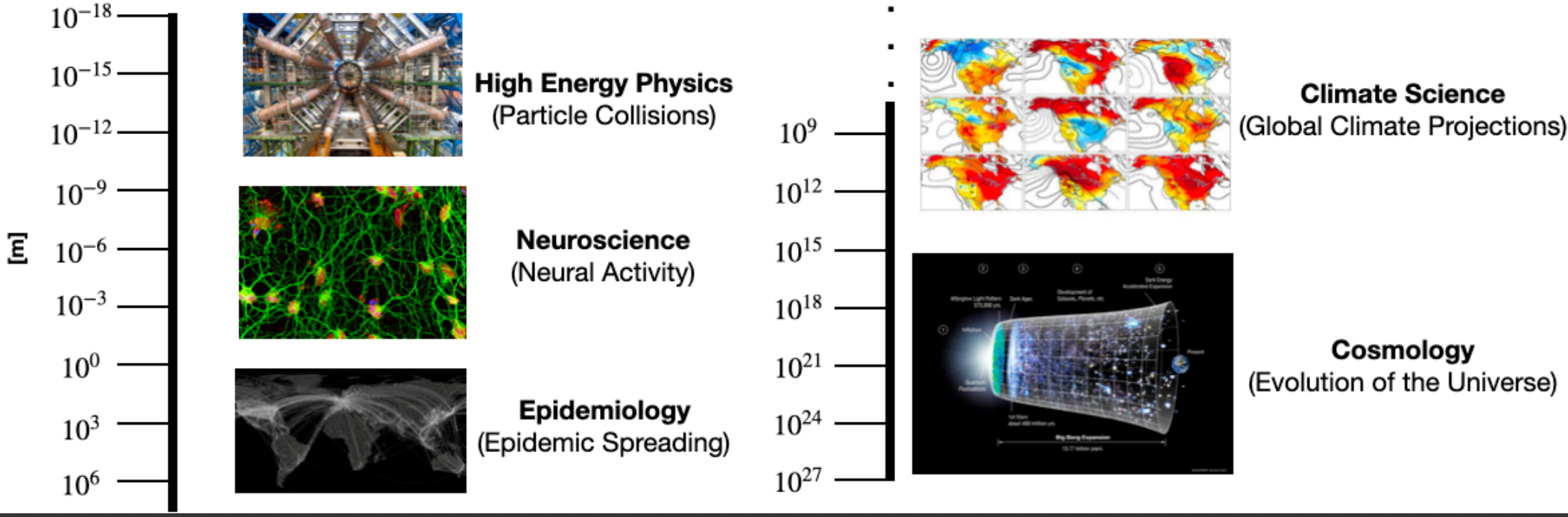
Likelihood-Free Frequentist Inference

Ann B. Lee
Department of Statistics & Data Science / MLD
Carnegie Mellon University

Collaborators: Nic Dalmasso (JP Morgan); Tommaso Dorigo (INFN/Padova); Rafael Izbicki (UFSCar), Mikael Kuusela (CMU), Luca Masserano (CMU), and David Zhao (CMU)

Simulators are Ubiquitous in Science



Credit: Dalmasso (adapted from Cranmer et al, 2020)

- For many complex phenomena, the only meaningful model (theory) may be in the form of simulations.

Taxonomy of Different Types of Simulators

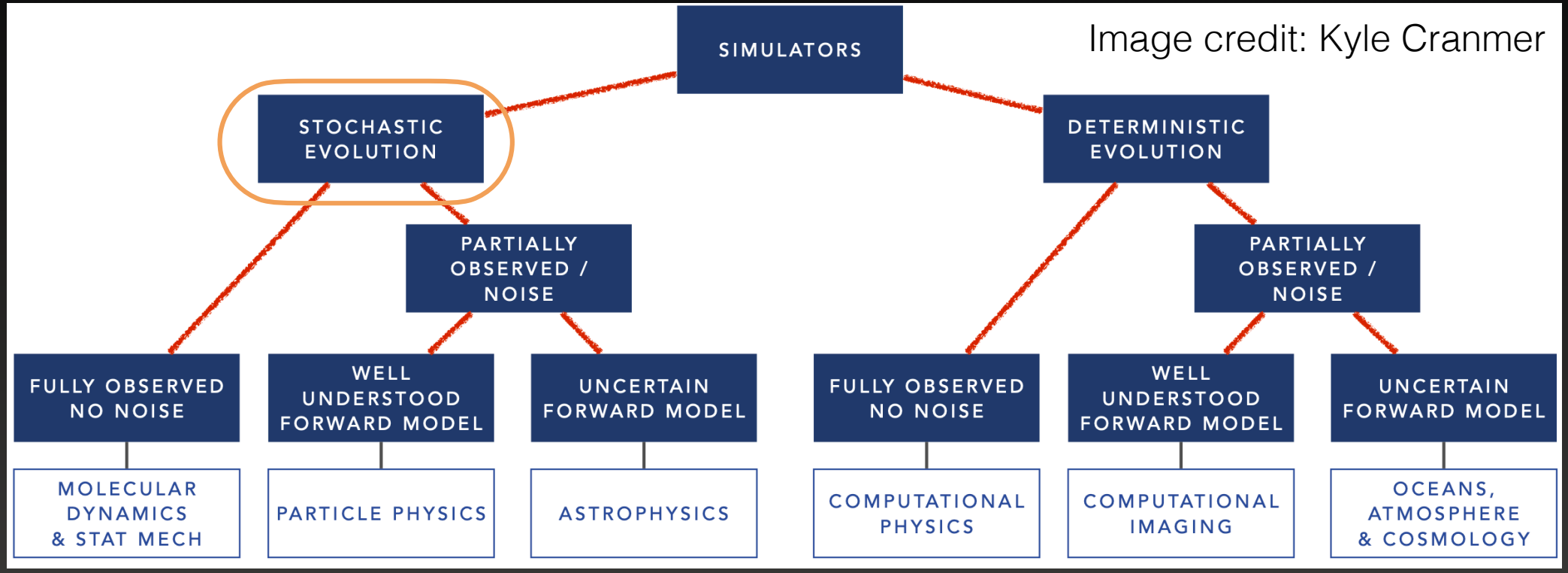


Image credit: Kyle Cranmer

- These simulators may be good at simulating observable data — but often poorly suited for the inverse problem of inferring the underlying scientific mechanisms associated with observed real-world phenomena.

Statistical Challenges for Complex Models

- Forward problem: Does data from the approximate model have the same distribution as high-fidelity (simulated or observed) data?
  - Ask if two distributions are different, and if so, how they differ in high dimensions (capture dependencies between all variables)?

