Beyond Neyman-Pearson

Peter D. Grünwald*

May 3, 2022

Abstract

A standard practice in statistical hypothesis testing is to mention the p-value alongside the accept/reject decision. We show the advantages of mentioning an e-value instead. With p-values, we cannot use an extreme observation (e.g. $\text{P} \ll \alpha$) for getting better frequentist decisions. With e-values we can, since they provide Type-I risk control in a generalized Neyman-Pearson setting with the decision task (a general loss function) determined post-hoc, after observation of the data — thereby providing a handle on 'roving α's'. When Type-II risks are taken into consideration, the only admissible decision rules in the post-hoc setting turn out to be e-value-based. We also propose to replace confidence intervals and distributions by the *e-posterior*, which provides valid post-hoc frequentist uncertainty assessments irrespective of prior correctness: if the prior is chosen badly, e-intervals get wide rather than wrong, suggesting the *e-posterior minimax* decision rule as a safer alternative for Bayes decisions. The resulting *quasi-conditional paradigm* addresses foundational and practical issues in statistical inference.

Dedicated to the memory of Sir David R. Cox (1924-2022).

1 Introduction

We perform a null hypothesis test with significance level α and we observe a p-value $\text{P} \ll \alpha$. Why aren't we allowed to say "we have rejected the null at level P"? While a continuous source of bewilderment to the applied scientist, professional statisticians understand the reason: to get a Type-I error probability guarantee of α — a cornerstone of the Neyman-Pearson (NP) theory of testing — we must set α in advance. But this immediately raises another question: why should the p-value be mentioned at all in scientific papers, next to the reject/accept decision for the pre-specified α (Berger, 2003, Hubbard, 2004)? The prevailing attitude is to accept this standard practice, on the grounds that it "provides more information" — as explicitly stated by, for example, Lehmann (1993), one of NP theory's main contributors. But this is problematic: there is nothing in NP theory to tell us what the decision-theoretic consequences of '$\text{P} \ll \alpha$' could be, whereas at the same time, the fundamental motivation behind NP theory *is* decision-theoretic: according to Neyman (1950), "[all of] *mathematical statistics deals with problems relating to performance characteristics of rules of inductive behavior* [i.e. decision rules] *based on random experiments*". There is no simple way though to translate observation of a P with $\text{P} \ll \alpha$ into better decisions: as is well-known and reviewed below (Equations (5) and (14)), intuitive and common decision-theoretic interpretations of

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

P $\ll \alpha$ are usually just wrong. We are therefore faced with a standard practice in NP testing that, according to NP theory, is not part of mathematical statistics! In fact, even just stating that P measures 'evidence against the null', without attaching direct decision-theoretic consequences to it, is highly problematic — as has been forcefully argued by many, p-values have properties that are at odds with any reasonable definition of 'evidence'; see e.g. (Royall, 1997) and the many references therein.

E as the New P We argue that this issue can be resolved, once and for all, *by mentioning e-values rather than* P*-values* next to the accept/reject decision. E-values (Vovk and Wang, 2021, Shafer, 2021, Grünwald et al., 2019, Ramdas et al., 2021) are a recently popularized alternative for P-values that are related to, but far more general than, likelihood ratios. Importantly, as reviewed below, for any NP test with the accept/reject-decision based on a P-value, the exact same test can be implemented by basing the decision on an e-value. Thus there is no a priori reason why one should accompany the decision of a NP test with a p-value rather than an e-value. But, in contrast to the p-value, the e-value has a clear decision-theoretic justification that remains valid if decision tasks are formulated *post-hoc*, i.e. after seeing, and in light of, the data. Concretely, after the result of a study has been published, and when new circumstances prevail, one conceivably might contemplate different actions, with different associated losses, than originally planned. For example, a study about vaccine efficacy (VE) in a pandemic may have 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 a very small p-value $\ll \alpha$, and at the same time the virus' reproduction rate may be much higher than anticipated. Based on *both* the observed data (summarized by P) *and* the changed circumstances, one might now contemplate a new action, vaccinate everyone over 40, with higher losses if the alternative is false and higher pay-offs if it is true. E-values can be used unproblematically for such a post-hoc formulated decision task; p-values cannot. A second example is simply the fact that scientific results are *published* and remain on record so as to be useful for future deployment. A company contemplating to produce medication X may find a publication about the efficacy of X that is a few years old, but was never acted upon. Consider the situation that the fact that the null (no efficacy) was rejected at the given α would not nearly be enough evidence to justify further investment, but in fact the observed p-value (or inverse e-value) was $\ll \alpha$. How can this information be transferred into taking a rational decision about further investment? E-variables provide a handle on such problems that p-values do not.

From Testing to Estimation with Confidence: the e-posterior The P $\ll \alpha$ question has a counterpart in estimation with confidence intervals. Upon observing data from a parametric statistical model $\{P_\theta : \theta \in \Theta\}$, the question now becomes how to properly interpret the statement "$\theta \in \text{CS}_\alpha$", where CS_α is a $(1-\alpha)$-confidence set, usually an interval. The correct, basic interpretation only says that, when repeatedly performing studies, the true parameter will lie in CS_α in a fraction of about $1-\alpha$ studies. But practitioners want more, and indeed, CS's are often given an evidential interpretation — one outputs not one but a system of confidence intervals, one for each of a series of coefficients such as $80\%, 90\%, 95\%, 99\%$, and this, it is said "*summarizes what the data tell us about θ, given the model*" (Cox and Hinkley, 1974, page 227) or "the information about the parameter" (Lehmann, 1959). Such

an evidential interpretation is highly problematic though. Illustrations abound (Royall, 1997) and include the famous setting of Cox (1958) in which optimal (minimal width) confidence intervals may depend on an independent coin flip that is totally external to the experiment being performed. Interpretational problems concerning 'evidence' are sometimes dismissed as vague, but as we show in Example 2 and 3, they translate into serious *practical* problems once we deal with post-hoc determined decision tasks as above.

Our second main claim is that these issues can be resolved by replacing standard CS's by special CS's based, once again, on e-values — the recently popularized *anytime-valid* CSs (Darling and Robbins, 1967, Howard et al., 2021) being a special case. To see how, first note that standard CS systems as above can be conveniently represented by a single data-dependent *confidence distribution (CD)* on Θ (Schweder and Hjort, 2016); for some models this coincides with the Bayesian posterior in an *objective-Bayes* analysis (Berger, 2006). Similarly systems of e-value based CIs can be represented by a single data-dependent function on Θ that we will call an *e-posterior* (without the word 'distribution' attached, since technically it isn't). In Example 2 and 3 we show that using the CD to guide decisions against standard loss functions can have bad consequences if the loss function is chosen in a post-hoc, data-dependent way: the loss one expects to make, according to the confidence distribution, may be much smaller than the *actual* expected loss, which may even be infinite. In contrast, the loss one expects to make according to the e-posterior with the associated decision rule gives a correct upper-bound-in-expectation on the actual expected loss — no matter what the true parameter is.

The BIND Assumption underlying p-values and standard confidence intervals While so far we highlighted the problems with post-hoc determined loss functions, in the next section we show that decisions based on P's and CS's in an *intuitive manner* may already become unsafe as soon as the decision task involves a 'Type-I' loss function that can take on more than two values, even if this loss function *is* determined in advance. Essentially, we can only be sure that decisions based on P's and CS's are reliable if both (1) the loss function is binary-valued (B) and (2), it is determined in advance, or at least independently (IND) of the observed data. Thus, they really operate under a BIND (binary + independence) assumption. E-variables and -posteriors lead to decisions that remain safe if BIND is violated.

Note the phrase *used in an intuitive manner* though. By this we mean that CD's (or p-values) are used to make decisions *as if they were Bayesian posterior distributions (or probabilities)*, commensurate with the confidence distribution's close similarity to 'objective Bayes' posteriors (see (5), (7), (10). Perhaps we can translate p-values and CI's into decisions in a different way, that remains valid without the BIND assumption? It turns out that we can, but the only general way for doing so that we know of is to convert any given p-value to an e-value, and any confidence posterior to an e-posterior via *calibrators* (for example, (Shafer et al., 2011) show that $(1/\sqrt{\mathrm{P}} - 1$ is a calibrator). But then we could also have used E-methods directly, and this would often have resulted in tighter confidence intervals (e-posteriors based on problem-specific e-variables lead to CIs that are between 1.4–2 times as wide as standard CIs (Example 6); converting via calibrators can make them significantly wider (Grünwald et al., 2019, Section 7)).

1.1 History, Background and Contents of this Paper

Suppose we observe data Y taking values in some set $\mathcal{Y}$, the null hypothesis $\mathcal{H}_0$ being represented as a collection of distributions for Y. An e-value is the value of a special type of statistic called an e-variable. An e-variable is any *nonnegative* random variable $S = S(Y)$ that can be written as a function of the observed Y and that satisfies the inequality:

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

The e-variable's simplest application is in defining tests: the S-based hypothesis test at level α is defined to reject the null iff $S \geq 1/\alpha$. Since for any e-variable S, all $P \in \mathcal{H}_0$, by Markov's inequality, $P(S \geq 1/\alpha) \leq \alpha)$, with such a test we get a Type-I error guarantee of α.

E-variables first implicitly appear in the testing literature as a building block of nonnegative martingales in the work by H. Robbins and his students from the late 1960s (Darling and Robbins, 1967, Robbins, 1970), but there they were not studied as separate entities. As such, they have probably been originally introduced by Levin (of P vs NP fame) (1976), were independently re-discovered by Zhang et al. (2011) and were first analyzed by Shafer et al. (2011). Still, the concept mostly lay dormant until 2019, when interest in them suddenly exploded (Grünwald et al., 2019, Shafer, 2021, Henzi and Ziegel, 2021, Ramdas et al., 2021). In most of these papers though, they are treated in a sequential context; ours (with Vovk and Wang (2021), Wang and Ramdas (2020)) is one of the first to consider them nonsequentially. In the sequential setting, they distinguish themselves from p-values by allowing to preserve Type-I error guarantees under *optional continuation* — performing additional studies if previous studies had certain outcomes, and then combining the results; when e-values are extended to e-processes (Section 4) they can also deal with *optional stopping* (the difference between OS and OC is explained by Grünwald et al. (2019)). The idea of distinguishing between decision tasks presented for and after an observation was anticipated (in a completely different, nonstatistical context) by (Grünwald and Halpern, 2011, Section 3)

A different kind of Robustness Standard P and CS-based decision rely the BIND assumption; an assumption that will often be false or unverifiable at the time study results are published. E-values provide valid error and risk guarantees without making such assumptions, and are therefore *robust* tools for inference. But whereas 'robustness' usually refers to robust inference in the presence of outliers, or model structure or noise process misspecification, this is a different, much less studied form of robustness: robustness in terms of the actual decision task that the study results will be used to solve.

From Wald to Generalized Neyman-Pearson (GNP) Technically, to obtain frequentist guarantees without the BIND assumption we need to shift from errors and error probabilities to losses and risks. This idea goes back to (Wald, 1939), one of the most influential papers in the history of statistics: like Wald, we first re-formulate standard NP testing in terms of risks. But while Wald lets go off the Type-I/II error paradigm as soon as he allows for more than two actions, we stick with it and show that the e-value is then the natural statistic to base decisions upon, and remains so if the decision task is determined post-hoc. Thus, our *GNP* (*Generalized Neyman-Pearson*) Theory follows a path opened up by Wald but apparently not pursued further thereafter.

The Quasi-Conditional Paradigm & Related Work GNP and e-posterior based decision rules allow loss functions to be chosen in a fully 'conditional' manner (they can depend on the observed data in arbitrary ways), but their performance is evaluated unconditionally in terms of the sampling distribution. This *quasi-conditional stance* provides a middle ground between fully Bayesian and traditional Neyman-Pearson-Wald type methods and analysis. It involves priors, but inferences are (unconditionally) valid irrespective of their correctness — the priors encode 'hope' rather than 'belief' (Section 4). It is related to, but quite different from, *conditional frequentist* approaches (Kiefer, 1977, Berger et al., 1994); we elaborate on the difference in Section 2.3. There we also discuss relations to *inferential models* (Martin and Liu, 2015, Balch et al., 2019, Martin, 2021).

Contents Section 2.1 below introduces GNP testing with post-hoc loss functions. Section 2.2 illustrates how confidence intervals and distributions have difficulties with post-hoc decision functions. We then remark on how to properly interpret our findings in Section 2.3. After these high-level sections we give a detailed mathematical treatment of our results. First, Section 3 shows how all admissible decision rules in the *GNP* framework can be based on e-values. Section 4 treats the e-posterior and the new *e-posterior minimax* decision rule. We end by tying up loose ends — e.g. explaining how the theory can be extended to models with nuisance parameters — in the concluding Section 5. All longer mathematical derivations and proofs are delegated to the appendices.

2 High-Level Overview

2.1 Losses instead of Errors: the GNP setting

NP tells us to fix some α and then adopt the decision rule that, among all decision rules with Type-I error bounded by α, minimizes the Type-II error. In his seminal (1939) paper, Abraham Wald already suggested to re-interpret this procedure in terms of a nonnegative loss function $L(\cdot, \cdot)$, with $L(\theta, a)$ denoting the loss made by action a if θ is the true state of nature. In the basic NP setting, we have $\theta \in \{0, 1\}$ and $\mathcal{A} = \{0, 1\}$, $L(0, 1) > 0, L(1, 0) > 0$ and we may 'of course' (as Wald writes) set $L(0, 0) = L(1, 1) = 0$. In this formulation, the usual α-Type-I error guarantee is replaced by an *ℓ-Type-I risk guarantee*. Formally, we fix an ℓ in advance of observing the data and we say that decision rule δ, defined as a function from $\mathcal{Y}$ to $\mathcal{A}$, is *Type-I risk safe* if

$$\textsc{risk}(0, \delta) \leq \ell, \text{ where } \textsc{risk}(0, \delta) := \sup_{P_0 \in \mathcal{H}_0} \mathbf{E}_{Y \sim P_0}[L(0, \delta(Y))]. \tag{2}$$

Following NP again, with again 'error probability' replaced by 'risk', we now postulate that among all Type-I risk-safe decision rules, we ideally want to pick one that minimizes the *Type-II risk*, given by

$$\textsc{risk}(1, \delta) := \sup_{P_1 \in \mathcal{H}_1} \mathbf{E}_{Y \sim P_1}[L(1, \delta(Y))]. \tag{3}$$

(2) expresses that, whatever we do, we want to make sure that our risk (expected loss) under the null is no larger than ℓ. This kind of procedure may have most appeal if $L(1, 0) = \ell$: ℓ then represents the loss that we can trivially achieve, simply by accepting the null — usually this means taking no real action at all and perpetuating the status quo. By requiring the

Type-I risk guarantee (2), we then impose that the risk of our decision rule is not larger than our worst-case loss ℓ that we get if we perpetuate the status quo. Still, all our results below continue to hold if $L(1,0) \neq \ell$.

