Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting

Niccolò Dalmasso [1] **Rafael Izbicki** [2] **Ann B. Lee** [1]

Abstract

Parameter estimation, statistical tests and confidence sets are the cornerstones of classical statistics that allow scientists to make inferences about the underlying process that generated the observed data. A key question is whether one can still construct hypothesis tests and confidence sets with proper coverage and high power in a so-called likelihood-free inference (LFI) setting; that is, a setting where the likelihood is not explicitly known but one can forward-simulate observable data according to a stochastic model. In this paper, we present `ACORE` (Approximate Computation via Odds Ratio Estimation), a frequentist approach to LFI that first formulates the classical likelihood ratio test (LRT) as a parametrized classification problem, and then uses the equivalence of tests and confidence sets to build confidence regions for parameters of interest. We also present a goodness-of-fit procedure for checking whether the constructed tests and confidence regions are valid. `ACORE` is based on the key observation that the LRT statistic, the rejection probability of the test, and the coverage of the confidence set are conditional distribution functions which often vary smoothly as a function of the parameters of interest. Hence, instead of relying solely on samples simulated at fixed parameter settings (as is the convention in standard Monte Carlo solutions), one can leverage machine learning tools and data simulated in the neighborhood of a parameter to improve estimates of quantities of interest. We demonstrate the efficacy of `ACORE` with both theoretical and empirical results. Our implementation is available on `Github`.

[1]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, USA [2]Department of Statistics, Federal University of São Carlos, São Paulo, Brazil. Correspondence to: Niccolò Dalmasso <ndalmass@stat.cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1. Introduction

Parameter estimation, statistical tests and confidence sets are the cornerstones of classical statistics that relate observed data to properties of the underlying statistical model. Most frequentist procedures with good statistical performance (e.g., high power) require explicit knowledge of a likelihood function. However, in many science and engineering applications, complex phenomena are modeled by forward simulators that *implicitly* define a likelihood function: For example, given input parameters θ, a statistical model of our environment, climate or universe may combine deterministic dynamics with random fluctuations to produce synthetic data $\mathbf{X}$. Simulation-based inference without an explicit likelihood is called *likelihood-free inference* (LFI).

The literature on LFI is vast. Traditional LFI methods, such as Approximate Bayesian Computation (ABC; Beaumont et al. 2002; Marin et al. 2012; Sisson et al. 2018), estimate posteriors by using simulations sufficiently close to the observed data, hence bypassing the likelihood. More recently, several approaches that leverage machine learning algorithms have been proposed; these either directly estimate the posterior distribution (Marin et al., 2016; Chen & Gutmann, 2019; Izbicki et al., 2019; Greenberg et al., 2019) or the likelihood function (Izbicki et al., 2014; Thomas et al., 2016; Price et al., 2018; Ong et al., 2018; Lueckmann et al., 2019; Papamakarios et al., 2019). We refer the reader to Cranmer et al. (2019) for a recent review of the field.

A question that has not received much attention so far is whether one, in an LFI setting, can construct inference techniques with good frequentist properties. Frequentist procedures have nevertheless played an important role in many fields. In high energy physics for instance, classical statistical techniques (e.g., hypothesis testing for outlier detection) have resulted in discoveries of new physics and other successful applications (Feldman & Cousins, 1998; Cranmer, 2015; Cousins, 2018). Even though controlling type I error probabilities is important in these applications, most LFI methods do not have guarantees on validity or power. Ideally, a unified LFI approach should

- be *computationally efficient* in terms of the number of required simulations,
- handle *high-dimensional data* from different sources (without, e.g., predefined summary statistics),

- produce hypothesis tests and confidence sets that are *valid*; that is, have the nominal type I error or confidence level,
- produce hypothesis tests with *high power* or, equivalently, confidence sets with a small expected size,
- provide *diagnostics* for checking empirical coverage or for checking how well the estimated likelihood fits simulated data.

In this paper, we present ACORE (Approximate Computation via Odds Ratio Estimation), a frequentist approach to LFI, which addresses the above mentioned concerns.

Figure 2 summarizes the ACORE work structure: ACORE first compares synthetic data from the simulator F_θ to a reference distribution G by computing an "Odds Ratio". The odds ratio can be learnt with a probabilistic classifier, such as a neural network with a softmax layer, suitable for the data at hand. As we shall see, the estimated odds ratio is an approximation of the likelihood ratio statistic (Proposition 3.1). The ACORE test statistic (Equation 4), together with an estimate of the "Critical Value" (Algorithms 1 and 2), can be used for hypothesis testing or for finding a confidence set for θ. ACORE also includes "Diagnostics" (Section 3.3) for computing the empirical coverage of the constructed confidence set for θ.

At the heart of ACORE is the key observation that the likelihood ratio statistic, the critical value of the test, and the coverage of the confidence set are conditional distribution functions which often vary smoothly as a function of the (unknown) parameters of interest. Hence, instead of relying solely on samples simulated at fixed parameter settings (as is the convention in standard Monte Carlo solutions), one can leverage machine learning tools and data simulated in the neighborhood of a parameter to improve estimates of quantities of interest and decrease the total number of simulated data points. Our contribution is three-fold:

1. a new procedure for estimating the *likelihood ratio statistic*, which uses probabilistic classifiers and does not require repeated sampling at each θ or a separate interpolation or calibration step;
2. an efficient procedure for estimating the *critical value* that guarantees valid tests and confidence sets, based on quantile regression without repeated sampling at each θ;
3. a new *goodness-of-fit technique* for computing empirical coverage of constructed confidence sets as a function of the unknown parameters θ.

Finally, ACORE is simple and modular by construction. One can easily switch out, generalize or take pieces of the framework and apply it to any similar machine-learning-based LFI setting. The theoretical results of Section 3.1 hold for a general setting. In addition, given the vast arsenal of existing probabilistic classifiers developed in the literature, ACORE can be applied to many different types of complex data $\mathbf{X}$ (e.g., images, time series and functional data). In Section 4, we show empirical results connecting the power of the constructed hypothesis tests to the performance of the classifier. Note also that Algorithms 1 and 2 for estimating parametrized critical values apply to any hypothesis test on θ of the form of Equation 2 for any test statistic τ. The goodness-of-fit procedure in Section 3.3 for checking empirical coverage as a function of θ is also not tied to odds ratios.

1.1. Related Work

The problem of constructing confidence intervals with good frequentist properties has a long history in statistics (Neyman, 1937; Feldman & Cousins, 1998; Chuang & Lai, 2000). One of the earlier simulation-based approaches was developed in high energy physics (HEP) by Diggle & Gratton (1984); their proposed scheme of estimating the likelihood and likelihoood ratio statistic nonparametrically by histograms of photon counts would later become a key component in the discovery of the Higgs Boson (Aad et al., 2012). However, traditional approaches for building confidence regions and hypothesis tests in LFI rely on a series of Monte Carlo samples at each parameter value θ (Barlow & Beeston, 1993; Weinzierl, 2000; Schafer & Stark, 2009). Thus, these approaches quickly become inefficient with large or continuous parameter spaces. Traditional nonparametric approaches also have difficulties handling high-dimensional data without losing key information.

