Universal inference

Larry Wasserman[a,b,1,2], Aaditya Ramdas[a,1], and Sivaraman Balakrishnan[a,1]

[a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213; and [b]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

Contributed by Larry Wasserman, May 26, 2020 (sent for review December 26, 2019; reviewed by Peter Bühlmann and Robert Tibshirani)

We propose a general method for constructing confidence sets and hypothesis tests that have finite-sample guarantees without regularity conditions. We refer to such procedures as "universal." The method is very simple and is based on a modified version of the usual likelihood-ratio statistic that we call "the split likelihood-ratio test" (split LRT) statistic. The (limiting) null distribution of the classical likelihood-ratio statistic is often intractable when used to test composite null hypotheses in irregular statistical models. Our method is especially appealing for statistical inference in these complex setups. The method we suggest works for any parametric model and also for some nonparametric models, as long as computing a maximum-likelihood estimator (MLE) is feasible under the null. Canonical examples arise in mixture modeling and shape-constrained inference, for which constructing tests and confidence sets has been notoriously difficult. We also develop various extensions of our basic methods. We show that in settings when computing the MLE is hard, for the purpose of constructing valid tests and intervals, it is sufficient to upper bound the maximum likelihood. We investigate some conditions under which our methods yield valid inferences under model misspecification. Further, the split LRT can be used with profile likelihoods to deal with nuisance parameters, and it can also be run sequentially to yield anytime-valid *P* values and confidence sequences. Finally, when combined with the method of sieves, it can be used to perform model selection with nested model classes.

likelihood | testing | irregular models | confidence sequence

The foundations of statistics are built on a variety of generally applicable principles for parametric estimation and inference. In parametric statistical models, the likelihood-ratio test and confidence intervals obtained from asymptotically Gaussian estimators are the workhorse inferential tools for constructing hypothesis tests and confidence intervals. Often, the validity of these methods relies on large sample asymptotic theory and requires that the statistical model satisfy certain regularity conditions; see *Section 2* for precise definitions. When these conditions do not hold, there is no general method for statistical inference, and these settings are typically considered in an ad hoc manner. Here, we introduce a universal method which yields tests and confidence sets for any statistical model and has finite-sample guarantees.

We begin with some terminology. A parametric statistical model is a collection of distributions $\{P_\theta : \theta \in \Theta\}$ for an arbitrary set Θ. When the aforementioned regularity conditions hold, there are many methods for inference. For example, if $\Theta \subseteq \mathbb{R}^d$, the set

$$A_n = \left\{ \theta : 2 \log \frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\theta)} \leq c_{\alpha,d} \right\} \qquad [1]$$

is the likelihood-ratio confidence set, where $c_{\alpha,d}$ is the upper α quantile of a χ^2_d distribution, $\mathcal{L}$ is the likelihood function, and $\widehat{\theta}$ is the maximum-likelihood estimator (MLE). It satisfies the asymptotic coverage guarantee

$$P_{\theta^*}(\theta^* \in A_n) \to 1 - \alpha$$

as $n \to \infty$, where P_{θ^*} denotes the unknown true data-generating distribution.

Constructing tests and confidence intervals for irregular models—where the regularity conditions do not hold—is very difficult (1). An example is mixture models. In this case we observe $Y_1, \ldots, Y_n \sim P$ and we want to test

$$H_0 : \ P \in \mathcal{M}_{k_0} \ \text{ versus } \ H_1 : \ P \in \mathcal{M}_{k_1}, \qquad [2]$$

where $\mathcal{M}_k$ denotes the set of mixtures of k Gaussians, with an appropriately restricted parameter space Θ (see for instance ref. 2) and with $k_0 < k_1$. Finding a test that provably controls the type I error at a given level has been elusive. A natural candidate is to base the test on the likelihood-ratio statistic but this turns out to have an intractable limiting distribution (3). As we discuss further in *Section 3*, developing practical, simple tests for this pair of hypotheses is an active area of research (refs. 4–6 and references therein). However, it is possible that we may be able to compute an MLE using variants of the expectation–maximization (EM) algorithm. In this paper, we show that there is a remarkably simple test based on the MLE with guaranteed finite-sample control of the type I error. Similarly, we construct a confidence set for the parameters of a mixture model with guaranteed finite-sample coverage. These tests and confidence sets can in fact be used for any model. In regular statistical models (those for which the usual LRT is well behaved), our methods may not be optimal, although we do not yet fully understand how close to optimal they are beyond special cases (uniform, Gaussian). Our test is most useful in irregular (or singular) models for which valid tests are not known or require many assumptions. Going beyond parametric models, we show that our methods can be used for several nonparametric models as well and have a natural sequential analog.

1. Universal Inference

Let $Y_1, \ldots, Y_{2n}$ be an independent and identically distributed (i.i.d.) sample from a distribution P_{θ^*} which belongs to a

Significance

Most statistical methods rely on certain mathematical conditions, known as regularity assumptions, to ensure their validity. Without these conditions, statistical quantities like *P* values and confidence intervals might not be valid. In this paper we give a surprisingly simple method for producing statistical significance statements without any regularity conditions. The resulting hypothesis tests can be used for any parametric model and for several nonparametric models.

Author contributions: L.W., A.R., and S.B. performed research and wrote the paper.

Reviewers: P.B., Eidgenössche Technische Hochschule; and R.T., Stanford University.

Competing interest statement: L.W. and R.T. are coauthors on a manuscript written in 2015 and published in 2018.

[1]L.W., A.R., and S.B. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: larry@stat.cmu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1922664117/-/DCSupplemental.

First published July 6, 2020.

collection $(P_\theta : \theta \in \Theta)$. Note that θ^* denotes the true value of the parameter. Assume that each distribution P_θ has a density p_θ with respect to some underlying measure μ (for instance, the Lebesgue or counting measure).

A Universal Confidence Set. We construct a confidence set for θ^* by first splitting the data into two groups D_0 and D_1. For simplicity, we take each group to be of the same size n but this is not necessary. Let $\widehat{\theta}_1$ be any estimator constructed from D_1; this can be the MLE, a Bayes estimator that utilizes prior knowledge, a robust estimator, etc. Let

$$\mathcal{L}_0(\theta) = \prod_{i \in D_0} p_\theta(Y_i)$$

denote the likelihood function based on D_0. We define the split likelihood-ratio statistic (split LRS) as

$$T_n(\theta) = \frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta)}. \quad [3]$$

Then, the universal confidence set is

$$C_n = \left\{ \theta \in \Theta : \ T_n(\theta) \le \frac{1}{\alpha} \right\}. \quad [4]$$

Similarly, define the cross-fit LRS as

$$S_n(\theta) = (T_n(\theta) + T_n^{\text{swap}}(\theta))/2, \quad [5]$$

where T_n^{swap} is formed by calculating T_n after swapping the roles of D_0 and D_1. We can also define C_n with S_n in place of T_n.

Theorem 1. *C_n is a finite-sample valid $(1-\alpha)$ confidence set for θ^*, meaning that $P_{\theta^*}(\theta^* \in C_n) \ge 1 - \alpha$.*

If we did not split the data and $\widehat{\theta}_1$ was the MLE, then $T_n(\theta)$ would be the usual likelihood-ratio statistic and we would typically approximate its distribution using an asymptotic argument. For example, as mentioned earlier, in regular models, -2 times the log-likelihood-ratio statistic has, asymptotically, a χ^2_d distribution. But, in irregular models this strategy can fail. Indeed, finding or approximating the distribution of the likelihood-ratio statistic is highly nontrivial in irregular models. The split LRS avoids these complications.

Now we explain why C_n has coverage at least $1-\alpha$, as claimed by *Theorem 1*. We prove it for the version using T_n, but the proof for S_n is identical. Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}. Then,

$$\begin{aligned}
\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\psi)}{\mathcal{L}_0(\theta^*)}\right] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i\in D_0} p_\psi(Y_i)}{\prod_{i\in D_0} p_{\theta^*}(Y_i)}\right] \\
&= \int_A \frac{\prod_{i\in D_0} p_\psi(y_i)}{\prod_{i\in D_0} p_{\theta^*}(y_i)} \prod_{i\in D_0} p_{\theta^*}(y_i)\, dy_1 \cdots dy_n \\
&= \int_A \prod_{i\in D_0} p_\psi(y_i)\, dy_1 \cdots dy_n \\
&\le \prod_{i\in D_0} \left[\int p_\psi(y_i)\, dy_i\right] = 1.
\end{aligned}$$

Since $\widehat{\theta}_1$ is fixed when we condition on D_1, we have

$$\mathbb{E}_{\theta^*}[T_n(\theta^*) \mid D_1] = \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta^*)} \,\middle|\, D_1\right] \le 1. \quad [6]$$

Now, using Markov's inequality,

$$\begin{aligned}
P_{\theta^*}(\theta^* \notin C_n) &= P_{\theta^*}\left(T_n(\theta^*) > \frac{1}{\alpha}\right) \le \alpha \mathbb{E}_{\theta^*}[T_n(\theta^*)] \quad [7] \\
&= \alpha \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta^*)}\right] = \alpha \mathbb{E}_{\theta^*}\left(\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta^*)} \,\middle|\, D_1\right]\right) \\
&\le \alpha.
\end{aligned}$$

Remark 2: The parametric setup adopted above generalizes easily to nonparametric settings as long as we can calculate a likelihood. For a collection of densities $\mathcal{P}$, and a true density $p^* \in \mathcal{P}$, suppose we use D_1 to identify $\widehat{p}_1 \in \mathcal{P}$ and D_0 to calculate

$$T_n(p) = \prod_{i \in D_0} \frac{\widehat{p}_1(Y_i)}{p(Y_i)}.$$

We then define $C_n := \{p \in \mathcal{P} : T_n(p) \le 1/\alpha\}$, and our previous argument ensures that $P_{p^*}(p^* \in C_n) \ge 1 - \alpha$.