In a standard level-α-significance test, one rejects the null if $\text{P}(y)$, the p-value corresponding to data y, satisfies $\text{P}(y) \leq \alpha$. A corresponding decision rule in terms of loss functions is to reject the null whenever the observed $\text{P}(y)$ satisfies

$$\text{P}(y) \cdot L(0,1) \leq \ell. \tag{4}$$

We get exactly the same behaviour as for the standard α-test if we set $L(1,0) = \ell/\alpha$. For example, for $\alpha = 0.05$ we can set $\ell = 1$ and then $L(0,1) := 20$; then (4) tells us to reject the null if $p \leq 0.05$. The resulting decision rule will be Type-II error optimal among all decision rules that satisfy Type-I error probability ≤ 0.05 if and only if it is Type-II risk optimal among all decision rules that satisfy the Type-I risk bound (2): up till now it seems as we have merely reformulated standard NP theory.

Actions of Varying Intensity But now suppose we have *more than two* actions available. For example, consider four alternative actions: accept the null (retain the status quo), take mild action (e.g. vaccinate all people over 60), take more drastic action (vaccinate everyone over 40) and extreme action (vaccinate the whole population). Our first contribution is to consider this question, too, in terms of Type-I and Type-II risk and confidence — thereby taking a different direction than standard decision theory and in particular Wald (1939), who switches to non-Neyman-Pearsonian decision rules as soon as more than two actions are in play. For example, our action space could now be $\mathcal{A}_b = \{0,1,2,3\}$ with loss function $L_b(0,0) = 0, L_b(0,1) = 20\ell$, $L_b(0,2) = 100\ell$, $L_b(0,3) = 500\ell$ and $L_b(1,3) < L_b(1,2) < L_b(1,1) < L_b(1,0) = \ell$. In terms of p-values, the straightforward extension of (4) to this multi-action case would be to play action a where a is the largest value such that

$$\text{P}(y) \cdot L_b(0,a) \leq \ell. \tag{5}$$

But, assuming our p-value is strict so that it has a uniform distribution under the null, this gives a Type-I risk of

$$\mathbf{E}_{Y \sim P_0}[L_b(0, \delta(y))] = \left(\frac{1}{20} - \frac{1}{100}\right) \cdot 20\ell + \left(\frac{1}{100} - \frac{1}{500}\right) \cdot 100\ell + \frac{1}{500} \cdot 500\ell = 2.6\ell, \tag{6}$$

violating the guarantee we aimed to impose and showing that a naive p-value based procedure does not work. The problem gets exacerbated if we allow for more than four actions: in Appendix A.1) we show that the expected loss of the naive procedure (5) may go to ∞ as we add additional actions with $L_b(0,a)$ increasing and $L_b(1,a)$ decreasing in a. We also show that an obvious 'fix', namely modifying (5) to make sure that for each action a, $L_b(0,a)$ gets multiplied by exactly the probability that action a is taken, does not solve this issue.

Post-Hoc Loss Functions Allowing more than two actions is really just a warm-up to a further extension which arguably better models what often happens in, for example, medical practice: the post-hoc determination or modification of a decision task, after seeing the data and dependently on the data, such as described in the vaccine efficacy example in the introduction. That is, there is really un underlying class (whose definition may be unknowable)

of loss functions $L_b(\cdot,\cdot)$ with associated action spaces $\mathcal{A}_b$, and the decision-maker is posed a particular decision task $L_b(\cdot,\cdot)$ where b, indexing the loss actually used, is really the outcome of a random variable $B = b$, whose distribution may depend on the data in all kinds of ways (we give a precise formalization in Section 3.1). The actual $B = b$ that is presented is thus random and only fixed *after* the study result has become available; i.e. 'post-hoc'. Crucially, the process determining the actual value of B is typically murky; nobody knows exactly what loss function would have been considered in what alternative circumstances; one only knows the loss function finally arrived at.

Again, with p-values, we might be tempted to pick the largest action a such that (5) holds, where now b is really the (observed, known) outcome of random variable B whose definition is itself unknown. Now, even if for each b, L_b allows for only two actions, so that the problem superficially resembles the standard NP setting, using (5) can have disastrous consequences in the post-hoc setting, as the following example shows.

Example 1 Suppose there are three loss functions L_b, for $b \in \mathcal{B} = \{1, 2, 3\}$, with corresponding actions $\mathcal{A}_b = \{0, b\}$. We set $L_1(0,1) = 20\ell, L_2(0,2) = 100\ell, L_3(0,3) = 500\ell$, $L_b(0,0) = 0, L_b(1,0) := \ell$ for all $b \in \mathcal{B}$, and $L_b(1,b)$ decreasing in b. This is like the previous example, but rather than always being able to choose one among four actions, the very set of choices that is presented to the decision maker via setting $B = b$ might depend on the data Y or on external situations. One cannot rule out that this is done in an unfavourable manner — if the data suggest strong evidence then the policy developers (e.g. a pandemic outbreak management team) might only suggest actions with drastic consequences. Suppose, for example, that if $p > 0.02$, the decision-makers are presented loss L_1; if $0.001 < p \leq 0.02$ they are presented loss L_2; and if $p \leq 0.001$ they are presented loss L_3. Using (5), we then get (assuming again uniform P) a Type-I risk of

$$\mathbf{E}_{Y\sim P_0}[L(0, \delta(y))] = (0.05 - 0.02) \cdot 20\ell + (0.02 - 0.001) \cdot 100\ell + 0.001 \cdot 500\ell = 3 \cdot \ell.$$

As in (6) the resulting decision rule (5) is not Type-I risk safe, and again, the Type-I risk can even go to infinity with the number of potential actions.

E-Variables Reporting evidence as e-values (as defined by (1)) rather than p-values solves both the multiple action and post-hoc-loss issue identified above. In such a *Generalized Neyman-Pearson (GNP)* setting (precise definition in Section 3), we can simply pick any e-variable S we like and replace the decision rule (5) by: upon observing data $Y = y$ and loss function indexed by $B = b$,

$$\text{select the largest } a \text{ for which } S^{-1}(y) \cdot L_b(0,a) \leq \ell, \quad \text{i.e.} \quad L_b(0,a) \leq S(y) \cdot \ell, \tag{7}$$

where here and in the sequel we write (with minimal abuse of notation) $S(y)$ for the value that S takes upon observation $Y = y$, and we adopt the (in our setting harmless) convention that, for $u = 0$ and $v \geq 0$, $u^{-1}v := 0$ if $v = 0$ and $u^{-1}v = \infty$ if $v > 0$. For the original NP setting of two actions, this is simply the p-value based rule (5) with the p-value replaced by $1/S$, illustrating that *large* e-values correspond to evidence against the null. But in contrast to the p-value based rule, this one keeps being Type-I risk safe irrespective of the number of actions: as we show in Lemma 1 below, in contrast to p-values: no matter what e-variable S we take, no matter how many actions $\mathcal{A}$ contains, no matter the process determining the loss B, we have the Type-I risk guarantee (2).

As with p-values, many different e-variables can be defined for the same $\mathcal{H}_0$. An extreme choice is to start with a p-value P and to set $S^{\text{NP}(\alpha)} := (1/\alpha)$ if $\text{P} \leq \alpha$ and $S^{\text{NP}(\alpha)} = 0$ otherwise (Shafer, 2021). Clearly $\mathbf{E}_{Y \sim P_0}[S^{\text{NP}(\alpha)}] \leq \alpha(1/\alpha) = 1$ so $S^{\text{NP}(\alpha)}$ is an e-variable. In the case of a classical, 2-action NP problem as defined underneath (4), the test (7) based on e-variable $S = S^{\text{NP}(\alpha)}$ will lead to $a = 1$ (reject the null) exactly iff the classical NP test based on P does. This shows that any P-based NP test can also be arrived at using (7) with a special e-value: nothing is lost by replacing p-values with e-values. Still, in case there are more than 2 actions and/or post-hoc decisions, while preserving the ℓ-Type-I risk guarantee, decisions based on $S^{\text{NP}(\alpha)}$ may not be a very wise in the Type-II risk sense. For example, with the loss function used in (6) and $\alpha = 0.05$, we get that even for very small underlying P (i.e. extreme data), we will still choose action 1 whereas it seems more reasonable to select more extreme actions, minimizing Type-II loss, as the evidence against the null gets stronger. In case $\mathcal{H}_0 = \{P_0\}$ and $\mathcal{H}_1 = \{P_1\}$ are simple, this can be achieved by taking S to be a *likelihood ratio*: assuming P_j has density p_j relative to some μ,

$$S^{\text{LR}} := \frac{p_1(Y)}{p_0(Y)} \tag{8}$$

which are immediately seen to be e-variables ($\mathbf{E}_{P_0}[S^{\text{LR}}]] = \int p_0(y)(p_1(y)/p_0(y))d\mu = \int p_1(y)d\mu(y) = 1$), i.e. to satisfy (1). Extending likelihood-ratio based e-variable to composite $\mathcal{H}_0$ and $\mathcal{H}_1$ via the *reverse information projection* is the central topic of Grünwald et al. (2019); see also Section 5. We can compare $S^{\text{NP}(\alpha)}$ and S^{LR} if P underlying $S^{\text{NP}(\alpha)}$ is itself a monotonic function of the likelihood ratio S^{LR}, as it will be for the Neyman-Pearson test with optimal power. In the decision task above (5), when used in (7), $S^{\text{NP}(\alpha)}$ can, for each α, select at most 2 actions whereas S^{LR} leads to selection of action $0, 1, 2$ or 3 depending on the amount of evidence, at the price of imposing a larger threshold before any particular action is selected compared to the S^{α} that is optimal for that action (e.g. $S^{0.05}$ is optimal for action 1 in this sense).

Admissibility More generally, among all Type-I risk safe decision rules, we aim only for those that have *admissible* Type-II risk behaviour; we call a rule admissible if there exists no other decision rule that is never worse and sometimes strictly better. Lemma 1 below identifies the set of Type-II admissible decision rules as those that, for each y and b, follow (7) for some e-variable S. It has the flavour of a *complete class theorem* (Berger, 1985) showing that all reasonable GNP decisions may be based on e-variables.

The result does not indicate though which particular e-variable is Type-II risk *optimal* in any given situation — this is impossible to determine, because it depends on the definition of random variable B, which we assume to be unknown. In fact, in the above example, as long as P is a monotonic function of S^{LR}, S^{LR} and, for all $0 < \alpha < 1$, $S^{\text{NP}(\alpha)}$ are all admissible (Example 4). So what admissible e-variable to use? We do not have a complete answer to this question, but for many practical hypothesis testing problems we recommend the e-variables of Grünwald et al. (2019), that optimize the GRO (Growth-Rate optimality) criterion, related to good behaviour in an optional continuation setting. This criterion can be optimized for without knowing the definition of B; we refer to Grünwald et al. (2019) for details. For composite $\mathcal{H}_1$ and for estimation problems (see below) the construction of good e-variables/posteriors involves priors, but these have a 'hope' rather than 'belief' interpretation — see Section 4.

2.2 The Need for an E-Posterior

Now let us consider parametric models $\{P_\theta : \theta \in \Theta\}$. Any collection of p-value based Neyman-Pearson tests, one for each $\theta \in \Theta$ in the role of the null, can be 'inverted' to construct valid $1-\alpha$-confidence intervals, one for each α. Analogously, any collection of E-Variables $\{S_\theta : \theta \in \Theta\}$ with S_θ an e-variable for null $\{P_\theta\}$, can be used to construct a more robust (and wider) $1-\alpha$-confidence intervals. Just like it is tempting to interpret a 'system' of confidence intervals, one for each α, or a CD, as a type of 'posterior', one can also view the inverse $\bar{P}(\theta \mid Y) := S_\theta^{-1}(Y)$ as a type of posterior for parameter θ. The crucial difference is that this e-based posterior leads to valid (in a specific frequentist sense which we will define) inferences under a large class of decision-tasks that may be determined post-hoc, in a data-dependent fashion, whereas standard confidence intervals can only be used under the BIND assumption.

Example 2 Consider the normal location family: under P_θ the data are i.i.d. $\sim N(\theta, 1)$. With the standard (uniform, improper) 'objective Bayes' prior for this family and data $Y = x^n$, the posterior $W^\circ \mid Y = x^n$ 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 $1/n$ (Berger, 1985). In this case the objective Bayes posterior also coincides with the *fiducial* and the *confidence* distribution (CD) (Schweder and Hjort, 2016) based on x^n. These CD's are defined such that for each α, $[\theta_L, \theta_R]$ with θ_L the left-$\alpha/2$ quantile and θ_R the right-$\alpha/2$ quantile give the standard two-sided $(1-\alpha)$-confidence interval. The standard $(1-\alpha)$ Bayesian credible interval based on $w^\circ(\theta \mid x^n)$ and the standard $(1-\alpha)$-confidence interval therefore coincide. Analogously to how we connected Type-I error probability to Type-I risk, we can connect the validity of a $(1-\alpha)$ -confidence interval for a prespecified α to a risk guarantee of the form

$$\mathbf{E}_{Y\sim P_\theta}[B \cdot \mathbf{1}_{\theta \notin [\theta_L(Y), \theta_R(Y)]}] \leq \ell \quad \text{for some prespecified } \ell. \tag{9}$$

B, measuring how bad it is to make a mistake, may again depend on the data in potentially unknowable ways. The decision task is then to output a smallest possible $[\theta_L, \theta_R]$ for which (9) holds.

Based on the CD $w^\circ(\theta \mid Y)$, one would then presumably pick the smallest interval symmetric around $\hat{\theta}$ for which the Bayes posterior satisfies the required risk bound, i.e. the smallest $[\theta_L(Y), \theta_R(Y)]$ such that $|\theta_{\mathrm{L}}(Y) - \hat{\theta}| = |\theta_{\mathrm{R}}(Y) - \hat{\theta}|$ and

$$\mathbf{E}_{\bar{\theta}\sim W^\circ \mid Y=x^n}[b \cdot \mathbf{1}_{\bar{\theta} \notin [\theta_L(Y), \theta_R(Y)]}] \leq \ell \tag{10}$$

where b is the observed value taken by B; and for this interval, (10) holds with equality. The intuitive appeal for choosing this $[\theta_L(Y), \theta_R(Y)]$ is clear: (10) expresses that as a decision-maker one can expect the loss given the data to be bounded by ℓ; one simply wants to pick the smallest, most informative interval for which this holds true. Yet the *real* expectation of the loss may very well be different from (10) — it is given by

$$\mathbf{E}_{Y\sim P_{\theta^*}}\left[B(Y) \cdot \mathbf{1}_{\theta^* \notin [\theta_L(Y), \theta_R(Y)]}\right], \tag{11}$$

with θ^* indexing the true sampling distribution. This quantity may be much larger than ℓ (and hence than (10)) if B is dependent on Y. As an extreme example, fix any $\epsilon > 0$, If, whenever $Y \geq \epsilon$, we set $B := \ell/2F_0(-Y + \epsilon^2/Y)$ where F_0 is the CDF of a standard normal, then, under $\theta^* = 0$, (11) evaluates to ∞ (see Appendix A.1).