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

- Inverse problem: Suppose we have a forward model F_θ that implicitly encodes the relationship between parameter θ of interest (input) and high-dimensional observable data $\mathbf{X}$ (output).
  - Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, can we infer the true parameters θ with valid measures of uncertainty (confidence sets)?

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$

Statistical Methods for Comparing Distributions of High-Dimensional Data

Electronic Journal of Statistics
Vol. 13 (2019) 5253–5305
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1648

Global and local two-sample tests via regression

Ilmun Kim, Ann B. Lee, and Jing Lei

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

Validation of Approximate Likelihood and Emulator Models for Computationally Intensive Simulations

Niccolò Dalmasso,[1] Ann B. Lee,[1] Rafael Izbicki,[2] Taylor Pospisil,[3] Ilmun Kim,[1] Chieh-An Lin[4]

https://arxiv.org/abs/1905.11505 (AISTATS 2020)

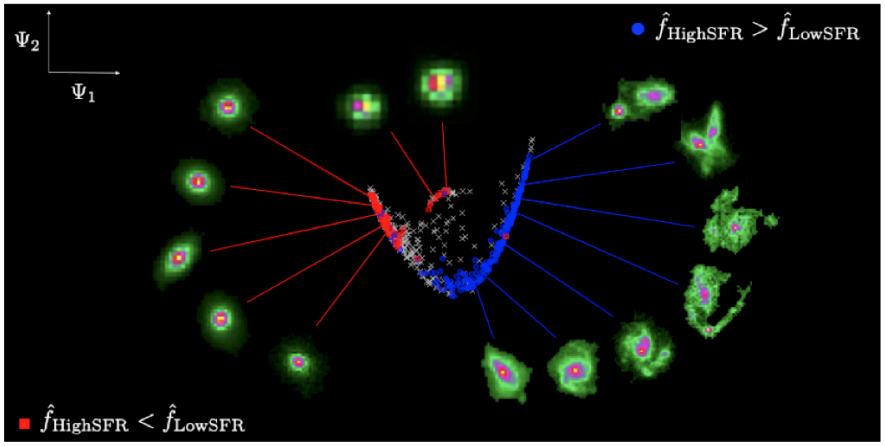


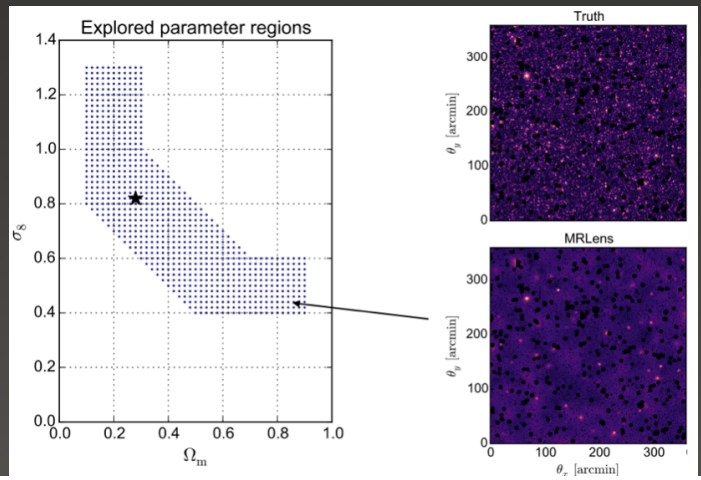
Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional m bl di

$$\text{Test} \quad H_0 : \widehat{\mathcal{L}}(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta$$
$$\text{versus} \quad H_1 : \widehat{\mathcal{L}}(\mathbf{x};\theta) \neq \mathcal{L}(\mathbf{x};\theta) \text{ for some } \mathbf{x} \in \mathcal{X} \text{ and } \theta \in \Theta$$

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS 471, 3273–3282 (2017) doi:10.1093/mnras/stx1807
Advance Access publication 2017 July 18

Local two-sample testing: a new tool for analysing high-dimensional astronomical data

P. E. Freeman,★ I. Kim and A. B. Lee
Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA



Statistical Challenges for Complex Models

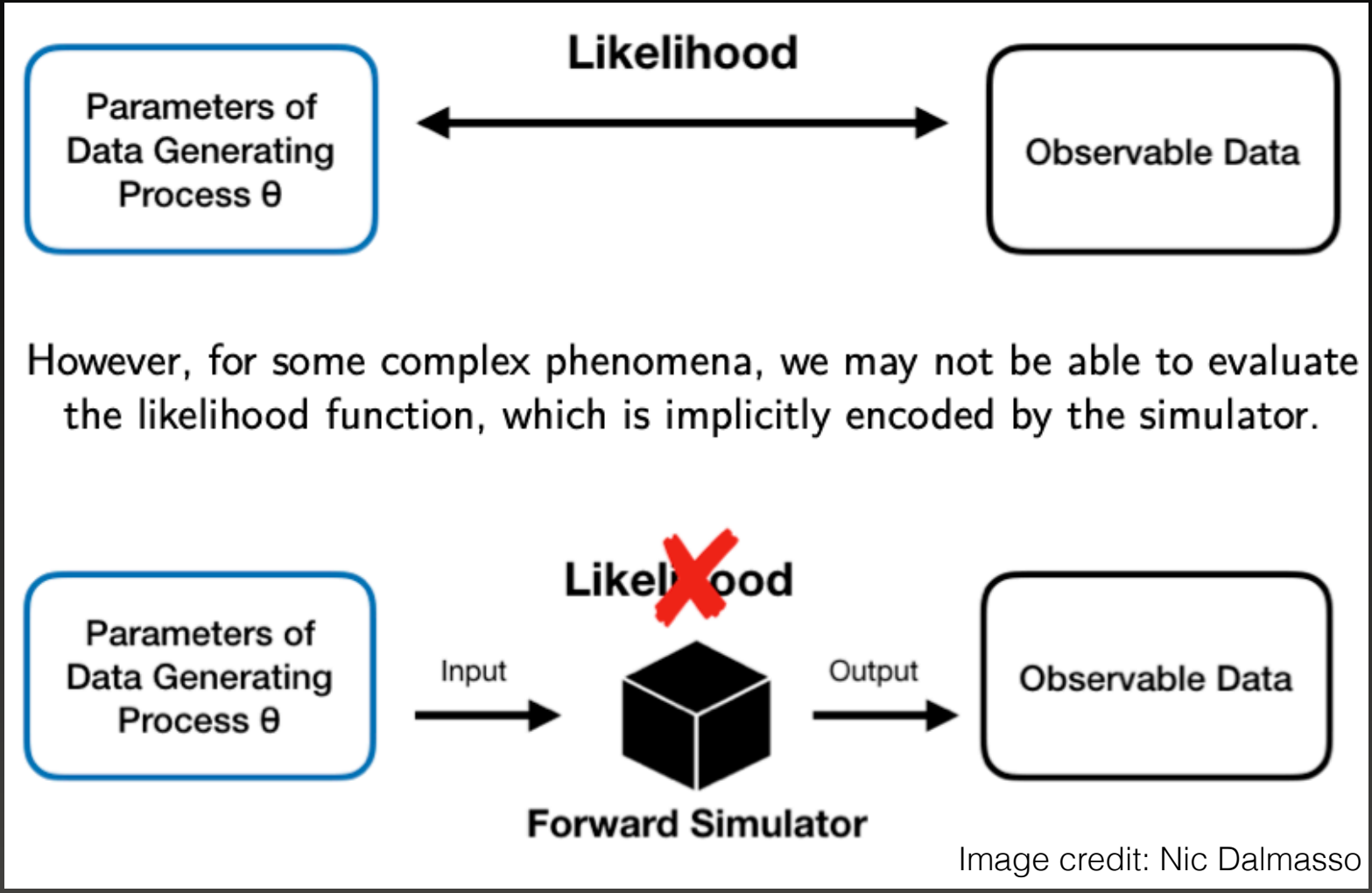
- Forward problem: Does data from the approximate model have the same distribution as high-fidelity (simulated or observed) data?
  - Ask if two distributions are different, and if so, how they differ in high dimensions (capture dependencies between all variables)?

