Regularized e-processes: anytime valid inference with knowledge-based efficiency gains

Ryan Martin[*]

November 4, 2024

Abstract

Classical statistical methods have theoretical justification when the sample size is predetermined. In applications, however, it's often the case that sample sizes aren't predetermined; instead, they're often data-dependent. Since those methods designed for static sample sizes aren't reliable when sample sizes are dynamic, there's been recent interest in *e-processes* and corresponding tests and confidence sets that are *anytime valid* in the sense that their justification holds up for arbitrary dynamic data-collection plans. But if the investigator has relevant-yet-incomplete prior information about the quantity of interest, then there's an opportunity for efficiency gain, but existing approaches can't accommodate this. The present paper offer a new, *regularized e-process* framework that features a knowledge-based, imprecise-probabilistic regularization with improved efficiency. A generalized version of Ville's inequality is established, ensuring that inference based on the regularized e-process remains anytime valid in a novel, knowledge-dependent sense. In addition, the proposed regularized e-processes facilitate possibility-theoretic uncertainty quantification with strong frequentist-like calibration properties and other desirable Bayesian-like features: satisfies the likelihood principle, avoids sure-loss, and offers formal decision-making with reliability guarantees.

Keywords and phrases: Credal set; decision-making; e-value; imprecise probability; inferential model; possibility theory; safety; uncertainty quantification.

1 Introduction

Most statistical and machine learning methods in the literature have theoretical justification that assumes the sample size is (or can safely be treated as) fixed, say, by the experimental or data-collection protocol. But in many applications this assumption is violated—e.g., decisions to stop/continue experimentation are made dynamically while monitoring the data—yet the same fixed-sample-size methods are used for data analysis. Use of methods when they lack theoretical justification raises doubts about their reliability. This lack of reliability is surely at least a contributor to the widely-publicized replication crisis in science (e.g., Baker 2016; Camerer et al. 2018; Nuzzo 2014).

[*]Email: `rgmarti3@ncsu.edu`

To address this concern, there's been a recent surge of effort to develop so-called *safe* or *anytime valid* statistical methods, i.e., methods that can preserve their reliability even when applied in cases where the data-collection process is dynamic; see the recent survey by Ramdas et al. (2023) and Section 2.2 below. The centerpiece of these new developments is a mathematical object called an *e-value*, *e-variable*, or *e-process*, and I'll focus throughout on the latter. What's important is that an e-process enjoys certain martingale-like properties, which implies a time-uniform probability bound, namely, *Ville's inequality*, that can be leveraged to construct test and confidence procedures with error rate control guarantees that hold no matter how the data-collection process is stopped. There are some guiding principles behind the construction of an e-process, and numerous different instantiations are now available in the literature (Section 2.2). Different anytime valid methods are typically compared in terms of their efficiency, i.e., power of tests, size of confidence intervals, etc. In terms of the e-process itself (rather than in terms of the methods derived from it), efficiency corresponds to how quickly it grows when evaluated at the "wrong" hypotheses: fast growth of the e-process is desirable because it means that what's "wrong" can be detected sooner, hence practitioners can make safe and reliable decisions with less time/exposure and fewer resources.

But e-processes have limits to how fast they can grow (e.g., Grünwald et al. 2024). What can be done to improve efficiency when there are no more e-process tweaks to be made? It's now statistical second-nature to leverage penalty functions or prior distributions for *regularization*, to encourage certain structure in estimates, usually for the sake of efficiency gain. Regularization is relatively easy when the goal is asymptotic consistency or calibration, since many different regularization strategies work in that sense. But anytime validity is a finite-sample property, so virtually any tweak made to an e-process would jeopardize the anytime validity property that motivated its use in the first place.

This paper considers the situation in which the investigator *knows something* about Θ, the relevant quantity of interest, before the data are in hand and ready to be analyzed. To keep the names and roles clear, "I" refers to me, this paper's author, and "You" refers to the investigator, a catch-all term for the individual(s) carrying out a study, analyzing data, etc. The seemingly simple phrase *knows something* is actually quite nuanced. On the one hand, I mean *know* in the strong sense of absolute certainty whereas, on the other hand, the *something* that's known could be so substantial that it renders data irrelevant, so miniscule that it means practically nothing, or, as is typical, somewhere in the middle of these two extremes. (To be clear, ignorance is a special type of knowledge, and it's covered by what I'll present herein.) According to Levi (1980), Your *corpus of knowledge* consists of the (lower and upper) probabilities that You'd assign to hypotheses about the uncertain Θ based on the information available to You at the relevant moment in time. The point I want to emphasize here is that Your corpus of knowledge is what it is at the time You call on it; without new information that warrants revising Your corpus, there's no justification for adding to or subtracting from it.

Modern data science research focuses on automation: developing tools that You can directly apply to reveal insights in Your data. But these automated methods can't accommodate what You know *a priori*, so they implicitly expect You to subtract from Your corpus of knowledge. Some might argue that this subtraction is warranted because Your subjective judgments have no place in scientific investigations. At the end of the day, however, You'll inevitably apply what You *know* to Your problem in some way, and rightfully so. Without statistical guidance on how to properly express and incorporate this knowledge, it may be done improperly, thereby creating systematic risk of errors, or at least doubts about reliability. This paper aims to offer such guidance.

The previous paragraphs betray my subjectivist perspective, but my proposal is different from what's been done before. The typical criticism of subjectivism is that it's focused more on what's "right" in principle than what "works" in practice.[1] That is, subjective Bayes has strong justification based on core principles, but practical obstacles often get in the way of achieving this principled ideal. By abandoning those principles, however, practitioners can often find simpler methods that give reasonable answers and can be justified based on their effectiveness. The apparent gap between what's "right" and what "works" is troubling, so I'm working to close it. My proposal first converts Your corpus of knowledge into what I'm calling a *regularizer*, and then combines this regularizer with Your favorite e-process to get a new knowledge-aware e-process for reliable and efficient inference and uncertainty quantification.

As I mentioned above, ignorance is a special type of knowledge—one where Your corpus indicates that You're unsure about everything—so my proposal isn't strictly for subjectivists. In fact, the results presented herein are general and cover both the ignorant and non-ignorant cases alike. To be crystal clear, I'm not asking You to concoct a non-trivial corpus of knowledge before You can follow my proposal. Instead, I'm offering an opportunity to use what You already know—which could be nothing or something—to improve upon well-established statistical methods.

The key insight is leveraging Your corpus of knowledge in two different ways. The first is *regularization*, which amounts to directly manipulating the data-driven e-process using what You know. Unless You're ignorant, Your corpus of knowledge casts a degree of doubt on some values θ of Θ. The regularizer appropriately discounts those θ values, thereby boosting the regularized e-process at those values. A bigger e-process means more θ values can be excluded from consideration, hence more efficiency. The second is introducing a generalized, *regularization-aware* notion of anytime validity with respect to which the regularized e-process is evaluated. Since You can't doubt what You know, You'd be willing to evaluate the performance of Your regularized e-process based on a metric that depends on what You know. Roughly, my proposal requires that the regularized e-process be anytime valid with respect to any (data, Θ)-joint distribution compatible with both the data-generating model and Your corpus of knowledge., the latter being expressed in terms of a credal set of "priors" for Θ. The standard definition of anytime validity corresponds to the special—and most restrictive—case where the aforementioned credal set contains all priors for Θ. By moving away from this most-restrictive case, regularization allows for efficiency gains without ruining anytime validity.

Then the chief technical development in the paper is the establishment of a generalized version of Ville's inequality for my proposed regularized e-process, one that depends in a certain way on Your corpus of knowledge. This requires considerable build-up, but the payoff is commensurate: provably reliable, non-asymptotic inference with (or without) regularization that features opportunities for improved efficiency. Beyond construction of test and confidence procedures, a full-blown framework for (possibility-theoretic) uncertainty quantification, which facilitates both calibrated, data-dependent (imprecise)

[1] For example, Speed (2016) writes: "If a statistical analysis is clearly shown to be effective [...], it gains nothing from being described as principled."

probabilistic reasoning about Θ, and reliable, data-driven risk assessments and decisions.

Is this knowledge-based, regularization-aware notion of anytime validity satisfactory justification for the use of Your more-efficient, regularized e-process? That it's satisfactory to You is automatic—it's based on *Your* corpus of knowledge. Whether the justification is satisfactory to others depends on if Your corpus of knowledge can withstand their scrutiny. This is a matter for epistemologists, pertaining to questions about justified credence (e.g., Dunn 2015; Pettigrew 2021; Tang 2016), so here is not the place to pursue this. My point is that harnessing the full power of data science is a collaborative effort between statisticians and subject matter experts, and my proposed knowledge-based regularized e-process framework aims to exemplify this relationship. Statisticians provide the best-available e-process and other technical know-how, subject matter experts provide the application-specific "justified credence," and together they (mathematically and figuratively) lift each other up to contribute more than each could individually.

The remainder of this paper is organized as follows. Following some background in Section 2 on e-processes and imprecise probability, Section 3 dives into my specific proposal to incorporate Your corpus of knowledge (encoded as an imprecise probability) into the data-driven, e-process-based analysis via regularization. After a concrete regularizer construction and some technical details on credal sets, Sections 3.5–3.6 present the imprecise-probabilistic generalization of Ville's inequality and the statistical implications, respectively. Some simple illustrations of the efficiency gains that can be realized are presented in Section 3.7. Focus shifts in Section 4 to broader uncertainty quantification about Θ, superficially similar to Bayesian inference. What I'm proposing here is a brand of *inferential model* (IM), where the familiar Bayesian probabilistic reasoning is replaced by a provably reliable form of possibilistic reasoning; basically, the difference is only in the calculus. My proposed e-possibilistic IM framework inherits the exact, anytime validity from the aforementioned regularized e-process, which implies that my uncertainty quantification is calibrated in a strong sense that implies reliability. Remarkably, in addition to these frequentist-like calibration properties, the e-possibilistic IM construction leads to a number of other desirable Bayesian-like properties. This includes: satisfying the likelihood principle, avoiding a Dutch book sure-loss, and von Neumann–Morganstern-style decision-making that features strong reliability guarantees. An illustration of the proposed framework, in the context of a simple clinical trial study, is presented in Section 5, where the main goal is to illustrate how background knowledge about the subject matter gets translated first into a partial prior and then into a regularizer. Some possible extensions of the proposed framework are mentioned in Section 6 and concluding remarks are made in Section 7. The appendix contains some additional technical details.

2 Background

2.1 Setup and notation

Start with a baseline probability space $(\mathbb{S}, \mathcal{A}, \mathsf{P})$, where $\mathcal{A}$ is a σ-algebra of subsets of $\mathbb{S}$ and P is a probability measure. Let $Z : \mathbb{S} \to \mathbb{Z}$ be a measurable function that takes values in a topological space $\mathbb{Z}$. The goal is to reliably answer certain questions about the uncertain P, or relevant features thereof, based on one or more data point Z. I'll write P for both the probability on $(\mathbb{S}, \mathcal{A})$ and the induced distribution of Z.

Next, let $Z_1, Z_2, \ldots$ denote independent and identically distributed (iid) copies of Z and, for each $n \geq 1$, write $Z^n = (Z_1, \ldots, Z_n)$. The iid assumption isn't necessary, but the advantage is that the joint distribution of $Z^\infty = (Z_1, Z_2, \ldots)$ is fully determined by the same P introduced above. Define the filtration $\mathcal{A}_n = \sigma(Z^n)$ to be the sequence of σ-algebras determined by the information accumulated after observation of Z^n. A stopping time N is an integer-valued random variable with the property that, for any integer n, the event $\{N = n\}$ is contained in $\mathcal{A}_n$. Intuitively, this means that the stopping time's value is determined solely by the process's history, not on aspects of the future.

For the statistical applications that I have in mind, P is *uncertain*. I'll be more specific about what "uncertain" means in Section 2.3; here I'm primarily focused on notation. To facilitate this notion of an "uncertain P," it'll help to introduce (the notation of) a model, namely, $\mathscr{P} = \{\mathsf{P}_\omega : \omega \in \mathbb{O}\}$, indexed by $\mathbb{O}$. Of course, this could be a familiar parametric model, but it could also be that $\mathbb{O}$ is in one-to-one correspondence with the set of all relevant probability distributions, so there's no loss of generality in my introducing the index ω. The advantage is that it's easier to describe mathematically an "uncertain P" by a more familiar type of uncertain variable Ω in the space $\mathbb{O}$. In any case, again, the transition from "uncertain P" to "uncertain Ω" is without loss of generality.

The goal then is to learn about the uncertain P or, equivalently, the uncertain Ω, based on observations $Z_1, Z_2, \ldots$ from the underlying process that depends on P or Ω. It's often the case that only certain features of the uncertain (P or) Ω are relevant to the application at hand—e.g., maybe one only needs to know about the mean survival time for patients on a particular treatment—so I'll define this relevant feature as $\Theta = f(\Omega)$, taking values in the possibility space $\mathbb{T} = f(\mathbb{O})$. To be clear, this feature-extracting function $f : \mathbb{O} \to \mathbb{T}$ is fixed by the context of the problem; it's only the input Ω that's uncertain, which makes the output Θ also uncertain. Since f could be the identity function, in which case $\Theta = \Omega$, this focus on learning about Θ based on data $Z_1, Z_2, \ldots$ is without loss of generality.

One relatively mild technical condition that I'll impose throughout is that

the function f is continuous.

I don't believe that continuity is necessary, but this greatly simplifies some of the developments in Section 3.4; there's also a relatively naive ways to bypass a continuity assumption, but this leads to some efficiency loss so I won't discuss it here. Besides, continuity is not a major practical restriction. For cases in which there is an underlying parametric model in consideration, the relevant quantity would either be the parameter itself, or some easily-interpretable feature thereof, e.g., one component of a parameter vector. In virtually all such cases, the function that determines this feature would be continuous. When no parametric model is assumed, so that the index $\Omega \in \mathbb{O}$ has no direct practical interpretation, the kinds of relevant features $\Theta = f(\Omega)$ are often solutions to certain optimization problems, e.g., risk minimizers. Most machine learning applications are of this type. Let $L_t : \mathbb{Z} \to \mathbb{R}$ be a practically meaningful loss function indexed by $t \in \mathbb{T}$, e.g., squared-error loss, and define the mapping $(\omega, t) \mapsto \mathsf{P}_\omega L_t$,the expected loss corresponding to t at ω. If this mapping is continuous, then Berge's *maximum theorem* (e.g., Aliprantis and Border 2006, Theorem 17.31) implies that the expected loss minimizer $f(\omega) := \arg\min_t \mathsf{P}_\omega L_t$ is continuous.

2.2 E-processes

Start by fixing a particular $\omega \in \mathbb{O}$. A sequence $(M^n : n \geq 0)$ is a *supermartingale*, relative to P_ω and the filtration $(\mathcal{A}_n)$, if $\mathsf{E}_\omega(M^n \mid \mathcal{A}_{n-1}) \leq M^{n-1}$, where E_ω denotes expected value with respect to P_ω. What I have in mind is that there's a function $M(\cdot)$ that maps $\mathbb{Z}$-valued sequences to numbers, and $M^n = M(Z^n)$ for each n, with $M^0 = M(\square)$ the value M takes when applied to the empty sequence $\square$. An ω-dependent supermartingale $(M_\omega^n : n \geq 0)$ is a *test supermartingale* for ω if it's non-negative and $M_\omega^0 \equiv 1$ (e.g., Shafer et al. 2011). A sequence $(\mathfrak{e}_\omega^n : n \geq 0)$ is an *e-process* for ω if it's non-negative and upper bounded by a test supermartingale for ω; see Ramdas et al. (2023) for a detailed survey. Again, $(\mathfrak{e}_\omega^n)$ is determined by a mapping $\mathfrak{e}_\omega(\cdot)$ applied to observations: $\mathfrak{e}_\omega^n = \mathfrak{e}_\omega(Z^n)$. Test supermartingales and, hence, e-processes satisfy two key properties: the first (e.g., Durrett 2010, Theorem 5.7.6) is

$$\mathsf{E}_\omega(M_\omega^N) \leq 1 \quad \text{all } \omega \in \mathbb{O}, \text{ all stopping times } N,$$

and the second, known as *Ville's inequality* (e.g., Shafer and Vovk 2019), is

$$\mathsf{P}_\omega(M_\omega^N \geq \alpha^{-1}) \leq \alpha \quad \text{all } \alpha \in (0,1], \text{ all stopping times } N.$$

Results of this type have important statistical implications (e.g., Shafer 2021). One is that, if $(\mathfrak{e}_\omega^n)$ is an e-process for ω, then the test that rejects the hypothesis "$\Omega = \omega$" based on data Z^n if and only if $\mathfrak{e}_\omega(Z^n) \geq \alpha^{-1}$ controls the frequentist Type I error at level α, regardless of what stopping rule might be used to terminate the data-collection process. This *anytime validity*, also termed *safety* in Grünwald et al. (2024), of the e-process-based tests is a major advancement beyond the classical tests that are valid only for predetermined n. Of course, if one has a collection of e-processes $(\mathfrak{e}_\omega^n)$, one for each $\omega \in \mathbb{O}$, then the above testing procedure can be inverted to construct a confidence set for Ω which inherits the same anytime validity property: the frequentist coverage probability is no less than the nominal level independent of the choice of stopping rule.

A practically important extension of the ideas presented above is to the case of composite hypotheses. For the applications I have in mind here, even the simple hypotheses are composite, since fixing the value θ of Θ would, in general, correspond to a set of ω values, i.e., $f^{-1}(\{\theta\}) = \{\omega \in \mathbb{O} : f(\omega) = \theta\}$. Fortunately, it's at least conceptually straightforward to accommodate this more general case: if an e-process $(\mathfrak{e}_\omega^n)$ is available for each $\omega \in \mathbb{O}$, then, with a minor abuse of notation, the following is an e-process for θ,

$$\mathfrak{e}_\theta^n = \inf_{\omega \in \mathbb{O}: f(\omega) = \theta} \mathfrak{e}_\omega^n, \quad \theta \in \mathbb{T}.$$

It follows immediately from the properties discussed above that

$$\sup_{\omega : f(\omega) = \theta} \mathsf{E}_\omega\{\mathfrak{e}_\theta(Z^N)\} \leq 1 \quad \text{all } \theta \in \mathbb{T}, \text{ all stopping times } N. \tag{1}$$

There's also a corresponding version of Ville's inequality, from which the anytime validity of the corresponding tests and confidence sets for Θ are derived:

$$\sup_{\omega \in \mathbb{O} : f(\omega) = \theta} \mathsf{P}_\omega\bigl\{\mathfrak{e}_\theta(Z^N) \geq \alpha^{-1}\bigr\} \leq \alpha \quad \text{all } \alpha \in (0,1], \text{ all stopping times } N. \tag{2}$$

My examples below are relatively simple, so I can focus on Savage–Dickey e-processes, i.e., Bayes factors that rely on "default" priors. This approach is quite general but difficulties can arise in more complex problems. Ramdas et al. (2023) give an extensive review of the various e-process constructions, but one I'd like to mention specifically is *universal inference*, first put forward in Wasserman et al. (2020). Universal e-processes typically correspond to likelihood ratios where one of the likelihoods is "cautiously maximized," following a general idea goes back at least to Wald (1947, Eq. 10.10). For recent applications to nonparametric problems, see Gangrade et al. (2023) and Dixit and Martin (2024). A variation for when there's no likelihood is proposed in Dey et al. (2024).

2.3 Imprecise probability

Certain aspects of imprecise probability theory will play a key role in this paper, so here I introduce the essential notation and concepts. The goal, recall, is to quantify uncertainty about Θ, to give a quantitative expression of Your corpus of knowledge.

How might You quantify uncertainty about Θ? The most familiar approach would introduce a probability, say, Q, supported on subsets of $\mathbb{T}$ and, then, for any relevant hypothesis "$\Theta \in H$," where $H \subseteq \mathbb{T}$, Your uncertainty about the truthfulness of this hypothesis is quantified via the credence $\mathsf{Q}(H)$, the Q-probability of H; henceforth I'll refer to both H and "$\Theta \in H$" as hypotheses about Θ. But it's easy to imagine situations in which You have limited information and, hence, can't precisely pin down the probability of H for each hypothesis H. The extreme case of *vacuous* (prior) information, i.e., You literally know nothing about Θ, is covered by the setup here. All You can say in this extreme case is that $\mathsf{Q}(H) \in [0,1]$ for all $H \not\in \{\varnothing, \mathbb{T}\}$, which is satisfied by all probabilities Q, so no single Q can be pinned down. More generally, You might be in between the two extremes of knowing nothing and knowing Q precisely. For example, suppose "You're 90% sure that Θ is bigger than 7." This information imposes various bounds on $\mathsf{Q}(H)$ depending on H, e.g., $\mathsf{Q}(H) \leq 0.1$ for all $H \subseteq (-\infty, 7]$, which are satisfied by many different probabilities Q, so this information fails to pin down a single Q. That the available information fails to pin down a single Q isn't grounds for You to ignore it.

When the information available about Θ is sufficient to pin down a single Q, then Your uncertainty quantification is *precise*; otherwise, it's *imprecise*. Mathematically, these can be described by corresponding *precise* and *imprecise probabilities*, respectively. Don't be mislead by the everyday use of the words "precise" and "imprecise"—precise probability isn't superior to imprecise probability. Your available information about Θ is what it is, so artificially embellishing on that information to avoid addressing the inherent imprecision doesn't make Your uncertainty quantification better. The goal is to represent Your corpus of knowledge as faithfully as possible; if that requires imprecision, then so be it.

One fairly general way to define imprecise probabilities is via a *credal set* $\mathscr{Q}$, i.e., a non-empty, closed, and convex collection of probabilities Q on $\mathbb{T}$ (e.g., Levi 1980, Sec. 4.2). The intuition is that $\mathscr{Q}$ encodes Your knowledge about Θ, so the collection $\mathscr{Q}(H) = \{\mathsf{Q}(H) : \mathsf{Q} \in \mathscr{Q}\}$ of credences quantifies uncertainty about the truthfulness of H. Natural summaries of this collection include the lower and upper bounds:

$$\underline{\mathsf{Q}}(H) = \inf\{\mathsf{Q}(H) : \mathsf{Q} \in \mathscr{Q}\} \quad \text{and} \quad \overline{\mathsf{Q}}(H) = \sup\{\mathsf{Q}(H) : \mathsf{Q} \in \mathscr{Q}\}. \tag{3}$$

Thanks to the structure in $\mathscr{Q}$, the lower and upper probabilities are linked together through a relationship called *conjugacy*, i.e., $\underline{\mathsf{Q}}(H) = 1 - \overline{\mathsf{Q}}(H^c)$ for all H. In what follows, I'll focus almost exclusively on upper probabilities. Following de Finetti, it is common to interpret $(\underline{\mathsf{Q}}, \overline{\mathsf{Q}})$ in terms of buying/selling prices for gambles:

$$\begin{aligned} \underline{\mathsf{Q}}(H) &= \text{Your supremum buying price for } \$1(\Theta \in H) \\ \overline{\mathsf{Q}}(H) &= \text{Your infimum selling price for } \$1(\Theta \in H). \end{aligned} \tag{4}$$

That is, You'd be willing to buy a ticket that pays \$1 if "$\Theta \in H$" from me for no more than $\$\underline{\mathsf{Q}}(H)$ and, similarly, You'd be willing to sell a ticket to me that pays \$1 if "$\Theta \in H$" to me for no less than $\$\overline{\mathsf{Q}}(H)$. Intuitively, You shouldn't be willing to buy a gamble for a price higher than You're willing to sell it for—that would put You at risk of a sure loss—so the setup would be unsatisfactory if it didn't rule out this case. Indeed, the assumed structure of the credal set implies that $\underline{\mathsf{Q}}(H) \leq \overline{\mathsf{Q}}(H)$ for all H. Moreover, $\overline{\mathsf{Q}}$ is not just a derivative of $\mathscr{Q}$, via (3), they're in one-to-one correspondence:

$$\mathscr{Q} = \{\mathsf{Q} : \mathsf{Q}(H) \leq \overline{\mathsf{Q}}(H) \text{ for all measurable } H\}. \tag{5}$$

That is, $\overline{\mathsf{Q}}$ is exactly the upper envelope of $\mathscr{Q}$ and, consequently, credal set-driven imprecise probabilities are *coherent* (e.g., Walley 1991), generalizing the notion put forth by de Finetti, Savage, Ramsey, and others in the context of precise probability. While these betting-based notions are compelling, that interpretation isn't necessary to apply the theory. One can—and here I will—follow Shafer (1976) and others by interpreting $\underline{\mathsf{Q}}(H)$ and $\overline{\mathsf{Q}}(H)$ as Your quantitative degree of confidence/support/belief and degree of plausibility, respectively, in the truthfulness of the claim "$\Theta \in H$."