LFI has recently benefited from using powerful machine learning tools like deep learning to estimate likelihood functions and likelihood ratios for complex data. Successful application areas include HEP (Guest et al., 2018), astronomy (Alsing et al., 2019) and neuroscience (Gonçalves et al., 2019). ACORE has some similarities to the work of Cranmer et al. (2015) which also uses machine learning methods for frequentist inference in an LFI setting. Other elements of ACORE such as leveraging the ability of ML algorithms to smooth over parameter space turning a density ratio estimate into a supervised classification problem have also previously been used in LFI settings: Works that smooth over parameter space include, e.g., Gaussian processes (Frate et al., 2017; Leclercq, 2018) and neural networks (Baldi et al., 2016). Works that turn a density ratio into a classification problem include applications to generative models (see Mohamed & Lakshminarayanan 2016 for a review), and Bayesian LFI (Thomas et al., 2016; Gutmann et al., 2018; Dinev & Gutmann, 2018; Hermans et al., 2019). Finally, like ACORE, Thornton et al. (2017) explores frequentist guarantees of confidence regions; however those regions are built under a Bayesian framework.

Novelty. What distinguishes the ACORE approach from other related work is that it uses an efficient procedure for estimating the (i) likelihood ratio, (ii) critical values and (iii) coverage of confidence sets across the entire parameter space, without the need for an extra interpolation or calibration step (as in traditional Monte Carlo solutions and more recent ML approaches). To the best of our knowledge, (ii) and (iii) are entirely novel in the LFI literature. In contrast to other methods that estimate (i), ACORE does not make parametric assumptions or require an additional calibration step, and it can accommodate all types of hypotheses. We provide theoretical guarantees on our procedures in terms of power and validity (Section 3.1; proofs in Supplementary Material C). We also offer a scheme for how to choose ML algorithms and the number of simulations so as to have good power properties and valid inference in practice.

Notation. Let F_θ with density f_θ represent the stochastic forward simulator for a sample point $\mathbf{X} \in \mathcal{X}$ at parameter $\theta \in \Theta$. We denote i.i.d "observable" data from F_θ by $\mathcal{D} = \left\{\mathbf{X}_1^{\text{obs}}, \ldots, \mathbf{X}_n^{\text{obs}}\right\}$, and the actually observed or measured data by $D = \left\{\mathbf{x}_1^{\text{obs}}, \ldots, \mathbf{x}_n^{\text{obs}}\right\}$. The likelihood function $\mathcal{L}(\mathcal{D}; \theta) = \prod_{i=1}^n f_\theta(\mathbf{X}_i^{\text{obs}})$.

2. Statistical Inference in a Traditional Setting

We begin by reviewing elements of traditional statistical inference that play a key role in ACORE.

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 (Neyman, 1937): Suppose that for each possible value $\theta_0 \in \Theta$, there is a level α test δ_{θ_0} of

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

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 $\mathcal{D} = D$, now define $R(D)$ as 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

$$\mathbb{P}\left[\theta_0 \in R(\mathcal{D}) \mid \theta = \theta_0\right] \geq 1 - \alpha$$

for all $\theta_0 \in \Theta$. That is, $R(\mathcal{D})$ defines a $(1-\alpha)$ *confidence set* for θ. Similarly, we can define a test with a desired significance level from 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

$$H_0 : \theta \in \Theta_0 \;\text{ versus }\; H_1 : \theta \in \Theta_1, \tag{2}$$

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

$$\Lambda(\mathcal{D}; \Theta_0) = \log \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\mathcal{D}; \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\mathcal{D}; \theta)}, \tag{3}$$

the LRT of hypotheses (2) rejects H_0 when $\Lambda(D; \Theta_0) < C$ for some constant C.

Figure 1 illustrates the construction of confidence sets for θ from level α likelihood ratio tests (1). The critical value for each such test δ_{θ_0} is $C_{\theta_0} = \{C : \mathbb{P}\left[\Lambda(\mathcal{D}; \theta_0) < C \mid \theta = \theta_0\right] = \alpha\}$.

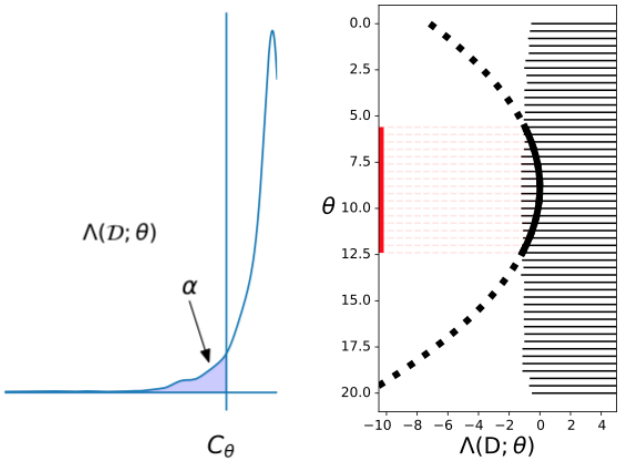

Figure 1. Constructing confidence intervals from hypothesis tests. *Left:* For each $\theta \in \Theta$, we find the critical value C_θ that rejects the null hypothesis $H_{0,\theta}$ at level α; that is, C_θ is the α-quantile of the distribution of the likelihood ratio statistic $\Lambda(\mathcal{D}; \theta)$ under the null. *Right:* The horizontal lines represent the acceptance region for each $\theta \in \Theta$. Suppose we observe data $\mathcal{D} = D$. The confidence set for θ (indicated with the red line) consists of all θ-values for which the observed test statistic $\Lambda(D; \theta)$ (indicated with the black curve) falls in the acceptance region.

3. ACORE: Approximate Computation via Odds Ratio Estimation

In a likelihood-free inference setting, we cannot directly evaluate the likelihood ratio statistic. Here we describe the details of how a simulation-based approach (ACORE, Figure 2) can lead to hypothesis tests and confidence sets with good frequentist properties.

3.1. Hypothesis Testing via Odds Ratios

We start by simulating a labeled sample for computing odds ratios. The estimated odds ratio then defines a new test statistic that we use in place of the unknown likelihood ratio statistic.

