 = 1/2$ the uniform prior, and for LR e-collections. For the former, we can replace the 'max' by a '+' using **C0** as in (26) and we get:

$$L(0,\delta(y)) \cdot \frac{p_0(y)}{\frac{1}{2}p_0(y) + \frac{1}{2}p_1(y)} + L(1,\delta(y)) \cdot \frac{p_1(y)}{\frac{1}{2}p_0(y) + \frac{1}{2}p_1(y)} =$$

$$2 \cdot L(0,\delta(y)) \cdot w_{\mathrm{U}}(0 \mid y) + 2 \cdot L(1,\delta(y)) \cdot w_{\mathrm{U}}(1 \mid y) = 2 \cdot \mathbf{E}_{\theta \sim W_{\mathrm{U}}|y} \left[L(\theta,\delta(y)) \right] \tag{28}$$

where $w_\mathrm{U}(j \mid y) = \frac{p_j(y)}{p_0(y)+p_1(y)}$ is the posterior based on uniform prior w_U with $w_\mathrm{U}(0) = w_\mathrm{U}(1) = 1/2$, showing that our bound is always within a factor 2 of the best Bayes-with-uniform-prior bound and hence within a factor of less than 2 of the standard frequentist minimax bound for this setting, obtainable by a procedure that depends on the specific L of interest. For the LR e-variables, we similarly get, refining (13),

$$L(0,\delta(y)) \cdot \frac{p_0(y)}{p_1(y)} + L(1,\delta(y)) \cdot \frac{p_1(y)}{p_0(y)} =$$
$$L(0,\delta(y)) \cdot \frac{w_\mathrm{U}(0 \mid y)}{w_\mathrm{U}(1 \mid y)} + L(1,\delta(y)) \cdot \frac{w_\mathrm{U}(1 \mid y)}{w_\mathrm{U}(0 \mid y)} = \mathbf{E}_{\theta \sim w_\mathrm{U}|y}\left[\frac{1}{w_\mathrm{U}(\delta(y) \mid y)} \cdot L(\theta,\delta(y)) \right]. \quad (29)$$

We note that for $\delta := \delta^{\mathrm{MM}}$ and the symmetric case $L(0,1) = L(1,0)$, we choose action 1 iff $p_1(y) \geq p_0(y)$ so that $w_\mathrm{U}(\delta^{\mathrm{MM}}(y) \mid y) \geq 1/2$: the LR bound (29) dominates the uniform-prior Savage-Dickey bound (28), and $1/w_\mathrm{U}(\delta^{\mathrm{MM}}(y) \mid y)$ in (29) is a *luckiness factor*, bounded by 2: if we happen to assign high posterior to the selected hypothesis, we are 'lucky' and our loss assessment will get tighter; still, it is valid irrespective of the value of this posterior. If asymmetric losses are observed, $1/w_\mathrm{U}(\delta^{\mathrm{MM}}(y) \mid y)$ can in general be both larger and smaller than 2, and the LR risk assessment bound does not dominate the Savage-Dickey one.

3.2 Conditionalist-Frequentist E-Specifications — not E-Processes

We first continue the *Berger-Brown-Wolpert* (BBW) simple-vs.-simple hypothesis testing setting of Example 5, later extending it to more general conditionalist frequentist inference. Let $V := p_1(Y)/p_0(Y)$. We will assume that (a) V is strictly increasing in Y, which (b) has a (Lebesgue) density under P_0 and P_1. Both assumptions are merely for simplicity; for example (a) can be dropped by directly analyzing V instead of Y, i.e. 'the likelihood ratio of the likelihood ratio'.

Fix some w with $0 < w < 1$. Since, for any y^*, the sum of both integrals below is 1, there must be a specific y^* such that

$$\int_{y<y^*} ((1-w)p_0(y) + wp_1(y))dy = w \; ; \; \int_{y \geq y^*} ((1-w)p_0(y) + wp_1(y))dy = 1 - w, \quad (30)$$

so that, upon setting

$$S_0^{\mathrm{BBW}(w)}(y) = \mathbf{1}_{y \geq y^*} \frac{(1-w)p_0(y) + wp_1(y)}{(1-w)p_0(y)} \; ; \; S_1^{\mathrm{BBW}(w)}(y) = \mathbf{1}_{y < y^*} \frac{(1-w)p_0(y) + wp_1(y)}{wp_1(y)} \quad (31)$$

we can verify that $\mathbf{E}_{Y \sim P_\theta}[S_\theta^{\mathrm{BBW}(w)}] = 1$ for $\theta \in \{0,1\}$ so that $\{S_\theta^{\mathrm{BBW}(w)} : \theta \in \{0,1\}\}$ is an e-collection. The corresponding e-posteriors only give useful risk assessment bounds (5) in quite restricted situations, however. To be sure, the bound (5) holds (as it must by Proposition 1), but it must be trivial for all decision rules except one as soon as we assume — as will be the case for every Wald-Neyman-Pearson loss, i.e. with $L(0,1) > 0, L(1,0) > 0$. The reason is that $\bar{P}^{\mathrm{BBW}(w)}(\theta = 1 \mid y) = \infty$ if $y \geq y^*$ and $\bar{P}^{\mathrm{BBW}(w)}(\theta = 0 \mid y) = \infty$ if $y < y^*$: for every loss function L and decision rule δ there will be y such that $\bar{R}(y,\delta(y)) = \infty$, unless $L(0,0) \leq 0, L(1,1) \leq 0$ and δ is the $\mathrm{BBW}(w)$-*decision rule*, which satisfies $\delta^{\mathrm{BBW}(w)}(y) = 1$ iff $y \geq y^*$. This is, in fact, the only decision rule that BBW themselves study. Note that it must then also be the case that the minimax decision rule δ^{MM} based on $\bar{P}^{\mathrm{BBW}(w)}$ must coincide with $\delta^{\mathrm{BBW}(w)}$, and thus

does not depend on the loss L of interest. This indicates that, for many loss functions, it will not be a very good rule — it is simply the only rule for which the risk assessment (5) is nontrivial (and then evaluates to (14)). To be sure, BBW provide a result implying that if the loss function L satisfies **C0**, then we may set w as a function of this loss function so that the rule $\delta^{\mathrm{BBW}(w)} = \delta^{\mathrm{MM}}$ is to some extent reasonable after all in the following sense: it coincides with the Bayes decision rule $\delta^{\mathrm{BAYES}}(y)$ based on prior W with $w(1) = w$ as long as $y \notin [y^-, y^+]$ for some specific values $y^- \le y^+$. They call $[y^-, y^+]$ the *no-decision region* and notice that in practice it is often quite small.

This support for $\delta^{\mathrm{BBW}(w)}$ notwithstanding, we feel that the pure Savage-Dickey (with $w = 1/2$) and LR e-posteriors should be preferred: at the small price of the slightly larger risk assessments (28) and (29), they (a) allow us to use these assessments for general decision rules rather than the potentially defective (since independent of b) $\delta^{\mathrm{BBW}(w)}$, and (b) they provide e-processes rather than just e-specifications (it is easy to see that the likelihood ratios (22) corresponding to BBW e-specifications have $q_\theta^{[n]}$ depending on n — BBW e-variables do not remain e-variables under optional stopping, generalized Savage-Dickey e-variables do).