A Universal Hypothesis Test. Now we turn to hypothesis testing. Let $\Theta_0 \subset \Theta$ be a possibly composite null set and consider testing

$$H_0 : \theta^* \in \Theta_0 \quad \text{versus} \quad \theta^* \notin \Theta_0. \quad [8]$$

The alternative above can be replaced by $\theta^* \in \Theta_1$ for any $\Theta_1 \subseteq \Theta$ or by $\theta^* \in \Theta_1 \backslash \Theta_0$. One way to test this hypothesis is based on the universal confidence set in Eq. **4**. We simply reject the null hypothesis if $C_n \bigcap \Theta_0 = \emptyset$. It is straightforward to see that if this test makes a type I error, then the universal confidence set must fail to cover θ^*, and so the type I error of this test is at most α.

We present an alternative method that is often computationally (and possibly statistically) more attractive. Let $\widehat{\theta}_1$ be any estimator constructed from D_1, and let

$$\widehat{\theta}_0 := \operatorname*{argmax}_{\theta \in \Theta_0} \mathcal{L}_0(\theta)$$

be the MLE under H_0 constructed from D_0. Then the universal test, which we call the split likelihood-ratio test (split LRT), is defined as

$$\text{reject } H_0 \text{ if } U_n > 1/\alpha, \text{ where } U_n = \frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\widehat{\theta}_0)}. \quad [9]$$

Similarly, we can define the cross-fit LRT as

$$\text{reject } H_0 \text{ if } W_n > 1/\alpha, \text{ where } W_n = \frac{U_n + U_n^{\text{swap}}}{2}, \quad [10]$$

where, as before, U_n^{swap} is calculated like U_n after swapping the roles of D_0 and D_1.

Theorem 3. *The split and cross-fit LRTs control the type I error at α; i.e., $\sup_{\theta^* \in \Theta_0} P_{\theta^*}(U_n > 1/\alpha) \le \alpha$.*

The proof is straightforward. We prove it for the split LRT, but once again the cross-fit proof is identical. Suppose that H_0 is true and $\theta^* \in \Theta_0$ is the true parameter. By Markov's inequality, the type I error is

$$\begin{aligned}
P_{\theta^*}(U_n > 1/\alpha) &= P_{\theta^*}\left(\mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\widehat{\theta}_0) > 1/\alpha\right) \\
&\le \alpha \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\widehat{\theta}_0)}\right] \overset{\text{(i)}}{\le} \alpha \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta^*)}\right] \overset{\text{(ii)}}{\le} \alpha.
\end{aligned}$$

Above, inequality (i) uses the fact that $\mathcal{L}_0(\widehat{\theta}_0) \ge \mathcal{L}_0(\theta^*)$ which is true when $\widehat{\theta}_0$ is the MLE, and inequality (ii) follows by conditioning on D_1 as argued earlier in Eq. **7**.

Remark 4. We may drop the use of $\Theta, \Theta_0, \Theta_1$ above and extend the split LRT to a general nonparametric setup. Both tests can be used to test any null $H_0 : p^* \in \mathcal{P}_0$ against any alternative $H_1 : p^* \in \mathcal{P}_1$. Importantly, no parametric assumption is needed on $\mathcal{P}_0, \mathcal{P}_1$, and no relationship is imposed whatsoever between $\mathcal{P}_0, \mathcal{P}_1$. As before, use D_1 to identify $\widehat{p}_1 \in \mathcal{P}_1$, use D_0 to calculate the MLE $\widehat{p}_0 \in \mathcal{P}_0$, and define $U_n = \prod_{i \in D_0} \frac{\widehat{p}_1(Y_i)}{\widehat{p}_0(Y_i)}$.

We call these procedures universal to mean that they are valid in finite samples with no regularity conditions. Constructions like this are reminiscent of ideas used in sequential settings where an estimator is computed from past data and the likelihood is evaluated on current data; we expand on this in *Section 7*.

We note in passing that another universal set is the following. Define $C = \left\{\theta : \int_\Theta \mathcal{L}(\psi) d\Pi(\psi) / \mathcal{L}(\theta) \leq 1/\alpha\right\}$, where $\mathcal{L}$ is the full likelihood (from all of the data) and Π is any prior. This also has the same coverage guarantee but requires specifying a prior and doing an integral. In irregular or nonparametric models, the integral will typically be intractable.

Perspective: Poor Man's Chernoff Bound. At first glance, the reader may worry that Markov's inequality seems like a weak tool to use, resulting in an underpowered conservative test or confidence interval. However, this is not the right perspective. One should really view our proof as using a "poor man's Chernoff bound."

For a regular model, we would usually compare the log-likelihood ratio to the $(1-\alpha)$ quantile of a χ^2 distribution (with degrees of freedom related to the difference in dimensionality of the null and alternate models). Instead, we compare the log-split-likelihood ratio to $\log(1/\alpha)$, which scales like the $(1-\alpha)$ quantile of a χ^2 distribution with one degree of freedom.

In any case, instead of finding the asymptotic distribution of $\log U_n$ (usually having a moment-generating function, like a χ^2), our proof should be interpreted as using the simpler but non-trivial fact that $\mathbb{E}_{\theta^*}[e^{\log(U_n)}] \leq 1$. Hence we are really using the fact that $\log U_n$ has an exponential tail, just as an asymptotic argument would.

A true Chernoff-style bound for a χ^2 random variable would have bounded $\mathbb{E}_{\theta^*}[e^{a \log(U_n)}]$ by an appropriate function of a and then optimized over the choice of $a > 0$ to obtain a tight bound. Our methods correspond to choosing $a = 1$, leading us to call the technique a poor man's Chernoff bound. The key point is that our methods should be viewed as using Markov's inequality on the exponential of the random variable of interest.

Perspective: In-Sample versus Out-of-Sample Likelihood. We may rewrite the universal set as

$$C_n = \left\{\theta \in \Theta : \; 2 \log \frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta)} \leq 2 \log(1/\alpha)\right\}.$$

For a regular model, it is natural to compare the above expression to the usual LRT-based set A_n from Eq. **1**. At first, it may visually seem like the LRT-based set uses the threshold $c_{\alpha,d}$, while the universal set uses $2\log(1/\alpha)$ which is much smaller in high dimensions. However, a key point to keep in mind is that comparing the numerators of the test statistics in both cases, the classical likelihood-ratio set uses an in-sample likelihood and the split LRS confidence set uses an out-of-sample likelihood. Hence, simply comparing the thresholds does not suffice to draw a conclusion about the relative sizes of the confidence sets. We next check that for regular models, the size of the universal set indeed shrinks at the right rate.

2. Sanity Check: Regular Models

Although universal methods are not needed for well-behaved models, it is worth checking their behavior in these cases. We expect that C_n would not have optimal size but we would hope that it still shrinks at the optimal rate. We now confirm that this is true.

Throughout this example we treat the dimension as a fixed constant before subsequently turning our attention to an example where we more carefully track the dependence of the confidence set diameter on dimension. In this and subsequent sections we use standard stochastic order notation for convergence in probability o_p and boundedness in probability O_p (7). We make the following regularity assumptions (see for instance ref. 7 for a detailed discussion of these conditions):

1) The statistical model is identifiable; i.e., for any $\theta \neq \theta^*$ it is the case that $P_\theta \neq P_{\theta^*}$. The statistical model is differentiable in quadratic mean (DQM) at θ^*; i.e., there exists a function s_{θ^*} such that

$$\int \left[\sqrt{p_\theta} - \sqrt{p_{\theta^*}} - \frac{1}{2}(\theta - \theta^*)^T s_{\theta^*} \sqrt{p_{\theta^*}}\right]^2 d\mu = o(\|\theta - \theta^*\|^2), \text{ as } \theta \to \theta^*.$$

2) The parameter space $\Theta \subset \mathbb{R}^d$ is compact, and the log-likelihood is a smooth function of θ; i.e., there is a measurable function ℓ with $\sup_\theta P_\theta \ell^2 < \infty$ such that for any $\theta_1, \theta_2 \in \Theta$

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \ell(x) \|\theta_1 - \theta_2\|.$$

3) A consequence of the DQM condition is that the Fisher information matrix

$$I(\theta^*) := \mathbb{E}_{\theta^*}[s_{\theta^*} s_{\theta^*}^T]$$

is well defined, and we assume it is nondegenerate.

Under these conditions the optimal confidence set has (expected) diameter $O(1/\sqrt{n})$. Our first result shows that the same is true of the universal set, provided that the initial estimate $\widehat{\theta}_1$ is $\sqrt{n}$ consistent; i.e., $\|\widehat{\theta}_1 - \theta^*\| = O_p(1/\sqrt{n})$. Under the conditions of our theorem, this consistency condition is satisfied when $\widehat{\theta}_1$ is the MLE but our result is more generally applicable.