As in precise probability, it's natural to extend probabilities to expected values, which is done via Lebesgue's integration theory. In imprecise probability, there is the same desire to extend the theory to cover more general uncertain quantities—what are commonly referred to as *gambles*—and this leads to lower and upper expected values or, in the spirit of de Finetti, to lower and upper previsions (e.g., Troffaes and de Cooman 2014; Walley 1991). Given $g : \mathbb{T} \to \mathbb{R}$, the lower and upper expected value is defined as

$$\underline{\mathsf{Q}}\, g = \inf_{\mathsf{Q} \in \mathscr{Q}} \mathsf{E}^{\Theta \sim \mathsf{Q}}\{g(\Theta)\} \quad \text{and} \quad \overline{\mathsf{Q}}\, g = \sup_{\mathsf{Q} \in \mathscr{Q}} \mathsf{E}^{\Theta \sim \mathsf{Q}}\{g(\Theta)\}, \tag{6}$$

i.e., the lower and upper limits of the usual expected values of $g(\Theta)$ with respect to probability distributions Q in the credal set $\mathscr{Q}$. According to the definition above, computation of the lower and upper expectations appears to require optimization over a potentially infinite-dimensional space $\mathscr{Q}$, which is intimidating. In certain cases, however, there are equivalent formulations, which often have computationally simpler formulas. Roughly, the *upper envelope theorem* in Troffaes and de Cooman (2014, Theorem 4.38) connects the bounds in (6) to the so-called *natural extension* of $\overline{\mathsf{Q}}$, and then their Theorem 6.14 links the natural extension to *Choquet integration* (Choquet 1954). I'll present the Choquet integral formula for one special case below; further details are given in Appendix A.

An important special case of imprecise probability, which is my main focus in what follows, is *possibility theory*. This theory goes back at least to Shackle (1961) and is intertwined with the theory of fuzzy sets (e.g., Zadeh 1965, 1978). The seminal text on the subject is Dubois and Prade (1988) and I recommend Dominik Hose's PhD thesis (Hose 2022) for a modern review; specific details about possibility theory for statistics

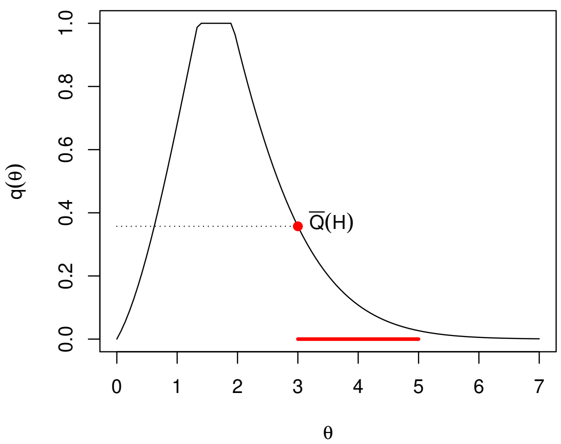


Figure 1: A possibility contour q, the hypothesis of interest $H = [3,5]$ (red), and the corresponding upper probability $\overline{\mathsf{Q}}(H)$ determined by optimization as in (7).

can be found in, e.g., Dubois (2006) and Martin (2021a, 2022b). What distinguishes possibility theory from other brands of imprecise probability is that it's determined by a real-valued function on $\mathbb{T}$. This is analogous to precise probability theory where probabilities are evaluated by integrating a density function; in possibility theory, the calculus is optimization rather than integration. Define a contour function $q : \mathbb{T} \to [0,1]$ such that $\sup_{\theta \in \mathbb{T}} q(\theta) = 1$, the condition analogous to a probability density function integrating to 1. Then the corresponding upper probability $\overline{\mathsf{Q}}$ is defined via optimization as

$$\overline{\mathsf{Q}}(H) = \sup_{\theta \in H} q(\theta), \quad H \subseteq \mathbb{T}. \tag{7}$$

A simple illustration of a possibility contour q and of the formula (7) is shown in Figure 1.

I'll end this background section with two unique possibility-theoretic details. The first is a simple formula for $\overline{\mathsf{Q}}$'s extension to a possibility-theoretic upper expected value. If $g : \mathbb{T} \to \mathbb{R}$ is a suitable non-negative function and $\overline{\mathsf{Q}}$ a possibility measure determined by contour q as described above, then the corresponding possibilistic upper expected value (see Appendix A) is given by

$$\overline{\mathsf{Q}}\, g = \int_0^1 \Big\{ \sup_{\theta : q(\theta) > s} g(\theta) \Big\}\, ds. \tag{8}$$

The key observation is that, by definition, the left-hand side of (8) is the supremum of ordinary Q-expected values over an infinite-dimensional $\mathscr{Q}$, whereas the right-hand side of (8) is a one-dimensional (Riemann) integral of a $\dim(\mathbb{T})$-dimensional supremum.

Second, there is a simple and elegant characterization of the aforementioned credal set $\mathscr{Q}$ associated with the possibility measure $\overline{\mathsf{Q}}$ determined by contour q. Indeed, Couso et al. (2001) and Dubois et al. (2004) have shown that

$$\mathsf{Q} \in \mathscr{Q} \iff \mathsf{Q}\{q(\Theta) \leq \alpha\} \leq \alpha \;\text{ for all } \alpha \in [0,1]. \tag{9}$$

Statistician readers are sure to notice the similarity to familiar notions like p-values: the probability Q belongs to the credal set if and only if the random variable $q(\Theta)$ satisfies the mathematical properties of a p-value as a function of $\Theta \sim \mathsf{Q}$.

3 Regularized e-processes

3.1 Objective

This section develops a framework of *regularized e-processes* that offers knowledge-based efficiency gains compared the basic e-processes surveyed in Section 2.2 above. The point is that one generally can't make non-trivial changes to an existing e-process without jeopardizing its anytime validity properties, so care is needed. My strategy is to relax the definition of "anytime valid" in a principled, knowledge-aware manner that allows for greater flexibility and, in turn, creates opportunities for improved efficiency.

Recall that regularization is predicated on the availability of genuine, relevant information about the quantities of interest. My proposal for knowledge-based e-process regularization is, roughly, to make use of exactly the information that's available in Your corpus of knowledge, no more and no less, and then combine this with an existing e-process such that the combination tends to be larger and yet is still anytime valid in the relaxed sense. More specifically, my proposal is two-fold:

1. Encode Your corpus of knowledge, which is initially expressed as an imprecise probability as in Section 2.3, in the form of a *regularizer* and then (easily) combine it with any data-driven e-process as discussed in Section 2.2 above, and
2. Instead of requiring e-processes to be anytime valid relative to a vacuous corpus of knowledge, I'll ask for a weaker version that's relative to the Your prior knowledge, and I'll offer a generalization of Ville's inequality that's satisfied by a collection of e-processes that tend to be larger, hence more efficient, than those in Section 2.2.

3.2 Prior information and regularization

Suppose Your knowledge about Θ takes the form of a (possibly vacuous) imprecise probability determined by a credal set $\mathscr{Q}$ or, equivalently, an upper probability $\overline{\mathsf{Q}}$. Introduce a function $\rho : \mathbb{T} \to [0, \infty]$ such that $\rho(\theta)$ is interpreted as the weight of evidence (relative to $\mathscr{Q}$ or $\overline{\mathsf{Q}}$) against the basic proposition "$\Theta = \theta$." This is different from the more familiar statistical notions of evidence because here there's no data. Not just any function ρ can be a measure of evidence, and the more familiar notions of evidence provide some guidance as to what should be expected here. Good (1950, Ch. 6.7), for example, proves a now-well-known result (which he attributes to Turing) that, in certain cases, the expectation of the Bayes factor weighing evidence against a true proposition is equal to 1. More generally, this "expected value is no more than 1" property is key to all the recent e-value developments; see Equation (1) above. To be a proper *regularizer*, the function ρ must be similarly bounded in (upper-)expectation.

Definition 1. Let $\overline{\mathsf{Q}}$ be an upper probability/prevision on $\mathbb{T}$. A function $\rho : \mathbb{T} \to [0, \infty]$ is a regularizer if its $\overline{\mathsf{Q}}$-upper expected value is bounded by 1, i.e., if $\overline{\mathsf{Q}}\rho \leq 1$.

This is rather abstract, so I'll give some concrete descriptions shortly; see, also, Section 3.3 and Appendix B. For now, I'll offer just some general comments. First, any non-negative function ρ that's bounded above by 1 is a regularizer, but such a trivial choice offers no meaningful regularization and, therefore, has no value. There's another way to see this: if the prior is fully vacuous, i.e., if You literally know nothing about Θ *a priori*, so that $\overline{\mathsf{Q}}\rho = \sup_{\theta \in \mathbb{T}} \rho(\theta)$, then choosing $\rho \leq 1$ is the only choice that satisfies the condition in Definition 1. So, a *trivial* choice of ρ that's upper-bounded by 1 matches the no-prior-information case and, hence, can't offer any meaningful regularization. Henceforth, except for the vacuous prior case where $\rho \equiv 1$ is the clear choice of regularizer, I'll focus on *non-trivial* regularizers, i.e., those that aren't upper-bounded by 1. Then the general goal is, for Your given $\overline{\mathsf{Q}}$, to find a function ρ which can take values possibly much larger than 1 while still satisfying $\overline{\mathsf{Q}}\rho \leq 1$. Second, if there are two regularizers ρ and ρ' with $\rho'(\theta) \leq \rho(\theta)$ for all θ, then the dominant or *larger* regularizer ρ is preferred. Formally, I'll say that a regularizer is *admissible* if $\overline{\mathsf{Q}}\rho = 1$, that is, if it can't be made larger in any substantive way without violating Definition 1.

Towards a more concrete understanding of what it takes to achieve the condition in Definition 1, consider the case of finite $\mathbb{T}$. Then Your prior credal set $\mathscr{Q}$ can be interpreted as a collection of probability mass functions, and

$$\overline{\mathsf{Q}}\rho = \sup_{\mathsf{Q} \in \mathscr{Q}} \underbrace{\sum_{\theta \in \mathbb{T}} \rho(\theta)\, \mathsf{Q}(\{\theta\})}_{\mathsf{Q}\rho := \mathsf{E}^{\Theta \sim \mathsf{Q}}\{\rho(\Theta)\}} \leq \sum_{\theta \in \mathbb{T}} \rho(\theta) \sup_{\mathsf{Q} \in \mathscr{Q}} \mathsf{Q}(\{\theta\}).$$

Since the functional $\mathsf{Q} \mapsto \mathsf{Q}\rho$ is linear and the domain $\mathscr{Q}$ is closed and convex, it's well-known (related to the Krein–Milman theorem, e.g., Bauer 1958) that the supremum is attained on the extreme points of $\mathscr{Q}$. Then it's clear that, for each (sub-)probability mass function η on $\mathbb{T}$, the following function ρ_η is a regularizer:

$$\rho_\eta(\theta) = \frac{\eta(\theta)}{\sup_{\mathsf{Q} \in \mathscr{Q}} \mathsf{Q}(\{\theta\})}, \quad \theta \in \mathbb{T}.$$

Two special cases deserve mention. First, if $\mathscr{Q}$ is a singleton, then all the admissible regularizers would be of the form ρ_η above, for some probability mass function η. Second, if $\mathscr{Q}$ is vacuous, i.e., if it contains all the probability mass functions on $\mathbb{T}$, then the denominator is constant equal to 1 and, since η can't exceed 1, neither can ρ_η, hence, it's trivial. So, clearly, it's a bad idea for You to plug a "vacuous prior" into the regularizer construction. If Your prior information is genuinely vacuous, then just use $\rho \equiv 1$, so that Your regularized e-process (10) matches the original, unregularized e-process.

The quality of the regularizer is determined by how much of a knowledge-based boost it offers to an existing e-process $\mathfrak{e}_\theta(\cdot)$ for Θ, as described in Section 2.2. My proposal is to define a *regularized e-process* by combining ρ and $(\mathfrak{e}_\theta : \theta \in \mathbb{T})$ as follows:

$$\mathfrak{e}^{\text{reg}}(\cdot, \theta) = \rho(\theta) \times \mathfrak{e}_\theta(\cdot), \quad \theta \in \mathbb{T}. \tag{10}$$

The intuition is that, if θ is incompatible with prior knowledge, so that $\rho(\theta) \gg 1$, then $\mathfrak{e}^{\text{reg}}(\cdot, \theta) \gg \mathfrak{e}_\theta(\cdot)$, which creates an opportunity for increased efficiency. But non-trivially manipulating the original e-process in an effort to boost efficiency will surely jeopardize anytime validity in the sense of Section 2.2, hence care is needed. In Section 3.5 below, I'll establish my main claim: in a certain sense, the regularized e-process enjoys efficiency gains without jeopardizing anytime validity. This "certain sense" involves what I'll argue is a meaningful imprecise-probabilistic relaxation of anytime validity as in Section 2.2.

Finally, there are justifications for my choice to define the regularized e-process in (10) via multiplication. One nice property is that it's updating-coherent in a Bayesian-like sense, i.e., "today's posterior is tomorrow's prior." To see this, let $z^n \equiv z^{1:n}$ be the data, processed as $\mathfrak{e}_\theta(z^{1:n})$, and $\rho(\theta)$ the regularizer. Combining these according to (10) gives the regularized e-process $\mathfrak{e}^{\text{reg}}(z^n, \theta)$. Now, suppose that more data $z^{n:(n+m)}$ becomes available. Since e-processes are typically combined via multiplication, I can write

$$\mathfrak{e}^{\text{reg}}(z^{1:(n+m)}, \theta) = \mathfrak{e}_\theta(z^{1:(n+m)}) \times \rho(\theta) = \mathfrak{e}_\theta(z^{n:(n+m)}) \times \underbrace{\mathfrak{e}_\theta(z^{1:n}) \times \rho(\theta)}_{\mathfrak{e}^{\text{reg}}(z^{1:n},\theta)}.$$

That is, the regularized e-process based on $z^{1:n}$ becomes the updated, old-data-dependent version of the regularizer that's combined with a new-data-dependent e-process according to the rule (10). There's at least one more principled justification for the product (10), similar to a result in Vovk and Wang (2021); see Appendix C.

3.3 Special case: possibilistic priors

Suppose Your prior knowledge about Θ is encoded by a possibility measure with contour function q; see Appendix B for others. Then Your $\overline{\mathsf{Q}}$ is defined via optimization as in (7), and upper expectation is as in (8). This is the simplest imprecise probability model, and simplicity is virtue: Your prior knowledge about Θ is necessarily limited, so a more expressive imprecise probability model used to describe it makes elicitation more difficult. The claim is that experts can offer statements like: "I'd not be surprised at all if $\Theta = a$, I'd be a little surprised if $\Theta = b$, and I'd be very surprised if $\Theta = c$." (*Surprise* is due to Shackle 1961.) Then the qualitative statement above could be made quantitative by introducing a possibility contour q with $q(a) = 1$, $q(b)$ smaller, $q(c)$ very small, etc. This is often how penalties and priors are chosen. For example, a Bayesian might take the least-surprising value to be the prior mode and then choose a density that's a decreasing function of surprise. But there's a drastic difference between a probability density and a possibility contour, even if they have similar shapes: the former determines a precise probability thereby adding artificial information to Your corpus of knowledge.

Returning to the regularization details, recall that, for the regularizer ρ to be effective, it needs to take values (potentially much) larger than 1 at θ values which are suitably incompatible with the prior information. In the present context, "incompatibility" would correspond to a small q value, so it makes sense that $\rho(\theta)$ be large when $q(\theta)$ is small. As a first attempt, this could be accomplished by taking ρ to be the reciprocal of q. But q is no more than 1, so such a ρ would never be less than 1 and hence it can't be a regularizer in the sense of Definition 1. Apparently q is too small for the reciprocal to work, so it needs to be inflated first. Following Shafer et al. (2011, Sec. 6), define a *calibrator* as a non-decreasing function $\gamma : [0, 1] \to (0, \infty]$ such that

$$\int_0^1 \frac{1}{\gamma(s)} \, ds = 1. \tag{11}$$

Of course, γ can taken to be the reciprocal of any probability density function supported on $[0, 1]$. Then I'll define the *regularizer* in terms of the calibrated version of q:

$$\rho(\theta) = \{\gamma \circ q(\theta)\}^{-1}, \quad \theta \in \mathbb{T}. \tag{12}$$

The role played by the calibrator is to inflate the contour "just enough."

Proposition 1. *Let* $\overline{\mathsf{Q}}$ *be the possibility measure determined by the contour function* q*. Then* ρ *defined in* (12) *is a regularizer, i.e., has* $\overline{\mathsf{Q}}$*-upper expectation bounded by 1.*

Proof. By definition of the $\overline{\mathsf{Q}}$-upper expectation,

$$\overline{\mathsf{Q}}\,\rho = \int_0^1 \Bigl\{ \sup_{\theta: q(\theta) > s} \rho(\theta) \Bigr\}\, ds = \int_0^1 \Bigl\{ \sup_{\theta: q(\theta) > s} \frac{1}{\gamma \circ q(\theta)} \Bigr\}\, ds.$$

Since γ is non-decreasing,

$$q(\theta) > s \implies \frac{1}{\gamma \circ q(\theta)} \leq \frac{1}{\gamma(s)},$$

and from here the claim follows immediately from (11). □

I'll follow Vovk and Wang (2021, Appendix B) and suggest use of (the reciprocal of) a suitable beta mixture of beta density functions, which takes the form

$$\gamma(u) = \frac{u\,(-\log u)^{1+\kappa}}{\kappa \times \texttt{igamma}(-\log u, 1+\kappa)}, \quad u \in [0,1], \quad \kappa > 0,$$

where $\texttt{igamma}(z, \alpha) = \int_0^z t^{\alpha-1}\, e^{-t}\, dt$ is the incomplete gamma function. A particular advantage of the above calibrator is how rapidly it vanishes as $u \to 0$, which is directly related to how severely the corresponding regularizer penalizes those θ values incompatible with the prior information. The gamma defined in the above display is the calibrator that I'll use for my illustrations in Section 3.7 and elsewhere.

Although the context and form here are different, the message above is a familiar one to those who have experience with e-processes, etc. Indeed, Equation (9) implies that, roughly, $q(\Theta)$ is a p-value relative to $\overline{\mathsf{Q}}$. And it's well-known (e.g., Sellke et al. 2001; Vovk 1993) that one must calibrate a p-value first before its reciprocal can be an e-value in the sense of having expected value bounded by 1.

3.4 Induced upper joint distributions

A bit more background is needed before I can get into the regularized e-process's properties. Recall that $\Theta = f(\Omega) \in \mathbb{T}$ is a function of $\Omega \in \mathbb{O}$. Since $\mathbb{T}$ is defined as the image of $\mathbb{O}$, the map $f : \mathbb{O} \to \mathbb{T}$ is trivially surjective. There are applications in which f would be a bijection, e.g., when f is the identity mapping, so that Θ and Ω are equivalent in some sense. But it's typical that $\Theta = f(\Omega)$ is a lower-dimensional feature of Ω. Whether f is or isn't a bijection becomes relevant when considering how knowledge about Θ translates to knowledge about the primitive Ω from which $\Theta = f(\Omega)$ is derived. The *push-forward* operation that takes a probability for Ω to a corresponding probability for $\Theta = f(\Omega)$ is well-defined, whereas the reverse *pull-back* operation is not unique when f is not a bijection. Details concerning pull-back measures are somewhat technical, but these aren't particularly relevant here, so I give only the key points.

Given a probability distribution Q for Θ on $\mathbb{T}$, the relevant question is under what conditions does there exist a corresponding probability distribution, say, R, for Ω on $\mathbb{O}$ such that the distribution of $f(\Omega)$ is Q. The classical result of Varadarajan (1963, Lemma 2.2) establishes the existence of an R corresponding to Q under very mild conditions, easily met in the statistical applications I have in mind here. In fact, for a given Q, there's a non-empty class $\mathscr{R}_{\mathsf{Q}}$ of such distributions, i.e.,

$$\mathscr{R}_{\mathsf{Q}} = \{\mathsf{R} : \text{if } \Omega \sim \mathsf{R}, \text{ then } f(\Omega) \sim \mathsf{Q}\}, \tag{13}$$

and, moreover, under the conditions imposed here, this class is easily shown to satisfy the properties of a credal set.

Proposition 2. *For any given* Q *and continuous* $f : \mathbb{O} \to \mathbb{T}$*, the collection* $\mathscr{R}_{\mathsf{Q}}$ *in* (13) *is non-empty, convex, and closed with respect to the weak topology.*

Proof. Take any fixed $\mathsf{Q} \in \mathscr{Q}$. Non-emptiness of $\mathscr{R}_{\mathsf{Q}}$ is a consequence of the classical results mentioned above. To prove convexity, take two elements R_1 and R_2 in $\mathscr{R}_{\mathsf{Q}}$ and a constant $\tau \in [0,1]$, and then define the mixture $\mathsf{R}^{\text{mix}} = (1-\tau)\mathsf{R}_1 + \tau\mathsf{R}_2$; the goal is to show that $\mathsf{R}^{\text{mix}} \in \mathscr{R}_{\mathsf{Q}}$. If $\Omega \sim \mathsf{R}^{\text{mix}}$, then for any event H in $\mathbb{T} = f(\mathbb{O})$,

$$\begin{aligned} \mathsf{R}^{\text{mix}}\{f(\Omega) \in H\} &= (1-\tau)\,\mathsf{R}_1\{f(\Omega) \in H\} + \tau\,\mathsf{R}_2\{f(\Omega) \in H\} \\ &= (1-\tau)\,\mathsf{Q}(H) + \tau\,\mathsf{Q}(H) \\ &= \mathsf{Q}(H). \end{aligned}$$

This implies that $\mathsf{R}^{\text{mix}} \in \mathscr{R}_{\mathsf{Q}}$ if R_1 and R_2 are and, therefore, that $\mathscr{R}_{\mathsf{Q}}$ is convex. Finally, to prove that $\mathscr{R}_{\mathsf{Q}}$ is closed with respect to the weak topology, consider a sequence $(\mathsf{R}_t : t \geq 1)$ in $\mathscr{R}_{\mathsf{Q}}$ with a weak limit R_∞; the goal is to show that $\mathsf{R}_\infty \in \mathscr{R}_{\mathsf{Q}}$. By the continuous mapping theorem, if $\Omega_t \sim \mathsf{R}_t$, then $f(\Omega_t) \to f(\Omega_\infty)$ in distribution as $t \to \infty$. But the distribution of $f(\Omega_t)$ is Q for all t and, consequently, the distribution of $f(\Omega_\infty)$ is also Q. This implies R_∞ is contained in $\mathscr{R}_{\mathsf{Q}}$ and, hence, the latter collection is closed. □

An advantage to this credal set characterization is that it allows for the construction of a coherent upper joint distribution for (Z^N, Ω) based solely on the model $\mathscr{P}$ and Your prior information in $\mathscr{Q}$ about Θ. Furthermore, this construction is conceptually straightforward: just maximize over all the ordinary joint distributions for (Z^N, Ω) in the above-defined credal set. Specifically, let $\overline{\mathbf{P}}$ denote this induced upper joint distribution for (Z^N, Ω), which is defined as via its upper expectation as

$$\overline{\mathbf{P}}\, g = \sup_{\mathsf{Q} \in \mathscr{Q}} \sup_{\mathsf{R} \in \mathscr{R}_{\mathsf{Q}}} \underbrace{\mathsf{E}^{\Omega \sim \mathsf{R}}\big[\mathsf{E}^{Z \sim \mathsf{P}_\Omega}\{g(Z^N, \Omega)\}\big]}_{\text{joint distribution expectation}}, \tag{14}$$

where $g : \mathbb{Z}^\infty \times \mathbb{O} \to \mathbb{R}$ is a suitable function, and the highlighted term is just the usual expectation of $g(Z^N, \Omega)$ with respect to the joint distribution determined by the (marginal) distribution R for Ω and the (now-interpreted-as-a-conditional) distribution P_ω for Z^N, given $\Omega = \omega$. As mentioned in Section 2.3, $\overline{\mathbf{P}}$-upper probabilities correspond to the right-hand side above with the appropriate indicator function plugged in for g. It's with respect to this upper joint distribution that the regularized e-process has desirable properties, as I demonstrate next.