BBW's Original, Conditionalist-Frequentist Interpretation   For general p_0 and p_1 as above, we may partition the sample space $\mathcal{Y}$ into regions $\{\mathcal{Y}_\alpha : \alpha \in [0,1]\}$ where $\mathcal{Y}_\alpha = \{y_\alpha^+, y_\alpha^-\}$ are chosen such that $P_0(Y \ge y_\alpha^+) = \alpha$ and $P_1(Y \le y_\alpha^-) = \beta(\alpha)$ for some function $\beta : [0,1] \to [0,1]$. We may think of observing $Y = y$ as first observing α so that $y \in \mathcal{Y}_\alpha$, and only then observing $Y = y$ itself. We may then, by construction, after observing α, perform *a conditional frequentist test* with Type-I and Type-II 'conditional' error probabilities α and $\beta(\alpha)$, by rejecting $\theta = 0$ if $Y = y_\alpha^+$ and accepting if $Y = y_\alpha^-$. The BBW approach was originally conceived as a specific way to construct such a partition, with the special property that for any α we have that $y_\alpha^+ \ge y^*, y_\alpha^- \le y^*$ (so that δ^{MM} invariably rejects if y_α^+ is observed for some α) and

$$\alpha = \frac{(1-w)p_0(y_\alpha^+)}{(1-w)p_0(y_\alpha^+) + wp_1(y_\alpha^+)}, \quad \beta(\alpha) = \frac{wp_1(y_\alpha^-)}{(1-w)p_0(y_\alpha^-) + wp_1(y_\alpha^-)}.$$

These are exactly our BBW e-posteriors $\bar{P}^{\mathrm{BBW}(w)}(0 \mid y_\alpha^+)$ and $\bar{P}^{\mathrm{BBW}(w)}(1 \mid y_\alpha^-)$ — BBW in fact call them conditional error probabilities, because that is how they derived them, and note that they are formally identical to Bayesian posteriors; the interpretation in terms of e-posteriors is ours.

Kiefer and Vovk Confidence Estimators   Essentially the same analysis as for BBW conditional error probabilities also applies to general *conditionalist frequentist e-variables*, based on Kiefer's and Vovk's conditional confidence estimators, to which we now turn: every such confidence estimator gives a valid e-specification, hence e-posteriors for each n, but in general these will not be e-processes, and they can evaluate to infinity, hence only giving useful, nontrivial risk assessments for a restricted set of loss functions and decision rules — although this set may not be as restricted as in the BBW case, where only a single rule, $\delta^{\mathrm{BBW}(w)}$ gave nontrivial risk assessments.

Consider a partition $\{\mathcal{Y}_\alpha : \alpha \in [0,1]\}$ of $\mathcal{Y}$ together with a function $\mathrm{CI}(y)$ that maps each $y \in \mathcal{Y}$ into a subset of Θ. In case Θ is a continuum, we usually take $\mathrm{CI}(y)$ to be an interval, so that $\mathrm{CI}(y) = [\theta_L, \theta_R]$ with endpoints θ_L and θ_R determined by y. We say that $(\{\mathcal{Y}_\alpha : \alpha \in [0,1]\}, \mathrm{CI})$ define a *Kiefer-type conditional frequentist confidence set estimator* if

for all $\theta \in \Theta$, all $\alpha \in [0, 1]$,

$$P_\theta(\theta \notin \mathrm{CI}(Y) \mid Y \in \mathcal{Y}_\alpha) \le \alpha.$$

Such confidence estimators were first considered by Kiefer (1976). If $\beta(\alpha) = \alpha$, the BBW approach can be reinterpreted as a Kiefer-type confidence estimator: we set $\mathcal{Y}_\alpha$ as above and

$$\mathrm{CI}(y) = \{\delta^{\mathrm{BBW}(w)}(y)\} \text{ so that } \mathrm{CI}(y_\alpha^+) = \{1\}, \mathrm{CI}(y_\alpha^-) = \{0\}. \tag{32}$$

If $\beta(\alpha) \ne \alpha$ then the BBW approach does not give a Kiefer-type confidence estimator. However, Vovk (1993) provided a notion of confidence estimator that generalizes both BBW's conditional error probabilities and Kiefer-type confidence estimators: for any function $\mathrm{CI}(y)$ that maps each $y \in \mathcal{Y}$ into a subset of Θ combined with any function $\hat{\alpha}(Y)$ taking values in $[0, 1]$, we say that CI and $\hat{\alpha}$ define a *Vovk-type confidence set estimator* if for all θ,

$$\mathbf{E}_{Y \sim P_\theta}\left[\mathbf{1}_{\theta \notin \mathrm{CI}(Y)} \cdot \frac{1}{\hat{\alpha}(Y)}\right] \le 1. \tag{33}$$

Vovk (1993) defines confidence set estimators in this partition-free manner, motivated by a game-theoretic interpretation. He gives a (straightforward) proof showing that any confidence set estimator in the sense of Kiefer is also a confidence set estimator in the sense of Vovk, with $\hat{\alpha}(y)$ set to the α such that $y \in \mathcal{Y}_\alpha$. Similarly, if we set CI as in (32) and, for each $\alpha \in [0, 1]$, $\hat{\alpha}(y_\alpha^+) := \alpha, \hat{\alpha}(y_\alpha^-) = \beta(\alpha)$, then it is easily seen that (33) hold with equality for both $\theta \in \{0, 1\}$, i.e. we obtain a Vovk-type confidence estimator.

Now, if $(\mathrm{CI}(Y), \hat{\alpha}(Y))$ is a confidence set estimator in sense of Vovk, it directly follows from (33) that for all θ, $S_\theta = \mathbf{1}_{\theta \notin \mathrm{CI}(Y)}1/\hat{\alpha}(Y)$ is an e-variable, so $\bar{P}(\theta \mid y)$ is an e-posterior. Just as in the specific BBW scenario, such a posterior may become ∞.

4 Condition Zero and Capped Posteriors

Consider Example 2 again, the normal location family with the squared error loss (18) as in Example 7. If we base our risk assessment directly on the Savage-Dickey e-posterior as defined there, we get bounds that, while correct, are not always sharp. This is due to the fact that, as seen from Figure 1, $\bar{P}(\theta \mid Y) = S_\theta^{-1}(Y)$ can become very high for θ close to $\hat{\theta}$ — a milder form of the issues with the confidence-estimator based e-variables of the previous section, for which it can become even infinite. There, the issues were avoided if the loss function satisfied Condition Zero. It turns out that a weakened version of this condition is sufficient to avoid this problem, for decision problems more general than testing. We now introduce a new concept, and weaken and generalize Condition Zero with this in mind:

Definition 1 [Capped E-Posterior; Condition Zero] *For any given e-posterior $\bar{P}(\theta \mid y)$, its* capped *version $\bar{P}^{\mathrm{CAP}}(\theta \mid y)$ is defined by setting, for all $\theta \in \Theta$, $y \in \mathcal{Y}$: $\bar{P}^{\mathrm{CAP}}(\theta \mid y) := \min\{1, \bar{P}(\theta \mid y)\}$. Let δ be a decision rule. We say that δ satisfies* **C0***, i.e. Condition Zero, if*

$$\sup_{\theta \in \Theta} \bar{P}(\theta \mid y)L(\theta, \delta(y)) = \sup_{\theta \in \Theta} \bar{P}^{\mathrm{CAP}}(\theta \mid y)L(\theta, \delta(y)) \tag{34}$$

for every $y \in \mathcal{Y}$. We say that the loss function L *itself satisfies Condition Zero if there exists a decision rule δ such that (34) holds.*

Our original Condition Zero in simple-vs.-simple testing was defined as a condition on a given loss function L; the present condition is strictly weaker as a condition on loss functions and generalizes the idea to decision rules δ. Why would we want to impose such a condition?

First note that setting $S_\theta = 1$ for all $\theta \in \Theta$ gives a trivial, yet valid, e-collection; it provides no indication at all about θ based on data y. This suggests that, when using a nontrivial collection, based on data y, we cannot make any inferences about θ for which $S_\theta(y) < 1$ (i.e. $\bar{P}(\theta \mid y) > 1$): our e-variable provides less evidence against θ than what can be obtained trivially: in a sense, y gives us no information at all for making inferences about θ. It is therefore of interest to restrict to decision rules satisfying the new Condition Zero. For such problems, our risk assessment will *never* be based on evaluating $\bar{P}(\theta \mid y)$ at a value of θ for which $\bar{P}(\theta \mid y) > 1$, i.e. for which it really does not give any information about this θ. Also, it may be argued that the capped e-posterior is a more intuitive tool than the e-posterior: for any $\alpha \geq 0$, the set $\{\theta \in \Theta : \bar{P}(\theta \mid y) \geq \alpha\}$ is a $1 - \alpha$-confidence interval (Grünwald, 2022) (see Figure 1, middle panel), whereas for sets containing θ with $\bar{P}(\theta \mid y) \geq 1$, no clear interpretation seems to exist. Thus, by insisting on **C0** one really insists on inferences that can be based on the more intuitive $\bar{P}^{\text{CAP}}$ while still satisfying (5).