Simulating a labeled sample. Let G be a distribution with larger support than F_θ for all $\theta \in \Theta$. The distribution G

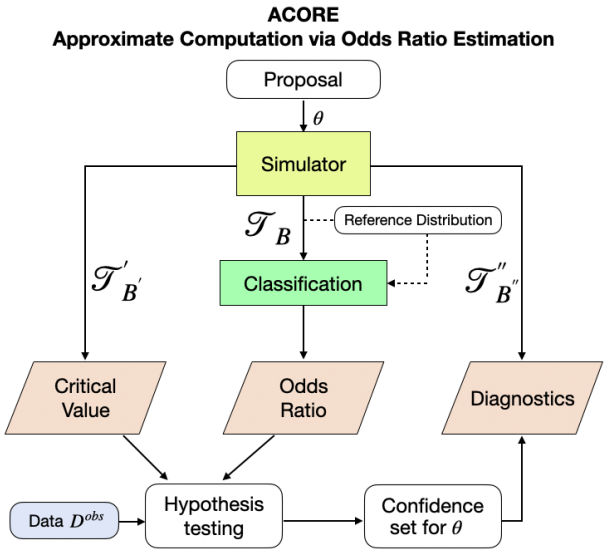

Figure 2. Schematic diagram of ACORE. The simulator provides synthetic observable data $\mathcal{T}_B$ for learning a parametrized odds ratio via probabilistic classification. The simulator also generates a separate sample $\mathcal{T}'_{B'}$ for learning critical values as a function of $\theta \in \Theta$. Once data $\mathcal{D}^{\text{obs}}$ are observed, the odds ratio can be used to construct hypothesis tests or confidence sets for θ. ACORE provides diagnostics for computing the empirical coverage of constructed confidence sets as a function of the (unknown) parameter θ. The three main parts of ACORE (critical value, odds ratio, diagnostics) are separate modules. Each module leverages machine learning methods in the training phase and is amortized, i.e., they perform inference on new data without having to be retrained.

could for example be a dominating distribution which it is easy to sample from. We use F_θ and G to simulate a labeled training sample $T_B = \{\theta_i, \mathbf{x}_i, y_i\}_{i=1}^B$ for estimating odds ratios. The random sample $\mathcal{T}_B = \{\theta_i, \mathbf{X}_i, Y_i\}_{i=1}^B$ is identically distributed as $(\theta, \mathbf{X}, Y)$, where the parameters $\theta \sim r_\Theta$ (a fixed proposal distribution over Θ), the "label" $Y \sim \text{Ber}(p)$ (a Bernoulli distribution with known p with Y independent of θ), $\mathbf{X}|\theta, Y = 1 \sim F_\theta$ and $\mathbf{X}|\theta, Y = 0 \sim G$. That is, the label Y_i is the indicator that the sample point $\mathbf{X}_i$ was generated from F_θ rather than G. We call G a "reference distribution" as we are comparing F_θ for different θ with this distribution. For all our experiments in this work we use p=1/2; other choices could account for computational differences in sampling from F_θ versus G. (Algorithm 3 in Supplementary Material A summarizes our procedure.)

Odds ratios. For fixed $\mathbf{x}$, we define the *odds* at θ as

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

and the *odds ratio* at $\theta_0, \theta_1 \in \Theta$ as

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

One way of interpreting the odds $\mathbb{O}(\theta, \mathbf{X})$ is to regard it as a measure of the chance that $\mathbf{X}$ was generated from F_θ. That is, a large odds $\mathbb{O}(\theta, \mathbf{x})$ reflects the fact that it is plausible that $\mathbf{x}$ was generated from F_θ (rather than G). Thus, $\mathbb{OR}(\mathbf{x}; \theta_0, \theta_1)$ measures the plausibility that $\mathbf{x}$ was generated from θ_0 rather than θ_1. When testing (2), we therefore reject H_0 if $\sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^n \log\left(\mathbb{OR}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right) < C$, for some constant C. By Bayes rule, this is just the likelihood ratio test of (2).

Hypothesis testing in an LFI setting. In an LFI setting, we cannot directly evaluate the likelihood ratio statistic (3). The advantage of rewriting the LRT in terms of odds ratios is that we can forward-simulate a labeled training sample $\mathcal{D}_B$, as described above, and then use a probabilistic classifier (suitable for the data at hand) to efficiently estimate the odds ratios $\mathbb{OR}(\mathbf{x}; \theta_0, \theta_1)$ for all $\theta_1, \theta_2 \in \Theta$: The probabilistic classifier compares data from the forward simulator F_θ with data from the reference distribution G and returns a parametrized odds estimate $\widehat{\mathbb{O}}(\mathbf{x}; \theta)$, which is a function of $\theta \in \Theta$. We can directly compute the odds ratio estimate $\widehat{\mathbb{OR}}(\mathbf{x}; \theta_0, \theta_1)$ at any two values $\theta_1, \theta_2 \in \Theta$ from $\widehat{\mathbb{O}}(\mathbf{x}; \theta)$. There is no need for a separate training step.

We reject H_0 if the ACORE test statistic defined as

$$\tau(\mathcal{D}; \Theta_0) := \sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^n \log\left(\widehat{\mathbb{OR}}(\mathbf{X}_i^{\text{obs}}; \theta_0, \theta_1)\right) \tag{4}$$

is small enough for observed data $\mathcal{D} = D$. If the probabilities learned by the classifier are well estimated, τ is exactly the likelihood ratio statistic:

Proposition 3.1 (Fisher Consistency)**.** *If $\widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{x}) = \mathbb{P}(Y = 1|\theta, \mathbf{x})$ for every θ and $\mathbf{x}$, then the ACORE test statistic (4) is the likelihood ratio statistic (Equation 3).*

Estimating the critical value. A key question is how to *efficiently* estimate the critical value of a test. In this section we consider a single composite null hypothesis $H_0 : \theta \in \Theta_0$. (The setting for constructing confidence sets by testing (1) for all $\theta_0 \in \Theta$ is discussed in Section 3.2.). Suppose that we reject the null hypothesis if the test statistic (4) is smaller than some constant C. To achieve a test with a desired level of significance α, we need (for maximum power) the largest C that satisfies

$$\sup_{\theta \in \Theta_0} \mathbb{P}\left(\tau(\mathcal{D}; \Theta_0) < C \mid \theta\right) \leq \alpha. \tag{5}$$

However, we cannot explicitly compute the critical value C or the rejection probability as we do not know the distribution of the test statistic τ.

Simulation-based approaches are often used to compute rejection probabilities and critical values in lieu of large-sample theory approximations. Typically, such simulations

compute a separate Monte Carlo simulation at each fixed $\theta \in \Theta_0$ on, e.g., a fine enough grid on θ. That is, the convention is to rely solely on sample points generated at fixed θ to estimate the rejection probabilities $\mathbb{P}(\tau(\mathcal{D}; \Theta_0) < C|\theta)$. Here we propose to estimate the critical values C for all $\theta \in \Theta_0$ and significance levels $\alpha \in [0, 1]$ simultaneously. At the heart of our approach is the key observation that the rejection probability $\mathbb{P}(\tau(\mathcal{D}; \Theta_0) < C|\theta)$ is a conditional cumulative distribution function, which in many settings varies smoothly as a function of θ and C. Thus, similar to how we estimate odds for the ACORE statistic, one can use data generated in the neighborhood of θ to improve estimates of our quantities of interest at any θ. This is what a quantile regression implicitly does to estimate C.

