Anytime valid and asymptotically optimal statistical inference driven by predictive recursion

Vaidehi Dixit* and Ryan Martin†

October 10, 2023

Abstract

Distinguishing two classes of candidate models is a fundamental and practically important problem in statistical inference. Error rate control is crucial to the logic but, in complex nonparametric settings, such guarantees can be difficult to achieve, especially when the stopping rule that determines the data collection process is not available. In this paper we develop a novel e-process construction that leverages the so-called predictive recursion (PR) algorithm designed to rapidly and recursively fit nonparametric mixture models. The resulting PRe-process affords anytime valid inference uniformly over stopping rules and is shown to be efficient in the sense that it achieves the maximal growth rate under the alternative relative to the mixture model being fit by PR. In the special case of testing for a log-concave density, the PRe-process test is computationally simpler and faster, more stable, and no less efficient compared to a recently proposed anytime valid test.

Keywords and phrases: e-process; mixture model; nonparametric; test martingale; universal inference.

1 Introduction

Arguably the most fundamental problem in mathematical statistics is that of distinguishing between two classes of candidate models based on observed data. When the two classes of models are *simple*, i.e., there are just two distinct probability distributions being compared, then the seminal work of Neyman and Pearson (1933) settles the question on how to optimally distinguish the two. More specifically, Neyman and Pearson proved that the most powerful test to distinguish between the two candidate distributions is based on the magnitude of the likelihood ratio. Beyond the test's statistical properties, the *law of likelihood* (e.g., Edwards 1992; Hacking 1976; Royall 1997) offers principled justification for carrying out the comparison between the two models in this fashion. In real applications, however, it is rare for the relevant comparison to be between two simple hypotheses, so the Neyman–Pearson lemma is generally insufficient to settle these practical questions. That is, when either of the two classes are *composite*, i.e., consisting

*Department of Statistics, University of Missouri, vdixit@missouri.edu

†Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu

of more than one probability distribution, then it is no longer clear how to define the likelihood ratio and to operationalize the law of likelihood.

To overcome this difficulty, there are two common strategies to quantify the "likelihood" of a composite hypothesis: the classical likelihood ratio testing, like that covered by Wilks's theorem (Wilks 1938), maximizes the likelihood over the respective models, whereas the Bayesian approach averages the likelihood with respect to suitable priors. In either case, judgments are made as above by considering the magnitude of the corresponding (marginal) likelihood ratio. For various reasons, however, simply thresholding these likelihood ratios is not fully satisfactory: classical strategies to calibration are based on sampling distribution calculations that assume a specific data-generating process, and so the reliability statements lack robustness to the kind of departures from these assumptions common in practice, e.g., data-peeking. And since neither of these likelihood ratios generally meet the conditions necessary (see below) to achieve a sort of universal calibration independent of the data-generating process, there is a need for new approaches.

To meet this need, there has been a recent surge of interest in testing procedures that are both valid and efficient under optional stopping; see Ramdas et al. (2022a) for a survey of these rapidly-moving developments. The basic building blocks are *e-values*, which in some cases lead to *test martingales* (Shafer et al. 2011) and other kinds of *e-processes*. These also have close connections to betting interpretations of probability (Shafer 2021) and, more generally, to *game-theoretic probability* (Shafer and Vovk 2019). A practical challenge, however, is that the mathematical definition of an e-process is not enough to determine what specifically to do in applications, i.e., there are so many e-processes for the data analyst to choose from. Two general strategies for the construction of e-values (which are sometimes e-processes) are *universal inference* (Dunn et al. 2023; Gangrade et al. 2023; Wasserman et al. 2020) and *reverse information projection* (Grünwald 2022, 2023; Grünwald et al. 2019). The former approach has different variants, but these can be computationally demanding in sequential settings and/or lacking in statistical efficiency. The latter strategy is efficiency-motivated, so much so that it can be difficult to apply in complex, nonparametric problems. So there is a need for an e-process construction that is simple—both conceptually and computationally—and makes efficient use of the available data. The present paper aims to fill this gap.

As indicated above, the familiar forms of likelihood ratios fail to be *anytime valid* in the sense of controlling errors independent of a data-generating process, stopping rule, etc. It turns out, however, as explained in Section 2.2 below, e-processes themselves are effectively likelihood ratios. Since the null model is determined by the problem itself, the e-process construction boils down to specification of the non-null model likelihood. To understand the considerations that affect this choice of non-null model likelihood, return again to the classical case covered by Neyman–Pearson. The most powerful test to distinguish the true distribution from a (false) simple null makes use of the true likelihood. Beyond this classical case, it is only natural for the likelihood in the e-process construction to match that true likelihood as closely as possible. Of course, the data analyst does not know the true likelihood, so his/her choice needs to be sufficiently flexible to adapt to what the data says about the underlying distribution. Mixture models are known to be quite flexible, so our proposal here is to build this non-null model likelihood using a nonparametric mixture model. Fitting such a mixture model can be computationally demanding using classical methods, but here we employ a fast, recursive

algorithm called *predictive recursion*, or PR for short (e.g., Newton 2002; Newton et al. 1998; Newton and Zhang 1999); see Section 2.3 for a brief review of PR. With this flexible and computationally efficient strategy to construct a non-null model likelihood, we define a suitable e-process, which we call the *PRe-process*, that can be used for anytime valid and efficient statistical inference.

Thanks to the recursive form of PR's mixture model fit, it immediately follows that the proposed PRe-process is upper-bounded by a genuine likelihood ratio, which is a test martingale under the null and hence has expected value upper-bounded by 1; see Theorem 1. Then the anytime validity, i.e., valid hypothesis tests and confidence regions uniformly over stopping rules that drive the data collection process, is an immediate consequence of Ville's inequality; see Corollaries 1–2. With finite-sample anytime validity guaranteed, we then turn our attention to the question of efficiency in Section 3.3. There, in Theorem 2, we establish that, under certain regularity conditions, the PRe-process achieves the optimal growth rate for true distributions in the alternative relative to the posited mixture model. And since our PRe-process proposal can itself be viewed as a variant of universal inference, our growth rate result also sheds light on the efficiency of universal inference more generally.

Two quick illustrations are presented in Section 4—one testing for a monotone density and the other testing a parametric null—where it is shown that the empirical growth rate of the PRe-process closely matches the asymptotically optimal growth rate established in Theorem 2. Then, in Section 5, in the context of testing for a log-concave density, we show that the PRe-process is no less efficient than the e-process-based test in Gangrade et al. (2023), more efficient than the e-value-based test proposed in Dunn et al. (2021), and computationally much more efficient than both the methods. Finally, some concluding remarks are made in Section 6 and technical details can be found in Appendix A.

2 Background

2.1 Notation

Suppose the data consists of a sequence $X_1, X_2, \ldots$ of random variables taking values in $\mathbb{X}$; write $X^n = (X_1, \ldots, X_n) \in \mathbb{X}^n$ for the first n entries in this sequence, with $n \geq 1$. Let $\mathscr{A}$ denote the σ-algebra of measurable subsets of $\mathbb{X}^\infty$ and write $\mathscr{A}_k = \sigma(X^k)$, for the filtration, the sequence of (sub-)σ-algebras generated by X^k, $k \geq 1$. An integer-valued random variable N will be called a *stopping time* if $\{N \leq n\} \in \mathscr{A}_n$.

Consider a collection $\mathscr{P}$ of candidate joint distributions for $X^\infty = (X_1, X_2, \ldots)$ on $\mathscr{A}$ and let $P = P^\infty$ denote a generic distribution in $\mathscr{P}$. Write $\mathscr{P}_0$ and $\mathscr{P}_1$, both subsets of $\mathscr{P}$, for the null and alternative hypotheses, respectively. Throughout, we use upper-case letters for distributions/probability measures and the corresponding lower-case letters for the associated density (with respect to some common dominating measure on $\mathbb{X}$). For example, if P_0 is a member of $\mathscr{P}_0$, then p_0 is its corresponding density.