3.5 Regularized Ville's inequality & anytime validity

My claim is that the regularized e-process satisfies a prior knowledge-dependent versions of property (1) and of Ville's inequality (2). This leads to a corresponding "regularized" version of anytime validity which, among other things, is discussed in Section 3.6 below.

Theorem 1. *Suppose the available prior information is encoded in the upper probability* $\overline{\mathsf{Q}}$*, which determines an upper joint distribution* $\overline{\mathbf{P}}$ *as in* (14)*. If* ρ *is a regularizer in the sense of Definition 1 relative to* $\overline{\mathsf{Q}}$*, then for any e-process* $\{\mathfrak{e}_\theta(\cdot) : \theta \in \mathbb{T}\}$*, the corresponding regularized version* $\mathfrak{e}^{\text{reg}}$ *in* (10) *satisfies*

$$\overline{\mathbf{P}}\big[\mathfrak{e}^{\text{reg}}\{Z^N, f(\Omega)\}\big] \leq 1, \quad \textit{for all stopping times } N, \tag{15}$$

and the following regularized Ville's inequality holds:

$$\overline{\mathbf{P}}\big[\mathfrak{e}^{\text{reg}}\{Z^N, f(\Omega)\} > \alpha^{-1}\big] \leq \alpha \quad \textit{all } \alpha \in [0,1], \textit{ all stopping times } N. \tag{16}$$

Proof. The first claim is just a direct computation using (14) when the function $g = \mathfrak{e}^{\text{reg}}$ factors as $g(\cdot, \omega) = \rho(f(\omega))\, \mathfrak{e}_{f(\omega)}(\cdot)$:

$$\begin{aligned}
\overline{\mathbf{P}}\, \mathfrak{e}^{\text{reg}} &= \sup_{\mathsf{Q} \in \mathscr{Q}} \sup_{\mathsf{R} \in \mathscr{R}_{\mathsf{Q}}} \mathsf{E}^{\Omega \sim \mathsf{R}}\big[\mathsf{E}^{Z \sim \mathsf{P}_\Omega}\{\mathfrak{e}^{\text{reg}}(Z^N, f(\Omega))\}\big] \\
&= \sup_{\mathsf{Q} \in \mathscr{Q}} \sup_{\mathsf{R} \in \mathscr{R}_{\mathsf{Q}}} \mathsf{E}^{\Omega \sim \mathsf{R}}\big[\rho(f(\Omega))\, \underbrace{\mathsf{E}^{Z \sim \mathsf{P}_\Omega}\{\mathfrak{e}_{f(\Omega)}(Z^N)\}}_{\leq 1, \text{ by } (1)}\big] \\
&\leq \sup_{\mathsf{Q} \in \mathscr{Q}} \sup_{\mathsf{R} \in \mathscr{R}_{\mathsf{Q}}} \mathsf{E}^{\Omega \sim \mathsf{R}}\big[\rho\{f(\Omega)\}\big] \\
&= \sup_{\mathsf{Q} \in \mathscr{Q}} \mathsf{E}^{\Theta \sim \mathsf{Q}}\{\rho(\Theta)\} \\
&\leq 1,
\end{aligned}$$

where the last inequality follows by definition of the regularizer ρ, and the penultimate equality follows by definition of $\mathscr{R}_{\mathsf{Q}}$: if $\Omega \sim \mathsf{R} \in \mathscr{R}_{\mathsf{Q}}$, then the distribution of $f(\Omega)$ is the same as that of Θ under Q. The second claim, the regularized Ville's inequality in (16), follows from the first claim and an application of Markov's inequality inside the upper-probability calculation. If I write "$(Z, \Omega) \sim \mathsf{P}_\bullet \otimes \mathsf{R}$" to represent the joint distribution of (Z, Ω) under the model where $\Omega \sim \mathsf{R}$ and $(Z \mid \Omega = \omega) \sim \mathsf{P}_\omega$, then:

$$\begin{aligned}
\overline{\mathbf{P}}\big[\mathfrak{e}^{\text{reg}}\{Z^N, f(\Omega)\} > \alpha^{-1}\big] &= \sup_{\mathsf{Q} \in \mathscr{Q}} \sup_{\mathsf{R} \in \mathscr{R}_{\mathsf{Q}}} \mathsf{P}^{(Z,\Omega) \sim \mathsf{P}_\bullet \otimes \mathsf{R}}\big[\mathfrak{e}^{\text{reg}}\{Z^N, f(\Omega)\} > \alpha^{-1}\big] \\
&\leq \sup_{\mathsf{Q} \in \mathscr{Q}} \sup_{\mathsf{R} \in \mathscr{R}_{\mathsf{Q}}} \alpha\, \mathsf{E}^{\Omega \sim \mathsf{R}}\big[\mathsf{E}^{Z \sim \mathsf{P}_\Omega}\{\mathfrak{e}^{\text{reg}}(Z^N, f(\Omega))\}\big] \\
&= \alpha\, \overline{\mathbf{P}}\, \mathfrak{e}^{\text{reg}} \\
&\leq \alpha,
\end{aligned}$$

where the first "$\leq$" above is by the usual Markov's inequality and the last line is by the claim (15) proved above. □

The above result is related to the"uniformly-randomized Markov inequality" in Ramdas and Manole (2023, Theorem 1.2). They prove that, if X and U are independent random variables, with $X \geq 0$ and U stochastically no smaller than $\mathsf{Unif}(0,1)$, then $\mathsf{P}(X/U \geq a^{-1}) \leq a\,\mathsf{E}(X)$ for $a > 0$. The key observation is that the random variable X is being boosted by the division by $U \leq 1$ but the same Markov inequality bound holds, thereby sharpening the usual Markov inequality claims. The connection between their randomized Markov inequality and the result in Theorem 1 and is most clear in the special case where the regularizer ρ is defined in terms of the possibility contour q as in Section 3.3. In such a case, recall that the random variable $q(\Theta)$ is no smaller than $\mathsf{Unif}(0,1)$ when $\Theta \sim \mathsf{Q}$, for any prior Q in the credal set $\mathscr{Q}$. Since the calibration step $\gamma \circ q(\Theta)$ boosts it further, the proposed regularization is simply dividing the usual e-process by a no-smaller-than-uniform random variable.

3.6 Statistical implications

Theorem 1's most important take-away is its implication concerning safe, anytime valid inference. Indeed, when relevant, non-vacuous prior information is available, Theorem 1 shows how that knowledge can be used to enhance an existing e-process in such a way that anytime validity is preserved but efficiency is generally gained. This enhancement is achieved through the incorporation of a regularizer as in (10) that inflates and deflates the original e-process when the latter is large and small, respectively. Specifically, if the goal is to test "$\Theta \in H$," then the regularized e-process-based testing procedure

$$\text{reject "}\Theta \in H\text{" based on data } z^n \iff \inf_{\theta \in H} \mathfrak{e}^{\text{reg}}(z^n, \theta) > \alpha^{-1}$$

will be anytime valid (relative to Your prior knowledge) in the sense that

$$\begin{aligned}\overline{\mathbf{P}}(\text{Type I error}) &:= \overline{\mathbf{P}}\{f(\Omega) \in H \text{ and test of `}f(\Omega) \in H\text{' rejects}\} \\ &= \overline{\mathbf{P}}\Big\{f(\Omega) \in H \text{ and } \inf_{\theta \in H} \mathfrak{e}^{\text{reg}}(Z^N, \theta) > \alpha^{-1}\Big\} \\ &\leq \overline{\mathbf{P}}\big\{\mathfrak{e}^{\text{reg}}(Z^N, f(\Omega)) > \alpha^{-1}\big\} \\ &\leq \alpha,\end{aligned}$$

where the last line follows by (16). This is a non-trivial generalization of the familiar frequentist Type I error control so it warrants some detailed comments. In the statistics literature, the prevailing viewpoint is that the quantity of interest Θ is a fixed unknown, so the statement "$\Theta \in H$" is absolutely true for some H and absolutely false for others, but nothing more can be said. This aligns with the vacuous-prior case where, with the exception of the trivial hypotheses $\varnothing$ and $\mathbb{T}$, Your corpus of knowledge makes exactly the same statements about every hypothesis: $\underline{\mathsf{Q}}(H) = 0$ and $\overline{\mathsf{Q}}(H) = 1$. That is, there's no evidence supporting the truthfulness of either "$\Theta \in H$" or "$\Theta \notin H$." With this extreme form of prior knowledge, Your corresponding regularizer is $\rho \equiv 1$—so that $\mathfrak{e}^{\text{reg}}(\cdot, \theta) =$

$\mathfrak{e}_\theta(\cdot)$—and $\overline{\mathbf{P}}$ simplifies[2] to yield the following:

$$\begin{aligned}\overline{\mathbf{P}}(\text{Type I error}) &:= \overline{\mathbf{P}}\{f(\Omega) \in H \text{ and test of `}f(\Omega) \in H\text{' rejects}\} \\ &= \sup_{\omega: f(\omega) \in H} \mathsf{P}_\omega\Big\{\inf_{\theta \in H} \mathfrak{e}_\theta(Z^N) > \alpha^{-1}\Big\}.\end{aligned}$$

Then the anytime Type I error control property of the original e-process-based test, derived from (2), is recovered as a special case of Theorem 1. Moreover, the familiar and intuitive "supremum-over-the-null" is simply what appears when the imprecise-probabilistic framework formulated here is applied to the special vacuous-prior case. Beyond this extreme case, the different H's have varying degrees of belief/plausibility and this is taken into account in the evaluation of the test via the conjunction: "$f(\Omega) \in H$" *and* "test of '$f(\Omega) \in H$' rejects." In particular, monotonicity of $\overline{\mathbf{P}}$ implies that $\overline{\mathbf{P}}(\text{Type I error}) \leq \overline{\mathsf{Q}}(H)$, which means that the test doesn't have to be particularly good for those H with small $\overline{\mathsf{Q}}(H)$ to control the (generalized) Type I error rate. If (generalized) Type I error control is easy for those *a priori* implausible H's, then that gives the e-process an opportunity to focus its effort on handling the plausible H's more efficiently.

Similarly, a nominal $100(1-\alpha)\%$ confidence region for Θ, based on the regularized e-process applied to data z^n, is

$$C_\alpha^{\text{reg}}(z^n) = \big\{\theta \in \mathbb{T} : \mathfrak{e}^{\text{reg}}(z^n, \theta) \leq \alpha^{-1}\big\},$$

i.e., the collection of simple null hypotheses that the regularized e-process-based test would not reject at level α. Then the non-coverage (upper) probability is

$$\overline{\mathbf{P}}\{C_\alpha^{\text{reg}}(Z^N) \not\ni f(\Omega)\} = \overline{\mathbf{P}}\{\mathfrak{e}^{\text{reg}}(Z^N, f(\Omega)) > \alpha^{-1}\} \leq \alpha, \tag{17}$$

which implies that the corresponding lower probability of coverage, i.e., of $C_\alpha^{\text{reg}}(Z^N) \ni f(\Omega)$, is bounded from below by $1-\alpha$. This justifies calling C_α^{reg} a (generalized) "anytime $100(1-\alpha)\%$ confidence region." Like in the discussion above, this is generally different from the usual notion of coverage probability of a confidence set. In the case of vacuous prior information, however, the regularizer is $\rho \equiv 1$, the corresponding confidence region is $C_\alpha(z^n) = \{\theta \in \mathbb{T} : \mathfrak{e}_\theta(z^n) \leq \alpha^{-1}\}$, and inequality (17) reduces to

$$\sup_\omega \mathsf{P}_\omega\{C_\alpha^{\text{reg}}(Z^N) \not\ni f(\omega)\} \leq \alpha.$$

This, of course, is exactly the familiar anytime coverage probability guarantees offered by e-process-based confidence sets, via (2).

Finally, I want to (re-)emphasize that there's more to this proposal than anytime validity preservation: regularization creates opportunities for efficiency gain. Recall that Your knowledge encoded in $\mathscr{Q}$ or $\overline{\mathsf{Q}}$ is "real"—no reason to include it otherwise!—so You expect $\rho(\Theta)$ to be small; see Definition 1. Therefore, in the absence of data, if $\rho(\theta)$ is large, then You'd infer that "$\Theta = \theta$" isn't true. So, You'd be satisfied if θ's present in

[2]This simplification may not be immediately obvious, so here's an explanation. Expectation of a random variable is linear in the distribution and, therefore, it's supremum over a closed and convex set of distributions is attained at the extremes, on the boundary. If the prior is vacuous, so that the credal set contains all distributions, then the boundary consists of point mass distributions. Hence, the supremum expectation over all distributions is attained at a point mass distribution.

the unregularized e-process confidence region $C_\alpha(z^n)$ happened to be absent from Your $C_\alpha^{\text{reg}}(z^n)$ if it's because a large regularizer $\rho(\theta)$ value knocked them out. It's the removal of θ's that are sufficiently incompatible with the prior knowledge that leads to the improved efficiency. While it's not generally true that $C_\alpha^{\text{reg}}(z^n)$ is contained in $C_\alpha(z^n)$ for all α and (almost) all z^n, the regularized confidence regions do tend to be narrower than their unregularized counterpart, as demonstrated in Section 3.7 below.

3.7 Efficiency gains

The goal here is to illustrate the proposed regularized e-process framework, in particular, to demonstrate how and to what extent inference based on a given e-process experiences efficiency gains through the incorporation of partial prior information and regularization. For this purpose, I'll take a simple model and a simple e-process so that effort can be focused on understanding the possibilistic prior information, the corresponding regularization, and the effect this has in terms of efficiency.

To set the scene, consider a normal mean model with known variance equal to 1. I'm assuming a parametric model, hence the uncertain model index, Ω, and the quantity of interest, Θ, are the same. So I'll write $\mathscr{P} = \{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$, where $\mathsf{P}_\theta = \mathsf{N}(\theta, 1)$ and $\mathbb{T} = \mathbb{R}$. For the e-process, I'll consider the Bayes factor

$$\mathfrak{e}_\theta(\cdot) = \frac{\int L_t(\cdot)\, \xi(t)\, dt}{L_\theta(\cdot)},$$

where L_θ is the usual Gaussian likelihood function and, in the numerator, ξ is a mixing probability density function on $\mathbb{T}$. In what follows, I'll take ξ to be the $\mathsf{N}(0, v)$ density function, with $v = 10$. This is a version of the so-called Savage–Dickey e-process as described in, e.g., Grünwald (2023). The integration can be done in closed-form, so

$$\mathfrak{e}_\theta(z^n) = (nv+1)^{1/2} \exp\bigl\{-\tfrac{n}{2}(\theta - \bar{z}_n)^2 + \tfrac{1}{2}\bigl(\tfrac{n}{nv+1}\bigr)\, \bar{z}_n^2\bigr\}, \quad \theta \in \mathbb{T}.$$

In words, the e-process is a ratio of the (Bayesian marginal) likelihood under the model where Θ is different from θ—where "different from θ" is quantified by ξ—to the likelihood under the model where Θ equals θ. It's important to point out that ξ isn't genuine prior information, it's just a (default) choice that's made to define the e-process. The mixing distributions may be relatively diffuse, hence my choice of ξ's variance as $v = 10$.

The examples that follow consider three different forms of partial prior information and the corresponding regularization of the above simple e-process. I must emphasize that *I'm not recommending off-the-shelf use of the regularizers presented below.* While I stand by my proposed framework, it's Your responsibility to determine what, if any, relevant prior information about Θ is available and how to quantify it. I'm in no position to say what You should or shouldn't believe about the quantity of interest in Your application. So, all that I'll do here is offer examples of prior knowledge that might be available in an application, to show how that information can be encoded as a possibility contour and converted into a regularizer and to demonstrate the effects that incorporation of this information can have on the results, in particular, as it pertains to efficiency.

I'll consider three different kinds of incomplete prior information. All three are indexed by a parameter K, although the meaning of K will be different in each use case.

Consequently, the corresponding regularized e-processes will not be comparable between the three types of prior information, but they will be comparable as K varies within the types. Here, $K \in \{0.1, 0.2, 0.4, 0.8\}$ and the three types of prior information I'll consider are below; first an explanation in words and then a mathematical description.

1. "You expect that $|\Theta| \leq K$." This is pretty clear in words but, mathematically, this corresponds to a credal set $\mathscr{Q}$ that contains exactly those distributions Q such that $\mathsf{E}^{\Theta \sim \mathsf{Q}}|\Theta| \leq K$. Of course, this includes certain Gaussian, uniform, and even some heavier-tailed priors. As shown in Dubois et al. (2004) and elsewhere, this prior information can be described mathematically via the possibility contour

$$q(\theta) = 1 \wedge K|\theta|^{-1}, \quad \theta \in \mathbb{T}.$$

2. "You're at most $100K/5\%$ sure that $|\Theta| > 2K$." This means that You can't rule out the possibility of $|\Theta| \leq 2K$, but that You judge the probability of the event $|\Theta| > 2K$ to be upper-bounded by $K/5$. Mathematically, this can be described easily by the possibility contour

$$q(\theta) = \tfrac{K}{5}\, 1(|\theta| > 2K) + 1(|\theta| \leq 2K), \quad \theta \in \mathbb{T}.$$

   I'm using the K-dependent weights only so that the different lines in the plots shown in Figure 3 don't overlap on a large part of the range of θ values.

3. "For each confidence level $c \in [0,1]$, You're $100(1-c)\%$ sure that Θ is in the central $1-c$ interval corresponding to a $\mathsf{N}(0,K)$ distribution." Note that this is different from, and considerably weaker than, a belief that Θ is $\mathsf{N}(0,K)$ distributed. In fact, this just means that $\mathsf{N}(0,K)$ is a (most diffuse) element of the prior credal set $\mathscr{Q}$. This is described mathematically by the probability-to-possibility transform of the dominating Gaussian distribution:

$$q(\theta) = \mathsf{P}^{\Theta \sim \mathsf{N}(0,K)}\{\mathtt{dnorm}_K(\Theta) \leq \mathtt{dnorm}_K(\theta)\} = 1 - \mathtt{pchisq}(\theta^2/K),$$

   where $\mathtt{dnorm}_K$ is the $\mathsf{N}(0,K)$ density and $\mathtt{pchisq}$ is the $\mathsf{ChiSq}(1)$ CDF.

The first two kinds of prior information are rather weak, only based on some very basic judgments about a first moment and the probability of a single event. The third kind of prior is much more thorough and nuanced, so the reader should be expecting that this kind of partial prior information will have much stronger regularization effect.

Figures 2, 3, and 4 plot the regularized (and unregularized) log-transformed e-process as a function of θ for the three types of priors, respectively, and for three different values of the observed sample mean $\bar{z}$ based on a sample of size $n = 5$:

- $\bar{z} = 0.25$ is pretty consistent with all the priors;
- $\bar{z} = 0.5$ is at least marginally inconsistent with the priors, and;
- $\bar{z} = 1$ is rather inconsistent with the priors.

The black line corresponds to the unregularized e-process, and the four colored lines correspond to the different values of K; the dashed horizontal line corresponds to $-\log 0.05 \approx 3$, which is the cutoff that determines the (regularized) e-process's 95% confidence interval. Not surprisingly, given that prior Types 1–2 are rather weak, the effect of regularization is difficult to detect. There is a small efficiency gain in Figures 2–3, more so in the latter, since the prior does more than strictly constrain Θ to an interval around the origin. But the more forceful prior Type 3 has a much stronger regularization effect. The take-away message is that, as expected, the stronger the prior information You have, the more efficient of inference (e.g., tighter confidence intervals) You can justify.

Another perspective on the efficiency gain concerns the number of sample points required before a false hypothesis can be rejected. Here I'll consider the hypothesis "$\Theta = 0.7$" and track the (log) regularized e-process's growth as the sample size increases. The fewer number of sample points required to reject the false hypothesis, the more efficient the procedure is. I'll use the same Type 3 prior as above, so both the data and the prior are inconsistent with the hypothesis. Figure 5 plots the path $n \mapsto \text{avg}\{\log \mathfrak{e}^{\text{reg}}(z^n, 0.7)\}$, where the average is taken over 1000 replications of the data-generating process. Clearly, the most informative prior ($K = 0.1$), makes it possible to detect the discrepancy of 0.7 much earlier, whereas the other priors, which don't offer as strong of regularization, take longer to reach the desired conclusion. Except for the weakest prior information ($K = 0.8$), regularization results in greater efficiency; see Section 7.

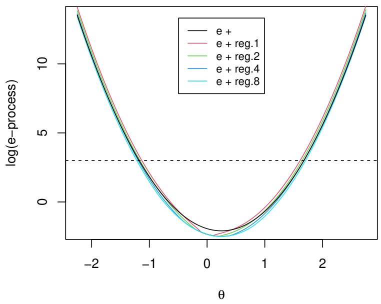
4 Regularized e-uncertainty quantification

4.1 Objective

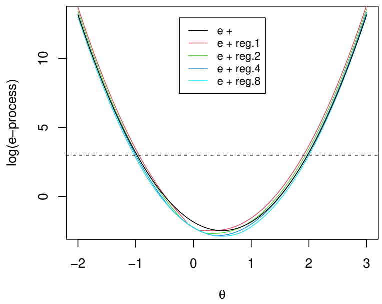
So far, I've treated *inference* as the construction of suitable test and confidence set procedures. In this section, however, I have a more ambitious goal reliable, data-driven, uncertainty quantification about the uncertain Θ. Following Fisher and others, my goal here is to uncover what Efron called the "Holy Grail of statistics."

It's a mathematical fact that uncertainty arising in statistical inference generally cannot be quantified reliably using ordinary probability theory. Indeed, the *false confidence theorem* (Balch et al. 2019) establishes that, without a genuine prior distribution, every probabilistic quantification of uncertainty—Bayes, fiducial, etc.—tends to assign high confidence, i.e., high posterior probability, to certain false hypotheses (Martin 2019, 2021a, 2024b). False confidence creates an unacceptable risk of unreliability, or "systematically misleading conclusions" (Reid and Cox 2015). Fortunately, well-developed alternatives to probability theory already exist, such as Dempster–Shafer theory (Dempster 2008; Shafer 1976), possibility theory (Dubois and Prade 1988), and the theory of lower previsions (Troffaes and de Cooman 2014; Walley 1991). *Inferential models* (IMs)—as originally developed in Martin and Liu (2013, 2015)—corresponds to a new framework for statistical reasoning, which is necessary to avoid false confidence, etc.