Theorem 5. *Suppose that $\widehat{\theta}_1$ is a $\sqrt{n}$-consistent estimator of θ^*. Under the assumptions above, the split LRT confidence set has diameter $O_p(\sqrt{\log(1/\alpha)/n})$.*

A proof of this result is in *SI Appendix*. At a high level, to bound the diameter of the split LRT set it suffices to show that for any θ sufficiently far from θ^*, it is the case that

$$\frac{\mathcal{L}_0(\theta)}{\mathcal{L}_0(\widehat{\theta}_1)} \leq \alpha.$$

To establish this, note that we can write this condition as

$$\log \frac{\mathcal{L}_0(\theta)}{\mathcal{L}_0(\theta^*)} + \log \frac{\mathcal{L}_0(\theta^*)}{\mathcal{L}_0(\widehat{\theta}_1)} \leq \log(\alpha).$$

Bounding the first term requires showing if we consider any θ sufficiently far from θ^*, its likelihood is small relative to the likelihood of θ^*. We build on the work of Wong and Shen (8) who provide uniform upper bounds on the likelihood ratio under technical conditions which ensure that the statistical model is not too big. Conversely, to bound the second term we need to argue that if $\widehat{\theta}_1$ is sufficiently close to θ^*, then it must be the case that their likelihoods cannot be too different. This in turn follows by exploiting the DQM condition.

Analyzing the Nonparametric Split LRT. While our previous result focused on the diameter of the split LRT set in parametric problems, similar techniques also yield results in the nonparametric case. In this case, since we have no underlying parameter space, it will be natural to measure the diameter of our confidence set in terms of some metric on probability distributions. We consider bounding the diameter of our confidence set in the Hellinger metric. Formally, for two distributions P and Q the (squared) Hellinger distance is defined as

$$H^2(P, Q) = \frac{1}{2}\int \left(\sqrt{dP} - \sqrt{dQ}\right)^2.$$

We will also require the use of the χ^2 divergence given by

$$\chi^2(P, Q) = \int \left(\frac{dP}{dQ} - 1\right)^2 dQ,$$

assuming that P is absolutely continuous with respect to Q. Roughly, and analogous to our development in the parametric case, to bound the diameter of the split LRT confidence set, we need to ensure that our statistical model $\mathcal{P}$ is not too large and further that our initial estimate $\widehat{p}_1$ is sufficiently close to p^*.

To measure the size of $\mathcal{P}$ we use its Hellinger bracketing entropy. Denote by $\log N(u, \mathcal{F})$ the Hellinger bracketing entropy of the class of distributions $\mathcal{F}$ where the bracketing functions are separated by at most u in the Hellinger distance (we refer to ref. 8 for a precise definition). We suppose that the bracketing entropy of $\mathcal{P}$ is not too large; i.e., for some $\epsilon_n > 0$ we have that for some constant $c > 0$,

$$\int_{\epsilon_n^2}^{\epsilon_n} \sqrt{\log(N(u, \mathcal{P}))}\, du \le c\sqrt{n}\epsilon_n^2. \qquad [11]$$

Although we do not explore this in detail, we note in passing that the smallest value ϵ_n for which the above condition is satisfied provides an upper bound on the rate of convergence of the nonparametric MLE in the Hellinger distance (8). To characterize the quality of $\widehat{p}_1$ we use the χ^2 divergence. Concretely, we suppose that

$$\chi^2(p^*, \widehat{p}_1) \le O_p(\eta_n^2). \qquad [12]$$

Theorem 6. *Under conditions Eqs.* **11** and **12**, *the split LRT confidence set has Hellinger diameter upper bounded by* $O_p(\eta_n + \epsilon_n + \sqrt{\log(1/\alpha)/n})$.

Comparing LRT to Split LRT for the Multivariate Normal Case. In the previous calculation we treated the dimension of the parameter space as fixed. To understand the behavior of the method as a function of dimension in the regular case, suppose that $Y_1, \ldots, Y_n \sim N_d(\theta, I)$, where $\theta \in \mathbb{R}^d$. Recalling that we use $c_{\alpha,d}$ and z_α to denote the upper α quantiles of the χ^2_d and standard Gaussian, respectively, the usual confidence set for θ based on the LRT is

$$\begin{aligned} A_n &= \left\{\theta :\ \|\theta - \overline{Y}\|^2 \le \frac{c_{\alpha,d}}{n}\right\} \\ &= \left\{\theta :\ \|\theta - \overline{Y}\|^2 \le \frac{d + \sqrt{2d}z_\alpha + o(\sqrt{d})}{n}\right\}, \end{aligned}$$

where the second form follows from the normal approximation of the χ^2_d distribution. For the universal set, we use the sample average from D_1 as our initial estimate $\widehat{\theta}_1$. Denoting the sample means $\overline{Y}_1$ and $\overline{Y}_0$ we see that

$$C_n = \left\{\theta :\ \log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) \le \log(1/\alpha)\right\},$$

which is the set of θ such that

$$-\left(\frac{n}{2}\right)\frac{\|\overline{Y}_0 - \overline{Y}_1\|^2}{2} + \left(\frac{n}{2}\right)\frac{\|\theta - \overline{Y}_0\|^2}{2} \le \log\left(\frac{1}{\alpha}\right).$$

In other words, we may rewrite

$$C_n = \left\{\theta :\ \|\theta - \overline{Y}_0\|^2 \le \frac{4}{n}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}.$$

Next, note that $\|\overline{Y}_0 - \overline{Y}_1\|^2 = O_p(d/n)$, so both sets have radii $O_p(d/n)$. Precisely, the squared radius R_n^2 of C_n is

$$\begin{aligned} R_n^2 &\overset{d}{=} \frac{4\log(1/\alpha) + 4\chi^2_d}{n} \\ &\overset{d}{=} \frac{4\log(1/\alpha) + 4\ d + \sqrt{32\ d}\ Z + O_p(\sqrt{d})}{n}, \end{aligned}$$

where Z is an independent standard Gaussian. So both their squared radii share the same scaling with d and n, and for large d and constant α, the squared radius of C_n is about 4 times larger than that of A_n.

3. Examples

Mixture Models. As a proof of concept, we do a small simulation to check the type I error and power for mixture models. Specifically, let $Y_1, \ldots, Y_{2n} \sim P$, where $Y_i \in \mathbb{R}$. We want to distinguish the hypotheses in Eq. **2**. For this brief example, we take $k_0 = 1$ and $k_1 = 2$.

Finding a test that provably controls the type I error at a given level has been elusive. A natural candidate is the likelihood-ratio statistic but, as mentioned earlier, this has an intractable limiting distribution. To the best of our knowledge, the only practical test for the above hypothesis with a tractable limiting distribution is the EM test due to ref. 4. This very clever test is similar to the likelihood-ratio test except that it includes some penalty terms and requires the maximization of some of the parameters to be restricted. However, the test requires choosing some tuning parameters and, more importantly, it is restricted to one-dimensional problems. There is no known confidence set for mixture problems with guaranteed coverage properties. Another approach is based on the bootstrap (5) but there is no proof of the validity of the bootstrap for mixtures.

Fig. 1 shows the power of the test when $n = 200$ and $\widehat{\theta}_1$ is the MLE under the full model $\mathcal{M}_2$. The true model is taken to be $(1/2)\phi(y; -\mu, 1) + (1/2)\phi(y; \mu, 1)$, where ϕ is a normal density with mean μ and variance 1. The null corresponds to $\mu = 0$. We take $\alpha = 0.1$ and the MLE is obtained by the EM algorithm, which we assume converges on this simple problem. Understanding the local and global convergence (and nonconvergence) of the EM algorithm to the MLE is an active research area but is beyond the scope of this paper (refs. 9–11 and references therein). As expected, the test is conservative with type I error near 0 but has reasonable power when $\mu > 1$.

Fig. 1 also shows the power of the bootstrap test (5). Here, the P value is obtained by bootstrapping the LRS under the estimated null distribution. As expected, this has higher power than the universal test since it does not split the data. In this simulation, both tests control the type I error, but unfortunately the bootstrap test does not have any guarantee on the type I error, even asymptotically. The lower power of the universal test is the price paid for having a finite-sample guarantee. It is also worth noting that the bootstrap test requires running the EM algorithm for each bootstrap sample while the universal test requires only one EM run.