Example 10 [Two-Point prior for exponential families: Condition Zero holds]
The generalized Savage-Dickey two-point prior-based e-posterior that we employed for the Gaussian location family in Example 8 gives rise to e-collections such that the corresponding minimax $\delta^{\text{MM}}(y) = \hat{\theta}(y)$, also satisfy Condition Zero whenever the b-weighted squared error (18) is used: independently of n, $\bar{P}(\theta \mid y) \approx 0.68$ for the maximizing θ, as was established below (24). For exponential families, we show in the Supplementary Material (Section A, inside the proof of Theorem 1) that the same holds if the model $\{P_\theta : \theta \in \Theta^\circ\}$ is restricted to an (arbitrary) compact subset of the parameter space and n is large enough. We do not know if the result can be extended beyond exponential families, since its proof makes heavy use of their properties.

In practice, the DM may be presented with a loss function for which Condition Zero does not hold, for any reasonable decision rule, and the high-risk problem looms. What to do in such a case? There are two alternative solutions. As to the first (we describe the second underneath Example 11), one can (sometimes vastly) improve risk assessments by slightly modifying any given e-variable S_θ to another e-variable S_θ' that provides almost as much evidence against θ as S_θ whenever S_θ provides more evidence against y than the trivial 1 (i.e. when $S_\theta(y) > 1$), and almost as much as the trivial $S = 1$ otherwise. The corresponding $\bar{P}'(\theta \mid y) = S_\theta'^{-1}(y)$ will then be similar to $\bar{P}(\theta \mid y)$ if $\bar{P}(\theta \mid y) \leq 1$ and bounded by a small constant otherwise. Three straightforward ways of doing this are provided by setting S_θ' equal to any of the following options:

$$S_\theta^*(y) := \frac{\max\{S_\theta(y), 1\}}{\mathbf{E}_{Y \sim P_\theta}[\max\{S_\theta(Y), 1\}]} \quad , \quad S_\theta^\circ(y) := \frac{1}{2} \cdot \max\{1, S_\theta(y)\}$$

$$S_\theta^{[\gamma]}(y) := (1 - \gamma) + \gamma S_\theta(y) \text{ for } \gamma = 1/2. \tag{35}$$

One immediately verifies that all three options produce e-variables: $S^{[\gamma]}$ produces one for all γ, because mixtures of e-variables are e-variables. Using that the expectation in S_θ^* is bounded by 2 (to see this, upper bound max by a sum), we see that S° is dominated by both S_θ^* and $S_\theta^{[1/2]}$ hence must also be an e-variable and we also see that for all three choices, we

have $\sup_\theta \bar{P}'(\theta \mid y) \le 2\sup(S_\theta^\circ(y))^{-1} = 2$. S_θ^* may be the most direct solution, but it can turn an e-process into an e-specification that is not an e-process anymore, whereas $S_\theta^{[1/2]}$ preserves the e-process property. Since $S_\theta^{[1/2]}$ dominates S_θ°, we prefer it; it is depicted in Figure 1.

Example 11 [Improving Pure Savage-Dickey for Exponential Families] We illustrate the idea above for the case that $\bar{P}(\theta \mid y) = 1/S_\theta(y)$ is the Savage-Dickey e-posterior based on an arbitrary prior W on the Gaussian location family. Let $\bar{P}^{[\gamma]}(\theta \mid y) := 1/S_\theta^{[\gamma]}(y)$ be the corresponding 'dampened' e-posterior as defined in (35). In the Supplementary Material, Section C we describe this setting in more detail and extend it to 1-dimensional exponential families, also showing (in Proposition 3) that for any $0 < \gamma \le 1$ the MLE $\hat{\theta}$ is the $\bar{P}^{[\gamma]}$-e-posterior minimax estimator irrespective of the weight b in the loss (18). Now for simplicity let W be a Gaussian prior with some fixed precision λ as in Example 2 and with prior mean $\theta_0 = 0$. We also show in the Supplementary Material (Proposition 4), that, for the choice $\gamma = 1/2$ that we advocated above, the bound (15) and (16) hold with

$$\bar{R}(\delta^{\mathrm{MM}}) = \bar{R}(Y, \hat{\theta}(Y)) = \frac{2}{n}\left(\frac{1}{2}\log\frac{n+\lambda}{\lambda} + \frac{n\lambda}{n+\lambda}\hat{\theta}(Y)^2\right), \tag{36}$$

losing just a $\log n$ factor compared to the e-posterior that achieves the standard frequentist minimax bound up to a constant factor, and that we considered in Example 8. Again we extend this result to general 1-dimensional families with smooth and strictly positive prior densities (and again we do not know if it extends to more general 1-dimensional families). In contrast, for the original Savage-Dickey e-posterior $\bar{P}(\theta \mid y)$ we get a much larger $\bar{R}(\delta^{\mathrm{MM}}) \ge n^{-1/2}$, as can be seen by evaluating $\cdot\bar{P}(\theta \mid y)(\theta - \hat{\theta})^2$ at $\theta = \hat{\theta} + n^{-1/2}$. By 'dampening' the posterior we improved the factor of order $n^{-1/2}$ to a factor of order $(\log n)n^{-1}$.

As a second solution to the high-risk problem, we can alternatively *modify* our risk assessment (5) for decision problems and rules that do not satisfy **C0**. We then find that if we replace the e-posterior in the risk assessment $\bar{R}(\delta)$ as in (5) by its capped version, the risk assessment itself may become invalid (i.e. $\bar{R}(\delta)$ does not satisfy (15)/(16)), but, for arbitrary data-dependent decision problems, $2\bar{R}(\delta)$ does remain valid:

Proposition 2 *Set, for arbitrary multi-loss decision problem and rule δ, not necessarily satisfying* **C0***,*

$$\bar{R}^{\mathrm{CAP}}(Y, A) := \sup_{\theta \in \Theta} \bar{P}^{\mathrm{CAP}}(\theta \mid y) \cdot L(\theta, a) \quad, \quad \bar{R}^{\mathrm{CAP}}(\delta) := \bar{R}^{\mathrm{CAP}}(Y, \delta(Y)). \tag{37}$$

Then (15) and (16) hold with $2\bar{R}^{\mathrm{CAP}}(\delta)$ in the role of $\bar{R}(\delta)$.

Proof: Calculating $\bar{R}^{\mathrm{CAP}}(y, a)$ based on the capped posterior is equivalent to calculating $\bar{R}(y, a)$ based on the modified posterior $\bar{P}'(\theta \mid y) = (S_\theta^\circ(y))^{-1}$ as defined in (35) and multiplying the answer by 2. Since $\bar{P}'(\theta \mid y)$ is a standard e-posterior and no requirements on δ are made, we can use Proposition 1 based on $\bar{P}'(\theta \mid y)$ to get the required result. $\square$

This solution to the high-risk problem of some e-collections shows that, for some collections, the original e-posterior in its capped version may not be so bad after all. This solution may in fact be preferable, since it does not interfere with the use of e-values in optional continuation: we can keep multiplying e-values as explained in the next section, and get a factor of 2 in the bound for the final product. In contrast, if we use any of the modifications in (35), we must modify the e-variable at each study, making the final (i.e. when we stop adding studies) product of e-values potentially much smaller.

5 Additional Background and Discussion

The General Definition of E-Values   The definition of e-collection, e-posterior, risk assessment (5) and e-posterior minimax decision rule readily generalize to models with nuisance parameters, composite hypothesis testing and nonparametric settings. All are covered by the following extended definition: let $\mathcal{P}$ be a set of distributions for random variable Y and consider a parameter (vector) of interest defined by a function $\delta : \mathcal{P} \to \Theta$ with $\Theta \subset \mathbb{R}^k$ for some $k > 0$. We now call $\{S_\theta : \theta \in \Theta\}$ an *e-collection* for $\mathcal{P}$ if for all $\theta \in \Theta$, $S_\theta = S_\theta(Y)$ is a nonnegative function of Y and for all $P \in \mathcal{P}$ with $\delta(P) = \theta$, we have the corresponding analogue of (3), i.e. for *all* such P:

$$\mathbf{E}_P[S_\theta] \le 1. \tag{38}$$

The e-posterior $\bar{P}(\theta \mid y)$ relative to such a collection is still defined as $S_\theta^{-1}(y)$, and the definition of risk assessment (5) and e-posterior minimax rule (6) remains unchanged. It is easily checked that the results that validate the e-posterior risk assessment and minimax rule, Proposition 1 and Proposition 2, still hold in this more general setting.