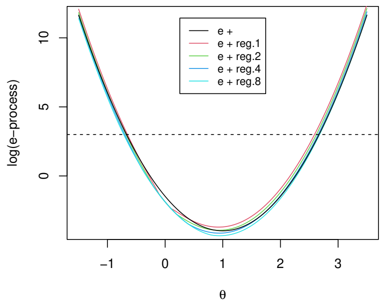
In this section I'll first give a brief overview of the recent formulation of *possibilistic IMs*, which is simple and more powerful. Then I'll discuss a new and insightful twist on the possibilistic IM formulation that involves e-processes (Martin 2024a). Partial prior knowledge can also be incorporated, leading to a framework of regularized e-process-based uncertainty quantification. The corresponding e-possibilistic IM and its reliability



(a) $\bar{z} = 0.25$

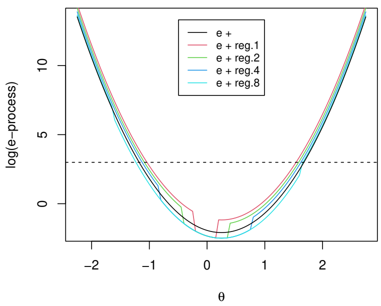


(b) $\bar{z} = 0.5$

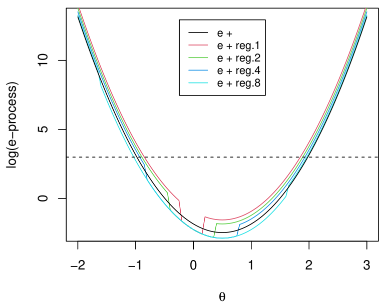


(c) $\bar{z} = 1$

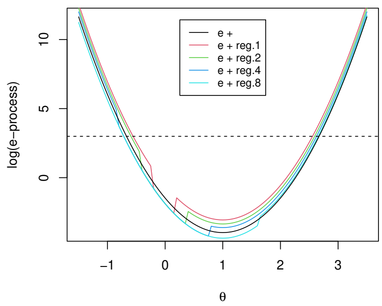
Figure 2: Plot of $\theta \mapsto \mathfrak{e}^{\text{reg}}(z^n, \theta)$ for three different data sets z^n based on prior Type 1.



(a) $\bar{z} = 0.25$

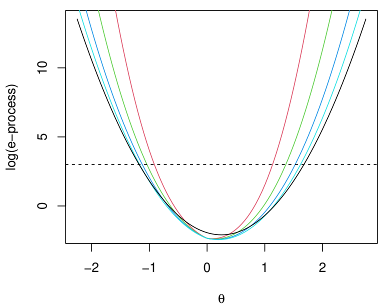


(b) $\bar{z} = 0.5$

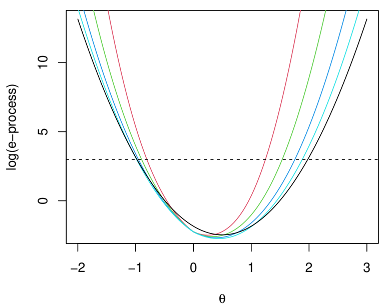


(c) $\bar{z} = 1$

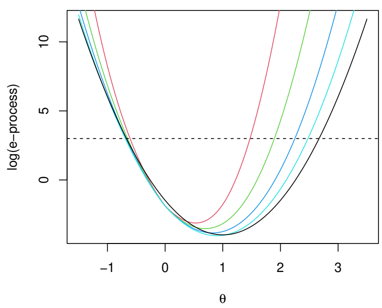
Figure 3: Plot of $\theta \mapsto \mathfrak{e}^{\text{reg}}(z^n, \theta)$ for three different data sets z^n based on prior Type 2.



(a) $\bar{z} = 0.25$



(b) $\bar{z} = 0.5$



(c) $\bar{z} = 1$

Figure 4: Plot of $\theta \mapsto \mathfrak{e}^{\text{reg}}(z^n, \theta)$ for three different data sets z^n based on prior Type 3.

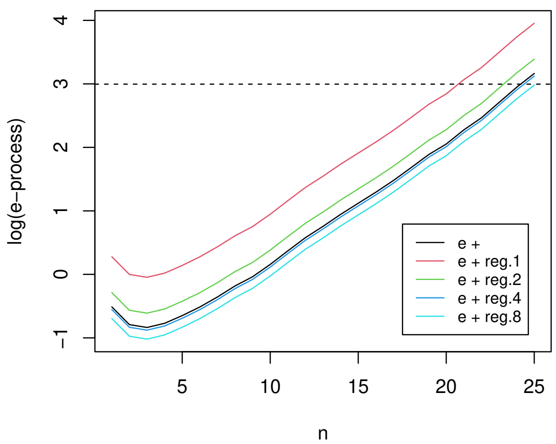


Figure 5: Plot of $n \mapsto \text{avg}\{\log \mathfrak{e}^{\text{reg}}(z^n, 0.7)\}$ for the prior of Type 3 when data are generated from a standard normal distribution.

properties are developed in Section 4.4, and certain unexpected-but-welcome behavioral consequences of this reliability in the context are established in Section 4.5.

4.2 Possibilistic IMs

My current efforts (Martin 2022b, 2023, 2024a) focus on the construction of *possibilistic IMs*, where the IM output takes the mathematical form of a possibility measure as reviewed briefly in Section 2.3. The specific proposal put forward in the above references starts with defining the IM's possibility contour, based on observed data $Z^n = z^n$, as

$$\pi_{z^n}(\theta) = \sup_{\omega: f(\omega)=\theta} \mathsf{P}_\omega\{r(Z^n, \theta) \leq r(z^n, \theta)\}, \quad \theta \in \mathbb{T},$$

where $r(z^n, \theta)$ provides a ranking of the parameter value θ in terms of its compatibility with z—large values indicate higher compatibility. For example, in Martin (2022b, 2023), I recommended taking r to be the relative profile likelihood

$$R(z^n, \theta) = \frac{\sup_{\omega: f(\omega)=\theta} L_{z^n}(\omega)}{\sup_\omega L_{z^n}(\omega)}, \quad \theta \in \mathbb{T},$$

with $\omega \mapsto L_{z^n}(\omega)$ the likelihood function corresponding to the "iid P_ω" model. Then the possibilistic IM's upper probability output is, as in (7), given by optimization:

$$\overline{\Pi}_{z^n}(H) = \sup_{\theta \in H} \pi_{z^n}(\theta), \quad H \subseteq \mathbb{T}. \tag{18}$$

The interpretation is that a small $\overline{\Pi}_{z^n}(H)$ means there's strong evidence in data z^n against the truthfulness of H. That is, a small upper probability assigned to H implies doubt and, therefore, I'd be inclined to "reject" H. But how small is "small"? Clearly some calibration of the IM's numerical output is needed in order for this line of reasoning to be reliable, i.e., to ensure that I don't systematically doubt true hypotheses or buttress false hypotheses. The possibilistic IM's calibration property is

$$\sup_{\omega: f(\omega) \in H} \mathsf{P}_\omega\{\overline{\Pi}_{Z^n}(H) \leq \alpha\} \leq \alpha, \quad \alpha \in [0,1], \quad H \subseteq \mathbb{T}. \tag{19}$$

In words, (19) says that the IM assigning small upper probability values to true hypotheses about Θ is itself a small probability event. Since there's an explicit link between the two notions of "small," this offers the calibration needed to avoid systematically erroneous inferences. Compare this to what Walley (2002) calls the "fundamental frequentist principle." The analogous property concerning the IM's lower probabilities—recall the conjugacy relationship, $\underline{\Pi}_{z^n}(H) = 1 - \overline{\Pi}_{z^n}(H^c)$, from Section 2.3—reads as follows:

$$\sup_{\omega: f(\omega) \notin H} \mathsf{P}_\omega\{\underline{\Pi}_{Z^n}(H) \geq 1 - \alpha\} \leq \alpha, \quad \alpha \in [0,1], \quad H \subseteq \mathbb{T}.$$

This latter result says that the IM assigning large lower probability—or "confidence"—to false hypotheses is a small-probability event; now it should be clear why I say that the IM is safe from the risk of false confidence.

It's based on the property (19) that I say the IM is *valid* and, in turn, that uncertainty quantification based on the IM output is reliable. From here, it's straightforward to establish that test and confidence procedures derived from the IM output achieve the usual frequentist error rate control guarantees. But there's more to the IM output than the statistical procedures derived from it, and the interested reader can check out, e.g., Martin (2022b) and Cella and Martin (2023) for more details.

4.3 Connection to e-processes

The discussion above focused on the case of a fixed sample size. It's well-known, however, that the fixed-n sampling distribution properties—such as validity in (19)—generally fail when the stopping rule N is data-dependent. If the stopping rule N that's used is *known*, then this can be easily incorporated into the IM construction such that validity is achieved. The trouble, of course, is that the N employed in the data-collection process is often *unknown*, e.g., the original investigators might not explain all the study details in their paper. Martin (2024a) suggested a conceptually simple work-around that interprets N as a "nuisance parameter." Then the general rules in Martin (2023) for handling nuisance parameters immediately suggest a new IM contour

$$\pi_{z^n}(\theta) = \sup_N \sup_{\omega: f(\omega) = \theta} \mathsf{P}_\omega\{r(Z^N, \theta) \leq r(z^n, \theta)\}, \quad \theta \in \mathbb{T},$$

where the outermost supremum is over all the stopping rules in consideration (that are consistent with the observation "$N = n$"). Of course, direct computation of the right-hand side above can be challenging, depending on the complexity of the set of stopping rules in consideration. It's here that e-processes come in handy.

Recall that the ranking function r in the possibilistic IM construction is quite flexible. One alternative to the proposal above is to take r as the reciprocal of an e-process:

$r(z^n, \theta) = \mathfrak{e}_\theta(z^n)^{-1}$. In that case, the usual Ville's inequality gives

$$\begin{aligned}
\pi_{z^n}(\theta) &= \sup_N \sup_{\omega: f(\omega)=\theta} \mathsf{P}_\omega\{r(Z^N, \theta) \leq r(z^n, \theta)\} \\
&= \sup_N \sup_{\omega: f(\omega)=\theta} \mathsf{P}_\omega\{\mathfrak{e}_\theta(Z^N) \geq \mathfrak{e}_\theta(z^n)\} \\
&\leq 1 \wedge \mathfrak{e}_\theta(z^n)^{-1}.
\end{aligned}$$

Set $\pi^{\mathfrak{e}}_{z^n}(\theta) = 1 \wedge \mathfrak{e}_\theta(z^n)^{-1}$ to be the upper bound, so that

$$\pi_{z^n}(\theta) \leq \pi^{\mathfrak{e}}_{z^n}(\theta) := 1 \wedge \mathfrak{e}_\theta(z^n)^{-1}, \quad \theta \in \mathbb{T}.$$

This upper bound corresponds to what Grünwald (2023) calls his capped *e-posterior*. But there are two further observations about this bound that are worth noting. First, the bound itself is (typically[3]) a possibility contour. Consequently, there's a corresponding possibilistic IM, with coherent upper probability $\overline{\Pi}^{\mathfrak{e}}_{z^n}$ defined via optimization as in (18). This determines an *e-possibilistic IM* for uncertainty quantification about Θ. Second, that $\pi^{\mathfrak{e}}_{z^n}$ is an upper bound of the contour π_{z^n}, and that the latter determines an anytime valid IM, immediately implies that the IM corresponding to the former is anytime valid too. This discussion is summarized in the following theorem.

Theorem 2. *Given an e-process $\mathfrak{e}$, the corresponding e-possibilistic IM with upper probability determined by optimization of the contour $\pi^{\mathfrak{e}}_{z^n}$, i.e.,*

$$\overline{\Pi}^{\mathfrak{e}}_{z^n}(H) = \sup_{\theta \in H} \pi^{\mathfrak{e}}_{z^n}(\theta), \quad H \subseteq \mathbb{T},$$

is anytime valid in the sense that

$$\sup_{\omega: f(\omega) \in H} \mathsf{P}_\omega\{\overline{\Pi}^{\mathfrak{e}}_{Z^N}(H) \leq \alpha\} \leq \alpha, \quad \text{all } \alpha \in [0,1], \text{ all } N, \text{ all } H \subseteq \mathbb{T}. \tag{20}$$

Proof. Since the IM with contour $\pi^{\mathfrak{e}}$ dominates that with contour π defined earlier, the anytime validity of the latter implies that of the former. For concreteness, however, I'll give a direct proof of anytime validity of $\overline{\Pi}^{\mathfrak{e}}$.

For any data set z^n, the IM's possibility measure output $H \mapsto \overline{\Pi}_{z^n}(H)$ is monotone. Therefore, if ω is such that $f(\omega) \in H$, then

$$\overline{\Pi}^{\mathfrak{e}}_{z^n}(H) \geq \overline{\Pi}^{\mathfrak{e}}_{z^n}(\{f(\omega)\}) = \pi^{\mathfrak{e}}_{z^n}(f(\omega)),$$

and, consequently, for any $\alpha \in [0,1]$,

$$\overline{\Pi}_{z^n}(H) \leq \alpha \implies \pi^{\mathfrak{e}}_{z^n}(f(\omega)) \leq \alpha \iff \mathfrak{e}_{f(\omega)}(z^n) \geq \alpha^{-1},$$

where the right-most property is by definition of the IM's contour function in terms of the e-process's reciprocal. It follows that

$$\sup_{\omega: f(\omega) \in H} \mathsf{P}_\omega\{\overline{\Pi}^{\mathfrak{e}}_{Z^N}(H) \leq \alpha\} \leq \sup_{\omega: f(\omega) \in H} \mathsf{P}_\omega\{\mathfrak{e}_{f(\omega)}(Z^N) \geq \alpha^{-1}\} \leq \alpha,$$

with the last inequality due to (2), thus completing the proof. □

[3] $\pi^{\mathfrak{e}}_{z^n}$ fails to be a possibility contour at a given z^n if and only if $\mathfrak{e}_\theta(z^n)$ is strictly greater than 1 for all θ. But e-processes have expected value upper-bounded by 1, so it'd be exceptionally rare, although not impossible, for z^n to not be particularly compatible with any θ, so that $\mathfrak{e}_\theta(z^n)$ is everywhere greater than 1 and, hence, $\pi^{\mathfrak{e}}_{z^n}$ is everywhere below 1. This doesn't affect the statistical properties, only the interpretation of the IM; see Section 4.5.1 below.

As discussed in Section 4.2, this calibration property is important because, without it, there'd be no meaningful justification for any particular interpretation of the numerical values an e-possibilistic IM assigns to various inputs (z^n, H). But in light of Theorem 2, the same reliability-guaranteeing "no-false-confidence" statement made above for the simple possibilistic IM also holds here for the new e-possibilistic IM.

One important consequence of anytime validity is that one can derive test procedures from the IM and these will inherit, automatically, the desired frequentist properties. Indeed, for a given hypothesis $H \subset \mathbb{T}$ about the uncertain Θ, and a given significance level $\alpha \in [0,1]$, the test procedure

$$\text{reject "}\Theta \in H\text{" if and only if } \overline{\Pi}^{\mathfrak{e}}_{Z^N}(H) \leq \alpha$$

controls the frequenist Type I error rate at the specified level α. This result is "obvious" because $\overline{\Pi}^{\mathfrak{e}}_{Z^N}(H)$ can be small if and only if $\mathfrak{e}_\theta(Z^N)$ is large for all $\theta \in H$: in particular,

$$\overline{\Pi}^{\mathfrak{e}}_{Z^N}(H) \leq \alpha \iff \inf_{\theta \in H} \mathfrak{e}_\theta(Z^N) \geq \alpha^{-1}.$$

If the hypothesis is true, i.e., if $\Theta = f(\Omega) \in H$, then the right-most event in the above display implies $\mathfrak{e}_{f(\Omega)}(Z^N) \geq \alpha^{-1}$, which, by Ville's inequality (2), has P_Ω-probability no more than α, uniformly over stopping times N.

For the case of vacuous prior information, there's a consequence of anytime validity (20) that deserves mention; this property was actually used in the proof of Theorem 2. For the more general prior information cases to be considered below, this new property is actually stronger than (20), but here the two are equivalent.

Corollary 1. *Given an e-process $\mathfrak{e}$, the corresponding e-possibilistic IM is strongly anytime valid in the sense that its contour satisfies*

$$\sup_{\omega \in \mathbb{O}} \mathsf{P}_\omega\big[\pi^{\mathfrak{e}}_{Z^N}\{f(\omega)\} \leq \alpha\big] \leq \alpha, \quad \textit{all } \alpha \in [0,1], \textit{ all } N. \tag{21}$$

Proof. Apply Theorem 2 with H running over all the singletons in $\mathbb{T} = f(\mathbb{O})$. □

Two comments related to Corollary 1 are in order. First, similar to the points about the construction of anytime valid testing procedures, it's straightforward to construct confidence sets too. Given a significance level $\alpha \in [0,1]$, a $100(1-\alpha)\%$ confidence set derived from the e-process-based IM is the α-level set defined by the contour:

$$C^{\mathfrak{e}}_\alpha(Z^N) = \{\theta \in \mathbb{T} : \pi^{\mathfrak{e}}_{Z^N}(\theta) > \alpha\}.$$

Then strong anytime validity implies that this is, indeed, a genuine anytime valid confidence set in the sense that

$$\sup_{\omega \in \mathbb{O}} \mathsf{P}_\omega\{C^{\mathfrak{e}}_\alpha(Z^N) \not\ni f(\omega)\} \leq \alpha, \quad \text{all stopping times } N.$$

But, by definition of $\pi^{\mathfrak{e}}$, it's easy to see that $C^{\mathfrak{e}}_\alpha$ can be re-expressed as

$$C^{\mathfrak{e}}_\alpha(Z^N) = \{\theta \in \mathbb{T} : \mathfrak{e}_\theta(Z^N) < \alpha^{-1}\},$$

which the reader will recognize as the (unregularized) confidence set determined by the e-process $\mathfrak{e}$. Then the above non-coverage probability bound follows immediately from the relevant properties (i.e., Ville's inequality) of $\mathfrak{e}$ presented above.

Second, to see why this strong anytime validity (21) is "stronger" than anytime validity (20), note that monotonicity of possibility measures implies

$$\overline{\Pi}^{\mathfrak{e}}_{Z^N}(H) \geq \overline{\Pi}^{\mathfrak{e}}_{Z^N}(\{f(\omega)\}) \equiv \pi^{\mathfrak{e}}_{Z^N}(f(\omega)) \quad \text{for all } H \text{ with } H \ni f(\omega). \tag{22}$$

Interestingly, the above holds even for hypotheses H that depend on data Z^N in some way, e.g., if an adversary with access to the data Z^N is trying to dupe the statistician by selecting a "most difficult" hypothesis H post hoc. This suggests that strong anytime validity is equivalent to a *uniform-in-hypotheses* version of anytime validity, and the following corollary states this explicitly; see, also, Cella and Martin (2023). Note the difference between (23) below and (20) above: the former has varying H on the *inside* of the probability statement whereas the latter has varying H on the *outside*.

Corollary 2. *The e-process-based possibilistic IM constructed above is uniformly anytime valid in the sense that*

$$\sup_{\omega \in \mathbb{O}} \mathsf{P}_\omega\{\overline{\Pi}^{\mathfrak{e}}_{Z^N}(H) \leq \alpha \text{ for some } H \text{ with } H \ni f(\omega)\} \leq \alpha, \quad \text{all } \alpha \in [0,1], \text{ all } N. \tag{23}$$

Proof. The proof is based on the following observation:

$$\overline{\Pi}^{\mathfrak{e}}_{Z^N}(H) \leq \alpha \text{ for some } H \text{ with } H \ni f(\omega) \iff \pi^{\mathfrak{e}}_{Z^N}(f(\omega)) \leq \alpha. \tag{24}$$

The "$\Longleftarrow$" direction is obvious since $H := \{f(\omega)\}$ is a hypothesis that contains $f(\omega)$. The "$\Longrightarrow$" is similarly obvious by the monotonicity property as stated in (22). Since Corollary 1 says the right-most event in the above display has probability no more than α, uniformly in ω and in N, the same must be true of the equivalent left-most event. □

The take-away message is that the e-possibilistic IM constructed here offers reliable, anytime valid uncertainty quantification that's safe not just against arbitrary-but-fixed hypotheses as in (20), but also against possibly adversarial or otherwise data-dependent hypotheses as in (23). This is important for at least two reasons:

- Investigators feeling "publish-or-perish" pressures might succumb to temptations to explore for hypotheses that are incompatible with their data, to secure a statistically significant result. A commitment to using the e-possibilistic IM for assessing hypotheses prevents these sociological temptations from causing harm to science.
- As Mayo (2018) articulates, investigators yearn for more from their data than null hypothesis significance testing. Conclusions about the "null hypothesis" mark the beginning of the scientific investigation, not the end, so it's imperative that investigators can also probe however they like for hypotheses that might be compatible with their data, without risking unreliability. Multiplicity corrections can't accommodate this kind of probing; for more on this, see Cella and Martin (2023).

Up to now, I've not commented on why the *possibilistic* IM formulation is specifically appropriate. To me, the most compelling justification comes from the uniform validity result in Corollary 2. The equivalence (24) in the proof of Corollary 2 holds for all imprecise-probabilistic IMs, but what I'm calling the "contour" $\pi^{\mathfrak{e}}$ has properties it the possibilistic cases that it doesn't have under other imprecise-probabilistic formulations. For a generic, data-dependent upper probability $\overline{\Pi}_{Z^N}$ on $\mathbb{T}$, the function $\theta \mapsto \overline{\Pi}_{Z^N}(\{\theta\})$ doesn't completely determine $\overline{\Pi}_{Z^N}$ and, moreover, it can happen that $\theta \mapsto \overline{\Pi}_{Z^N}(\{\theta\})$ is always small, no matter what Z^N is. Then the probability (with respect to the sampling distribution of Z^N) that it's less than some $\alpha < 1$ could be large—perhaps even equal to 1. It's unique to the possibilistic framework that the contour fully determines the upper probability and, moreover, takes values arbitrarily close to 1. Without this special structure, strong validity and, hence, uniform validity can't be attained. If it could be attained by some other, non-possibilistic IM construction, then Lemma 1 in Martin (2022b) says that there's a possibilistic IM that's no worse in terms of efficiency. The conclusion is that, if strong validity and the safety it offers is a priority, which it is to me, then the possibilistic formulation is without loss of generality/efficiency.

4.4 Partial priors and regularization

Section 3 showed how prior knowledge can be encoded as a regularizer and then combined with a given e-process such that the resulting regularized e-process is anytime valid in a relaxed sense. The next step is to flip this regularized e-process into a regularized e-possibilistic IM and to establish the corresponding anytime validity properties in the context of uncertainty quantification. The remainder of this subsection details these next steps. Section 4.5 below deals with some more nuanced behavioral properties.

Following the developments in Section 4.3 above, define the regularized e-possibilistic IM contour function for Θ, given z^n, as

$$\pi_{z^n}^{\mathfrak{e}\times\rho}(\theta) = 1 \wedge \mathfrak{e}^{\text{reg}}(z^n, \theta)^{-1}, \quad \theta \in \mathbb{T}. \tag{25}$$

The corresponding possibility lower and upper probabilities are denoted as $\underline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}$ and $\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}$, respectively, with the latter defined via optimization and the former via conjugacy, as usual. By definition of $\pi^{\mathfrak{e}\times\rho}$ above, the level sets

$$C_\alpha^{\mathfrak{e}\times\rho}(z^n) = \{\theta \in \mathbb{T} : \pi_{z^n}^{\mathfrak{e}\times\rho}(\theta) > \alpha\}, \quad \alpha \in [0,1],$$

are identical to the regularized e-process-based confidence regions C_α^{reg} in Section 3.6. Therefore, the sets $C_\alpha^{\mathfrak{e}\times\rho}$ inherit the same non-coverage upper probability bound as in (17), hence can be referred to as (generalized) "anytime valid confidence regions."

Theorem 3 below generalizes the strong validity property in Corollary 1 for the unregularized e-possibilistic IM. It's an immediate consequence of Theorem 1 and the definition of $\pi^{\mathfrak{e}\times\rho}$. Recall the upper joint distribution $\overline{\mathbf{P}}$ for (Z^N, Ω), defined in (14) above.