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \sim F \quad \text{and} \quad \mathbf{X}_1^*, \ldots, \mathbf{X}_n^* \sim F^*$$

- Inverse problem: Suppose we have a forward model F_θ that implicitly encodes the relationship between parameter θ of interest (input) and high-dimensional observable data $\mathbf{X}$ (output).
  - Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, can we infer the true parameters θ with valid measures of uncertainty (confidence sets)?

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$

What is Likelihood-Free Inference?



- Inference on parameters in the second setting is called likelihood-free inference (LFI).

Classical LFI: Approximate Bayesian Computation (ABC)

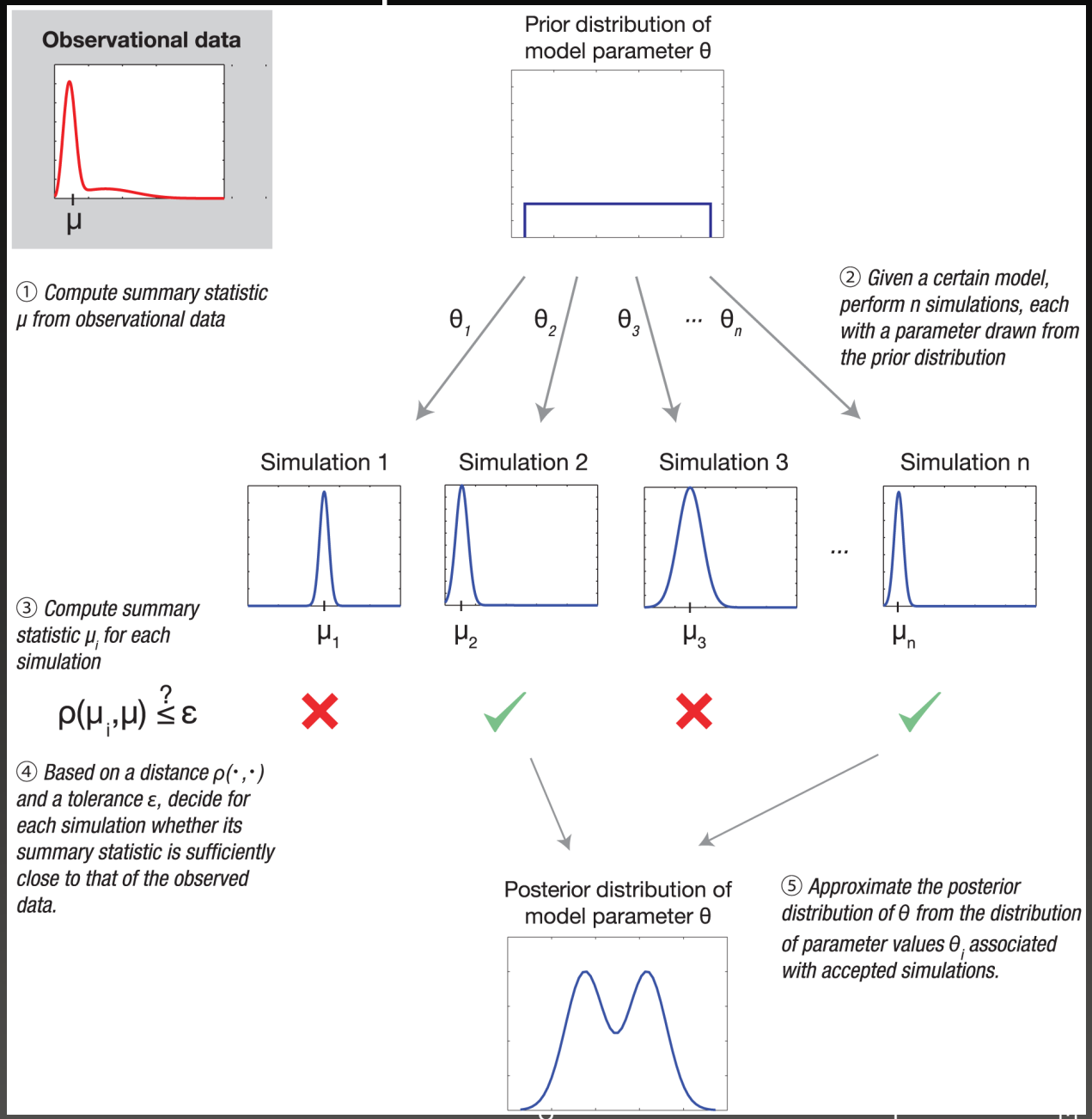


Image credit: Sunnaker et al. 2013

Changing LFI Landscape

- More recent developments use ML algorithms to directly estimate key inferential quantities from simulated data

$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim \pi(\theta), \mathbf{X} \sim F_\theta$$

  - Posteriors, $f(\boldsymbol{\theta}|x)$ [e.g., Papamakarios et al, 2016; Lueckmann et al, 2016; Izbicki et al, 2019; Greenberg et al, 2019]
  - Likelihoods, $f(x|\boldsymbol{\theta})$ or $f(x|\boldsymbol{\theta})/g(x)$ [e.g., Izbicki et al, 2014; Thomas et al, 2016; Durkan et al, 2020; Brehmer et al., 2020]
  - Likelihood ratios, $f(x|\boldsymbol{\theta}_1)/f(x|\boldsymbol{\theta}_2)$ [e.g, Cranmer et al, 2015; Thomas et al, 2016; Hermans et al, 2020; Durkan et al, 2020; Brehmer et al, 2020]

- These new training-based approaches provide "amortized" inference. Can handle complex high-dimensional data without relying on summary statistics.