Algorithm 1 outlines the details of the procedure for estimating C. In brief, we use a training sample $\mathcal{T}'_{B'} = \{(\theta_i, \tau_i)\}_{i=1}^{B'}$ (independent of $\mathcal{T}_B$) to estimate the α-conditional quantile $c_\alpha(\theta)$ defined by $\mathbb{P}(\tau \leq c_\alpha(\theta) \mid \theta) = \alpha$. Let $\widehat{c}_\alpha(\theta)$ be the estimate of $c_\alpha(\theta)$ from a quantile regression of τ on θ. By (5), our estimate of the critical value C is $\widehat{C} = \inf_{\theta \in \Theta_0} \widehat{c}_\alpha(\theta)$. As we shall see, even if the odds are not well estimated, tests and confidence regions based on estimated odds are still valid as long as the thresholds are well estimated. Next we show that the sample size B' in Algorithm 1 controls the type I error (Theorem 3.3), whereas the training sample size B for estimating odds is related to the power of the test (Theorem 3.4).

Algorithm 1 Estimate the critical value C for a level-α test of composite hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$

Require: stochastic forward simulator F_θ; sample size B' for training quantile regression estimator; r_{Θ_0} (a fixed proposal distribution over the null region Θ_0); test statistic τ; quantile regression estimator; desired level $\alpha \in (0, 1)$

Ensure: estimated critical value $\widehat{C}$

Set $\mathcal{T}' \leftarrow \emptyset$
for i in {1,...,B'} **do**
  Draw parameter $\theta_i \sim r_{\Theta_0}$
  Draw sample $\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n} \overset{iid}{\sim} F_{\theta_i}$
  Compute test statistic $\tau_i \leftarrow \tau((\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}); \Theta_0)$
  $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(\theta_i, \tau_i)\}$
end for
Use $\mathcal{T}'$ to learn parametrized function $\widehat{c}_\alpha(\theta) := \widehat{F}^{-1}_{\tau|\theta}(\alpha|\theta)$ via quantile regression of τ on θ
return $\widehat{C} \leftarrow \inf_{\theta \in \Theta_0} \widehat{c}_\alpha(\theta)$

Theoretical guarantees. We denote convergence in probability and in distribution by $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\text{Dist}}$, respectively. We start by showing that our procedure leads to valid hypothesis tests (that is, tests that control the type I error probability) as long as B' in Algorithm 1 is large enough. In order to do so, we assume that the quantile regression estimator used in Algorithm 1 to estimate the critical values is consistent in the following sense:

Assumption 3.2. *Let $\hat{F}_{B'}(\cdot|\theta)$ be the estimated cumulative distribution function of the test statistic τ conditional on θ based on a sample size B', and let $F(\cdot|\theta)$ be true conditional distribution. For every $\theta \in \Theta_0$, assume that the quantile regression estimator is such that*

$$\sup_{t \in \mathbb{R}} |\hat{F}_{B'}(t|\theta) - F(t|\theta)| \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} 0$$

Under some conditions, Assumption 3.2 holds for instance for quantile regression forests (Meinshausen, 2006).

Next we show that, for every fixed training sample size B in Algorithm 3, Algorithm 1 yields a valid hypothesis test as $B' \to \infty$. The result holds even if the likelihood ratio statistic is not well estimated.

Theorem 3.3. *Let $C_{B,B'} \in \mathbb{R}$ be the critical value of the test based on the statistic $\tau = \tau_B$ for a training sample size B with critical value chosen according to Algorithm 1 for a fixed $\alpha \in (0, 1)$. If the quantile estimator satisfies Assumption 3.2 and $|\Theta| < \infty$, then*

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

where C^*_B is such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}(\tau_B \leq C^*_B|\theta) = \alpha.$$

Finally we show that as long as the probabilistic classifier is consistent and the critical values are well estimated (which holds for large B' according to Theorem 3.3), the power of the ACORE test converges to the power of the LRT as B grows.

Theorem 3.4. *Let $\widehat{\phi}_{B,C_B}(\mathcal{D})$ be the test based on the statistic $\tau = \tau_B$ for a labeled sample size B with critical value $C_B \in \mathbb{R}$.[1] Moreover, let $\phi_{C^*}(\mathcal{D})$ be the likelihood ratio test with critical value $C^* \in \mathbb{R}$.[2] If, for every $\theta \in \Theta$,*

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

where $|\Theta| < \infty$, and $\widehat{C}_B$ is such that $\widehat{C}_B \xrightarrow[B \longrightarrow \infty]{\text{Dist}} C^*$, then, for every $\theta \in \Theta$,

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

3.2. Confidence Sets

To construct a confidence set for θ, we use the equivalence of tests and confidence sets (Section 2): Suppose that we

[1] That is, $\widehat{\phi}_{B,C_B}(\mathcal{D}) = 1 \iff \tau_B(\mathcal{D}; \Theta_0) < C_B$.
[2] That is, $\phi_{C^*}(\mathcal{D}) = 1 \iff \Lambda(\mathcal{D}; \Theta_0) < C^*$.

for every $\theta_0 \in \Theta$ can find the critical value C_{θ_0} of a test of (1) with type I error no larger than α. The random set

$$R(\mathcal{D}) = \{\theta_0 \in \Theta \,|\tau(\mathcal{D}; \theta_0) \geq C_{\theta_0}\},$$

then defines a $(1-\alpha)$ confidence region for θ.

However, rather than repeatedly running Algorithm 1 for each null hypothesis $\Theta_0 = \{\theta_0\}$ separately, we estimate all critical values C_{θ_0} (for different $\theta_0 \in \Theta$) simultaneously. Algorithm 2 outlines our procedure. Again, we use quantile regression to learn a parametrized function C_{θ_0}. The whole procedure for computing confidence sets via ACORE is summarized in Algorithm 4 in Supplementary Material B. Theorem 3.3 implies that the constructed confidence set has the nominal $1-\alpha$ confidence level as $B' \to \infty$. The size of the confidence set depends on the training sample size B and the classifier.