2.2 E-processes

An *e-value*, E, for a distribution P_0, is a non-negative function of the observable data such that $\mathsf{E}_{P_0}(E) \leq 1$. When the expected value is equal to 1, then we can interpret E as

the realized payoff for a bet of \$1 against P_0; see, e.g., Shafer (2021). With batch data, say, X, and models that have densities, the non-negativity and expected value constraint implies that $x \mapsto q(x) := E(x)\, p_0(x)$ is a density too; therefore, $E(x) = q(x)/p_0(x)$ must be a likelihood ratio. With this connection to likelihood ratios, E can be understood as a measure of the evidence—"e" stands for evidence—in the data against P_0, relative to the alternative Q. The connection between the e-value and testing is now clear: if the observed value of E is large, then one would be inclined to reject the hypothesis P_0. A Bayes factor is an example of an e-value in the case of a simple hypothesis involving a single P_0. On the other hand, likelihood ratios of the form "$\sup_p p(x)/p_0(x)$" commonly used in mathematical statistics are not e-values.

The situation described above is impractically simple. Indeed, it is rare that $\mathscr{P}_0$ is a singleton and often data does not come in a batch. *Test martingales* (Shafer et al. 2011) offer a solution to the non-batch data problem. A test martingale is just a martingale (M_n), with respect to the distribution P_0, adapted to the filtration $(\mathscr{A}_n)$ with $M_0 \equiv 1$; consequently, $\mathsf{E}_P(M_n) = 1$ for all n and, by Doob's optional stopping theorem, $\mathsf{E}_P(M_N) = 1$ for all stopping times N. Test martingales can also be expressed as likelihood ratios, i.e., if Q is absolutely continuous with respect to P_0, then

$$M_n = \frac{q(X^n)}{p_0(X^n)} = \prod_{i=1}^n \frac{q(X_i \mid X^{i-1})}{p_0(X_i \mid X^{i-1})}, \quad n \geq 1. \tag{1}$$

So then the construction of a test martingale boils down to the choice of Q for the numerator. To handle composite $\mathscr{P}_0$ cases, one can construct a collection of test martingales $(M_n^{P_0})$ indexed by n and by $P_0 \in \mathscr{P}_0$. This has the same likelihood ratio representation as above but, mathematically, the choice of Q could vary with P_0. From here, one can define an *e-process* (E_n) to be a sequence of non-negative random variables with $E_n \leq M_n^{P_0}$ for all n and all $P_0 \in \mathscr{P}_0$. An example of this would be $E_n = \inf_{P_0 \in \mathscr{P}_0} M_n^{P_0}$ and, if it happened that the Q in (1) does not depend on P_0, then this simplifies to

$$E_n = \frac{q(X^n)}{\sup_{P_0 \in \mathscr{P}_0} p_0(X^n)}, \quad n \geq 1.$$

This is a common construction of an e-process in applications, e.g., this is how the variant of universal inference in Section 8 of Wasserman et al. (2020) works, but it is not the only way; see Ramdas et al. (2022a) for more discussion on the various alternatives, in particular, the proposals in Grünwald et al. (2019) and in Waudby-Smith and Ramdas (2022). We focus in this paper on e-processes that have the form above, i.e., only depend on a choice of Q representing the alternative.

Why all the fuss about e-processes? There are some very powerful results that tightly control the probabilistic behavior of these processes. We are referring to the classical result known as *Ville's inequality* (e.g., Shafer and Vovk 2019, Sec. 9.1). In our present case, this says that if M_n is a test martingale for $\mathscr{P}_0$, then

$$\sup_{P_0 \in \mathscr{P}_0} P_0\big(\text{there exists } n \geq 1 \text{ such that } M_n \geq \alpha^{-1}\big) \leq \alpha, \quad \text{all } \alpha \in (0,1),$$

or, equivalently (Howard et al. 2021, Lemma 3),

$$\sup_{P_0 \in \mathscr{P}_0} P_0(M_N \geq \alpha^{-1}) \leq \alpha, \quad \text{for all } \alpha \in (0,1) \text{ and all } N,$$

where N is a stopping time. Of course, the same holds true for any e-process E_n bounded by M_n. The implications of this are far-reaching: it allows for the construction of statistical procedures—hypothesis tests and confidence sets—that are *anytime valid* in the sense that the reliability claims hold (basically) no matter when or how the investigator decides to conclude their study and perform the statistical analysis, thereby offering some additional control or *safety* compared to procedures that only control the Type I error rate for a particular sampling scheme. For example, a test that rejects $\mathscr{P}_0$ based on data X^n when $E_n \geq \alpha^{-1}$ will control the Type I error rate *even if the investigator peeked at the data* when deciding whether to conclude the study at time n. More details on this are given in Ramdas et al. (2022a) and Section 3 below.

2.3 Predictive recursion

Here and in what follows, we will focus on the case where $X_1, X_2, \ldots$ are iid $\mathbb{X}$-valued random variables; possible extensions beyond the iid case will be discussed in Section 6. Following the notation above, a general mixture model for the (common) marginal distribution defines the density function as

$$q^{\Psi}(x) = \int_{\mathbb{U}} p_u(x) \, \Psi(du), \quad x \in \mathbb{X}, \tag{2}$$

where p_u is the density corresponding to a distribution P_u, with $\{P_u : u \in \mathbb{U}\} \subseteq \mathscr{P}$, and Ψ is a mixing distribution defined on (a suitable σ-algebra of subsets of) the indexing space $\mathbb{U}$. Such models are incredibly flexible, so they are often used in contexts where the mixture structure itself is not directly relevant, e.g., in density estimation applications. It is common to think of $x \mapsto p_u(x)$ as a kernel in some restricted parametric family, like Gaussian, but that is not necessary here; indeed, $\mathbb{U}$ could just be a generic indexing of the entire model $\mathscr{P}$. But even if $\{P_u : u \in \mathbb{U}\}$ is a relatively narrow parametric family, it is well known that mixtures thereof, as in (2), are still incredibly flexible.

The goal here is to fit the mixture model (2) to the observed data $X^n = (X_1, \ldots, X_n)$. One common strategy is nonparametric maximum likelihood, as discussed in Laird (1978), Lindsay (1995), and others. Another common strategy is to introduce a prior distribution for Ψ, e.g., a Dirichlet process prior (e.g., Ferguson 1973; Ghosal 2010; Lo 1984), and carry out a corresponding nonparametric Bayesian analysis. While there are advantages to sticking with the classical approaches, the complexity of these models creates serious computational challenges. In particular, since tailored Monte Carlo methods are required, one cannot capitalize on the Bayesian coherent updating property—"today's posterior is tomorrow's prior"—when data are processed sequentially. Maximum likelihood estimation faces similar challenges in the sequential data context. A novel alternative was developed by Michael Newton and collaborators in the late 1990s, namely, a fast, recursive algorithm called *predictive recursion*, or PR for short (e.g., Newton 2002), successfully applied in various large-scale inference settings, e.g., Tao et al. (1999), Newton et al. (2001), Martin and Tokdar (2012), Tansey et al. (2018), and Woody et al. (2022).

Here we provide a quick review of the PR algorithm and its properties; for more details, see Martin (2021). Start with two user-specified inputs: a weight sequence $(w_i :$

$i \geq 1) \subset (0,1)$ that satisfies

$$\sum_{i=1}^{\infty} w_i = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} w_i^2 < \infty, \tag{3}$$

and an initial guess Ψ_0 of Ψ supported on the index space $\mathbb{U}$. Then the PR algorithm updates the initial guess along the data sequence according to the rule

$$\hat{\Psi}_i(du) = (1 - w_i)\, \hat{\Psi}_{i-1}(du) + w_i \, \frac{p_u(X_i)\, \hat{\Psi}_{i-1}(du)}{\int_{\mathbb{U}} p_v(X_i)\, \hat{\Psi}_{i-1}(dv)}, \quad u \in \mathbb{U}, \quad i \geq 1. \tag{4}$$

When the data is X^n, the PR estimator $\hat{\Psi}_n$ is returned. Note that PR is recursive in the sense that one only needs $\hat{\Psi}_n$ and the new data point X_{n+1} to get the new $\hat{\Psi}_{n+1}$. This implies that the update $\hat{\Psi}_n \to \hat{\Psi}_{n+1}$ and, therefore, the update $\hat{q}_{x^n} \to \hat{q}_{x^{n+1}}$ is an $O(1)$ operation. Moreover, this leads naturally to a plug-in estimator of the mixture density

$$\hat{q}_{x^n}(x) := q^{\hat{\Psi}_n}(x) = \int_{\mathbb{U}} p_u(x)\, \hat{\Psi}_n(du), \quad x \in \mathbb{X}.$$

Large-sample consistency properties of the PR estimators have been explored in Ghosh and Tokdar (2006), Martin and Ghosh (2008), Tokdar et al. (2009), Martin and Tokdar (2009, 2011), and Dixit and Martin (2023b). The property most relevant to our developments here will be explained below in Section 3.

