Likelihood-free frequentist inference: bridging classical statistics and machine learning for reliable simulator-based inference[*]

Niccolò Dalmasso[†,1], Luca Masserano[†,2],
David Zhao[2], Rafael Izbicki[3], Ann B. Lee[2]

[1]*Department of Statistics and Data Science, Carnegie Mellon University,*
e-mail: niccolo.dalmasso@gmail.com

[2]*Department of Statistics and Data Science, Machine Learning Department, Carnegie Mellon University,*
e-mail: masserano.luca@gmail.com,
e-mail: dzhaoism@gmail.com,
e-mail: annlee@andrew.cmu.edu

[3]*Department of Statistics, Federal University of São Carlos,*
e-mail: rafaelizbicki@gmail.com

Abstract: Many areas of science rely on simulators that implicitly encode intractable likelihood functions of complex systems. Classical statistical methods are poorly suited for these so-called likelihood-free inference (LFI) settings, especially outside asymptotic and low-dimensional regimes. At the same time, popular LFI methods — such as Approximate Bayesian Computation or more recent machine learning techniques — do not necessarily lead to valid scientific inference because they do not guarantee confidence sets with nominal coverage in general settings. In addition, LFI currently lacks practical diagnostic tools to check the actual coverage of computed confidence sets across the entire parameter space. In this work, we propose a modular inference framework that bridges classical statistics and modern machine learning to provide (i) a practical approach for constructing confidence sets with near finite-sample validity at any value of the unknown parameters, and (ii) interpretable diagnostics for estimating empirical coverage across the entire parameter space. We refer to this framework as *likelihood-free frequentist inference* (LF2I). Any method that defines a test statistic can leverage LF2I to create valid confidence sets and diagnostics without costly Monte Carlo or bootstrap samples at fixed parameter settings. We study two likelihood-based test statistics (`ACORE` and `BFF`) and demonstrate their performance on high-dimensional complex data. Code is available at https://github.com/lee-group-cmu/lf2i.

MSC2020 subject classifications: Primary 62F25; secondary 62G08, 62P35.
Keywords and phrases: Likelihood-free inference, simulator-based infer-

1. Introduction

Hypothesis testing and uncertainty quantification are the hallmarks of scientific inference. Methods that achieve good statistical performance (e.g., high power) often rely on being able to explicitly evaluate a likelihood function, which relates parameters of the data-generating process to observed data. However, in many areas of science and engineering, complex phenomena are modeled by forward simulators that *implicitly* define a likelihood function. For example,[1] given input parameters θ from some parameter space Θ, a stochastic model F_θ may encode the interaction of atoms or elementary particles, or the transport of radiation through the atmosphere or through matter in the Universe by combining deterministic dynamics with random fluctuations and measurement errors, to produce synthetic data $\mathbf{X}$.

Simulation-based inference with an intractable likelihood is commonly referred to as *likelihood-free inference* (LFI). The most well-known approach to LFI is Approximate Bayesian Computation (ABC; see [6, 63, 81] for a review). These methods use simulations sufficiently close to the observed data $D = \left\{\mathbf{x}_1^{\mathrm{obs}}, \ldots, \mathbf{x}_n^{\mathrm{obs}}\right\}$ to infer the underlying parameters, or more precisely, the posterior distribution $p(\theta|D)$. Recently, the arsenal of LFI methods has been expanded with new machine learning algorithms (such as neural density estimators) that instead use the output from simulators as training data. The objective here is to learn a "surrogate model" or *approximation* of the likelihood $p(D|\theta)$ or posterior $p(\theta|D)$. The surrogate model, rather than the simulations themselves, is then used for inference. Machine-learning (ML) based methods have revolutionized LFI in terms of the complexity and dimensionality of the problems that can be tackled (see [23] for a recent review). Nevertheless, neither ABC nor ML-based LFI approaches guarantee confidence sets with frequentist coverage, which are crucial to ensure reliability of downstream scientific conclusions. Suppose that we have a high-fidelity simulator F_θ, which implicitly encodes the likelihood, and that we observe data $\mathcal{D}$ of finite sample size n. We address two open challenges in LFI:

i) The first challenge is finding practical procedures for constructing a $(1 - \alpha)$ confidence set $R(\mathcal{D})$ with nominal coverage[2]

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in R(\mathcal{D})\right) = 1 - \alpha, \tag{1}$$

---

[1]**Notation.** Let F_θ represent the stochastic forward model for a sample point $\mathbf{X} \in \mathcal{X}$ at parameter $\theta \in \Theta$. We refer to F_θ as a "simulator", as the assumption is that we can sample data from the model. We denote i.i.d "observable" data from F_θ by $\mathcal{D} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, and the actually observed or measured data by $D = \left\{\mathbf{x}_1^{\mathrm{obs}}, \ldots, \mathbf{x}_n^{\mathrm{obs}}\right\}$. The likelihood function is defined as $\mathcal{L}(D; \theta) = \prod_{i=1}^n p(\mathbf{x}_i^{\mathrm{obs}}|\theta)$, where $p(\cdot|\theta)$ is the density of F_θ with respect to a fixed dominating measure ν , which could be the Lebesgue measure.

[2]We use the notation $\mathbb{P}_{\mathcal{D}|\theta}(\cdot)$ to emphasize the fact that $\mathcal{D}$ is random, but θ is fixed.

where $\alpha \in (0, 1)$, *regardless of* the true value of the unknown parameter $\theta \in \Theta$ and of the number of observations n. Monte Carlo and bootstrap procedures are computationally infeasible for continuous parameter spaces Θ, and large-sample theory does not apply when, e.g., $n = 1$. The latter $n = 1$ scenario is very common in, e.g., large astronomical surveys where each object (e.g., galaxy or star) has a different parameter value θ and may only be measured once.

ii) The second challenge is finding practical and interpretable procedures to check that the empirical coverage of the constructed sets $R(\mathcal{D})$ is indeed close to (and no smaller than) $1 - \alpha$ for *any* $\theta \in \Theta$ (again, without resorting to costly Monte Carlo simulations at fixed parameter settings on a fine grid in parameter space Θ [18, Section 13]). Local validity across the entire parameter space is essential for reliable scientific inference because the scientist does not actually know what the true value of θ is for the object of interest.

Novelty and significance In this paper, we introduce a fully modular statistical framework that addresses both problems above. We refer to the general approach as *likelihood-free frequentist inference* (LF2I)[3]. LF2I is fully nonparametric and targets modern scientific applications, involving, e.g, high-dimensional data of different modalities, intractable likelihood models, and/or small sample sizes. Section 7.1 describes how LF2I is related to other work in this area.

At the heart of LF2I is the *Neyman construction of confidence sets*, albeit applied to a setting where the test statistic's distribution is unknown. Frequentist confidence sets and their equivalence to hypothesis tests have a long history in statistics [36, 69, 70]. While classical statistical procedures have significantly impacted fields like high-energy physics (see Section 7.1), most simulator-based methods lack theoretical guarantees for confidence sets beyond low-dimensional data and large-sample assumptions [35]. Implementing the Neyman construction for LFI is challenging not only because one cannot evaluate the likelihood, but also because one needs to test null hypotheses across the entire parameter space. While Monte Carlo and bootstrap methods estimate critical values and p-values from a batch of simulations at each null value θ_0 [62, 87], they become computationally infeasible for high-dimensional parameters. As a result, practical implementations might rely on parametric assumptions or asymptotic theory [71, 92]. For instance, it is often assumed that the likelihood-ratio (LR) statistic follows a χ^2 distribution, but this does not hold for irregular models or small sample sizes [4, 46, 53]. This work seeks to quickly and accurately estimate critical values and coverage across the parameter space without knowing the test statistic distribution or relying on large-sample approximations.

The key insight behind LF2I is that the main quantities of interest in frequentist statistical inference — test statistics, critical values, p-values and coverage of the confidence set — are *distribution functions indexed by the (unknown) parameter θ*, which generally vary smoothly over the parameter space Θ. As a result, one can leverage machine learning methods and data simulated in the neighborhood of a parameter to improve estimates of quantities of interest with

---

[3]Code is available as a Python package at https://github.com/lee-group-cmu/lf2i

fewer total simulations. Figure 1 illustrates the general LF2I inference machinery, which is composed of three modular branches with separate functionalities:

i) The test statistic branch (Figure 1 center and Section 3.2) uses a simulated set $\mathcal{T}$ to estimate a test statistic $\lambda(\mathcal{D}; \theta_0)$ for testing $H_{0,\theta_0} : \theta = \theta_0$ versus $H_{1,\theta_0} : \theta \neq \theta_0$. We study the theoretical and empirical performance of LF2I confidence sets derived from likelihood-based test statistics learned via the odds function $\mathbb{O}(\mathbf{X}; \theta)$ (Equation 7).

ii) The calibration branch (Figure 1 left and Section 3.3) uses a left-out set $\mathcal{T}'$ to estimate critical values C_{θ_0} for every level-α test of H_{0,θ_0} via quantile regression of the estimated test statistic $\lambda(\mathcal{D}; \theta_0)$ on $\theta_0 \in \Theta$. Once we have estimated the quantile function $\widehat{C}_{\theta_0}$ indexed by θ_0, we can directly construct Neyman confidence sets

$$\widehat{R}(\mathcal{D}) := \left\{ \theta \in \Theta \,\middle|\, \lambda(\mathcal{D}; \theta) \geq \widehat{C}_\theta \right\} \tag{2}$$

that have approximate $(1 - \alpha)$ finite-n coverage for every value of $\theta \in \Theta$. LF2I with critical values is amortized, meaning that once trained it can be evaluated on an arbitrary number of observations D. Alternatively, we can estimate p-values $p(D; \theta_0)$ for every test at $\theta = \theta_0$ with observed data D.

iii) The diagnostics branch (Figure 1 right and Section 3.4) uses a validation set $\mathcal{T}''$ to assess the empirical coverage $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{R}(\mathcal{D}))$ of the constructed confidence sets $\widehat{R}(\mathcal{D})$ across the parameter space by regressing the indicator variable $W := \mathbb{I}(\lambda(\mathcal{D}; \theta) \geq \widehat{C}_\theta)$ on θ. The diagnostics branch is not part of the inference procedure itself. Its purpose is to provide an independent assessment of local (instance-wise) coverage of the final constructed confidence sets.

The LF2I approach was first introduced in a conference proceeding [26]. This preliminary version — ACORE (Approximate Computation via Odds Ratio Estimation) — uses a test statistic that maximizes odds over the parameter space. In this follow-up paper, we analyze the statistical and computational properties of LF2I, while also introducing a new test statistic — the Bayesian Frequentist Factor (BFF) — which is the Bayes Factor [49, 50] treated as a frequentist test statistic. We show that the validity of LF2I only depends on calibration, whereas its power depends on the test statistic's definition and its estimation quality. In addition to new theoretical results in Section 4, we compare LF2I with approaches using Monte Carlo methods or Wilks' theorem (Section 6.1), and we illustrate how our diagnostics can help scientists in choosing the best tool to handle nuisance parameters (Section 6.2). Finally, we construct confidence sets given a high-dimensional particle physics simulation where ABC approaches are neither computationally feasible nor valid (Section 6.3).
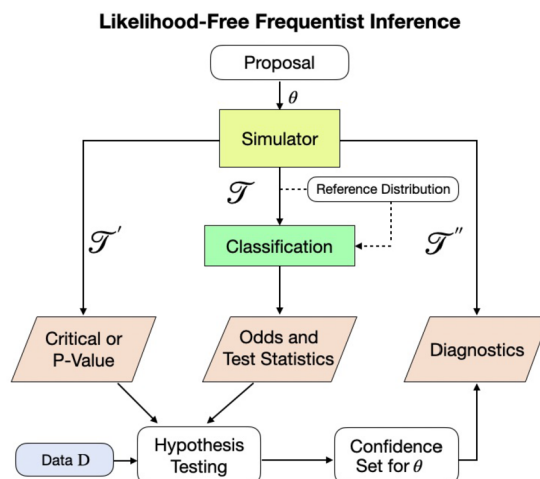
Likelihood-Free Frequentist Inference

FIG 1. ***The three-branch fully modular framework for likelihood-free frequentist inference (LF2I). Center branch:*** *Draw a sample $\mathcal{T}$ of size B from the simulator to estimate an arbitrary test statistic $\lambda(\mathcal{D}; \theta)$. Here we show how to do so by estimating the likelihood via the odds function $\mathbb{O}(\mathbf{X}; \theta)$.* ***Left branch:*** *Draw a second sample $\mathcal{T}'$ of size B' to estimate the critical values C_θ or p-values $p(\mathcal{D}; \theta)$ for all $\theta \in \Theta$.* ***Left + Center:*** *Once data D are observed, we can construct confidence sets $\widehat{R}(D)$ with finite-n validity according to Equation 12.* ***Right branch:*** *The LF2I diagnostics branch independently checks whether the coverage $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{R}(\mathcal{D}))$ of the confidence set is indeed correct across the entire parameter space.*

2. Statistical inference in a traditional setting

We now review the Neyman construction of confidence sets and the definitions of likelihood ratio and Bayes factor, before moving on to the details of the LF2I framework and its two instances, `ACORE` and `BFF`.

Equivalence of tests and confidence sets A classical approach to constructing a confidence set for an unknown parameter $\theta \in \Theta$ is to invert a series of hypothesis tests [70]. Suppose that for each possible value $\theta_0 \in \Theta$, there is a level-α test δ_{θ_0} of

$$H_{0,\theta_0} : \theta = \theta_0 \quad \text{versus} \quad H_{1,\theta_0} : \theta \neq \theta_0. \tag{3}$$

That is, a test δ_{θ_0} where the type I error (the probability of erroneously rejecting a true null hypothesis H_{0,θ_0}) is no larger than α. For observed data $\mathcal{D} = D$, let $R(D)$ be the set of all parameter values $\theta_0 \in \Theta$ for which the test δ_{θ_0} does not reject H_{0,θ_0}. Then, by construction, the random set $R(\mathcal{D})$ satisfies

$$\mathbb{P}_{\mathcal{D}|\theta} \left(\theta \in R(\mathcal{D}) \right) \geq 1 - \alpha \quad \forall \theta \in \Theta,$$

which makes it a $(1 - \alpha)$ *confidence set* for θ. Similarly, we can define tests with a desired significance level by inverting a confidence set with a certain coverage.

Likelihood ratio test   A general form of hypothesis tests that often leads to high power is the likelihood ratio test (LRT). Consider testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1, \tag{4}$$

where $\Theta_1 = \Theta \setminus \Theta_0$. For the *likelihood ratio (LR) statistic*,

$$\mathrm{LR}(\mathcal{D}; \Theta_0) = \log \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\mathcal{D}; \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\mathcal{D}; \theta)}, \tag{5}$$

the LRT of hypotheses (4) rejects H_0 when $\mathrm{LR}(D; \Theta_0) < C$ for some constant C. Figure 2 illustrates the construction of confidence sets for θ from level α likelihood ratio tests (3). The critical value for each such test δ_{θ_0} is $C_{\theta_0} = \sup \left\{ C : \mathbb{P}_{\mathcal{D}|\theta_0} \left(\mathrm{LR}(\mathcal{D}; \theta_0) < C \right) \leq \alpha \right\}$.

Bayes factor   Let π be a probability measure over the parameter space Θ. The Bayes factor [49, 50] for comparing the hypothesis $H_0 : \theta \in \Theta_0$ to its complement, the alternative H_1, is the ratio of the marginal likelihood of the two hypotheses:

$$\mathrm{BF}(\mathcal{D}; \Theta_0) \equiv \frac{\mathbb{P}(\mathcal{D}|H_0)}{\mathbb{P}(\mathcal{D}|H_1)} = \frac{\int_{\Theta_0} \mathcal{L}(\mathcal{D}; \theta) d\pi_0(\theta)}{\int_{\Theta_1} \mathcal{L}(\mathcal{D}; \theta) d\pi_1(\theta)}, \tag{6}$$

where π_0 and π_1 are the restrictions of π to the parameter regions Θ_0 and $\Theta_1 = \Theta_0^c$, respectively. The Bayes factor is often used as a Bayesian alternative to significance testing, as it quantifies the change in the odds in favor of H_0 when going from the prior to the posterior: $\frac{\mathbb{P}(H_0|\mathcal{D})}{\mathbb{P}(H_1|\mathcal{D})} = \mathrm{BF}(\mathcal{D}; \Theta_0) \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$.

3. Likelihood-free frequentist inference via odds estimation

In the typical LFI setting, we cannot directly evaluate the likelihood ratio $\mathrm{LR}(\mathcal{D}; \Theta_0)$ or even the likelihood $\mathcal{L}(\mathcal{D}; \theta)$. In this work, we describe a version of LF2I that is based on odds estimation. We assume that we have access to (i) a forward simulator F_θ to draw observable data, (ii) a reference distribution G that does not depend on θ, with larger support than F_θ for all $\theta \in \Theta$, and (iii) a probabilistic classifier to discriminates samples from F_θ and G.