This discrepancy between what one *believes* will happen according to a posterior and what actually will happen has repercussions for Neyman's interpretation of statistics as inductive behaviour. To illustrate, imagine a decision-maker who is confronted with such a decision problem many times (each time j the underlying $\theta_{(j)}$ with $Y_{(j)} \sim P_{\theta_{(j)}}$ and the sample size $n_{(j)}$ and the importance function $B_{(j)}$ may be different). Then, based on (10) one would think to have, by the law of large numbers, the guarantee that, almost surely,

$$\limsup_{j\to\infty} \frac{1}{m} \sum_{j=1}^{m} B_{(j)} \cdot \mathbf{1}_{\theta \notin [\theta_L(Y_{(j)}), \theta_R(Y_{(j)})]} \leq \ell. \tag{12}$$

Unfortunately however, this statement is likely false if in reality there is a dependence between $B_{(j)}$ and $Y_{(j)}$ (the problem, may, for example, become a lot more relevant once one knows that $\hat{\theta}$ (and hence presumably θ) lies in a particular region of interest). Assume for example that there is a sequence of independent studies $Y_{(1)}, Y_{(2)}, \ldots$, all of which are of the form $Y_{(j)} = (X_{(j),1})$ and thus have sample size 1 (we may think of the $Y_{(j)}$ as z-scores summarizing studies of varying sample size). Instantiate $\epsilon = 0.01$ and set $B_{(j)} := \ell/2F_0(-Y_{(j)} + \epsilon^2/Y_{(j)})$ as above if $Y_{(j)} \geq \epsilon$ and $B_{(j)} = 0$ otherwise. Suppose that the $Y_{(j)}$ are all independently sampled from the same $\theta^* = 0$. Here is a sample of 15 corresponding $B_{(j)}$'s (generated by R):

$$1.15, 0, 3.44, 1.09, 1.91, 4.17, 10.40, 1.11, 0, 0, 1.47, 1.31, 0, 0, 0 \tag{13}$$

(while the sequence looks rather innocuous, the definition of B and the fact that θ^* stays the same over time may still feel artificial; see Section 2.3). Then, using (11), with $\theta_{\mathrm{L}}, \theta_{\mathrm{R}}$ chosen by (10), the limit in (12) will go to ∞ rather than to ℓ. The first reaction may be to require the decision-maker to model the dependency between Y and B. But the precise relation may be unknowable, and then it is not clear how to do this. This general discrepancy between posterior expectations and what can actually be expected fully disappears if we base our decisions on e-posteriors rather than confidence distributions/objective Bayes posteriors (Example 8).

E-posterior Minimax Decision Rules: beyond the Type-I/II Dichotomy The application above was still implicitly within the Type-I and Type-II risk paradigm, the Type-II loss being the *width* of the reported confidence interval. However, now that we look at estimation rather than testing, other types of loss functions and decision criteria do suggest themselves. In Section 4.2 we introduce the *E-Posterior Minimax Decision Rule*, which leads to decisions that minimize expected-loss bounds for standard loss functions such as squared error loss, where these 'luckiness' bounds (Example 5) hold irrespective of whether the prior assumptions encoded in the E-posterior hold or not, and the bounds can be evaluated based on one's sample.

Example 3 Consider the confidence distribution $W^\circ \mid X^n$ of the previous example again. 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 and on the specific study one is in in unknown ways. Take the Gaussian location family and set

$$L_b(\theta, a) = b \cdot (\theta - a)^2.$$

For simplicity, as in Example 2, we take $n = 1$ fixed, so that $\hat{\theta}(Y) = Y = X_1$. The idea is that we are presented a decision task with loss function $L_B(\theta, a)$. Above we assume that the

loss itself is linearly related to B; our approach extends to arbitrary relations (in general we might even deal with some b indicating a squared loss problem whereas other b indicating, say, an absolute loss problem), but linearity allows for a simple treatment since, if B is treated as independent of the data, then the Bayes optimal action based on $w^\circ(\theta \mid y)$ is simply the MLE, $\hat{\theta}$ irrespective of the observed B. Based on the this posterior, i.e. ignoring potential dependencies between Y and B, the loss we *think* we make upon observing Y and B, is given by $\mathbf{E}_{\bar{\theta}\sim W^\circ|Y}[L_B(\bar{\theta},\hat{\theta})] = \mathbf{E}_{\bar{Y}\sim P_{\hat{\theta}}}[L_B(\bar{Y},\hat{\theta})]$, so that the average of the losses we expect to make, in several studies, is given by

$$\mathbf{E}_{Y\sim P_{\theta^*}}[\mathbf{E}_{\bar{Y}\sim P_{\hat{\theta}(Y)}}[L_{B(Y)}(\bar{Y},\hat{\theta}(Y))]]$$

whereas the loss we *actually* make on average is $\mathbf{E}_{Y\sim P_{\theta^*}}[L_B(\theta^*,Y)]$. In Appendix A.1 we show that one can define B in such a way that the first expression is finite and the second expression is infinite. The numbers B one will actually be presented with in an i.i.d. sequence of studies of size 1 (as in the previous example) may again not reveal that something 'adversarial' is going on — here are 15 i.i.d. realizations of B (generated by R using the prescription of Example 9):

$$1.04, 1.62, 1.73, 1.07, 1.67, 1.17, 1.26, 1.00, 1.184, 1.02, 4.08, 1.60, 1.01, 1.07, 1.57.$$

In Example 9 we show that the e-posterior minimax decision rule also leads one to adopt the MLE here, but the discrepancy between 'true' and 'believed' expectation will disappear.

2.3 On the Proper Interpretation of These Results

The Quasi-Conditional Paradigm Assume we have a prior on $\mathcal{H}_0$ and $\mathcal{H}_1$ and priors W_0 and W_1 on the distributions inside these hypotheses. We can then use Bayes' theorem to calculate the Bayes posterior $P(\mathcal{H}_0 \mid Y)$ based on data Y and then define the *conditional Type-I error probability* to be simply equal to this posterior, $\hat{\alpha}_{|y} := P(H_0 \mid Y = y)$, implying that

$$\begin{array}{l}\text{"for any fixed } \alpha_0\text{, among all studies with } \hat{\alpha}_{|Y} \geq \alpha_0\text{,}\\ \text{we make a Type-I error at most a fraction } \alpha_0 \text{ of the time"}.\end{array} \tag{14}$$

Such a fully conditional statement, with post-hoc determined $\hat{\alpha}_{|Y}$, is only correct if the priors can be fully trusted. It would definitely be incorrect if we set $\hat{\alpha}_{|Y}$ either to a p-value or an e-value based on Y. Still, with e-values, as we have seen, we can define the decision task to be any arbitrary function of the data —even a function unknown to the statistician — and still get valid frequentist inference. Thus, we may term our approach *quasi-conditional*: it is fully conditional (arbitrary dependency on data possible) in terms of the decision task presented, yet it is unconditional in the sense that the performance of the approach is evaluated in expectation over all possible data, and not conditionally.

This sets it apart from all main existing approaches: the Bayesian approach allows full conditionality in terms of both evaluation and the decision task presented (it is perfectly alright if the decision task is decided upon in light of the data) — which comes at the price of strong assumptions. The classical NP approach and general frequentist decision theory is neither conditional in terms of evaluation nor in terms of decision task — the latter can be set freely, but it has to be set in advance or at least independently of the data. And finally, although our work has been inspired by classical *conditional frequentist* approaches (Kiefer,

1977, Berger et al., 1994, Berger, 2003) the latter are quite different in that they condition on some coarsening $\mathcal{C}$ of the data when evaluating procedures (giving '$\mathcal{C}$-conditional error probabilities') but do not allow the decision task to be set in arbitrary data-dependent ways. We plan to provide a more detailed comparison to these latter approaches elsewhere.

The quasi-conditional stance has more in common with *inferential models (IMs)* of R. Martin and collaborators (Martin and Liu, 2015, Balch et al., 2019, Martin, 2021). Like we do in the present work (Example 2 and 3) they point out discrepancies between what one would expect to happen (or think to happen with high probability) according to a Bayesian posterior and what can be expected to happen according to the unknown, true distribution and provided IMs as a safer alternative for a fiducial or 'objective Bayes' posterior. Unlike our inferential posteriors, the specific IMs proposed by (Martin and Liu, 2015) still work under the BIND assumption and thus may not provide reliable inferences if BIND does not hold. But it may very well be that other IMs (IMs constitute a family of methods rather than a single method) that essentially behave like e-posteriors as well; finding out if this is the case is a major goal for future work, as well as placing both IMs and the present work in the context of *safe probability*, a method for expressing clearly what decision tasks inference method can be safely used for Grünwald (2018).

3 Detailed Treatment of GNP Decision Problems

3.1 Type-I Risk Safety and Compatibility

Definition 1 *A* multi-loss decision task *is a tuple* $(\mathcal{B}, \Gamma, \{(\mathcal{A}_b, L_b : \Gamma \times \mathcal{A}_b \to \mathbb{R}_0^+) : b \in \mathcal{B}\})$ *with, for each* $b \in \mathcal{B}$, L_b *representing a loss function mapping* state of nature $\gamma \in \Gamma$ *and* action a_b *in* action space $\mathcal{A}_b$ *to the loss* $L_b(\gamma, a_b)$.

Relative to any given multi-loss decision task and random variable Y *taking values in set* $\mathcal{Y}$, *a* multi-loss decision rule *is defined to be any set of functions* $\{\delta_b : b \in \mathcal{B}\}$, *with* $\delta_b(y)$ *denoting the* $a \in \mathcal{A}_b$ *picked when loss function* L_b *is presented and* $Y = y$ *is observed.*

Definition 2 *A* GNP (Generalized Neyman-Pearson) decision task for testing *is a multi-loss decision task with* $\Gamma \in \{0, 1\}$, *accompanied by a set of* maximally acceptable risks ℓ_b, *one for each* $b \in \mathcal{B}$, *i.e. a collection* $\{(L_b(0, \cdot), L_b(1, \cdot), \ell_b, \mathcal{A}_b) : b \in \mathcal{B}\}$, *such that for all* $b \in \mathcal{B}$,

- $\mathcal{A}_b$ *is a subset of* $\mathbb{R}_0^+$ *containing* 0; *the* Type-I loss $L(0, \cdot) : \mathcal{A}_b \to \mathbb{R}_0^+$ *is an increasing function of* $a \in \mathcal{A}_b$, *while the* Type-II loss $L_b(1, \cdot) : \mathcal{A}_b \to \mathbb{R}$ *is strictly decreasing in* a *and* $\ell_b \in \mathbb{R}_0^+$.

Relative to any given GNP decision task and null hypothesis $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ *and random variable* Y *taking values in* $\mathcal{Y}$ *we further define:*

- *Let* δ *be any decision rule and let* $S = S(Y)$ *be any e-variable. We call* δ compatible with S *if we have* $L_b(0, \delta_b(y)) \leq \ell_b S(y)$ *for all* $y \in \mathcal{Y}$.

As in the examples in Section 2, we may usually assume $L_b(0, 0) = 0$ for all $b \in \mathcal{B}$. In practice, ℓ_b would be set by the decision-maker and indicate the maximum acceptable Type-I risk in case the decision task would be restrict to $\mathcal{B}' = \{b\}$, i.e. L_b would be presented irrespective of the data. For simplicity in interpreting the results we set $\ell_b := \ell$ for some fixed ℓ for all $b \in \mathcal{B}$ in all examples until Section 4.2, when the use of having ℓ_b depend on b will become clear.

Let us first consider a concrete setting in which a policy maker observes not just Y but also some other, random variable U expressing 'side-information' that is independent of Y and distributed according to some distribution Q. Based on (U, Y), the actual loss function L_B with index B to be used is decided using some additional, conditional distribution $R \mid U, Y$. Importantly, Q, R and the definition of U may be unknown to both statisticians and policymakers.

Now let $\mathcal{B}$ and $\mathcal{U}$ be countable. Let Δ_U be the set of all distributions on $\mathcal{U}$, and let $\Delta_{B|U,Y}$ be the set of all conditional distributions on $\mathcal{B}$ given U, Y, i.e. each $R \in \Delta_{B|U,Y}$ provides, for each $u \in \mathcal{U}$ and each $y \in \mathcal{Y}$, a distribution $R(B = \cdot \mid U = u, = y)$ on $\mathcal{B}$. For a given GNP decision task and null hypothesis $\mathcal{H}_0$, we say that decision rule δ is *Type-I risk-safe* if we have

$$\text{For all } Q \in \Delta_U, \text{ all } R \in \Delta_{B|U,Y}, \text{ all } P \in \mathcal{H}_0\text{: } \mathbf{E}_{Y\sim P}\mathbf{E}_{U\sim Q}\mathbf{E}_{B\sim R|U,Y}\left[\frac{L_B(0,\delta_B(Y))}{\ell_B}\right] \leq 1, \tag{15}$$

where we note that if ℓ_b is the same for all b we can replace division by ℓ_B by putting ℓ on the right. We may think of R as the distribution that a policy maker, or an adversary, or 'society' implicitly chooses to decide on the particular decision task to be solved once the outcome of the study is known. In practice, the space $\mathcal{U}$ of 'values' that U can take and the definition of R may be unknowable, but this need not concern us: the details of U, Q and R are irrelevant, since (15) is equivalent to

$$\sup_{P\in\mathcal{H}_0} \mathbf{E}_{Y\sim P}\left[\sup_{b\in\mathcal{B}} \frac{L_b(0,\delta_b(Y))}{\ell_b}\right] \leq 1. \tag{16}$$

To see the equivalence of (15) and (16), note that (16) $\Rightarrow$ (15) is obvious; the reverse implication follows by taking as distribution R in (15) the one which, for each $y \in \mathcal{Y}$ and $u \in \mathcal{U}$, puts probability 1 on the b achieving $\sup_{b\in\mathcal{B}} \frac{L_b(0,\delta_b(Y))}{\ell_b}$, assuming the supremum is achieved. If it is not achieved, the result follows by taking a limit of b's towards the supremum.

Given this equivalence we shall take (16), which avoids all tenuous assumptions about existence of countable sample spaces and random variables as definition of Type-I risk safety from now on.