Algorithm 2 [Many Simple Null Hypotheses] Estimate the 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

Require: stochastic forward simulator F_θ; sample size B' for training quantile regression estimator; r (a fixed proposal distribution over the full parameter space Θ); test statistic τ; quantile regression estimator; desired level $\alpha \in (0,1)$

Ensure: estimated critical values $\widehat{C}_\theta$ for all $\theta = \theta_0 \in \Theta$

1: Set $\mathcal{D}' \leftarrow \emptyset$
2: **for** i in {1,...,B'} **do**
3: Draw parameter $\theta_i \sim r$
4: Draw sample $\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n} \overset{iid}{\sim} F_{\theta_i}$
5: Compute test statistic $\tau_i \leftarrow \tau((\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}); \theta_i)$
6: $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\theta_i, \tau_i)\}$
7: **end for**
8: Use $\mathcal{D}'$ to learn parametrized function $\widehat{C}_\theta := \widehat{F}^{-1}_{\tau|\theta}(\alpha|\theta)$ via quantile regression of τ on θ **return** $\widehat{C}_{\theta_0} \leftarrow \widehat{F}^{-1}_{\tau|\theta_0}(\alpha|\theta_0)$

3.3. Evaluating Empirical Coverage for All Possible Values of θ

After the parametrized ACORE statistic and the critical values have been estimated, it is important to check whether the resulting confidence sets indeed are valid or, equivalently, if the resulting hypothesis tests have the nominal significance level. We also want to identify regions in parameter space where we clearly overcover. That is, the two main questions are: (i) do the constructed confidence sets satisfy

$$\mathbb{P}\left[\theta_0 \in R(\mathcal{D}) \mid \theta = \theta_0\right] \geq 1-\alpha,$$

for every $\theta_0 \in \Theta$, and (ii) how close is the actual coverage to the nominal confidence level $1-\alpha$? To answer these questions, we propose a goodness-of-fit procedure where we draw B'' new samples from the simulator given θ, construct a confidence set for each sample, and then check which computed regions include the "true" θ. More specifically: we generate a set $\mathcal{T}''_{B''} = \{(\theta'_1, \mathcal{D}'_1), \ldots, (\theta'_{B''}, \mathcal{D}'_{B''})\}$, where $\theta'_i \sim r_\Theta$ and $\mathcal{D}'_i$ is a sample of size n of i.i.d. observable data from $F_{\theta'_i}$. We then define

$$W_i := \mathbb{I}\left(\theta'_i \in R(\mathcal{D}'_i)\right),$$

where $R(\mathcal{D}'_i)$ is the confidence set for θ for data $\mathcal{D}'_i$. If R has the correct coverage, then

$$\mathbb{P}(W_i = 1 | \theta_i) \geq 1-\alpha.$$

We can estimate the probability $\mathbb{P}(W_i = 1|\theta_i)$ using any probabilistic classifier; some methods also provide confidence bands that assess the uncertainty in estimating this quantity (Eubank & Speckman, 1993; Claeskens et al., 2003; Krivobokova et al., 2010). By comparing the estimated probability to $1-\alpha$, we have a diagnostic tool for checking how close we are to the nominal confidence level over the entire parameter space Θ. See Figure 3 for an example.

Finally note that our procedure parametrizes the coverage of the confidence set as a function of the true parameter value. This is in contrast to other goodness-of-fit techniques (e.g., Cook et al. 2006; Bordoloi et al. 2010; Talts et al. 2018; Schmidt et al. 2019) that only check for *marginal* coverage, i.e., $n^{-1}\sum_{i=1}^n W_i \geq 1-\alpha$.

4. Toy Examples

We consider two examples where the true likelihood is known. In the first example, the forward simulator F_θ follows a Poisson$(100+\theta)$ distribution similar to the signal-background model in Section 5. In the second example, we consider a Gaussian mixture model (GMM) with two unit-variance Gaussians centered at $-\theta$ and θ, respectively. In both examples, $n = 10$, the proposal distribution r_Θ is a uniform distribution, and the reference distribution G is a normal distribution. Table 1 summarizes the set-up.

	Poisson Example	*GMM Example*
r_Θ	Unif$(0, 20)$	Unif$(0, 10)$
F_θ	Poisson$(100+\theta)$	$\frac{1}{2}\mathcal{N}(-\theta, 1) + \frac{1}{2}\mathcal{N}(\theta, 1)$
G	$\mathcal{N}(110, 15^2)$	$\mathcal{N}(0, 5^2)$
True θ	$\theta_0 = 10$	$\theta_0 = 5$

Table 1. Set-up for the two toy examples.

First we investigate how the power of ACORE and the size of the derived confidence sets depend on the performance of the classifier used in the odds ratio estimation (Section 3.1). We consider three classifiers: multilayer perceptron (MLP), nearest neighbor (NN) and quadratic discriminant analysis

Poisson Example

B	*Classifier*	*Cross Entropy Loss*	*Average Power*	*Size of Confidence Set [%]*
100	MLP	0.87 ± 0.27	0.24	75.9 ± 19.3
	NN	0.76 ± 0.15	0.29	71.6 ± 19.7
	QDA	$\mathbf{0.66 \pm 0.02}$	**0.41**	$\mathbf{60.0 \pm 15.6}$
500	MLP	0.69 ± 0.01	0.35	$\mathbf{65.9 \pm 20.4}$
	NN	0.67 ± 0.01	0.38	$\mathbf{62.9 \pm 15.8}$
	QDA	$\mathbf{0.64 \pm 0.01}$	**0.47**	$\mathbf{54.2 \pm 9.4}$
1,000	MLP	0.69 ± 0.01	0.37	$\mathbf{63.3 \pm 19.8}$
	NN	0.66 ± 0.01	0.44	$\mathbf{56.9 \pm 15.9}$
	QDA	$\mathbf{0.64 \pm 0.01}$	**0.50**	$\mathbf{51.3 \pm 7.7}$
-	**Exact**	$\mathbf{0.64 \pm 0.01}$	**0.54**	$\mathbf{45.0 \pm 4.9}$

GMM Example

B	*Classifier*	*Cross Entropy Loss*	*Average Power*	*Size of Confidence Set [%]*
100	MLP	$\mathbf{0.39 \pm 0.03}$	**0.88**	$\mathbf{14.1 \pm 4.7}$
	NN	0.81 ± 0.31	0.42	58.4 ± 23.3
	QDA	0.64 ± 0.02	0.15	85.3 ± 21.1
500	MLP	$\mathbf{0.35 \pm 0.01}$	**0.90**	$\mathbf{12.1 \pm 2.4}$
	NN	0.45 ± 0.05	0.57	44.3 ± 24.1
	QDA	0.62 ± 0.01	0.15	84.9 ± 19.9
1,000	MLP	$\mathbf{0.35 \pm 0.01}$	**0.90**	$\mathbf{12.1 \pm 2.5}$
	NN	0.41 ± 0.02	0.77	$\mathbf{24.9 \pm 15.9}$
	QDA	0.62 ± 0.01	0.12	88.1 ± 18.0
-	**Exact**	$\mathbf{0.35 \pm 0.01}$	**0.92**	$\mathbf{9.5 \pm 2.0}$

Table 2. Results for Poisson example (left) and GMM example (right). The tables show the cross entropy loss, power (averaged over θ) and size of ACORE confidence sets for different values of B and for different classifiers. These results are based on 100 repetitions; the numbers represent the mean and one standard deviation. The best results in each setting are marked in bold-faced; we see that the classifier with the lowest cross entropy loss (a quantity that is easily computed in practice) is linked with the highest average power and the smallest confidence set. As B increases, the best ACORE values approach the values for the exact LRT, listed in the bottom row in red color. (The QDA for the GMM example does not improve with increasing B because the quadratic classifier cannot separate F_θ and G in a mixed distribution with three modes, hence breaking the assumption of Theorem 3.4.) All nine probabilistic classifiers yield valid 90% confidence regions according to our diagnostics; see Table 3.