3.1. Estimating an odds function across the parameter space

We start by generating a labeled sample $\mathcal{T} = \{(\theta_i, \mathbf{X}_i, Y_i)\}_{i=1}^B$ to compare data from F_θ with data from the reference distribution G. Here, $\theta \sim \pi_\Theta$ (a proposal distribution over Θ), the "label" $Y \sim \mathrm{Ber}(p)$, $\mathbf{X}|(\theta, Y = 1) \sim F_\theta$ and $\mathbf{X}|(\theta, Y = 0) \sim G$. We then define the odds at θ and fixed $\mathbf{x}$ as

$$\mathbb{O}(\mathbf{x}; \theta) := \frac{\mathbb{P}(Y = 1|\theta, \mathbf{x})}{\mathbb{P}(Y = 0|\theta, \mathbf{x})}. \tag{7}$$
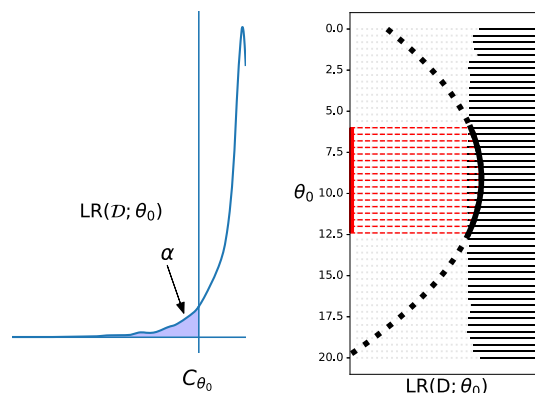
FIG 2. *Neyman construction of confidence sets by inverting hypothesis tests. **Left:** For each $\theta_0 \in \Theta$, we find the critical value C_{θ_0} that rejects the null hypothesis H_{0,θ_0} at level α; that is, C_{θ_0} is the α-quantile of the distribution of the test statistic under the null (a likelihood ratio $LR(\mathcal{D}; \theta_0)$ in this case). **Right:** The horizontal solid lines represent acceptance regions for each $\theta_0 \in \Theta$. Suppose we observe data D. The confidence set for θ (red vertical solid line) consists of all θ_0-values for which the observed test statistic $LR(D; \theta_0)$ (black curve) falls in the acceptance region.*

One way of interpreting $\mathbb{O}(\mathbf{x}; \theta)$ is to regard it as a measure of the chance that $\mathbf{x}$ was generated from F_θ rather than from G. That is, a large odds $\mathbb{O}(\mathbf{x}; \theta)$ reflects the fact that it is plausible that $\mathbf{x}$ was generated from F_θ (instead of G). We call G a "reference distribution" as we are comparing F_θ for different θ with this distribution. Equation 7 is equivalent to the likelihood $p(\mathbf{x}|\theta)$ up to a normalization constant, as shown in [26, Proposition 3.1]. The odds function $\mathbb{O}(\mathbf{X}; \theta)$ with $\theta \in \Theta$ as a parameter can be estimated with a probabilistic classifier, such as a neural network with a softmax layer, suitable for the data at hand. Algorithm 3 in Appendix A summarizes our procedure for simulating a labeled sample $\mathcal{T}$. For all experiments in this paper, we use p=1/2 and $G = F_{\mathbf{X}}$, where $F_{\mathbf{X}}$ is the (empirical) marginal distribution of F_θ with respect to π_Θ.

3.2. *Test statistics based on odds*

For testing $H_{0,\Theta_0} : \theta \in \Theta_0$ versus all alternatives $H_{1,\Theta_0} : \theta \notin \Theta_0$, we consider two test statistics: ACORE and BFF. Both statistics are based on $\mathbb{O}(\mathbf{X}; \theta)$, but whereas ACORE eliminates the parameter θ by maximization, BFF averages over the parameter space.

3.2.1. *ACORE by maximization*

The ACORE statistic [26] for testing Equation 3 is given by

$$\Lambda(\mathcal{D}; \Theta_0) := \log \frac{\sup_{\theta \in \Theta_0} \prod_{i=1}^{n} \mathbb{O}(\mathbf{X}_i; \theta)}{\sup_{\theta \in \Theta} \prod_{i=1}^{n} \mathbb{O}(\mathbf{X}_i; \theta)}$$

$$= \sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i; \theta_0, \theta_1)\right), \tag{8}$$

where the odds ratio

$$\mathbb{OR}(\mathbf{x}; \theta_0, \theta_1) := \frac{\mathbb{O}(\mathbf{x}; \theta_0)}{\mathbb{O}(\mathbf{x}; \theta_1)} \tag{9}$$

at $\theta_0, \theta_1 \in \Theta$ measures the plausibility that a fixed $\mathbf{x}$ was generated from θ_0 rather than θ_1. We use $\widehat{\Lambda}(\mathcal{D}; \Theta_0)$ to denote the ACORE statistic based on $\mathcal{T}$ and estimated odds $\widehat{\mathbb{O}}(\mathbf{X}; \theta_0)$. When $\widehat{\mathbb{O}}(\mathbf{X}; \theta_0)$ is well-estimated for every θ and $\mathbf{X}$, $\widehat{\Lambda}(\mathcal{D}; \Theta_0)$ is the same as the $\mathrm{LR}(\mathcal{D}; \Theta_0)$ in Equation 5 [26, Proposition 3.1].

3.2.2. *BFF by averaging*

Because the ACORE statistics in Equation 8 involves taking the supremum (or infimum) over Θ, it may not be practical in high dimensions. Hence, in this work, we propose an alternative statistic for testing (3) based on averaged odds:

$$\tau(\mathcal{D}; \Theta_0) := \frac{\int_{\Theta_0} \prod_{i=1}^{n} \mathbb{O}(\mathbf{X}_i; \theta_0) d\pi_0(\theta)}{\int_{\Theta_0^c} \prod_{i=1}^{n} \mathbb{O}(\mathbf{X}_i; \theta) d\pi_1(\theta)}, \tag{10}$$

where π_0 and π_1 are the restrictions of the proposal distribution π to the parameter regions Θ_0 and Θ_0^c, respectively. Let $\widehat{\tau}(\mathcal{D}; \Theta_0)$ denote estimates based on $\mathcal{T}$ and $\widehat{\mathbb{O}}(\theta_0; \mathbf{x})$. If the probabilities learned by the classifier are well estimated, then the estimated averaged odds statistic $\widehat{\tau}(\mathcal{D}; \Theta_0)$ is exactly the Bayes factor:

Proposition 1 (Fisher consistency)
Assume that, for every $\theta \in \Theta$, G dominates ν. If $\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{x}) = \mathbb{P}(Y = 1|\theta, \mathbf{x})$ for every θ and $\mathbf{x}$, then $\widehat{\tau}(\mathcal{D}; \Theta_0)$ is the Bayes factor $BF(\mathcal{D}; \Theta_0)$.

In this paper, we are using the Bayes factor as a frequentist test statistic. Hence, our term *Bayes Frequentist Factor (BFF)* statistic for τ and $\widehat{\tau}$.

3.3. *Fast construction of Neyman confidence sets*

Instead of a costly MC or bootstrap hypothesis test of $H_0 : \theta = \theta_0$ at each θ_0 on a fine grid (see, e.g., [62] and [87]), we draw only one sample $\mathcal{T}'$ of size B'. We then estimate either the critical value C_{θ_0} via quantile regression (Section 3.3.1), or the p-value $p(D; \theta_0)$ via probabilistic classification (Section 3.3.2), for all $\theta_0 \in \Theta$ simultaneously. In Supplementary Material H[4], we propose a practical strategy to choose the number of simulations B' and the learning algorithm.

---

[4] Available at https://lucamasserano.github.io/data/LF2I_supplementary_material.pdf

---

Algorithm 1 Estimate critical values C_{θ_0} for a level-α test of $H_{0,\theta_0} : \theta = \theta_0$ vs. $H_{1,\theta_0} : \theta \neq \theta_0$ for all $\theta_0 \in \Theta$ simultaneously.

Input: simulator F_θ; number of simulations B'; π_Θ (fixed proposal distribution over the parameter space); test statistic λ; quantile regression estimator; level $\alpha \in (0, 1)$
Output: estimated critical values $\widehat{C}_{\theta_0}$ for all $\theta_0 \in \Theta$

1: Set $\mathcal{T}' \leftarrow \emptyset$
2: **for** i in $\{1, \ldots, B'\}$ **do**
3:      Draw parameter $\theta_i \sim \pi_\Theta$
4:      Draw sample $\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n} \overset{iid}{\sim} F_{\theta_i}$
5:      Compute test statistic $\lambda_i \leftarrow \lambda((\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}); \theta_i)$
6:      $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(\theta_i, \lambda_i)\}$
7: **end for**
8: Use $\mathcal{T}'$ to learn the conditional quantile function $\widehat{C}_\theta := \widehat{F}_{\lambda|\theta}^{-1}(\alpha|\theta)$ via quantile regression of λ on θ
9: **return** $\widehat{C}_{\theta_0}$

3.3.1. The critical value via quantile regression

Algorithm 1 describes how to use quantile regression (e.g., [54, 68]) to estimate the critical value C_{θ_0} for a level-α test of (3) as a function of $\theta_0 \in \Theta$. To test a composite null hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, we use the cutoff $\widehat{C}_{\Theta_0} := \inf_{\theta \in \Theta_0} \widehat{C}_\theta$. Although we originally proposed the calibration procedure for ACORE, the same scheme leads to a valid test (control of type I error as the number of simulations $B' \to \infty$) for *any* test statistic λ (Theorem 8). Remarkably, this holds even if the test statistic is not well estimated. Note that in practice, we observe that the number of simulations B' needed to achieve correct coverage is usually much lower relative to B, the number of simulations needed to estimate the test statistic. In addition, Algorithm 1 does not rely on the observed data D and is therefore amortized, meaning that once the test statistic and critical values have been estimated, we can compute confidence sets for any new data set without the need to retrain the model.

3.3.2. The p-value via probabilistic classification

If the data D are observed beforehand, then given any test statistic λ we can alternatively compute p-values for each hypothesis $H_{0,\theta_0} : \theta = \theta_0$, that is,

$$p(D; \theta_0) := \mathbb{P}_{\mathcal{D}|\theta_0} \left(\lambda(\mathcal{D}; \theta_0) < \lambda(D; \theta_0) \right). \tag{11}$$

The p-value $p(D; \theta_0)$ can be used to test hypothesis and create confidence sets for any desired level α. As detailed in Algorithm 5, we can estimate it simultaneously for all $\theta \in \Theta$ by drawing a training sample $\mathcal{T}' = \{(Z_1, \theta_1), \ldots, (Z_{B'}, \theta_{B'})\}$ and using the random variable $Z := \mathbb{I}(\lambda(\mathcal{D}; \theta) < \lambda(D; \theta))$ as a label for each θ. To test the composite null hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$,

---

Algorithm 2 Estimate empirical coverage $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{R}(\mathcal{D}))$, for all $\theta \in \Theta$

Input: simulator F_θ; number of simulations B''; π_Θ (fixed proposal distribution over parameter space); test statistic λ; level α; critical values $\widehat{C}_\theta$; probabilistic classifier
Output: estimated coverage $\widehat{\mathbb{P}}_{\mathcal{D}|\theta}(\theta \in \widehat{R}(\mathcal{D}))$ for all $\theta \in \Theta$

1: Set $\mathcal{T}'' \leftarrow \emptyset$
2: **for** i in $\{1, \ldots, B''\}$ **do**
3:     Draw parameter $\theta_i \sim \pi_\Theta$
4:     Draw sample $\mathcal{D}_i := \{\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}\} \overset{iid}{\sim} F_{\theta_i}$
5:     Compute test statistic $\lambda_i \leftarrow \lambda(\mathcal{D}_i; \theta_i)$
6:     Compute indicator variable $W_i \leftarrow \mathbb{I}\left(\lambda_i \geq \widehat{C}_{\theta_i}\right)$
7:     $\mathcal{T}'' \leftarrow \mathcal{T}'' \cup \{(\theta_i, W_i)\}$
8: **end for**
9: Use $\mathcal{T}''$ to learn $\widehat{\mathbb{P}}_{\mathcal{D}|\theta'}(\theta' \in \widehat{R}(\mathcal{D}))$ across Θ by regressing W on θ
10: **return** $\widehat{\mathbb{P}}_{\mathcal{D}|\theta}(\theta \in \widehat{R}(\mathcal{D}))$

---

we use

$$\widehat{p}(D; \Theta_0) := \sup_{\theta \in \Theta_0} \widehat{p}(D; \theta).$$

Note that there is a key computational difference between estimating p-values versus estimating critical values. The p-value is a function of both θ and the observed sample D itself. As a result, Algorithm 5 has to be repeated for each observed D, making the computation of p-values non-amortized.

3.3.3. Amortized confidence sets

Finally, we construct an approximate confidence region for θ by taking the set

$$\widehat{R}(D) = \left\{\theta \in \Theta \,\middle|\, \lambda(D; \theta) \geq \widehat{C}_\theta\right\}, \tag{12}$$

or, alternatively,

$$\widehat{R}(D) = \{\theta \in \Theta \,|\, \widehat{p}(D; \theta) > \alpha\}. \tag{13}$$

See Algorithm 6 in Appendix C for details. As shown in [26, Theorem 3.3], the random set $\widehat{R}(\mathcal{D})$ has nominal $(1 - \alpha)$ coverage as $B' \to \infty$ regardless of the observed sample size n. As noted in Section 3.3.1, the confidence set in Equation 12 is fully *amortized*, meaning that once we have $\lambda(\mathcal{D}; \theta)$ and $\widehat{C}_\theta$ as a function of $\theta \in \Theta$, we can perform inference on new data without retraining.

3.4. Diagnostics: checking coverage across the parameter space

The LF2I framework has a separate module ("Diagnostics" in Figure 1) for evaluating "local" goodness-of-fit in different regions of the parameter space Θ. This estimates the coverage probability $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{R}(\mathcal{D}))$ of confidence sets

$\widehat{R}(\mathcal{D})$ across the parameter space via probabilistic classification. As detailed in Algorithm 2, we first generate a set of size B'' from the simulator: $\mathcal{T}'' = \{(\theta_1, \mathcal{D}_1), \ldots, (\theta_{B''}, \mathcal{D}_{B''})\}$. Then, for each sample $\mathcal{D}_i$, we check whether or not the test statistic λ_i is larger than the estimated critical value $\widehat{C}_{\theta_i}$ (the output from Algorithm 1). This is equivalent to computing a binary variable W_i for whether or not the "true" value θ_i falls within the confidence set $\widehat{R}(\mathcal{D}_i)$ (Equation 12). Recall that the computations of the test statistic and the critical value are amortized, meaning that we do not retrain algorithms to estimate these two quantities. The final step is to estimate empirical coverage as a function of θ by using W as a label for each θ. This estimation requires a new fit, but after training the probabilistic classifier, we can evaluate the estimated coverage anywhere in parameter space Θ.

This diagnostic procedure locates regions in parameter space where estimated confidence sets might under- or over-cover; see Figures 3, 4 and 6 for examples. Note that standard goodness-of-fit techniques for conditional densities [9, 16, 79, 84] only check for marginal coverage over Θ.

4. Theoretical guarantees

We now prove consistency of the critical value and p-value estimation methods (Algorithms 1 and 5, respectively) and provide theoretical guarantees for the power of BFF. We refer the reader to Appendix D for a proof for finite Θ that the power of ACORE converges to the power of LRT as B grows (Theorem 7).

In this section, $\mathbb{P}_{\mathcal{D}, \mathcal{T}'|\theta}$ denotes the probability integrated over both $\mathcal{D} \sim F_\theta$ and $\mathcal{T}'$, whereas $\mathbb{P}_{\mathcal{D}|\theta}$ denotes integration over $\mathcal{D} \sim F_\theta$ only. For notational ease, we do not explicitly state again (inside the parentheses of the same expression) that we condition on θ.

4.1. Critical value estimation

We start by showing that our procedure for choosing critical values leads to valid hypothesis tests (that is, tests that control the type I error probability), as long as the number of simulations B' in Algorithm 1 is sufficiently large. We assume that the null hypothesis is simple, that is, $\Theta_0 = \{\theta_0\}$ — which is the relevant setting for the Neyman construction of confidence sets in the absence of nuisance parameters. See Theorem 8 in Appendix F for results for composite null hypotheses.

We assume that the quantile regression estimator described in Section 3.3.1 is consistent in the following sense:

Assumption 1 (Uniform consistency) *Let $F(\cdot|\theta)$ be the cumulative distribution function of the test statistic $\lambda(\mathcal{D}; \theta_0)$ conditional on θ, where $\mathcal{D} \sim F_\theta$. Let $\widehat{F}_{B'}(\cdot|\theta)$ be the estimated distribution function indexed by θ, implied by a quantile regression with a sample $\mathcal{T}'$ of B' simulations $\mathcal{D} \sim F_\theta$. Assume that*

the quantile regression estimator is such that

$$\sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta_0) - F(\lambda|\theta_0)| \xrightarrow[B' \longrightarrow \infty]{P} 0.$$

Assumption 1 holds, for instance, for quantile regression forests [68].