E-Variable Compatibility determines Type-I Risk Safety and vice versa In NP Theory, Type-I error guarantees come first — we look for an optimal decision rule among all rules that have the Type-I error guarantee. Analogously, we first restrict our search for 'good' decision rules to those that are Type-I risk safe for the given decision task. The following observation shows that compatibility with an e-variable is the same as Type-I risk safety, thereby explaining the importance of e-variables to (generalized) NP testing:

Proposition 1 *Fix an arbitrary GNP decision task. For every decision rule δ defined relative to this problem:*

1. *For every e-variable S on $\mathcal{Y}$: if δ is compatible with S, then δ is Type-I risk safe.*
2. *Suppose that δ is Type-I risk safe. Set $S(y) := \sup_{b\in\mathcal{B}} L_b(0,\delta(y))/\ell_b$. Then (I) S is an e-variable, and δ is compatible with this S, and, (II), for all $b \in \mathcal{B}$, $L_b(0,0)/\ell_b \leq \inf_{y\in\mathcal{Y}} S(y)$. As a consequence, for arbitrary given δ, we have: if there exists no e-variable S on $\mathcal{Y}$ such that δ is compatible with S, then δ is not Type-I Risk safe.*

Proof: Both Part 1 and Part 2(I) are immediate from the definition. For Part 2(II), use that $L_b(0,a)$ is increasing in a, so that

$$\frac{L_b(0,0)}{\ell_b} \leq \inf_{y\in\mathcal{Y}} \frac{L_b(0,\delta_b(y))}{\ell_b} \leq \inf_{y\in\mathcal{Y}} \sup_{b'\in\mathcal{B}} \frac{L_{b'}(0,\delta_{b'}(y))}{\ell_{b'}} = \inf_{y\in\mathcal{Y}} S(y).$$

□

3.2 Type-II Admissibility and GNP Decision Rules

To move to Type-II risk, we must specify an alternative $\mathcal{H}_1$, like $\mathcal{H}_0$ a set of distributions for Y. For any given decision rule δ° relative to a given GNP decision task, any $\mathcal{H}_0$ and $\mathcal{H}_1$ and given Q and R as defined above, we define the Type-II risk as

$$\text{RISK}(1,\delta^\circ) := \sup_{P_1\in\mathcal{H}_1} \mathbf{E}_{Y\sim P_1}\mathbf{E}_{U\sim Q}\mathbf{E}_{B\sim R|U,Y}[L_B(1,\delta_B^\circ(Y))] \tag{17}$$

We call a decision rule δ° *Type-II strictly better* than decision rule δ if:

For all $P_1 \in \mathcal{H}_1$, $Q \in \Delta_U$, $R \in \Delta_{B|U,Y}$:

$$\mathbf{E}_{Y\sim P_1}\mathbf{E}_{U\sim Q}\mathbf{E}_{B\sim R|U,Y}[L_B(1,\delta_B^\circ(Y))] \leq \mathbf{E}_{Y\sim P_1}\mathbf{E}_{U\sim Q}\mathbf{E}_{B\sim R|U,Y}[L_B(1,\delta_B(Y))] \tag{18}$$

For some $P_1 \in \mathcal{H}_1$, some $Q \in \Delta_U$, some $R \in \Delta_{B|U,Y}$:

$$\mathbf{E}_{Y\sim P_1}\mathbf{E}_{U\sim Q}\mathbf{E}_{B\sim R|U,Y}[L_B(1,\delta_B^\circ(Y))] < \mathbf{E}_{Y\sim P_1}\mathbf{E}_{U\sim Q}\mathbf{E}_{B\sim R|U,Y}[L_B(1,\delta_B(Y))] \tag{19}$$

We call a decision rule δ *Type-II risk-admissible* if it is Type-I risk-safe and there is no other Type-I risk-safe decision rule δ° that is strictly better in the sense above. Clearly this definition is in the same spirit as standard admissibility definitions in classical statistical decision theory, and the lemma below is in the spirit of a *complete class theorem* (Berger, 1985) expressing that in searching for good decision rules in GNP problems we can restrict ourselves to those based on e-variables via the *maximal decision rule*, which we now define:

For given e-variable S, the *maximal* decision rule δ^* relative to S upon observing $B = b$ and $Y = y$, is given by:

$$\delta_b^*(y) := \text{the largest } a \in \mathcal{A}_b \text{ such that } L_b(0,a) \leq \ell_b S(y). \tag{20}$$

To state Lemma 1 we need two additional definitions: we call an e-variable S *sharp* if for some $P \in \mathcal{H}_0$, $\mathbf{E}_P[S] = 1$. Let $B : \mathcal{Y} \to \mathcal{B}$ be some function. We call decision rule δ^* defined relative to e-variable S as in (20) *B-sharp* if $\delta^*_{B(y)}(y)$ satisfies (20) a.s. with equality, i.e. for all $P \in \mathcal{H}_0$,

$$L_{B(Y)}(0,\delta_{B(Y)}(Y)) = \ell_{B(Y)} \cdot S(Y), \text{ with } P\text{-probability 1.}$$

For a sharp e-variable S, there cannot be another e-variable that gives uniformly more evidence against the null, i.e. that is almost surely not smaller but with positive probability strictly larger. For a B-sharp decision rule against a sharp e-variable S, there cannot be another decision rule with a.s. uniformly smaller Type-II losses under L_B that is still Type-I risk safe.

Lemma 1 *Suppose that all $P \in \mathcal{H}_0 \cup \mathcal{H}_1$ have full support $\mathcal{Y}$. Suppose further that (I) for some $c \geq 0$, for all $b \in \mathcal{B}$, $L_b(0,0) \leq c\ell_b$; and (II) for all $b \in \mathcal{B}$, $\mathcal{A}_b$ is either finite, or $L_b(0,a)$ is continuous in a with either (IIa) $\sup_{a\in\mathcal{A}_b} L_b(0,a) = \infty$ or (IIb) $\sup_{a\in\mathcal{A}_b} L_b(0,a) = \max_{a\in\mathcal{A}_b} L_b(0,a)$ is achieved. Then:*

1. *For any e-variable S that is bounded below by c, (20) has a unique solution δ^*; and this maximal δ^* is compatible with S (and hence by Proposition 1, Type-I risk safe).*

2. *All Type-II risk admissible decision functions are essentially of the form (20) relative to some e-variable S, in the sense that if δ is Type-II risk admissible, then there exists an S such that, with δ^* a maximal rule as in (20), for all $P \in \mathcal{H}_0 \cup \mathcal{H}_1$, $b \in \mathcal{B}$: $P(\delta_b(Y) = \delta_b^*(Y)) = 1$.*

3. *Suppose that $\mathcal{B}$ contains the special value* TRIV *with $\mathcal{A}_{\text{TRIV}} = \{0\}$ and $L_{\text{TRIV}}(0,0) = L_{\text{TRIV}}(1,0) = 0$; ℓ_{TRIV} can be set to any value ≥ 0. Suppose there exists a sharp e-variable S and a function $B : \mathcal{B} \to \mathcal{Y}$ such that the maximal δ^* defined by (20) exists and is B-sharp relative to this S. Then δ^* is admissible.*

The condition in Part 1 that S is bounded below by $c \geq 0$ is automatically satisfied in the natural setting that for all $b \in \mathcal{B}$, $L_b(0,0) = 0$. If we have $L_b(0,0) > 0$ (this will become relevant in Section 4.2) then we can modify any given S to a 'dampened' version $S^{[1/2]} := (1/2) + (1/2)S$. Clearly, $S^{[1/2]}$ is still an e-variable, and the condition would now automatically hold if we set $\ell_b = 2L_b(0,0)$.

The second part illustrates that we can restrict our search for admissible decision rules to the ones that are maximal for some e-variable S. The third part illustrates that such admissible decision rules do exist, assuming that the decision maker *may* be presented a trivial decision task, in which no choice is available — which can also be interpreted as a post-hoc cancellation of the real decision task.

Example 4 Consider a simple vs. simple testing problem with $\mathcal{H}_0 = \{P_0\}$ and $\mathcal{H}_1 = \{P_1\}$. Let $\text{P}(Y)$ be a strict p-value, i.e. $P_0(\text{P} \leq \alpha) = \alpha$ for $\alpha \in [0,1]$, that is monotonically decreasing in the likelihood ratio $S^{\text{LR}}(Y)$; use of such a p-value is standard in Neyman-Pearson testing with continuous-valued outcome spaces. Consider the following variation of Example 1: $\mathcal{B} = \mathbb{R}^+ \cup \{\text{TRIV}\}$ with trivial loss function L_{TRIV} as defined above, and for $b \in \mathbb{R}^+$, $\mathcal{A}_b = \{0, b\}$ and $L_b(0,0) = 0, L_b(0,b) = b$ and we set $L_b(1,0) := \ell_b := 1$ for all $b \in \mathbb{R}^+$. Consider for all $0 < \alpha < 1$, the decision rule δ^* as in (20) relative to e-variable $S^{\text{NP}(\alpha)}$. When presented with loss function L_b, this decision rule always plays 0 if $b > 1/\alpha$. If $b \leq 1/\alpha$, it plays b if $b \leq S^{\text{NP}(\alpha)}$ (i.e. if $S^{\text{NP}(\alpha)} = 1/\alpha$, i.e. if $\text{P} \leq \alpha$) and 0 otherwise (i.e. if $S^{\text{NP}(\alpha)} = 0$, i.e. if $\text{P} > \alpha$). By Part 3 of the lemma above, this decision rule is admissible for all α. The conditions are easily verified by noting that the e-variable is trivially sharp, and taking $B(y) = 1/\alpha$ to be constant. Then δ^* becomes B-sharp relative to $S^{\text{NP}(\alpha)}$.

Now consider the δ^* as in (20) based on the likelihood ratio e-variable S^{LR}. When presented L_b, this decision rule plays b if $b \leq S^{\text{LR}}$ and 0 otherwise. This decision rule is admissible as well: to verify the conditions of Part 3 of the lemma above, note that, again, S^{LR} is sharp. To see this, define $B(y) := S^{\text{LR}}(y)$ (corresponding to the situation in which an adversary always poses the highest-loss decision rule at which one would still reject the 0). Then δ^* is B-sharp relative to S.

4 E-Posteriors for General Decision Problems

We now let $\{P_\theta : \theta \in \Theta\}$ be a statistical model. For simplicity we assume that all parameters in $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$ are of interest; the case with nuisance parameters is deferred to Section 5. Let $Y, \mathcal{Y}$ be as before and let $\mathcal{S} = \{S_\theta : \theta \in \Theta\}$ be a collection such that for each

$\theta \in \Theta$, $S_\theta = S_\theta(Y)$ is an e-variable relative to null hypothesis $\mathcal{H}_0 = \{P_\theta\}$. The *e-posterior* corresponding to $\mathcal{S}$ is defined simply by setting, for all $y \in \mathcal{Y}$, $\bar{P}(\theta \mid y) := S_\theta^{-1}(y)$, with conventions about division by 0 as underneath (7).

Although it is not required by the definition above, all e-posteriors we encounter in our examples are actually based on e-processes, a more general notion than e-variable that arises when all the $P_0 \in \mathcal{H}_0$ determine the distribution of a discrete random process $\{X_i : i \in \mathbb{N}\}$ with each X_i taking values in a set $\mathcal{X}$. Such e-process-based e-posteriors (although not under this name) play a major role in existing work (such as (Howard et al., 2021) and the other references below) on *anytime-valid confidence sequences*, and reviewing this work seems the best way to introduce them.

E-Processes and Anytime-Valid Confidence We observe a random vector $Y = X^\tau$ for some stopping time τ defined relative to some filtration $(\sigma(V^n) : n \in \mathbb{N}\}$ where V_i is a coarsening of $X^i = (X_1, \ldots, X_i)$ (usually we just take $V_i = X_i$ but other choices are sometimes convenient (Grünwald et al., 2019)). An *e-process* (Ramdas et al., 2021) is a function $s : \bigcup_{n \in \mathbb{N}} \mathcal{X}^n \to \mathbb{R}_0^+$ such that for *every* stopping time τ (defined relative to the filtration above), $s(X^\tau)$ is an e-variable. We set $Y := X^\tau$ for some τ whose precise definition may be unknown; in practice we only observe that $\tau = n, Y = X^n$. This information is sufficient to calculate the e-variable $S := S(Y) = s(X^\tau)$, since for all n we must have $\tau = n \Rightarrow s(X^\tau) = s(X^n)$. Thus, the e-posteriors below may invariably be viewed in two ways: based on a collection $\mathcal{S}$ of e-variables for fixed n, but also as based on a collection $\mathcal{S}'$ of e-processes, turned into a collection of e-variables by the stopping rule τ. Since for all x^n for which $\tau = n$, '$\tau = n; X^\tau = x^n$' is the same event as $X^n = x^n$, we can and will abbreviate $\bar{P}(\theta \mid \tau = n; X^\tau = x^n)$ to $\bar{P}(\theta \mid x^n)$, as is also a common abbreviation for the standard Bayes posterior. We then have $\bar{P}(\theta \mid x^n) = s^{-1}(x^n)$.

A direct consequence of $\bar{P}(\theta \mid y)$ being an e-posterior is that under any stopping time τ, for all $\theta \in \Theta$,

$$P_\theta(\bar{P}(\theta \mid X^\tau) \leq \alpha) = P_\theta\left(s_\theta(X^\tau) \geq \frac{1}{\alpha}\right) \leq \alpha, \tag{21}$$

where we used that $s_\theta(X^\tau)$ is an e-variable and then Markov's inequality. (21) expresses that $(\text{CS}_{\alpha,n} : n \in \mathbb{N})$ with $\text{CS}_{\alpha,n} = \{\theta \in \Theta : \bar{P}(\theta \mid X^n) \geq \alpha\}$ is a $1 - \alpha$ *anytime-valid confidence sequence*. We observe that data collection is stopped at some n (i.e. $\tau = n$) and we observe data $X^n = x^n$, and we output $\text{CS}_{\alpha,n}$. We can be sure that $P_\theta(\theta \in \text{CS}_{\alpha,n}) \geq \alpha$ irrespective of the definition of τ; in particular we may not know this definition, as will often be the case in practice. The use of e-processes $\{S_\theta : \theta \in \Theta\}$ for constructing such AV confidence sets is well-known; by re-casting their inverse as e-posteriors we highlight their reliable usability in the more complex decision problems of Section 4.1 and 4.2 that *do not involve an a priori fixed* α. Before we move to these we give some examples of e–posteriors, based on well-known constructions for anytime-valid confidence sequences.

Prior-based e-posteriors One simple type of e-posterior directly relates to standard Bayesian posteriors: let W be a distribution on Θ. We can define a P_θ-e-process by setting $S_\theta(y) = \frac{p_W(y)}{p_\theta(y)}$ where p_W is the Bayes marginal density $p_W(y) := \int p_\theta(y) dW(\theta)$. We denote the corresponding e-posterior by $\bar{P}_{[W]}$. In case W has density w, we have

$$\bar{P}_{[W]}(\theta \mid y) = \frac{p_\theta(y)}{p_W(y)} = \frac{w(\theta \mid y)}{w(\theta)} \tag{22}$$

where $w(\theta \mid y)$ is the standard Bayes posterior density of θ given y. We now explore this particular e-posterior for the Gaussian location family and later, in Example 9 for general 1-dimensional exponential families. Our aim here is simplicity in illustration — prior-to-posterior ratios for much more complex models such as Gaussian processes were earlier explicitly used by Waudby-Smith and Ramdas (2020) and Neiswanger and Ramdas (2021).