(QDA). For different values of B (sample size for estimating odds ratios), we compute the binary cross entropy (a measure of classifier performance), the power as a function of θ, and the size of the constructed confidence set. Table 2 summarizes results based on 100 repetitions. (To compute the critical values in Algorithm 2, we use quantile gradient boosted trees and a large enough sample size $B' = 5000$ to guarantee 90% confidence sets; see Supplementary Material D.) The last row of the table shows the best attainable cross entropy loss (Supplementary Material F), the confidence set size and power for the true likelihood function. For all 18 settings, the computation of one ACORE confidence set takes between 10 to 30 seconds on a single CPU.[3] A full breakdown of the runtime of ACORE confidence sets can be found in Supplementary Material I.

For each setting with fixed B, the best classifier according to cross entropy loss achieves the highest power and the smallest confidence set.[4] Moreover, as B increases, the best values (marked in bold-faced) get closer to those of the true likelihood (marked in red). The cross-entropy loss is easy to compute in practice. Our results indicate that minimizing the cross-entropy loss is a good rule of thumb for achieving ACORE inference results with desirable statistical properties.

Next we illustrate our goodness-of-fit procedure (Section 3.3) for checking the coverage of the constructed confidence sets across the parameter space Θ. To pass our goodness-of-fit test, we require the nominal coverage to be within two standard deviations of the estimated coverage for all parameter values. Figure 3 shows the estimated coverage with logistic regression for the Poisson example with $B = 1000$ (the training sample size for estimating odds via QDA) and three different values of B' (the training sample size for estimating the critical value C via gradient boosted quantile regression). As expected (Theorem 3.3), the estimated coverage gets closer to the nominal 90% confidence level as B' increases. We can use these diagnostic plots to choose B'. For instance, here $B' = 500$ is large enough for ACORE to achieve good coverage. (See Supplementary Material D for a detailed analysis of this example.)

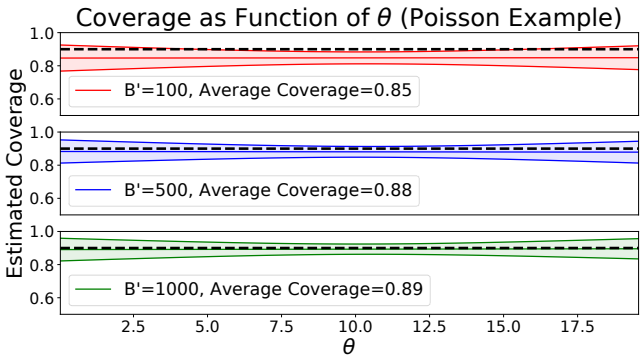

Figure 3. Estimated coverage as a function of θ in the Poisson example for ACORE with different values of B'. The mean and one standard deviation prediction intervals are estimated via logistic regression. Our diagnostics show that $B' = 500$ is large enough to achieve the nominal confidence level $1 - \alpha = 0.9$. (We here use $n = 10$, a QDA classifier with $B = 1000$ and gradient boosted quantile regression).

[3] More specifically, an 8-core Intel Xeon 3.33GHz X5680 CPU.

[4] In traditional settings, high power has been shown to lead to a small expected interval size under certain distributional assumptions (Pratt, 1961; Ghosh, 1961).

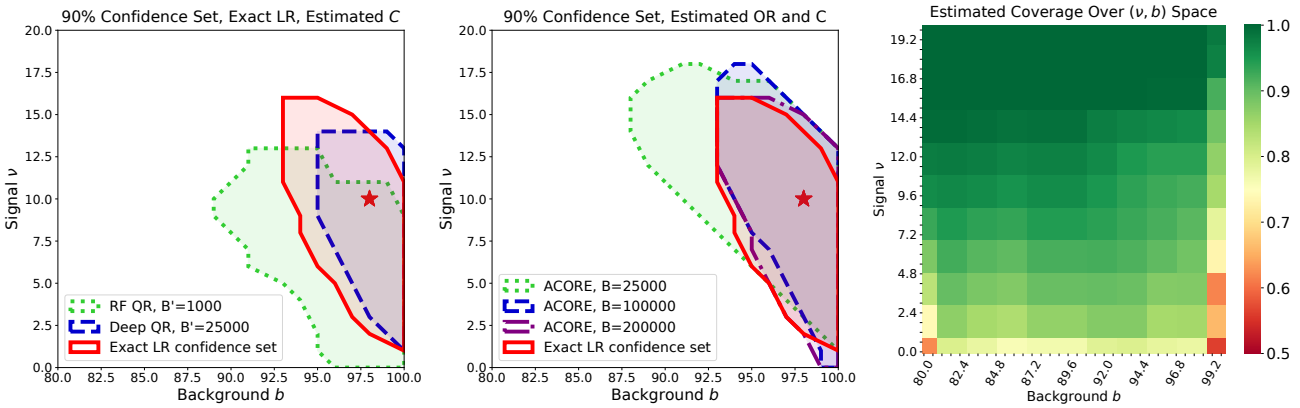

Figure 4. Signal detection HEP example. *Left*: 90% confidence sets computed with the exact likelihood ratio statistic. Estimating critical values can however be challenging, as highlighted by the differences in the results for two different quantile regression (QR) algorithms and sample sizes: Random Forest QR at $B' = 1000$ (green dotted) versus Deep QR at $B' = 25000$ (blue dashed). Our goodness-of-fit procedure can be used to select the best method in a principled way. (The red contour shows the exact LR confidence set, and the red star is at the true parameter setting.) *Center*: 90% confidence sets when using ACORE to estimate both odds ratios and critical values. This is the LFI setting. Our proposed strategy for choosing ACORE components selects a 5-layer deep neural network with $B = 100000$; this yields a confidence set (dashed blue) close to the exact LR set (solid red). Increasing B does not show a noticeable improvement (dash-dotted purple), whereas decreasing B makes estimates worse (dotted green). *Right*: Heat map of the estimated coverage for a confidence set that did *not* pass our goodness-of-fit diagnostic. The overall coverage of the confidence set is correct (91.8% vs. the 90% nominal confidence level), but the set clearly undercovers in low-signal and high-background regions.

Supplementary Materials G and H include a comparison between ACORE and Monte Carlo Gaussian Process (MC GP) interpolation (Frate et al., 2017) and calibrated neural nets classifiers (CARL, Cranmer et al. 2015), respectively. Our results show that MC-based GP interpolation provides a better approximation of the likelihood ratio when the simulated data are approximately Gaussian (as in the Poisson example). However, when the parametric assumptions are not valid (as in the GMM example), MC-based GP fails to approximate the likelihood ratio regardless of the number of available simulations. For both examples, CARL leads to lower power and larger confidence intervals than ACORE. See Tables 4 and 5 for details.