Next, we show that Algorithm 1 yields a valid hypothesis test as $B' \to \infty$.

Theorem 1 *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on an absolutely continuous statistic $\lambda(\mathcal{D}; \theta_0)$ chosen according to Algorithm 1 for a fixed $\alpha \in (0, 1)$. If the quantile estimator satisfies Assumption 1, then, for every $\theta_0, \theta \in \Theta$,*

$$\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}(\lambda(\mathcal{D}; \theta_0) \leq C_{B'}) \xrightarrow[B' \longrightarrow \infty]{a.s.} \alpha,$$

where $\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}$ denotes the probability integrated over $\mathcal{D} \sim F_{\theta_0}$ and conditional on the random variable $C_{B'}$.

If the convergence rate of the quantile regression estimator is known (Assumption 2), Theorem 2 provides a finite-B' guarantee on how far the type I error of the test will be from the nominal level.

Assumption 2 (Convergence rate of the quantile regression estimator) *Using the notation of Assumption 1, assume that the quantile regression estimator is such that*

$$\sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta_0) - F(\lambda|\theta_0)| = O_P\left(\left(\frac{1}{B'}\right)^r\right)$$

for some $r > 0$.

Theorem 2 *With the notation and assumptions of Theorem 1, and if Assumption 2 also holds, then,*

$$|\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}(\lambda(\mathcal{D}; \theta_0) \leq C_{B'}) - \alpha| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

4.2. P-value estimation

Next we show that the p-value estimation method described in Section 3.3.2 is consistent. The results shown here apply to any test statistic λ. That is, these results are not restricted to BFF.

We assume consistency in the sup norm of the regression method used to estimate the p-values:

Assumption 3 (Uniform consistency) *The regression estimator used in Equation 11 is such that*

$$\sup_{\theta \in \Theta_0} |\widehat{\mathbb{E}}_{B'}[Z|\theta] - \mathbb{E}[Z|\theta]| \xrightarrow[B' \longrightarrow \infty]{a.s.} 0.$$

Examples of estimators that satisfy Assumption 3 include [7, 38, 41, 57].

The next theorem shows that the p-values obtained according to Algorithm 5 converge to the true p-values. Moreover, the power of the tests obtained using the estimated p-values converges to the power one would obtain if the true p-values could be computed.

Theorem 3 *Under Assumption 3 and if $p(\mathcal{D}; \Theta_0)$ is an absolutely continuous random variable then, for every $\theta \in \Theta$,*

$$\widehat{p}(D; \Theta_0) \xrightarrow[B' \longrightarrow \infty]{a.s.} p(D; \Theta_0)$$

and

$$\mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \Theta_0) \le \alpha) \xrightarrow[B' \longrightarrow \infty]{} \mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \Theta_0) \le \alpha).$$

The next corollary shows that as $B' \longrightarrow \infty$, the tests obtained using the p-values from Algorithm 5 have size α.

Corollary 1 *Under Assumption 3 and if F_θ is continuous for every $\theta \in \Theta$ and $p(\mathcal{D}; \Theta_0)$ is an absolutely continuous random variable, then*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \Theta_0) \le \alpha) \xrightarrow[B' \longrightarrow \infty]{} \alpha.$$

Under stronger assumptions about the regression method, it is also possible to derive rates of convergence for the estimated p-values.

Assumption 4 (Convergence rate of the regression estimator) *The regression estimator is such that*

$$\sup_{\theta \in \Theta_0} |\widehat{\mathbb{E}}[Z|\theta] - \mathbb{E}[Z|\theta]| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

for some $r > 0$.

Examples of regression estimators that satisfy Assumption 4 can be found in [29, 41, 83, 94].

Theorem 4 *Under Assumption 4,*

$$|p(D; \Theta_0) - \widehat{p}(D; \Theta_0)| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

4.3. Power of `BFF`

In this section, we provide convergence rates for `BFF` and show that its power relates to the integrated squared error

$$\mathcal{L}(\widehat{\mathbb{O}}, \mathbb{O}) := \int \left(\widehat{\mathbb{O}}(\mathbf{x}; \theta) - \mathbb{O}(\mathbf{x}; \theta)\right)^2 dG(\mathbf{x}) d\pi(\theta), \tag{14}$$

which measures how well we are able to estimate the odds function.

We assume that we are testing a simple hypothesis $H_{0,\theta_0} : \theta = \theta_0$, where θ_0 is fixed, and that $G(\mathbf{x})$ is the marginal distribution of $X \sim F_\theta(\mathbf{x})$ with respect to $\pi(\theta)$. We also assume that $\mathbf{x}$ contains all observations; that is, $\mathbf{X} = \mathcal{D}$. In this case, the denominator of the average odds is

$$
\begin{aligned}
\int_\Theta \mathbb{O}(\mathbf{x}, \theta) d\pi(\theta) &= \int_{\Theta_1} \frac{p \cdot p(\mathbf{x}|\theta)}{(1-p)g(\mathbf{x})} d\pi(\theta) \\
&= \frac{p}{1-p} \int_\Theta \frac{p(\mathbf{x}|\theta)}{\int_\Theta p(\mathbf{x}|\theta) d\pi(\theta)} d\pi(\theta) = \frac{p}{1-p},
\end{aligned}
\tag{15}
$$

where g is the density of G with respect to ν and therefore there is no need to estimate the denominator in Equation 10.

We also assume that the odds and estimated odds are both bounded away from zero and infinity:

Assumption 5 (Bounded odds and estimated odds) *There exists $0 < m$, $M < \infty$ such that for every $\theta \in \Theta$ and $\mathbf{x} \in \mathcal{X}$, $m \leq \mathbb{O}(\mathbf{x}; \theta), \widehat{\mathbb{O}}(\mathbf{x}; \theta) \leq M$.*

Finally, we assume that the CDF of the power function of the test based on the BFF statistic τ in Equation 10 is smooth in a Lipschitz sense:

Assumption 6 (Smooth power function) *For every $\theta_0 \in \Theta$, the cumulative distribution function of $\tau(\mathcal{D}; \theta_0)$, F_τ, is Lipschitz with constant C_L, i.e., for every $x_1, x_2 \in \mathbb{R}$, $|F_\tau(x_1) - F_\tau(x_2)| \leq C_L |x_1 - x_2|$.*

With these assumptions, we can relate the odds loss with the probability that the outcome of BFF is different from the outcome of the test based on the Bayes factor:

Theorem 5 *For fixed $c \in \mathbb{R}$, let $\phi_{\tau;\theta_0}(\mathcal{D}) = \mathbb{I}(\tau(\mathcal{D}; \theta_0) < c)$ and $\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = \mathbb{I}(\widehat{\tau}_B(\mathcal{D}; \theta_0) < c)$ be the testing procedures for testing $H_{0,\theta_0} : \theta = \theta_0$ based on τ and $\widehat{\tau}_B$, respectively. Under Assumptions 5-6, for every $0 < \epsilon < 1$ and $\theta \in \Theta$,*

$$
\int \mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})) d\pi(\theta_0) \leq \frac{2MC_L \cdot \sqrt{L(\widehat{\mathbb{O}}, \mathbb{O})}}{\epsilon} + \epsilon,
$$

where T denotes the realized training sample $\mathcal{T}$ and $\mathbb{P}_{\mathcal{D}|\theta,T}$ is the probability measure integrated over the observable data $\mathcal{D} \sim F_\theta$, but conditional on the train sample used to create the test statistic.

Theorem 5 demonstrates that, on average (over $\theta_0 \sim \pi$), the probability that hypothesis tests based on the BFF statistic versus the Bayes factor lead to different conclusions is bounded by the integrated odds loss. This result is valuable because the integrated odds loss is easy to estimate in practice, and hence provides us with a practically useful metric. For instance, the integrated odds loss can serve as a natural criterion for selecting the "best" statistical model out of a set of candidate models with different classifiers, for tuning model hyperparameters, and for evaluating model fit.

Next, we provide rates of convergence of the test based on `BFF` to the test based on the Bayes factor. We assume that the chosen probabilistic classifier has the following rate of convergence:

Assumption 7 (Convergence rate of the probabilistic classifier) *The probabilistic classifier trained with* $\mathcal{T}$, $\widehat{\mathbb{P}}(Y = 1|\mathbf{x}, \theta)$ *is such that*

$$\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{\mathbb{P}}(Y = 1|\mathbf{x}, \theta) - \mathbb{P}(Y = 1|\mathbf{x}, \theta)\right)^2 dH(\mathbf{x}, \theta)\right] = O\left(B^{-\kappa/(\kappa+d)}\right),$$

for some $\kappa > 0$ *and* $d > 0$, *where* $H(\mathbf{x}, \theta)$ *is a measure over* $\mathcal{X} \times \Theta$.

Typically, κ relates to the smoothness of $\mathbb{P}$, while d relates to the number of covariates of the classifier — in our case, the number of parameters plus the number of features. In Supplementary Material I, we provide some examples where Assumption 7 holds.

We also assume that the density of the product measure $G \times \pi$ is bounded away from infinity.

Assumption 8 (Bounded density) $H(\mathbf{x}, \theta)$ *dominates* $H' := G \times \pi$, *and the density of* H' *with respect to* H, *denoted by* h', *is such that there exists* $\gamma > 0$ *with* $h'(\mathbf{x}, \theta) < \gamma$, $\forall \mathbf{x} \in \mathcal{X}, \theta \in \Theta$.

If the probabilistic classifier has the convergence rate given by Assumption 7, then the average probability that hypothesis tests based on the `BFF` statistic versus the Bayes factor goes to zero has the rate given by the following theorem.

Theorem 6 *Let* $\phi_{\tau;\theta_0}(\mathcal{D})$ *and* $\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})$ *be as in Theorem 5. Under Assumptions 5-8, there exists* $K' > 0$ *such that, for any* $\theta \in \Theta$,

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}))d\pi(\theta_0) \leq K'B^{-\kappa/(4(\kappa+d))}.$$

Corollary 2 *Under Assumptions 5-8, there exists* $K' > 0$ *such that, for any* $\theta \in \Theta$,

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = 1)d\theta_0 \geq \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1)d\theta_0 - K'B^{-\kappa/(4(\kappa+d))}.$$

Corollary 2 tells us that the average power of the `BFF` test is close to the average power of the exact Bayes factor test. This result also implies that `BFF` converges to the most powerful test in the Neyman-Person setting, where the Bayes factor test is equivalent to the LRT.

5. Handling nuisance parameters

In most applications, we only have a small number of parameters that are of primary interest. The other parameters in the model are usually referred to

as nuisance parameters. In this setting, we decompose the parameter space as $\Theta = \Phi \times \Psi$, where Φ contains the parameters of interest, and Ψ contains nuisance parameters. Our goal is to construct a confidence set for $\phi \in \Phi$. To guarantee frequentist coverage by Neyman's inversion technique, however, one needs to test null hypotheses of the form $H_{0,\phi_0} : \phi = \phi_0$ by comparing the test statistics to the cutoffs $\widehat{C}_{\phi_0} := \inf_{\psi \in \Psi} \widehat{C}_{(\phi_0, \psi)}$ (Section 3.3.1). That is, one needs to control the type I error at each ϕ_0 for *all* possible values of the nuisance parameters. Computing such infimum can be numerically unwieldy, especially if the number of nuisance parameters is large [86, 96]. Below we propose approximate schemes for handling nuisance parameters:

In ACORE, we use a hybrid resampling or "likelihood profiling" method [14, 34, 80] to circumvent unwieldy numerical calculations as well as to reduce computational cost. For each ϕ (on a fine grid over Φ), we first compute the "profiled" value

$$\widehat{\psi}_\phi = \arg\max_{\psi \in \Psi} \prod_{i=1}^{n} \widehat{\mathbb{O}}\left(\mathbf{x}_i^{\mathrm{obs}}; (\phi, \psi)\right),$$

which (because of the odds estimation) is an approximation of the maximum likelihood estimate of ψ at the parameter value ϕ for observed data D. By definition, the estimated ACORE test statistic for the hypothesis $H_{0,\phi_0} : \phi = \phi_0$ is exactly given by $\widetilde{\Lambda}(\mathcal{D}; \phi_0) = \widehat{\Lambda}(\mathcal{D}; (\phi_0, \widehat{\psi}_{\phi_0}))$. However, rather than comparing this statistic to $\widehat{C}_{\phi_0}$, we use the hybrid cutoff

$$\widehat{C}'_{\phi_0} := \widehat{F}^{-1}_{\widehat{\Lambda}(\mathcal{D};\phi_0)\big|(\phi_0, \widehat{\psi}_{\phi_0})}\left(\alpha \,\Big|\, \phi_0, \widehat{\psi}_{\phi_0}\right), \tag{16}$$

where $\widehat{F}^{-1}$ is obtained via a quantile regression as in Algorithm 1, but using a training sample $\mathcal{T}'$ generated at *fixed* $\widehat{\psi}_{\phi_0}$ (that is, we run Algorithm 1 with the proposal distribution $\pi'((\phi, \psi)) \propto \pi(\phi) \times \delta_{\widehat{\psi}_\phi}(\psi)$, where $\delta_{\widehat{\psi}_\phi}(\psi)$ is a point mass distribution at $\widehat{\psi}_\phi$). Alternatively, one can compute the p-value

$$\widehat{p}(D; \phi_0) := \widehat{\mathbb{E}}\left[\mathbb{I}\left(\widehat{\Lambda}\left(\mathcal{D}; \phi_0\right) < \widehat{\Lambda}\left(D; \phi_0\right)\right)\,\Big|\, \phi_0, \widehat{\psi}_{\phi_0}\right] \tag{17}$$

via probabilistic classification as in Algorithm 5, but with $\mathcal{T}'$ simulated at fixed $\widehat{\psi}_{\phi_0}$ (that is, we run Algorithm 5 with the proposal distribution $\pi'((\phi, \psi)) \propto \pi(\phi) \times \delta_{\widehat{\psi}_\phi}(\psi)$. Hybrid methods do not always control α, but they are often a good approximation that lead to robust results [1, 76]. We refer to ACORE approaches based on Equation 16 or Equation 17 as "h-ACORE" approaches.

In contrast to ACORE, the BFF test statistic averages (rather than maximizes) over nuisance parameters. Hence, instead of adopting a hybrid resampling scheme to handle nuisance parameters, we approximate p-values and critical values, in what we refer to as "h-BFF", by using the marginal model of the data $\mathcal{D}$ at a parameter of interest ϕ:

$$\widetilde{\mathcal{L}}(D; \phi) = \int_{\psi \in \Psi} \mathcal{L}(D; \theta)\, d\pi(\psi).$$

We implement such a scheme by first drawing the train sample $\mathcal{T}'$ from the entire parameter space $\Theta = \Phi \times \Psi$, and then applying quantile regression (or probabilistic classification) using ϕ only.

Algorithm 7 details our construction of `ACORE` and `BFF` confidence sets when calibrating critical values under the presence of nuisance parameter (construction via p-value estimation is analogous). In Section 6.2, we demonstrate how our diagnostics branch can shed light on whether or not the final results have adequate frequentist coverage.

6. Experiments

We analyze the empirical performance of the LF2I framework under different problem settings: unknown null distribution of (known) test statistic (Section 6.1); nuisance parameters (Section 6.2); intractable likelihood and high-dimensional data (Section 6.3).

We use the cross-entropy loss (Eq. 24) when estimating the odds function in Equation 7 and the empirical coverage probability as in Section 3.4 via probabilistic classification. Moreover, we use the pinball loss [54] when estimating critical values as in Section 3.3.1 via quantile regression.

6.1. Gaussian mixture model: unknown null distribution

A common practice in LFI is to first estimate the likelihood and then assume that the LR statistic is approximately χ^2 distributed according to Wilks' theorem [31]. However, in settings with small sample sizes or irregular statistical models, such approaches may lead to confidence sets with incorrect coverage; it is often difficult to identify exactly when that happens, and then know how to recalibrate the confidence sets. (See [4] for a discussion of all conditions needed for Wilks' theorem to apply, which are often not realized in practice.)