Example 5 [The Normal Location Family - smooth prior] 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 $\hat{\theta}$ the MLE $(\sum_{i=1}^n x_i)/n$, and posterior variance $1/(n+\lambda)$, i.e. with density

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

so that, by (22),

$$\bar{P}_{[W]}(\theta \mid y) = \sqrt{\frac{n+\lambda}{\lambda}} \cdot e^{-\frac{n+\lambda}{2}(\theta-\breve{\theta})^2 + \frac{\lambda}{2}\cdot(\theta-\theta_0)^2} = \sqrt{\frac{n+\lambda}{\lambda}} \cdot e^{-\frac{n}{2}(\theta-\hat{\theta})^2 + \frac{1}{2}\cdot\frac{n\lambda}{n+\lambda}\cdot(\hat{\theta}-\theta_0)^2} \tag{23}$$

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. Note that $\bar{P}_W(\theta \mid Y) \leq \alpha$ iff $\theta \in \overline{\mathrm{CS}}_{\alpha,n}$ with

$$\overline{\mathrm{CS}}_{\alpha,n} = \left\{ \theta \in \mathbb{R} : (\theta - \hat{\theta})^2 \geq \frac{2}{n} \cdot \left(-\log \alpha + \frac{1}{2} \log \frac{n+\lambda}{\lambda} \right) + \frac{\lambda}{n+\lambda} \cdot (\hat{\theta} - \theta_0)^2 \right\}. \tag{24}$$

We see that $\bar{P}(\theta \mid y)$ plays a role analogous not to a density but to a *tail probability* $\int_{\theta' \geq \theta} w(\theta \mid y)$ of a Bayesian posterior/CD, which is why we write $\bar{P}$ with capital P.

A crucial difference between Bayesian posteriors and prior-based e-posteriors is that the likelihood ratio defining the latter is allowed to depend on the θ in the numerator. Thus, it is o.k. (since its inverse defines e-processes) to work with $\bar{P}_{[W_\theta]}$ defined by $\bar{P}_{[W_\theta]}(\theta|y) := p_\theta(y)/p_{[W_\theta]}(y)$ with W_θ a prior whose definition depends on θ. If, for example, we take as W_θ the Gaussian with mean $\theta_0 := \theta$ and variance λ, then (24) reduces to

$$\overline{\mathrm{CS}}_{\alpha,n} = \left\{ \theta \in \mathbb{R} : (\theta - \hat{\theta})^2 \geq \frac{n+\lambda}{n} \cdot \frac{2}{n} \cdot \left(-\log \alpha + \frac{1}{2} \log \frac{n+\lambda}{\lambda} \right). \right\} \tag{25}$$

Luckiness: E-posteriors vs. Bayesian posteriors as Hope vs. Belief The confidence sequence $\mathrm{CS}_{\alpha,n} = \mathbb{R} \setminus \overline{\mathrm{CS}}_{\alpha,n}$ defined by (24) depends on the data Y via the MLE $\hat{\theta} = \hat{\theta}(Y)$: the closer $\hat{\theta}$ to the mean θ_0 of prior W, the narrower and hence 'better' the interval. Still, $\mathrm{CS}_{\alpha,n}$ is valid irrespective of whether the prior W is in any sense 'correct' or representative. We may thus say that whereas in Bayesian inference, W encodes *belief*, in e-posterior applications W encodes something weaker which might be called *hope*: if the data is well-aligned with the prior (MLE $\hat{\theta}$ close to θ_0) our bounds improve, but they are valid irrespective of which $\theta \in \Theta$ generates the data. Bounds such as (24) are called *luckiness bounds* (if one is lucky, the bound will be good) and this way of thinking about data-dependent bounds was pioneered

in the PAC-Bayesian literature (Shawe-Taylor and Williamson, 1997, Grünwald and Mehta, 2019). From (25) we see that in this particular example, the data dependency disappears if we use prior W_{θ_0}; note that $\text{CS}_{\alpha,n}$ in (24) will be narrower than the one based on (25) in the 'lucky' event that $|\hat{\theta} - \theta_0|$ is within a constant times $(\log n)/n$.

If we compare the size of a standard confidence interval to (24), we see that the latter is wider by a factor of order $\log n$. We can use a different θ-dependent prior W in which this logarithmic blow-up is replaced by a small constant factor if, again, we are *lucky*, but in a different sense. This approach works using a discrete prior that anticipates the α that one will probably be interested in, to be denoted α^*, and the sample size n one plans for or hopes to achieve, denoted as n^*. We now present e-posteriors for Gaussian location families leading to anytime-valid CS's that are within a constant factor of the standard width as long as the actual n and α are close to the anticipated n^* and α^*. Crucially though, if the actual values are not equal to the anticipated ones (by early or late stopping or optional continuation), the CI is still valid, unlike the case for a standard confidence interval; and its width deteriorates gracefully as the discrepancy ratio c as defined in (27) moves away from 1. In Appendix A.2 we point out how the construction can be generalized to 1-dimensional exponential families.

Example 6 **[θ-dependent discrete prior]** Fix *anticipated* sample size n^* and confidence level $0 < \alpha^* < 1$. For each θ we define $\theta^- < \theta$ and $\theta^+ > \theta$ to satisfy

$$\frac{1}{2} n^* (\theta - \theta^+)^2 = \frac{1}{2} n^* (\theta - \theta^-)^2 = -\log \frac{\alpha^*}{2}. \tag{26}$$

Now define the e-variable $S_\theta(y) = (1/2)\frac{p_{\theta^-}(y)}{p_\theta(y)} + (1/2)\frac{p_{\theta^+}(y)}{p_\theta(y)}$. We will use it for actual sample sizes n and levels $0 < \alpha < 1$ that are not necessarily equal to the hoped-for n^* and α^*. In Appendix A.2 we show that a sufficient condition for $S_\theta(Y) \geq \alpha^{-1}$ is that

$$\left(\theta - \hat{\theta}\right)^2 \geq \frac{1}{2} \cdot \left(\frac{-\log(\alpha/2)}{n}\right) \cdot \left(c^{1/2} + c^{-1/2}\right)^2 \quad \text{with} \quad c := \frac{n^*/(-\log(\alpha^*/2))}{n/(-\log(\alpha/2))} \tag{27}$$

Note that this corresponds to the rejection region at level α for the test based on S_θ, which is thus seen to have width $|\theta - \hat{\theta}| \asymp 1/\sqrt{n}$, of the same order as the region for the standard Neyman-Pearson test, with a factor depending on how well aligned n, n^*, α and α^* are. S_θ gives the e-posterior promised above: we set $\bar{P}_{[n^*,\alpha^*]}(\theta \mid Y) := 1/S_\theta(Y)$.

In practice, one would use $\bar{P}_{[n^*,\alpha^*]}$ if one is quite sure that n will be close to n^*, and $\bar{P}_{[W]}$ for small λ otherwise. Combining ideas from Example 5 and 6 one can also get e-posteriors like those in Example 5 that do not have strong hopes about n^*, but with a log log -dependence rather than a log-dependence n; see the techniques of 'stitching' (Howard et al., 2021) or 'switching' (Van der Pas and Grünwald, 2018). This, however, comes at the cost of worse multiplicative constants. While the e-processes of Example 5 and Example 6 and the variations we just mentioned are well-known in E-Value circles, we now continue with their novel interpretations that, we feel, justifies the terminology *e-posterior*.

4.1 E-posteriors, Confidence and Type-I Risk

Our first use of the e-posterior is for deriving confidence statements in the post-hoc setting. For this, we modify Definition 2 as follows:

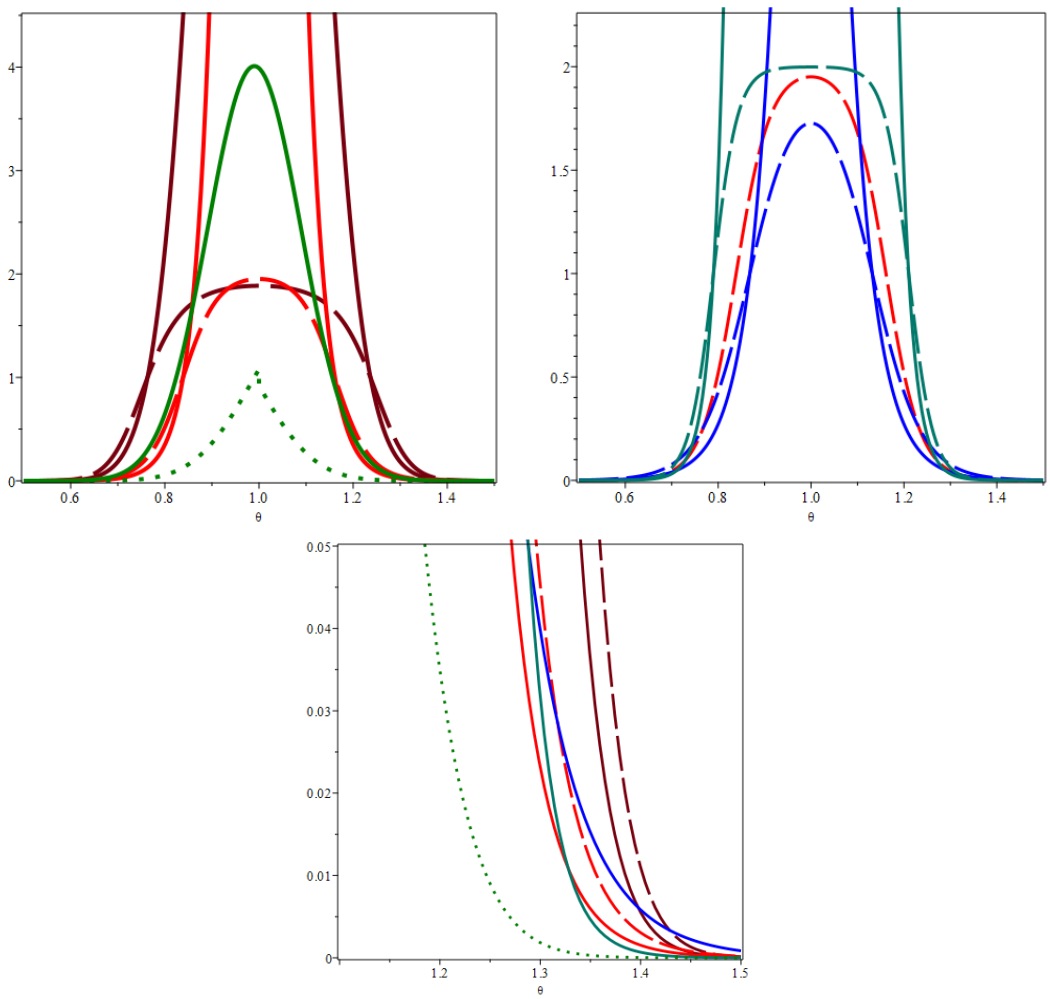

Figure 1: *(top left)* the red-brown solid curve is the e-posterior $\bar{P}_{[W]}(\theta \mid y)$ for the normal location family based on a sample $y = (x_1, \ldots, x_{100})$ with $\hat{\theta}(y) = 1$ and prior W with $\lambda = 1$ and $\theta_0 = 0$ as in (23) (it reaches a maximum of about 15). The light-red solid curve is the discrete e-posterior $\bar{P}_{[n^*,\alpha^*]}$ for $\alpha^* = 0.05$ and with a well-aligned $n^* = n = 100$. The dashed lines are the corresponding 1/2-dampened e-posteriors as used in Example 9; their maximum is by design bounded by 2. For comparison, the solid green curve shows the standard posterior $w(\theta \mid y)$ (compared to the confidence distribution $w^\circ(\theta \mid y)$ of Section 2.2 it is slightly tilted towards $\theta = 0$ but otherwise visually indistinguishable). More informatively, the dotted green curve shows the tail area of the standard posterior, $\int_{\theta':|\theta'-\hat{\theta}|\geq\theta} w(\theta|y)d\theta$. *(top right)* the light-red dashed curve is as on the left (note the different scale), the dampened discrete e-posterior $\bar{P}^{[1/2]}_{[100,0.05]}$. For comparison we show $\bar{P}_{[200,0.05]}$ (in blue) and $\bar{P}_{[50,0.05]}$ (in blue-green) with misaligned n^* (dashed lines are 1/2-dampened versions). *(bottom)* Focus on the right tail, with the same legend as before ($\bar{P}^{[1/2]}_{[200,0.05]}$ and $\bar{P}^{[1/2]}_{[50,0.05]}$ lie about 0.05 to the right of their undampened counterparts and are not shown; neither is the Bayes posterior density). Note that the Bayes posterior tail area is 0.05 at $\hat{\theta} + 1.96/\sqrt{n} \approx 1.2$, as is to be expected from a standard confidence/credible interval; the discrete posterior (light red) reaches 0.05 at $\hat{\theta} + 2.72/\sqrt{n} \approx 1.3$, in accordance with Example 8. It is seen that for confidence-statements as in Section 4.1, the un-dampened posteriors are to be preferred.

Definition 3 *A* GNP decision task for estimation *relative to given parametric model* $\{P_\theta : \theta \in \Theta\}$ *is a multi-loss decision task with* $\Gamma = \Theta \cup \{\text{II}\}$*, accompanied by a set of maximally acceptable risks* ℓ_b*, one for each* $b \in \mathcal{B}$*, i.e. a collection* $\{(\mathcal{A}_b, L_b : \Theta \times \mathcal{A}_b \to \mathbb{R}_0^+, L_b(\text{II}, \cdot) : \mathcal{A}_b \to \mathbb{R}_0^+, \ell_b) : b \in \mathcal{B}\}$ *such that for all* $b \in \mathcal{B}$*, we have* $\ell_b \in \mathbb{R}_0^+$*. Relative to any given such task and and decision rule* δ *we further define:*

- *Let* $\bar{P}(\theta \mid Y)$ *be any e-posterior relative to* $\{P_\theta : \theta \in \Theta\}$*. We call* δ compatible with $\bar{P}$ *if we have* $\bar{P}(\theta \mid y) \cdot L_b(\theta, \delta_b(y)) \leq \ell_b$ *for all* $y \in \mathcal{Y}, b \in \mathcal{B}, \theta \in \Theta$*.*
- *We call* δ *Type-I risk safe if*

$$\sup_{\theta \in \Theta} \mathbf{E}_{Y \sim P_\theta} \left[\sup_{b \in \mathcal{B}} \frac{L_b(0, \delta_b(Y))}{\ell_b} \right] \leq 1. \tag{28}$$

Note that there is no alternative $\mathcal{H}_1$ any more, but there still is a Type-II loss, now denoted as $L_b(\text{II}, a)$, the Type-II risk being defined as in (17) but with the supremum over $\{P_\theta : \theta \in \Theta\}$ (i.e. as if $\mathcal{H}_0 = \mathcal{H}_1$). By the same reasoning as in Proposition 1, we get once again that every decision rule that is compatible with some e-posterior $\bar{P}(\theta \mid Y)$ is Type-I risk safe and vice versa. Again we aim for a decision rule minimizing Type-II risk under a constraint of the Type-I risk. However, the set $\mathcal{A}_b$ is now not necessarily embedded in $\mathbb{R}$ any more, hence we cannot impose that $L_b(\theta, a)$ is 'increasing' or $L_b(\text{II}, a)$ is 'decreasing' in a. Rather than seeking for a 'maximal' decision rule as in (20) we will therefore simply look, among all compatible decision rules for some given e-posterior, for one that has small Type-II loss.