5. Signal Detection in High Energy Physics

In order to apply ACORE, we need to choose four key components: (i) a probabilistic classifier, (ii) a training sample size B for learning odds ratios, (iii) a quantile regression algorithm, and (iv) a training sample size B' for estimating critical values. We propose the following practical strategy to choose such components:

1. Use the cross entropy loss to select the classifier and B (as seen in Section 4, a small cross entropy corresponds to higher power and a smaller confidence set);
2. Then use our goodness-of-fit procedure (Section 3.3) to select the quantile regression method and B'.

We illustrate ACORE and this strategy on a model described in Rolke et al. (2005) and Sen et al. (2009) for a high energy physics (HEP) experiment. In this model, particle collision events are counted under the presence of a background process b. The goal is to assess the intensity ν of a signal (i.e., an event which is not part of the background process). The observed data D consist of $n = 10$ realizations of $\mathbf{X} = (N, M)$, where $N \sim \text{Poisson}(b+\nu)$ is the number of events in the signal region, and $M \sim \text{Poisson}(b)$ is the number of events in the background (control) region. (We use a uniform proposal distribution r_Θ and a Gaussian reference distribution G.) This model is a simplified version of a real particle physics experiment where the true likelihood function is not known.

Figure 4 illustrates the role of B, B', and our goodness-of-fit procedure when estimating confidence sets. (For details, see Supplementary Material E.) In the *left* panel, we use the true LR statistic to show that, even if the LR is available, estimating the critical value C well still matters. Our goodness-of-fit diagnostic provides a principled way of choosing the best quantile regression (QR) method and the best sample size B' for estimating C. In this example, random forest QR does not pass our goodness-of-fit test; it also leads to a confidence region quite different from the exact one. Deep QR, which passes our test, gives a more accurate region estimate. In the *center* panel, we use ACORE to estimate both the odds ratio and the critical value C (this is the LFI setting). If we choose B by identifying when the cross entropy loss levels off, we would choose $B = 100000$. Decreasing B leads to a worse cross-entropy loss and, as the

figure shows, also a larger confidence region. Increasing B beyond our selected sample size does not lead to substantial gains. The *right* panel illustrates how our goodness-of-fit procedure can be used to identify regions in parameter space where a constructed confidence set is not valid. The heat map refers to an example which did *not* pass our goodness-of-fit procedure. While the overall (marginal) coverage is at the right value, our diagnostic procedure (for estimating coverage as a *function* of ν and b) is able to identify under-coverage in low-signal and high-background regions. That is, for a valid confidence set, one needs to better estimate the critical value C by, e.g., using a different quantile regression estimator or by increasing B' (either uniformly over the parameter space or by an active learning scheme which increases the number of simulations at parameter settings where one undercovers).

6. Conclusions

In this paper we introduce ACORE, a framework for carrying out frequentist inference in LFI settings. ACORE is well suited for settings with costly simulations, as it efficiently estimates test statistics and critical values across the entire parameter space. We provide a new goodness-of-fit procedure for estimating coverage of constructed confidence sets for all possible parameter settings. Even if the likelihood ratio is not well estimated, ACORE provides valid inference as long as hypothesis tests and confidence sets pass our goodness-of-fit procedure (albeit at the cost of having less power and larger sets). We provide practical guidance on how to choose the smallest number of simulations to guarantee powerful and valid procedures.

Future studies will investigate the effect of G and r_Θ on performance, as well as how ACORE scales with increasing (a) feature space dimension and (b) parameter space dimension. Because we utilize ML methods to efficiently estimate odds ratio and critical values (Algorithms 3 and 1), performance in (a) will depend on the convergence rates of the chosen probabilistic classifier and quantile regression method. For (b), scaling relies on having an efficient search algorithm; this search is challenging for all likelihood-based methods. Common solutions include gradient-free optimization methods, such as Nelder-Mead (Nelder & Mead, 1965) and Bayesian optimization (Snoek et al., 2012), and approximation techniques, such as profile likelihoods (Murphy & Vaart, 2000) and hybrid resampling (Chuang & Lai, 2000; Sen et al., 2009). Such approaches can potentially be integrated into ACORE. The ACORE framework can also be adapted to accommodate test statistics such as the Bayes factor (Kass & Raftery, 1995). In addition to Bayes factors, we will investigate choosing the number of simulations B via sequential testing and likelihood goodness-of-fit tests such as Dalmasso et al. (2020). In addition, we will consider extending the ACORE framework to include other statistical quantities in the likelihood ratio estimation process by, for example, adapting the regression on likelihood ratio (ROLR) score in Brehmer et al. (2020b). Finally, we will include a theoretical study of how the power of the ACORE test relates to classifier performance.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments and suggestions. ND is grateful to Tudor Manole, Alan Mishler, Aleksandr Podkopaev and the STAMPS research group for insightful discussions. RI is grateful for the financial support of FAPESP (2019/11321-9) and CNPq (306943/2017-4). The early stages of this research were supported in part by the National Science Foundation under DMS-1520786.

References

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A., Abdinov, O., Aben, R., Abi, B., Abolins, M., and et al. 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):1–29, Sep 2012. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.08.020. URL http://dx.doi.org/10.1016/j.physletb.2012.08.020.

Alsing, J., Charnock, T., Feeney, S., and Wand elt, B. Fast likelihood-free cosmology with neural density estimators and active learning. *Monthly Notice of the Royal Astronomical Society*, 488(3):4440–4458, Sep 2019. doi: 10.1093/mnras/stz1960.

Baldi, P., Cranmer, K., Faucett, T., Sadowski, P., and Whiteson, D. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76 (5):235, Apr 2016. ISSN 1434-6052. doi: 10.1140/epjc/s10052-016-4099-4. URL https://doi.org/10.1140/epjc/s10052-016-4099-4.

Barlow, R. and Beeston, C. Fitting using finite monte carlo samples. *Computer Physics Communications*, 77(2):219 – 228, 1993. ISSN 0010-4655. doi: https://doi.org/10.1016/0010-4655(93)90005-W. URL http://www.sciencedirect.com/science/article/pii/001046559390005W.

Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Bordoloi, R., Lilly, S. J., and Amara, A. Photo-z performance for precision cosmology. *Monthly Notices of the Royal Astronomical Society*, 406(2):881–895, 2010.

Brehmer, J., Kling, F., Espejo, I., and Cranmer, K. MadMiner, 2019. URL https://github.com/diana-hep/madminer.

Brehmer, J., Kling, F., Espejo, I., and Cranmer, K. MadMiner: Machine learning-based inference for particle physics. *Comput. Softw. Big Sci.*, 4(1):3, 2020a. doi: 10.1007/s41781-020-0035-2.

Brehmer, J., Louppe, G., Pavez, J., and Cranmer, K. Mining gold from implicit models to improve likelihood-free inference. *Proc. Nat. Acad. Sci.*, 117(10):5242–5249, 2020b. doi: 10.1073/pnas.1915980117.