Changing LFI Landscape

- More recent developments use ML algorithms to directly estimate key inferential quantities from simulated data

$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim \pi(\theta), \mathbf{X} \sim F_\theta$$

  - Posteriors, f(**θ**|x) [e.g., Papamakarios et al, 2016; Lueckmann et al, 2016; Izbicki et al, 2019; Greenberg et al, 2019]
  - Likelihoods, f(x|**θ**) or f(x|**θ**)/g(x) [e.g., Izbicki et al, 2014; Thomas et al, 2016; Durkan et al, 2020; Brehmer et al., 2020]
  - Likelihood ratios, f(x|$\boldsymbol{\theta}_1$)/f(x|$\boldsymbol{\theta}_2$) [e.g, Cranmer et al, 2015; Thomas et al, 2016; Hermans et al, 2020; Durkan et al, 2020; Brehmer et al, 2020]

- These new training-based approaches provide "amortized" inference. Can handle complex high-dimensional data without a prior dimension reduction.

What Might be Missing in the LFI/ML Literature?

- The construction of frequentist hypothesis tests and confidence sets with correct coverage has a long history in statistics (Fisher 1925; Neyman 1935), with the equivalence between valid tests and confidence sets formalized by Neyman.

- Classical statistical analysis has also played a key role in HEP and for the discovery of new physics. A few examples…

  - Feldman/Cousins, "*Unified approach to the classical statistical analysis of small signals*", Phys. Rev. D, 1998

  - Cowan/Cranmer/Gross/Vitells, "*Asymptotic formulae for likelihood-based tests of new physics*", Eur.Phys.J.C. 2011

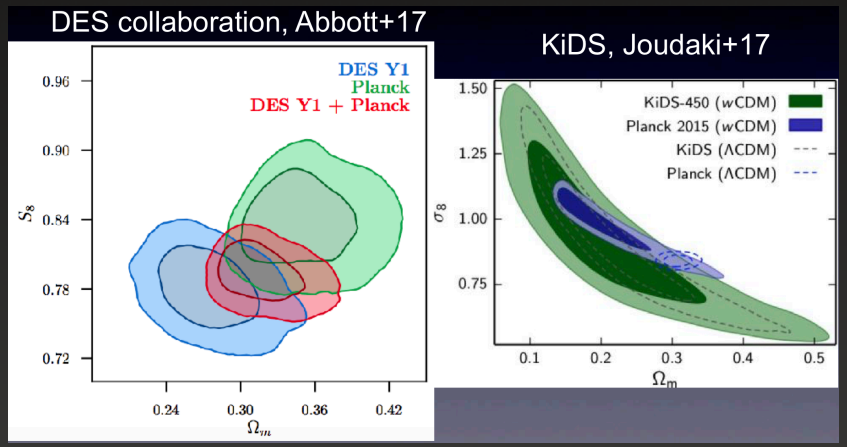
  - The ATLAS and CMS Collaborations and the LHC Higgs Combination Group, "*Procedure for the LHC Higgs boson search combination in Summer 2011*", 2011 CMS/ATL Tech Report.

What Might be Missing in the LFI-ML Literature?

- Nevertheless: Inferential tools with finite-sample guarantees of freq. coverage have not received much attention in the recent LFI-ML Literature.

- Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, want to infer the true parameters θ with valid measures of uncertainty.

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$



$$\{(\theta_1, \mathbf{X}_1), (\theta_2, \mathbf{X}_2), \ldots, (\theta_B, \mathbf{X}_B)\}, \text{ where } \theta \sim \pi(\theta),\ \mathbf{X} \sim F_\theta$$

Predictive Approach Can Be Very Powerful, But One Needs to Correct for Bias

[New project with Luca Masserano, Dr. Tommaso Dorigo, Dr. Mikael Kuusela]

Data coming from Dorigo et al. (2020): ~ 400'000 **simulated muons** with true incoming energy sampled uniformly between 100 and 2000 GeV.

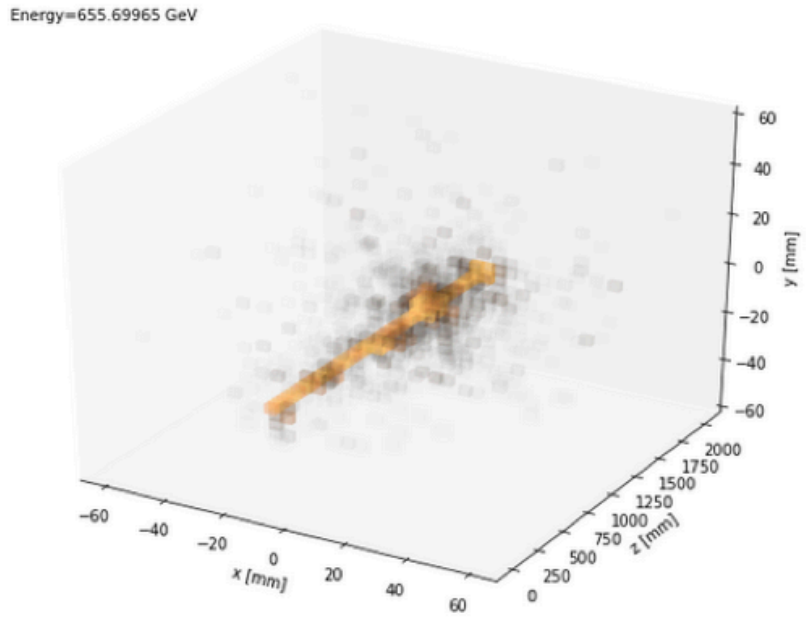


Figure 4: Muon entering the calorimeter in z direction.