Model Selection Using Sieves. Sieves are a general approach to nonparametric inference. A sieve (12) is a sequence of nested

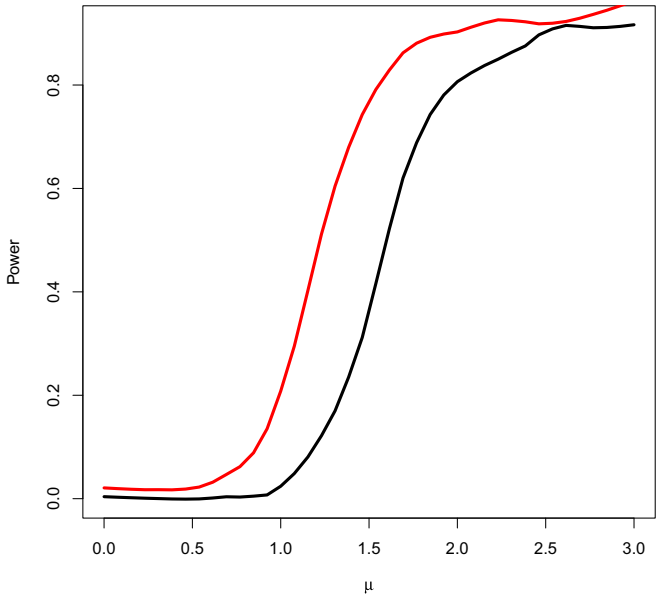

Fig. 1. The plot shows the power of the universal/bootstrap (black/red) tests for a simple Gaussian mixture, as the mean-separation μ varies ($\mu = 0$ is the null). The sample size is $n = 200$ and the target level is $\alpha = 0.1$.

models $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \cdots$. If we assume that the true density p^* is in $\mathcal{P}_j$ for some (unknown) j, then universal testing can be used to choose the model. One possibility is to test $H_j : p^* \in \mathcal{P}_j$ one by one for $j = 1, 2, \ldots$. We reject H_j if

$$\prod_{i \in D_0} \frac{\widehat{p}_{j+1}(Y_i)}{\widehat{p}_j(Y_i)} > 1/\alpha,$$

where $\widehat{p}_j$ is the MLE in model $\mathcal{P}_j$. Then we take $\widehat{j}$ to be the first j such that H_j is not rejected and proclaim that $p^* \in \mathcal{P}_j$ for some $j \geq \widehat{j}$. Even though we test multiple different hypotheses and stop at a random $\widehat{j}$, this procedure still controls the type I error, meaning that

$$P_{p^*}(p^* \in \mathcal{P}_{\widehat{j}-1}) \leq \alpha,$$

meaning that our proclamation is correct with high probability. The reason we do not need to correct for multiple testing is because a type I error can occur only once we have reached the first j such that $p^* \in \mathcal{P}_j$.

A simple application is to choose the number of mixture components in a mixture model, as discussed in the previous example. Here are some other interesting examples in which the aforementioned ideas yield valid tests and model selection using sieves: 1)testing the number of hidden states in a hidden Markov model (the MLE is computable using the Baum–Welch algorithm), 2) testing the number of latent factors in a factor model, and 3) testing the sparsity level in a high-dimensional linear model $Y = X\beta + \epsilon$ (under $H_0 : \beta$ is k sparse, the MLE corresponds to best-subset selection).

Whenever we can compute the MLE (specifically, the likelihood it achieves), then we can run our universal test, and we can do model selection using sieves. We will later see that an upper bound of the maximum likelihood suffices and is sometimes achievable by minimizing convex relaxations of the negative log-likelihood.

Nonparametric Example: Shape-Constrained Inference. A density p is log-concave if $p = e^g$ for some concave function g. Consider testing $H_0 :$ p is log-concave versus $H_1 :$ p is not log-concave. Let $\mathcal{P}_0$ be the set of log-concave densities and let $\widehat{p}_0$ denote the nonparametric maximum-likelihood estimator over $\mathcal{P}_0$ computed using D_0 (13) which can be computed in polynomial time (14). Let $\widehat{p}_1$ be any nonparametric density estimator such as the kernel density estimator (15) fitted on D_1. In this case, the universal test is to reject H_0 when

$$\prod_{i \in D_0} \frac{\widehat{p}_1(Y_i)}{\widehat{p}_0(Y_i)} > \frac{1}{\alpha}.$$

To the best of our knowledge this is the first test for this problem with finite-sample guarantee. Under the assumption that $p \in \mathcal{P}_0$, the universal confidence set is

$$C_n = \left\{ p \in \mathcal{P}_0 : \prod_{i \in D_0} p(Y_i) \geq \alpha \prod_{i \in D_0} \widehat{p}_1(Y_i) \right\}.$$

While the aforementioned test can be efficiently performed, the set C_n may be hard to explicitly represent, but we can check whether a distribution $p \in C_n$ efficiently.

Positive Dependence (Multivariate Total Positivity of Order 2). The split LRT solves a variety of open problems related to testing for a general notion of positive dependence called multivariate total positivity of order 2 (MTP_2) (16). The convex optimization problem of maximum-likelihood estimation in Gaussian models under total positivity was recently solved (17), but in ref. 17, example 5.8 and the following discussion, they state that the testing problem is still open. Given data from a multivariate distribution p, consider testing $H_0 : p$ is Gaussian MTP_2 against $H_1 : p$ is Gaussian (or an even more general alternative). Since proposition 2.2 in ref. 17 shows that the MLE under the null can be efficiently calculated, our universal test is applicable.

In fact, calculating the MLE in any MTP_2 exponential family is a convex optimization problem (ref. 18, theorem 3.1), thus making a test immediately feasible. As a particularly interesting special case, ref. 18, section 5.1 provides an algorithm for computing the MLE for MTP_2 Ising models. Testing $H_0 : p$ is Ising MTP_2 against $H_1 : p$ is Ising is stated as an open problem in ref. 18, section 6, and is solved by our universal test. (We remark that even though the MTP_2 MLE is efficiently computable, evaluating the maximum likelihood in the Ising case may still take $O(2^d)$ time for a d-dimensional problem.)

Finally, MTP_2 can be combined with log-concavity, uniting shape constraints and dependence. General existence and uniqueness properties of the MLE for totally positive log-concave densities have been recently derived (19), along with efficient algorithms to compute the MLE. Our methods immediately yield a test for $H_0 : p$ is MTP_2 log-concave against $H_1 : p$ is log-concave.

All of the above models were singular, and hence the LRS has been hard to study. In some cases, its asymptotic null distribution is known to be a weighted sum of χ^2 distributions, where the weights are rather complicated properties of the distributions (usually unknown to the practitioner). In contrast, the split LRT is applicable without assumptions, and its validity is nonasymptotic.

Independence versus Conditional Independence. Consider data that are trivariate vectors of the form (X_{1i}, X_{2i}, X_{3i}) which are modeled as trivariate normal. The goal is to test $H_0 :$ X_1 and X_2 are independent versus $H_1 :$ X_1 and X_2 are independent given X_3. The motivation for this test is that this problem arises in the construction of causal graphs. It is surprisingly difficult to test these nonnested hypotheses. Indeed, Guo and Richardson (20) study carefully the subtleties of the problem and they show

that the limiting distribution of the LRS is complicated and cannot be used for testing. They propose a new test based on a concept called envelope distributions. Despite the fact that the hypotheses are nonnested, the universal test is applicable and can be used quite easily for this problem. Further, one can also flip H_0 and H_1 and test for conditional independence in the Gaussian setting as well. We leave it to future work to compare the power of the universal test and the envelope test.

Cross-Fitting Can Beat Splitting: Uniform Distribution. In all previous examples, the split LRT is a reasonable choice. However, in this example, the cross-fit approach easily dominates the split approach. Note that this is a case where we would not recommend our universal tests since there are well-studied standard confidence intervals in this model. The example is just meant to bring out the difference between the split and cross-fit approaches.

Suppose that p_θ is the uniform density on $[0,\theta]$. Let us take $\widehat{\theta}_1$ to be the MLE from D_1. Thus, $\widehat{\theta}_1$ is the maximum of the data points in D_1. Now $\mathcal{L}_0(\theta)=\theta^{-n}I(\theta\geq\widehat{\theta}_0)$, where $\widehat{\theta}_0$ is the maximum of the data points in D_0. It follows that $C_n=[0,\infty)$ whenever $\widehat{\theta}_1<\widehat{\theta}_0$ which happens with probability 1/2. The set C_n has the required coverage but is too large to be useful. This happens because the densities have different support. A similar phenomenon occurs when testing $H_0:\theta\leq A$ versus $H_1:\theta\in\mathbb{R}^+$ for some fixed $A>0$, but not when testing against $H_1:\theta>A$. One can partially avoid this behavior by choosing $\widehat{\theta}_1$ to not be the MLE. However, the simplest way to avoid the degeneracy is to use the cross-fit approach, where we swap the roles of D_0 and D_1, and average the resulting test statistics. Exactly one of two test statistics will be 0, and hence the average will be nonzero. Further, it is easy to show that this test and resulting interval are rate optimal, losing a constant factor due to data splitting over the standard tests and interval constructions. In more detail, the classical (exact) pivotal $1-\alpha$ confidence interval for θ is $C'_{2n}=[\widehat{\theta},\widehat{\theta}(1/\alpha)^{1/(2n)}]$, where $\widehat{\theta}$ is the maximum of all of the data points. On the other hand, for $\widehat{\theta}_1,\widehat{\theta}_0$ defined above, assuming without loss of generality that $\widehat{\theta}_0\leq\widehat{\theta}_1$ a direct calculation shows that the cross-fit interval takes the form $C_n=[\widehat{\theta}_0,\widehat{\theta}_1(2/\alpha)^{1/n}]$. Ignoring constants, both these intervals have expected length $O(\theta\log(1/\alpha)/n)$.

4. Derandomization

The universal method involves randomly splitting the data and the final inferences will depend on the randomness of the split. This may lead to instability, where different random splits produce different results; in a related context, this has been called the "*P*-value lottery" (21).