The Original Interpretation of E-Variables   E-variables, while being implicit in earlier work going back at least to Darling and Robbins (1967), were largely unknown until 2019, when the concept was given a name and was popularized and extensively developed by papers such as (Grünwald et al., 2019, Shafer, 2021, Vovk and Wang, 2021). In the mean time they have been studied in a wide variety of contexts (Henzi and Ziegel, 2021, Ren and Barber, 2022, Bates et al., 2022) and even had an international workshop devoted to them. Originally, they were used as tools to extend traditional Neyman-Pearson tests to situations with *optional continuation* while keeping Type-I error guarantees: suppose one observes a sequence of studies concerning the same null hypothesis $\{P \in \mathcal{P} : \delta(P) = \theta_0\}$. One can then multiply the e-values derived from the individual studies. Broadly speaking, the product will be a new e-variable, irrespective of whether the decision to perform a subsequent study depends on previous study outcomes (the 'continuation rule'), and irrespective of the rule used to decide when to stop performing new studies. Since, for any fixed α, the probability, under the null, that an e-variable gets larger than $1/\alpha$, is bounded by α, one can use e-values, combined by multiplication, as an alternative to p-values in a setting in which optional continuation of experiments is allowed (note that this is always possible, whereas optional stopping *within* a study is only possible if we deal with e-processes rather than e-variables, as in Section 3).

Composite Testing   In light of this original motivation, much initial work on e-values focused on hypothesis testing in Neyman-Pearson (NP) style, but under optional continuation. Just as in the original NP paradigm, Type-I error of decision rules was required to be bounded by a pre-given significance level α, and other desiderata such as small Type-II error only came second. This is different from the treatment of simple-vs.-simple testing in the present paper. Here, as in Bayesian and minimax approaches, we do allow asymmetric losses ($L(0,1) \ne L(1,0)$), but these potentially different losses appear in the risk assessment (5) in a symmetric fashion after all.

Existing papers on e-values within the original paradigm in fact emphasize composite hypotheses: the theoretical papers Grünwald et al. (2019), Pérez-Ortiz et al. (2022) and the practical (Turner et al., 2021, Turner and Grünwald, 2022) consider situations in which the null has nuisance parameters (such as, for example, the variance in the t-test or the

proportion in 2×2 tables). For such cases (the null $\mathcal{P} = \{P_\lambda : \lambda \in \Lambda\}$ is characterized by a (potentially) high-dimensional parameter vector λ; the parameter of interest δ is of (potentially) much smaller dimension, often a scalar), Grünwald et al. (2019) offer a generic method to construct e-variables as in (38) for such situations via the *reverse information projection*. This method is employed by all papers mentioned above to construct e-variables, with the explicit goal of providing Type-I error bounds under optional continuation in the Neyman-Pearson style; but we can equally use them to provide e-posterior risk assessments and minimax rules with general loss functions as in this paper — in our examples above, we only used 'simple' e-collections for ease of illustration. Of particular interest is the paper (Grünwald, 2022), which may be viewed as a companion to the present one. It also extends the use of e-variables to deal with post-hoc specified loss functions but it sticks to a variation of the Neyman-Pearson tenet: there is a distinction between Type-I risks (expected losses), on which a hard constraint is imposed (similar to the significance level α as a constraint on Type-I error *probability* in traditional testing), and Type-II risks, which are minimized subject to this constraint.

Nonparametrics The same generalization (38) covers nonparametric settings. For example, in Cox's celebrated proportional hazards model, the $P \in \mathcal{P}$ specify an underlying continuous time process with dynamics conditional on some covariate vector $\bar{x}$. While the full likelihood of this process is intractable, Ter Schure et al. (2021) show that, for any vector θ of covariate coefficients, the ratio between any two partial likelihoods defined relative to the same set of covariates constitute an e-process. This implies that all the results on e-posterior risk assessment of the present paper also pertain to ratios of Cox's partial likelihoods. In fact, for any $\mathcal{P}$ and any $\theta \in \Theta$, any ratio between conditional, marginal or partial likelihoods that has the same distribution under all distributions in $\mathcal{P}$ with $\delta(P) = \theta$ is trivially an e-variable, and can be used to define an e-collection. We regard all this as an additional advantage of e-posteriors over Bayesian ones: strictly speaking, a Bayesian needs to specify the *full* likelihood for each distribution in $\mathcal{P}$ in order to get a valid a posterior — as has been argued by e.g. Robins and Wasserman (2012), this may be viewed as 'overkill' and poses practical (and unnecessary) difficulties for Bayesian nonparametric inferences especially when the samples are small. In contrast, Waudby-Smith and Ramdas (2022) provides an e-collection which can be used to learn the mean of a sequence of bounded random variables based on a small sample without *any* assumptions on the underlying random variables except boundedness and existence of a common mean.

Conclusion and Future Work We have presented the idea of representing uncertainty by e-posteriors based on e-collections. While e-collections already existed, the idea to use them in risk assessments such as (5) and to define e-posterior minimax rules as in (6) is, to our knowledge, completely new. There are of course many open questions regarding this quasi-conditional methodology. The main one is perhaps: *what e-collection should be used in practice?* If one takes as one's goal "collecting as much evidence as possible in subsequent studies against a single $\theta \in \Theta$", and one is able to formulate a prior on Θ representing a 'best guess' (note that Θ does not necessarily index a set of distributions; rather a set of properties) then there is a good case to be made for the GRO (growth-rate optimal) e-variable as defined by Grünwald et al. (2019), which is always well-defined; we refer to that paper for extensive discussion. For simple nulls, the GRO e-variable is always of pure Savage-Dickey form. However, if we are interested not in testing but in estimation and want an e-posterior

over the full range of $\theta \in \Theta$, then the e-posterior based on the GRO e-variables may not be intuitively optimal, as we saw in this paper: the pure Savage-Dickey e-variable based on a normal prior W in the Gaussian location family (which is GRO relative to W) gives worse risk assessments than the one based on the two-point prior as in Example 3. We do not have a general recommendation (in the style of 'reference priors' for Bayesians) for e-collections, and obviously more research is needed here. This paper is really just a first, exploratory one, comparing various possibilities such as two-point, capped- and dampened e-collections. A second related important question is: in what situations is the bound (16) sufficiently tight to be practically relevant? Essentially, if we are in the setting that (bookie, or policy makers) provide B such that $B \cdot \bar{R}(\delta)$ is constant or varies only very little over y, then the bound (15) looks like a standard expected risk bound. If $\bar{R}(\delta)$ can vary highly depending on the data, then the bounds (15) and (16) are still valid, but the supremum in (16) may make them rather weak. Can we get better bounds in such cases? And what if we let B also vary with θ and a? Then, upon learning $\{B_{\theta,a} : \theta \in \Theta, a \in \mathcal{A}\}$, DM might want to re-consider the originally planned decision rule δ, significantly complicating the analysis. Related to this is the question what loss functions L correspond most tightly to real-world problems at all. In our hypothesis testing examples, we considered asymmetric loss functions which seems reasonable, but in our estimation example, we used the squared error loss — which is mathematically convenient and incentives one to output the posterior mean; but in practical situations one may be more interested in very different, hard-to-formalize, problem-dependent losses. All in all, there is plenty of room for future work here!

References

Michael Scott Balch, Ryan Martin, and Scott Ferson. Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A*, 475(2227):20180565, 2019.

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

S. Bates, M. I. Jordan, M. Sklar, and J. Soloff. Principal-agent hypothesis testing. *arXiv:2205.06812*, 2022.

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

J. Berger, J. Bernardo, and D. Sun. Objective bayesian inference and its relationship to frequentism. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*. Blackwell, 2022.