[Kieseler et al., July 2021 arXiv:2107.02119]

1. Bias

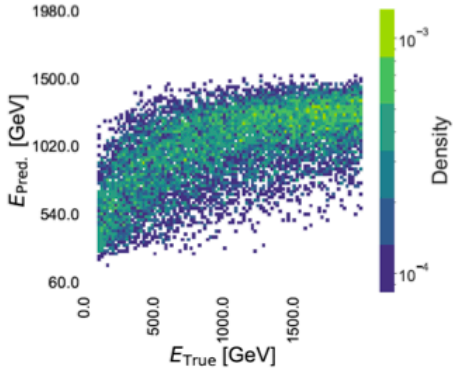


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

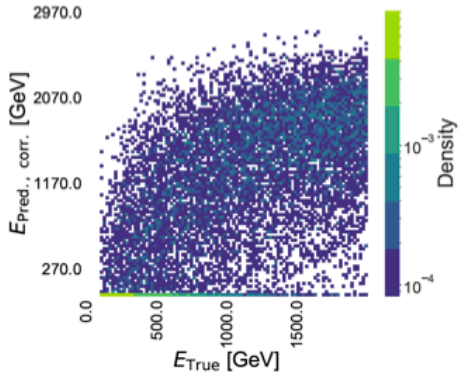


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

$$\mathbb{E}[\theta|X] \neq \theta^{\star}$$

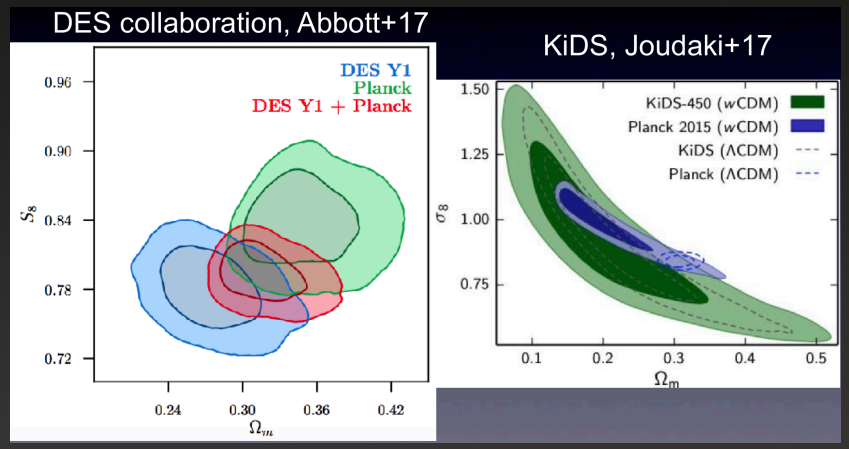
Source: Dorigo et al 2020.
Slide credit: Luca Masserano

How About Frequentist LFI Approaches?

- Statistical tests and confidence sets are the hallmarks of scientific inference. However: Practical inferential tools with finite-sample guarantees of freq. coverage have not received much attention.

  - Given observed data $D=\{\mathbf{X}_1,.. \mathbf{X}_n\}$, want to infer the true parameters θ with valid measures of uncertainty.

$$\mathbb{P}[\theta \in R(\mathcal{D})] \geq 1 - \alpha$$



- Most approaches that estimate likelihoods or likelihood ratios
  - rely on asymptotic assumptions (Wilks 1938) for downstream inference
  - do not assess quality of inference across entire parameter space

Unified Inference Machinery for Frequentist LFI

- Bridges ML with classical statistics to provide:

  (i) valid inference: confidence sets and hypothesis tests with finite-sample guarantees (Type I error control and power)

  (ii) practical diagnostics: check actual coverage across entire parameter space

- Goal: Modular procedures with theoretical guarantees.
  - Can accommodate different types of high-dimensional data
  - Compatible with any test statistic (including LR statistics; but more generally also statistics based on ML/forward predictions)

Preliminary work presented at ICML 2021 by Nic Dalmasso

https://arxiv.org/abs/2002.10399

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

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 procedure 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

More recent preprint (July 2021)

https://arxiv.org/abs/2107.03920

Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning in Simulation and Uncertainty Quantification

Niccolò Dalmasso *† NICCOLO.DALMASSO@GMAIL.COM

David Zhao *† DAVIDZHAO@STAT.CMU.EDU

Rafael Izbicki ‡ RAFAELIZBICKI@GMAIL.COM

Ann B. Lee † ANNLEE@STAT.CMU.EDU

Abstract

Many areas of science make extensive use of computer simulators that implicitly encode likelihood functions of complex systems. Classical statistical methods are poorly suited for these so-called likelihood-free inference (LFI) settings, outside the asymptotic and low-dimensional regimes. Although new machine learning methods, such as normalizing flows, have revolutionized the sample efficiency and capacity of LFI methods, it remains an open question whether they produce reliable measures of uncertainty.

In this paper, we present a statistical framework for LFI that unifies classical statistics with modern machine learning to: (1) construct frequentist confidence sets and hypothesis tests with finite-sample guarantees of nominal coverage (type I error control) and power, and (2) provide rigorous diagnostics for assessing empirical coverage over the entire parameter space. We refer to our framework as *likelihood-free frequentist inference* (LF2I). Any method that estimates a test statistic, such as the likelihood ratio, can be plugged into our

Equivalence of Tests and Confidence Sets

Key ingredients:

- data $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$
- a test statistic, such as the likelihood ratio statistic $\mathrm{LR}(\mathcal{D}; \theta_0)$
- an α-level critical value $C_{\theta_0,\alpha}$

Reject the null hypothesis H_0 if $\mathrm{LR}(\mathcal{D}; \theta_0) < C_{\theta_0,\alpha}$