The Gaussian mixture model (GMM) is a classical example where the LR statistic is known but its null distribution is unknown in finite samples. Indeed, the development of valid statistical methods for GMM is an active area of research [12, 25, 66, 78, 90]. Here we consider a one-dimensional Normal mixture with unknown mean but known unit variance:

$$X \sim 0.5N(\theta, 1) + 0.5N(-\theta, 1),$$

where the parameter of interest $\theta \in \Theta = [0, 5]$. In this example, the LRT statistic is not estimated but computed exactly. The goal is to analyze three different approaches for estimating the critical value C_{θ_0} of a level-α LRT of the hypothesis test $H_{0,\theta_0} : \theta = \theta_0$, for different $\theta_0 \in \Theta$, in a setting where we have removed potential effects of estimation errors in the test statistic:

- "LR with Monte Carlo samples", where we draw 1000 simulations at each point θ_0 on a fine grid over Θ and take C_{θ_0} to be the $1 - \alpha$ quantile of the
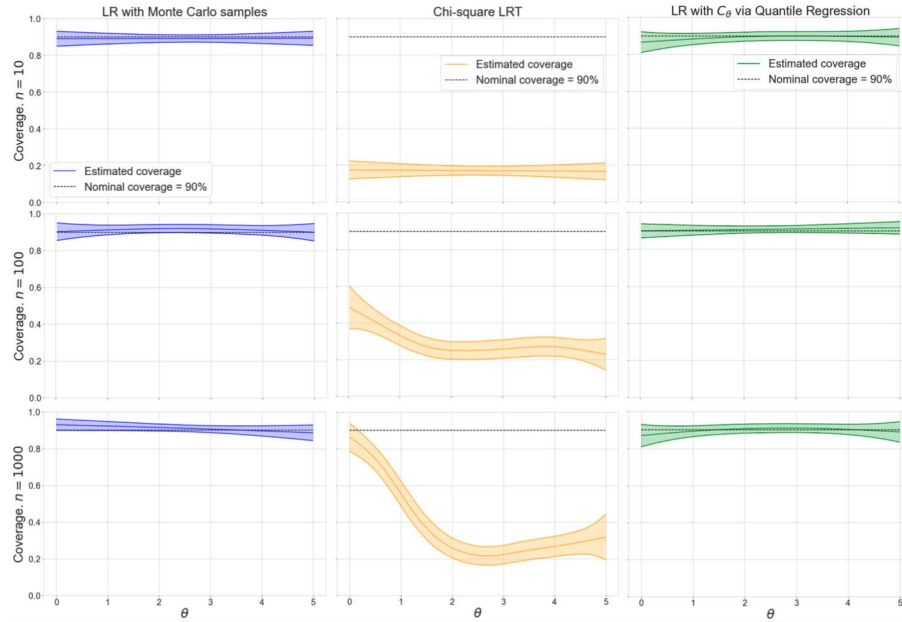
FIG 3. **GMM with unknown null distribution**. *Each panel shows the estimated coverage across the parameter space of 90% confidence sets for θ. Rows represent experiments with different observed sample sizes: $n = 10, 100, 1000$ (top, center, bottom). Columns represent three different approaches. **Left:** "LR with Monte Carlo samples" achieves nominal coverage everywhere but is computationally expensive, especially in higher dimensions. **Center:** "Chi-square LRT" clearly under-covers, i.e. confidence sets are not valid even for large n, other than at $\theta = 0$ where the mixture collapses to one Gaussian. **Right:** "LR with C_{θ_0} via quantile regression" returns finite-sample confidence sets with the nominal coverage of 90% for all values of θ, but using a total of 1000 simulations, instead of a MC sample of 1000 simulations at each grid point.*

distribution of the LR statistic, computed using the MC samples at each fixed θ_0. This approach is often just referred to as MC hypothesis testing.

- "Chi-square LRT", where we *assume* that $-2\mathrm{LR}(\mathcal{D}; \theta_0) \sim \chi_1^2$, and hence take $-2C_{\theta_0}$ to be the same as the upper α quantile of a χ_1^2 distribution.
- "LR with C_{θ_0} via quantile regression", where we estimate C_{θ_0} via quantile regression (Algorithm 1) based on a total of $B' = 1000$ simulations of size n sampled uniformly on Θ.

We then construct confidence sets by inverting the hypothesis tests, and finally assess their conditional coverage with the diagnostic branch of the LF2I framework (Algorithm 2 with $B'' = 1000$).

Figure 3 shows LF2I diagnostics for the three different approaches when the observed sample size (i.e., the number of observations from each unknown θ) is $n = 10, 100, 1000$. Confidence sets from "Chi-square LRT" are clearly not valid at any n, which shows that Wilks' theorem does not apply in this setting. The only exception arises when n is large enough and θ approaches 0, in which case
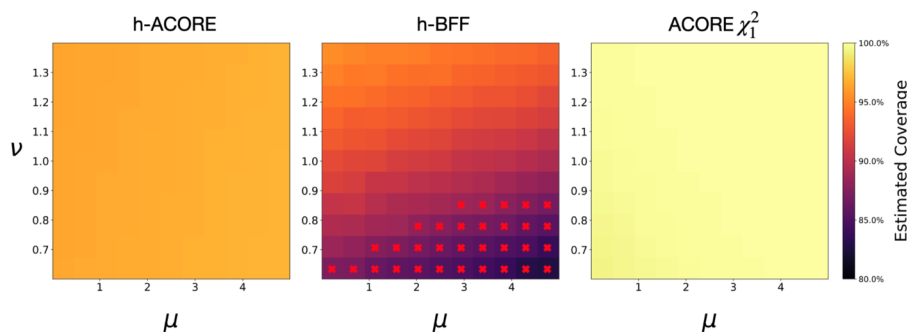
FIG 4. ***Poisson counting experiment with nuisance parameters.*** *The diagnostics branch provides guidance as to which LFI approach to use for the problem at hand by pinpointing regions of the parameter space Θ where inference is unreliable. The panels show empirical coverage as a function of both μ, the parameter of interest, and ν, the nuisance parameter. Nominal coverage is* 90%. **Left:** *h-ACORE, which uses profiled likelihoods, is overly conservative in terms of actual coverage ($\approx$ 96%) across Θ.* **Center:** *h-BFF, which marginalizes over ν, under-covers in several regions (red crosses).* **Right:** *ACORE χ_1^2, which uses cutoffs from the chi-square distribution, has almost no constraining power, yielding empirical coverage close to* 100% *everywhere.*

the mixture reduces to a *unimodal* Gaussian whose LR statistic has a known limiting distribution (see bottom center panel of Figure 3). On the other hand, "LR with C_{θ_0} via quantile regression" returns valid finite-sample confidence sets with conditional coverage equivalent to "LR with Monte Carlo samples". A key difference between the LF2I and MC methods is that the LF2I results are based on 1000 samples in total, whereas the MC results are based on 1000 MC samples at each θ_0 on a grid. The latter approach quickly becomes intractable in higher parameter dimensions and larger scales.

In Appendix E, we show that critical values are clearly non-constant across the parameter space, which also provides insight as to why assumptions of a pivotal test statistic (e.g., a χ^2-distributed test statistic asymptotically, or calibration based on a single point in the parameter space [89]) do not yield correct coverage. Supplementary Material J gives details on the specific quantile regressor (for Algorithm 1) and probabilistic classifier (for Algorithm 2) used in Figure 3, and presents extensions of the above experiments to confidence sets via p-value estimation and asymmetric mixtures.

6.2. *Poisson counting experiment: nuisance parameters and diagnostics*

Hybrid methods, which maximize or marginalize over nuisance parameters, do not always control the type I error of statistical tests. For small sample sizes, there is no theorem as to whether profiling or marginalization of nuisance parameters will give better frequentist coverage for the parameter of interest [18,

Section 12.5.1]. In addition, most practitioners consider a thorough check of frequentist coverage to be impractical [18, Section 13]. In this example, we apply the hybrid schemes from Section 5 to a high-energy physics (HEP) counting experiment [19, 20, 21, 42, 61] with nuisance parameters, which is a simplified version of a real particle physics experiment where the true likelihood function is not known. We illustrate how our diagnostics can guide the analyst and provide insight into which method to choose for the problem at hand.

Consider a "Poisson counting experiment" where particle collision events are counted under the presence of both an uncertain background process and a (new) signal process. The goal is to estimate the signal strength. To avoid identifiability issues, the background rate is estimated separately by counting the number of events in a control region where the signal is believed to be absent. Hence, the observable data $\mathbf{X} = (N_b, N_s)$ contain two measurements, where $N_b \sim \text{Pois}(\nu \tau b)$ is the number of events in the control region, and $N_s \sim \text{Pois}(\nu b + \mu s)$ is the number of events in the signal region. Our parameter of interest is the signal strength μ, whereas the scaling factor for the background ν is a nuisance parameter. The hyper-parameters s and b indicate the nominally expected counts from signal and backgrounds, and τ describes the relationship in measurement time between the two processes. We treat the three hyper-parameters as known with values $s = 15$, $b = 70$, $\tau = 1$, respectively. The hyper-parameters move the model away from the Gaussian limiting regime and make the relationship between data and parameters more complicated [42].

We compare the hybrid methods h-ACORE and h-BFF with ACORE χ_1^2 (which uses cutoffs from the chi-square distribution). We learn the odds using a QDA classifier with $B = 100,000$ and estimate critical values for the hybrid methods via quantile gradient boosted trees with $B' = 10,000$. We evaluate the different methods on a separate set of size $B'' = 1,000$ by estimating coverage and measuring the length of confidence sets for each of the simulated samples.

Figure 4 shows the estimated coverage as a function of both μ and ν. Confidence sets are considered to be valid when they achieve the nominal coverage level regardless of the true value of *both* the parameter of interest and the nuisance parameters. Both h-ACORE and ACORE χ_1^2 are overly conservative across the whole parameter space, while h-BFF under-covers in regions of high signal strength and low background. These results are consistent with the length of the corresponding confidence sets shown in Figure 5: h-ACORE and ACORE χ_1^2 are overly conservative, with the former being almost uninformative for the majority of evaluation samples. On the other side, while h-BFF seems to provide tighter parameter constraints, their length can be trusted only in regions where the method has coverage at least equal to the nominal level. Our LF2I diagnostic branch can pinpoint the regions of the parameter space where inference is reliable or not.
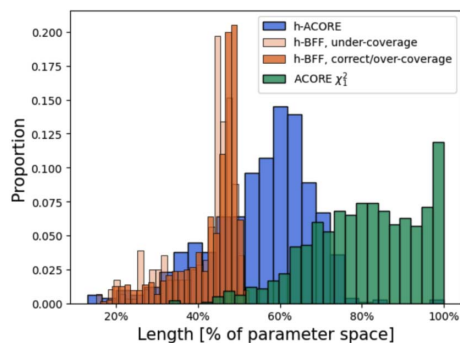
Fig 5. **Constraining power.** *Relative size of the confidence sets constructed in Section 6.2.* `ACORE` χ_1^2 *and* `h-ACORE` *yield the widest intervals (they are indeed overly conservative according to Figure 4).* `h-BFF` *provides tighter confidence sets, but their size cannot be trusted when the method under-covers. LF2I diagnostics can identify the parameter regions where the approach is not valid (red crosses in Figure 4). The dark-orange histogram reports* `h-BFF` *results after removing those points.*

6.3. *Muon energy estimation: intractable likelihood and high-dimensional data*

We now showcase LF2I on a high-energy physics application with intractable likelihood and very high-dimensional data. The goal is to estimate the energy of muons using a high-granularity calorimeter in a particle collider experiment. Muons are subatomic particles that have proven to be excellent probes of new physical phenomena: their detection and measurement has enabled several crucial discoveries in the last few decades, including the discovery of the Higgs boson [2, 5, 11, 15, 43]. Traditionally, the energy of a muon is determined from the curvature of its trajectory in a magnetic field, but curvature-based measurements have proven to be insufficiently precise at high energies. Recently, muon energy measurements based on their radiative losses in a dense, finely segmented calorimeter (Figure 6, left) have been shown to be a feasible alternative [30, 53].

In this application, the dimensionality of one data point $\mathbf{x}$ (a 3D image) is of the order of $\approx 50{,}000$ and the observed sample size is $n = 1$ (as each unique data point is the output of one experiment with a specific parameter of interest θ). In total, we have available 886,716 3D "image" inputs $\mathbf{x}$ with corresponding scalar muon energies θ. The data are obtained by accurately mimicking particle showers with GEANT4 [3], a high-fidelity simulator that has been calibrated for decades and is trusted to incorporate all the dynamics of the Standard Model of particle physics. The data are available at [52].

The scientific goal of this experiment is to quantify whether a high-granularity calorimeter would better constrain the energy of a muon (that is, lead to smaller confidence sets) than, for example, a detector that only measures the total en-
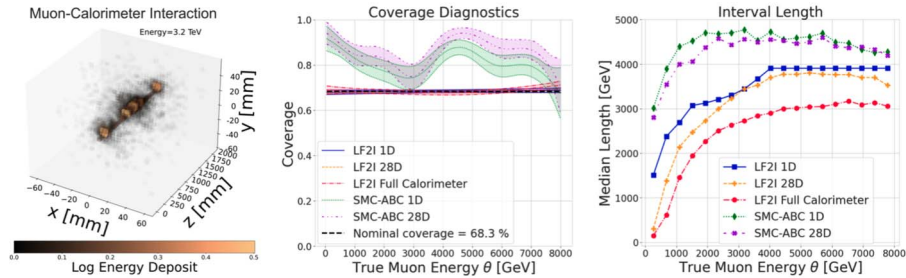
FIG 6. **Muon energy estimation.** *LF2I guarantees nominal coverage and yields smaller confidence intervals relative to SMC-ABC.* **Left:** *Data point example of a muon with incoming energy $\theta \approx 3.2$ TeV entering a calorimeter with $32 \times 32 \times 50$ cells.* **Center:** *LF2I (blue, orange, red in the right two panels) achieves coverage at the nominal level (68.3%), whereas SMC-ABC (green and purple) is consistently over-covering across the parameter space.* **Right:** *Median lengths of constructed intervals. While being extremely computationally intensive, SMC-ABC has also the least constraining power regardless of the data set used. SMC-ABC on the full calorimeter data is not reported as it was computationally infeasible to run.*

ergy of the incoming particle. To answer this question, we consider nested versions of the same energy measurement, where the inputs to our algorithms are of increasing dimensionality: (i) a 1D input which is equal to the sum over all the cells of the calorimeter (for each muon with deposited energy $E > 0.1$ GeV); (ii) 28 custom features extracted from the spatial and energy information of the calorimeter cells (see [53]); and (iii) the full calorimeter measurement, $\mathbf{x} \in \mathbb{R}^{51,200}$. We then construct LF2I confidence sets for each data point using BFF. On the full calorimeter data, we learn the odds function through a convolutional neural network classifier derived from the regressor proposed in [53], and estimate critical values via quantile gradient boosted trees. For the 1D and 28D data sets, we instead learn odds through a gradient boosting classifier. In both cases, we use approximately 83% of the data to learn the odds function ($B = 738,930$) and 14% to estimate critical values ($B' = 123,155$). For comparison, we also include results from SMC-ABC [82], a popular LFI algorithm from the Approximate Bayesian Computation literature. To provide a fair assessment of the results, SMC-ABC uses all the simulations that LF2I exploits separately (i.e., $B + B' = 862,085$). The remaining data points ($B'' = 24,631$) are used for validation and diagnostics of both methods.

Figure 6 (center) shows that LF2I with the BFF test statistic achieves the nominal level of coverage (68.3%) regardless of the data set used. This is consistent with Theorem 1: as long as the quantile regression is well estimated, LF2I confidence sets are guaranteed to be valid at the nominal $(1 - \alpha)$ level regardless of how well the test statistic is estimated. On the other hand, SMC-ABC is overly conservative with credible intervals that strongly over-cover across the whole parameter space. As to constraining power (interval length), Figure 6 (right) shows that SMC-ABC credible intervals are significantly wider than LF2I con-

fidence sets for both the 1D and 28D data sets (running SMC-ABC on the 51,200-dimensional full calorimeter data was computationally infeasible, and we were not able to report the results). Finally, note how the amount of information in the data directly influences the size of LF2I confidence sets: going from the 1D data set to the full calorimeter leads to noticeably smaller confidence intervals, and hence higher constraining power.

Remark on validity and computational cost   SMC-ABC does not have the right coverage, because the goal of ABC is to construct Bayesian credible regions and not valid confidence sets; see, e.g., [45] for other examples of SMC-ABC under- or over-covering. Furthermore, note that (i) LF2I is amortized: once training is done, confidence sets can be efficiently computed on an arbitrary number of observations without having to retrain the algorithms; and (ii) there is no need for a prior dimension reduction of the data (that is, we can directly input the three-dimensional image). Specifically, LF2I required approximately 10 and 5 CPU minutes on an AMD's EPYC 7763 machine to train the odds classifier and the quantile regressor respectively, and less than a second to obtain confidence intervals all at once for all observations (in this example, unique 24,631 "test" muons) regardless of their dimensionality. In contrast, SMC-ABC required approximately 1 CPU hour for *each* observation even for the lower-dimensional 1D and 28D data sets.

7. Conclusions and discussion

Validity   Our proposed LF2I methodology leads to frequentist confidence sets and hypothesis tests with finite-sample guarantees (when there are no nuisance parameters). *Any* existing or new test statistic – that is, not only estimates of the LR or BF statistics – can be plugged into our framework to create tests that control type I error. The implicit assumption is that the null distribution of the test statistic varies smoothly in parameter space. If that condition holds, then we can efficiently leverage quantile regression methods to construct valid confidence sets by a Neyman inversion of simple hypothesis tests, without having to rely on asymptotic results.