James O. Berger and Alessandra Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96(453):174–184, 2001.

J.O. Berger and R.L. Wolpert. *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, CA, 2nd edition, 1988.

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.

J.M. Bernardo and A.F.M Smith. *Bayesian Theory*. Wiley, Chichester, 1994.

C Brownie and J Kiefer. The ideas of conditional confidence in the simplest setting. *Communications in Statistics-Theory and Methods*, 6(8):691–751, 1977.

D.A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings National Academy of Sciences*, 58(1):66, 1967.

Sarrat Dass and James Berger. Unified conditional frequentist and bayesian testing of composite hypotheses. *Scandinavian Journal of Statistics*, 2003.

W. Edwards, H. Lindman, and L.J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242, 1963.

I. J. Good. C383. a comment concerning optional stopping. *Journal of Statistical Computation and Simulation*, 39(3):191—192, 1991.

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

P. Grünwald and N. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings of the Thirtieth Conference on Algorithmic Learning Theory (ALT) 2019*, 2019.

P. Grünwald and T. van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

P. Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing, 2019. arXiv preprint arXiv:1906.07801 . Accepted pending minor modifications to Journal of the Royal Statistical Society, Series B.

P. D. Grünwald. Viewing all models as "probabilistic". In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pages 171–182, 1999.

P.D. Grünwald and J.Y. Halpern. Making decisions using sets of probabilities: Updating, time consistency, and calibration. *Journal of Artificial Intelligence Research (JAIR)*, 42:393–426, 2011.

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

Peter Grünwald. Beyond Neyman-Pearson. *arXiv:2205.00901*, 2022.

Allard Hendriksen, Rianne de Heide, and Peter Grünwald. Optional stopping with bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, 16(3):961–989, 2021.

Alexander Henzi and Johanna F. Ziegel. Valid sequential inference on probability forecast performance. *arXiv preprint arXiv:2103.08402*, 2021.

Ralf Herbrich and Robert C. Williamson. Algorithmic luckiness. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

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

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

Jack Kiefer. Admissibility of conditional confidence procedures. *The Annals of Statistics*, pages 836–865, 1976.

Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified Gaussian processes. In *Algorithmic Learning Theory*, pages 963–982. PMLR, 2021.

J. Neyman. *First Course in Probability and Statstics*. Henry Holt and Company, New York, 1950.

Oscar Oelrich, Shutong Ding, Måns Magnusson, Aki Vehtari, and Mattias Villani. When are Bayesian model probabilities overconfident? *arXiv preprint arXiv:2003.04026*, 2020.

Samuel Pawel, Alexander Ly, and Eric-Jan Wagenmakers. Evidential calibration of confidence intervals. *arXiv:2206.12290*, 2022.

Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne de Heide, and Peter Grünwald. E-statistics, group invariance and anytime valid testing. *arXiv:2208.07610*, 2022.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.

Zhimei Ren and Rina Foygel Barber. Derandomized knockoffs: Leveraging e-values for false discovery rate control. *arXiv:2205.15461*, 2022.

J. Robins and L. Wasserman. Robins and Wasserman respond to a Nobel prize winner, 2012. URL https://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/. Blog post on the *Normal Deviate* blog.

Richard Royall. *Statistical evidence: a likelihood paradigm*. Chapman and Hall, 1997.

T. Schweder and N. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, 2016.

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

Glenn Shafer. Testing by betting: a strategy for statistical and scientific communication (with discussion and response). *Journal of the Royal Statistic Society A*, 184(2):407–478, 2021.

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' 98)*, pages 2–9, Nashville, Tennessee, 1997.

Botond Szabó, Aad W Van Der Vaart, and JH van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

J. Ter Schure, M.F. Perez-Ortiz, A. Ly, and P. Grünwald. The safe log rank test: Error control under continuous monitoring with unlimited horizon. *arXiv:1906.07801*, 2021.

Rosanne Turner and Peter Grünwald. Anytime-valid confidence intervals for contingency tables and beyond. *arXiv:2203.09785*, 2022.

Rosanne Turner, Alexander Ly, and Peter Grünwald. Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *arXiv:2106.02693*, 2021.

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).

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 2021.

Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10:299–326, 1939.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2022. (accepted, to appear with discussion).

D. Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks, 1991.

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.

Supplementary Material

In this Supplementary Material, in Section A we provide details and proofs concerning the two-point prior-based generalized Savage-Dickey e-variables for the Gaussian location family and its extension to one-dimensional exponential families. In the final Section C we provide details and proofs concerning the 'pure' Savage-Dickey e-variables for Gaussian location and one-dimensional exponential families. In between in Section B we show how the MLE is e-posterior minimax both for Gaussian location with two-point prior and for general exponential families with pure Savage-Dickey.

A Two-point prior-based E-Collections

A.1 A Theorem for One Dimensional Exponential Families

Here we present Theorem 1, which extends the result (24) for two-point prior-based e-collections of Example 8 to 1-dimensional exponential families. More precisely, we let $\{P_\theta : \theta \in \Theta\}$ be any regular (Barndorff-Nielsen, 1978) 1-dimensional exponential family given in any diffeomorphic parameterization (for concreteness, take the mean-value one) and extended to n outcomes by independence. We write the KL divergence between two members of the family defined on a single outcome as $D(\theta'\|\theta)$. We denote by $\hat{\theta}(y)$ the MLE based on data $y = x^n$, which is known to be unique and equal to the empirical average $n^{-1} \sum_{i=1}^{n} \phi(X_i)$, with ϕ the sufficient statistic, whenever this average lies in Θ, which for regular families is an open set. Suppose we observe $Y = x^n$ with $\hat{\theta}(Y) \in \Theta$. We now use as our loss function

$$L(\theta, \breve{\theta}) \coloneqq 2 \cdot D(\breve{\theta}\|\theta),$$

which is parameterization independent; but with the mean-value parameterization, in the Gaussian location family, $2D(\breve{\theta}\|\theta) = (\theta - \breve{\theta})^2$ becomes the squared error loss as in Example 7; of course the factor 2 is merely for mathematical convenience in the proofs.

As we show below (42), for every $C > 0$, every $\theta \in \Theta$, there exist $\theta^- < \theta < \theta^+$ such that, consistent with the Gaussian case,

$$D(\theta\|\theta^+) = \frac{C}{n^*} \ , \ D(\theta\|\theta^-) = \frac{C}{n^*}. \tag{39}$$

As in the Gaussian example, we take $C = 1$, and we construct, for each n, the e-variables $S_\theta^{[n]}(y) = \frac{(1/2)p_{\theta^-}(y)+(1/2)p_{\theta^+}(y)}{p_\theta(y)}$, obtaining again a generalized Savage-Dickey e-process. Based on the standard Taylor expansion

$$D(\theta^*\|\theta) = \frac{1}{2}I(\theta^\circ)(\theta^* - \theta)^2 \tag{40}$$

where $I(\cdot)$ is the Fisher information, which must hold exactly for some θ° in between θ^* and θ, we may expect risk assessments based on this e-variable to behave similarly as in the Gaussian case. Indeed, below we first prove the result (24) of the main text for the Gaussian case, after which we proceed to prove, heavily using (40), the following extension to regular 1-dimensional exponential families:

Theorem 1 *Let Θ° be an arbitrary compact subinterval of Θ. Consider restricted model $\{P_\theta : \theta \in \Theta^\circ\}$ and the collection of e-processes $\{S_\theta : \theta \in \Theta^\circ\}$. Then, for every $0 < c \leq 1$,*

there is a constant C^* such that whenever $\hat{\theta} \in \Theta$, and $c \le n^*/n \le 1/c$, (24) still holds up to a vanishing $1 \pm O(n^{-1/2})$ factor. That is, there is C^* such that, for all n, uniformly for all x^n with $\hat{\theta}(x^n) \in \Theta^\circ$,

$$\max_{\theta \in \Theta^\circ} L(\theta, \hat{\theta}(x^n)) \cdot \bar{P}(\theta \mid x^n) = \frac{1}{n} \cdot \left(\frac{n^*}{n} \right) \cdot \exp^{n/n^*} \cdot f_n \cdot 0.53 \ldots \overset{\text{if } n^*=n}{=} \frac{1}{n} \cdot f_n \cdot 1.45 \ldots \quad (41)$$

where $1 - \sqrt{C^*/n} \le f_n \le 1 + \sqrt{C^*/n}$.