We can get rid of or reduce the variability of our inferences, at the cost of more computation by using many splits, while maintaining validity of the method. The key property that we used in both the universal confidence set and the split LRT is that $\mathbb{E}_{\theta^*}[T_n]\leq 1$, where $T_n=\mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\widehat{\theta})$. Imagine that we obtained B such statistics $T_{n,1}\ldots,T_{n,B}$ with the same property. Let

$$\overline{T}_n=B^{-1}\sum_{j=1}^{B}T_{n,j}.$$

Then we still have that $\mathbb{E}_{\theta^*}[\overline{T}_n]\leq 1$ and so inference using our universal methods can proceed using the combined statistic $\overline{T}_n$. Note that this is true regardless of the dependence between the statistics.

Using the aforementioned idea, we can immediately design natural variants of the universal method:

- K-fold. We can split the data once into $2\leq K\leq n$ folds. Then repeat the following K times: Use $K-1$ folds to calculate $\widehat{\theta}_1$ and evaluate the likelihood ratio on the last fold. Finally, average the K statistics. Alternatively, we could use onefold to calculate $\widehat{\theta}_1$ and evaluate the likelihood on the other $K-1$ folds.
- Subsampling. We do not need to split the data just once into K folds. We can repeat the previous procedure for repeated random splits of the data into K folds. We expect this to reduce variance that arises from the algorithmic randomness.
- All splits. We can remove all algorithmic randomness by considering all possible splits. While this is computationally infeasible, the potential statistical gains are worth studying.

We remark that all these variants allow a large amount of flexibility. For example, in cross-fitting, $\widehat{\theta}_1$ need not be used the same way in both splits: It could be the MLE on one split, but a Bayesian estimator on another split. This flexibility could be useful if the user does not know which variant would lead to higher power in advance and would like to hedge across multiple natural choices. Similarly, in the K-fold version, if a user is confused whether to evaluate the likelihood ratio on onefold or on $K-1$ folds, then the user can do both and average the statistics.

Of course, with such flexibility comes the risk of an analyst cherry picking the variant used after looking at which form of averaging results in the highest LR (this would correspond to taking the maximum instead of the average of multiple variants), but this is a broader issue. For this reason (and this reason alone), the cross-fitting LRT proposed initially may be a useful default in practice, since it is both conceptually and computationally simple. We have already seen that (twofold) cross-fit inference improves over split inference drastically in the case of the uniform distribution discussed in the previous section. We leave a more detailed theoretical and empirical analysis of the power of these variants to future work.

5. Extensions

Profile Likelihood and Nuisance Parameters. Suppose that we are interested in some function $\psi=g(\theta)$. Let

$$B_n=\left\{\psi:\ C_n\bigcap g^{-1}(\psi)\neq\emptyset\right\},$$

where we define $g^{-1}(\psi)=\{\theta:g(\theta)=\psi\}$. By construction, B_n is a $1-\alpha$ confidence set for ψ. Defining the profile-likelihood function

$$\mathcal{L}_0^\dagger(\psi)=\sup_{\theta:\ g(\theta)=\psi}\mathcal{L}_0(\theta), \qquad [13]$$

we can rewrite B_n as

$$B_n=\left\{\psi:\ \frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0^\dagger(\psi)}\leq\frac{1}{\alpha}\right\}. \qquad [14]$$

In other words, the same data-splitting idea works for the profile likelihood too. As a particularly useful example, suppose $\theta=(\theta_u,\theta_n)$, where θ_n is a nuisance component; then we can define $g(\theta)=\theta_u$ to obtain a universal confidence set for only the component θ_u we care about.

Upper Bounding the Null Maximum Likelihood. Computing the MLE and/or the maximum likelihood (under the null) is sometimes computationally hard. Suppose one could come up with a relaxation F_0 of the null likelihood $\mathcal{L}_0$. This should be a proper relaxation in the sense that

$$\max_\theta F_0(\theta)\geq\max_\theta\mathcal{L}_0(\theta).$$

For example, $\mathcal{L}_0$ may be defined as $-\infty$ outside its domain, but F_0 could extend the domain. As another example, instead of minimizing the negative log-likelihood which could be nonconvex and hence hard to minimize, we could minimize a convex relaxation. In such settings, define

$$\widehat{\theta}_0^F := \operatorname*{argmax}_\theta F_0(\theta).$$

If we define the test statistic

$$T_n' := \frac{\mathcal{L}_0(\widehat{\theta}_1)}{F_0(\widehat{\theta}_0^F)},$$

then the split LRT may proceed using T_n' instead of T_n. This is because $F_0(\widehat{\theta}_0^F) \geq \mathcal{L}_0(\widehat{\theta}_0)$, and hence $T_n' \leq T_n$.

One particular case when this would be useful is the following. While discussing sieves, we had mentioned that testing the sparsity level in a high-dimensional linear model involves solving the best subset selection problem, which is nondeterministic polynomial-time hardness in the worst case. There exist well-known quadratic programming relaxations that are more computationally tractable. Another example is testing whether a random graph is a stochastic block model, for which semidefinite relaxations of the MLE are well studied (22); similar situations arise in communication theory (23) and angular synchronization (24).

The takeaway message is that it suffices to upper bound the maximum likelihood to perform inference.

Robustness via Powered Likelihoods. It has been suggested by some authors (25–29) that inferences can be made robust by replacing the likelihood $\mathcal{L}$ with the power likelihood $\mathcal{L}^\eta$ for some $0 < \eta < 1$. Note that

$$\mathbb{E}_\theta\left[\left(\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta)}\right)^\eta \;\middle|\; D_1\right] = \prod_{i\in D_0}\int p_{\widehat{\theta}_1}^\eta(y_i)p_\theta^{1-\eta}(y_i)dy_i \leq 1,$$

and hence all of the aforementioned methods can be used with the robustified likelihood as well. (The last inequality follows because the η-Renyi divergence is nonnegative.)

Smoothed Likelihoods. Sometimes the MLE is not consistent or it may not exist since the likelihood function is unbounded, and a (doubly) smoothed likelihood has been proposed as an alternative (30). For simplicity, consider a kernel $k(x, y)$ such that $\int k(x, y)dy = 1$ for any x, for example a Gaussian or Laplace kernel. For any density p_θ, let its smoothed version be denoted

$$\widetilde{p}_\theta(y) := \int k(x, y)p_\theta(x)dx.$$

Note that $\widetilde{p}_\theta$ is also a probability density. Denote the smoothed empirical density based on D_0 as

$$\widetilde{p}_n := \frac{1}{|D_0|}\sum_{i\in D_0} k(X_i, \cdot).$$

Define the smoothed maximum-likelihood estimator as the Kullback–Leibler (KL) projection of $\widetilde{p}_n$ onto $\{\widetilde{p}_\theta\}_{\theta\in\Theta_0}$,

$$\widetilde{\theta}_0 := \arg\min_{\theta\in\Theta_0} K(\widetilde{p}_n, \widetilde{p}_\theta),$$

where $K(P, Q)$ denotes the KL divergence between P and Q. If we define the smoothed likelihood on D_0 as

$$\widetilde{\mathcal{L}}_0(\theta) := \prod_{i\in D_0}\exp\int k(X_i, y)\log\widetilde{p}_\theta(y)dy,$$

then it can be checked that $\widetilde{\theta}_0$ maximizes the smoothed likelihood; that is, $\widetilde{\theta}_0 = \arg\max_{\theta\in\Theta_0}\widetilde{\mathcal{L}}_0(\theta)$. As before, let $\widehat{\theta}_1 \in \Theta$ be any estimator based on D_1. The smoothed split LRT is defined analogous to Eq. **9** as

$$\text{reject } H_0 \text{ if } \widetilde{U}_n > 1/\alpha, \text{ where } \widetilde{U}_n = \frac{\widetilde{\mathcal{L}}_0(\widehat{\theta}_1)}{\widetilde{\mathcal{L}}_0(\widetilde{\theta}_0)}. \qquad [15]$$

We now verify that the smoothed split LRT controls type I error. First, for any fixed $\psi \in \Theta$, we have

$$\begin{aligned}
\mathbb{E}_{\theta^*}\left[\frac{\widetilde{\mathcal{L}}_0(\psi)}{\widetilde{\mathcal{L}}_0(\widetilde{\theta}_0)}\right] &\overset{\text{(i)}}{\leq} \mathbb{E}_{\theta^*}\left[\frac{\widetilde{\mathcal{L}}_0(\psi)}{\widetilde{\mathcal{L}}_0(\theta^*)}\right] \\
&= \prod_{i\in D_0}\int \exp\left(\int k(x, y)\log\frac{\widetilde{p}_\psi(y)}{\widetilde{p}_{\theta^*}(y)}dy\right)p_{\theta^*}(x)dx \\
&\overset{\text{(ii)}}{\leq} \int\left(\int k(x, y)\frac{\widetilde{p}_\psi(y)}{\widetilde{p}_{\theta^*}(y)}dy\right)p_{\theta^*}(x)dx \\
&= \int\left(\frac{\int k(x, y)p_{\theta^*}(x)dx}{\widetilde{p}_{\theta^*}(y)}\right)\widetilde{p}_\psi(y)dy \\
&= \int\widetilde{p}_\psi(y)dy = 1.
\end{aligned}$$

Above, step (i) is because $\widetilde{\theta}_0$ maximizes the smoothed likelihood, and step (ii) follows by Jensen's inequality. An argument mimicking Eqs. **6** and **7** completes the proof. As a last remark, similar to the unsmoothed case, note that upper bounding the smoothed maximum likelihood under the null also suffices.