Nuisance parameters and diagnostics   For small sample sizes, no theorem guarantees whether profiling or marginalizing nuisance parameters will provide better frequentist coverage for the parameter of interest [18, Section 12.5.1]. It is generally believed that hybrid resampling methods return approximately valid confidence sets, but that a rigorous check of validity is infeasible when the true solution is not known. Our diagnostic branch presents practical tools for assessing empirical coverage across the entire parameter space (including nuisance parameters). After seeing the results, one can decide which method is most appropriate for the application at hand. For example, in the Poisson counting experiment of Section 6.2, LF2I diagnostics revealed that h-BFF (which averages the estimated odds over nuisance parameters) returned smaller confidence

intervals, but at the cost of under-covering in some regions of the parameter space.

Power Statistical power is the hardest property to achieve in practice in LFI. This is the area where we foresee that most statistical and computational advances will take place. As shown theoretically in Theorem 5 and empirically in Supplementary Material K, the power (or size) of LF2I confidence sets depends not only on the theoretical properties of the (exact) test statistics, but is also influenced by how precisely we are able to estimate it. In the case of `ACORE` and `BFF`, the latter can be divided in (i) how well we are able to estimate the likelihood or odds function (a statistical estimation error), and (ii) how accurate are the integration or maximization procedures we use (a purely numerical error); see Supplementary Material H for a more precise breakdown of the sources of error in LF2I confidence sets, particularly for `ACORE` and `BFF`. Machine learning offers exciting possibilities on both fronts. For example, with regards to (i), [10] offers compelling evidence that one can can dramatically improve estimates of the likelihood $p(\mathbf{x}|\theta)$ for $\theta \in \Theta$, or the likelihood ratio $p(\mathbf{x}|\theta_1, \theta_2)$ for $\theta_1, \theta_2 \in \Theta$, by a "mining gold" approach that extracts additional information from the simulator about the latent process. Future work could incorporate such an approach into the LF2I framework, with the calibration and diagnostic branches as separate modules.

Other test statistics Our work presents also another new direction for LF2I: So far frequentist LFI methods have been estimating either likelihoods or likelihood ratios, and then often relying on asymptotic properties of the LR statistic. We note that there are settings where it may be easier to either estimate the posterior $p(\theta|\mathbf{x})$ rather than the likelihood $p(\mathbf{x}|\theta)$, or alternatively to obtain point estimates for parameters directly via predictions algorithms. Because the LF2I framework is agnostic to which algorithms we use to construct the test statistic itself, we can potentially leverage methods that estimate the conditional mean $\mathbb{E}[\theta|\mathbf{x}]$ and variance $\mathbb{V}[\theta|\mathbf{x}]$ to construct frequentist confidence sets and hypothesis tests for θ with finite-sample guarantees. For example, [65] uses $T = \frac{(\mathbb{E}[\theta|\mathbf{x}] - \theta_0)^2}{\mathbb{V}[\theta|\mathbf{x}]}$, which in some scenarios corresponds to the Wald statistic for testing $H_{0,\theta_0} : \theta = \theta_0$ against $H_{1,\theta_0} : \theta \neq \theta_0$ [88], as an attractive alternative to get LF2I confidence sets from prediction algorithms and posterior estimators.

See Appendices A-F for proofs and details on the algorithms, and refer to the separate Supplementary Material file[5] for additional experiments and results referenced in the main text.

7.1. Related work

Classical statistical inference in high-energy physics (HEP) LF2I is inspired by pioneering work in HEP that adopted classical hypothesis tests and

---

[5]Available at https://lucamasserano.github.io/data/LF2I_supplementary_material.pdf

Neyman confidence sets for the discovery of new physics [1, 11, 21, 22, 35]. Our work grew from the discussion in HEP regarding theory and practice, and open problems such as how to efficiently construct Neyman confidence sets for general settings [21], how to assess coverage across the parameter space without costly Monte Carlo simulations [18], and how to choose hybrid techniques in practice [17]. This paper proposes a general approach to solve the above-mentioned open problems with a modular framework that can be adapted to fit the data at hand.

Universal inference Recently, [90] proposed a "universal" inference test statistic for constructing valid confidence sets and hypothesis tests with finite-sample guarantees without regularity conditions. The assumptions are that the likelihood $\mathcal{L}(\mathcal{D}; \theta)$ is known and that one can compute the maximum likelihood estimator (MLE). Our LF2I framework does *not* require a tractable likelihood, but it assumes that we have regression methods that can estimate the chosen test statistic and its critical values. In tractable likelihood settings where both universal inference and LF2I apply, the LF2I approach leads to more powerful tests than universal inference (see, e.g., Figure 11 in Supplementary Material).

Simulation-based calibration of Bayesian posterior distributions In Bayesian inference, the posterior distribution $\pi(\theta|\mathbf{x})$ is fundamental for quantifying uncertainty about the parameter θ given the data $\mathbf{x}$. Recent methods have been developed to assess the quality of estimated posterior distributions; that is, assessing whether an estimate $\widehat{\pi}(\theta|\mathbf{x})$ is consistent with the posterior distribution $\pi(\theta|\mathbf{x})$ implied by the assumed prior and likelihood [27, 28, 55, 58, 95]. *The calibration in LF2I is fundamentally different:* Even if posteriors are calibrated in the sense that $\widehat{\pi}(\theta|\mathbf{x}) = \pi(\theta|\mathbf{x})$ for every $\mathbf{x}$ and θ, confidence sets derived from it will not necessarily have the correct empirical coverage (according to Eq. 1). LF2I is agnostic to the choice of the test statistic (for instance, whether the test statistic is formed from likelihoods or posteriors [65]), and provides guarantees of how well we are able to constrain the true parameters of interest regardless of the choice of the prior or proposal distribution $\pi(\theta)$.

Likelihood-free inference via machine learning Recent LFI methods have been using simulators output as training data to learn surrogate models for inference; see [23] for a review. These techniques use synthetic data simulated across the parameter space to directly estimate key quantities, such as:

1. *posteriors* $p(\theta|\mathbf{x})$ [8, 13, 39, 48, 60, 64, 72, 77];
2. *likelihoods* $p(\mathbf{x}|\theta)$ [33, 40, 51, 59, 67, 74, 75, 91, 93]; or
3. *density ratios*, such as the likelihood-to-marginal ratio $p(\mathbf{x}|\theta)/p(\mathbf{x})$ [32, 44, 47, 85],[6] the likelihood ratio $p(\mathbf{x}|\theta_1)/p(\mathbf{x}|\theta_2)$ for $\theta_1, \theta_2 \in \Theta$ [10, 24] or the profile-likelihood ratio [42].[7]

---

[6]In 2014, Izbicki et al. approximate likelihoods for high-dimensional data (such as 2D images) via density ratios [47, Equation 5] and kernel methods, building on Izbicki's PhD thesis work on spectral series approaches to high-dimensional nonparametric inference. The kernel approximate likelihood approach was later superseded by neural SBI approaches.

[7]`ACORE` and `BFF` are based on estimating the odds $\mathbb{O}(\mathbf{X}; \theta)$ at $\theta \in \Theta$ (Equation 7); this is a

Recently, there have also been works that directly predict parameters θ of intractable models using neural networks [37, 56] (that is, they do not estimate posteriors, likelihoods or density ratios). In addition, new methods such as normalizing flows [73] and other neural density estimators are revolutionizing LFI in terms of sample efficiency and capacity, and will continue to do so.

Nonetheless, although the goal of LFI is inference on the unknown parameters θ, it remains an open question whether a given LFI algorithm produces reliable measures of uncertainty, as current methods lack guarantees of local (instance-wise) validity and power for a finite number of observations. They also have no practical diagnostics to assess local coverage across the parameter space. Our framework can be used in combination with any LFI approach that relies on a test statistic (such as the LRT) to provide both local coverage and diagnostics. Finally, thanks to the modular structure of LF2I, the diagnostic branch can be used separately to evaluate whether other approaches (like ABC and posterior methods that return credible regions) have good frequentist coverage, and in cases where they do not, LF2I can identify regions of the parameter space of over- or under-confidence.

Appendix A: Estimating odds

Algorithm 3 shows how to create the training set $\mathcal{T}$ for estimating odds. Out of the total number of simulations B, a proportion p is generated by the stochastic forward simulator F_θ at different parameter values θ, while the rest is sampled from a reference distribution G. Note that G can be any distribution that dominates F_θ. If G is the marginal distribution $F_{\mathbf{x}}$ and $n = 1$, then computations for `BFF` are simplified because its denominator equals one. Algorithm 4 shows how to sample from the marginal distribution $F_{\mathbf{x}}$. In practice, if the data is pre-simulated, one can sample from the (empirical) marginal using permutations to break the relationship between θ and $\mathbf{X}$ for $\mathbf{X} \sim G = F_{\mathbf{x}}$.

Appendix B: Estimating p-values

Given observed data D and a test statistic λ, we can compute p-values $p(D; \theta_0) := \mathbb{P}_{\mathcal{D}|\theta_0}(\lambda(\mathcal{D}; \theta_0) < \lambda(D; \theta_0))$ for each hypothesis $H_{0,\theta_0} : \theta = \theta_0$. Algorithm 5 describes how to estimate such p-values for all $\theta_0 \in \Theta$ simultaneously.

Appendix C: Constructing confidence sets

Algorithm 6 details the construction of LF2I confidence sets with `ACORE` and `BFF` as defined in Section 3 (the algorithm based on p-value estimation is analogous).

"likelihood-to-marginal ratio" approach, which estimates a one-parameter function as in the original paper by [47]. The likelihood ratio $\mathbb{OR}(\mathbf{X}; \theta_0, \theta_1)$ at $\theta_0, \theta_1 \in \Theta$ (Equation 9) is then computed from the odds function, without the need for an extra estimation step.

Algorithm 3 Generate a labeled sample of size B for estimating odds.

Input: simulator F_θ; reference distribution G; proposal distribution π_Θ over parameter space; number of simulations B; parameter p of Bernoulli distribution
Output: labeled training sample $\mathcal{T}$

1: Set $\mathcal{T} \leftarrow \emptyset$
2: **for** i in $\{1, ..., B\}$ **do**
3:      Draw parameter value $\theta_i \sim \pi_\Theta$
4:      Draw $Y_i \sim \text{Ber}(p)$
5:      **if** $Y_i == 1$ **then**
6:          Draw sample $\mathbf{X}_i \sim F_{\theta_i}$
7:      **else**
8:          Draw sample $\mathbf{X}_i \sim G$
9:      **end if**
10:     $\mathcal{T} \leftarrow \mathcal{T} \cup (\theta_i, \mathbf{X}_i, Y_i)$
11: **end for**
12: **return** $\mathcal{T} = \{\theta_i, \mathbf{X}_i, Y_i\}_{i=1}^B$

Algorithm 4 Sample from the marginal distribution $G = F_{\mathbf{X}}$.

Input: simulator F_θ; proposal distribution π_Θ over parameter space
Output: sample $\mathbf{X}_i$ from the marginal distribution $F_{\mathbf{X}}$

1: Draw parameter value $\theta_i \sim \pi_\Theta$
2: Draw sample $\mathbf{X}_i \sim F_{\theta_i}$
3: **return** $\mathbf{X}_i$

Algorithm 5 Estimate p-values $p(D; \theta_0)$ given observed data D for a level-α test of $H_{0,\theta_0} : \theta = \theta_0$ vs. $H_{1,\theta_0} : \theta \neq \theta_0$, for all $\theta_0 \in \Theta$ simultaneously

---

Input: observed data D; simulator F_θ; number of simulations B'; π_Θ (fixed proposal distribution over the parameter space Θ); test statistic λ; probabilistic classifier
Output: estimated p-value $\widehat{p}(D; \theta)$ for all $\theta = \theta_0 \in \Theta$

1: Set $\mathcal{T}' \leftarrow \emptyset$
2: **for** i in $\{1, \ldots, B'\}$ **do**
3:      Draw parameter $\theta_i \sim \pi_\Theta$
4:      Draw sample $\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n} \overset{iid}{\sim} F_{\theta_i}$
5:      Compute test statistic $\lambda_i \leftarrow \lambda((\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}); \theta_i)$
6:      Compute indicator $Z_i \leftarrow \mathbb{I}(\lambda_i < \lambda(D; \theta_i))$
7:      $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(\theta_i, Z_i)\}$
8: **end for**
9: Use $\mathcal{T}'$ to learn the p-value function $\widehat{p}(D; \theta)$ using Z as the label for each θ
10: **return** $\widehat{p}(D; \theta_0)$

Algorithm 7 details the construction of the (hybrid) `ACORE` and `BFF` confidence sets defined in Section 5 for the general setting with nuisance parameters. Note that the first chunk on estimating the odds and the last chunk with Neyman inversion are the same for `ACORE` and `BFF`. Furthermore, the test statistics are

---

Algorithm 6 Construct $(1-\alpha)$ confidence set for θ (no nuisance parameters).

Input: simulator F_θ; proposal distribution π over Θ; parameter p of Bernoulli; number of simulations B (test statistic); number of simulations B' (critical values); probabilistic classifier; observations $D = \{\mathbf{x}_i^{\mathrm{obs}}\}_{i=1}^n$; level $\alpha \in (0,1)$; size of evaluation grid over parameter space n_{grid}; test statistic λ (`ACORE` or `BFF`)

Output: θ evaluation points in confidence set $\widehat{R}(D)$

1: **// Estimate odds**
2: Generate labeled sample $\mathcal{T}$ according to Algorithm 3
3: Apply probabilistic classifier to $\mathcal{T}$ to learn $\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{X})$, for all $\theta \in \Theta$ and $\mathbf{X} \in \mathcal{X}$
4: Let the estimated odds $\widehat{\mathbb{O}}(\mathbf{X};\theta) \leftarrow \frac{\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{X})}{\widehat{\mathbb{P}}(Y=0|\theta,\mathbf{X})}$
5:
6: **// Compute cut-offs for `ACORE` or `BFF`**
7: **if** $\lambda == $ `ACORE` **then**
8:     Let $\lambda(\mathcal{D};\theta) \leftarrow \widehat{\Lambda}(\mathcal{D};\theta)$ be the `ACORE` statistic (Equation 8) with estimated odds
9: **else if** test_stat $==$ BFF **then**
10:     Let $\lambda(\mathcal{D};\theta) \leftarrow \widehat{\tau}(\mathcal{D};\theta)$ be the `BFF` statistic (Equation 10) with estimated odds
11: **end if**
12: Learn critical values $\widehat{C}_\theta$ according to Algorithm 1
13:
14: **// Confidence sets for θ via Neyman inversion**
15: Initialize confidence set $\widehat{R}(D) \leftarrow \emptyset$
16: Let L_Θ be a lattice over Θ with n_{grid} elements
17: **for** $\theta_0 \in \mathrm{L}_\Theta$ **do**
18:     **if** $\lambda(D;\theta_0) \geq \widehat{C}_{\theta_0}$ **then**
19:         $\widehat{R}(D) \leftarrow \widehat{R}(D) \cup \{\theta_0\}$
20:     **end if**
21: **end for**
22: **return** confidence set $\widehat{R}(D)$

---

the same whether or not there are nuisance parameters.

Appendix D: Theoretical guarantees of power for `ACORE` with calibrated critical values

Next, we show, for finite Θ, that as long as the probabilistic classifier is consistent and the critical values are well estimated (which holds for large B' according to Theorem 8), the power of the `ACORE` test converges to the power of the LRT as B grows.

Theorem 7 *For each $C \in \mathbb{R}$, let $\widehat{\phi}_{B,C}(\mathcal{D})$ be the test based on the `ACORE` statistic $\widehat{\Lambda}_B$ with critical value C[8] for number of simulations B in Algorithm 3. Moreover, let $\phi_C(\mathcal{D})$ be the likelihood ratio test with critical value C. If, for every $\theta \in \Theta$, the probabilistic classifier is such that*

$$\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{X}) \xrightarrow[B\longrightarrow\infty]{P} \mathbb{P}(Y=1|\theta,\mathbf{X}),$$

---

[8]That is, $\widehat{\phi}_{B,C}(\mathcal{D}) = 1 \iff \widehat{\Lambda}_B(\mathcal{D};\Theta_0) < C$.

Algorithm 7 Construct confidence set for ϕ with (approximate) coverage $1 - \alpha$ under the presence of nuisance parameters.