Thus, even though now $\hat{\theta}$ may not be precisely equal to the e-posterior minimax estimator δ^{MM} any more, we may use it as our decision rule anyway, and then still get, for each compact sub-model Θ°, risk bounds that, at least for sufficiently large n, are of the same order as standard minimax frequentist risk bounds.

A.2 Proofs for Example 8, 10 in main text, and Theorem 1 above

Below we prove all results that were stated in the main text for the Gaussian location family and 1-dimensional exponential families. We will freely, without further reference, use standard terminology and results concerning exponential families (all to be found in (Barndorff-Nielsen, 1978)), but we first do highlight two slightly nonstandard results that we will use repeatedly. The first (easily proved using steepness of regular exponential families) is that for each fixed $\theta' \in \Theta$, $D(\theta' \| \theta)$ is a continuous function of $\theta \in \Theta$ satisfying

$$\sup_{\theta < \theta'} D(\theta' \| \theta) = \sup_{\theta > \theta'} D(\theta' \| \theta) = \infty. \quad (42)$$

The second result we need is the *KL robustness property* (Grünwald, 2007) that holds for all regular exponential families: for any fixed $y = x^n$ such that $\hat{\theta}(y)$ is well-defined, any $\theta \in \Theta$ and any prior W on Θ (W does not need to have a density), we have:

$$\frac{p_\theta(y)}{p_W(y)} = \exp(-nD(\hat{\theta} \| \theta) + D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)})), \quad (43)$$

where P_W is the Bayes mixture distribution as defined in Example 1, and, for any distribution P underlying random process $X_1, X_2, \ldots$, we denote the marginal distribution of the first n outcomes as $P^{(n)}$.

Proof of (24) in Example 8 and Claim in Example 10 Fix $C > 0$. By (42) above, for general regular 1-dimensional exponential families, there exist $\theta^- < \theta < \theta^+$ such that (39) holds.

Now define, for $i = 1, 2, \ldots$, $x^i \in \mathbb{R}^i$, the e-processes $S_\theta^{\mathrm{L}}(x^i) = \frac{p_{\theta^-}(x^i)}{p_\theta(x^i)}$ and $S_\theta^{\mathrm{R}}(x^i) = \frac{p_{\theta^+}(x^i)}{p_\theta(x^i)}$ and $S_\theta \coloneqq (1/2)S_\theta^{\mathrm{L}} + (1/2)S_\theta^{\mathrm{R}}$. While these e-processes provide valid e-variables at each sample size i we have that when evaluated at sample size n (on which their definition depends), S_θ^{L} coincides with the *uniformly most powerful Bayes factor* for a 1-sided test at sample size n and level $\alpha^*/2$, for any n and α^* such that $C^- = -\log(\alpha^*/2)$, for $\mathcal{H}_0 = \{P_\theta\}$ vs. $\mathcal{H}_1 = \{P_{\theta'} : \theta > \theta\}$; similarly for S_θ^{LR} with C^+ and $\mathcal{H}_1 = \{P_{\theta'} : \theta' < \theta\}$ respectively (Johnson, 2013).

We first focus on the Gaussian location family for which $D(\theta' \| \theta) = (1/2)(\theta' - \theta)^2$. $\bar{P}(\theta \mid y)$ can be written by (43) with W a prior putting mass $1/2$ on θ^+ and $1/2$ on θ^-. For the Gaussian

location family, (43) then gives:

$$\bar{P}(\theta \mid y) = \frac{p_\theta(y)}{\frac{1}{2}p_{\theta^-}(y) + \frac{1}{2}p_{\theta^+}(y)} = \frac{2e^{-nD(\hat\theta\|\theta)}}{e^{-nD(\hat\theta\|\theta^-)} + e^{-nD(\hat\theta\|\theta^+)}}$$

$$= \frac{2e^{-(n/2)(\hat\theta-\theta)^2}}{e^{-(n/2)\cdot(\hat\theta-\theta+\sqrt{2C/n})^2} + e^{-(n/2)(\hat\theta-\theta-\sqrt{2C/n})^2}} =$$

$$= \frac{2e^{C\cdot(n/n^*)}}{e^{-\sqrt{n}\cdot(\hat\theta-\theta)\sqrt{2C(n/n^*)}} + e^{\sqrt{n}\cdot(\hat\theta-\theta)\sqrt{2C(n/n^*)}}} \tag{44}$$

We will assume that, for any observed y, the MLE $\hat\theta = \hat\theta(y)$ is chosen as action (we will derive that this is the e-posterior minimax action later on, in Proposition 3). Thus, to calculate the risk assessment bound (5) we need to find the maximum over θ of

$$(\hat\theta - \theta)^2 \bar{P}(\theta \mid y) \tag{45}$$

We find, using (44), and the fact that the maximum (45) must be achieved at the same θ as the maximum of $2Cn(n/n^*)(\hat\theta - \theta)^2 \bar{P}(\theta \mid y)$, that the maximizing θ_n satisfies

$$\sqrt{2Cn \cdot \frac{n}{n^*}(\theta_n - \hat\theta)^2} = x^* \quad \text{where} \quad x^* \text{ maximizes } \frac{x^2}{\exp(x) + \exp(-x)}, \tag{46}$$

a function with a unique maximum, found numerically to be achieved at $x^* = \pm 2.065338138969\ldots$. (45) then becomes,

$$\frac{1}{n} \cdot \frac{(x^*)^2}{2C(n/n^*)} \cdot \frac{2e^{C(n/n^*)}}{e^{-x^*} + e^{x^*}} \tag{47}$$

We want to find the C giving the best possible bounds for the choice $n = n^*$, i.e. when the anticipated sample size turns out correct. This can be found by minimizing (47) for $n = n^*$: by differentiation we find $C = 1$, corresponding to the choice in Example 8. We then find (still at $n = n^*$) that $\theta_n = \hat\theta \pm x^*/\sqrt{2n} \approx \hat\theta \pm 1.46/\sqrt{n}$. Similarly, $\bar{P}(\theta \mid y)$ for the maximizing θ_n at $n = n^*$ is given by $2e^1/(e^{x^*} + e^{-x^*}) = 0.6783206434\ldots$. The desired maximum (45), at general n, is then given, using (47) with $C = 1$, by $(1/n)(n^*/n)e^{n/n^*}x^{*2}/(e^{x^*} + e^{-x^*}) = (1/n)(n^*/n)e^{n/n^*} \cdot 0.53222078$, which, if $n = n^*$, evaluates to $(1/n) \cdot 1.446729604\ldots$ and (24) follows.

Proof of Theorem 1 above and generalized Claim in Example 10 For simplicity we only consider the case with $n = n^*$ here; the extension to $c \le n/n^* \le 1/c$ can then be derived analogously to the Gaussian case. We now let Θ represent the mean-value parameterization. As in the main text, we set θ^+ and θ^- as in (39) above with $C = 1$, which defines $\bar{P}(\theta \mid y)$. We fix an arbitrary compact subinterval Θ° of Θ. We need to bound, for arbitrary fixed $\hat\theta \in \Theta^\circ$ the maximum over $\theta \in \Theta^\circ$ of

$$2 \cdot D(\hat\theta \mid \theta)\bar{P}(\theta \mid y). \tag{48}$$

We let $\beta(\theta)$ denote the canonical parameter corresponding to θ. Note that, for $g(\theta^*) = D(\theta^*\|\theta) - D(\theta^*\|\theta^-)$, we have $g'(\theta^*) = \beta(\theta^-) - \beta(\theta) < 0$ constant in θ^* (here we used that β is monotonically increasing in θ). In particular, if $\hat\theta$ and θ are both contained in Θ°, then $g'(\hat\theta) \le -C^-$ where $C^- = \min_{\theta\in\Theta^\circ} \beta(\theta) - \beta(\theta^-) > 0$, and then, for $\hat\theta = \theta - a$, $a > 0$:

$$\bar{P}(\theta \mid y) \le 2e^{-n(D(\hat\theta\|\theta)-D(\hat\theta\|(\theta)^-))} = 2e^{-n(D(\theta-a\|\theta)-D(\theta-a\|\theta^-)}$$

$$\le 2e^{nD(\theta\|\theta^-)-anC^*} = 2e^{C-anC^-}.$$

28

Analogously, if $\hat{\theta}, \theta \in \Theta^\circ$, $\hat{\theta} = \theta + a$, $a > 0$, we can derive $\bar{P}(\theta \mid y) \leq 2e^{C-anC^+}$ for $C^+ = \min_{\theta \in \Theta^\circ} \beta(\theta^+) - \beta(\theta) > 0$. It follows that with $C^\pm = \min\{C^-, C^+\}$, there exists another constant C' such that, if $\theta, \hat{\theta} \in \Theta^\circ$, then the risk assessment at θ satisfies

$$2 \cdot D(\hat{\theta} \| \theta) \bar{P}(\theta \mid y) \leq bC'a^2 e^{C-|a|nC^\pm}. \tag{49}$$