The above display can be interpreted as a data-driven predictive density for X_{n+1}, given X^n, not unlike the more familiar Bayesian predictive distribution sequence. The combination of being recursive and producing a predictive density explains the name *predictive recursion*. A point that will be especially relevant below is that PR also produces a joint marginal density for X^n—"marginal" in the sense that the mixing distribution Ψ has been integrated out—and it has a multiplicative form, à la Wald (1947, Eq. 10.10). That is, denoting this joint marginal density by $\hat{q}^{\text{PR}}(\cdot)$, it satisfies

$$\hat{q}^{\text{PR}}(x^n) = \hat{q}_{x^{n-1}}(x_n)\, \hat{q}^{\text{PR}}(x^{n-1}) = \prod_{i=1}^{n} \hat{q}_{x^{i-1}}(x_i), \quad n \geq 1, \quad x^n \in \mathbb{X}^n. \tag{5}$$

It is PR's flexibility and the multiplicative form of its joint marginal density, that makes it especially suitable for anytime valid nonparametric inference.

3 PRe-processes

3.1 Construction

As before, we have null and alternative hypotheses, $\mathscr{P}_0$ and $\mathscr{P}_1$, respectively. The proposal here is to construct a suitable marginal likelihood under the alternative by mixing over a specified class $\{P_u : u \in \mathbb{U}\}$ of distributions in $\mathscr{P}_1$—or perhaps in $\mathscr{P}_0$. One strategy would be to introduce a prior distribution supported on $\mathbb{U}$ and then find the corresponding Bayesian marginal likelihood. The challenge is that fitting a sufficiently flexible Bayesian mixture model to accommodate the "nonparametric" aspects of the

applications we have in mind would be computationally demanding. Here, instead, we make use of the PR algorithm reviewed in Section 2.3 above.

Let Ψ be a mixing probability distribution supported on the specified index set $\mathbb{U}$ and consider a basic mixture model with density

$$q(x) = \int_{\mathbb{U}} p_u(x)\, \Psi(du), \quad x \in \mathbb{X}.$$

This model can be fit to data X^n using the PR algorithm in (4) above, depending on a user-specified initial guess Ψ_0 and weights $(w_i : i \geq 1)$. We are not directly interested in estimation of the mixing distribution Ψ; our focus is on the joint marginal density for X^n that is produced as a by-product, namely, $\hat{q}^{\text{PR}}(X^n)$ as given in (5).

$$\hat{q}^{\text{PR}}(X^n) = \hat{q}_{X^{n-1}}(X_n)\, \hat{q}^{\text{PR}}(X^{n-1}) = \prod_{i=1}^{n} \hat{q}_{X^{i-1}}(X_i).$$

Note, again, the same key multiplicative form in Wald (1947, Eq. 10.10) that was the driving force behind the formulation in Section 8 of Wasserman et al. (2020). Moreover, the PR update from $\hat{q}^{\text{PR}}(X^{n-1})$ to $\hat{q}^{\text{PR}}(X^n)$ is an $O(1)$ computation, compared to the $O(n)$ computation one can expect with the "non-anticipatory" maximum likelihood estimation strategies employed in variants of universal inference (e.g., Gangrade et al. 2023).

With this, we define the following PR-driven e-process, i.e., *PRe-process*,

$$E_n^{\text{PR}} = E^{\text{PR}}(X^n; \mathscr{P}_0) = \frac{\hat{q}^{\text{PR}}(X^n)}{\sup_{p \in \mathscr{P}_0} p(X^n)}, \quad n \geq 1. \tag{6}$$

Intuitively, since likelihood tends to favor the true hypothesis, if H_0 is true (resp. false), then E_n^{PR} ought to be small (resp. large). This intuition suggests a test, i.e.,

$$\text{reject } H_0 \text{ if and only if } E_n^{\text{PR}} \text{ is large.}$$

We will be more specific about what it means to be "large" and the corresponding properties of the proposed test in Section 3.2 below.

Our proposal can be compared to two of the now-standard procedures for constructing e-processes reviewed in Ramdas et al. (2022a). Specifically, our proposal is, on the one hand, like the basic mixture method in that it forms a marginal likelihood under the alternative via a suitable mixture model, which we think is intuitively appealing. To achieve both the flexibility and the appealing intuition, the mixture model needs to be nonparametric, which would pose computational challenges for traditional likelihood-based methods. But the PR algorithm can easily and efficiently handle this major computational challenge since, again, the PRe-process updates when a new data point arrives are $O(1)$! So the PRe-process is like the (no data-splitting) variants of universal inference (Wasserman et al. 2020, Sec. 8), just with a specific focus on flexible mixture model alternatives with a fast, recursive updating scheme for fast, anytime valid inference. It is also statistically efficient, as we will demonstrate in Section 3.3.

Our proposal can also be compared to that in Grünwald et al. (2019). While our choice to use "$\sup_{p \in \mathscr{P}_0} p(X^n)$" in the denominator of (6) is a natural one, it is not the only option. Grünwald et al. propose a strategy based on the *reverse information projection*

(RIP), an idea which can be traced back at least to Li and Barron (2000), which amounts to replacing "$\sup_{p\in\mathscr{P}_0} p(X^n)$" in the denominator with the likelihood at a fixed—but strategically chosen—density, say, $p_0^\dagger$ in the convex hull, $\mathrm{co}(\mathscr{P}_0)$, of $\mathscr{P}_0$. In our context, they might propose to make inference based on the ratio

$$E_n^{\text{PR+RIP}} := \hat{q}^{\text{PR}}(X^n) \,/\, p_0^\dagger(X^n),$$

where $P_0^\dagger$ is the reverse information projection of $\hat{Q}^{\text{PR}}$ onto $\mathscr{P}_0$, i.e.,

$$\inf_{P_0\in\mathrm{co}(\mathscr{P}_0)} K(\hat{Q}^{\text{PR}}, P_0) = K(\hat{Q}^{\text{PR}}, P_0^\dagger).$$

This strategy has powerful motivation and a number of nice properties concerning maximal growth rate, etc. However, the complexity of the applications we have in mind, along with the complexity in PR's learning process raise some important and non-trivial questions: first, how to compute the required $P_0^\dagger$ and, second, whether $E_n^{\text{PR+RIP}}$ does actually define a proper e-process. We leave these questions open for future investigation.

3.2 Validity

Of course, we cannot refer to the quantity defined in (6) as an "e-process" without showing that it satisfies the required properties. Theorem 1 establishes the basic e-process property from which all the relevant statistical properties (Corollaries 1–2) follow.

Theorem 1. *Consider a model $\mathscr{P}$ for the iid data sequence $X_1, X_2, \ldots$. For a model $\mathscr{P}_0 \subset \mathscr{P}$ to be tested, the PRe-process defined in* (6) *is an e-process.*

Proof. For any fixed n, since the supremum over $\mathscr{P}_0$ appears in the denominator of (6), the following inequality is immediate:

$$E_n^{\text{PR}} = E^{\text{PR}}(X^n; \mathscr{P}_0) \leq E^{\text{PR}}(X^n; \{p_0\}) = \frac{\hat{q}^{\text{PR}}(X^n)}{p_0(X^n)}, \quad \text{for all } P_0 \in \mathscr{P}_0.$$

The upper bound is a collection of test martingales indexed by $P_0 \in \mathscr{P}_0$ and, therefore, (E_n^{PR}) is an e-process under the null $\mathscr{P}_0$. □

From this basic validity theorem, we can deduce a number of more directly interpretable statistical results. In particular, suitably-defined testing and confidence procedures control frequentist error rates under optional stopping.