Input: simulator F_θ; proposal distribution π over $\Theta = \Phi \times \Psi$; parameter p of Bernoulli; number of simulations B (test statistic); number of simulations B' (critical values); probabilistic classifier; observations $D = \{\mathbf{x}_i^{\mathrm{obs}}\}_{i=1}^n$; level $\alpha \in (0, 1)$; size of evaluation grid over parameter space, n_{grid}; test statistic λ (`ACORE` or `BFF`)
Output: ϕ evaluation points in confidence set $\widehat{R}(D)$

1: // **Estimate odds**
2: Generate labeled sample $\mathcal{T}$ according to Algorithm 3
3: Apply probabilistic classifier to $\mathcal{T}$ to learn $\widehat{\mathbb{P}}(Y = 1 | \theta, \mathbf{X}), \forall \theta = (\phi, \psi) \in \Theta, \mathbf{X} \in \mathcal{X}$
4: Let the estimated odds $\widehat{\mathbb{O}}(\mathbf{X}; \theta) \leftarrow \frac{\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{X})}{\widehat{\mathbb{P}}(Y=0|\theta,\mathbf{X})}$
5:
6: // **Compute (hybrid) critical values for `h-ACORE` or `h-BFF`**
7: **if** $\lambda ==$ `ACORE` **then**
8:      Let $\widehat{\psi}_\phi \leftarrow \arg\max_{\psi \in \Psi} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{x}_i^{\mathrm{obs}}; (\phi, \psi))$ for every ϕ
9:      Let $\lambda(\mathcal{D}; \phi) \leftarrow \widehat{\Lambda}\left(\mathcal{D}; (\phi, \widehat{\psi}_\phi)\right)$ be `ACORE` (Equation 8) with estimated odds
10:      Generate $\mathcal{T}'$ as in Algorithm 1 using the proposal $\pi'((\phi, \psi)) \propto \pi(\phi) \times \delta_{\widehat{\psi}_\phi}(\psi)$
11:      Learn $\widehat{C}_\phi = \widehat{F}^{-1}_{\lambda(\mathcal{D};\phi)\big|\left(\phi, \widehat{\psi}_\phi\right)}(\alpha)$ for every ϕ as in Algorithm 1 using $\mathcal{T}'$
12: **else if** $\lambda ==$ `BFF` **then**
13:      Let $\pi_\Psi(\psi)$ be the restriction of proposal distribution π over Ψ
14:      Let $\lambda(\mathcal{D}; \phi) \leftarrow \widehat{\tau}(\mathcal{D}; \phi)$ be the `BFF` statistic (Equation 10) with estimated odds
15:      Learn $\widehat{C}_\phi = \widehat{F}^{-1}_{\lambda(\mathcal{D};\phi)\big|(\phi)}(\alpha)$ for every ϕ (no ψ) as in Algorithm 1
16: **end if**
17:
18: // **Confidence sets for ϕ via Neyman inversion**
19: Initialize confidence set $\widehat{R}(D) \leftarrow \emptyset$
20: Let L_Φ be a lattice over Φ with n_{grid} elements
21: **for** $\phi_0 \in \mathrm{L}_\Phi$ **do**
22:      **if** $\lambda(D; \phi_0) \geq \widehat{C}_{\phi_0}$ **then**
23:          $\widehat{R}(D) \leftarrow \widehat{R}(D) \cup \{\phi_0\}$
24:      **end if**
25: **end for**
26: **return** confidence set $\widehat{R}(D)$

---

where $|\Theta| < \infty$, and $\widehat{C}_B$ is chosen such that $\widehat{C}_B \xrightarrow[B \longrightarrow \infty]{D} C$ for a given $C \in \mathbb{R}$, then, for every $\theta \in \Theta$,

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\phi}_{B,\widehat{C}_B}(\mathcal{D}) = 1\right) \xrightarrow[B \longrightarrow \infty]{} \mathbb{P}_{\mathcal{D}|\theta}\left(\phi_C(\mathcal{D}) = 1\right).$$

Proof Because $\widehat{\mathbb{P}}(Y = 1 | \theta, \mathbf{X}) \xrightarrow[B \longrightarrow \infty]{P} \mathbb{P}(Y = 1 | \theta, \mathbf{X})$, it follows directly from the properties of convergence in probability that for every $\theta_0, \theta_1 \in \Theta$

$$\sum_{i=1}^n \log\left(\widehat{\mathbb{OR}}(\mathbf{X}_i^{\mathrm{obs}}; \theta_0, \theta_1)\right) \xrightarrow[B \longrightarrow \infty]{P} \sum_{i=1}^n \log\left(\mathbb{OR}(\mathbf{X}_i^{\mathrm{obs}}; \theta_0, \theta_1)\right).$$

The continuous mapping theorem implies that

$$\widehat{\Lambda}_B(\mathcal{D};\Theta_0) \xrightarrow[B\longrightarrow\infty]{P} \sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\mathrm{obs}};\theta_0,\theta_1)\right),$$

and therefore $\widehat{\Lambda}_B(\mathcal{D};\Theta_0)$ converges in distribution to $\sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\mathrm{obs}};\theta_0,\theta_1)\right)$. Now, from Slutsky's theorem,

$$\widehat{\Lambda}_B(\mathcal{D};\Theta_0) - \widehat{C}_B \xrightarrow[B\longrightarrow\infty]{D} \sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\mathrm{obs}};\theta_0,\theta_1)\right) - C.$$

It follows that

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\phi}_{B,\widehat{C}_B}(\mathcal{D}) = 1\right) = \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\Lambda}_B(\mathcal{D};\Theta_0) - \widehat{C}_B \leq 0\right)$$

$$\xrightarrow[B\longrightarrow\infty]{} \mathbb{P}_{\mathcal{D}|\theta}\left(\sup_{\theta_0\in\Theta_0} \inf_{\theta_1\in\Theta} \sum_{i=1}^{n} \log\left(\mathbb{OR}(\mathbf{X}_i^{\mathrm{obs}};\theta_0,\theta_1)\right) - C \leq 0\right)$$

$$= \mathbb{P}_{\mathcal{D}|\theta}\left(\phi_C(\mathcal{D}) = 1\right),$$

where the last equality follows from Proposition 1. ∎

Appendix E: Analysis of critical values for experiments 6.1 and 6.2

In this section we visualize how critical values vary across the parameter space Θ for the experiments of Sections 6.1 and 6.2. Figure 7 compares critical values for the exact LRT of the Gaussian Mixture Model (GMM) example, where the distribution of the test statistic is unknown, using three different methods:

i) The first approach is to compute cutoffs via Monte Carlo (MC) simulations at fixed values of θ. These critical values can be considered the "ground truth", since for this one-dimensional example we were able to use a high-resolution grid and large batches at each grid point. Unfortunately, MC quickly becomes infeasible if the dimensionality of the parameter space increases. In addition, a scientist cannot adopt MC samples in practical settings, where one only has access to a pre-determined data set and not to the simulator itself.

ii) The second approach is to assume that the cutoff is (asymptotically) constant across the parameter space. Here we have computed cutoffs assuming that Wilks' theorem holds and that the limiting distribution is a χ^2-distribution, which is not the case. Indeed, the bottom central panel of Figure 3 shows that the χ^2-approximation achieves correct coverage only when $\theta = 0$ (i.e., when the GMM collapses to one Gaussian).

iii) The third approach is to compute the critical values of the (known) test
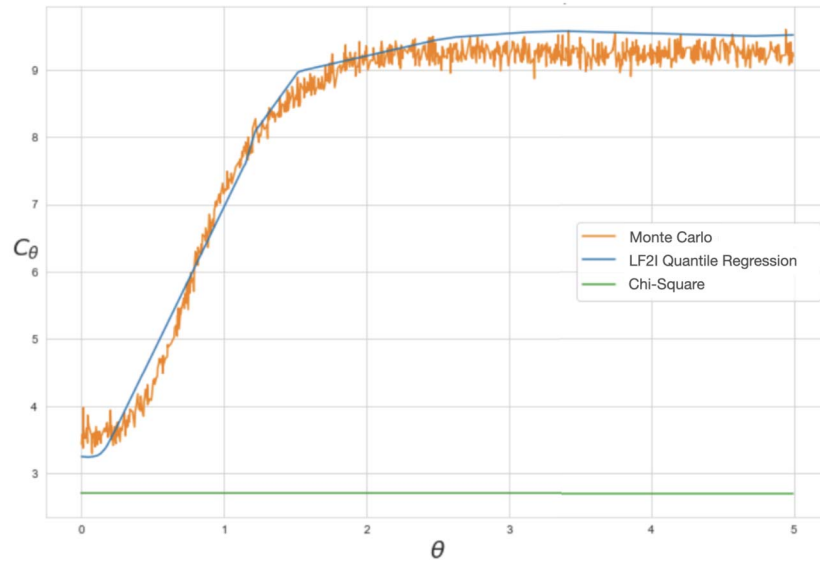
FIG 7. **Comparison of critical values** *obtained via* Monte Carlo, *the* Chi-Square *asymptotic assumption of Wilks' Theorem, and* LF2I Quantile Regression, *for the GMM example of Section* 6.1.

statistic via quantile regression (QR). With a very small calibration set (0.1% of the total simulations used for the MC approach), QR is able to approximate the quantile surface and achieve nominal coverage for all values of θ (see Figure 3).

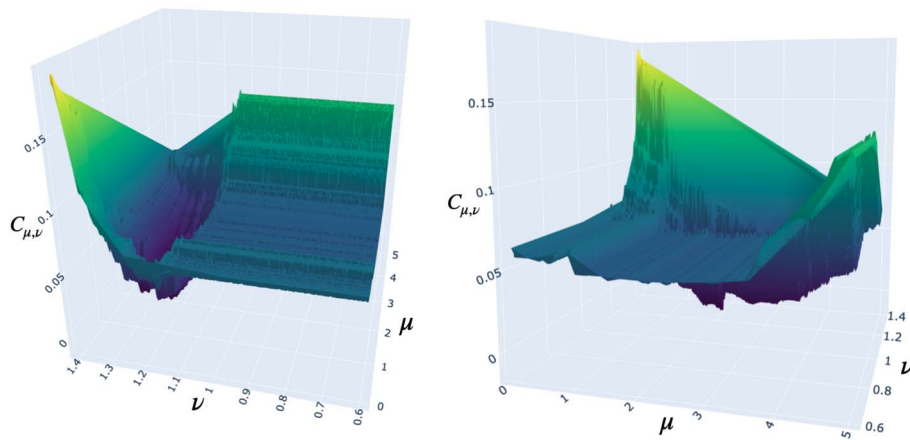FIG 8. **Critical values of** `h-ACORE` **estimated via quantile regression** *as a function of the parameter of interest μ and the nuisance parameter ν, for the example of Section* 6.2. *The figures show the same 2D surface from two different angles.*

Figure 8 shows similar results for the HEP example of Section 6.2; here we visualize the the critical values of `h-ACORE` (estimated via LF2I) as a function of the parameter of interest μ and the nuisance parameter ν. Again, we see evidence that the quantile surface is far from being constant, and that the test statistic is not pivotal. Hence, there is a need for a quantile regression that adapts to the varying distribution of the test statistic.

Appendix F: Additional proofs

Proof [Proof of Proposition 1] Because ν dominates F_θ, G also dominates F_θ. Let $f(\mathbf{x}|\theta)$ be the density of F_θ with respect to G. By Bayes rule,

$$\mathbb{O}(\mathbf{x};\theta) := \frac{\mathbb{P}(Y=1|\theta,\mathbf{x})}{\mathbb{P}(Y=0|\theta,\mathbf{x})} = \frac{f(\mathbf{x}|\theta)p}{(1-p)}.$$

If $\widehat{\mathbb{P}}(Y=1|\theta,\mathbf{x}) = \mathbb{P}(Y=1|\theta,\mathbf{x})$, then $\widehat{\mathbb{O}}(\mathbf{x};\theta_0) = \mathbb{O}(\mathbf{x};\theta_0)$. Therefore,

$$
\begin{aligned}
\widehat{\tau}(\mathcal{D};\Theta_0) :=& \frac{\int_{\Theta_0} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\mathrm{obs}};\theta)d\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\mathrm{obs}};\theta)d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n \mathbb{O}(\mathbf{X}_i^{\mathrm{obs}};\theta)d\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n \mathbb{O}(\mathbf{X}_i^{\mathrm{obs}};\theta)d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n \frac{f(\mathbf{X}_i^{\mathrm{obs}}|\theta)p}{(1-p)}d\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n \frac{f(\mathbf{X}_i^{\mathrm{obs}}|\theta)p}{(1-p)}d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n f(\mathbf{X}_i^{\mathrm{obs}}|\theta)d\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n f(\mathbf{X}_i^{\mathrm{obs}}|\theta)d\pi_1(\theta)}
\end{aligned}
$$

Moreover, the chain rule implies that $f(\mathbf{x}|\theta) = p(\mathbf{x}|\theta)h(\mathbf{x})$, where $h(\mathbf{x}) := \frac{d\nu}{dG}(\mathbf{x})$. It follows that

$$
\begin{aligned}
\widehat{\tau}(\mathcal{D};\Theta_0) =& \frac{\int_{\Theta_0} \prod_{i=1}^n f(\mathbf{X}_i^{\mathrm{obs}}|\theta)d\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n f(\mathbf{X}_i^{\mathrm{obs}}|\theta)d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n p(\mathbf{X}_i^{\mathrm{obs}}|\theta)h(\mathbf{X}_i^{\mathrm{obs}})d\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n p(\mathbf{X}_i^{\mathrm{obs}}|\theta)h(\mathbf{X}_i^{\mathrm{obs}})d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n p(\mathbf{X}_i^{\mathrm{obs}}|\theta)d\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n p(\mathbf{X}_i^{\mathrm{obs}}|\theta)d\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \mathcal{L}(\mathcal{D};\theta)d\pi_0(\theta)}{\int_{\Theta_1} \mathcal{L}(\mathcal{D};\theta)d\pi_1(\theta)} \\
=& \mathrm{BF}(\mathcal{D};\Theta_0).
\end{aligned}
$$

∎

Proof [Proof of Theorem 1] By definition, for every fixed $c_{B'}$, $\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}(\lambda(\mathcal{D};\theta_0) \leq c_{B'}) = F(c_{B'}|\theta_0)$. It follows that the random variable $\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}(\lambda(\mathcal{D};\theta_0) \leq C_{B'}) = F(C_{B'}|\theta_0)$. Moreover, by construction, $\alpha = \widehat{F}_{B'}(C_{B'}|\theta_0)$. It follows that

$$
\begin{aligned}
|\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}(\lambda(\mathcal{D};\theta_0) \leq C_{B'}) - \alpha| &= |F(C_{B'}|\theta_0) - \alpha| \\
&= |F(C_{B'}|\theta_0) - \widehat{F}_{B'}(C_{B'}|\theta_0)| \\
&\leq \sup_{\lambda \in \mathbb{R}} |F(\lambda|\theta_0) - \widehat{F}_{B'}(\lambda|\theta_0)| \xrightarrow[B' \to \infty]{P} 0.
\end{aligned}
$$

The result follows from the fact that convergence in probability to a constant implies almost sure convergence. ∎

Proof [Proof of Theorem 2] The proof follows from applying the convergence rate to the last equation in the proof of Theorem 1. ∎

Assumption 9 (Uniform consistency in θ and λ) *Let $\widehat{F}_{B'}(\cdot|\theta)$ be the estimated cumulative distribution function of the test statistic $\lambda(\mathcal{D};\Theta_0)$ conditional on θ based on a sample $\mathcal{T}'$ with size B' implied by the quantile regression, and let $F(\cdot|\theta)$ be its true distribution given θ. Assume that the quantile regression estimator is such that*

$$
\sup_{\theta \in \Theta_0, \lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B' \to \infty]{P} 0.
$$

This assumption holds, for instance, for quantile regression forests [68] under additional assumptions (see Proposition 2).

Proposition 2 *If, for every $\theta \in \Theta_0$, the quantile regression estimator is such that*

$$
\sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B' \to \infty]{P} 0 \tag{18}
$$

and either

- $|\Theta| < \infty$ *or,*
- Θ *is a compact subset of $\mathbb{R}^d$, and the function $g_{B'}(\theta) = \sup_{t \in \mathbb{R}} |\widehat{F}_{B'}(t|\theta) - F(t|\theta)|$ is almost surely continuous in θ and strictly decreasing in B',*

then Assumption 9 holds.