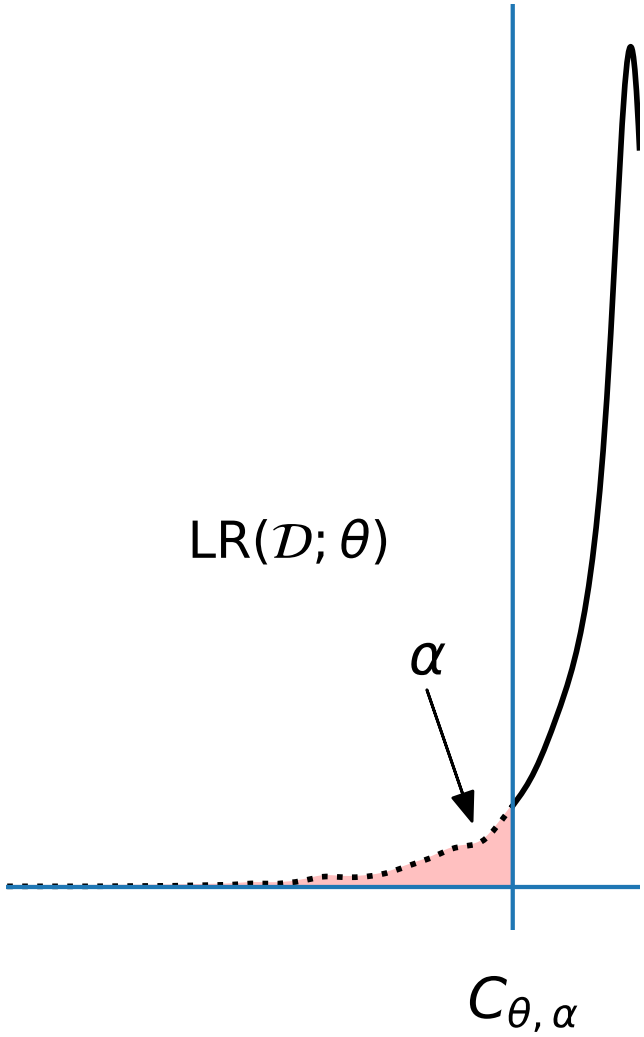
Theorem (Neyman 1937)

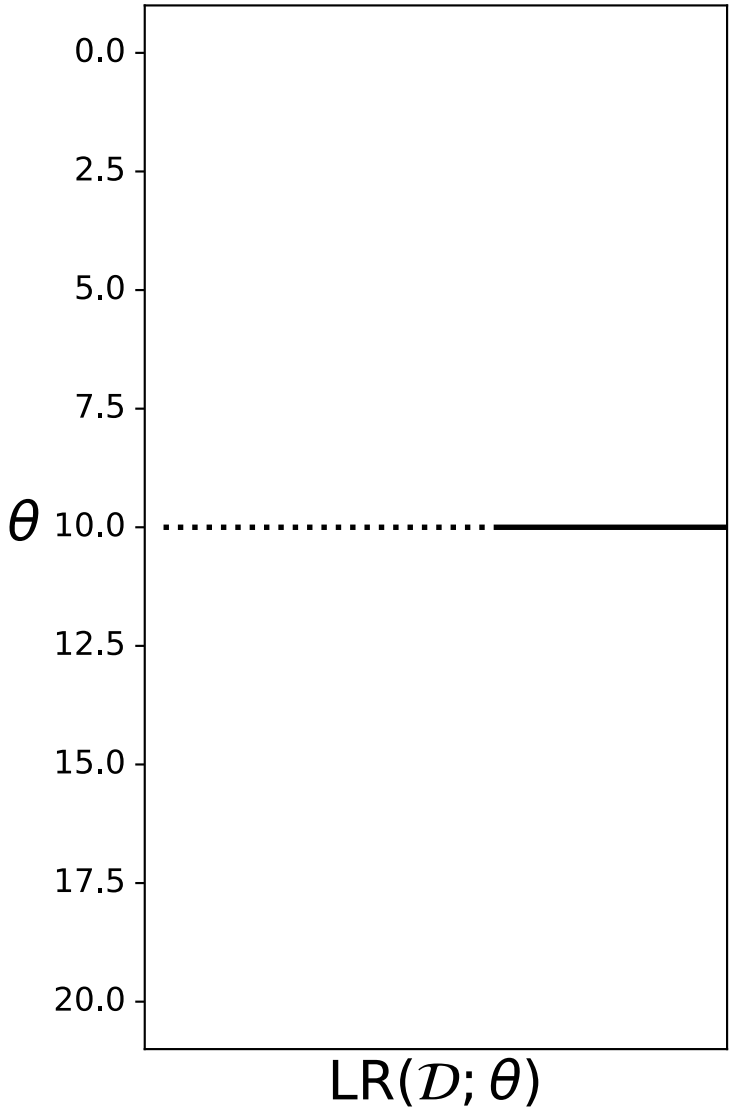
Constructing a $1-\alpha$ confidence set for θ is equivalent to testing

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

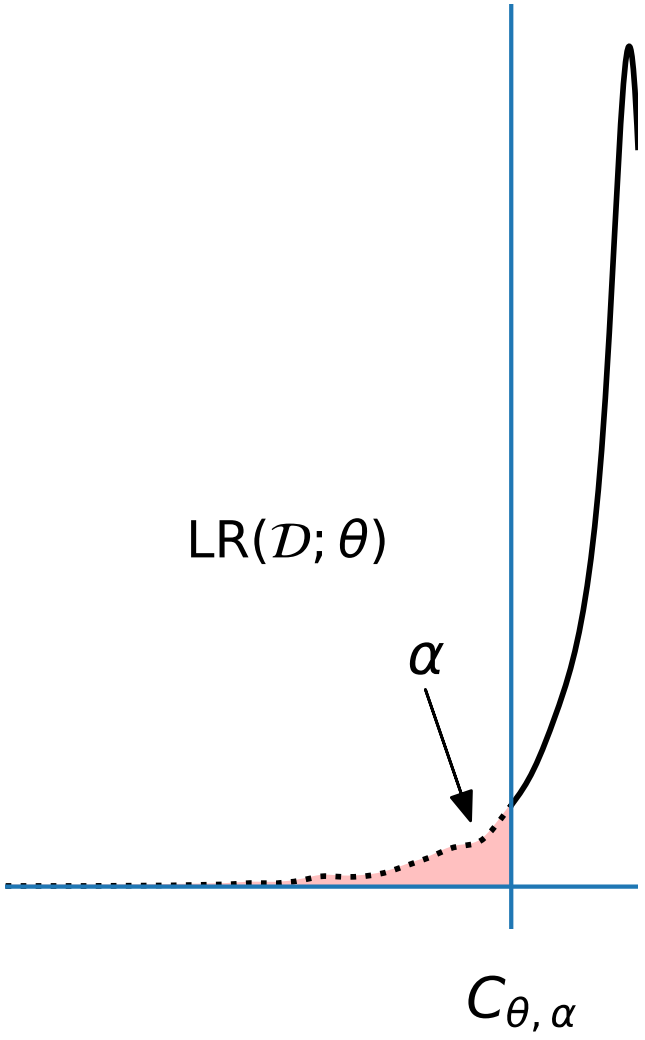
for every θ_0 in the parameter space.