Writing $|a| = \sqrt{K/n}$, we see that for large enough K, the right-hand side is smaller than $1/n$. We thus find that there is a constant K such that, for all $\theta, \hat{\theta} \in \Theta^\circ$, if $|\hat{\theta} - \theta| \geq \sqrt{K/n}$, then the risk bound (49), and hence (48), is smaller than $1/n$. Without loss of generality we may assume $K \geq 1.46^2$ (the need for this condition will become clear later).

We now fix $K' > 0$ and set, for $\alpha \in [-K', K']$, for $\hat{\theta} \in \Theta^\circ$,

$$\hat{\theta}_\alpha := \hat{\theta} + \alpha \cdot \sqrt{\frac{2}{nI(\hat{\theta})}}. \tag{50}$$

where K' is chosen such that $\hat{\theta}_{-K'} \leq \hat{\theta} - \sqrt{K/n}$ and $\hat{\theta}_{K'} \geq \hat{\theta} + \sqrt{K/n}$ for all $\hat{\theta} \in \Theta^\circ$. Since $I(\theta)$ is bounded from below on Θ° by a positive constant, such a K' must exist (note we do not require that $\theta_\alpha \in \Theta^\circ$ for all $\alpha \in [-K', K]$). We now define $\Theta'_{\hat{\theta}} := \Theta^\circ \cap \{\hat{\theta}_\alpha : \alpha \in [-K', K]\}$. Our proof strategy for the remainder of the proof is showing that $\Theta'_{\hat{\theta}}$ is an interval around $\hat{\theta}$ that is small enough so that we can re-derive the results for the Gaussian case (up to a factor converging to 1) on this set, and then showing that the global maximum of (48) over Θ° is achieved on this set, so that only this set matters.

We have that $|I'(\theta)/I(\theta)|$, where I' is the derivative if I, is bounded by a constant on Θ°. A first-order Taylor approximation of $I(\theta)$ then implies that there exist a constant C° such that for all large enough n larger than some n°, for all $\hat{\theta} \in \Theta^\circ$, we have

$$\text{for all } \theta, \theta^\circ \in \Theta'_{\hat{\theta}} : 1 - \sqrt{\frac{C^\circ}{n}} \leq \frac{I(\theta)}{I(\theta^\circ)} \leq 1 + \sqrt{\frac{C^\circ}{n}}. \tag{51}$$

We consider, at sample size $n > n^\circ$ the risk assessment (5), i.e. the maximum over $\theta \in \Theta^\circ$ of (48) that we get for any fixed $\hat{\theta}(y) \in \Theta^\circ$. We first consider the maximum over $\theta \in \Theta'_{\hat{\theta}}$.

Within $\Theta'_{\hat{\theta}}$, we can perform essentially the same derivation as in (44), using the Taylor approximation (40) in the second step in both numerator and denominator and aggressively and repeatedly using (51) above to vary θ in $I(\theta)$. This (tediously but straightforwardly) gives, for $\alpha \in [-K', K']$ such that $\hat{\theta}_\alpha \in \Theta'$,

$$\bar{P}(\hat{\theta}_\alpha \mid y) = 2e^1 e^{O(n^{-1/2})} \cdot \frac{1}{e^{-2\alpha} + e^{2\alpha}} = 2e^1 e^{O(n^{-1/2})} \cdot \frac{1}{e^{-\sqrt{(2I(\hat{\theta})n}(\theta - \hat{\theta})} + e^{+\sqrt{(2I(\hat{\theta})n}(\theta - \hat{\theta})}}$$

$$D(\hat{\theta} \| \hat{\theta}_\alpha) = (1 + O(n^{-1/2}))(\hat{\theta} - \hat{\theta}_\alpha)^2 I(\hat{\theta})$$

which holds uniformly for all $\hat{\theta} \in \Theta^\circ$ and all $\alpha \in [-K', K']$ such that $\hat{\theta}_\alpha \in \Theta'_{\hat{\theta}}$, i.e. we can pick the same constant in $O(n^{-1/2})$ for all such $\hat{\theta}$ and α. But this means that for all large enough n, on the set Θ', we can repeat the same reasoning as in (46) in the Gaussian case up to a factor of order $1 + O(n^{-1/2})$: since this factor converges to 0, we must have that, uniformly for all $\hat{\theta} \in \Theta$, for all n larger than some n_0, the maximum will be above $(1/n) \cdot 1.44$ and will be achieved by one or both elements inside a set $\Theta_{\max} = \{\hat{\theta} + \gamma_n^+/\sqrt{n}, \hat{\theta} - \gamma_n^-/\sqrt{n}\} \cap \Theta'$, where both γ_n^+ and γ_n^- converge to $1.46\ldots$, at which point $\bar{P}(\theta_n \mid y) \approx 0.68$. Note that by requiring

$K \geq 1.46^2$ we made sure K' is large enough so that both choices for the maximum are in $\{\hat{\theta}_\alpha : \alpha \in [-K', K]\}$; at least one of these two choices for θ_n then also lies in Θ' for n larger than some n_0, where n_0 can be chosen uniformly, independently of $\hat{\theta}$.

Thus, the risk (48) maximized over $\theta \in \Theta'_{\hat{\theta}} \subset \Theta^\circ$ is bounded from below by $1.44/n$, and the maximum over Θ° must be at least as large. But (49) now gives that the maximum of (48) over $\theta \in \Theta^\circ \setminus \Theta'_{\hat{\theta}}$ is at most $1/n$ which is smaller than the maximum risk within $\Theta'_{\hat{\theta}}$. Therefore the global maximum is achieved within $\Theta'_{\hat{\theta}}$ and, just as for the Gaussian location case, is given by $(1/n) \cdot 1.45\ldots$, which is what we had to prove (the extension to $n^* \neq n$ mentioned in (41) is now trivial to add; the various 'for all large enough n' qualifications we imposed are absorbed into C^*). As a by-product, we also find that, since $\bar{P}(\theta_n \mid y) \ll 1$, Condition Zero holds automatically in this setting.

B Establishing that the MLE is e-posterior minimax

The proposition below shows, by setting $\theta' := \hat{\theta}$, that the MLE $\hat{\theta}$ is e-posterior minimax both for the Gaussian location family with the two-point prior (Example 8) and for general exponential families with a pure Savage-Dickey e-variable with arbitrary prior (Example 11).