Proof If $|\Theta| < \infty$, the union bound and Equation 18 imply that

$$
\sup_{\theta \in \Theta_0} \sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B' \to \infty]{P} 0. \tag{19}
$$

Similarly, by Dini's theorem, Equation 19 also holds if Θ is a compact subset of $\mathbb{R}^d$, and the function $g_{B'}(\theta)$ is continuous in θ and strictly decreasing in B'. ∎

Theorem 8 *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on a absolutely continuous statistic $\lambda(\mathcal{D}; \Theta_0)$ chosen according to Algorithm 1 for a fixed $\alpha \in (0,1)$. If the quantile estimator satisfies Assumption 9, then*

$$C_{B'} \xrightarrow[B' \longrightarrow \infty]{P} C^*,$$

where C^* is such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \Theta_0) \leq C^*) = \alpha.$$

Proof Assumption 9 implies that

$$\sup_{\theta \in \Theta_0} |\widehat{F}_{B'}^{-1}(\alpha|\theta) - F^{-1}(\alpha|\theta)| \xrightarrow[B' \longrightarrow \infty]{P} 0.$$

The result then follows from the fact that

$$0 \leq |C_{B'} - C^*| = |\sup_{\theta \in \Theta_0} \widehat{F}_{B'}^{-1}(\alpha|\theta) - \sup_{\theta \in \Theta_0} F^{-1}(\alpha|\theta)|$$
$$\leq \sup_{\theta \in \Theta_0} |\widehat{F}_{B'}^{-1}(\alpha|\theta) - F^{-1}(\alpha|\theta)|,$$

and thus

$$|C_{B'} - C^*| \xrightarrow[B' \longrightarrow \infty]{P} 0.$$

∎

Lemma 1 *Let $g_1, g_2, \ldots$ be a sequence of random functions such that $g_i : \mathcal{Z} \longrightarrow \mathbb{R}$, and let Z be a random quantity defined over $\mathcal{Z}$, independent of the random functions. Assume that $g(Z)$ is absolutely continuous with respect to the Lebesgue measure. If, for every $z \in \mathcal{Z}$,*

$$g_m(z) \xrightarrow[m \longrightarrow \infty]{a.s.} g(z),$$

then

$$g_m(Z) \xrightarrow[m \longrightarrow \infty]{\mathcal{L}} g(Z).$$

Proof Fix $y \in \mathbb{R}$ and let $A_y = \{z \in \mathcal{Z} : g(z) \neq y\}$. Notice that $\mathbb{P}(Z \in A_y) = 1$. Moreover, the almost sure convergence of $g_m(z)$ implies its convergence in distribution. It follows that for every $z \in A_y$,

$$\lim_m \mathbb{P}(g_m(z) \leq y) = \mathbb{P}\left(g(z) \leq y\right). \tag{20}$$

Now, using Equation 20 and Lebesgue's dominated convergence theorem, notice that

$$
\begin{aligned}
\lim_{m} \mathbb{P}(g_m(Z) < y) &= \lim_{m} \int_{\mathcal{Z}} \mathbb{P}(g_m(Z) < y | Z = z) d\mathbb{P}_Z(z) \\
&= \int_{\mathcal{Z}} \lim_{m} \mathbb{P}(g_m(Z) < y | Z = z) d\mathbb{P}_Z(z) \\
&= \int_{A_z} \lim_{m} \mathbb{P}(g_m(z) < y) d\mathbb{P}_Z(z) = \int_{A_z} \mathbb{P}(g(z) < y) d\mathbb{P}_Z(z) \\
&= \int_{\mathcal{Z}} \mathbb{P}(g(Z) < y | Z = z) d\mathbb{P}_Z(z) = \mathbb{P}(g(Z) < y),
\end{aligned}
$$

which concludes the proof. ∎

Proof [Proof of Theorem 3] Assumption 3 implies that, for every D,

$$
\begin{aligned}
0 \le |\widehat{p}(D; \Theta_0) - p(D; \Theta_0)| &= |\sup_{\theta \in \Theta_0} \widehat{p}(D; \theta) - \sup_{\theta \in \Theta_0} p(D; \theta)| \\
&\le \sup_{\theta \in \Theta_0} |\widehat{p}(D; \theta) - p(D; \theta)| \xrightarrow[B' \longrightarrow \infty]{\text{a.s.}} 0,
\end{aligned}
$$

and therefore $\widehat{p}(D; \Theta_0)$ converges almost surely to $p(D; \Theta_0)$. It follows from Lemma 1 that $\widehat{p}(\mathcal{D}; \Theta_0)$ converges in distribution to $p(\mathcal{D}; \Theta_0)$. Conclude that

$$
\begin{aligned}
\mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \Theta_0) \le \alpha) = F_{\widehat{p}(\mathcal{D}; \Theta_0) | \theta}(\alpha) &\xrightarrow[B' \longrightarrow \infty]{} F_{p(\mathcal{D}; \Theta_0) | \theta}(\alpha) \\
&= \mathbb{P}_{\mathcal{D} | \theta}(p(D; \Theta_0) \le \alpha),
\end{aligned}
$$

where F_Z denotes the cumulative distribution function of the random variable Z. ∎

Proof [Proof of Corollary 1] Fix $\theta \in \Theta$. Because F_θ is continuous, the definition of $p(\mathcal{D}; \theta)$ implies that its distribution is uniform under the null. Thus $\mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \theta) \le \alpha) = \alpha$. Theorem 3 therefore implies that

$$
\mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \theta) \le \alpha) \xrightarrow[B' \longrightarrow \infty]{} \mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \theta) \le \alpha) = \alpha. \tag{21}
$$

Now, for any $\theta \in \Theta_0$, uniformity of the p-value implies that

$$
\begin{aligned}
\mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \Theta_0) \le \alpha) = \mathbb{P}_{\mathcal{D} | \theta}\left(\sup_{\theta_0 \in \Theta_0} p(\mathcal{D}; \theta_0) \le \alpha\right) &\le \mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \theta) \le \alpha) \\
&= \alpha.
\end{aligned}
$$

Conclude from Theorem 3 that

$$
\mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\widehat{p}(\mathcal{D}; \Theta_0) \le \alpha) \xrightarrow[B' \longrightarrow \infty]{} \mathbb{P}_{\mathcal{D} | \theta}(p(\mathcal{D}; \Theta_0) \le \alpha) \le \alpha. \tag{22}
$$

The conclusion follows from putting together Equations 21 and 22. ∎

Proof [Proof of Theorem 4]

$$|\widehat{p}(D;\Theta_0) - p(D;\Theta_0)| = |\sup_{\theta \in \Theta_0} \widehat{p}(D;\theta) - \sup_{\theta \in \Theta_0} p(D;\theta)|$$

$$\leq \sup_{\theta \in \Theta_0} |\widehat{p}(D;\theta) - p(D;\theta)|$$

$$= O_P\left(\left(\frac{1}{B'}\right)^r\right),$$

where the last line follows from Assumption 4 ∎

Lemma 2 *Under Assumption 5, for every* $\theta, \theta_0 \in \Theta$

$$\mathbb{E}^2_{\mathcal{D}|\theta,T}\left[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|\right] \leq M^2 \int (\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0))^2 dG(\mathbf{x}).$$

Proof For every $\theta \in \Theta$,

$$\mathbb{E}^2_{\mathcal{D}|\theta,T}[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|]$$

$$= \left(\int |\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)| \; dF(\mathbf{x}|\theta)\right)^2$$

$$= \left(\int |\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0)| \; dF(\mathbf{x}|\theta)\right)^2$$

$$= \left(\int |\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0)|\mathbb{O}(\mathbf{x};\theta)dG(\mathbf{x})\right)^2$$

$$\leq \left(\int (\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0)^2 dG(\mathbf{x})\right)\left(\int \mathbb{O}^2(\mathbf{x};\theta)dG(\mathbf{x})\right),$$

where the last inequality follows from Cauchy-Schwarz. Assumption 5 implies that

$$\int \mathbb{O}^2(\mathbf{x};\theta)dG(\mathbf{x}) \leq M^2,$$

from which we conclude that

$$\mathbb{E}^2_{\mathcal{D}|\theta,T}[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|] \leq M^2 \int (\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0))^2 dG(\mathbf{x}).$$

∎

Lemma 3 *For fixed* $c \in \mathbb{R}$, *let* $\phi_{\tau;\theta_0}(\mathcal{D}) = \mathbb{I}(\tau(\mathcal{D};\theta_0) < c)$ *and* $\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = \mathbb{I}(\widehat{\tau}_B(\mathcal{D};\theta_0) < c)$ *be the testing procedures for testing* $H_{0,\theta_0} : \theta = \theta_0$ *obtained using* τ *and* $\widehat{\tau}_B$. *Under Assumptions 5-6, for every* $0 < \epsilon < 1$,

$$\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})) \leq \frac{2MC_L \cdot \sqrt{\int (\mathbb{O}(\mathbf{x};\theta_0) - \widehat{\mathbb{O}}(\mathbf{x};\theta_0))^2 dG(\mathbf{x})}}{\epsilon} + \epsilon.$$

Proof [Proof of Lemma 3] It follows from Markov's inequality and Lemma 2 that with probability at least $1 - \epsilon$, $\mathcal{D}$ is such that

$$|\tau(\mathcal{D}; \theta_0) - \widehat{\tau}(\mathcal{D}; \theta_0)| \leq \frac{M \cdot \sqrt{\int (\mathbb{O}(\mathbf{x}; \theta_0) - \widehat{\mathbb{O}}(\mathbf{x}; \theta_0))^2 dG(\mathbf{x})}}{\epsilon} \qquad (23)$$

Now we upper bound $\mathbb{P}_{\mathcal{D}|\theta, T}(\phi_{\tau; \theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}; \theta_0}(\mathcal{D}))$. Define A as the event that Eq. 23 happens and let $h(\theta_0) := \int (\mathbb{O}(\mathbf{x}; \theta_0) - \widehat{\mathbb{O}}(\mathbf{x}; \theta_0))^2 dG(\mathbf{x})$. Then:

$$\mathbb{P}_{\mathcal{D}|\theta, T}(\phi_{\tau; \theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}; \theta_0}(\mathcal{D}))$$
$$\leq \mathbb{P}_{\mathcal{D}|\theta, T}(\phi_{\tau; \theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}; \theta_0}(\mathcal{D}), A) + \mathbb{P}_\theta(A^c)$$
$$\leq \mathbb{P}_{\mathcal{D}|\theta, T}(\mathbb{I}(\tau(\mathcal{D}; \theta_0) < c) \neq \mathbb{I}(\widehat{\tau}(\mathcal{D}; \theta_0) < c), A) + \epsilon$$
$$\leq \mathbb{P}_{\mathcal{D}|\theta, T}\left(c - \frac{M \cdot \sqrt{h(\theta_0)}}{\epsilon} < \tau(\mathcal{D}; \theta_0) < c + \frac{M \cdot \sqrt{h(\theta_0)}}{\epsilon}\right) + \epsilon$$

Assumption 6 then implies that

$$\mathbb{P}_{\mathcal{D}|\theta, T}(\phi_{\tau; \theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}; \theta_0}(\mathcal{D})) \leq \frac{K' \cdot \sqrt{h(\theta_0)}}{\epsilon} + \epsilon$$

where $K' = 2MC_L$, which concludes the proof. ∎

Proof [Proof of Theorem 5] Follows directly from Lemma 3 and Jensen's inequality. ∎

Lemma 4 *Under Assumptions 5-8, there exists $C > 0$ such that*

$$\mathbb{E}_{\mathcal{T}}\left[L(\widehat{\mathbb{O}}, \mathbb{O})\right] \leq CB^{-\kappa/((\kappa+d))}.$$

Proof Let $\widehat{p} = \widehat{\mathbb{P}}(Y = 1|\mathbf{x}, \theta)$ and $p = \mathbb{P}(Y = 1|\mathbf{x}, \theta)$ be the probabilistic classifier and true classification function, respectively, on the training sample T. Let $h(y) = \frac{y}{1-y}$ for $0 < y < 1$. A Taylor expansion of h implies that

$$(h(\widehat{p}) - h(p))^2 = (h(p) + R_1(\widehat{p}) - h(p))^2 = R_1(\widehat{p})^2,$$

where $R_1(\widehat{p}) = h'(\xi)(\widehat{p} - p)$ for some ξ between p and $\widehat{p}$. Also note that due to Assumption 5,
$$\exists a > 0 \text{ s.t. } p, \widehat{p} > a, \; \forall x \in \mathcal{X}, \theta \in \Theta.$$
Thus,

$$\mathbb{E}_{\mathcal{T}}\left[\iint (h(\widehat{p}) - h(p))^2 dG(\mathbf{x}) d\pi(\theta)\right]$$

$$= \mathbb{E}_{\mathcal{T}} \left[\iint \frac{1}{(1-\xi)^4} \left(\widehat{p} - p \right)^2 dG(\mathbf{x}) d\pi(\theta) \right]$$

$$\leq \frac{1}{(1-a)^4} \mathbb{E}_{\mathcal{T}} \left[\iint \left(\widehat{p} - p \right)^2 dG(\mathbf{x}) d\pi(\theta) \right]$$

$$= \frac{1}{(1-a)^4} \mathbb{E}_{\mathcal{T}} \left[\int \left(\widehat{\mathbb{P}}(Y=1|\mathbf{x},\theta) - \mathbb{P}(Y=1|\mathbf{x},\theta) \right)^2 h'(\mathbf{x},\theta) dH(\mathbf{x},\theta) \right]$$

$$\leq \frac{\gamma}{(1-a)^4} \mathbb{E}_{\mathcal{T}} \left[\int \left(\widehat{\mathbb{P}}(Y=1|\mathbf{x},\theta) - \mathbb{P}(Y=1|\mathbf{x},\theta) \right)^2 dH(\mathbf{x},\theta) \right]$$

$$= O\left(B^{-\kappa/(\kappa+d)} \right).$$

∎

Proof [Proof of Theorem 6] It follows from Theorem 5 that

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})) d\pi(\theta_0)$$

$$= \mathbb{E}_{\mathcal{T}} \left[\int \mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})) d\pi(\theta_0) \right]$$

$$\leq \frac{2MC_L \cdot \mathbb{E}_{\mathcal{T}} \left[\sqrt{L(\widehat{\mathbb{O}}, \mathbb{O})} \right]}{\epsilon} + \epsilon$$

$$\leq \frac{2MC_L \cdot \sqrt{\mathbb{E}_{\mathcal{T}} \left[L(\widehat{\mathbb{O}}, \mathbb{O}) \right]}}{\epsilon} + \epsilon,$$

where the last step follows from Jensen's inequality. It follows from this and Lemma 4 that

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})) d\pi(\theta_0) \leq \frac{KB^{-\kappa/(2(\kappa+d))}}{\epsilon} + \epsilon,$$

where $K = 2MC_L\sqrt{C}$. Notice that taking $\epsilon^* = \sqrt{K}B^{-\kappa/(4(\kappa+d))}$ optimizes the bound and gives the result. ∎

Proof [Proof of Corollary 2] The result follows from noticing that

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = 1) \geq \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1) - \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})),$$

and therefore

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = 1) d\theta_0 \geq \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1) d\theta_0$$

$$- \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})) d\theta_0$$

$$\geq \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1)d\theta_0 - K'B^{-\kappa/(4(\kappa+d))},$$

where the last inequality follows from Theorem 6. ∎

Appendix G: Loss functions

In this work, we use the cross-entropy loss to train probabilistic classifiers. Consider a sample point $\{\theta, \mathbf{x}, y\}$ generated according to Algorithm 3. Let p be a Bernoulli(y) distribution, and q be a Ber $\left(\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{x})\right) = \text{Ber}\left(\frac{\widehat{\mathbb{O}}(\mathbf{x};\theta)}{1+\widehat{\mathbb{O}}(\mathbf{x};\theta)}\right)$ distribution. The *cross-entropy* between p and q is given by

$$\mathcal{L}_{\text{CE}}(\widehat{\mathbb{O}}; \{\theta, \mathbf{x}, y\}) = -y \log\left(\frac{\widehat{\mathbb{O}}(\mathbf{x};\theta)}{1+\widehat{\mathbb{O}}(\mathbf{x};\theta)}\right) - (1-y)\log\left(\frac{1}{1+\widehat{\mathbb{O}}(\mathbf{x};\theta)}\right)$$
$$= -y\log\left(\widehat{\mathbb{O}}(\mathbf{x};\theta)\right) + \log\left(1+\widehat{\mathbb{O}}(\mathbf{x};\theta)\right). \quad (24)$$

For every $\mathbf{x}$ and θ, the expected cross-entropy $\mathbb{E}[L_{\text{CE}}(\widehat{\mathbb{O}}; \{\theta, \mathbf{x}, y\})]$ is minimized by $\widehat{\mathbb{O}}(\mathbf{x};\theta) = \mathbb{O}(\mathbf{x};\theta)$. If the probabilistic classifier attains the minimum of the cross-entropy loss, then the estimated `ACORE` statistic $\widehat{\Lambda}(\mathcal{D};\Theta_0)$ will be equal to the likelihood ratio statistic in Equation 5, as shown in [26]. Similarly, as stated in Proposition 1, at the minimum, the estimated `BFF` statistic $\widehat{\tau}(\mathcal{D};\Theta_0)$ is equal to the Bayes factor in Equation 6.

Acknowledgments

References

[1] AAD, G., ABAJYAN, T., ABBOTT, B., ABDALLAH, J., ABDEL KHALEK, S., ABDELALIM, A. A., ABDINOV, O., ABEN, R., ABI, B., ABOLINS, M. et al. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B* **716** 1–29. https://doi.org/10.1016/j.physletb.2012.08.020

[2] AAD, G., ABAJYAN, T., ABBOTT, B., ABDALLAH, J., KHALEK, S. A., ABDELALIM, A. A., ABEN, R., ABI, B., ABOLINS, M., ABOUZEID, O. et al. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B* **716** 1–29.

[3] AGOSTINELLI, S., ALLISON, J., AMAKO, K. a., APOSTOLAKIS, J., ARAUJO, H., ARCE, P., ASAI, M., AXEN, D., BANERJEE, S., BARRAND, G. et al. (2003). GEANT4—a simulation toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** 250–303.

[4] ALGERI, S., AALBERS, J., MORÅ, K. D. and CONRAD, J. (2019). Searching for new physics with profile likelihoods: Wilks and beyond. *arXiv preprint arXiv:1911.10237*.