Conditional Likelihood for Non-i.i.d. Data. Our presentation so far has assumed that the data are drawn i.i.d. from some distribution under the null. However, this is not really required (even under the null) and was assumed for expositional simplicity. All that is needed is that we can calculate the likelihood on D_0 conditional on D_1 (or vice versa). For example, this could be tractable in models involving sampling without replacement from an urn with $M \gg n$ balls. Here θ could represent the unknown number of balls of different colors. Such hypergeometric sampling schemes result in non-i.i.d. data, but conditional on one subset of data (for example how many red, green, and blue balls were sampled from the urn in that subset), one can evaluate the conditional likelihood of the second half of the data and maximize it, rendering it possible to apply our universal tests and confidence sets.

6. Misspecification and Convex Model Classes

There are some natural examples of convex model classes (31, 32), including 1) all mixtures (potentially infinite) of a set of base distributions, 2) distributions with the first moment specified/bounded and possibly other moments bounded (e.g., first moment equals zero, second moment bounded by one), 3) the set of (coordinate-wise) monotonic densities with the same support, 4) unimodal densities with the same mode, 5) densities that are symmetric about the same point, 6) distributions with the same median or multiple quantiles (e.g., median = 0, 0.9 quantile = 2), 7) the set of all K-tuples $(P_1, \ldots, P_K)$ of distributions satisfying a fixed partial stochastic ordering (e.g., all triplets (P_1, P_2, P_3) such that $P_1 \preceq P_2$ and $P_1 \preceq P_3$, where $\preceq$ is the usual stochastic ordering), and 8) the set of convex densities with the same support. Some cases like 6) and 7) also result in weakly closed convex sets, as does case 2) for a specified mean. (Several of these examples also apply in discrete settings such as constrained multinomials.)

It is often possible to calculate the MLE over these convex model classes using convex optimization; for example see refs. 33 and 34 for case 7). This renders our universal tests and confidence sets immediately applicable. However, in this special case, it is also possible to construct additional tests, and the universal confidence set has some nontrivial guarantees if the model is misspecified.

Model Misspecification. Suppose the data come from a distribution Q with density $q \notin \mathcal{P}_\Theta \equiv \{p_\theta\}_{\theta\in\Theta}$, meaning that the model is misspecified and the true distribution does not belong to the considered model. In this case, what does the universal set C_n defined in Eq. **4** contain? We will answer this question when the set of measures/densities $\mathcal{P}_\Theta$ is convex. Define the Kullback–Leibler divergence of q from $\mathcal{P}_\Theta$ as

$$K(q, \mathcal{P}_\Theta) := \inf_{\theta\in\Theta} K(q, p_\theta).$$

Following definition 4.2 in Li's (31) PhD thesis, a function $p^* \equiv p^*_{q\to\Theta}$ is called the reversed information projection (RIPR) of q onto $\mathcal{P}_\Theta$ if for every sequence p_n with $K(q, p_n) \to K(q, \mathcal{P}_\Theta)$, we have $\log p_n \to \log p^*$ in $L^1(Q)$. Theorem 4.3 in ref. 31 proves that p^* exists and is unique, satisfies $K(q, p^*) = K(q, \mathcal{P}_\Theta)$, and

$$\forall\theta\in\Theta,\ \mathbb{E}_{Y\sim q}\left[\frac{p_\theta(Y)}{p^*(Y)}\right] \le 1. \qquad [16]$$

The above statement can be loosely interpreted as "if the data come from $q \notin \mathcal{P}_\Theta$, its RIPR p^* will have higher likelihood than any other model in expectation." We discuss this condition further at the end of this subsection.

It might be reasonable to ask whether the universal set contains p^*. For various technical reasons (detailed in ref. 31) it is not the case, in general, that p^* belongs to the collection $\mathcal{P}_\Theta$. Since the universal set considers densities in $\mathcal{P}_\Theta$ only by construction, it cannot possibly contain p^* in general. However, when p^* is a density in $\mathcal{P}_\Theta$, then it is indeed covered by our universal set.

Proposition 7. *Suppose that the data come from $q \notin \mathcal{P}_\Theta$. If $\mathcal{P}_\Theta$ is convex and there exists a density $p^* \in \mathcal{P}_\Theta$ such that $K(q, p^*) = \inf_{\theta\in\Theta} K(q, p_\theta)$, then we have $P_q(p^* \in C_n) \ge 1 - \alpha$.*

The proof is short. Examining the proof of *Theorem 1*, we must simply verify that for each $i \in D_0$, we have

$$\mathbb{E}_q\left[\frac{p_{\widehat{\theta}_1}(Y_i)}{p^*(Y_i)}\right] \le 1,$$

which follows from Eq. **16**. Here is a heuristic argument for why Eq. **16** holds when $p^* \in \mathcal{P}_\Theta$. For any $\theta \in \Theta$, note that $K(q, \mathcal{P}_\Theta) = K(q, p^*) = \min_{\alpha\in[0,1]} K(q, \alpha p^* + (1-\alpha)p_\theta)$ since $\mathcal{P}_\Theta$ is convex. The Karush–Kuhn–Tucker condition for this optimization problem is that gradient with respect to α is negative at $\alpha = 1$ (the minimizer). Exchanging derivative and integral immediately yields Eq. **16**. This argument is formalized in ref. 31, chap. 4.

An Alternate Split LRT (RIPR Split LRT). We return back to the well-specified case for the rest of this paper. First note that the fact in Eq. **16** can be rewritten as

$$\forall\theta\in\Theta,\ \mathbb{E}_{Y\sim p_\theta}\left[\frac{q(Y)}{p^*(Y)}\right] \le 1, \qquad [17]$$

which is informally interpreted as "if the data come from p_θ, then any alternative $q \notin \mathcal{P}_\Theta$ will have lower likelihood than its RIPR p^* in expectation." This motivates the development of an alternate RIPR split LRT to test composite null hypotheses that is defined as follows. As before, we divide the data into two parts, D_0 and D_1, and let $\widehat{\theta}_1 \in \Theta_1$ be any estimator found using only D_1. Now, define p_0^* to be the RIPR of $p_{\widehat{\theta}_1}$ onto the null set $\{p_\theta\}_{\theta\in\Theta_0}$. The RIPR split LRT rejects the null if

$$R_n \equiv \prod_{i\in D_0} \frac{p_{\widehat{\theta}_1}(Y_i)}{p_0^*(Y_i)} > 1/\alpha.$$

The main difference from the original MLE split LRT is that earlier we ignored $\widehat{\theta}_1$ and simply calculated the MLE $\widehat{\theta}_0$ under the null based on D_0.

Proposition 8. *If $\{p_\theta\}_{\theta\in\Theta}$ is a convex set of densities, then $\sup_{\theta_0\in\Theta_0} P_{\theta_0}(R_n > 1/\alpha) \le \alpha$.*

The fact that p_0^* is potentially not an element of $\{p_\theta\}_{\theta\in\Theta_0}$ does not matter here. The validity of the test follows exactly the same logic as the MLE split LRT, observing that Eq. **17** implies that for any true $\theta^* \in \Theta_0$, we have

$$\mathbb{E}_{p_{\theta^*}}\left[\frac{p_{\widehat{\theta}_1}(Y_i)}{p_0^*(Y_i)}\right] \le 1.$$

Without sample splitting and with a fixed alternative distribution, the RIPR LRT has been recently studied (35). When $\mathcal{P}_\Theta$ is convex and the RIPR split LRT is implementable, meaning that it is computationally feasible to find the RIPR or evaluate its likelihood, then this test can be more powerful than the MLE split LRT. Specifically, if the RIPR is actually a density in the null set, then

$$R_n = \prod_{i\in D_0} \frac{p_{\widehat{\theta}_1}(Y_i)}{p_0^*(Y_i)} \ge \prod_{i\in D_0} \frac{p_{\widehat{\theta}_1}(Y_i)}{p_{\widehat{\theta}_0}(Y_i)} = U_n,$$

since $\widehat{\theta}_0$ maximizes the denominator among null densities. Because of the restriction to convex sets, and since there exist many more subroutines to calculate the MLE over a set than to find the RIPR, the MLE split LRT is more broadly applicable than the RIPR split LRT.

7. Anytime *P* Values and Confidence Sequences

Just like the sequential likelihood-ratio test (36) extends the LRT, the split LRT has a simple sequential extension. Similarly, the confidence set can be extended to a "confidence sequence" (37).

Suppose the split LRT failed to reject the null. Then we are allowed to collect more data and update the test statistic (in a particular fashion) and check if the updated statistic crosses $1/\alpha$. If it does not, we can further collect more data and reupdate the statistic, and this process can be repeated indefinitely. Importantly we do not need any correction for repeated testing; this is primarily because the statistic is upper bounded by a nonnegative martingale. We describe the procedure next in the case when each additional dataset is of size one, but the same idea applies when we collect data in groups.