Corollary 1. *For a null hypothesis $H_0 : P \in \mathscr{P}_0$ and alternative hypothesis $H_1 : P \in \mathscr{P}_1$, define the following testing procedures indexed by $\alpha \in [0, 1]$:*

$$T_\alpha(X^n) = \begin{cases} 1 & \text{if } E^{\text{PR}}(X^n; \mathscr{P}_0) \geq \alpha^{-1} \\ 0 & \text{otherwise.} \end{cases}$$

That is, the above test rejects H_0 if and only if the PRe-process exceeds α^{-1}. Then, under the setup of Theorem 1, the aforementioned test controls the frequentist Type I error at the designated level, i.e., for any stopping rule N,

$$\sup_{P_0\in\mathscr{P}_0} P_0\{T_\alpha(X^N) = 1\} \leq \alpha.$$

Proof. By definition, $T_\alpha(X^N) = 1$ if and only if $E^{\text{PR}}(X^N; \mathscr{P}_0) \geq \alpha^{-1}$. Theorem 1 together with Ville's inequality implies that the latter event has P_0-probability no more than α for all $P_0 \in \mathscr{P}_0$, which proves the claim. □

Given null and alternative models, $\mathscr{P}_0$ and $\mathscr{P}_1$, respectively, one can define a corresponding "anytime p-value"

$$\pi^{\text{PR}}(X^n; \mathscr{P}_0) = E^{\text{PR}}(X^n; \mathscr{P}_0)^{-1}.$$

It follows that, for all stopping times N,

$$\sup_{P \in \mathscr{P}_0} P\{\pi^{\text{PR}}(X^N; \mathscr{P}_0) \leq \alpha\} \leq \alpha, \quad \alpha \in [0,1].$$

So, as expected, the anytime-valid test proposed in Corollary 1 is equivalent to a test that rejects when the above p-value is no more than α.

By considering singleton null hypotheses $\mathscr{P}_0 = \{P_0\}$ and varying P_0, the above p-value defines a function $P_0 \mapsto \pi^{\text{PR}}(X^n; \{P_0\})$ which, as explained in Martin (2023), determines a full data-dependent imprecise probability distribution supported on $\mathscr{P}$. This, in turn, determines a valid *inferential model* (e.g., Martin 2022; Martin and Liu 2013, 2015) that provides reliable imprecise-probabilistic uncertainty quantification about the unknown P, but we will not comment further on this here.

One familiar and particularly relevant consequence of the aforementioned reliability is the construction of an anytime-valid confidence set. Let $\phi : \mathscr{P} \to \Phi$ be a map that extracts a relevant feature from $P \in \mathscr{P}$. For example, $\phi(P) = \int_{\mathbb{X}} x \, P(dx)$ is the mean of P, $\phi(P) = \inf\{x : P((-\infty, x]) \geq \tau\}$ is the τ^{th} quantile of P, and $\phi(P) = P$ is the distribution P itself. A valid confidence set for $\phi(P)$ is given below.

Corollary 2. *Given a relevant feature map $\phi : \mathscr{P} \to \Phi$, define the following data-dependent subset of Φ:*

$$C_\alpha(X^n) = \big\{\phi(P) : \pi^{\text{PR}}(X^n; \{P\}) > \alpha\big\}, \quad \alpha \in [0,1].$$

Under the setup of Theorem 1, for any stopping time N, the set $C_\alpha(X^N)$ is an anytime-valid $100(1-\alpha)\%$ confidence set in the sense that

$$\sup_{P \in \mathscr{P}} P\{C_\alpha(X^N) \not\ni \phi(P)\} \leq \alpha.$$

Proof. By definition, $C_\alpha(X^N) \not\ni \phi(P)$ if and only if $E^{\text{PR}}(X^N; \{P\}) \geq \alpha^{-1}$. Theorem 1 together with Ville's inequality implies that the latter event has P-probability no more than α, which proves the claim. □

3.3 Asymptotic growth rate optimality

As discussed above, the fact that E_n^{PR} is an e-process ensures its *validity*, i.e., that it provides frequentist error rate control. Validity, however, is a property relative to the distribution under the null H_0. For the e-process procedure to be *efficient*, one needs to consider its properties under the alternative. In particular, when the alternative is true, we want the e-value to be large, at least asymptotically, so that we will be inclined to

correctly reject the false null hypothesis. Hence, we aim to show that the PRe-process grows at the fastest possible rate, at least asymptotically. Towards this, consider the case where $P^\star \notin \mathscr{P}_0$ determines the true distribution of the data X^∞.

The main result in Martin and Tokdar (2011) states that, under some conditions (Appendix A below), with $P^\star$-probability 1,

$$n^{-1} \log \hat{q}^{\text{PR}}(X^n) = n^{-1} \log p^\star(X^n) - K(P^\star, \mathscr{Q}) + o(1), \quad n \to \infty, \tag{7}$$

where $K(P^\star, Q)$ is the Kullback–Leibler divergence of Q from $P^\star$ and

$$K(P^\star, \mathscr{Q}) = \inf_{Q \in \mathscr{Q}} K(P^\star, Q), \tag{8}$$

with $\mathscr{Q} = \text{co}(\{P_u : u \in \mathbb{U}\})$, the set of all mixtures having densities q^Ψ with the mixing distribution Ψ free to vary. Since minimizing Kullback–Leibler divergence $Q \mapsto K(P^\star, Q)$ over $\mathscr{Q}$ is what the maximum likelihood estimator aims to do, we can conclude from (7) that PR is effectively maximizing the mixture model likelihood asymptotically.

Next, concerning the null model $\mathscr{P}_0$, we assume that

$$\liminf_{n \to \infty} n^{-1} \log\{p^\star(X^n) / \hat{p}_{0,n}(X^n)\} \geq K(P^\star, \mathscr{P}_0), \quad \text{with } P^\star\text{-probability 1}, \tag{9}$$

where $\hat{p}_{0,n}(X^n) = \sup_{p \in \mathscr{P}_0} p(X^n)$ is the maximum likelihood estimator under the null $\mathscr{P}_0$. Of course, if $\mathscr{P}_0 = \{P_0\}$ is a singleton and $K(P^\star, P_0) < \infty$, then it follows from the standard law of large numbers that

$$n^{-1} \log\{p^\star(X^n) / p_0(X^n)\} \to K(P^\star, P_0) \quad \text{as } n \to \infty, \text{ with } P^\star\text{-probability 1}.$$

More generally, for a composite $\mathscr{P}_0$, a uniform law of large numbers establishes (9)—with equality and "lim inf" replaced by "lim"—if $\mathscr{P}_0$ is a Glivenko–Cantelli class (e.g., Kosorok 2008; van de Geer 2000; van der Vaart 1998; van der Vaart and Wellner 1996). Sections 4–5 consider two nonparametric problems: testing monotonicity and testing log-concavity. That these form suitable Glivenko–Cantelli classes is demonstrated in van der Vaart (1998, Example 19.11) and Doss and Wellner (2016, Theorem 3.1), respectively.

Putting everything together, we have that, with $P^\star$-probability 1, the PRe-process (6) satisfies the following asymptotic inequality:

$$\begin{aligned} n^{-1} \log E_n^{\text{PR}} &= n^{-1} \log\{\hat{q}^{\text{PR}}(X^n) / \hat{p}_{0,n}(X^n)\} \\ &= n^{-1} \log\{p^\star(X^n) / \hat{p}_{0,n}(X^n)\} + n^{-1} \log\{\hat{q}^{\text{PR}}(X^n) / p^\star(X^n)\} \\ &\geq K(P^\star, \mathscr{P}_0) - K(P^\star, \mathscr{Q}) + o(1). \end{aligned}$$

This effectively proves the following theorem; sufficient conditions for (7) and the detailed proof are provided in Appendix A below.