Theorem 3. *Given an e-process $\mathfrak{e}$ and regularizer ρ, the corresponding regularized e-possibilistic IM with contour* (25) *is strongly anytime valid in the sense that it satisfies*

$$\overline{\mathbf{P}}\big[\pi_{Z^N}^{\mathfrak{e}\times\rho}\{f(\Omega)\} \le \alpha\big] \le \alpha, \quad \text{all } \alpha \in [0,1], \text{ all } N. \tag{26}$$

This is a generalization of Corollary 1 because, if the prior information is vacuous and $\rho \equiv 1$, then $\pi_{Z^N}^{\mathfrak{e}\times\rho} \equiv \pi_{Z^N}^{\mathfrak{e}}$ and the upper joint distribution $\overline{\mathbf{P}}$ reduces to the supremum of the P_ω-probabilities over all $\omega \in \mathbb{O}$ as in (21).

Corollary 3. *For the same regularized e-possibilistic IM considered in Theorem 3, the following hold for all thresholds $\alpha \in [0,1]$ and all stopping times N:*

$$\overline{\mathbf{P}}\big\{\textit{there exists } H \textit{ with } H \ni f(\Omega) \textit{ and } \overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(H) \leq \alpha\big\} \leq \alpha$$
$$\overline{\mathbf{P}}\big\{H \ni f(\Omega) \textit{ and } \overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(H) \leq \alpha\big\} \leq \alpha, \quad H \subseteq \mathbb{T}.$$

The interpretation is, as before, in terms of reliability. The second line says that, for any fixed H, the upper probability that H is true *and* the IM assigns it small upper probability is small. The first line is stronger, it says that the probability the IM assigns small upper probability to any true hypothesis about $f(\Omega)$ is small. To make sense of this, suppose data z^n is observed and, for a relevant hypothesis, it happens that $\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H)$ is smaller than a threshold that You deem to be "sufficiently small." Then Corollary 3 offers justification for You to draw the inference that H is false—if it were true, then it would be a "rare event" that the IM assigned it such a small upper probability.

I'll end this subsection on a slightly different note. The following comments apply to both the regularized and unregularized e-possibilistic IM, but I'll focus attention on the regularized version since the unregularized version is a special case. Recall that e-processes are typically likelihood ratios of some sort. In that case, the IM output $(\underline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}, \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho})$ only depends on the data z^n through the likelihood function and no other specifics concerning a "model" are used for drawing inferences; e.g., the threshold is α not some α-quantile of a model-dependent sampling distribution. This implies that the e-possibilistic IM satisfies both the *likelihood principle* (e.g., Basu 1975; Berger and Wolpert 1984; Birnbaum 1962) and the strong frequentist-like properties derived above. This is remarkable because the likelihood principle is Bayesians' turf:

> *I, myself, came to take... Bayesian statistics... seriously only through recognition of the likelihood principle.* (Savage—discussion of Birnbaum 1962)

Here I've shown that it's *possible* to quantify uncertainty in a Bayesian-like way that satisfies the likelihood principle without sacrificing frequentist-like reliably.

4.5 From uncertainty quantification to behavior

4.5.1 Anytime validity implies no-sure-loss (and more)

When data $Z^N = z^n$ is fixed, the e-possibilistic IM typically defines an imprecise probability model where $\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}$ has the mathematical properties of a possibility measure. Provided that the function $\theta \mapsto \pi_{z^n}^{\mathfrak{e}\times\rho}(\theta)$ defined in (25) isn't bounded away from 1, which is typically the case, the function $H \mapsto \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H)$ determined by maximizing a possibility contour function over the set $H \subseteq \mathbb{T}$ as in (7) is a genuine possibility measure. This implies that its credal set is non-empty and, in turn, that the IM output is *coherent* in the sense of Walley (1991, Sec. 2.5). For the present purposes, it's enough to understand coherence as a stronger version of *no-sure-loss*, which can be described as follows. Suppose beliefs concerning the uncertain value Θ are assessed via the buying/selling prices they consider acceptable for certain gambles about Θ. Let z^n be one of the aforementioned typical data sets and let Your buying/selling prices be determined by the IM output $(\underline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}, \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho})$

according to the interpretation in (4). Then coherence of Your IM implies that

$$\sup_{\theta \in \mathbb{T}} \sum_{k=1}^{K} \{\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H_k) - 1(\theta \in H_k)\} \geq 0 \quad \text{for all } K \text{ and } (H_1, \ldots, H_K) \text{ combos.}$$

To get the intuition, suppose the above condition fails. Then there exists a combination K and $(H_1, \ldots, H_K)$, and a sufficiently small $\delta > 0$, such that

$$\sup_{\theta \in \mathbb{T}} \sum_{k=1}^{K} \bigl[\{\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H_k) + \delta\} - 1(\theta \in H_k)\bigr] < 0.$$

Since $\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H_k)$ is, by definition, Your infimum selling price for the gamble $1(\Theta \in H_k)$, the transactions where You accept payment of $\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H_k) + \delta$ dollars for $\$1(\Theta \in H_k)$ for each k are all acceptable to You. But then the above display reveals a troubling result: somehow, by only making transactions that are acceptable *a priori*, You end up with negative total earnings regardless of what value the uncertain Θ takes on. This indicates a severe shortcoming in Your pricing scheme; fortunately, the e-possibilistic IM is typically free of this internal inconsistency.

I said "typically" several times in the above paragraph, and the explanation here is exactly the same in Section 4.3 when this matter first came up. The IM output $\overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}$ would fail to be a possibility measure if and only if $\pi_{z^n}^{\mathfrak{e}\times\rho}$ defined in (25) was bounded away from 1 on $\mathbb{T}$; I've been calling that function a "possibility contour" but that's only legitimate if $\sup_{\theta \in \mathbb{T}} \pi_{z^n}^{\mathfrak{e}\times\rho}(\theta) = 1$. For $\pi_{z^n}^{\mathfrak{e}\times\rho}(\theta)$ to be strictly less than 1, and hence the IM output determines an incoherent imprecise probability, would require that $\theta \mapsto \mathfrak{e}^{\text{reg}}(z^n, \theta)$ also be bounded strictly greater than 1. But (15) establishes that $\overline{\mathbf{P}}(\mathfrak{e}^{\text{reg}}) \leq 1$, i.e., $\mathfrak{e}^{\text{reg}}(Z^N, \Theta)$ "tends" to be less than 1, so a data set z^n could indeed be called atypical if it were such that $\mathfrak{e}^{\text{reg}}(z^n, \theta)$ were strictly greater than 1 for all θ.

In addition the fixed-data behavioral considerations, it's natural to interpret $(\mathscr{Q}, z^n) \mapsto (\underline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}, \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho})$ as a rule by which "prior" information is *updated* in light of data (z^n) to a "posterior" quantification of uncertainty. More familiar notions of imprecise-probabilistic updating include generalized Bayes rule (Miranda and de Cooman 2014; Walley 1991) and Dempster's rule (e.g., Cuzzolin 2021; Shafer 1976). With this "updating rule" interpretation comes further questions about the IM's ability to protect You from sure loss, etc. What's different here is that there's a temporal component: can I force You into transactions such that, no matter what data is observed, You lose money? If so, then there's a serious issue with Your assessments. Mathematically, this *sure loss* property—see, e.g., Walley (1991, Sec. 2.4.1) and Gong and Meng (2021, Def. 3.3)—corresponds to existence of a hypothesis $H \subset \mathbb{T}$ such that

$$\sup_{z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) < \underline{\mathsf{Q}}(H) \quad \text{or} \quad \inf_{z^n} \underline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) > \overline{\mathsf{Q}}(H). \tag{27}$$

For intuition, consider the first of the above two inequalities. If this inequality holds, then, for any pair of positive numbers (ε, δ), You'd be willing to buy the gamble $\$1(\Theta \in H)$ from me for $\$\{\underline{\mathsf{Q}}(H) - \varepsilon\}$ and then sell the same gamble back to me, after observing z^n, for $\$\{\sup_{z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) + \delta\}$. No matter whether $\Theta \in H$ or $\Theta \notin H$, the payoff You receive from this sequence of transactions is

$$\Big\{\sup_{z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) + \delta\Big\} - \{\underline{\mathsf{Q}}(H) - \varepsilon\} = \underbrace{\Big\{\sup_{z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) - \underline{\mathsf{Q}}(H)\Big\}}_{<0, \text{ by (27)}} + (\varepsilon + \delta).$$

Since both individual transactions are acceptable to You, and there exists pairs (ε, δ) such that You net payoff is strictly negative, it follows that You can be made a sure loser.

Fortunately, such extremes are easily prevented. In fact, the possibilistic IM avoids an even less severe internal inconsistency, which I call *one-sided contraction*:

$$\sup_{z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) < \overline{\mathsf{Q}}(H) \quad \text{or} \quad \inf_{z^n} \underline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) > \underline{\mathsf{Q}}(H). \tag{28}$$

One-sided contraction is less concerning than sure-loss, but still problematic. To see this, suppose that the first of the two inequalities in (28) holds for a given H. If I want to buy $\$1(\Theta \in H)$ from You *a priori*, then it's apparent that I can wait until data z^n is revealed and purchase the gamble for a lower price. This doesn't imply that You lose money, only that You're systematically giving up opportunities to earn more money. This can't happen in the familiar, precise Bayesian case: since the expected value of a posterior probability is the prior probability, it's impossible for each posterior/conditional probability to be less than the prior/marginal probability.

The stronger notion of (two-sided) *contraction* corresponds to replacing the "or" in the above display with "and," and the dual notion of *dilation* corresponds to flipping both inequalities and replacing "or" with "and." Both contraction and dilation are problematic in their own respects, but contraction is generally more serious. Of course, if an e-possibilistic IM avoids one-sided contraction, as the next theorem establishes, then it necessarily avoids both sure-loss and (two-sided) contraction.

Theorem 4. *The e-possibilistic IM avoids one-sided contraction, i.e., there are no H such that* (28) *holds for the updating rule* $(\mathscr{Q}, z^n) \mapsto (\underline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}, \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho})$.

Proof. Fix any hypothesis $H \subseteq \mathbb{T}$. I'll focus on proving that the first inequality in (28) doesn't hold, i.e., that

$$\sup_{n,z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) \geq \overline{\mathsf{Q}}(H). \tag{29}$$

The version involving lower probabilities is proved similarly. To start, note that

$$\inf_{n,z^n} \mathfrak{e}_\theta(z^n) \leq \sup_N \sup_{\omega: f(\omega)=\theta} \mathsf{E}_\omega\{\mathfrak{e}_\theta(Z^N)\} \leq 1,$$

where the first inequality is because an average is never smaller than the minimum, and the second inequality by (1). That is, for any θ, there exists (n, z^n) such that $\mathfrak{e}_\theta(z^n)$ is no more than 1. Now write out the left-hand of (29) as follows:

$$\sup_{n,z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) = \sup_{n,z^n} \sup_{\theta\in H} \pi_{z^n}^{\mathfrak{e}\times\rho}(\theta) = 1 \wedge \Big\{\inf_{n,z^n} \inf_{\theta\in H} \rho(\theta)\, \mathfrak{e}_\theta(z^n)\Big\}^{-1}.$$

It follows from the "first observation" above that

$$\sup_{n,z^n} \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}(H) \geq 1 \wedge \Big\{\inf_{\theta\in H} \rho(\theta)\Big\}^{-1}.$$

Since $\rho(\theta) \geq \{\inf_{\theta \in H} \rho(\theta)\} \times 1(\theta \in H)$, it's easy to see that

$$\Bigl\{\inf_{\theta \in H} \rho(\theta)\Bigr\} \times \overline{\mathsf{Q}}(H) \leq \overline{\mathsf{Q}}\, \rho \leq 1,$$

where the first inequality is by monotonicity of the upper expectation and the second by definition of the regularizer ρ. Plugging this bound into that above gives

$$\sup_{n, z^n} \overline{\Pi}_{z^n}^{\mathfrak{e} \times \rho}(H) \geq 1 \wedge \overline{\mathsf{Q}}(H) = \overline{\mathsf{Q}}(H),$$

which completes the proof of the theorem. □

An even stronger internal consistency property, a similarly temporal version of *coherence*, might be desired. Walley (1991, Sec. 6.5.2) presents two conditions that are necessary and sufficient for the kind of coherence that I have in mind here. The first of those conditions is precisely no-one-sided-contraction, so it'd not be out of line to say that the e-possibilistic IM is *half-coherent* based on the result in Theorem 4. The second of Walley's conditions is more complicated and not worth writing down here. Since there are cases in which the only coherent prior-to-posterior updating rule is generalized Bayes (e.g., Walley 1991, Sec. 6.5.4), it's not possible for the proposed IM to be coherent in general. There are cases, however, where the IM is coherent: in the fixed sample size case, I argued in Martin (2022a, Sec. 3.3) that coherence holds provided that "$z \mapsto \overline{\Pi}_z(H)$" is continuous for each H. The fact that the sample size here isn't fixed creates some technical complications, so I'll postpone this investigation into coherence for now.

4.5.2 Anytime valid decision-making

For a generic action space $\mathbb{A}$, let $\ell_a(\theta) \geq 0$ represent the (non-negative[4]) loss associated with taking action $a \in \mathbb{A}$ when the relevant state of the world is $\theta \in \mathbb{T}$. Under ideal, perfect-information settings, where the true state of the world Θ is known, the best decision would correspond to choosing $a^\star$ to minimize the loss $a \mapsto \ell_a(\Theta)$; note that the case of maximizing utility can be recovered from this by defining utility as, say, $u_a = -\ell_a$. Unfortunately, Θ is uncertain, so the aforementioned minimization exercise can't be carried out. But an e-possibilistic IM $(\underline{\Pi}_{z^n}^{\mathfrak{e} \times \rho}, \overline{\Pi}_{z^n}^{\mathfrak{e} \times \rho})$ is available for data-driven uncertainty quantification, so I propose to mimic the von Neumann–Morganstern proposal and evaluate the corresponding upper and lower expected loss:

$$\overline{\Pi}_{z^n}^{\mathfrak{e} \times \rho}(\ell_a) = \int_0^1 \Bigl\{ \sup_{\theta : \pi_{z^n}^{\mathfrak{e} \times \rho}(\theta) \geq s} \ell_a(\theta) \Bigr\}\, ds \qquad \text{and} \qquad \underline{\Pi}_{z^n}^{\mathfrak{e} \times \rho}\, \ell_a = -\overline{\Pi}_{z^n}^{\mathfrak{e} \times \rho}(-\ell_a).$$

Since the loss is non-negative, the upper and lower expectations surely exist, though they could be $+\infty$ and $-\infty$, respectively, at least for some a. They're both finite at a given a if $\theta \mapsto \ell_a(\theta)$ is previsable with respect to $(\underline{\Pi}_{z^n}^{\mathfrak{e} \times \rho}, \overline{\Pi}_{z^n}^{\mathfrak{e} \times \rho})$, e.g., if the loss is bounded. In any case, these two functionals together determine the IM's assessment of the quality of

[4]Non-negativity of the loss is not necessary, it's just that the theory below is more complicated otherwise. If the loss is bounded away from $-\infty$, then a positive constant can be added to it to make it non-negative without affecting the shape, the expected loss minimizer, etc.

action a. Since the goal is to make the loss small in some sense, it's reasonable to define the "best" action, given data z^n, as the minimizer of the upper expected loss:

$$\hat{a}(z^n) := \arg\min_{a \in \mathbb{A}} \overline{\Pi}_{z^n}^{\mathfrak{e} \times \rho}(\ell_a). \tag{30}$$

Of course, if there are multiple minimizers, then any one could be chosen. Further details can be found in Denœux (2019) and Martin (2021b).

For an illustration, reconsider the Gaussian example from Section 3.7. Recall the Savage–Dickey e-process given above,

$$\mathfrak{e}_\theta(z^n) = (nv+1)^{1/2} \exp\bigl\{-\tfrac{n}{2}(\theta - \bar{z}_n)^2 + \tfrac{1}{2}\bigl(\tfrac{n}{nv+1}\bigr)\, \bar{z}_n^2\bigr\}, \quad \theta \in \mathbb{T},$$

where $v > 0$ is a variance hyperparameter that can take any value; in my numerical examples here and above, I'll take $v = 10$. Starting with vacuous prior information, so that $\rho \equiv 1$, the e-possibilistic IM's contour $\pi_{z^n}^{\mathfrak{e}}$ is plotted in Figure 6. I take the loss function to be squared error: $\ell_a(\theta) = (\theta - a)^2$, also plotted in Figure 6 for several values of a. It's easy to see, based on the symmetry of both the e-process and the loss, that the minimizer of the upper expected loss is $\hat{a}(z^n) = n^{-1} \sum_{i=1}^n z_i$, the sample mean—no surprise. For reasons that will be apparent below, it's of interest to evaluate that upper expected loss at the aforementioned minimizer:

$$\begin{aligned} \overline{\Pi}_{z^n}^{\mathfrak{e}}\, \ell_{\hat{a}(z^n)} &= \int_0^1 \Bigl\{ \sup_{\theta: \pi_{z^n}^{\mathfrak{e}}(\theta) \geq s} \ell_{\hat{a}(z^n)}(\theta) \Bigr\}\, ds \\ &= n^{-1}\bigl\{2 + \log(nv+1) + \bigl(\tfrac{n}{nv+1}\bigr)\bar{z}_n^2\bigr\}. \end{aligned}$$

(The detailed calculation is presented in Appendix A.) Note that, for fixed v, as a function of n, the asymptotic order of the upper expected loss is $n^{-1} \log n$. This is within a log-factor of the usual rate n^{-1} associated with the Bayes or frequentist minimax risks—the log-factor is the price one pays for the benefits of anytime validity.

Sticking with the same Gaussian illustration, but this time consider regularization corresponding to the Type 3 partial prior described in Section 3.7, the strongest of the three partial priors consider there. Figure 7(a) shows the possibility contours for the four different values of the partial prior hyperparameter as in Section 3.7, along with the vacuous prior contour (same as in Figure 6). As expected, the regularized contours are a bit more concentrated than the unregularized contour, with more concentration corresponding to the stronger prior knowledge. From this picture, it's pretty clear that the regularized e-possibilistic IM's risk-minimizing action is closer to 0 than that for the unregularized possibilistic IM. The introduction of regularization complicates some of the calculations, so I opt for a numerical solution. Figure 7(b) shows plots of the e-possibilistic IM's risk assessment as a function of the action a, again for different values of the partial prior hyperparameter. This confirms the intuition that the more the partial prior supports "Θ near 0," the closer the IM's risk-minimizing rule will be to 0.

At the end of the day, the IM offers a different assessment of the quality of various actions compared to, say, a Bayesian assessment that relies on a precise posterior distribution. What does this new framework for assessment have to offer? Given the strong reliability properties of the IM itself, as presented in Sections 4.3–4.4, it makes sense to ask if these carry over to the formal decision-making problem. If so, then that

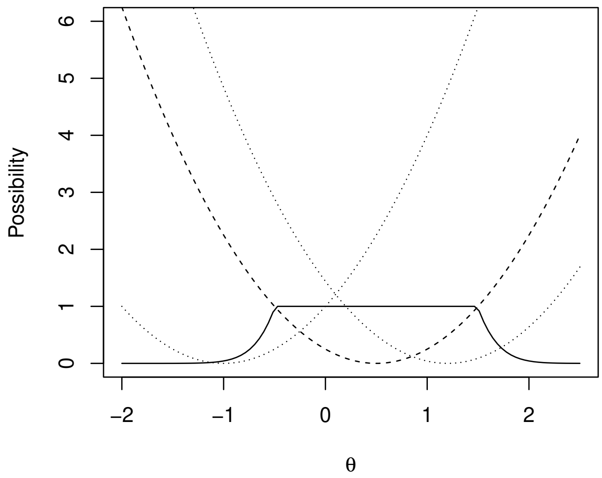


Figure 6: Plot of the (unregularized) e-possibilistic IM's contour (solid), based on data z^n with $n = 5$ and $\bar{z}_n = 0.5$. Plots of the squared error loss function for several different values of a are overlaid (dotted lines), including for $\hat{a} = \bar{z}_n$ (dashed line).

is a genuinely novel, e-possibilistic IM-specific advancement over the existing Bayesian results. The main result is presented next, a generalization of the corresponding result in Grünwald (2023) for what he calls a capped e-posterior.

Theorem 5. *Given a loss function $\ell \geq 0$, the the regularized e-possibilistic IM's assessment of actions satisfies*

$$\overline{\mathbf{P}}\Big\{\sup_{a \in \mathbb{A}} \frac{\ell_a(f(\Omega))}{\overline{\Pi}_{Z^N}^{\mathfrak{e} \times \rho}(\ell_a)}\Big\} \leq 1 \quad \textit{for all } N. \tag{31}$$

In the special case of vacuous partial prior information, (31) *specializes to*

$$\sup_{\omega \in \mathbb{O}} \mathsf{E}_\omega \Big\{\sup_{a \in \mathbb{A}} \frac{\ell_a(f(\omega))}{\overline{\Pi}_{Z^N}^{\mathfrak{e}}(\ell_a)}\Big\} \leq 1 \quad \textit{for all } N.$$

Proof. Write Θ instead of $f(\Omega)$ for now. First, observe that the integrand

$$s \mapsto \sup\{\ell_a(\theta) : \pi_{Z^N}^{\mathfrak{e} \times \rho}(\theta) \geq s\}$$

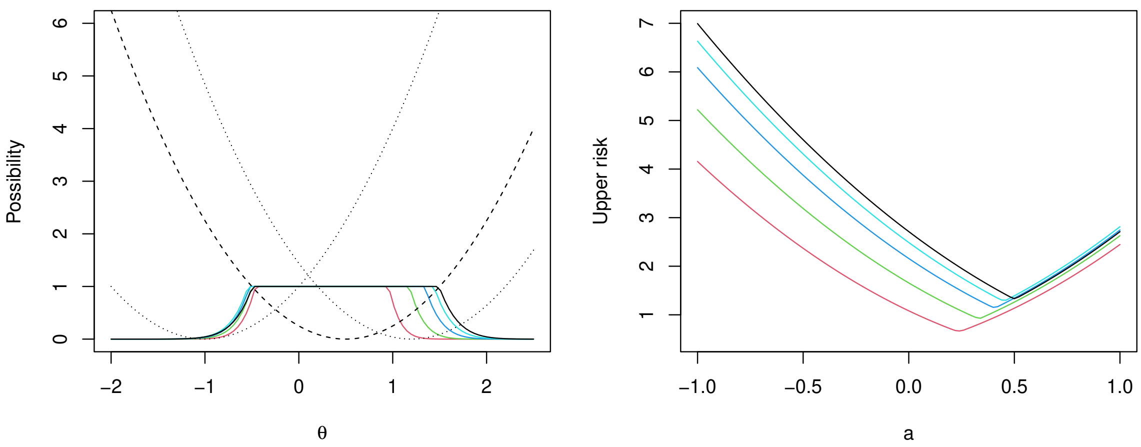
in $\overline{\Pi}_{Z^N}^{\mathfrak{e} \times \rho}(\ell_a)$ is non-decreasing. Next, I proceed by considering two separate cases.