[5] AUGUSTIN, J.-E., BOYARSKI, A. M., BREIDENBACH, M., BULOS, F., DAKIN, J., FELDMAN, G., FISCHER, G., FRYBERGER, D., HANSON, G., JEAN-MARIE, B. et al. (1974). Discovery of a Narrow Resonance in e+ e-Annihilation. *Physical Review Letters* **33** 1406.

[6] BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.

[7] BIERENS, H. J. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association* **78** 699–707. MR0721221

[8] BLUM, M. G. and FRANÇOIS, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and computing* **20** 63–73. MR2578077

[9] BORDOLOI, R., LILLY, S. J. and AMARA, A. (2010). Photo-z performance for precision cosmology. *Monthly Notices of the Royal Astronomical Society* **406** 881–895.

[10] BREHMER, J., LOUPPE, G., PAVEZ, J. and CRANMER, K. (2020). Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences* **117** 5242–5249. https://doi.org/10.1073/pnas.1915980117 MR4225023

[11] CHATRCHYAN, S., KHACHATRYAN, V., SIRUNYAN, A. M., TUMASYAN, A., ADAM, W., AGUILO, E., BERGAUER, T., DRAGICEVIC, M., ERÖ, J., FABJAN, C. et al. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B* **716** 30–61.

[12] CHEN, J. and LI, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics* **37** 2523–2542. MR2543701

[13] CHEN, Y. and GUTMANN, M. U. (2019). Adaptive Gaussian Copula ABC. In *Proceedings of Machine Learning Research* (K. CHAUDHURI and M. SUGIYAMA, eds.). *Proceedings of Machine Learning Research* **89** 1584–1592. PMLR.

[14] CHUANG, C.-S. and LAI, T. L. (2000). HYBRID RESAMPLING METHODS FOR CONFIDENCE INTERVALS. *Statistica Sinica* **10** 1–33. MR1742099

[15] COLLABORATION, C. et al. (1995). Observation of top quark production in Pbar-P collisions. *arXiv preprint hep-ex/9503002*.

[16] COOK, S. R., GELMAN, A. and RUBIN, D. B. (2006). Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics* **15** 675-692. https://doi.org/10.1198/106186006X136976 MR2291268

[17] Cousins, R. D. (2006). Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature. In *Statistical Problems In Particle Physics, Astrophysics And Cosmology* 75–85. World Scientific.

[18] Cousins, R. D. (2018). Lectures on Statistics in Theory: Prelude to Statistics in Practice.

[19] Cousins, R. D., Linnemann, J. T. and Tucker, J. (2008). Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **595** 480–501.

[20] Cowan, G. (2012). Discovery sensitivity for a counting experiment with back- ground uncertainty. *Technical Report.*

[21] Cowan, G., Cranmer, K., Gross, E. and Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C* **71**. https://doi.org/10.1140/epjc/s10052-011-1554-0

[22] Cranmer, K. (2015). Practical Statistics for the LHC. *arXiv e-prints* arXiv:1503.07622.

[23] Cranmer, K., Brehmer, J. and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences* **117** 30055–30062. https://doi.org/10.1073/pnas.1912789117 MR4263287

[24] Cranmer, K., Pavez, J. and Louppe, G. (2015). Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. *arXiv preprint arXiv:1506.02169.*

[25] Dacunha-Castelle, D. and Gassiat, E. (1997). Testing in locally conic models, and application to mixture models. *ESAIM: Probability and Statistics* **1** 285–317. MR1468112

[26] Dalmasso, N., Izbicki, R. and Lee, A. (2020). Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting. In *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 2323–2334. PMLR, Virtual.

[27] Dey, B., Newman, J. A., Andrews, B. H., Izbicki, R., Lee, A. B., Zhao, D., Rau, M. M. and Malz, A. I. (2021). Re-calibrating Photometric Redshift Probability Distributions Using Feature-space Regression. *arXiv preprint arXiv:2110.15209.*

[28] Dey, B., Zhao, D., Newman, J. A., Andrews, B. H., Izbicki, R. and Lee, A. B. (2022). Calibrated predictive distributions via diagnostics for conditional coverage. *arXiv preprint arXiv:2205.14568.*

[29] Donoho, D. L. (1994). Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probability Theory and Related Fields* **99** 145–170. MR1278880

[30] Dorigo, T., Guglielmini, S., Kieseler, J., Layer, L. and Strong, G. C. (2022). Deep Regression of Muon Energy with a K-Nearest

Neighbor Algorithm. *arXiv preprint arXiv:2203.02841*.

[31] DRTON, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics* **37** 979–1012. https://doi.org/10.1214/07-aos571 MR2502658

[32] DURKAN, C., MURRAY, I. and PAPAMAKARIOS, G. (2020). On Contrastive Learning for Likelihood-free Inference. In *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. SINGH, eds.). *Proceedings of Machine Learning Research* **119** 2771–2781. PMLR.

[33] FASIOLO, M., WOOD, S. N., HARTIG, F. and BRAVINGTON, M. V. (2018). An extended empirical saddlepoint approximation for intractable likelihoods. *Electron. J. Statist.* **12** 1544–1578. https://doi.org/10.1214/18-EJS1433 MR3806432

[34] FELDMAN, G. (2000). Multiple measurements and parameters in the unified approach Technical Report, Technical Report, Talk at the FermiLab Workshop on Confidence Limits.

[35] FELDMAN, G. J. and COUSINS, R. D. (1998). Unified approach to the classical statistical analysis of small signals. *Physical Review D* **57** 3873–3889. https://doi.org/10.1103/physrevd.57.3873

[36] FISHER, R. A. (1925). *Statistical Methods for Research Workers*, 11th ed. rev. ed. Oliver and Boyd: Edinburgh.

[37] GERBER, F. and NYCHKA, D. (2021). Fast covariance parameter estimation of spatial Gaussian process models using neural networks. *Stat* **10** e382. MR4318995

[38] GIRARD, S., GUILLOU, A. and STUPFLER, G. (2014). Uniform strong consistency of a frontier estimator using kernel regression on high order moments. *ESAIM: Probability and Statistics* **18** 642–666. MR3334007

[39] GREENBERG, D., NONNENMACHER, M. and MACKE, J. (2019). Automatic Posterior Transformation for Likelihood-Free Inference. In *Proceedings of the 36th International Conference on Machine Learning* (K. CHAUDHURI and R. SALAKHUTDINOV, eds.). *Proceedings of Machine Learning Research* **97** 2404–2414. PMLR, Long Beach, California, USA.

[40] GUTMANN, M. U. and CORANDER, J. (2016). Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. *Journal of Machine Learning Research* **17** 1-47. MR3555016

[41] HARDLE, W., LUCKHAUS, S. et al. (1984). Uniform consistency of a class of regression function estimators. *The Annals of Statistics* **12** 612–623. MR0740915

[42] HEINRICH, L. (2022). Learning Optimal Test Statistics in the Presence of Nuisance Parameters. *arXiv preprint arXiv:2203.13079*.

[43] HERB, S., HOM, D., LEDERMAN, L., SENS, J., SNYDER, H., YOH, J., APPEL, J., BROWN, B., BROWN, C., INNES, W. et al. (1977). Observation of a dimuon resonance at 9.5 GeV in 400-GeV proton-nucleus collisions. *Physical Review Letters* **39** 252.

[44] HERMANS, J., BEGY, V. and LOUPPE, G. (2020). Likelihood-free MCMC with Amortized Approximate Ratio Estimators. *arXiv preprint arXiv:1903.04057*.

[45] HERMANS, J., DELAUNOY, A., ROZET, F., WEHENKEL, A. and LOUPPE, G. (2021). Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*.

[46] HO, M., FARAHI, A., RAU, M. M. and TRAC, H. (2021). Approximate Bayesian Uncertainties on Deep Learning Dynamical Mass Estimates of Galaxy Clusters. *The Astrophysical Journal* **908** 204.

[47] IZBICKI, R., LEE, A. and SCHAFER, C. (2014). High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (S. KASKI and J. CORANDER, eds.). *Proceedings of Machine Learning Research* **33** 420–429. PMLR, Reykjavik, Iceland.

[48] IZBICKI, R., LEE, A. B. and POSPISIL, T. (2019). ABC–CDE: Toward Approximate Bayesian Computation With Complex High-Dimensional Data and Limited Simulations. *Journal of Computational and Graphical Statistics* 1–20. https://doi.org/10.1080/10618600.2018.1546594 MR4007735

[49] JEFFREYS, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. *Mathematical Proceedings of the Cambridge Philosophical Society* **31** 203–222. https://doi.org/10.1017/S030500410001330X

[50] JEFFREYS, H. (1961). *Theory of probability*, 3rd ed. ed. Clarendon Press Oxford. MR0187257

[51] JÄRVENPÄÄ, M., GUTMANN, M. U., VEHTARI, A. and MARTTINEN, P. (2021). Parallel Gaussian Process Surrogate Bayesian Inference with Noisy Likelihood Evaluations. *Bayesian Anal.* **16** 147–178. https://doi.org/10.1214/20-BA1200 MR4194277

[52] KIESELER, J., STRONG, G. C., CHIANDOTTO, F., DORIGO, T. and LAYER, L. Preprocessed dataset for "Calorimetric measurement of multi-TeV muons via deep regression", August 2021. *URL https://doi.org/10.5281/zenodo* **5163817**.

[53] KIESELER, J., STRONG, G. C., CHIANDOTTO, F., DORIGO, T. and LAYER, L. (2022). Calorimetric Measurement of Multi-TeV Muons via Deep Regression. *The European Physical Journal C* **82** 1–26.

[54] KOENKER, R., CHERNOZHUKOV, V., HE, X. and PENG, L. (2017). *Handbook of quantile regression*. CRC press. MR3728340

[55] LEMOS, P., COOGAN, A., HEZAVEH, Y. and PERREAULT-LEVASSEUR, L. (2023). Sampling-based accuracy testing of posterior estimators for general inference. *arXiv preprint arXiv:2302.03026*.

[56] LENZI, A., BESSAC, J., RUDI, J. and STEIN, M. L. (2021). Neural networks for parameter estimation in intractable models. *arXiv preprint arXiv:2107.14346*. MR4578090

[57] LIERO, H. (1989). Strong uniform consistency of nonparametric regression function estimates. *Probability theory and related fields* **82** 587–614. MR1002902

[58] LINHART, J., GRAMFORT, A. and RODRIGUES, P. L. (2023). L-C2ST: Local Diagnostics for Posterior Approximations in Simulation-Based Infer-

ence. *arXiv preprint* arXiv:2306.03580.

[59] LUECKMANN, J.-M., BASSETTO, G., KARALETSOS, T. and MACKE, J. H. (2019). Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference* 32–53. [MR3980813](MR3980813)

[60] LUECKMANN, J.-M., GONCALVES, P. J., BASSETTO, G., ÖCAL, K., NONNENMACHER, M. and MACKE, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 1289–1299. Curran Associates, Inc.

[61] LYONS, L. (2008). Open statistical issues in Particle Physics. *The Annals of Applied Statistics* **2** 887 – 915. https://doi.org/10.1214/08-AOAS163 [MR2516798](MR2516798)

[62] MACKINNON, J. G. (2009). Bootstrap hypothesis testing. *Handbook of computational econometrics* **183** 213.

[63] MARIN, J.-M., PUDLO, P., ROBERT, C. and RYDER, R. (2012). Approximate Bayesian computational methods. *Statistics and Computing* **22** 1167-1180. https://doi.org/10.1007/s11222-011-9288-2 [MR2992292](MR2992292)

[64] MARIN, J.-M., RAYNAL, L., PUDLO, P., RIBATET, M. and ROBERT, C. (2016). ABC random forests for Bayesian parameter inference. *Bioinformatics (Oxford, England)* **35**. https://doi.org/10.1093/bioinformatics/bty867

[65] MASSERANO, L., DORIGO, T., IZBICKI, R., KUUSELA, M. and LEE, A. (2023). Simulator-Based Inference with WALDO: Confidence Regions by Leveraging Prediction Algorithms and Posterior Estimators for Inverse Problems. In *International Conference on Artificial Intelligence and Statistics* 2960–2974. PMLR.

[66] MCLACHLAN, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **36** 318–324.

[67] MEEDS, E. and WELLING, M. (2014). GPS-ABC: Gaussian process surrogate approximate Bayesian computation. *arXiv preprint* arXiv:1401.2838.

[68] MEINSHAUSEN, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research* **7** 983-999. [MR2274394](MR2274394)

[69] NEYMAN, J. (1935). On the Problem of Confidence Intervals. *Ann. Math. Statist.* **6** 111–116. https://doi.org/10.1214/aoms/1177732585

[70] NEYMAN, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **236** 333–380.

[71] NEYMAN, J. and PEARSON, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika* **20A** 175–240.

[72] PAPAMAKARIOS, G. and MURRAY, I. (2016). Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation. In *Advances in Neural Information Processing Systems* (D. LEE, M. SUGIYAMA,

U. Luxburg, I. Guyon and R. Garnett, eds.) **29** 1028–1036. Curran Associates, Inc.

[73] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. and Lakshminarayanan, B. (2021). Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research* **22** 1-64. MR4253750

[74] Papamakarios, G., Sterratt, D. and Murray, I. (2019). Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows. In *The 22nd International Conference on Artificial Intelligence and Statistics* 837–848.

[75] Picchini, U., Simola, U. and Corander, J. (2020). Adaptive MCMC for synthetic likelihoods and correlated synthetic likelihoods. *arXiv preprint arXiv:2004.04558*.

[76] Qian, X., Tan, A., Ling, J. J., Nakajima, Y. and Zhang, C. (2016). The Gaussian CL_s method for searches of new physics. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **827** 63-78.

[77] Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L. and Köthe, U. (2020). BayesFlow: Learning Complex Stochastic Models With Invertible Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 1-15. https://doi.org/10.1109/TNNLS.2020.3042395 MR4516681

[78] Redner, R. (1981). Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions. *The Annals of Statistics* **9** 225–228. MR0600553

[79] Schmidt, S. J., Malz, A. I., Soo, J. Y. H., Almosallam, I. A., Brescia, M., Cavuoti, S., Cohen-Tanugi, J., Connolly, A. J., DeRose, J., Freeman, P. E., Graham, M. L., Iyer, K. G., Jarvis, M. J., Kalmbach, J. B., Kovacs, E., Lee, A. B., Longo, G., Morrison, C. B., Newman, J. A., Nourbakhsh, E., Nuss, E., Pospisil, T., Tranin, H., Wechsler, R. H., Zhou, R., Izbicki, R. and Collaboration), T. L. D. E. S. (2020). Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society* **499** 1587-1606. https://doi.org/10.1093/mnras/staa2799

[80] Sen, B., Walker, M. and Woodroofe, M. (2009). ON THE UNIFIED METHOD WITH NUISANCE PARAMETERS. *Statistica Sinica* **19** 301–314. MR2487891

[81] Sisson, S. A., Fan, Y. and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation.* Chapman and Hall/CRC. MR3889281

[82] Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104** 1760–1765. MR2301870

[83] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* 1040–1053. MR0673642

[84] TALTS, S., BETANCOURT, M., SIMPSON, D., VEHTARI, A. and GELMAN, A. (2018). Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv preprint arXiv:1804.06788.*

[85] THOMAS, O., DUTTA, R., CORANDER, J., KASKI, S. and GUTMANN, M. U. (2021). Likelihood-Free Inference by Ratio Estimation. *Bayesian Anal.* Advance publication. https://doi.org/10.1214/20-BA1238 [MR4377135](MR4377135)

[86] VAN DEN BOOM, W., REEVES, G. and DUNSON, D. B. (2020). Approximating posteriors with high-dimensional nuisance parameters via integrated rotated Gaussian approximation. *Biometrika.* https://doi.org/10.1093/biomet/asaa068 [MR4259131](MR4259131)

[87] VENTURA, V. (2010). Bootstrap tests of hypotheses. In *Analysis of parallel spike trains* 383–398. Springer.

[88] WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society* **54** 426–482. [MR0012401](MR0012401)

[89] WARNE, D. J., MACLAREN, O. J., CARR, E. J., SIMPSON, M. J. and DROVANDI, C. (2024). Generalised likelihood profiles for models with intractable likelihoods. *Statistics and Computing* **34** 50. [MR4671704](MR4671704)

[90] WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences* **117** 16880–16890. [MR4242731](MR4242731)

[91] WILKINSON, R. (2014). Accelerating ABC methods using Gaussian processes. In *Artificial Intelligence and Statistics* 1015–1023.

[92] WILKS, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.* **9** 60–62. https://doi.org/10.1214/aoms/1177732360

[93] WOOD, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102-4. https://doi.org/10.1038/nature09319

[94] YANG, Y., BHATTACHARYA, A. and PATI, D. (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753.*

[95] ZHAO, D., DALMASSO, N., IZBICKI, R. and LEE, A. B. (2021). Diagnostics for conditional density models and bayesian inference algorithms. In *Uncertainty in Artificial Intelligence* 1830–1840. PMLR.

[96] ZHU, Y., SHEN, X. and PAN, W. (2020). On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association* **115** 217-230. [MR4078458](MR4078458)