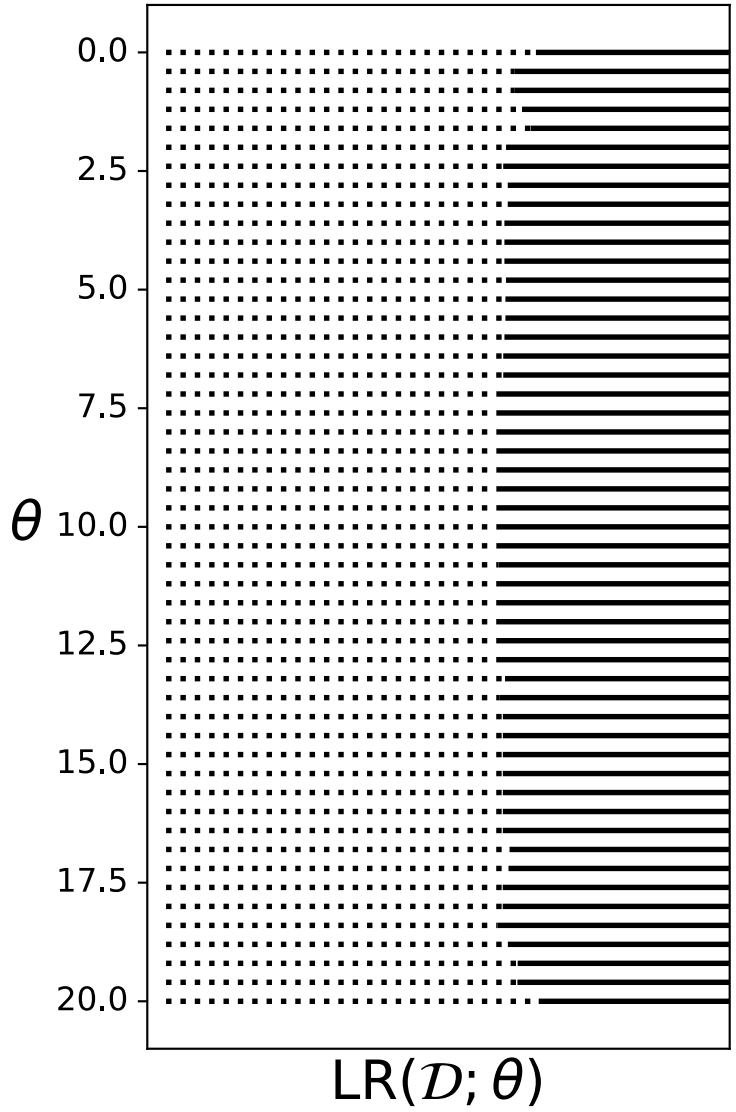
1. Fixed θ. Find the rejection region for test statistic Λ.