The Running MLE Sequential LRT. Consider the following, more standard, sequential testing/estimation setup. We observe an i.i.d. sequence $Y_1, Y_2, \ldots$ from P_{θ^*}. We want to test the hypothesis in Eq. **8**. Let $\widehat{\theta}_{1,t-1}$ be any nonanticipating estimator based on the first $t-1$ samples, for example the MLE, $\operatorname{argmax}_{\theta\in\Theta_1} \prod_{i=1}^{t-1} p_\theta(Y_i)$, or a regularized version of it to avoid misbehavior at small sample sizes. Denote the null MLE as

$$\widehat{\theta}_{0,t} = \underset{\theta \in \Theta_0}{\operatorname{argmax}} \prod_{i=1}^{t} p_\theta(Y_i).$$

At any time t, reject the null and stop if

$$M_t := \frac{\prod_{i=1}^{t} p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^{t} p_{\widehat{\theta}_{0,t}}(Y_i)} > 1/\alpha.$$

This test is computationally expensive: We must calculate $\widehat{\theta}_{1,t-1}$ and $\widehat{\theta}_{0,t}$ at each step. In some cases, these may be quick to calculate by warm starting from $\widehat{\theta}_{1,t-2}$ and $\widehat{\theta}_{0,t-1}$. For example, the updates can be done in constant time for exponential families, since the MLE is often a simple function of the sufficient statistics. However, even in these cases, the denominator takes time $O(t)$ to recompute at step t.

The following result shows that with probability at least $1-\alpha$, this test will never stop under the null. Let τ_θ denote the stopping time when the data are drawn from P_θ, which is finite only if we stop and reject the null.

Theorem 9. *The running MLE LRT has type I error at most* α, *meaning that* $\sup_{\theta^* \in \Theta_0} P_{\theta^*}(\tau_{\theta^*} < \infty) \le \alpha$.

The proof involves the simple observation that under the null, M_t is upper bounded by a nonnegative martingale L_t with initial value one. Specifically, define the (oracle) process starting with $L_0 := 1$ and

$$L_t := \frac{\prod_{i=1}^{t} p_{\widehat{\theta}_{i-1}}(Y_i)}{\prod_{i=1}^{t} p_{\theta^*}(Y_i)} \equiv L_{t-1} \frac{p_{\widehat{\theta}_{t-1}}(Y_t)}{p_{\theta^*}(Y_t)}. \qquad [18]$$

Note that under the null, we have $M_t \le L_t$ because $\widehat{\theta}_{0,t}$ and θ^* both belong to Θ_0, but the former maximizes the null likelihood (denominator). Further, it is easy to verify that L_t is a nonnegative martingale with respect to the natural filtration $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$. Indeed,

$$\begin{aligned}\mathbb{E}_{\theta^*}[L_t | \mathcal{F}_{t-1}] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i=1}^{t} p_{\widehat{\theta}_{i-1}}(Y_i)}{\prod_{i=1}^{t} p_{\theta^*}(Y_i)} \,\middle|\, \mathcal{F}_{t-1} \right] \\ &= L_{t-1} \mathbb{E}_{\theta^*}\left[\frac{p_{\widehat{\theta}_{t-1}}(Y_t)}{p_{\theta^*}(Y_t)} \,\middle|\, \mathcal{F}_{t-1} \right] = L_{t-1},\end{aligned}$$

where the last equality mimics Eq. **6**. To complete the proof, we note that the type I error of the running MLE LRT is simply bounded as

$$\begin{aligned}P_{\theta^*}(\exists t \in \mathbb{N} : M_t > 1/\alpha) &\le P_{\theta^*}(\exists t \in \mathbb{N} : L_t > 1/\alpha) \\ &\overset{\text{(i)}}{\le} \mathbb{E}_{\theta^*}[L_0] \cdot \alpha = \alpha,\end{aligned}$$

where step (i) follows by Ville's inequality (38, 39), a time-uniform version of Markov's inequality for nonnegative supermartingales.

Naturally, this test does not have to start at $t = 1$ when only one sample is available, meaning that we can set $M_0 = M_1 = \cdots = M_{t_0} = 1$ for the first t_0 steps and then begin the updates. Similarly, t need not represent the time at which the tth sample was observed; it can just represent the tth recalculation of the estimators (there may be multiple samples observed between $t-1$ and t).

Anytime-Valid *P* Values. We can also get a P value that is uniformly valid over time. Specifically, both $p_t = 1/M_t$ and $\bar{p}_t = \min_{s \le t} 1/M_s$ may serve as P values.

Theorem 10. *For any random time* T, *not necessarily a stopping time*, $\sup_{\theta^* \in \Theta_0} P_{\theta^*}(\bar{p}_T \le x) \le x$ *for* $x \in [0,1]$.

The aforementioned property is equivalent to the statement that under the null $P(\exists t \in \mathbb{N} : \bar{p}_t \le \alpha) \le \alpha$, and its proof follows by substitution immediately from the previous argument. Naturally $\bar{p}_t \le p_t$, but from the perspective of designing a level α test they are equivalent, because the first time that p_t falls below α is also the first time that $\bar{p}_t$ falls below α. The term "anytime-valid" is used because, unlike typical P values, these are valid at (data-dependent) stopping times or even random times chosen post hoc. Hence, inference is robust to "peeking," optional stopping, and optional continuation of experiments. Such anytime P values can be inverted to yield confidence sequences, as described below.

Confidence Sequences. A confidence sequence for θ^* is an infinite sequence of confidence intervals that are all simultaneously valid. Such confidence intervals are valid at arbitrary stopping times and also at other random data-dependent times that are chosen post hoc. In the same setup as above, but without requiring a null set Θ_0, define the running MLE likelihood-ratio process

$$R_t(\theta) := \frac{\prod_{i=1}^{t} p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^{t} p_\theta(Y_i)}.$$

Then, a confidence sequence for θ^* is given by

$$C_t := \{\theta : R_t(\theta) \le 1/\alpha\}.$$

In fact, the running intersection $\bar{C}_t = \bigcap_{s \le t} C_s$ is also a confidence sequence; note that $\bar{C}_t \subseteq C_t$.

Theorem 11. C_t *and* $\bar{C}_t$ *are confidence sequences for* θ^*, *meaning that* $P_{\theta^*}(\exists t \in \mathbb{N} : \theta^* \notin \bar{C}_t) \le \alpha$. *Equivalently,* $P_{\theta^*}(\theta^* \in C_\tau) \ge 1 - \alpha$ *for any stopping time* τ, *and also* $P_{\theta^*}(\theta^* \in C_T) \ge 1 - \alpha$ *for any arbitrary random time* T.

The proof is straightforward. First, note that $\theta^* \notin \bar{C}_t$ for some t if and only if $\theta^* \notin C_t$ for some t. Hence,

$$P_{\theta^*}(\exists t \in \mathbb{N} : \theta^* \notin C_t) = P_{\theta^*}(\exists t \in \mathbb{N} : R_t(\theta^*) > 1/\alpha) \le \alpha,$$

where the last step uses, as before, Ville's inequality for the martingale $R_t(\theta^*) \equiv L_t$ from Eq. **18**. The fact that the other two statements in *Theorem 11* are equivalent to the first one follows from recent work (40).

Duality. It is worth remarking that confidence sequences are dual to anytime P values, just like confidence intervals are dual to standard P values, in the sense that a $(1-\alpha)$ confidence sequence can be formed by inverting a family of level α sequential tests (each testing a different point in the space), and a level α sequential test for a composite null set Θ_0 can be obtained by checking whether the $(1-\alpha)$ confidence sequence intersects the null set Θ_0.

In fact, our constructions of p_t and C_t (without running minimum/intersection) obey the same property: $p_t < \alpha$ only if $C_t \cap \Theta_0 = \emptyset$, and the reverse implication follows if Θ_0 is closed. To see the forward implication, assume that there exists some element $\theta' \in C_t \cap \Theta_0$. Since $\theta' \in C_t$, we have $R_t(\theta') \le 1/\alpha$. Since $\theta' \in \Theta_0$, we have $\inf_{\theta^* \in \Theta_0} R_t(\theta^*) \le 1/\alpha$. This last condition can be restated as $M_t \le 1/\alpha$, which means that $p_t \ge \alpha$.

It is also possible to obtain an anytime P value from a family of confidence sequences at different α, by defining p_t as the smallest α for which $C_t \equiv C_t(\alpha)$ intersects Θ_0.

Extensions. All of the extensions from *Section 5* extend immediately to the sequential setting. One can handle nuisance parameters using profile likelihoods; this for example leads to sequential

t tests (for the Gaussian family, with the variance as a nuisance parameter), which also yield confidence sequences for the Gaussian mean with unknown variance. Non-i.i.d. data, such as in sampling without replacement, can be handled using conditional likelihoods, and robustness can be increased with powered likelihoods. In these situations, the corresponding underlying process L_t may not be a martingale, but a supermartingale. Also, as before, we may also use upper bounds on the maximum likelihood at each step (perhaps minimizing convex relaxations of the negative log-likelihood) or smooth the likelihood if needed.