1. Case: $\mathfrak{e}^{\text{reg}}(Z^N, \Theta) > 1$. In this case, $\pi_{Z^N}^{\mathfrak{e} \times \rho}(\Theta) = 1$ so, by the monotonicity property mentioned above, the pair (Z^N, Θ) is such that

$$\sup_{\theta : \pi_{Z^N}^{\mathfrak{e} \times \rho}(\theta) \geq s} \ell_a(\theta) \geq \ell_a(\Theta) \quad \text{for all } s \in [0, 1].$$

   Of course, if the integrand is bounded below, then the integral, over all of $[0, 1]$, is also bounded below by the same value. Therefore,

$$\overline{\Pi}_{Z^N}^{\mathfrak{e} \times \rho}(\ell_a) \geq \ell_a(\Theta),$$



(a) Contour

(b) IM's risk assessment

Figure 7: Panel (a): Plots of the (regularized) e-possibilistic IM's contour, unregularized and regularized, based on data z^n with $n = 5$ and $\bar{z}_n = 0.5$; colors correspond to those in Figure 4. Plots of the squared error loss function for several different values of a are overlaid (dotted lines), including for $\hat{a} = \bar{z}_n$ (dashed line). Panel (b): Plots of the e-possibilistic IM's risk assessment $a \mapsto \overline{\Pi}_{z^n}^{\mathfrak{e}\times\rho}\,\ell_a$ for five different partial priors—the vacuous prior (black) is minimized at $\bar{z}_n = 0.5$ while the others shrink toward 0.

which implies

$$\frac{\ell_a(\Theta)}{\overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(\ell_a)} \leq 1 < \mathfrak{e}^{\text{reg}}(Z^N, \Theta). \tag{32}$$

2. Case: $\mathfrak{e}^{\text{reg}}(Z^N, \Theta) \leq 1$. In this case, $\pi_{Z^N}^{\mathfrak{e}\times\rho}(\Theta) \leq 1$, so I can lower-bound the upper expected loss by truncating the range of integration as follows:

$$\begin{aligned} \overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(\ell_a) &= \int_0^1 \Big\{ \sup_{\theta:\pi_{z^n}^{\mathfrak{e}\times\rho}(\theta)\geq s} \ell_a(\theta) \Big\}\, ds \\ &\geq \int_0^{\pi_{Z^N}^{\mathfrak{e}\times\rho}(\Theta)} \Big\{ \sup_{\theta:\pi_{z^n}^{\mathfrak{e}\times\rho}(\theta)\geq s} \ell_a(\theta) \Big\}\, ds \\ &\geq \pi_{Z^N}^{\mathfrak{e}\times\rho}(\Theta)\, \ell_a(\Theta), \end{aligned}$$

where the last inequality is again by the monotonicity property highlighted above. On rearranging, and using the fact that $\pi_{Z^N}^{\mathfrak{e}\times\rho}(\Theta) = \mathfrak{e}_\Theta(Z^N)\,\rho(\Theta)$ in this case, I get

$$\frac{\ell_a(\Theta)}{\overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(\ell_a)} \leq \mathfrak{e}^{\text{reg}}(Z^N, \Theta),$$

which is the same bound as in (32).

Next, it's clear that the common bound derived in the two separate cases above holds uniformly in the actions, i.e.,

$$\sup_{a\in\mathbb{A}} \frac{\ell_a(\Theta)}{\overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(\ell_a)} \leq \mathfrak{e}^{\text{reg}}(Z^N, \Theta).$$

Plugging $f(\Omega)$ back in for Θ, taking $\overline{\mathbf{P}}$-expectation on both sides, and then applying (15), establishes the bound in (31). $\square$

Several comments are in order. The first concerns the interpretation of Theorem 5: it says that there is no (possibly data-dependent) action a such that the e-possibilistic IM's assessment of a doesn't tend to be overly optimistic (i.e., significantly smaller) than that of an oracle who knows the true state of the world Ω. To understand this, consider the oracle and IM assessments as functions of the actions:

$$a \mapsto \ell_a(\Theta) \quad \text{and} \quad a \mapsto \overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(\ell_a).$$

Imagine a plot of these two functions. The former, the oracle assessment, will take small values for a near the unattainable "best" action $a^\star = \arg\min_a \ell_a(\Theta)$. Since the data's informativeness is limited, it'd be unrealistic to expect that the IM's assessment would tend to be small near $a^\star$ too. Consequently, if the IM's assessment of some a is much more optimistic than the oracle's, then it's likely that this a is different from $a^\star$. And if the IM's assessment is favoring actions away from $a^\star$, then there'd be a risk of suffering a large loss by taking actions suggested by the IM, e.g., $\hat{a}$ in (30). Since Theorem 5 excludes this possibility, I claim the IM-based assessment is reliable.

Second,the uniform bound on an expectation can be turned into an uniform probability bound. In particular,

$$\overline{\mathbf{P}}\Big\{\sup_{a\in\mathbb{A}} \frac{\ell_a(f(\Omega))}{\overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(\ell_a)} \geq \alpha^{-1}\Big\} \leq \alpha \quad \text{for all } \alpha \in [0,1], \text{ all } N. \tag{33}$$

The above says that existence of a triple (Z^N, Θ, a) such that the IM's risk assessment is drastically smaller than the oracle's is a $\overline{\mathbf{P}}$-rare event. This helps drive home the point that Theorem 5 implies the IM's risk assessment is reliable.

Third, the supremum on the inside of the $\overline{\mathbf{P}}$-expectation implies that the fixed a can be replaced by any data-dependent a, i.e., $a(Z^N)$, leading to

$$\overline{\mathbf{P}}\Big\{\frac{\ell_{a(Z^N)}(f(\Omega))}{\overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}(\ell_{a(Z^N)})}\Big\} \leq 1 \quad \text{for all } N.$$

In particular, the above holds with IM-based rule $\hat{a}(Z^N)$ in (30). With this focus on a single, but data-dependent action, say, $\hat{a}(Z^N)$, there's a bit more that can be said concerning the probability bound (33). Indeed, (33) can be unwrapped as

$$\overline{\mathbf{P}}\Big\{\ell_{\hat{a}(Z^N)}(f(\Omega)) \geq \alpha^{-1}\overline{\Pi}_{Z^N}^{\mathfrak{e}\times\rho}\,\ell_{\hat{a}(Z^N)}\Big\} \leq \alpha.$$

That is, it's a $\overline{\mathbf{P}}$-rare event that the realized loss $\ell_{\hat{a}(Z^N)}(\Theta)$—corresponding to the uncertain (Z^N, Θ)—incurred by taking the IM's suggested action $\hat{a}(Z^N)$ is a large multiple of the IM's internal, data-dependent assessment of the risk.

To conclude this section, I'll briefly compare the setup and result here to that in Grünwald (2023). His capped e-posterior (Definition 1) is exactly my unregularized e-possibilistic IM contour, but he doesn't use it to develop a possibility-theoretic approach to uncertainty quantification and inference. Instead, he proceeds in somewhat of an ad hoc manner, proposing the follow *e-risk assessment* of an action a,

$$\mathrm{ER}_a(Z^N) = \sup_\theta \bigl\{ \ell_a(\theta)\, \pi_{Z^N}^{\mathfrak{e}}(\theta) \bigr\}.$$

The right-hand side above does weight the risk relative to the capped e-posterior, but since it's not a Choquet integral, the direct connection with the von Neumann–Morganstern theory is lost. But he does establish a result analogous to that in Theorem 5 above: using my notation, Grünwald's Proposition 2 says that, under a condition,

$$\sup_{\omega \in \Omega} \mathsf{E}_\omega \Bigl\{ \frac{\ell_{\tilde{a}(Z^N)}(f(\omega))}{2 \times \mathrm{ER}_{\tilde{a}(Z^N)}(Z^N)} \Bigr\} \leq 1,$$

where $\tilde{a}(Z^N)$ is the minimizer of the e-risk assessment above. One clear difference between my result and Grünwald's is the factor "2" in the above display. But how do the various risk assessments compare? Figure 8 suggests the following relationships:

$$\mathrm{ER}_a(z^n) \leq \overline{\Pi}_{z^n}^{\mathfrak{e}}\, \ell_a \leq 2 \times \mathrm{ER}_a(z^n), \quad a \in \mathbb{A}. \tag{34}$$

The first inequality holds in some generality,[5] but I've not been able to confirm the second inequality beyond just with plots like the one here. If the second inequality also holds, then my results are stronger than Grünwald's in the sense that it's easier to control a ratio if his denominator is bigger than mine. The other difference is that his result holds only under what he calls "Condition 0," but I won't discuss this here. The point of this comparison is simply that the results obtained here and in Grünwald (2023) are similar, but my e-possibilistic IM result holds with and without regularization, and is all done within a proper, coherent, possibility-theoretic framework.

5 Application

Statistical, scientific, and ethical challenges emerge in the context of comparing two medical treatments when one treatment is potentially far superior to the other. The goal is to demonstrably prove that the one treatment is superior, so that the inferior treatment can be safely removed from use, thereby saving/improving lives. On the one hand, if one treatment is far superior, then, with a sufficient number of experimental units, this should be easy to prove statistically. On the other hand, if researchers have a hunch that one treatment is superior, then it's borderline unethical—the Hippocratic Oath says "Do no harm"—to continue with the planned randomized trial that exposes some patients to the inferior treatment with much higher mortality rates. This context clearly highlights the importance of (a) accommodating adaptive data-collection schemes (rather than traditional, fixed-sample-size designs) and (b) using reliable and efficient statistical methods from which justifiable conclusions can be drawn while minimizing the exposure of patients to a potentially inferior treatment.

One well-known study is presented in Ware (1989). His investigation concerned persistent pulmonary hypertension in newborns. The standard treatment for many years,

[5] That is, if $\theta \mapsto \ell_a(\theta)$ is convex and if the level sets $\{\theta : \pi_{z^n}^{\mathfrak{e}}(\theta) \geq s\}$, for $s \in [0,1]$ are convex, then the first inequality in (34) holds.

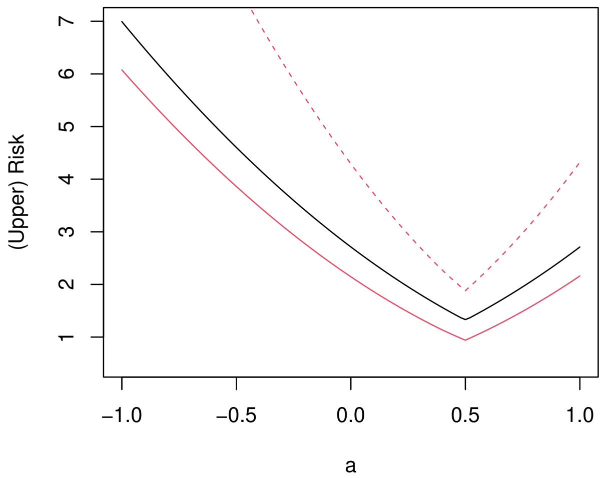


Figure 8: Plots of the (unregularized) e-possibilistic IM's risk assessment $a \mapsto \overline{\Pi}^{\mathfrak{e}}_{z^n}\,\ell_a$ (black) and Grünwald's two risk assessments $a \mapsto \text{ER}_a(z^n)$ (red) and $a \mapsto 2 \times \text{ER}_a(z^n)$ (red dashed) based on data z^n with $n = 5$ and $\bar{z}_n = 0.5$.

called the conventional medical therapy (CMT), had a mortality rate of at least 80%. But by 1985, a new potential treatment had emerged, namely, extracorporeal membrane oxygenation (ECMO), for which reported survival rates were at least 80%. Unfortunately, the empirical support for this latter claim was rather thin: only one randomized trial had been performed and, in that trial, only one patient was assigned to CMT. Despite ECMO's strong performance in these studies, the combination of its statistically-limited support and its potential for side-effects gave some medical researchers pause. Ware's paper describes a trial that he and his colleagues designed and carried out, one that's based on randomization (within blocks) until a prespecified number of deaths—namely, 4—are observed in *each group*. Ware's data is as follows:

$$\text{CMT: 10 patients with 4 deaths} \qquad \text{ECMO: 9 patients with 0 deaths.} \tag{35}$$

I'll focus my analysis below on the data in (35). Note that they didn't actually carry out the design as planned: in particular, they didn't continue following patients until a fourth death in the ECMO group was observed.[6] On-the-fly decisions like this are quite common, hence the importance of anytime valid statistical procedures.

Assuming independence, and that the uncertain survival probabilities—denoted by Θ_{cmt} and Θ_{ecmo}—are constant across patients, the likelihood function based on (35) is

$$L_{z^n}(\theta) \propto \theta_{\text{cmt}}^6 \, (1 - \theta_{\text{cmt}})^4 \, \theta_{\text{ecmo}}^9, \qquad \theta = (\theta_{\text{cmt}}, \theta_{\text{ecmo}}),$$

where "z^n" is just the symbol I'll use for the data in (35). For the e-process, I'll use a slightly modified version of the recommendation in Turner and Grünwald (2023), which

[6]It's worth noting that Ware's team did carry out a second, non-randomized phase, in which another 20 patients were given ECMO and only 1 died. But there were other changes made in the patient admission process, which created some concerns about homogeneity between the first and second phases of the trial. As such, like Ware and also Walley, I'm excluding the Phase 2 data from my analysis.

corresponds to a Bayes factor for testing a generic point null:

$$\mathfrak{e}_\theta(z^n) = \frac{\hat\theta_{\text{cmt},\beta}^6 \,(1-\hat\theta_{\text{cmt},\beta})^4 \,\hat\theta_{\text{ecmo},\beta}^9}{\theta_{\text{cmt}}^6 \,(1-\theta_{\text{cmt}})^4 \,\theta_{\text{ecmo}}^9}, \qquad \theta = (\theta_{\text{cmt}}, \theta_{\text{ecmo}}),$$

where

$$\hat\theta_{\text{cmt},\beta} = \frac{6+\beta}{10+2\beta} \quad \text{and} \quad \hat\theta_{\text{ecmo},\beta} = \frac{9+\beta}{9+2\beta},$$

with $\beta = 0.18$ as suggested in Turner and Grünwald (2023, Sec. 3). The results of my analysis based on this e-process will be presented below; see, e.g., Figure 9(a).

The reader can glean from the above discussion that, while the data in this particular study might be rather limited, there is some "prior information" available based on background knowledge, historical data, etc. But what to do with it? Of course, the frequentist analysis in Ware (1989) formally ignores prior information; informally, however, the available prior information is incorporated in ad hoc ways. Bayesian solutions face the problem that the prior information is insufficient to determine a single, precise prior distribution for Θ. A common strategy in such cases is a prior sensitivity analysis based on a few selected priors that are "consistent" in some sense with the scientific information available; see Kass and Greenhouse (1989) and Kass (1992) for such an analysis in the CMT/ECMO study. Clearly, none of these priors are "right," so the best one can hope for is that the sensitivity analysis reveals that the posterior isn't sensitive to the prior. But that happens only when the data is at least moderately informative, in which case, the likelihood alone is enough—the prior and Bayesian analysis isn't really needed. Walley (1996, Sec. 5) offers a generalized Bayesian analysis based on his so-called imprecise Dirichlet model. Understandably, Walley too virtually ignores the prior information but, rather than literally ignoring it, he models the near-ignorance by a prior credal set that contains all independent beta distributions for $(\Theta_{\text{cmt}}, \Theta_{\text{ecmo}})$. My interest is in cases where the data is limited, like in the present CMT/ECMO study, and the goal is to leverage, rather than ignore, real-but-necessarily-incomplete prior information to *strengthen* the analysis and justification for its conclusions.

Next, I'll present my encoding of the available partial prior information—with minimal embellishment on the information given in Ware (1989)—as a possibility distribution to be used in my subsequent analysis. To be clear, I have no expertise in persistent pulmonary hypertension in newborns specifically or pediatrics generally. As such, my analysis is sure to be overly simplistic and, for that reason, is only for illustrative purposes. That said, I think my partial prior formulation, and the conclusions suggested by the subsequent analysis, are quite reasonable. There's bound to be some degree of subjectivity in these considerations, but that's nothing to be afraid of; subjectivity isn't a dirty word. In high-stakes contexts involving uncertainty, the reason we give more credibility to experts than to novices is that we trust the former's insights and judgments more than that latter's, not because the one knows more objective facts than the other. One of my non-technical goals in this paper is to emphasize both the need for and potential benefits from sound subjective judgments in statistical analyses. So, please feel free to disagree with my prior encoding, analysis, etc. and put forward your own; I welcome the opportunity to learn.

There are two statements—"≤ 0.2" and "≥ 0.8"—that stand out in Ware's report, which are best interpreted as "prior limits" for Θ_{cmt} and Θ_{ecmo}, respectively. What's missing are quantitative statements about the confidence or degree of belief in these limits, but there are qualitative statements made in the text. This is where some subjective judgment becomes necessary. As Ware explains, there's reasonably strong support for the prior limit "≥ 0.8" for Θ_{ecmo}, based on historical data. Importantly, Ware argues that the study inclusion protocol, etc. are such that the patients in these previous studies and those in his study are more-or-less homogeneous. The support for the prior limit "≤ 0.2" for Θ_{cmt}, on the other hand, is considerably weaker. The trouble is that there's only one case involving a patient that could've received ECMO but was randomly assigned to CMT instead; this casts doubt on whether the difference between survival rates in prior studies can be solely attributed to the CMT/ECMO treatment. Based on my interpretation of Ware's report, the confidence I assign to the two prior limits are as follows:

- 90% confident in "≥ 0.8" for Θ_{ecmo}, and
- 50% confident in "≤ 0.3" for Θ_{cmt}.

Aside from the difference in confidence levels, which I'll explain shortly, notice that I stretch out the limit for Θ_{cmt} a bit; this is because Ware used "0.2" as a sort of prior mode for Θ_{cmt}, so it might be prudent to have the mode closer to the middle of the range to which I assign some non-trivial degree of confidence. The confidence levels above are based on my judgment that researchers are quite confident in the performance of ECMO but far less—say, roughly *half* as—confident about CMT on similar populations of patients. Mathematically, I opt to encode this vaguely-stated quantification of uncertainty as a possibility measure for $\Theta = (\Theta_{\text{cmt}}, \Theta_{\text{ecmo}})$. Marginally, the two contours are

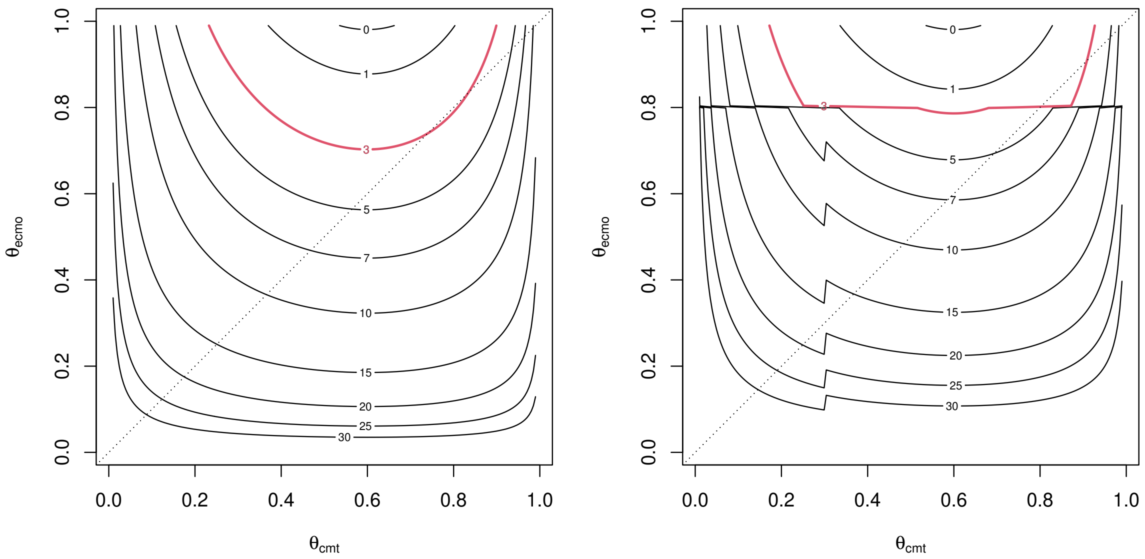
$$\begin{aligned} q_{\text{ecmo}}(\theta_{\text{ecmo}}) &= 0.1 + 0.9 \cdot 1(\theta_{\text{ecmo}} \geq 0.8) \\ q_{\text{cmt}}(\theta_{\text{cmt}}) &= 0.5 + 0.5 \cdot 1(\theta_{\text{cmt}} \leq 0.3). \end{aligned}$$

The aforementioned "confidence levels" correspond to the properties

$$\overline{\mathsf{Q}}_{\text{cmt}}(\Theta_{\text{cmt}} \leq 0.3) = 0.5 \quad \text{and} \quad \underline{\mathsf{Q}}_{\text{ecmo}}(\Theta_{\text{ecmo}} \geq 0.8) = 0.9.$$

Following Walley and other authors, I'll treat Θ_{cmt} and Θ_{ecmo} as independent *a priori* which, for reasons I won't explain here, justifies taking a joint possibility contour q, and corresponding possibility measure $\overline{\mathsf{Q}}$, for $\Theta = (\Theta_{\text{cmt}}, \Theta_{\text{ecmo}})$ as the product of the marginal possibility contours presented above. This corresponds to a piecewise constant function on the unit square, taking value 1 on the rectangle $[0, 0.3] \times [0.8, 1]$ in the $(\Theta_{\text{cmt}}, \Theta_{\text{ecmo}})$ space, and other smaller values in the three other rectangles. While others might disagree to some extent with the particular confidence levels that I chose, namely, 0.5 and 0.9, I believe that this possibilistic prior is consistent with the information about CMT vs. ECMO based on Ware's presentation. In particular, the credal set determined by $\overline{\mathsf{Q}}$ contains independent products of certain beta distributions, among others.

First, plots of the unregularized and regularized e-processes are shown in Figure 9; for the regularized e-process, I'm using the same calibrator γ as described in Section 3.3. As expected, the e-process contour bottoms out at the value corresponding to the simple sample proportions, namely, $(0.6, 1)$. As θ moves away from that minimizer, the e-process increases and creates contours of a shape determined by the data and the specific e-process formulation. Note that the unregularized e-process contours are smooth, whereas those of the regularized version's are rough in some places; this roughness is due to the



(a) Unregularized (b) Regularized

Figure 9: Plots of the unregularized and regularized e-processes based on Ware's CMT/ECMO data. Heavy red lines mark the corresponding 95% confidence sets.

relatively large jump discontinuity in the prior contour there. The heavy red line marks the confidence sets C_α and C_α^{reg} for $\alpha = 0.05$, and there are two notable observations. First, as it pertains to Θ_{ecmo}, for which prior information is relatively strong, the limits are much tighter in the regularized case compared to the unregularized. Second, for Θ_{cmt}, the limits are a bit looser for the regularized case compared to the unregularized. The latter point might seem disappointing, but this is exactly what should happen: prior knowledge that doesn't fully agree with data ought to result in more conservative inference.

A relevant scientific question is if EMCO is a more effective treatment than CMT. One way to answer this is to formulate a hypothesis test of "$\Theta_{\text{ecmo}} \leq \Theta_{\text{cmt}}$" versus "$\Theta_{\text{ecmo}} > \Theta_{\text{cmt}}$." The diagonal line through the two plots in Figure 9 represents the boundary between these two propositions and, since the red contour curves intersect with the diagonal line, the corresponding e-process-based tests cannot reject the hypothesis "$\Theta_{\text{ecmo}} \leq \Theta_{\text{cmt}}$" at level $\alpha = 0.05$; but they could reject at a slightly smaller α levels. For comparison, Ware presents an analysis that using Fisher's exact test, leading to a p-value of 0.054, which is completely consistent with my conclusions based on Figure 9(a). My results are based on procedures proved to be anytime valid, so this provides additional comfort given that Fisher's exact test is not.