Chen, Y. and Gutmann, M. U. Adaptive gaussian copula ABC. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1584–1592. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/chen19d.html.

Chuang, C.-S. and Lai, T. L. Hybrid resampling methods for confidence intervals. *Statistica Sinica*, 10(1):1–33, 2000. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306697.

Claeskens, G., Van Keilegom, I., et al. Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics*, 31(6):1852–1884, 2003.

Cook, S. R., Gelman, A., and Rubin, D. B. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15 (3):675–692, 2006. doi: 10.1198/106186006X136976.

Cousins, R. D. Lectures on Statistics in Theory: Prelude to Statistics in Practice. *arXiv e-prints*, art. arXiv:1807.05996, Jul 2018.

Cranmer, K. Practical Statistics for the LHC. *arXiv e-prints*, art. arXiv:1503.07622, Mar 2015.

Cranmer, K., Pavez, J., and Louppe, G. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *arXiv e-prints*, art. arXiv:1911.01429, Nov 2019.

Dalmasso, N., Lee, A., Izbicki, R., Pospisil, T., Kim, I., and Lin, C.-A. Validation of approximate likelihood and emulator models for computationally intensive simulations. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3349–3361, Online, 26–28 Aug 2020. PMLR. URL http://proceedings.mlr.press/v108/dalmasso20a.html.

Diggle, P. J. and Gratton, R. J. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46 (2):193–227, 1984. ISSN 00359246. URL http://www.jstor.org/stable/2345504.

Dinev, T. and Gutmann, M. U. Dynamic Likelihood-free Inference via Ratio Estimation (DIRE). *arXiv e-prints*, art. arXiv:1810.09899, Oct 2018.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. Wiley, New York, 2 edition, 2001. ISBN 978-0-471-05669-0.

Eubank, R. L. and Speckman, P. L. Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301, 1993.

Feldman, G. J. and Cousins, R. D. Unified approach to the classical statistical analysis of small signals. *Physical Review D*, 57(7):3873–3889, Apr 1998. ISSN 1089-4918. doi: 10.1103/physrevd.57.3873. URL http://dx.doi.org/10.1103/PhysRevD.57.3873.

Frate, M., Cranmer, K., Kalia, S., Vand enberg-Rodes, A., and Whiteson, D. Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes. *arXiv e-prints*, art. arXiv:1709.05681, Sep 2017.

Ghosh, J. K. On the relation among shortest confidence intervals of different types. *Calcutta Statistical Association Bulletin*, 10(4):147–152, 1961. doi: 10.1177/0008068319610404.

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., Greenberg, D. S., and Macke, J. H. Training deep neural density estimators to identify mechanistic models of neural dynamics. *bioRxiv*, 2019. doi: 10.1101/838383.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2404–2414, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/greenberg19a.html.

Guest, D., Cranmer, K., and Whiteson, D. Deep Learning and Its Application to LHC Physics. *Annual Review of Nuclear and Particle Science*, 68(1):161–181, Oct 2018. doi: 10.1146/annurev-nucl-101917-021019.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, Mar 2018. ISSN 1573-1375. doi: 10.1007/s11222-017-9738-6. URL https://doi.org/10.1007/s11222-017-9738-6.

Hermans, J., Begy, V., and Louppe, G. Likelihood-free mcmc with amortized approximate likelihood ratios, 2019.

Izbicki, R., Lee, A., and Schafer, C. High-dimensional density ratio estimation with extensions to approximate likelihood computation. In *Artificial Intelligence and Statistics*, pp. 420–429, 2014.

Izbicki, R., Lee, A. B., and Pospisil, T. ABC–CDE: Toward approximate bayesian computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, pp. 1–20, 2019. doi: 10.1080/10618600.2018.1546594.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572.

Krivobokova, T., Kneib, T., and Claeskens, G. Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, 105(490):852–863, 2010.

Leclercq, F. Bayesian optimization for likelihood-free cosmological inference. *Physical Review D*, 98(6):063511, Sep 2018. doi: 10.1103/PhysRevD.98.063511.

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 32–53, 2019.

MacKay, D. J. C. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002. ISBN 0521642981.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

Marin, J.-M., Raynal, L., Pudlo, P., Ribatet, M., and Robert, C. ABC random forests for bayesian parameter inference. *Bioinformatics (Oxford, England)*, 35, 05 2016. doi: 10.1093/bioinformatics/bty867.

Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Mohamed, S. and Lakshminarayanan, B. Learning in Implicit Generative Models. *arXiv e-prints*, art. arXiv:1610.03483, Oct 2016.

Murphy, S. A. and Vaart, A. W. V. D. On profile likelihood. *Journal of the American Statistical Association*, 95(450): 449–465, 2000. doi: 10.1080/01621459.2000.10474219.

Nelder, J. A. and Mead, R. A Simplex Method for Function Minimization. *The Computer Journal*, 7 (4):308–313, 01 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308. URL https://doi.org/10.1093/comjnl/7.4.308.

Neyman, J. 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(767):333–380, 1937. ISSN 00804614. URL http://www.jstor.org/stable/91337.

Ong, V. M. H., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. Variational bayes with synthetic likelihood. *Statistics and Computing*, 28(4):971–988, Jul 2018. ISSN 1573-1375. doi: 10.1007/s11222-017-9773-3.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 837–848, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pratt, J. W. Length of confidence intervals. *Journal of the American Statistical Association*, 56(295):549–567, 1961. doi: 10.1080/01621459.1961.10480644.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018. doi: 10.1080/10618600.2017.1302882.

Rolke, W. A., López, A. M., and Conrad, J. Limits and confidence intervals in the presence of nuisance parameters.

Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 551(2):493 – 503, 2005. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2005.05.068.

Schafer, C. M. and Stark, P. B. Constructing confidence regions of optimal expected size. *Journal of the American Statistical Association*, 104(487):1080–1089, 2009. doi: 10.1198/jasa.2009.tm07420.

Schmidt, S., Malz, A., and et al. Evaluation of probabilistic photometric redshift estimation approaches for LSST. *(In Preparation) Under internal review by LSST-DESC*, 2019.

Sen, B., Walker, M., and Woodroofe, M. On the unified method with nuisance parameters. *Statistica Sinica*, 19 (1):301–314, 2009. ISSN 10170405, 19968507.

Sisson, S. A., Fan, Y., and Beaumont, M. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, 2018.

Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 2951–2959. Curran Associates, Inc., 2012.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Thomas, O., Dutta, R., Corander, J., Kaski, S., and Gutmann, M. U. Likelihood-free inference by ratio estimation. *arXiv e-prints*, art. arXiv:1611.10242, Nov 2016.

Thornton, S., Li, W., and ge Xie, M. An effective likelihood-free approximate computing method with statistical inferential guarantees, 2017.

Weinzierl, S. Introduction to monte carlo methods, 2000.