Such confidence sequences have been developed under very general nonparametric, multivariate, matrix, and continuous-time settings using generalizations of the aforementioned supermartingale technique (39–41). The connections between anytime-valid P values, e values, safe tests, peeking, confidence sequences, and the properties of optional stopping and continuation have been explored recently (35, 40, 42, 43). The connection to the present work is that when run sequentially, our universal (MLE or RIPR) split LRT yields an anytime-valid P value, an e value, and a safe test, which can be inverted to form universal confidence sequences and are valid under optional stopping and continuation, and these are simply because the underlying process of interest is bounded by a nonnegative (super)martingale. This line of research began over 50 y ago by Darling and Robbins (37), Robbins (44), Robbins and Siegmund (45), and Lai (46, 47). In fact, for testing point nulls, the running MLE (or nonanticipating) martingale was suggested in passing by Wald (ref. 48, equation 10:10) and analyzed in depth by refs. 45 and 49 where connections were shown to the mixture sequential probability-ratio test. These ideas have been utilized in changepoint detection for both point nulls (50) and composite nulls (51).

8. Conclusion

Inference based on the split likelihood-ratio statistic (and variants) leads to simple tests and confidence sets with finite-sample guarantees. Our methods are most useful in problems where standard asymptotic methods are difficult/impossible to apply, such as complex composite null testing problems or nonparametric confidence sets. Going forward, we intend to run simulations in a variety of models to study the power of the test and the size of the confidence sets and study their optimality in special cases. We do not expect the test to be rate optimal in all cases, but it might have analogous properties to the generalized LRT. It would also be interesting to extend these methods (like the profile-likelihood variant) to semiparametric problems where there are a finite-dimensional parameter of interest and an infinite-dimensional nuisance parameter.

9. Data Availability

Due to space constraints, we have relegated technical details of the proofs of *Theorems* 5 and *6* to *SI Appendix*. There are no additional data, protocols, or code associated with this paper.

ACKNOWLEDGMENTS. We thank Caroline Uhler and Arun K. Kuchibhotla for references to open problems in shape-constrained inference and Ryan Tibshirani for suggesting the relaxed-likelihood idea. We are grateful to Bin Yu, Hue Wang and Marco Molinaro for helpful feedback which motivated parts of *Section 6*. We thank the reviewers and Dennis Boos for helpful suggestions and Len Stefanski for pointing us to work on smoothed likelihoods.

1. M. Drton, Likelihood ratio tests and singularities. *Ann. Stat.* **37**, 979–1012 (2009).
2. R. Redner, Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Stat.* **9**, 225–228 (1981).
3. D. Dacunha-Castelle, E. Gassiat, Testing in locally conic models, and application to mixture models. *ESAIM Probab. Stat.* **1**, 285–317 (1997).
4. J. Chen, P. Li, Hypothesis test for normal mixture models: The EM approach. *Ann. Stat.* **37**, 2523–2542 (2009).
5. G. J. McLachlan, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat.* **36**, 318–324 (1987).
6. P. Chakravarti, S. Balakrishnan, L. Wasserman, Gaussian mixture clustering using relative tests of fit. arXiv:1910.02566 (7 October 2019).
7. A. W. Van der Vaart, *Asymptotic Statistics* (Cambridge University Press, 2000), vol. 3.
8. W. H. Wong, X. Shen, Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Stat.* **23**, 339–362 (1995).
9. S. Balakrishnan, M. J. Wainwright, B. Yu, Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Stat.* **45**, 77–120 (2017).
10. J. Xu, D. J. Hsu, A. Maleki, "Global analysis of expectation maximization for mixtures of two Gaussians" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016), vol. 29, pp. 2676–2684.
11. C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, M. I. Jordan, "Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2016), vol. 29, pp. 4116–4124.
12. X. Shen, W. H. Wong, Convergence rate of sieve estimates. *Ann. Stat.* **22**, 580–615 (1994).
13. M. Cule, R. Samworth, M. Stewart, Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. B* **72**, 545–607 (2010).
14. B. Axelrod, I. Diakonikolas, A. Stewart, A. Sidiropoulos, G. Valiant, "A polynomial time algorithm for log-concave maximum likelihood via locally exponential families" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), vol. 32, pp. 7723–7735.
15. B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Routledge, 2018).
16. S. Karlin, Y. Rinott, Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10**, 467–498 (1980).
17. S. Lauritzen, C. Uhler, P. Zwiernik, Maximum likelihood estimation in Gaussian models under total positivity. *Ann. Stat.* **47**, 1835–1863 (2019).
18. S. Lauritzen, C. Uhler, P. Zwiernik, Total positivity in structured binary distributions. arXiv:1905.00516 (1 May 2019).
19. E. Robeva, B. Sturmfels, N. Tran, C. Uhler, Maximum likelihood estimation for totally positive log-concave densities. arXiv:1806.10120 (26 June 2018).
20. F. Guo, T. S. Richardson, On testing marginal versus conditional independence. arXiv:1906.01850 (5 June 2019).
21. N. Meinshausen, L. Meier, P. Bühlmann, P-values for high-dimensional regression. *J. Am. Stat. Assoc.* **104**, 1671–1681 (2009).
22. A. A. Amini, E. Levina, On semidefinite relaxations for the block model. *Ann. Stat.* **46**, 149–179 (2018).
23. J. Dahl, B. H. Fleury, L. Vandenberghe, "Approximate maximum-likelihood estimation using semidefinite programming" in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, 2003), **vol. 6**, pp. VI–721.
24. A. S. Bandeira, N. Boumal, A. Singer, Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Math. Program.* **163**, 145–167 (2017).
25. R. Royall, T. S. Tsou, Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. *J. R. Stat. Soc. B* **65**, 391–404 (2003).
26. P. Grünwald, "The safe Bayesian" in *International Conference on Algorithmic Learning Theory* (Springer, Berlin, Germany, 2012), pp. 169–183.
27. C. Holmes, S. Walker, Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104**, 497–503 (2017).
28. P. Grünwald, T. Van Ommen, Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12**, 1069–1103 (2017).
29. J. W. Miller, D. B. Dunson, Robust Bayesian inference via coarsening. *J. Am. Stat. Assoc.* **114**, 1113–1125 (2019).
30. B. Seo, B. G. Lindsay, A universally consistent modification of maximum likelihood. *Stat. Sin.* **23**, 467–487 (2013).
31. Q. J. Li, "Estimation of mixture models," PhD thesis, Yale University, New Haven, CT (1999).
32. P. D. Hoff, Nonparametric estimation of convex models via mixtures. *Ann. Stat.* **31**, 174–200 (2003).
33. H. Brunk, W. Franck, D. Hanson, R. Hogg, Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *J. Am. Stat. Assoc.* **61**, 1067–1080 (1966).
34. R. L. Dykstra, C. J. Feltz, Nonparametric maximum likelihood estimation of survival functions with a general stochastic ordering and its dual. *Biometrika* **76**, 331–341 (1989).
35. P. Grunwald, R. de Heide, W. Koolen, Safe testing. arXiv:1906.07801 (18 June 2019).
36. A. Wald, Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16**, 117–186 (1945).
37. D. Darling, H. Robbins, Confidence sequences for mean, variance, and median. *Proc. Natl. Acad. Sci. U.S.A.* **58**, 66–68 (1967).
38. J. Ville, *Étude Critique de la Notion de Collectif* (Gauthier-Villars, Paris, France, 1939).

39. S. R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon, Time-uniform Chernoff bounds via nonnegative supermartingales. *Probab. Surv.* **17**, 257–317 (2020).
40. S. R. Howard, A. Ramdas, J. McAuliffe, J. Sekhon, Uniform, nonparametric, non-asymptotic confidence sequences. arXiv:1810.08240 (18 October 2018).
41. S. R. Howard, A. Ramdas, Sequential estimation of quantiles with applications to A/B-testing and best-arm identification. arXiv:1906.09712 (24 June 2019).
42. R. Johari, P. Koomen, L. Pekelis, D. Walsh, *Peeking at A/B Tests: Why It Matters, and What to Do about It* (ACM Press, 2017), pp. 1517–1525.
43. G. Shafer, A. Shen, N. Vereshchagin, V. Vovk, Test martingales, Bayes factors and p-values. *Stat. Sci.* **26**, 84–101 (2011).
44. H. Robbins, Statistical methods related to the law of the iterated logarithm. *Ann. Math. Stat.* **41**, 1397–1409 (1970).
45. H. Robbins, D. Siegmund, "A class of stopping rules for testing parametric hypotheses" in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Univ. California, Berkeley, CA, 1970–1972), vol. 4, pp. 37–41. (1972).
46. T. L. Lai, On confidence sequences. *Ann. Stat.* **4**, 265–280 (1976).
47. T. L. Lai, Boundary crossing probabilities for sample sums and confidence sequences. *Ann. Probab.* **4**, 299–312 (1976).
48. A. Wald, *Sequential Analysis* (Courier Corporation, 1947).
49. H. Robbins, D. Siegmund, The expected sample size of some tests of power one. *Ann. Stat.* **2**, 415–436 (1974).
50. G. Lorden, M. Pollak, Nonanticipating estimation applied to sequential analysis and changepoint detection. *Ann. Stat.* **33**, 1422–1454 (2005).
51. A. Vexler, Martingale type statistics applied to change points detection. *Commun. Stat. Theor. Methods* **37**, 1207–1224 (2008).