Proposition 3 *Consider a regular exponential family given in its mean-value parameter space as above. Let $\theta' \in \Theta$ and let $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ be a continuous function such that $\lim_{x \to \infty} x f(x) = 0$. We have*

$$\min_a \max_{\theta \in \Theta} D(a \| \theta) \cdot f(D(\theta' \| \theta))$$

is achieved by $a = \theta'$. In particular,

1. *Suppose that $\hat{\theta} = \hat{\theta}(y) \in \Theta$ and consider the dampened e-posterior $\bar{P}^{[\gamma]}(\theta \mid y)$ for any $0 < \gamma \leq 1$ and relative to any prior W. It can be written as $f(D(\hat{\theta} \| \theta))$ for a function f of the required type.*

2. *For the normal location family, the e-posterior based on the two-point discrete prior as in (44) can also be written as $f(D(\hat{\theta} \| \theta))$ for a function f of the required type.*

Proof: Let $g(a) := \max_{\theta \in \Theta} D(a \| \theta) \cdot f(D(\theta' \| \theta))$. First consider $a = \theta'$. It follows from (42) that the maximum over θ in the definition of $g(a) = g(\theta')$ is achieved by some $\theta^*_- < \theta'$ and also by some $\theta^*_+ > \theta'$. Now consider $a \neq \theta'$. We must show that $g(a) > g(\theta')$. If $a > \theta'$, we have that $D(a \| \theta^*_-) > D(\theta' \| \theta^*_-)$ so

$$g(a) \geq D(a \| \theta^*_-) \cdot f(D(\theta' \| \theta^*_-)) > D(\theta' \| \theta^*_-) \cdot f(D(\theta' \| \theta^*_-)) = g(\theta').$$

The case $a < \theta'$ goes similarly, with θ^*_+ replacing θ^*_-. This establishes the first result.

As to (1), the case with $\gamma = 1$ now follows directly from (43). For $\gamma < 1$, use the fact that $1/((1 - \gamma)x^{-1} + \gamma)$ is increasing in x.

As to (2): using (44), and once more that $D(\hat{\theta} \| \theta) = (1/2)(\hat{\theta} - \theta)^2$, and considering separately the cases that $\hat{\theta} > \theta$ and $\hat{\theta} < \theta$, we find that $\bar{P}(\theta \mid y) = f(D(\hat{\theta} \| \theta))$, where f is of the required form. $\square$

C Pure Savage-Dickey based E-Collections for Gaussian and Exponential Families

Let $\bar{P}^{[\gamma]}(\theta \mid y)$ be the 'dampened e-posterior' as in the main text, based on arbitrary prior W on Θ. In Proposition 3 above we showed that for any $0 < \gamma \leq 1$ the MLE $\hat{\theta}$ is the $\bar{P}^{[\gamma]}$-e-posterior minimax estimator irrespective of b. Below we further, via Proposition 4 below, show that, for the choice $\gamma = 1/2$, the risk assessment bound (5) holds with

$$\bar{R}(y, \hat{\theta}(y)) = \frac{2}{n} D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)}) \tag{52}$$

for all n such that the expression on the right is larger than 1 — which will be the case for all but the smallest n. Here we used the notation $D(P_\theta^{(n)} \| P_W^{(n)})$ for the KL divergence between distribution P_θ and Bayes marginal P_W, both defined on n outcomes. For the normal location family with prior with mean 0 and precision λ, we have the exact expression

$$D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)}) = \frac{1}{2} \log \frac{n+\lambda}{\lambda} + \frac{n\lambda}{n+\lambda} \hat{\theta}^2 \tag{53}$$

which together with (52) gives (36) in the main text. (53) is found by using the fact that $D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)}) = \ln \bar{P}_W(\hat{\theta} \mid y)$, an identity which follows from (43) and holds for general regular exponential families. If these have continuous prior w, we get, for $\hat{\theta}$ in any compact subset of the parameter space, the expression (Chapter 8 of Grünwald (2007))

$$D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)}) = \frac{1}{2} \log \frac{n}{2\pi} - \log \frac{w(\hat{\theta})}{I(\hat{\theta})^{1/2}} + o(1),$$

with $I(\hat{\theta})$ the Fisher information at $\hat{\theta}$, allowing us to generalize (36) to general exponential families.

We proceed to derive (52). We start with a proposition that is applicable more generally than just for KL loss:

Proposition 4 *Let $\Theta \subset \mathbb{R}$, $b \in \mathbb{R}^+$ and let $L : \Theta \times \Theta \to \mathbb{R}_0^+$ be a loss function and let $\breve{\theta} : \mathcal{Y} \to \Theta$ be an estimator. Fix some $y \in \mathcal{Y}$ and let $\breve{\theta} := \breve{\theta}(y)$. Consider an e-posterior $\bar{P}(\theta \mid y)$ with $\bar{P}_{\sup} := \sup_{y \in \mathcal{Y}, \theta \in \Theta} \bar{P}(\theta \mid y)$. and let $\bar{P}'$ be an upper bound on $\bar{P}$ up to a factor $C_{\sup}$ i.e. for all $\theta \in \Theta$, $y \in \mathcal{Y}$, $\bar{P}(\theta \mid y) \leq C_{\sup} \bar{P}'(\theta \mid y)$. Fix some $\theta_L, \theta_R \in \Theta$ (depending on $\breve{\theta}$) with $\theta_L < \theta_R$ so that both:*

  1. for all $\theta \geq \theta_R$, $\bar{P}'(\theta \mid y) \leq 1$ and $\bar{P}'(\theta \mid y) L(\theta, \breve{\theta})$ is decreasing in θ.

  2. for all $\theta \leq \theta_L$, $\bar{P}'(\theta \mid y) \leq 1$ and $\bar{P}'(\theta \mid y) L(\theta, \breve{\theta})$ is increasing in θ.

Then our main risk assessment bound (5) holds for $\bar{P}(\theta \mid y)$ with

$$\bar{R}(y, \breve{\theta}(y)) = b(y) \cdot \max\{C_{\sup}, \bar{P}_{\sup}\} \cdot \max_{\theta \in [\theta_L, \theta_R]} L(\theta, \breve{\theta}). \tag{54}$$

Proof: For $\theta \leq \theta_L$, $\bar{P}(\theta \mid y) L(\theta, \breve{\theta}) \leq \bar{P}'(\theta \mid y) C_{\sup} L(\theta, \breve{\theta}) \leq C_{\sup} \bar{P}'(\theta_L \mid y) L(\theta_L, \breve{\theta}) \leq C_{\sup} L(\theta_L, \breve{\theta})$. Analogously for $\theta \geq \theta_R$, we have $\bar{P}(\theta \mid y) L(\theta, \breve{\theta}) \leq C_{\sup} L(\theta_R, \breve{\theta})$. Finally for $\theta \in [\theta_L, \theta_R]$, we have $\bar{P}(\theta \mid y) L(\theta, \breve{\theta}) \leq \bar{P}_{\sup} \cdot \max_{\theta \in [\theta_L, \theta_R]} \cdot L(\theta, \breve{\theta})$. The result follows. $\square$

We now use Proposition 4 to show the bound (52). Assume the setting of that bound. We set $\breve{\theta}$ to the MLE and $L(\theta, \breve{\theta}) \coloneqq D(\breve{\theta}\|\theta)$. We first apply Proposition 4 with $\bar{P}'(\theta|y) \coloneqq \bar{P}(\theta|y)$ (and $C_{\sup} = 1$) set to the dampened e-posterior $\bar{P}^{[1/2]}$ relative to a prior W (independent of θ). Because of the dampening with $\gamma = 1/2$, we know that $\bar{P}_{\sup} \leq 2$. We will apply the proposition with $\theta_L < \theta_R$ such that $\bar{P}^{[1/2]}(\theta_L \mid y) = \bar{P}^{[1/2]}(\theta_R \mid y) = 1$. These must exist (use (42)), and by (43) they satisfy

$$D(\hat{\theta}\|\theta_L) = D(\hat{\theta}\|\theta_R) = \frac{D(P^{(n)}_{\hat{\theta}}\|P^{(n)}_W)}{n}.$$

To verify the conditions of Proposition 4, we will show, using (43), that

$$D(\hat{\theta}\|\theta)\exp(-nD(\hat{\theta}\|\theta) + D(P^{(n)}_{\hat{\theta}}\|P^{(n)}_W)) \tag{55}$$

is increasing for $\theta < \theta_L$ and decreasing for $\theta > \theta_R$. For this, setting $C = D(P^{(n)}_{\hat{\theta}}\|P^{(n)}_W)$, it is sufficient to show that $g(u) \coloneqq u\exp(-nu + C)$ is decreasing if $u \geq D(\hat{\theta}_L\|\theta_L)$, i.e. if $u \geq C/n$. Differentiation gives that $g(u)$ is decreasing if $u > 1/n$, so Proposition 4 can be applied if $C \geq 1$ and then (54) gives (52).