Although the setting is more general, we only consider its instantiation to inference of confidence intervals: Θ is an interval in $\mathbb{R}$ and $\mathcal{A}_b = \mathcal{A} := \{[\theta_L, \theta_R] : \theta_L, \theta_R \in \Theta, \theta_R \geq \theta_L\}$ for all $b \in \mathcal{B}$. Action $a := [\theta_L, \theta_R]$ represents a confidence interval and we set the Type II-criterion $L(\text{II}, [\theta_L, \theta_R]) = g(|\theta_R - \theta_L|)$ (independent of b) to be a strictly increasing function of its width such that the cost of 'abstaining' — which amounts to giving as interval the full Θ — is given by $g(\infty) := \ell$ for some fixed $\ell > 0$ (any function g that is strictly increasing with limit ℓ will do)*. We can set, as a simple example, with $a = [\theta_\text{L}, \theta_\text{R}]$, $L_\theta(\theta, a) = b \cdot \mathbf{1}_{\theta \notin a}$ so that b expresses, in a very simple way, how important the decision is. We choose $\ell_b := 1$ for all b. In the examples below we report a confidence interval that is symmetric around the MLE and that is compatible to an e-posterior in the sense of Definition 3, so that we have Type-I risk safety. Among all such intervals we report the smallest one, so that our Type-II loss (confidence width) is small as well. The width of the confidence interval a that we report based on this procedure when presented $B = b$ will increase logarithmically with b, as illustrated below. Alternatively, we could choose $\ell_b := b$ (the choice is up to us, decision makers); then the reported interval will not depend on the observed $B = b$, but the assessment of the Type-I risk by the e-posterior will become dependent on this b.

Example 7 [Normal Location Family, Continued] Suppose you observe $Y = y$, $B = b$. Then with the smooth-prior-based e-posterior as in Example 5, we get:

$$\bar{P}_{[W]}(\theta \mid Y) \cdot L_b(\theta, [\theta_\text{L}, \theta_\text{R}]) = b \cdot \mathbf{1}_{\theta < \theta_\text{L} \vee \theta > \theta_\text{R}} \cdot \sqrt{\frac{n+\lambda}{\lambda}} \cdot e^{-\frac{n}{2}(\theta - \hat{\theta})^2 - \frac{n\lambda(\hat{\theta} - \theta_0)^2}{2(n+\lambda)}}$$

*The idea can be extended in various ways to multivariate Θ by looking at volume instead of width, but we will not pursue this here.

which, from (24) is bounded by ℓ_b if

$$\theta_{\mathrm{L}} \leq \hat{\theta} - A \text{ and } \theta_{\mathrm{R}} \geq \hat{\theta} + A \text{ where } \; A = \sqrt{\frac{2}{n}} \cdot \sqrt{\log \frac{b}{\ell_b} + \frac{1}{2} \log \frac{n+\lambda}{\lambda} + \frac{n\lambda(\hat{\theta} - \theta_0)^2}{n+\lambda}}.$$

To minimize the Type-II risk, we set θ_{L} and θ_{R} to satisfy these with equality. Suppose we set $\ell_b := \ell$ irrespective of b. Then with this procedure, we have a guaranteed Type-I risk bounded by ℓ, with interval widths scaling as $\sqrt{\log b}$.

We can also use the e-posterior based on the discrete prior which anticipates (is optimized for) a particular n^* and α^*, while still giving valid bounds for other n and α, as in Example 6. For the normal location family, reasoning analogously as above and using (27) and substituting ℓ_b/b for α (so that now $c = (n^*/n) \cdot (\log(2b/(\ell_b))/(-\log(\alpha^*/2)))$, gives that $\bar{P}_{[n^*,\alpha^*]}(\theta \mid Y) \cdot L_b(\theta, [\theta_{\mathrm{L}}, \theta_{\mathrm{R}}])$ is bounded by ℓ_b if

$$\theta_{\mathrm{L}} \leq \hat{\theta} - A \text{ and } \theta_{\mathrm{R}} \geq \hat{\theta} + A \text{ where } \; A = \sqrt{\frac{2}{n}} \cdot \sqrt{\log \frac{2b}{\ell_b}} \cdot \left(\frac{c^{1/2} + c^{-1/2}}{2} \right). \tag{29}$$

and we may again choose θ_{L} and θ_{R} to satisfy this with equality. The closer n to the anticipated n^*, the smaller the confidence interval. This formula is valid for $b > \ell_b/2$ and useful for $b \geq \ell_b$. For, if $b < \ell_b$ then the desired risk bound is obtained trivially by any interval, including the empty one: the maximum accepted risk is ℓ_b and the maximum loss, b, would then be smaller.

Example 8 [Example 2, Cont.] Let us compare the confidence intervals obtained by (10) (i.e. based on the standard objective Bayes/CD $W^\circ \mid X^n$, assuming b to be a fixed constant) to e-posterior induced confidence intervals, using 'scale' $\ell = 1$ throughout. First consider a case that b is indeed fixed, say to 20. Then using (10) we get the standard 95% confidence interval $\hat{\theta} \pm 1.96/\sqrt{n}$. If we would anticipate this b and n (yet would still want valid inferences if b were generated by some other means or the stopping time would turn out different from n) we can apply the e-posterior $\bar{P}_{[n^*,\alpha^*]}$ with $n^* = n$ and $\alpha^* = \ell/2b = 0.025$ so that c in (29) becomes 1. We get that based on the e-posterior, we output $\hat{\theta} \pm \sqrt{2(\log 40)}/\sqrt{n} = \hat{\theta} \pm 2.72/\sqrt{n}$, so our CI is wider by a constant factor of about 1.4.

Now let us see what happens if we work with this $\bar{P}_{[n^*,\alpha^*]}$ (in particular, $\alpha^* = 0.05$, $n^* = n$, and we take $\ell = 1$) if in reality B is not fixed. From (29) we get, using these choices, that the e-posterior confidence interval upon observing $B = b$ is given by $[\hat{\theta} - A, \hat{\theta} + A]$ with $A = \sqrt{2/n}(\log(2b)/\sqrt{\log 40} + \sqrt{\log 40})$. By construction, this gives by equivalence of compatibility and Type-I risk safety, the desired bound $\mathbf{E}_{Y \sim P_\theta}[L_B(\theta, \delta_B(Y))] \leq 1$, which holds irrespective of the true θ. So, according to the e-posterior, we expect a loss bounded by ℓ, and we get a loss bounded by ℓ.

4.2 The E-Posterior Minimax Decision Rule

We now drop the NP Type-I/II risk dichotomy paradigm and re-consider the mechanism by which E-variables and posteriors provided risk bounds to discover alternative reasonable decision rules. For each $\theta \in \Theta, b \in \mathcal{B}, y \in \mathcal{Y}$, let $\ell_{b,y}$ be such that

$$\frac{L_b(\theta, \delta_b(y))}{\ell_{b,y}} \leq S_\theta(y) \text{ i.e. } \bar{P}(\theta \mid y) L_b(\theta, \delta_b(y)) \leq \ell_{b,y}, \text{ for all } y \in \mathcal{Y}. \tag{30}$$

This corresponds to choosing $\ell_{b,y}$ to get 'compatibility' in the sense of Definition 2 and 3 but now we allow the maximum-acceptable-risk $\ell_{b,y}$ to also depend on the data Y themselves. Just as in (16), we get the bound:

$$\mathbf{E}_{Y\sim P_\theta}\left[\frac{L_B(\theta,\delta_B(Y))}{\ell_{B,Y}}\right] \leq \mathbf{E}_{Y\sim P_\theta}\left[\sup_{b\in\mathcal{B}}\frac{L_b(\theta,\delta_b(Y))}{\ell_{b,Y}}\right] \leq 1. \tag{31}$$

If one knows δ and $\bar{P}$, one can employ the best bound $\ell'_{b,y}$ for which (30), and hence the resulting bound (31) holds: clearly $\ell'_{b,y} = \max_{\theta\in\Theta}\bar{P}(\theta \mid y)L_b(\theta,\delta_b(y))$. But this further suggests to use the decision rule $\bar{\delta}$ for which this bound is itself minimized, i.e. to pick a decision rule $\bar{\delta}$ satisfying, for all $y\in\mathcal{Y}$, $b\in\mathcal{B}$,

$$\max_{\theta\in\Theta}\bar{P}(\theta \mid y)\cdot L_b(\theta;\bar{\delta}_b(y)) = \min_{a\in\mathcal{A}_b}\max_{\theta\in\Theta}\bar{P}(\theta \mid y)\cdot L_b(\theta;a). \tag{32}$$

We call any such decision rule *e-posterior minimax*. Note that this rule does not require a secondary, Type-II loss criterion! In earlier applications of e-variables and posteriors, we imposed ℓ ourselves and it made sense to choose it independently of b, but now it is determined so as to give the best possible bounds so we want it to depend on b and even on y. We shall now establish that in our running examples, the MLE is e-posterior minimax optimal, and (a close approximation to) the corresponding bound $\ell'_{b,y}$ can be easily calculated.

If we try this directly with the e-posteriors designed in the previous section we do not always get sharp bounds, due to the high variability of $\bar{P}(\theta \mid Y) = S_\theta^{-1}(Y)$, as can be seen from the left panel in Figure 1. To (sometimes vastly) improve the bound, we can modify any given posterior $\bar{P}(\theta \mid y) = S_\theta^{-1}(y)$ by defining a new, *dampened* posterior $\bar{P}^{[\gamma]}(\theta \mid Y) = (S_\theta^{[\gamma]}(Y))^{-1}$ where $S_\theta^{[\gamma]}$ is itself an e-variable defined as $S_\theta^{[\gamma]} = (1-\gamma)+\gamma S_\theta$ for $0\leq\gamma<1$, akin to what we suggested underneath Lemma 1 for the case that $L(0,0)>0$. When giving bounds, for simplicity we will content ourselves with using $\gamma = 1/2$; this will ensure that the posterior can never become larger than 2.

Example 9 [One-Dimensional Exponential Families] Let $\{P_\theta : \theta\in\Theta\}$ be any given regular (Barndorff-Nielsen, 1978) 1-dimensional exponential family given in its mean-value parameterization and extended to n outcomes by independence. We write the KL divergence between two members of the family for a sequence of n outcomes as $D(P^{(n)}_{\theta'}\|P^{(n)}_\theta)$ and we abbreviate $D(P^{(1)}_{\theta'}\|P^{(1)}_\theta)$ to $D(\theta'\|\theta)$. We denote by $\hat{\theta}(y)$ the MLE based on data $y=x^n$, which is 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 is an open set. Suppose we observe $Y=x^n$ with $\hat{\theta}(Y)\in\Theta$. We use as our loss function $L_b(\theta,\breve{\theta}) := b\cdot D(\breve{\theta}\|\theta)$, with $b\in\mathcal{B}=\mathbb{R}_0^+$ — note that in the case of the Gaussian location family, $D(\hat{\theta}\|\theta) = (1/2)(\theta-\hat{\theta})^2$ becomes the squared error loss.

Consider first the e-posterior based on a smooth prior W as above Example 7. We fix some $0\leq\gamma<1$ and let $\bar{P}^{[\gamma]}_{[W]}(\theta \mid y)$ be the resulting dampened e-posterior. In Proposition 2 in the appendix we show that for any $0<\gamma\leq 1$ the MLE $\hat{\theta}$ is the $\bar{P}^{[\gamma]}_{[W]}$-e-posterior minimax estimator irrespective of b. We further, via Proposition 3, show that, for the choice $\gamma=1/2$, the bound (31) holds with

$$\ell_{b,y} = \frac{2b}{n}D(P^{(n)}_{\hat{\theta}}\|P^{(n)}_W) \tag{33}$$

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

$$D(P_{\hat\theta}^{(n)} \| P_W^{(n)}) = \frac{1}{2} \log \frac{n+\lambda}{\lambda} + \frac{n\lambda}{n+\lambda}\hat\theta^2 \text{ so that } \ \ell_b = \frac{2b}{n}\left(\frac{1}{2} \log \frac{n+\lambda}{\lambda} + \frac{n\lambda}{n+\lambda}\hat\theta^2\right)$$

which 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 (41) in the appendix 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 (Grünwald, 2007, Chapter 8)

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

with $I(\hat\theta)$ the Fisher information at $\hat\theta$. For the dampened discrete-prior based e-posterior $\bar{P}^{[\gamma]}_{[n^*,\alpha^*]}$ we have only derived results for the normal location family. For that family, Proposition 2 establishes that again the MLE is e-posterior minimax optimal irrespective of b and γ. In the appendix we show (below Proposition 3) that, for the choice $\gamma = 1/2$, the bound (31) holds if we take, with $c = (n^*/n) \cdot ((\log 2)/(-\log(\alpha^*/2)))$,

$$\ell_{b,y} = \ell_b = \frac{2b}{n} \cdot \frac{1}{\log 2} \cdot \left(c + c^{-1} + 2\right). \tag{34}$$

Of course, if n and b are fixed then for the MLE one can in fact get a better bound. But the bounds (33) and (34) will still hold if n were random, the outcome of some stopping time with unknown definition, and/or if a decision rule was used that may depend on y itself. In particular, returning to Example 3, if we use the e-posterior $\bar{P}_W^{[1/2]}$ or $\bar{P}_{[n^*,\alpha^*]}$, then, upon being presented $B = b$, we assess our uncertainty by stating

$$\frac{L_B(\theta, \hat\theta)}{\ell_{B,Y}} \leq 1$$

with $\ell_{b,y}$ given by (33) or (34), respectively. By (31), these assessments are correct 'on average', i.e. in expectation over Y. The fact that this is possible under arbitrary definitions of B is due to the — perhaps trivial but still — reason that the factor B in L_B and $\ell_{B,Y}$ cancels. In contrast, based on naively using a confidence or objective Bayes posterior, we would assess our uncertainty as in Example 3 by $\mathbf{E}_{\bar{Y} \sim P_{\hat\theta(Y)}}[L_{B(Y)}(\bar{Y}, \hat\theta(Y))]$, which as indicated there is not correct on average.

5 Loose Ends and Final Remarks

Composite H_0 and H_1, e-posteriors with nuisance parameters Whenever in this paper we gave a concrete example of an e-variable, it featured a simple null and alternative. We can still construct useful e-variables for general composite H_0 and H_1; Grünwald et al. (2019) is entirely devoted to developing methods for doing so based in the *Reverse Information Projection* (RIPr). These methods are readily extended to provide anytime-valid CS's for

models with nuisance parameters of the form $\{P_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}$. Here $\theta \in \Theta$ is the parameter (vector) of interest and $\gamma \in \Gamma$ is the nuisance parameter. Via the RIPr one designs a collection of e-processes $\mathcal{S} = \{S_\theta : \theta \in \Theta\}$, one for each $\theta \in \Theta$, each $S_{\theta'}$ being an e-process for the null hypothesis $\mathcal{H}_0 = \{P_{\theta',\gamma} : \gamma \in \Gamma\}$. This is done, for example, by Turner and Grünwald (2022) for the 2×2 contingency table setting. The e-processes S_θ can then be used, as in Section 4, to define an e-posterior by $\bar{P}(\theta \mid Y) := S_\theta^{-1}(Y)$, and the results of e.g. Turner and Grünwald (2022) can readily be applied to the confidence and e-posterior minimax applications of Section 4.1 and 4.2. Importantly, the e-posterior will now *only be defined on the parameters of interest, and not on the nuisance parameters.*

Why set an a priori α at all? We showed that, with e-values, we get valid risk bounds in post-hoc determined decision tasks, irrespective of any pre-set α. This raises the question whether we should set such an α at all. In fact we do not need to — hopefully making our approach acceptable to the many statisticians that are critical of significance testing (McShane et al., 2019). Still, if we want to, we can. For example, as researchers we might set ourselves an initial α for an initial study to be used in deciding whether a much larger and expensive study should even be contemplated (we could then in fact also combine the initial data and the data of the larger study by multiplying the corresponding e-variables (Grünwald et al., 2019)).