Theorem 2. *Consider $X_1, X_2, \ldots$ iid with distribution $P^\star$. If (7) and (9) hold, then with $P^\star$-probability 1, the PRe-process has asymptotic growth rate*

$$\Delta(P^\star; \mathscr{P}_0, \mathscr{Q}) = K(P^\star, \mathscr{P}_0) - K(P^\star, \mathscr{Q}). \tag{10}$$

That is, the PRe-process satisfies

$$E_n^{\text{PR}} \geq \exp\big[n\big\{\Delta(P^\star; \mathscr{P}_0, \mathscr{Q}) + o(1)\big\}\big], \quad n \to \infty, \quad P^\star\text{-probability 1}.$$

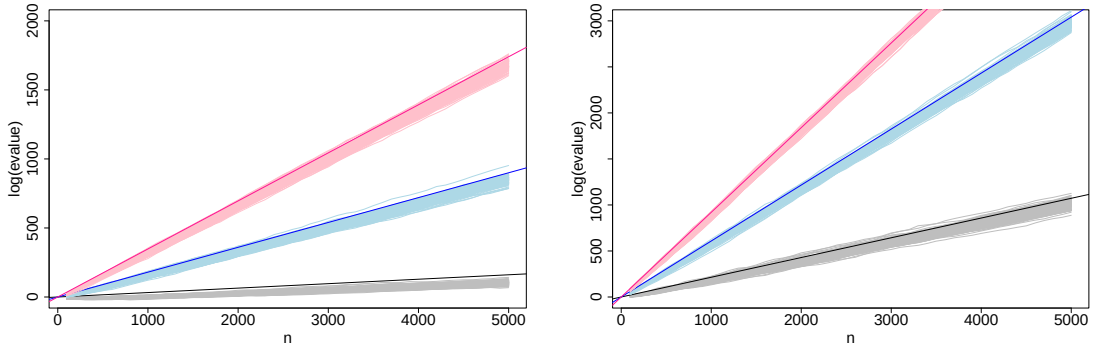
A few remarks concerning the above conclusions are in order. First, consider the case where $K(P^\star, \mathscr{P}_0) > 0$, which is our primary interest here. Our claim above is that the growth rate $\Delta(P^\star; \mathscr{P}_0, \mathscr{Q})$ is "optimal" which deserves some explanation. Start with the simple case of a simple-versus-simple test, i.e., $\mathscr{P}_0 = \{P_0\}$ and $\mathscr{P}_1 = \mathscr{Q} = \{P_1\}$. Then the usual likelihood ratio is an e-process and, if $P^\star = P_1$, then the growth rate is $K(P_1, P_0)$, which agrees with Δ in (10). That one cannot get a faster growth rate by replacing the alternative P_1 in the e-process by something else is a consequence of Gibbs's inequality (e.g., Shafer 2021). As Ramdas et al. (2022a) write: "in a simple versus simple test, the likelihood ratio is growth rate optimal." More generally, if we are committed to defining an e-process with "$\sup_{p_0 \in \mathscr{P}_0} p_0(X^n)$" in the denominator, then the leading term $K(P^\star, \mathscr{P}_0)$ in the growth rate (10) is unavoidable. In that case, the only input the user has in the construction is the choice of "$\hat{q}(X^n)$" in the numerator. If a commitment is made to choose "$\hat{q}$" in the convex hull $\mathscr{Q}$, then there is no choice of numerator that can achieve a larger growth rate than that in (10). This can also be compared to the growth rates in Grünwald et al. (2019). Indeed, if the true $P^\star$ was *known* and used as the "alternative" in the construction of their class of candidate e-values, then their solution is, by definition, growth rate optimal and its growth rate agrees with that in Theorem 2. The rate cannot be better when the true $P^\star$ is unknown, hence our PRe-process must be growth rate optimal too. Of course, other methods may have the same or similar asymptotic growth rate, but their rate cannot be faster than ours.

Second, suppose $K(P^\star, \mathscr{P}_0) = 0$ and $K(P^\star, \mathscr{Q}) > 0$, i.e., the null is definitively true. In this case, we actually want and expect the PRe-process to be small, even vanishing with n. Theorem 2 gets us close to this, the only snag is that it generally only gives a lower bound on the PRe-process, and a vanishing lower bound does not imply a vanishing PRe-process. However, if (9) holds with "lim inf" replaced by "lim" and "$\geq$" replaced by "=," as would often be the case, then Theorem 2 gives an exact asymptotic description of the behavior of E_n^{PR}. So, if the null is true, then indeed we get that the PRe-process is vanishing as $n \to \infty$, exactly as one would hope/expect. The same vanishing-under-the-null property holds more generally for non-negative martingales, as shown in Section 8.2 of Ramdas et al. (2022b). For the (literal) edge case where $K(P^\star, \mathscr{P}_0) = K(P^\star, \mathscr{Q}) = 0$, e.g., if $\mathscr{P}_0$ is the family of normal distributions and $\mathscr{Q}$ consists of mixtures thereof, then a $P^\star \in \mathscr{P}_0$ is also on the boundary of $\mathscr{Q}$, and Theorem 2 offers no guidance.

4 Illustrations

Here we consider two illustrations of the procedure described in Section 3: testing for monotonicity and testing a specific parametric null model. Validity of the PRe-process procedure is guaranteed by the general theory in Section 3.2, so the focus here is on efficiency, in particular, on comparing the empirical performance of the PRe-process procedure with that predicted by Theorem 2. So, in what follows, the data $X_1, X_2, \ldots$ will be generated from a distribution $P^\star$ that does not belong to $\mathscr{P}_0$.

To implement the PRe-process methodology, we need to compute the numerator and denominator of the ratio that defines E_n^{PR}. The denominator requires maximizing the likelihood function over the null, and for the examples we have in mind here, there are existing algorithms and software to carry this out. The numerator is where the PR algo-



(a) Testing monotonicity (b) Testing a parametric null

Figure 1: Plots of $\log E_n^{\text{PR}}$ versus n for 100 data sets of size $n = 5000$ under three settings of the true density in the alternative. Each case shows a line with slope based on the theoretical growth rate in Theorem 2. Panel (a) has $P^\star = \mathsf{Gamma}(\text{shape} = \gamma, \text{rate} = 1)$ with $\gamma = 2, 5, 10$ corresponding to gray, blue and pink bands, respectively, while Panel (b) has $P^\star$ a mixture of normals as in (11) with $\mu = 6, 10, 14$ corresponding to the gray, blue and pink bands respectively.

rithm is needed. It is well-known that location–scale Gaussian mixtures are quite flexible and this would be our general recommendation for the family $\mathscr{Q}$. There are exceptions, however, like the monotonicity example below where the data are supported on the positive half-line; in that case, of course, a Gaussian mixture would be inappropriate, but we can easily substitute in a similarly-flexible gamma mixture model. Going with the recommendation in Martin and Tokdar (2009), we take the weight sequence for PR as $w_i = (i+1)^{-0.67}$ for the examples here and in Section 5.

One important advantage of PR is its computational efficiency. The calculation in equation (6) can be done recursively as new observations become available. This aligns well with the "anytime" aspect of the solution. In both illustrations, the PRe-process is computationally efficient and the empirical performance closely matches the predictions made by the asymptotic efficiency results presented in Section 3.3.

Example 1. Monotone densities are often encountered in biomedical, engineering, and astronomy applications. The most common estimator of a monotone density is the nonparametric maximum likelihood estimator, also known as the *Grenander estimator* (e.g., Grenander 1956). Of course, there are many other estimators, including Bayesian estimators based on Dirichlet process (Salomond 2014) or Bernstein polynomial priors (Turnbull and Ghosh 2014), and even a method based on PR (Dixit and Martin 2023b). Similarly, a number of procedures are available for testing for monotonicity, including Hall and Heckman (2000), Ghosal et al. (2000), Chakraborty and Ghosal (2022).

Here we construct a new, anytime valid, PRe-process test for monotonicity. To compute the PRe-process, the Grenander estimator takes care of the denominator; in our case, we use the `Grenander` function in the R package `REBayes` (Koenker and Gu 2017). For the numerator, we use a mixture model with a gamma kernel, i.e., p_u is a gamma density with u consisting of the shape and rate/scale parameter pair. Then we use the PR algorithm to fit a mixture of p_u over $\mathbb{U} = [1, 15] \times [10^{-5}, 5]$ with the initial guess Ψ_0

taken to be a uniform distribution over $\mathbb{U}$ and weight sequence as given above.