For uncertainty quantification, the possibility contours, $\pi^{\mathfrak{e}}$ and $\pi^{\mathfrak{e}\times\rho}$, in this case look identical to those in Figure 9, just with different numerical labels on the contours. So, the confidence regions and test conclusions derived from $\pi^{\mathfrak{e}}$ and $\pi^{\mathfrak{e}\times\rho}$ give exactly the same results as those obtained from Figure 9. But there's more that can be done with the IM's possibilistic uncertainty quantification. First, consider a feature $\Delta = \Theta_{\text{ecmo}} - \Theta_{\text{cmt}}$, the difference in survival rates between EMCO and CMT. The so-called possibility-theoretic *extension principle* (e.g., Zadeh 1975) determines a (marginal) possibility contour for Δ

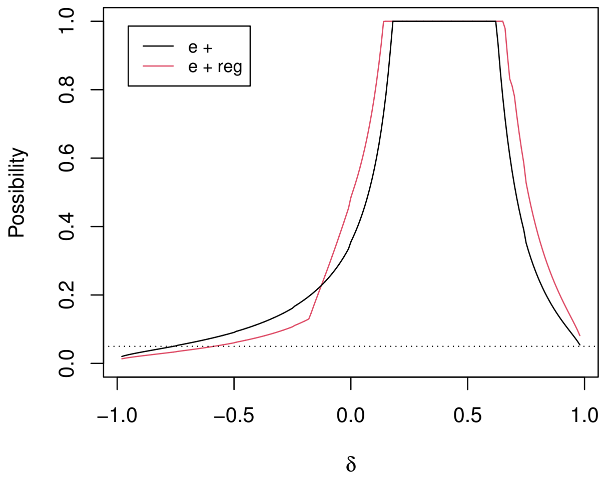


Figure 10: Plots of the regularized and unregularized e-possibilistic IM's marginal contour for $\Delta = \Theta_{\rm ecmo} - \Theta_{\rm cmt}$ based on Ware's CMT/ECMO data.

from that for Θ:

$$\phi_{z^n}^{\mathfrak{e}}(\delta) = \sup_{\theta \in [0,1]^2 : \, \theta_{\rm ecmo} - \theta_{\rm cmt} = \delta} \pi_{z^n}^{\mathfrak{e}}(\theta), \quad \delta \in [-1,1].$$

The regularized e-possibilistic marginal IM contour $\phi_{z^n}^{\mathfrak{e} \times \rho}$ is defined analogously. The corresponding upper probabilities $\overline{\Phi}_{z^n}^{\mathfrak{e}}$ and $\overline{\Phi}_{z^n}^{\mathfrak{e} \times \rho}$ are defined via optimization as usual. The dashed line at $\alpha = 0.05$ determines a marginal anytime valid confidence interval for Δ and, at least for the regularized version (in red), this more-or-less agrees with the generalized Bayes 95% credible interval presented in Walley (1996, Fig. 1). The message from this plot is that, while "$\Delta > 0$" is highly possible, nearly the entire range $[-1,1]$ is "sufficiently possible" based on Ware's data. Admittedly, the extension principle leads to conservative marginalization, but it's the most direct way to eliminate nuisance parameters while preserving anytime validity.

Finally, a formal decision-theoretic approach can be considered, using suitable lower and upper expected loss/utility based on the e-possibilistic IM output. Walley (1996) argues that of utmost importance—or *utility*—is a patient's survival. That is, the desired goal is for the patient to survive with the treatment they were given, all other factors are irrelevant. So, if Y is a binary indicator that the patient in question survives, then the utilities associated with CMT and ECMO are

$$u_{\rm cmt}(y) = u_{\rm ecmo}(y) = y, \quad y \in \{0,1\} \equiv \{\text{dies}, \text{survives}\}.$$

In this case, the expected utilities, given the respective treatments, are $\Theta_{\rm cmt}$ and $\Theta_{\rm ecmo}$. Since $(\Theta_{\rm cmt}, \Theta_{\rm ecmo})$ is uncertain, I can't simply pick which of CMT and ECMO has higher expected utility. But I can evaluate the lower and upper expectation of the difference between the two expected utilities, namely, $\Theta_{\rm ecmo} - \Theta_{\rm cmt}$, based on the full uncertainty quantification provided by the e-possibilistic IM. I've already obtained the marginal IM for the difference in expected utilities—which I previously denoted by Δ—so all that remains is to evaluate the lower and upper expectations. For the unregularized e-possibilistic IM, the upper and lower expectation of Δ are

$$\overline{\Phi}^{\mathfrak{e}}_{z^n}(\Delta) = \int_0^1 \sup\{\delta : \phi^{\mathfrak{e}}_{z^n}(\delta) > s\} \, ds$$
$$\underline{\Phi}^{\mathfrak{e}}_{z^n}(\Delta) = -\overline{\Phi}^{\mathfrak{e}}_{z^n}(-\Delta),$$

and analogously for the regularized version. Numerically, I get:

$$\big[\underline{\Phi}^{\mathfrak{e}}_{z^n}(\Delta), \overline{\Phi}^{\mathfrak{e}}_{z^n}(\Delta)\big] = [-0.042, 0.739] \quad \text{and} \quad \big[\underline{\Phi}^{\mathfrak{e}\times\rho}_{z^n}(\Delta), \overline{\Phi}^{\mathfrak{e}\times\rho}_{z^n}(\Delta)\big] = [0.086, 0.820].$$

That the first interval, corresponding to the unregularized version, contains 0 means that I can't rule out, based on Ware's data, that CMT is preferred to ECMO in terms of utility. For the regularized version, however, the interval is strictly to the right of the origin, which implies that ECMO is the preferred treatment in terms of expected utility, albeit just barely. For comparison, Walley reports his version the "lower and upper expectation of Δ" as $[0.152, 0.5]$, so he concludes that ECMO is the demonstrably better treatment. One has to decide whose analysis is more convincing, mine or Walley's, but my IM solution actually uses prior information and my expected utility offers a strong, decision-making reliability guarantee via Theorem 5.

6 Extensions

Here I'd like to briefly mention a few possible extensions of the formulation here. Some of these I more-or-less know how to do, and others I don't.

First, I've focused here on the case where, given Ω (or P), the data are iid. There's no conceptual difficulty in extending this to the case of independent but not iid, e.g., if data points depend on covariates. This case was treated informally in Section 5 where the treatment assignment—CMT versus ECMO—plays the role of a covariate. Dependent data sequences is a bigger challenge, a topic for future investigation.

Second, the literature on anytime valid inference—including this paper—has focused exclusively on the case where N can be literally anything (subject to the mathematical constraint that it be a stopping time). That's why the above results all have "for all N" appended to the statements. In applications, however, it's hard to imagine that literally *any* stopping rule is in consideration. If some stopping rules can be excluded, so that "for all N" can be replace by "for such-and-such N," then that creates an opportunity to leverage those constraints and improve the (regularized) e-process's efficiency. Even more generally, You might not know which N was actually used, but You might be able to quantify Your uncertainty with a possibility contour or other kind of imprecise probability. The relevant questions are if partial prior information about N can be incorporated; if so, then how; and then what kind of efficiency gains can be expected.

Finally, while the problem formulation is quite general, my illustrations have only considered cases involving uncertain Θ with fixed dimension. In anticipation of applications to high-dimensional problems, it's important to recognize that an efficiency gain can be realized only if the regularization is allowed to adapt to the dimension of Θ. In the classical statistics and machine learning literature, this is accommodated by tying the dimension to the sample size and then allowing the penalty/prior to depend explicitly on the sample size. This is a little awkward in the present context, for multiple reasons. Perhaps the biggest challenge is that the data are streaming, and the stopping rule is dynamic, so if dimension and sample size are linked, then it's as if the inference problem doesn't have a well-defined target. Another challenge is dealing with the fact that $\mathscr{Q}$ is assumed to be a *real* representation of Your partial knowledge and it might be awkward, at least initially, to think about knowledge depending on dimension. As I see it, however, the intuition behind standard penalties/priors—e.g., "most of the entries in Θ are near-zero"—is inherently imprecise/possibilistic and notions like "most" are relative to the number of things in question, which in this case is the dimension of Θ. This seems doable, but the question remains: specifically how should $\mathscr{Q}$ adapt to dimension?

7 Conclusion

This paper develops the new concept of and theory associated with *regularized e-processes*, from which I further develop reliable—i.e., anytime valid and efficient—inference and uncertainty quantification. On the technical side, the regularized e-process involves the combination of a fully data-driven e-process and prior knowledge about Θ that You, the investigator, might have. Importantly, Your prior information need not be—and typically won't be—sufficiently complete to pinpoint a single prior as a Bayesian analysis would require, so my proposal explicitly draws on aspects of imprecise probability theory. My proposed regularization discounts those values of Θ that are incompatible with the available prior knowledge, making it easier to "reject" such values compared to the purely data-driven e-process, hence the efficiency gains. The critical point, however, is that any such non-trivial discounting like described above jeopardizes the original e-process's inherent anytime reliability; therefore, the proposed regularization requires care.

By encoding the partial prior information rigorously as a coherent imprecise probability, I can build a collection of joint distributions for (data, Θ) compatible with the assumed model and available prior information. Then I generalize the now-familiar anytime validity property in a sound way, accounting for the available prior information, and similarly generalize Ville's inequality to prove that the proposed regularized e-process remains anytime valid in this more general sense. This reformulation offers mathematical justification for You to accept the efficiency gains offered by regularization.

From standard frequentist-style of inference based on (anytime) valid tests and confidence sets, I go on to develop a Bayesian-like framework for broader uncertainty quantification based on e-processes. This falls under the general umbrella of *inferential models*, or IMs. The proposed brand of uncertainty quantification is possibility-theoretic, which means that the proposed e-possibilistic IM operates with a different albeit simple and intuitive calculus: optimization instead of integration. Aside from these small technical differences, the approach proposed here at least superficially resembles Bayesian inference. What distinguishes the proposed e-possibilistic IM framework is that it exactly (not just asymptotically) enjoys the calibration necessary to ensure that the uncertainty quantification derived from it is reliable. This kind of reliability immediately implies that certain summaries of the e-possibilistic IM provide anytime valid test and confidence set procedures; it also implies more than the usual frequentist guarantees, e.g., reliable probing for hypotheses supported by data (Corollaries 2 and 3). Remarkably, the e-possibilistic IM construction leads to a number of other desirable properties, things often associated with Bayesian inference. This includes: satisfying the likelihood principle, avoiding sure-loss, and formal decision-making with strong reliability guarantees.

An illustration of the proposed framework in the context of a clinical study was presented in Section 5. While this was mainly just to illustrate what the proposed framework has to offer, the results were reasonable and computation was completely trivial. A next step is to consider other applications, of greater complexity, to build up experience and showcase what regularized e-processes can do. Since having incomplete prior knowledge is the norm, and not the exception, the proposal here should immediately replace the tools that practitioners are using on a daily basis, ones that require investigators to unjustifiably add to or delete from their corpus of knowledge.

A number of interesting and challenging questions remain open, I'll end here with a few remarks about three of them. First, to put the theory presented here into practice, You must convert what You know about the uncertain Θ into an imprecise probability. For example, in the clinical trial data illustration of Section 5, some degree of belief attached to bounds on Θ was converted into a possibility contour. The point is that You need to know (a) how to tease relevant quantitative details out of Your mostly qualitative corpus of knowledge and (b) how to properly map those relevant details to a suitable imprecise probability. On top of this, the second point, is that a possibilistic formulation of the prior information requires choice of a calibrator as I described in Section 3.3. The elicitation and encoding of partial prior information is an interesting question, more psychological than statistical, something that I plan to explore. A thorough comparison of how different calibrators perform in this context across different model/data settings is also needed

Second, the reader may have noticed in Section 3.7 that there are instances (i.e., model and data combinations) in which the regularized e-process based on weak prior information apparently gives slightly wider confidence limits than the unregularized e-process that assumes vacuous prior information. This is counter-intuitive because a vacuous prior encodes strictly less information than even a weak partial prior—my intuition was that regularization would always help if data and prior are compatible. That intuition holds if I could regularize via multiplying the e-process by the prior possibility contour's reciprocal. But this reciprocal, on its own, can't be a regularizer. So the relevant question is: for a given calibrator, is some prior information "too weak" that You're better off ignoring it than trying to pass it through a calibrator before regularization? It's hard for me to accept that relevant information should ever be ignored, so let me restate the previous question from a different perspective: should the choice of calibrator depend on the strength of the partial prior information and, if so, then how?

Acknowledgments

This work is partially supported by the U.S. National Science Foundation, grants SES–2051225 and DMS–2412628.

A Choquet integration

As briefly mentioned in Section 2.3, Choquet integration has an important role to play in this paper's developments. Basically, Choquet integration plays the same role in imprecise probability theory as Lebesgue integration does in ordinary/precise probability theory. That is, just as Lebesgue integration is the go-to mathematical framework for defining expectation with respect to probability measures, Choquet integration is the appropriate way to extend lower/upper probabilities, at least for the kind of models in consideration here, to more general lower/upper expectations. The path to making this connection isn't exactly direct, and the details are too involved to present here, but I think a relatively brief overview would be beneficial. My summary here is based primarily on details presented much more thoroughly and rigorously in Troffaes and de Cooman (2014).

Let $g : \mathbb{T} \to \mathbb{R}$ be a function, which I'll assume to be non-negative only for simplicity; if g can take both positive and negative values, then one apply the developments here to the difference between the positive and negative parts of g. If $\overline{\mathsf{Q}}$ is a general capacity—a normalized, monotone set function—supported on subsets of $\mathbb{T}$, then the Choquet integral of g with respect to $\overline{\mathsf{Q}}$ is defined as

$$\mathcal{I}_{\text{CHOQ}}(g) := \int_0^\infty \overline{\mathsf{Q}}\{\theta \in \mathbb{T} : g(\theta) \geq t\} \, dt, \tag{36}$$

where "$\int$" on the right-hand side is a Riemann integral, which is well-defined since the integrand is a monotone non-increasing function of t. In some sense, there's nothing particularly special about defining an "integral"—anyone can do it. The challenge is defining an integral that represents something relevant. In the present context, the most meaningful notion of an expected value of $g(\Theta)$ with respect to a coherent upper probability $\overline{\mathsf{Q}}$ (defined on all subsets H of $\mathbb{T}$) is the upper envelope

$$\overline{\mathsf{Q}}\, g = \sup_{\mathsf{Q} \in \mathscr{Q}} \mathsf{E}^{\Theta \sim \mathsf{Q}}\{g(\Theta)\},$$

which is the second expression given in (6); recall that $\mathscr{Q}$ here is the set of all probabilities dominated by $\overline{\mathsf{Q}}$, as in (5). How is this connected to the Choquet integral?

What links the Choquet integral above to the upper expectation is a deep result of Walley's, concerning the so-called *natural extension* of $\overline{\mathsf{Q}}$ from an upper probability to an upper expectation. In a purely mathematical sense, one may have a function f defined on a domain $\mathbb{X}$ with certain properties, and the relevant question is if f can be extended from $\mathbb{X}$ to a function $f^\star$ defined on a larger domain $\mathbb{X}^\star$, such that there's agreement on $\mathbb{X}$, i.e., $f^\star(x) = f(x)$ for $x \in \mathbb{X}$, and $f^\star$ maintains f's relevant properties on $\mathbb{X}^\star \setminus \mathbb{X}$. In the present context, the upper envelope can be viewed as a functional $1_H \mapsto \overline{\mathsf{Q}}\, 1_H := \overline{\mathsf{Q}}(H)$ defined on the collection $\{\theta \mapsto 1_H(\theta) : H \subseteq \mathbb{T}\}$ of indicator functions/gambles. This functional has a coherence property, by assumption, so the question is if it can be extended to a broader class of (bounded[7]) gambles without sacrificing coherence. Walley (1991, Ch. 3) answers this question in the affirmative, with what he calls the *natural extension*—the extension that imposes the least additional structure while preserving coherence. On the importance of natural extension, Walley (1991, p. 121–122) writes:

[7]Walley's developments focus on bounded gambles, but Part II of Troffaes and de Cooman (2014) generalizes Walley's results to certain unbounded gambles.

...natural extension may be seen as the basic constructive step in statistical reasoning; it enables us to construct new previsions from old ones.

The formula for the natural extension is rather complicated and not necessary for the present purposes. The relevant point here is that the *upper envelope theorem*[8] in Troffaes and de Cooman (2014, Theorem 4.38) links the upper expectation $\overline{\mathsf{Q}}\,g$ to Walley's natural extension of $\overline{\mathsf{Q}}$ to (bounded) gambles. Then they follow up (Troffaes and de Cooman 2014, Theorem 6.14) by linking the natural extension of $\overline{\mathsf{Q}}$ to the Choquet integral in (36). Therefore, the "$\mathcal{I}_{\text{CHOQ}}(g)$" notation can be dropped—the Choquet integral and the upper expectation $\overline{\mathsf{Q}}\,g$ are the same, so the latter notation is sufficient.

Then the formula given in (8) for the upper expectation with respect to a possibility measure $\overline{\mathsf{Q}}$ determined by contour q follows immediately—or at least *almost* immediately. Using the definition $\overline{\mathsf{Q}}(H)$ via optimization as in (7), the Choquet integral formula (36) above reduces to

$$\overline{\mathsf{Q}}\,g = \int_0^\infty \Big\{ \sup_{\theta : g(\theta) \geq t} q(\theta) \Big\} \, dt.$$

This expression looks similar to the formula given in (8), but it's not the same; the latter roughly has the roles of g and q in the above expression reversed. This apparent "symmetry" in the roles of g and q, and the corresponding alternative form of the Choquet integral as I advertised in (8), is established in Proposition 7.14 (and Proposition C.8) of Troffaes and de Cooman (2014) for the case of bounded g; this is generalized to certain unbounded gambles g in, e.g., their Proposition 15.42.

For a bit of practice with the possibility-theoretic Choquet integral, I'll first demonstrate that Choquet integration formula (8) for a possibility measure $\overline{\mathsf{Q}}$ reduces to the definition of upper probability (7), via optimization of the contour q, when the function g is an indicator, i.e., $g(\theta) = 1(\theta \in H)$ for some $H \subseteq \mathbb{T}$. For such a case, the integrand in (8) is given by

$$s \mapsto \sup_{\theta : q(\theta) \geq s} 1(\theta \in H) = \begin{cases} 1 & \text{if } s \leq \sup_{\theta \in H} q(\theta) \\ 0 & \text{otherwise.} \end{cases}$$

Then it's clear that

$$\begin{aligned} \overline{\mathsf{Q}}\,g &= \int_0^1 \sup_{\theta : q(\theta) > s} 1(\theta \in H) \, ds \\ &= \int_0^{\sup_{\theta \in H} q(\theta)} 1 \, ds + \int_{\sup_{\theta \in H} q(\theta)}^1 0 \, ds \\ &= \sup_{\theta \in H} q(\theta) \\ &= \overline{\mathsf{Q}}(H), \end{aligned}$$

as was to be shown. Warm-up complete.

[8] As the title of their book suggests, Troffaes and de Cooman (2014) focus almost exclusively on *lower* previsions, and what they prove is a *lower* envelope theorem. There is, however, an analogous result for the upper prevision and that's what I'm referring to here as the *upper envelope theorem*.

Next, I have a slightly more ambitious goal of verifying the formula below that was presented without proof in Section 4.5.2:

$$\overline{\Pi}^{\mathfrak{e}}_{z^n}\,\ell_{\hat{a}(z^n)} = \int_0^1 \Bigl\{ \sup_{\theta:\pi^{\mathfrak{e}}_{z^n}(\theta)\geq s} \ell_{\hat{a}(z^n)}(\theta) \Bigr\}\,ds = n^{-1}\bigl\{2 + \log(nv+1) + (\tfrac{n}{nv+1})\bar{z}_n^2\bigr\}.$$

In this case, the contour is $\pi^{\mathfrak{e}}_{z^n}(\theta) = 1 \wedge \mathfrak{e}_\theta(z^n)^{-1}$, where

$$\mathfrak{e}_\theta(z^n) = (nv+1)^{-1/2}\exp\bigl\{\tfrac{n}{2}(\theta-\bar{z}_n)^2 - \tfrac{1}{2}(\tfrac{n}{nv+1})\,\bar{z}_n^2\bigr\}, \quad \theta \in \mathbb{R}.$$

Then

$$\begin{aligned}
\pi^{\mathfrak{e}}_{z^n}(\theta) \geq s &\iff 1 \wedge \mathfrak{e}_\theta(z^n)^{-1} \geq s \\
&\iff \mathfrak{e}_\theta(z^n) \leq s^{-1} \\
&\iff (\theta-\bar{z}_n)^2 \leq n^{-1}\bigl\{2\log(s^{-1}) + \log(nv+1) + (\tfrac{n}{nv+1})\bar{z}_n^2\bigr\}.
\end{aligned}$$

It just so happens that $\ell_{\hat{a}(z^n)}(\theta) = (\theta-\bar{z}_n)^2$, and from this it's clear that

$$\sup_{\theta:\pi^{\mathfrak{e}}_{z^n}(\theta)\geq s} \ell_{\hat{a}(z^n)}(\theta) = n^{-1}\bigl\{2\log(s^{-1}) + \log(nv+1) + (\tfrac{n}{nv+1})\bar{z}_n^2\bigr\}.$$

So it remains to integrate the right-hand side above with respect to s over the interval $[0,1]$. Using the identity

$$\tfrac{d}{ds}(s - s\log s) = \log(s^{-1}),$$

and the fundamental theorem of calculus, it follows that

$$\begin{aligned}
\overline{\Pi}^{\mathfrak{e}}_{z^n}\,\ell_{\hat{a}(z^n)} &= \int_0^1 n^{-1}\bigl\{2\log(s^{-1}) + \log(nv+1) + (\tfrac{n}{nv+1})\bar{z}_n^2\bigr\}\,ds \\
&= n^{-1}\bigl\{2 + \log(nv+1) + (\tfrac{n}{nv+1})\bar{z}_n^2\bigr\},
\end{aligned}$$

as was to be shown.

B Other kinds of regularizers

In the main paper I focused exclusively on regularizers for the case where Your prior information is encoded as a possibility measure with contour q. I did so for two reasons: (a) I think the possibilistic formulation makes sense, and (b) it's relatively concrete to handle compared to other formulations. But this isn't the only option, so here I want to briefly mention a couple other strategies. This is absolutely not intended to be an exhaustive list of alternatives, nor is it my intention for this to be a tutorial on how to construct a regularizer in a non-possibilistic setup. My goal is simply to expose the reader to some other avenues to pursue if the possibilistic version I presented in the main paper isn't fully satisfactory; this also demonstrates my point that the possibilistic formulation is simpler and more concrete.

Arguably one of the most common imprecise probabilities models are the so-called *contamination classes*, *gross error models*, or *linear–vacuous mixtures*, often found in the literature on robust statistics (e.g., Huber 1973, 1981; Walley 1991, 2002; Wasserman

1990). This corresponds to a choice of centering probability Q_{cen} and a weight $\varepsilon \in (0,1)$. Then the credal set is given by

$$\mathscr{Q} = \bigl\{(1-\varepsilon)\,\mathsf{Q}_{\text{cen}} + \varepsilon\,\mathsf{Q} : \mathsf{Q} \text{ is any probability on } \mathbb{T}\bigr\}.$$

The basic idea is that You think the precise probability Q_{cen} is a pretty good assessment of Your uncertainty about Θ, but You don't fully trust the information that lead to this assessment; so ε is like the "chance You were misled," and if You were mislead, then literally anything might be true. It's relatively clear that the corresponding upper probability/prevision, say $\overline{\mathsf{Q}}$, has upper expectation of $\rho(\Theta)$ given by

$$\overline{\mathsf{Q}}\,\rho = (1-\varepsilon)\,\mathsf{Q}_{\text{cen}}\,\rho + \varepsilon \times \sup_{\theta \in \mathbb{T}} \rho(\theta).$$

Then the goal is to choose ρ such that the right-hand side above is (less than or) equal to 1. There's no simple formula for this, like in the case of possibilistic prior information, but the idea is take ρ such that it's relatively small where You "expect" Θ to be, i.e., in the support of Q_{cen}, and then somewhat large elsewhere. If, for example, Q_{cen} has bounded support, so that ρ can be defined independently on that support and elsewhere, then ρ can take values as large as $\varepsilon^{-1}(1-\varepsilon)m$ outside that support, where $m = \mathsf{Q}_{\text{cen}}\,\rho$.