How reasonable and relevant is the setting? We aim to provide an extension of NP theory with performance guarantees in an idealized setting that allows for dependency between the data and the loss. Of course, real life decision-making is murkier: loss functions are implicit and ill-defined; models are incorrect, protocols are not strictly followed, and so on — so how reasonable is our idealization? The rationale of our approach is that, with a method that provably works well in the idealized setting, there is at least some hope that it also performs reasonably well in the murkier real world; if a method is not suitable even in idealized settings, we would have no confidence at all of it doing anything reasonable in the real world. The same methodology underlies the original NP theory — the setting of Type I and Type II error control it deals with is an idealization. However, we would argue (following (Fisher, 1955, Edwards, 1984) and many others) that what it formalizes/idealizes is really the setting of industrial quality control rather than that of scientific inference, experimentation and accumulation of knowledge. We aim to formalize the latter instead (Ter Schure and Grünwald, 2021) — and then the BIND assumption of standard NP theory seems unrealistic; some account of dependency and optional continuation is needed, and e-values provide this. In reality the dependencies may not be as strong as in Example 2 and 3 — we adopted extreme cases there merely for illustrative purposes — but they will distort the validity of our conclusions.

Relatedly, thinking in terms of risks rather than error probabilities — as our approach requires — is difficult and practitioners will be tempted to think of e-values simply as 'evidence' or of e-posteriors simply as a notion of 'uncertainty' without directly contemplating risks, losses or actions. But this is perfectly fine: it would still be quite reassuring that they use notions of evidence and uncertainty that, *if* they were operationalized to make statements about actual decisions, would give risk bounds that remain valid without the often unrealistic BIND assumption. For if instead they follow the current practice of using p-values for evidence and confidence intervals for uncertainty while BIND does not hold, then it is simply

not clear what practical implication — and therefore, what real meaning — their statements really have.

Related and Future Work How are e-based decisions related to Bayesian inference? To Martin-Liu's inferential models? To game-theoretic probability (all our results can be interpreted in terms of betting games — see below)? To the likelihood principle (Edwards, 1984) and to Dawid's (1999) prequential principles? All of these comparisons are bound to lead to interesting insights and provide lots of opportunity for future work!

6 Acknowledgements

I am much indebted to the pioneering work by Vovk (1993) (which foreshadows the use of e-variables in a Neyman-Pearson setting) and the many subsequent works by Vovk and Shafer on testing by betting, as culminating in (Shafer, 2021) and the text book (Shafer and Vovk, 2019). In fact, the story told here can peraps be more straightforwardly told in terms of betting — I decided not to do this for the simple reason that the connection between testing and betting is not (yet) widely known.

References

M.S. Balch, R. Martin, and S. 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.

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

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

J.O. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

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.

David R Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958.

D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

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

A. Philip Dawid and Vladimir G. Vovk. Prequential probability: Principles and properties. *Bernoulli*, 5:125–162, 1999.

A.W.F. Edwards. *Likelihood*. Cambridge University Press, 1984.

R.A. Fisher. Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B*, 17:69—-78, 1955.

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, Rianne de Heide, and Wouter Koolen. Safe testing, 2019. arXiv preprint arXiv:1906.07801.

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. doi: https://doi.org/10.1016/j.jspi.2017.09.014.

Alexander Henzi and Johanna F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 2021.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *Annals of Statistics*, 2021.

R. Hubbard. Alphabet soup: Blurring the distinctions between p's and α's in psychological research. *Theory and Psychology*, 14(3):295–327, 2004.

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.

E.L. Lehmann. *Testing Statistical Hypotheses*. Wiley, first edition, 1959.

E.L. Lehmann. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249, 1993.

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

Ryan Martin. Inferential models and the decision-theoretic implications of the validity property. *arXiv preprint arXiv:2112.132472009*, 2021.

Ryan Martin and Chuanhai Liu. *Inferential models: reasoning with uncertainty*. CRC Press, 2015.

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

W. Neiswanger and A. Ramdas. Uncertainty quantification using martingales for misspecified Gaussian processes. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 33, pages 963–982, 2021.

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

S. van der Pas and P. Grünwald. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. *Statistica Sinica*, 28 (1):229–255, 2018.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M. Koolen. Testing exchangeability: fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 2021.

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

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

Judith ter Schure and Peter Grünwald. ALL-IN meta-analysis: breathing life into living systematic reviews. *arXiv preprint arXiv:2109.12141*, 2021.

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

G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, 2021. With Discussion.

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

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

John Shawe-Taylor and Robert C Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, 1997.

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

V.G. Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society, series B*, 55(2):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.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *arXiv preprint arXiv:2009.02824*, 2020.

I. Waudby-Smith and A. Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yanbao Zhang, Scott Glancy, and Emanuel Knill. Asymptotically optimal data analysis for rejecting local realism. *Physical Review A*, 84(6):062118, 2011.

A Details and Proofs

A.1 Details for Section 2.1

Unbounded expected loss based on (5) and an 'improvement' of (5) This issue is best illustrated by (but certainly not limited to) a discrete-valued p-value P that can take values $1, 1/2, 1/4, 1/8, \ldots, 1/2^k$ for some $k > 0$ and that is piecewise strict, i.e. it satisfies $P_0(\mathrm{P} \leq \alpha) = \alpha$ for $\alpha \in \{1, 1/2, \ldots, 1/2^k\}$. Consider a GNP decision task as in Section 3.1 with loss function satisfying $L_b(0, a) = 2a$, for $a \in \mathcal{A}_b = \{0, 1\ell, 2\ell, 4\ell, \ldots, 2^k\ell\}$. Based on (5), upon observing $\mathrm{P} = 2^{-c}$, one would take action 2^c. The resulting expected loss, analogously to (6), is given by $\sum_{c=1}^{k} 2 \cdot 2^{-c} 2^c = 2k$ which goes to ∞ as we make k larger — showing that the expected loss can be unbounded if we base decisions on (5). Now, instead of using (5) it may seem more reasonable to pick the largest a such that

$$\mathrm{Q}(y) \cdot L_b(0, a) \leq \ell, \tag{35}$$

where $\mathrm{Q}(y) = \mathrm{P}(y)/2$: with this modification, for each $a \in \mathcal{A}_b$, we end up multiplying $L_b(0, a)$ in (35) with exactly the probability that a will be selected (rather than, as in (5), with some potentially larger probability. For example, $a = 2^c$ will be selected if $\mathrm{Q}(y) = 2^{-c}$; this happens iff $\mathrm{P}(y) = 2^{-c+1}$, which happens with exactly probability 2^{-c}, so with probability $\mathrm{Q}(y)$). Yet still, using (35) leads to unbounded expected loss: in the above sample the expected loss is now k rather than $2k$, still growing linearly in k.

Example 10 **[Example 2, Details]** To construct a sequence of B's as in Example 2, we now fix some θ^* and some $\epsilon > 0$ and strictly positive function g_0 with $g_0(\epsilon) = \epsilon$ and $\lim_{y \to \infty} g_0(y) = 0$; and we set $g_{\theta^*}(y) := g_0(y - \theta^*)$ for $y = \theta^* + \epsilon = 1$.

Now take any $B(y)$ such that whenever $y \geq \theta^* + \epsilon$, then $B(y)$ is such that $\delta_{B(Y)}(Y)$ has as its left-end $\theta_{\mathrm{L}} = \theta^* + g_{\theta^*}(y)$ and, being symmetric around y, at the right-end $\theta_{\mathrm{R}} = y + (y - (\theta^* + g_{\theta^*}(y)))$: if $y = \theta^* + \epsilon$, the CI will be a single point at $\theta^* + \epsilon$; if y gets larger, the CI widens but no matter how large y, it never covers θ^*. The α_y corresponding to this interval must therefore satisfy $\alpha_y/2 = \int_{-\infty}^{\theta^* + g_{\theta^*}(y)} f_y(u)du$, where we by denote f_μ the density of a normal with variance 1 and mean μ, so $\alpha_y := 2F_y(\theta^* + g_{\theta^*}(y)) = 2F_0(\theta^* - y + g_{\theta^*}(y))$, where F_μ is the CDF of a normal with mean μ and variance 1. It follows that $B(y)$ must be equal to $\ell/\alpha_y = \ell/(2F_0(\theta^* - y + g_{\theta^*}(y)))$

In such a situation, if the data is actually sampled from θ^*, then the expected loss we

actually make can be calculated in steps as follows:

$$\mathbf{E}_{Y\sim P_{\theta^*}}[L_{B(Y)}(\theta^*,\delta(Y))] = \mathbf{E}_{Y\sim P_{\theta^*}}[B(Y)\cdot \mathbf{1}_{\theta^*\notin\delta(Y)}] \geq \mathbf{E}_{Y\sim P_{\theta^*}}[B(Y)\cdot \mathbf{1}_{Y\geq\theta^*+\epsilon}]$$
$$= \int_{\theta^*+\epsilon}^{\infty} f_{\theta^*}(y)\cdot\frac{\ell}{2\cdot F_0(\theta^*-y+g_{\theta^*}(y))}dy = \int_{\epsilon}^{\infty} f_0(y)\cdot\frac{\ell}{2\cdot F_0(g_0(y)-y))}dy$$
$$\geq \frac{\ell}{2}\cdot\int_{\epsilon}^{\infty}\exp\left(-\frac{y^2}{2}\right)\cdot\exp\left(\frac{(y-g_0(y))^2}{2}\right)(y-g_0(y))dy$$
$$\geq \ell\cdot\sqrt{\frac{\pi}{2}}\cdot\int_{\epsilon}^{\infty}\exp\left(-yg_0(y)\right)\cdot(y-g_0(y))dy,$$

where we used the standard result that, with P_0 denoting a standard normal distribution, $P_0(Y\geq c)\leq \exp(-c^2/2)/(c\cdot\sqrt{2\pi})$. Clearly the integral diverges for many choices of g_0 satisfying our requirements; for example, we can take $g_0(y)=\epsilon^2/y$ (which works for all $\epsilon>0$) or (if we want to make the probability of large B smaller) we can set $g_0(y)=\epsilon\cdot(\log(y+\exp(\epsilon)-\epsilon))/y$ if ϵ is set to 2; then $\exp(-yg_0(y))=(y+\exp(2)-2)^{-2}$. Let us take the former choice to see how a typical sample of the B's might look like. Without loss of generality, we take θ^* equal to 0, $\epsilon=0.01$ and $\ell=1$, and sample i.i.d. $B(Y_1), B(Y_2),\ldots$, and set $B=0$ ('decision-problem called off') whenever $y<\epsilon$. We then get (sample generated by R) the sample shown in (13).

Example 11 [Example 3, Details] Consider the Gaussian location family and take $L_b(\theta,a)=b\cdot(\theta-a)^2$ and $n=1$ as in Example 3. Fix some θ^* and set $B(y)=\exp((y-\theta^*)^2/2)g_{\theta^*}(y)$ where g_{θ^*} is some probability density that is symmetric around θ^*. Following Example 3, but now with B instantiated to the above, the loss one *thinks* one makes based on $w^\circ(\theta\mid Y)$, on average in several studies with true parameter θ^*, is given, using that $\hat\theta$ is the mean of $w(\theta\mid Y)$, which is a normal density with variance 1 (since $n=1$), by

$$\mathbf{E}_{Y\sim P_{\theta^*}}[\mathbf{E}_{\bar\theta\sim W|Y}[L_{B(Y)}(\bar\theta,\hat\theta(Y))]] = \mathbf{E}_{Y\sim P_{\theta^*}}\left[e^{(Y-\theta^*)^2/2}g_{\theta*}(Y)\mathbf{E}_{\bar\theta\sim W|Y}(\hat\theta(Y)-\bar\theta)^2\right] =$$
$$\mathbf{E}_{Y\sim P_{\theta^*}}\left[e^{(Y-\theta^*)^2/2}g_{\theta*}(Y)\right] = \frac{1}{\sqrt{2\pi}}\int g_{\theta^*}(y)dy = \frac{1}{\sqrt{2\pi}} \tag{36}$$

whereas the loss one *actually* makes on average is

$$\mathbf{E}_{Y\sim P_{\theta^*}}[L_B(\theta^*,Y)] = \mathbf{E}_{Y\sim P_{\theta^*}}\left[e(Y-\theta^*)^2/2g_{\theta^*}(Y)(\theta^*-Y)^2\right] = \frac{1}{\sqrt{2\pi}}\int g_{\theta^*}(y)(\theta^*-y)^2dy. \tag{37}$$

It is now easy to pick $g_{\theta^*}(y)$ such that the first expression is finite whereas the second is infinite. For example, suppose we take g_{θ^*} to be the distribution of $X-\theta^*$, where X has a Student's t-distribution with 3 degrees of freedom. Then $g^*(y)\asymp y^4$ so (37) will be infinite yet (36) is finite. The list of B's we showed in Example 3 is taken based on this g^*.

A.2 Details for Section 3

Proof of Lemma 1 We first prove another lemma:

Lemma 2 *Suppose that all $P\in\mathcal{H}_0$ have full support $\mathcal{Y}$. Suppose δ is Type-I risk safe and there exists a function $B:\mathcal{Y}\to\mathcal{B}$ and an e-variable S such that S is sharp and δ is B-sharp relative to S. Then δ is a.s. compatible with S, i.e. for all $P\in\mathcal{H}_0$, all $b\in\mathcal{B}$, $P(\delta_b(Y)\leq \ell_b S(Y))=1$.*

Proof: By Proposition 1, there must be some e-variable S' such that δ is compatible with S'. Suppose now that for the $P \in \mathcal{H}_0$ with $\mathbf{E}_P[S] = 1$, we have (a) $P(S \neq S') > 0$. By compatibility, for all $y \in \mathcal{Y}$, for the B above (b) : $S(y) = L_{B(y)}(0, \delta_{B(y)}(y))/\ell_{B(y)} \leq S'(y)$. But (b) implies $P(S \leq S' = 1)$ and (b)+(a) imply $P(S < S') > 0$. But then, by sharpness of S, S' cannot be an e-variable, so we have a contradiction; it follows that $P(S \neq S') = 0$; the result follows. □

Proof: [**of Lemma 1**] For (1): the existence of an $a \in \mathcal{A}_b$ such that $S^{-1}(y)L_b(0, a) \leq \ell_b$ is immediate from requirement (I); the existence of a largest such a by requirement (II). Compatibility is immediate.