For our experiments, the true distribution $P^\star$, which has a non-monotone density, is a gamma distribution with unit rate parameter and varying shape parameter. The idea is that, if the shape parameter is 1, then $P^\star$ would be exponential which is monotone; so as the shape parameter varies from 2, to 5, and to 10, the density becomes "less monotone." We generate 100 data sets of size 5000 from the true $P^\star$ and calculate the PRe-process E_n^{PR} at the increments $n = 100, 200, \ldots, 5000$. A plot of the $\log E_n^{\text{PR}}$ versus n is displayed in Figure 1(a), along with a reference line having slope $K(P^\star, \mathscr{P}_0)$. Note, first, that the growth rate increases as the true density gets "less monotone" and, second, that the PRe-process test attains the theoretical growth rate across all three scenarios.

Example 2. Next, consider a testing problem where the null lies on the boundary of the alternative. That is, the null hypothesis $\mathscr{P}_0$ is the Gaussian family indexed by their mean and variance parameters, while the alternative hypothesis $\mathscr{P}_1$ is the family of Gaussian mixtures. This is a challenging problem, as discussed in, e.g., Tokdar et al. (2010) and Tokdar and Martin (2021). Here we propose an anytime valid PRe-process test.

In this case, computation of the PRe-process's denominator is straightforward, since we have readily-available closed-form expressions for the maximum likelihood estimators under a normal model. For the numerator, let p_u denote a Gaussian kernel with u the mean and standard deviation pair, and apply the PR algorithm with support $\mathbb{U} = [-10, 20] \times [0.01, 3]$, and initial guess Ψ_0 uniform over $\mathbb{U}$. For our simulation, we generate data from a distribution $P^\star$, which has density

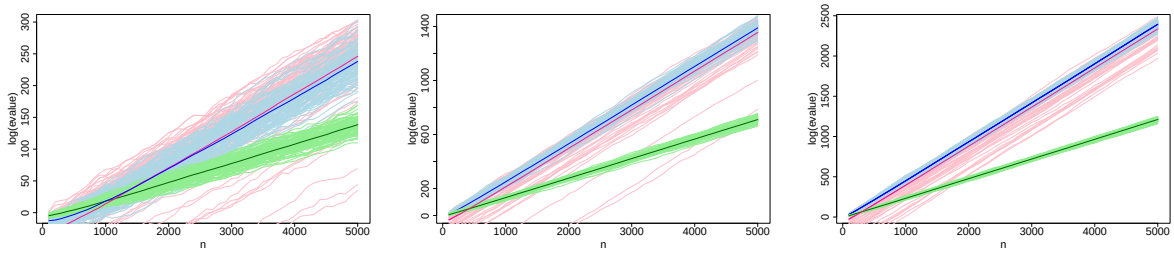
$$p^\star(x) = \tfrac{3}{4}\,\mathsf{N}(x \mid 0, 2) + \tfrac{1}{4}\,\mathsf{N}(x \mid \mu, 2), \quad x \in \mathbb{R}, \tag{11}$$

where, again, the distance between the two modes $\{0, \mu\}$ acts as a measure of the "degree of non-normality," with large distances corresponding to higher degree of non-normality. The three specific cases we consider here are $\mu = 6, 10, 14$. We generate 100 data sets of size 5000 under each configuration, calculate E_n^{PR} at increments $n = 100, 200, \ldots, 5000$. and plot $\log E_n^{\text{PR}}$ versus n in Figure 1(b), along with a reference line having slope $K(P^\star, \mathscr{P}_0)$. As expected, the slope increases in the "degree of non-normality," and the log-PRe-process paths closely follow this trend.

5 Comparison: testing log-concavity

The collection of log-concave densities on the real line includes the familiar Gaussian, logistic, and Laplace distribution, among others. Naturally, since log-concavity is a common structure, applications are abundant. Efficient maximum likelihood estimation techniques have been developed for estimating a log-concave density (e.g., Carpenter et al. 2018; Cule and Samworth 2010; Dümbgen and Rufibach 2011). As for valid testing of log-concavity, the literature is relatively scarce. Some recent advances have been made in Gangrade et al. (2023) and Dunn et al. (2021). Here, we present a test for log-concavity based on the PRe-process methodology described above.

For the PRe-process construction, first consider the denominator. Under the null model $\mathscr{P}_0$, we need the log-concave maximum likelihood estimator, which we get using the `logConDens` function from the R package `logcondens` (Dümbgen and Rufibach 2011). For the numerator, we consider a mixture model with a Gaussian kernel p_u, with u the



(a) $\mu = 6$ (b) $\mu = 10$ (c) $\mu = 14$

Figure 2: Plots of $\log E_n^{\mathrm{PR}}$ (blue), $\log E_n^{\mathrm{ULR}}$ (pink) and $\log E_n^{\mathrm{UI}}$ (green) versus n overlaid with the average line under each method based on the not-log-concave density $p^\star$ in (11). Panel (a) corresponds to $p^\star$ "closest" to log-concave while (c) is "farthest."

mean and standard deviation pair. For PR, we take $\mathbb{U} = [-10, 20] \times [0.01, 3]$, initial guess Ψ_0 a uniform distribution over $\mathbb{U}$, and weights w_i as above. An R package for PRe-process testing of log-concavity can be found at `https://github.com/vdixit005/PReprocess`. For Dunn et al.'s test statistic, which we denote by E_n^{UI}, we use the code available at `https://github.com/ RobinMDunn/LogConcaveUniv` with its default settings. For the test statistic proposed in Gangrade et al. (2023), which we denote by E_n^{ULR}, we follow their recommendation and use an iteratively fit Gaussian kernel density estimator.

For the simulation, we generate data from a bimodal distribution $P^\star$ with density in (11), where in this case μ acts as a measure of the "degree of non-log-concavity," i.e., if μ is close to 0, then $p^\star$ is only mildly non-log-concave; otherwise, $p^\star$ is more severely non-log-concave. The three cases we consider in our experiments are $\mu = 6, 10, 14$. We generate 100 data sets of size 5000 under each, calculate E_n^{PR}, E_n^{UI} and E_n^{ULR} at increments $n = 100, 200, \ldots, 5000$. Plots of $\log E_n$ versus n, along with a pointwise average over replications for each method are shown in Figure 2. The key takeaways are as follows. First, in terms of statistical efficiency, our PRe-process's growth rate is faster than Dunn et al.'s and no slower than Gangrade et al.'s in each scenario. Interestingly, the slopes of the n vs. $\log E_n^{\mathrm{UI}}$ lines in Figure 2 are roughly half that of the n vs. $\log E_n^{\mathrm{PR}}$ lines, which suggests that the loss of efficiency is due to the 50–50 data-splitting Dunn et al. employ. Second, in terms of stability, the log-PRe-process has considerably smaller variance than Gangrade et al.'s log-e-process. Third, in terms of computational efficiency, both E_n^{UI} and E_n^{ULR} need to process the entire observed data sequence each time a batch of data arrives because the previous calculation cannot be directly updated. The PRe-process, on the other hand, can be updated in $O(1)$ many steps when a new data point arrives.

6 Conclusion

In this paper we proposed a novel e-process construction based on the PR algorithm—a fast, recursive procedure for fitting flexible, nonparametric mixture models. It is shown that the corresponding PRe-process is a genuine e-process and, therefore, offers provable, finite-sample, anytime valid inference. Further, it is shown that the PRe-process attains the asymptotically optimal growth rate under the alternative relative to the posited mixture model fit by PR. Numerical results demonstrate that the PRe-process's empirical

performance closely agrees with the behavior predicted by the asymptotic growth rate optimality theorem and, moreover, that our PRe-process is no less efficient than a recently proposed, anytime valid test in Gangrade et al. (2023), but is more stable in terms of variance and computationally much more efficient.

The setup and theory presented here is general, but all of our illustrations consider only univariate data, i.e., testing for structure in a density function supported on the real line. The extension to the multivariate case is purely a computational problem—the bottleneck is PR's need to fit a mixture model over a relatively high-dimensional support. This is a challenge for the original PR algorithm where the normalization in (4) was based on quadrature. Recently, however, we developed a Monte Carlo-driven *PRticle filtering* algorithm (Dixit and Martin 2023a) that can readily accommodate multivariate mixture models. The incorporation of this Monte Carlo strategy into the PRe-process framework for anytime valid inference in multivariate settings is the focus of ongoing work.