A second kind of imprecise probability model is what Walley (1991, Sec. 2.9.4) calls the *constant odds ratio* model. Similar to that above, this is indexed by a precise probability distribution, which I'll denote again as Q_{cen}, and a weight $\tau \in (0,1)$. Walley explains this model in the context of a risky investment, where τ represents the rate at which You're taxed on said investment. Setup and context aside, the lower and upper probability of a hypothesis/event H under this model is

$$\underline{\mathsf{Q}}(H) = \frac{(1-\tau)\mathsf{Q}_{\text{cen}}(H)}{1-\tau\mathsf{Q}_{\text{cen}}(H)} \quad \text{and} \quad \overline{\mathsf{Q}}(H) = \frac{\mathsf{Q}_{\text{cen}}(H)}{1-\tau\mathsf{Q}_{\text{cen}}(H)}.$$

The name "constant odds ratio model" comes from the fact that the lower and upper odds of H versus H^c are

$$\frac{\underline{\mathsf{Q}}(H)}{\overline{\mathsf{Q}}(H^c)} = \frac{(1-\tau)\mathsf{Q}_{\text{cen}}(H)}{\mathsf{Q}_{\text{cen}}(H^c)} \quad \text{and} \quad \frac{\overline{\mathsf{Q}}(H)}{\underline{\mathsf{Q}}(H^c)} = \frac{\mathsf{Q}_{\text{cen}}(H)}{(1-\tau)\mathsf{Q}_{\text{cen}}(H^c)},$$

respectively, so the rato of lower to upper odds is constant in H:

$$\frac{\underline{\mathsf{Q}}(H)/\overline{\mathsf{Q}}(H^c)}{\overline{\mathsf{Q}}(H)/\underline{\mathsf{Q}}(H^c)} = \cdots = (1-\tau)^2.$$

The upper expectation of $\rho(\Theta)$ under this model is not so straightforward as for the contamination model above, but Walley shows that $\overline{\mathsf{Q}}\,\rho$ solves the equation $f(x) = 0$, where $f(x) = \tau\,\mathsf{Q}_{\text{cen}}(\rho - x)^+ + (1-\tau)\,(\mathsf{Q}_{\text{cen}}\,\rho - x)$, with $w^+ = \max(w, 0)$ the positive part of $w \in \mathbb{R}$. For a given ρ, one can numerically solve for x as a function of $(\tau, \mathsf{Q}_{\text{cen}}\,\rho)$, to obtain $\overline{\mathsf{Q}}\,\rho$. But choosing ρ such that this numerical solution is ≤ 1 requires care.

Other kinds of imprecise models can be considered, including monotone capacities (e.g., Huber 1973; Sundberg and Wagner 1992; Wasserman and Kadane 1990), belief functions (e.g., Dempster 1967; Denœux 1999; Shafer 1976), and probability-boxes (e.g., Destercke et al. 2008; Ferson et al. 2003), can be considered, and formulas for their upper expectations can be found in, e.g., Chapters 6–7 of Troffaes and de Cooman (2014).

C Justification of the product form in (10)

Here I'll present the previously-advertised justification for my choice to define the regularized e-process in (10) as a product of the regularizer and the original e-process. Frankly, this justification isn't much more compelling than (the analogy to Bayesian prior-to-posterior updating together with) the fact that multiplying the two ingredients is the most natural way to merge them. There are, however, some other reasonable options, e.g., averaging, so the result below adds some valuable insight. The reader may have ideas on how to strengthen this result in one way or another.

Throughout this section, to simplify notation, etc., I'll assume that f is the identity function, so that the quantity of interest Θ corresponds exactly with Ω. This helps because it allows me to drop f (and Ω) from the notation, to drop the set of pullback measures in the theoretical formulation, and to express the model as "P_θ," directly in terms of θ. I'm also going to drop the explicit mention of a generic stopping time, and write "Z" for the observable data. Since the main result of this section is conceptual in nature, these simplifications don't affect the take-away message.

Following Vovk and Wang (2021), define a function $(r, e) \mapsto m(r, e)$ to be a *re-merging function*—pronounced "R-E-merging" because it merges a **r**egularizer and an **e**-process—if, for given prior information $\overline{\mathsf{Q}}$ about $\Theta = f(\Omega)$ and a corresponding regularizer ρ, the merged variable $m(\rho, \mathfrak{e})$ satisfies two key properties:

- It treats the data as sovereign in the sense that

$$e \geq 1 \implies m(r, e) \geq r, \quad \text{for all } r \in (0, \infty), \tag{37}$$

  which, in words, means that if the data-dependent component, $\mathfrak{e}_\theta(\cdot)$, offers evidence that doesn't favor hypothesis "$\Theta = \theta$," then the merged e-process will show less support for that hypothesis than the prior information alone did.

- It's a regularized e-process in the sense that (15) holds for any input $\mathfrak{e}$, i.e.,

$$\overline{\mathbf{P}}\big[m\{\rho(\Theta), \mathfrak{e}_\Theta(Z)\}\big] \leq 1 \quad \text{for all e-processes } \mathfrak{e}. \tag{38}$$

Recall that $\overline{\mathbf{P}}$ is determined by the model and by the prior information, so $\overline{\mathbf{P}}$ is fixed by the context of the problem—all that's free to vary in these considerations is the input e-process $\mathfrak{e}$ and the merger function m. The class of re-merging functions is non-empty, since the product mapping is an re-merger. My claim is that, in a sense to be described below, the product merger in (10) is "best" among all the re-merging functions.

Continuing to follow Vovk and Wang (2021), I'll say that an re-merging function m weakly dominates another re-merging function m' if

$$(r, e) \in [1, \infty) \times [1, \infty) \implies m(r, e) \geq m'(r, e). \tag{39}$$

The idea is that, in the case where neither the data nor the prior show signs of compatibility with a given value θ, then merging based on m is more aggressive, i.e., shows no less evidence against θ, than merging based on m'. The complement to an e-process's any-time validity property is its efficiency, and efficiency requires that the e-process take large values when evidence is incompatible with a hypothesis in question; so, an re-merging function that returns larger regularized e-process values is preferred.

The present case differs in many ways from that in Vovk and Wang (2021), mainly due to the presence of the prior information, so I'll need some additional control. Towards this, I'll say that an re-merging function m $\overline{\mathbf{P}}$*-strictly weakly dominates* another re-merging function m' if m weakly dominates m' in the sense of (39) above, and if

$$\overline{\mathbf{P}}\big[m\{\rho(\Theta), \mathfrak{e}_\Theta(Z)\} > m'\{\rho(\Theta), \mathfrak{e}_\Theta(Z)\}, \\ \rho(\Theta) \geq 1, \, \mathfrak{e}_\Theta(Z) \geq 1\big] > 0, \quad \text{for all e-processes } \mathfrak{e}. \tag{40}$$

Vovk and Wang's "weak dominance" in (39) allows m and m' to be the same, so roughly all that (40) adds is that there exists a joint distribution for (Z, Θ)—corresponding to a prior Q in Your credal set $\mathscr{Q}$—with respect to which the "strict inequality" event on the right-hand side of (39) has positive probability. Therefore, $\overline{\mathbf{P}}$-strict weak dominance rules out the possibility that m and m' differ only in an insignificant way relative to $\overline{\mathbf{P}}$. In the special case where the prior information is vacuous, like in Vovk and Wang, strict weak dominance holds if m weakly dominates m' and if, for each e-process $\mathfrak{e}$, there exists a $\theta = \theta(\mathfrak{e})$ such that $\rho(\theta) \geq 1$ and

$$\mathsf{P}_\theta\big[m\{\rho(\theta), \mathfrak{e}_\theta(Z)\} > m'\{\rho(\theta), \mathfrak{e}_\theta(Z)\}, \, \mathfrak{e}_\theta(Z) \geq 1\big] > 0,$$

i.e., if roughly strict inequality holds with positive model probability for some θ.

In the case of vacuous prior information, I'll require $\rho \equiv 1$, as was suggested in the main paper. When the prior information is non-vacuous, the result below is restricted to *non-trivial* regularizers, which was loosely defined earlier as a regularizer that's not upper bounded by 1. Here, however, I need to be more specific about what non-trivial means: specifically, ρ is non-trivial (relative to the partial prior information) if $\overline{\mathsf{Q}}\{\rho(\Theta) > 1\} > 0$. In words, non-triviality means that there exists a probability Q such that $\rho(\Theta)$ isn't Q-almost surely upper bounded by 1. Finally, I'll also say that a regularizer is *admissible* if $\overline{\mathsf{Q}}\rho = 1$, that is, if it can't be made larger in any substantive way without violating Definition 1.

The following is similar to Proposition 4.2 in Vovk and Wang (2021) on what they refer to as *ie-merging* functions for merging independent e-values. I show that the product mapping isn't strictly weakly dominated by any other re-merging functions.

Proposition 3. *Given* $\overline{\mathsf{Q}}$*, fix a non-trivial, admissible regularizer* ρ*. Then the product rule* (10) *isn't* $\overline{\mathbf{P}}$*-strictly weakly dominated by any other re-merging function.*

Proof. The proof is by contradiction; that is, I'll assume that the product rule defined in (10) is $\overline{\mathbf{P}}$-strictly weakly dominated in the sense above and show that this leads to a contradiction. Let m denote this assumed-to-exist dominant re-merging function.

The assumed non-triviality of the regularizer ρ implies existence of points θ such that $\rho(\theta) > 1$ and such that $\rho(\theta) \leq 1$, and neither of these sets have $\overline{\mathsf{Q}}$-probability 0. Some of the θ's in the first set could have $\rho(\theta) = \infty$, but, those must have $\overline{\mathsf{Q}}$-probability 0 for, otherwise, the property that $\overline{\mathsf{Q}}\,\rho \leq 1$ would be violated; recall that the assumed admissibility of ρ means that $\overline{\mathsf{Q}}\,\rho = 1$.

From the assumed strict weak dominance of m, and from (37), I can deduce the following bound for generic inputs (r, e):

$$\begin{aligned} m(r,e) &= m(r,e)\,\big(1_{r<1,e<1} + 1_{r\geq 1,e<1} + 1_{r<1,e\geq 1} + 1_{r\geq 1,e\geq 1}\big) \\ &\geq m(r,e)\,1_{r<1,e<1} + m(r,e)\,1_{r\geq 1,e<1} + r\,1_{r<1,e\geq 1} + re\,1_{r\geq 1,e\geq 1}, \end{aligned}$$

where the "r" factor in the third term is by (37) and the "re" factor in the fourth term is by the assumed weak dominance. The first two terms depend on details of the particular choice of m on the respective ranges of (r, e), details that can't be controlled with only the information provided. I can apply the trivial non-negativity bound, however, and, from the above display, conclude that

$$m(r, e) \geq r\, 1_{r<1, e\geq 1} + re\, 1_{r\geq 1, e\geq 1}.$$

(In fact, with the input e-process to be constructed next, those two terms I lower-bounded by 0 would typically be equal to 0, so the above inequality isn't loose.) The not-strict equality in the above display is pointwise in (r, e), i.e., I can't rule out equality above for any given pair (r, e). But a pointwise lower bound isn't the goal—I'm aiming for a lower bound in upper expectation. For this latter goal, I'll apply (40) to flip the not-strict inequality "$\geq$" in the above display a strict inequality; more on this below.

Towards establishing a contradiction, I only need to produce one example of an input e-process such that the conclusion is problematic, and I'll do this with an incredibly simple e-process. Specifically, I define the input e-process $\mathfrak{e}^\star$ as follows:

- if θ is such that $\rho(\theta) < 1$, then $\mathfrak{e}^\star_\theta(Z) \equiv 1$, and
- if θ is such that $\rho(\theta) \geq 1$, then

  $$\mathfrak{e}^\star_\theta(Z) = \begin{cases} 2 & \text{if } Z \in \mathcal{E}_\theta \\ 0 & \text{otherwise,} \end{cases}$$

  where $\mathcal{E}_\theta$ is an event with $\mathsf{P}_\theta(\mathcal{E}_\theta) = \frac{1}{2}$.

It's easy to check that $\mathsf{E}_\theta\{\mathfrak{e}^\star_\theta(Z)\} = 1$ for all θ, so $\mathfrak{e}^\star$ is a genuine e-process. What's important about this particular e-process, as it pertains to the present proof, is that

$$1_{\rho(\theta)<1, \mathfrak{e}^\star_\theta \geq 1} = 1_{\rho(\theta)<1} \quad \text{and} \quad \mathfrak{e}^\star_\theta\, 1_{\rho(\theta)\geq 1, \mathfrak{e}^\star_\theta \geq 1} = \mathfrak{e}^\star_\theta\, 1_{\rho(\theta)\geq 1}.$$

Therefore, from the pointwise analysis above,

$$\begin{aligned} m\{\rho(\Theta), \mathfrak{e}^\star_\Theta(Z)\} &\overset{(+)}{\geq} \rho(\Theta)\, 1_{\rho(\Theta)<1, \mathfrak{e}^\star_\Theta(Z)\geq 1} + \mathfrak{e}^\star_\Theta(Z)\, \rho(\Theta)\, 1_{\rho(\Theta)\geq 1, \mathfrak{e}^\star_\Theta(Z)\geq 1} \\ &= \rho(\Theta)\, 1_{\rho(\Theta)<1} + \mathfrak{e}^\star_\Theta(Z)\, \rho(\Theta)\, 1_{\rho(\Theta)\geq 1}. \end{aligned}$$

The $(+)$ symbol above is to remind the reader that, while "$\geq$" holds pointwise, there's actually more that can be said. That is, in addition to "$\geq$" for every (z, θ) pair, there exists a joint distribution for (Z, Θ), compatible with the available prior information, such that strict inequality holds with positive probability. This implies that the inequality "$\geq$" highlighted with $(+)$ is *strict inequality* "$>$" in expectation with respect to the aforementioned joint distribution. Then

$$\begin{aligned} \overline{\mathbf{P}}\big[m\{\rho(\Theta), \mathfrak{e}^\star_\Theta(Z)\}\big] &> \overline{\mathbf{P}}\big\{\rho(\Theta)\, 1_{\rho(\Theta)<1} + \mathfrak{e}^\star_\Theta(Z)\, \rho(\Theta)\, 1_{\rho(\Theta)\geq 1}\big\} \\ &= \sup_{\mathsf{Q}\in\mathscr{Q}} \mathsf{E}^{(Z,\Theta)\sim \mathsf{P}_\bullet \otimes \mathsf{Q}}\big\{\rho(\Theta)\, 1_{\rho(\Theta)<1} + \mathfrak{e}^\star_\Theta(Z)\, \rho(\Theta)\, 1_{\rho(\Theta)\geq 1}\big\} \\ &= \overline{\mathsf{Q}}\big\{\rho(\Theta)\, 1_{\rho(\Theta)<1} + \rho(\Theta)\, 1_{\rho(\Theta)\geq 1}\big\} \qquad (41) \\ &= \overline{\mathsf{Q}}\, \rho \\ &= 1, \qquad (42) \end{aligned}$$

where (41) holds because $\mathfrak{e}^\star$ is an e-process relative to the model, i.e., $\mathsf{E}_\theta(\mathfrak{e}^\star_\theta) = 1$ for all θ, and (42) holds by the assumed admissibility of the regularizer ρ. Therefore, the defining property (38) of an re-merging fails for the chosen m; that is, I constructed an e-process $(\mathfrak{e}^\star)$ such that $\overline{\mathbf{P}}[m\{\rho(\Theta), \mathfrak{e}^\star_\Theta(Z)\}] > 1$. Since m was assumed to be re-merging, this creates the desired contradiction. So, I conclude that, as claimed, there's no re-merging function m that $\overline{\mathbf{P}}$-strictly weakly dominates the product rule in (10). □

References

Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis*. Springer, Berlin, third edition. A hitchhiker's guide.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454.

Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proc. Royal Soc. A*, 475(2227):2018.0565.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā Ser. A*, 37(1):1–71. Discussion and correspondance between Barnard and Basu.

Bauer, H. (1958). Minimalstellen von Funktionen und Extremalpunkte. *Arch. Math.*, 9:389–393.

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 6. Institute of Mathematical Statistics, Hayward, CA.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 57:269–326.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in *nature* and *science* between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Cella, L. and Martin, R. (2023). Possibility-theoretic statistical inference offers performance and probativeness assurances. *Internat. J. Approx. Reason.*, 163:109060.

Choquet, G. (1953–1954). Theory of capacities. *Ann. Inst. Fourier, Grenoble*, 5:131–295 (1955).

Couso, I., Montes, S., and Gil, P. (2001). The necessity of the strong α-cuts of a fuzzy set. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 9(2):249–262.

Cuzzolin, F. (2021). *The Geometry of Uncertainty*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339.

Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.*, 48(2):365–377.

Denœux, T. (1999). Reasoning with imprecise belief structures. *Internat. J. Approx. Reason.*, 20(1):79–111.

Denœux, T. (2019). Decision-making with belief functions: a review. *Internat. J. Approx. Reason.*, 109:87–110.

Destercke, S., Dubois, D., and Chojnacki, E. (2008). Unifying practical uncertainty representations. I. Generalized p-boxes. *Internat. J. Approx. Reason.*, 49(3):649–663.

Dey, N., Martin, R., and Williams, J. P. (2024). Generalized universal inference on risk minimizers. `arXiv:2402.00202`.

Dixit, V. and Martin, R. (2024). Anytime valid and asymptotically optimal inference driven by predictive recursion. *Biometrika*, to appear; `arXiv:2309.13441`.

Dubois, D. (2006). Possibility theory and statistical reasoning. *Comput. Statist. Data Anal.*, 51(1):47–69.

Dubois, D., Foulloy, L., Mauris, G., and Prade, H. (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliab. Comput.*, 10(4):273–297.

Dubois, D. and Prade, H. (1988). *Possibility Theory.* Plenum Press, New York.

Dunn, J. (2015). Reliability for degrees of belief. *Philos. Stud.*, 172(7):1929–1952.

Durrett, R. (2010). *Probability: Theory and Examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition.

Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D., and Sentz, K. (2003). Constructing probability boxes and Dempster–Shafer structures. Technical Report SAND2002–4015, Sandia National Laboratories.

Gangrade, A., Rinaldo, A., and Ramdas, A. (2023). A sequential test for log-concavity. `arXiv:2301.03542`.

Goldman, A. I. (1979). Reliabilism: What is justified belief? In Pappas, G. S., editor, *Justification and Knowledge*, page 11. Reidel, Dordrecht, Holland.

Gong, R. and Meng, X.-L. (2021). Judicious judgment meets unsettling updating: dilation, sure loss and Simpson's paradox. *Statist. Sci.*, 36(2):169–190.

Good, I. J. (1950). *Probability and the Weighing of Evidence.* Charles Griffin & Co., Ltd., London; Hafner Publishing Co., New York.

Grünwald, P., de Heide, R., and Koolen, W. (2024). Safe testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, to appear; `arXiv:1906.07801`.

Grünwald, P. D. (2023). The e-posterior. *Philos. Trans. Roy. Soc. A*, 381(2247):Paper No. 20220146, 21.

Hose, D. (2022). *Possibilistic Reasoning with Imprecise Probabilities: Statistical Inference and Dynamic Filtering.* PhD thesis, University of Stuttgart. `https://dominikhose.github.io/dissertation/diss_dhose.pdf`.

Huber, P. J. (1973). The use of Choquet capacities in statistics. *Bull. Inst. Internat. Statist.*, 45(book 4):181–191. With discussion by P. Gänssler, W. S. Hall, L. Schmetterer, Charles Stein, and G. N. Wilkinson and with reply by the author.

Huber, P. J. (1981). *Robust Statistics.* John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Kass, R. E. (1992). Bayes factors in practice. *Amer. Statist.*, 42(5):551–560.

Kass, R. E. and Greenhouse, J. B. (1989). Comment on 'Investigating therapies of potentially great benefit: ECMO' by Ware (1989). *Statist. Sci.*, 4(4):310–317.

Levi, I. (1980). *The Enterprise of Knowledge.* The MIT Press, Boston.

Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *Internat. J. Approx. Reason.*, 113:39–73.

Martin, R. (2021a). An imprecise-probabilistic characterization of frequentist statistical inference. `arXiv:2112.10904`.

Martin, R. (2021b). Inferential models and the decision-theoretic implications of the validity property. `arXiv:2112.13247`.

Martin, R. (2022a). Valid and efficient imprecise-probabilistic inference with partial priors, I. First results. `arXiv:2203.06703`.

Martin, R. (2022b). Valid and efficient imprecise-probabilistic inference with partial priors, II. General framework. `arXiv:2211.14567`.

Martin, R. (2023). Valid and efficient imprecise-probabilistic inference with partial priors, III. Marginalization. `arXiv:2309.13454`.

Martin, R. (2024a). A possibility-theoretic solution to Basu's Bayesian–frequentist via media. *Sankhya A*, to appear, `arXiv:2303.17425`.

Martin, R. (2024b). Which statistical hypotheses are afflicted by false confidence? In Bi, Y., Jousselme, A.-L., and Denoeux, T., editors, *BELIEF 2024*, volume 14909 of *Lecture Notes in Artificial Intelligence*, pages 140–149, Switzerland. Springer Nature.

Martin, R. and Liu, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313.

Martin, R. and Liu, C. (2015). *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing*. Cambridge University Press, Cambridge.

Miranda, E. and de Cooman, G. (2014). Lower previsions. In *Introduction to Imprecise Probabilities*, Wiley Ser. Probab. Stat., pages 28–55. Wiley, Chichester.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506:150–152.

Pettigrew, R. (2021). What is justified credence? *Episteme*, 18(1):16–30.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.*, 38(4):576–601.

Ramdas, A. and Manole, T. (2023). Randomized and exchangeable improvements of Markov's, Chebyshev's and Chernoff's inequalities. `arXiv:2304.02611`.

Ramsey, F. P. (1990). Knowledge. In Mellor, D. H., editor, *Frank P. Ramsey: Philosophical Papers*, pages 110–111. Cambridge University Press, Cambridge.

Reid, N. and Cox, D. R. (2015). On some principles of statistical inference. *Int. Stat. Rev.*, 83(2):293–308.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.*, 55(1):62–71.

Shackle, G. L. S. (1961). *Decision Order and Time in Human Affairs*. Cambridge University Press, Cambridge.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Shafer, G. (2021). Testing by betting: a strategy for statistical and scientific communication. *J. Roy. Statist. Soc. Ser. A*, 184(2):407–478.

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statist. Sci.*, 26(1):84–101.

Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ.

Speed, T. (2016). Terence's Stuff: Principles. *IMS Bulletin*, page 17.

Sundberg, C. and Wagner, C. (1992). Characterizations of monotone and 2-monotone capacities. *J. Theoret. Probab.*, 5(1):159–167.

Tang, W. H. (2016). Reliability theories of justified credence. *Mind*, 125(497):63–94.

Troffaes, M. C. M. and de Cooman, G. (2014). *Lower Previsions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Turner, R. J. and Grünwald, P. D. (2023). Exact anytime-valid confidence intervals for contingency tables and beyond. *Statist. Probab. Lett.*, 198:Paper No. 109835, 8.

Varadarajan, V. S. (1963). Groups of automorphisms of Borel spaces. *Trans. Amer. Math. Soc.*, 109:191–220.

Vovk, V. and Wang, R. (2021). E-values: calibration, combination and applications. *Ann. Statist.*, 49(3):1736–1754.

Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics. *J. Roy. Statist. Soc. Ser. B*, 55(2):317–351. With discussion and a reply by the author.

Wald, A. (1947). *Sequential Analysis*. John Wiley & Sons, Inc., New York.

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

Walley, P. (1996). Inferences from multinomial data: Learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B*, 58(1):3–57. With discussion and a reply by the author.

Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. *J. Statist. Plann. Inference*, 105(1):35–65.

Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Statist. Sci.*, 4(4):298–340. With comments and a rejoinder by the author.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. U.S.A.*, 117(29):16880–16890.

Wasserman, L. A. (1990). Prior envelopes based on belief functions. *Ann. Statist.*, 18(1):454–464.

Wasserman, L. A. and Kadane, J. B. (1990). Bayes' theorem for Choquet capacities. *Ann. Statist.*, 18(3):1328–1339.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. I. *Information Sci.*, 8:199–249.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28.