For (2): suppose that a decision rule δ is Type-II risk admissible. By definition it is also Type-I risk safe, hence by Proposition 1 it must be compatible with some e-variable S satisfying, for all $b \in \mathcal{B}$, $L(0, 0)/\ell_b \leq \inf_{y \in \mathcal{Y}} S(y)$. But then, by Part 1 of the Lemma, δ^* as in (20) is well-defined and, by definition of compatibility, for all $y \in \mathcal{Y}, b \in \mathcal{B}$ we must have $\delta_b(y) \leq \delta_b^*(y)$ with $\delta_b^*(y)$ as in (20) for that S, so that, for all $P \in \mathcal{H}_0$, for all $b \in \mathcal{B}$, we have $P(\delta_b(Y) \leq \delta_b^*(Y)) = 1$; but then if $P(\delta_b(Y) < \delta_b^*(Y)) > 0$ for some $b \in \mathcal{B}$, δ is not admissible by the fact that $L_b(1, a)$ is strictly increasing in a. It follows that $P(\delta_b = \delta_b^*) = 1$ for all $P \in \mathcal{H}_0$ hence also for $P \in \mathcal{H}_1$.

For (3), Let δ^* be given by (20) and let δ° be another Type-I risk safe decision rule. We will show that δ° cannot be strictly better than δ^*; this implies the result.

Suppose first (Case 3(a)) that for the given B for which δ^* is B-sharp relative to S, δ° is a.s. compatible with S 'on B', i.e.

$$\text{for all } P \in \mathcal{H}_0\text{:} \quad P(L_{B(Y)}(0, \delta^\circ_{B(Y)}(Y)) \leq \ell_{B(y)} \cdot S(y)) = 1.$$

Let $\mathcal{Y}'$ be any set with for all $y \in \mathcal{Y}'$, $L_{B(y)}(0, \delta^\circ_{B(y)}(y)) < \ell_{B(y)} \cdot S(y)$. If there exists such a set with $P(\mathcal{Y}') > 0$ for some $P \in \mathcal{H}_0$, then we also have $P(\mathcal{Y}') > 0$ for $P \in \mathcal{H}_1$ so under the given B, the Type-II risk of δ° is strictly larger (since $L_b(1, a)$ is strictly decreasing in a) than that of δ^*, so δ° is not strictly better than δ^*. Hence, for δ° to be strictly better than δ^*, we must have $P(\mathcal{Y}') = 0$ for all sets $\mathcal{Y}'$ as above, all $P \in \mathcal{H}_0$. But then by Lemma 2 above we have that δ° is a.s. compatible with S. By definition of δ^* we then have that for all $P \in \mathcal{H}_0 \cup \mathcal{H}_1$, all $b \in \mathcal{B}$: $P(\delta_b^\circ(Y) > \delta_b^*(Y)) = 0$, and hence δ^* is not strictly better than δ° in Case 3(a).

Now consider the alternative Case 3(b) that for some (hence all) $P \in \mathcal{H}_0$,

$$P(L_{B(Y)}(0, \delta^\circ_{B(Y)}(Y)) > \ell_{B(Y)} \cdot S(Y)) > 0. \tag{38}$$

In this case we must further have

$$P(L_{B(Y)}(0, \delta^\circ_{B(Y)}(Y)) \geq \ell_{B(Y)} \cdot S(Y)) < 1, \tag{39}$$

for suppose (39) does not hold, i.e. the probability is 1. We then have by sharpness of S that for some $P \in \mathcal{H}_0$, $\mathbf{E}_P[L_{B(Y)}(0, \delta^\circ_{B(Y)}(Y))/\ell_{B(Y)}] > \mathbf{E}_P[S(Y)] = 1$, violating the assumed Type-I risk safety of δ°.

(38) gives that there must be a set $\mathcal{Y}'$ with $P(\mathcal{Y}') > 0$ (and hence $P_1(\mathcal{Y}') > 0$) such that for all $y \in \mathcal{Y}'$,

$$L_{B(y)}(0, \delta^\circ_{B(y)}(y)) < \ell_{B(y)} \cdot S(y) = L_{B(y)}(0, \delta^*_{B(y)}(y)).$$

Now set $B'(y) := B(y)$ for all $y \in \mathcal{Y}'$ en $B'(y) :=$ TRIV for all $y \in \mathcal{Y} \setminus \mathcal{Y}'$. Since $L_b(0,a)$ is increasing in a and $L_b(1,a)$ is strictly decreasing in a, it follows that $\mathbf{E}_{P_1}[L_{B'}(1, \delta^\circ_{B'}(Y))] > \mathbf{E}_{P_1}[L_{B'}(1, \delta^*_{B'}(Y))]$ so δ° is not strictly better than δ^*. $\square$

Proofs for Section 4

Technical Preliminaries Consider a regular 1-dimensional exponential family $\{P_\theta : \theta \in \Theta\}$ given in its mean-value parameterization, as in the main text. We repeatedly use two results. The first (easily proved using steepness of regular exponential families (Barndorff-Nielsen, 1978)) is that for each fixed $\theta' \in \Theta$, $D(\theta'\|\theta)$ is a continuous function of $\theta \in \Theta$ satisfying

$$\sup_{\theta<\theta'} D(\theta'\|\theta) = \sup_{\theta>\theta'} D(\theta'\|\theta) = \infty. \tag{40}$$

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:

$$\frac{p_\theta(y)}{p_W(y)} = \exp(-nD(\hat{\theta}\|\theta) + D(P^{(n)}_{\hat{\theta}}\|P^{(n)}_W)). \tag{41}$$

The Extension of $\bar{P}_{[n^*,\alpha^*]}$ to General 1-d Exponential Families Fix anticipated n^* and α^*. By (40) above, for general regular 1-dimensional exponential families, there exist $\theta^- < \theta < \theta^+$ such that

$$n^* D(\theta\|\theta^+) = n^* D(\theta\|\theta^-) = -\log \alpha^*/2. \tag{42}$$

Now define the e-variable $S^{\mathrm{L}}_\theta(y) = \frac{p_{\theta^-}(y)}{p_\theta(y)}$ and $S^{\mathrm{R}}_\theta(y) = \frac{p_{\theta^+}(y)}{p_\theta(y)}$ and $S_\theta := (1/2)S^{\mathrm{L}}_\theta + (1/2)S^{\mathrm{R}}_\theta$. S^{L}_θ and S^{R}_θ coincide with the *uniformly most powerful Bayes factors* for a 1-sided test at sample size n^* and level $\alpha^*/2$ of $\mathcal{H}_0 = \{P_\theta\}$ vs. $\mathcal{H}_1 = \{P_{\theta'} : \theta > \theta\}$ and $\mathcal{H}_1 = \{P_{\theta'} : \theta' < \theta\}$ respectively (Johnson, 2013). Therefore we propose to define $\bar{P}_{n^*,\alpha^*}(\theta \mid y) := S_\theta(y)$ as a default 'discrete' (putting all its mass on θ^+ and θ^-) e-posterior for general exponential families. In the case of the Gaussian location family, $D(\theta'\|\theta) = (1/2)(\theta'-\theta)^2$ so the definition coincides with that in Example 6.

We next provide a bound specific to the normal location family case. Note that $\bar{P}_{[n^*,\alpha^*]}(\theta \mid y)$ can be written by (41) with W a prior putting mass 1/2 on θ^+ and 1/2 on θ^-. For the Gaussian location family, (41) then gives:

$$\begin{aligned}
\bar{P}_{[n^*,\alpha^*]}(\theta \mid y) &= \frac{p_\theta(y)}{\frac{1}{2}p_{\theta^-(y)} + \frac{1}{2}p_{\theta^+(y)}} = \frac{2e^{-nD(\hat{\theta}\|\theta)}}{e^{-nD(\hat{\theta}\|\theta^-)} + e^{-nD(\hat{\theta}\|\theta^+)}} \\
&= \frac{2e^{-(n/2)(\hat{\theta}-\theta)^2}}{e^{-(n/2)\cdot(\hat{\theta}-\theta+U)^2} + e^{-(n/2)(\hat{\theta}-\theta-U)^2}} = \frac{2e^{nU^2/2}}{e^{-n\cdot(\hat{\theta}-\theta)U} + e^{n\cdot(\hat{\theta}-\theta)U}} \\
&\leq 2e^{nU^2/2 - n|\hat{\theta}-\theta|U}
\end{aligned} \tag{43}$$

where $U = \sqrt{2(-\log(\alpha^*/2))/n^*}$. Using the final equation of (43) we see that a sufficient condition for $\bar{P}_{[n^*,\alpha^*]}(\theta \mid y) \leq \alpha$ is

$$n|\hat{\theta} - \theta|U \geq -\log(\alpha/2) + nU^2/2. \tag{44}$$

Straightforward rewriting shows that this is equivalent to (27).

Establishing that the MLE is e-posterior minimax, and the corresponding bounds Our results on the MLE being e-posterior minimax optimal rely on the following proposition:

Proposition 2 *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 $\lim_{x\to\infty} xf(x) = 0$. 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 $\bar{P}^{[\gamma]}_{[W]}(\theta \mid y)$ for any $0 < \gamma \leq 1$ and 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 dampened e-posterior $\bar{P}^{[\gamma]}_{[n^*,\alpha^*]}(\theta \mid y)$ 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 (40) 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 (41). For $\gamma < 1$, use the fact that $1/((1-\gamma)x^{-1} + \gamma)$ is increasing in x.

As to (2): using (43), and again 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, for $\gamma = 1$,

$$\bar{P}_{[n^*,\alpha^*]}(\theta \mid y) = 2e^{nA^2/2} \cdot \frac{1}{e^{-n\cdot\sqrt{2D(\hat{\theta}\|\theta)}} + e^{n\cdot\sqrt{2D(\hat{\theta}\|\theta)}}} = f(D(\hat{\theta}\|\theta)),$$

whhere f is of the required form. The result for $\gamma < 1$ then follows as above, using that $1/((1-\gamma)x^{-1} + \gamma)$ is increasing in x. □

We next evaluate the bounds in Example 9. We start with a proposition that may be used beyond the use of KL divergence as loss:

Proposition 3 *Let $\Theta \subset \mathbb{R}$, $b \in \mathbb{R}^+$ and let $L_b : \Theta \times \Theta \to \mathbb{R}_0^+$ be a loss function with $L_b(\theta, \theta') = bL_1(\theta, \theta')$ 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 $\bar{P}'(\theta \mid y)L_1(\theta, \breve{\theta})$ is decreasing in θ.*
2. *for all $\theta \leq \theta_L$, $\bar{P}'(\theta \mid y) \leq 1$ and $\bar{P}'(\theta \mid y)L_1(\theta, \breve{\theta})$ is increasing in θ.*

Then (30) holds for $\bar{P}(\theta \mid y)$ with

$$\ell_{b,y} = b \cdot \max\{C_{\sup}, \bar{P}_{\sup}\} \cdot \max_{\theta \in [\theta_L, \theta_R]} L_1(\theta, \breve{\theta}). \tag{45}$$

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

We now use Proposition 3 to show the bounds (33) and (34) of Example 9. Assume the setting of that example. We set $\breve{\theta}$ to the MLE and $L(\theta, \breve{\theta}) := D(\breve{\theta} \| \theta)$. We first apply Proposition 3 with $\bar{P}'(\theta|y) := \bar{P}(\theta|y)$ (and $C_{\sup} = 1$) set to the dampened e-posterior $p_W^{[1/2]}$ for a smooth prior W (independent of θ) to show (33). 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 $p_W^{[1/2]}(\theta_L \mid y) = p_W^{[1/2]}(\theta_R \mid y) = 1$. These must exist (use (40)) and by (41) they satisfy

$$D(\hat{\theta} \| \theta_L) = D(\hat{\theta} \| \theta_R) = \frac{D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)})}{n}.$$

To verify the conditions of Proposition 3, we will show, using (41), that

$$D(\hat{\theta} \| \theta) \exp(-n D(\hat{\theta} \| \theta) + D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)})) \tag{46}$$

is increasing for $\theta < \theta_L$ and decreasing for $\theta > \theta_R$. For this, setting $C = D(P_{\hat{\theta}}^{(n)} \| P_W^{(n)})$, it is sufficient to show that $g(u) := u \exp(-nu + C)$ 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 3 can be applied if $C \geq 1$ and then (45) gives (33).

We next apply Proposition 3 to the dampened discrete e-posterior $\bar{P} := \bar{P}_{[n^*, \alpha^*]}^{[1/2]}$ with the Gaussian location family to show (34). Again, by the dampening, $\bar{P}_{\sup} \leq 2$. We will use Proposition 3 with $\bar{P}'(\theta|y) = 2 \exp(nU^2/2 - n|\hat{\theta} - \theta|U)$ with U as in (43) ; $\bar{P}'$ can be seen to be an upper bound of $\bar{P}_{[n^*, \alpha^*]}$ by (43) and hence, since trivially $\bar{P}_{[n^*, \alpha^*]}^{[1/2]} \leq 2 \bar{P}_{[n^*, \alpha^*]}$, we have $\bar{P}' \leq C_{\sup} \bar{P}_{[n^*, \alpha^*]}^{[1/2]}$ with $C_{\sup} = 2$. Using (44) and (27) with $\alpha = 1$, we find that with $\theta_L' = \hat{\theta} - V'$, $\theta_R' = \hat{\theta} + V'$ and $c = (n^*/n) \cdot ((\log 2)/(-\log(\alpha^*/2)))$ and

$$V' = \sqrt{\frac{\log 2}{2n}} \cdot \left(c^{1/2} + c^{-1/2} \right)$$

we are guaranteed that $\bar{P}'(\theta \mid y) \leq 1$ for $\theta \leq \theta_L$ and $\theta \geq \theta_R$. Further, simple differentiation shows that $\bar{P}'(\theta \mid y)(\theta - \hat{\theta})^2$ is increasing at $\theta < \theta_L''$ and decreasing at $\theta > \theta_R''$ where $\theta_L'' = \hat{\theta} - V''$ and $\theta_R'' = \hat{\theta} + V''$ with

$$V'' = \frac{2}{nU} = \sqrt{\frac{2}{n}} \cdot \sqrt{\frac{n^*}{n(-\log(\alpha^*/2))}} = \sqrt{\frac{2}{(\log 2) \cdot n}} \cdot c^{1/2}.$$

Combining these two displays, we find that the conditions of Proposition 3 hold if we set $\theta_L = \hat{\theta} - V$, $\theta_R = \hat{\theta} + V$, $V = \sqrt{\frac{2}{(\log 2) \cdot n}} \cdot \left(c^{1/2} + c^{-1/2} \right)$. Using $\sup_{\theta \in [\theta_L, \theta_R]} D(\hat{\theta} \| \theta) = V^2/2$ and $\max\{C_{\sup}, p_{\sup}\} = 2$ in (45) now gives (34).