Acknowledgments

Thanks to Aaditya Ramdas for feedback on and corrections to a previous version. This work is partially supported by the U.S. National Science Foundation, SES–2051225.

A Technical details

A.1 Setup of Theorem 2

There are a few inputs that play key roles in the asymptotic properties of the PR algorithm. These include some user-specified inputs, namely, the family of kernel densities $\{p_u : u \in \mathbb{U}\}$ that determine the set of mixtures $\mathscr{Q}$, and PR's weights (w_i) and initial guess Ψ_0; two more relevant quantities that are mostly, if not entirely, out of the user's control are the true distribution $P^\star$, with density $p^\star$, and the corresponding Kullback–Leibler projection of $P^\star$ onto the set $\mathscr{Q}$ of mixtures. This latter projection is defined as the distribution $Q^\star$ with density $q^\star$ that satisfies

$$K(P^\star, Q^\star) = K(P^\star, \mathscr{Q}) := \inf_{Q \in \mathscr{Q}} K(P^\star, Q).$$

The existence of $Q^\star$ is ensured by various sets of conditions; see, e.g., Liese and Vajda (1987, Ch. 8). In particular, existence of $Q^\star$ is implied by Condition 1 below, as shown in Lemma 3.1 of Martin and Tokdar (2009). The following conditions, on which Theorem 2 is based, concern the basic properties of these individual inputs and their interplay.

Condition 1. $\mathbb{U}$ is compact and $u \mapsto p_u(x)$ is continuous for almost all x.

Condition 2. The PR weight sequence (w_i) satisfies (3).

Condition 3. The kernel density $p_u(x)$ satisfies

$$\sup_{u_1, u_2 \in \mathbb{U}} \int \Big\{ \frac{p_{u_1}(x)}{p_{u_2}(x)} \Big\}^2 p^\star(x)\, dx < \infty \tag{12}$$

Condition 4. The Kullback–Leibler projection $Q^\star$ in (8), with density $q^\star$, satisfies

$$\int \Big\{ \frac{p^\star(x)}{q^\star(x)} \Big\}^2 p^\star(x)\, dx < \infty. \tag{13}$$

Here we offer some explanation and intuition. First, compactness of the mixing distribution support $\mathbb{U}$ in Condition 1 is difficult to relax, but the fact that $\mathbb{U}$ can be taken arbitrarily large means that this imposes effectively no practical constraints on the user. Continuity of the kernel can be relaxed, but at the expense of a much more complicated condition; the reader interested in this can consult Equation (7) in Dixit and Martin (2023b) and the relevant discussion. Condition 2 says simply that the weights must be vanishing to ensure convergence but not too quickly since the algorithm needs an opportunity to learn; the requirement in (3) is just right for this. Condition 3 is non-trivial but holds for Gaussian and other exponential family distributions thanks to the compactness of $\mathbb{U}$ in Condition 1. Finally, Condition 4 concerns the quality of the mixture model itself. One cannot hope to achieve quality estimation/inference in any sense if the "best" member of the mixture model differs considerably from the true density $p^\star$. Equation (13) is just a particular way to say that $p^\star$ and $q^\star$ do not differ by too much.

Once the user makes his/her specification of the mixture model, Condition 3 determines a set of true densities $p^\star$ for which the PR algorithm will provide consistent estimation. Indeed, Theorem 1 in Martin and Tokdar (2011) states that, under Conditions 1–3, the PR estimator $\hat{q}_{X^n}$ satisfies $K(p^\star, \hat{q}_{X^n}) \to \inf_{q \in \mathscr{Q}} K(p^\star, q)$ with $P^\star$-probability 1 as $n \to \infty$. The user, of course, can vary the mixture model specification to tailor the PR algorithm toward what they expect $p^\star$ to look like. But we need more than consistency for our purposes here, and Conditions 4 further restricts the set of true densities to those for which the PR algorithm can give us the "more" that we need here.

A.2 Proof of Theorem 2

Theorem 2 follows immediately from (7) and (9) as explained in the main text. The goal here is to show that (7) holds under Conditions 1–4 as stated above. The argument closely follows that in Martin and Tokdar (2011), but we provide the details here for completeness since their context and notation is different.

The strategy of the proof is as follows. First, simplify the notation by writing $\hat{q}_{i-1}(X_i) = \hat{q}_{X^{i-1}}(X_i)$ for each i. Next, define the sequence of random variables

$$K_n = \frac{1}{n} \sum_{i=1}^n \log \frac{p^\star(X_i)}{\hat{q}_{i-1}(X_i)},$$

which might be interpreted as a sort of empirical Kullback–Leibler divergence. If $K^\star = K(P^\star, \mathscr{Q})$, then (7) is equivalent to $K_n \to K^\star$ with $P^\star$-probability 1 as $n \to \infty$. It is this latter claim that we will prove below, using a martingale strong law in Teicher (1998).

Towards this, define a sequence of random variables Z_i as,

$$Z_i = \log \frac{p^\star(X_i)}{\hat{q}_{i-1}(X_i)} - K(p^\star, \hat{q}_{i-1}), \quad i \geq 1.$$

Recall that $\mathscr{A}_{i-1} = \sigma(X^{i-1})$, so $\mathsf{E}(Z_i \mid \mathscr{A}_{i-1}) = 0$ and, therefore, $\{(Z_i, \mathscr{A}_i) : i \geq 1\}$ is a zero mean martingale sequence. For $q^\star$ as in (8), we have

$$\begin{aligned}
\mathsf{E}(Z_i^2 \mid \mathscr{A}_{i-1}) &\leq \int \left\{ \log \frac{p^\star(x)}{\hat{q}_{i-1}(x)} \right\}^2 p^\star(x)\, dx \\
&= \int \left\{ \log \frac{q^\star(x)}{\hat{q}_{i-1}(x)} + \log \frac{p^\star(x)}{q^\star(x)} \right\}^2 p^\star(x)\, dx \\
&\leq 2 \int \left\{ \log \frac{q^\star(x)}{\hat{q}_{i-1}(x)} \right\}^2 p^\star(x)\, dx + 2 \int \left\{ \log \frac{p^\star(x)}{q^\star(x)} \right\}^2 p^\star(x)\, dx \\
&= 2A_i + 2B
\end{aligned}$$

Of course, B is a finite constant according to Condition 4. To bound A_i let us first define $\mathbb{X}_0 = \{x : q^\star(x) < \hat{q}_{i-1}(x)\}$. Using basic properties of the logarithm, we get

$$\begin{aligned}
A_i &= \int \left\{ \log \frac{q^\star(x)}{\hat{q}_{i-1}(x)} \right\}^2 p^\star(x)\, dx \\
&= \int_{\mathbb{X}_0} \left\{ \log \frac{\hat{q}_{i-1}(x)}{q^\star(x)} \right\}^2 p^\star(x)\, dx + \int_{\mathbb{X}_0^c} \left\{ \log \frac{q^\star(x)}{\hat{q}_{i-1}(x)} \right\}^2 p^\star(x)\, dx \\
&\leq \int_{\mathbb{X}_0} \left\{ \frac{\hat{q}_{i-1}(x)}{q^\star(x)} - 1 \right\}^2 p^\star(x)\, dx + \int_{\mathbb{X}_0^c} \left\{ \frac{q^\star(x)}{\hat{q}_{i-1}(x)} - 1 \right\}^2 p^\star(x)\, dx \\
&\leq 2 + \int \left[\left\{ \frac{\hat{q}_{i-1}(x)}{q^\star(x)} \right\}^2 + \left\{ \frac{q^\star(x)}{\hat{q}_{i-1}(x)} \right\}^2 \right] p^\star(x)\, dx.
\end{aligned}$$

Since both $\hat{q}_{i-1}$ and $q^\star$ in the two numerators in the above display are mixtures of the kernel p_u, we can say that

$$A_i \leq 2 + 2 \sup_{u_1, u_2 \in \mathbb{U}} \int \left\{ \frac{p_{u_1}(x)}{p_{u_2}(x)} \right\}^2 p^\star(x)\, dx$$

which is bounded by Condition 3. Therefore, $\mathsf{E}(Z_i^2 \mid \mathscr{A}_{i-1})$ is bounded too. Switching from index "i" to the more natural "n," we have that

$$\frac{\mathsf{E}(Z_n^2 \mid \mathscr{A}_{n-1})}{n^2 (\log\log n)^{-1}} \lesssim n^{-2} (\log\log n) \to 0.$$

Therefore, by Markov's inequality, with $P^\star$-probability 1 we have,

$$\sum_{n=1}^{\infty} P^\star\Big(|Z_n| > \frac{n}{\log\log n} \;\Big|\; \mathscr{A}_{n-1} \Big) \lesssim \sum_{n=1}^{\infty} \frac{(\log\log n)^2}{n^2} < \infty.$$

From this, it follows by Corollary 2 of Teicher (1998) that $n^{-1} \sum_{i=1}^n Z_i \to 0$ with $P^\star$-probability 1. Therefore, also with $P^\star$-probability 1,

$$\left| K_n - \frac{1}{n} \sum_{i=1}^n K(p^\star, \hat{q}_{i-1}) \right| = \left| (K_n - K^\star) - \frac{1}{n} \sum_{i=1}^n \{ K(p^\star, \hat{q}_{i-1}) - K^\star \} \right| \to 0.$$

From Theorem 1 in Martin and Tokdar (2011) and Cesaro's theorem, we have $K_n - K^\star \to 0$ with $P^\star$-probability 1.

References

Carpenter, T., Diakonikolas, I., Sidiropoulos, A., and Stewart, A. (2018). Near-optimal sample complexity bounds for maximum likelihood estimation of multivariate log-concave densities. In *Conference On Learning Theory*, pages 1234–1262. PMLR.

Chakraborty, M. and Ghosal, S. (2022). Rates and coverage for monotone densities using projection-posterior. *Bernoulli*, 28(2):1093–1119.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist*, 4:254–270.

Dixit, V. and Martin, R. (2023a). A prticle filter algorithm for nonparametric estimation of multivariate mixing distributions. *Stat. Comput.*, 33(4):1–14.

Dixit, V. and Martin, R. (2023b). Revisiting consistency of a recursive estimator of mixing distributions. *Electron. J. Stat.*, 17(1):1007–1042.

Doss, C. R. and Wellner, J. A. (2016). Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Statist.*, 44(3):954–981.

Dümbgen, L. and Rufibach, K. (2011). logcondens: Computations related to univariate log-concave density estimation. *J. Stat. Softw.*, 39:1–28.

Dunn, R., Gangrade, A., Wasserman, L., and Ramdas, A. (2021). Universal inference meets random projections: a scalable test for log-concavity. `arXiv:2111.09254`.

Dunn, R., Ramdas, A., Balakrishnan, S., and Wasserman, L. (2023). Gaussian universal likelihood ratio testing. *Biometrika*, 110(2):319–337.

Edwards, A. W. F. (1992). *Likelihood*. Johns Hopkins University Press, Baltimore, MD, expanded edition. Revised reprint of the 1972 original.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.

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

Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics*, Camb. Ser. Stat. Probab. Math., pages 35–79. Cambridge Univ. Press, Cambridge.

Ghosal, S., Sen, A., and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.*, 28(4):1054–1082.

Ghosh, J. K. and Tokdar, S. T. (2006). Convergence and consistency of Newton's algorithm for estimating mixing distribution. In Fan, J. and Koul, H., editors, *Frontiers in Statistics*, pages 429–443. Imp. Coll. Press, London.

Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).

Grünwald, P. (2022). Beyond Neyman–Pearson. arXiv:2205.00901.

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

Grünwald, P., de Heide, R., and Koolen, W. (2019). Safe testing. arXiv:1906.07801.

Hacking, I. (1976). *Logic of Statistical Inference*. Cambridge University Press, Cambridge-New York-Melbourne, paperback edition.

Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.*, 28(1):20–39.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.*, 49(2):1055–1080.

Koenker, R. and Gu, J. (2017). Rebayes: an R package for empirical Bayes mixture methods. *J. Stat. Softw*, 82:1–26.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.*, 73(364):805–811.

Li, J. Q. and Barron, A. R. (2000). Mixture density estimation. In Solla, S., Leen, T., and Mueller, K.-R., editors, *Advances in Neural Information Processing Systems*, pages 279–285. MIT Press, Cambridge, Massachusetts.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. IMS, Haywood, CA.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.*, 12(1):351–357.

Martin, R. (2021). A survey of nonparametric mixing density estimation via the predictive recursion algorithm. *Sankhya B*, 83(1):97–121.

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

Martin, R. (2023). Comment on Waudby-Smith & Ramdas's "Estimating means of bounded random variables by betting". https://www4.stat.ncsu.edu/~rmartin/comment_rm.pdf.

Martin, R. and Ghosh, J. K. (2008). Stochastic approximation and Newton's estimate of a mixing distribution. *Statist. Sci.*, 23(3):365–382.

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.

Martin, R. and Tokdar, S. T. (2009). Asymptotic properties of predictive recursion: robustness and rate of convergence. *Electron. J. Stat.*, 3:1455–1472.

Martin, R. and Tokdar, S. T. (2011). Semiparametric inference in mixture models with predictive recursion marginal likelihood. *Biometrika*, 98(3):567–582.

Martin, R. and Tokdar, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*, 13(3):427–439.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F., and Tsui, K. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biology*, 8:37–52.

Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhyā Ser. A*, 64(2):306–322.

Newton, M. A., Quintana, F. A., and Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In Dey, D., Müller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, volume 133 of *Lecture Notes in Statist.*, pages 45–61. Springer, New York.

Newton, M. A. and Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86(1):15–26.

Neyman, J. and Pearson, E. S. (1933). On the problem of most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. London Ser. A*, 231:289–337.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2022a). Game-theoretic statistics and safe anytime-valid inference. `arXiv:2210.01948`.

Ramdas, A., Ruf, J., Larsson, M., and W., K. (2022b). Admissible anytime-valid sequential inference must rely on nonnegative martingales. `arXiv:2009.03167`.

Royall, R. M. (1997). *Statistical Evidence*, volume 71 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Salomond, J.-B. (2014). Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Stat.*, 8(1):1380–1404.

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.

Tansey, W., Oluwasanmi, K., Poldrack, R. A., and Scott, J. G. (2018). False discovery rate smoothing. *J. Amer. Statist. Assoc.*, 113(523):1156–1171.

Tao, H., Palta, M., Yandell, B. S., and Newton, M. A. (1999). An estimation method for the semiparametric mixed effects model. *Biometrics*, 55(1):102–110.

Teicher, H. (1998). Strong laws for martingale differences and independent random variables. *J. Theoret. Probab.*, 11(4):979–995.

Tokdar, S. T., Chakrabarti, A., and Ghosh, J. K. (2010). Bayesian nonparametric goodness of fit tests. In Cheh, M.-H., Dey, D., Müller, P., Sun, D., and Ye, K., editors, *Frontiers of Statistical Decision Making and Bayesian Analysis*, pages 185–194. Springer.

Tokdar, S. T. and Martin, R. (2021). Bayesian test of normality versus a Dirichlet process mixture alternative. *Sankhya B*, 83(1):66–96.

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009). Consistency of a recursive estimate of mixing distributions. *Ann. Statist.*, 37(5A):2502–2522.

Turnbull, B. C. and Ghosh, S. K. (2014). Unimodal density estimation using bernstein polynomials. *Comput. Stat. Data. Anal.*, 72:13–29.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

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

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

Waudby-Smith, I. and Ramdas, A. (2022). Estimating means of bounded random variables by betting. *J. Roy. Statist. Soc. B.*, to appear; `arXiv:2010.09686`.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist*, 9:60–62.

Woody, S., Padilla, O. H. M., and Scott, J. G. (2022). Optimal post-selection inference for sparse signals: a nonparametric empirical Bayes approach. *Biometrika*, 109(1):1–16.