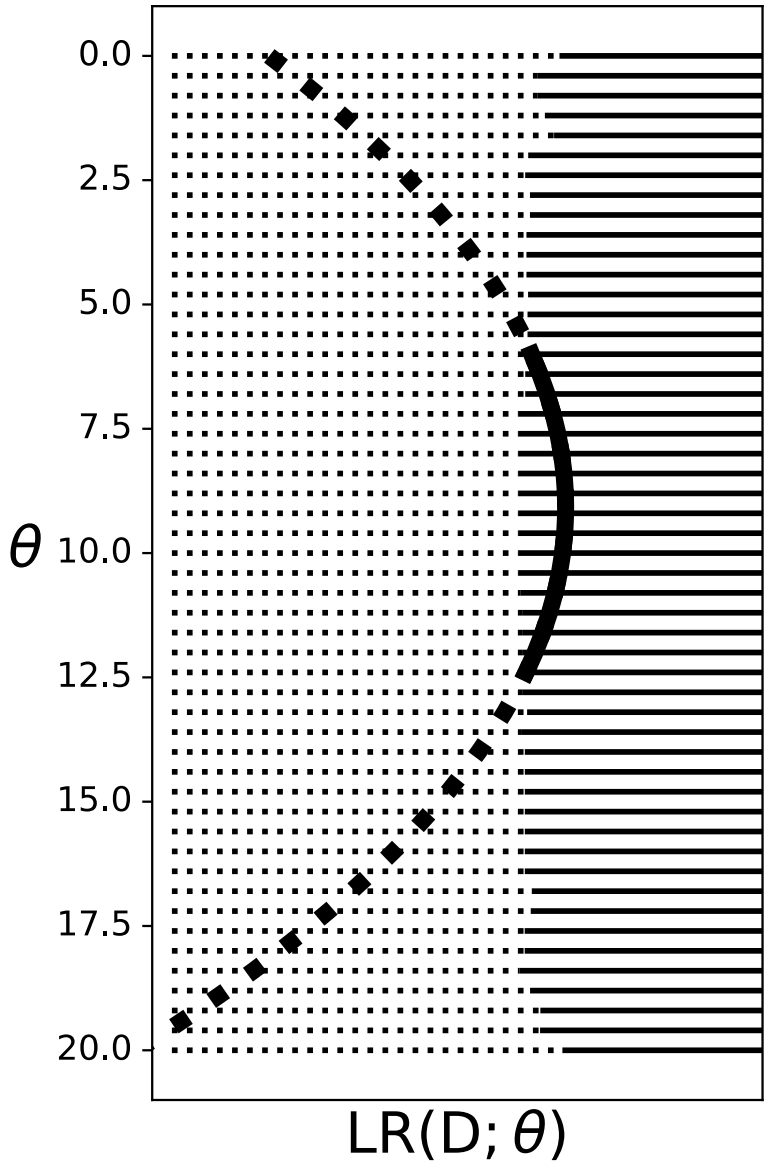


2. Repeat for every θ in parameter space.

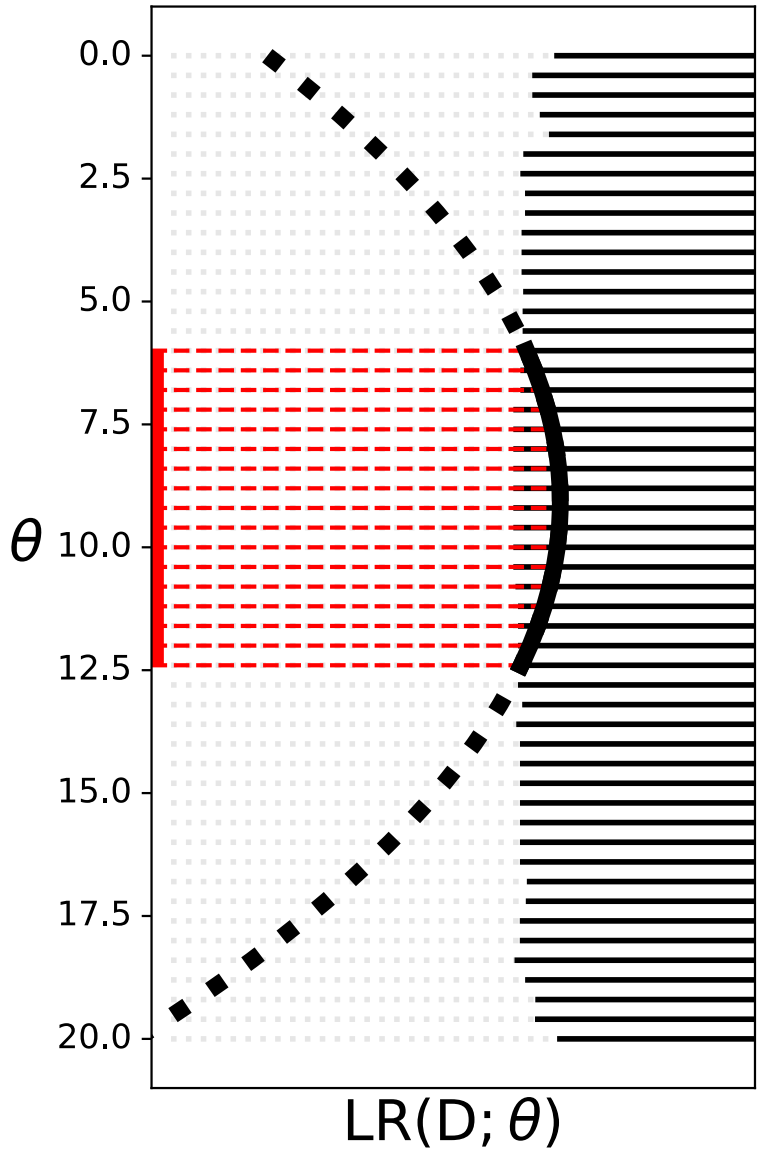




3. Observe data $\mathcal{D} = \mathrm{D}$. Calculate $\Lambda(\mathrm{D}; \theta)$.



4. Construct $(1 - \alpha)$ confidence set for θ.



Challenges

- Neyman construction itself. L. Lyons, "*Open Statistical Issues in Particle Physics*", AOAS 2008:

> However, in practice, it is very hard to use the Neyman frequentist construction when more than two or three parameters are involved: software to perform a Neyman construction efficiently in several dimensions would be most welcome. The

- Evaluation of frequentist coverage. R. Cousins: "*Lectures on Statistics in Theory: Prelude to Statistics in Practice*", arXiv:1807.05996.

> A complete, rigorous check of coverage considers a fine multi-D grid of *all* parameters, and *for each multi-D point in the grid*, generates an ensemble of toy MC pseudo-experiments, runs the full analysis procedure, and finds the fraction of intervals covering the μ_t of interest that was used for that ensemble. I.e., one calculates $P(\mu_t \in [\mu_1, \mu_2])$, and compares to C.L.

> *But*... the ideal of a fine grid is usually impractical.

How Do We Turn the Construction into Practical Procedures?

"Wrinkle": The Neyman construction requires one to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

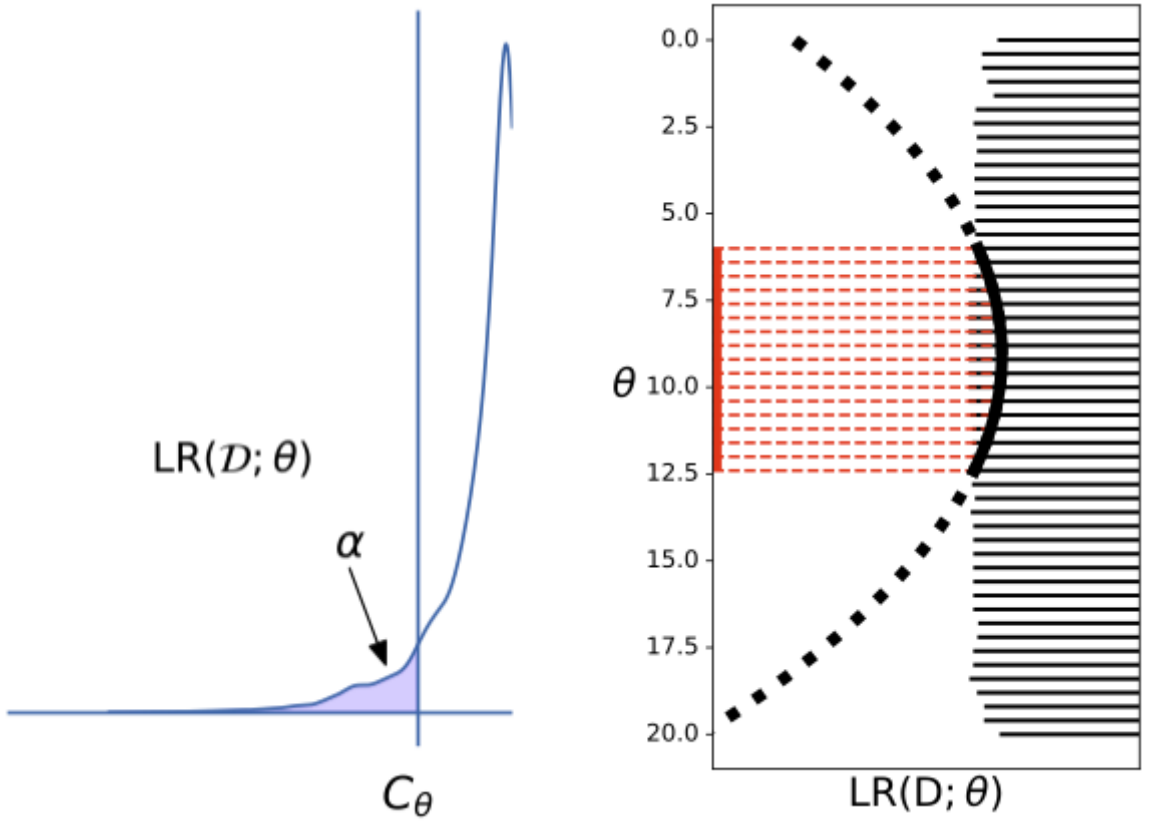
for *all* $\theta_0 \in \Theta$.

Key Realization: The main inferential quantities like

1. the **test statistic** $\tau(\mathcal{D}; \theta_0)$,
2. the **critical value** $C_{\theta_0,\alpha}$ or the p-value $p(D; \theta_0)$ of the test
3. the **coverage** $\mathbb{P}[\theta_0 \in R(\mathcal{D})]$ of the confidence set

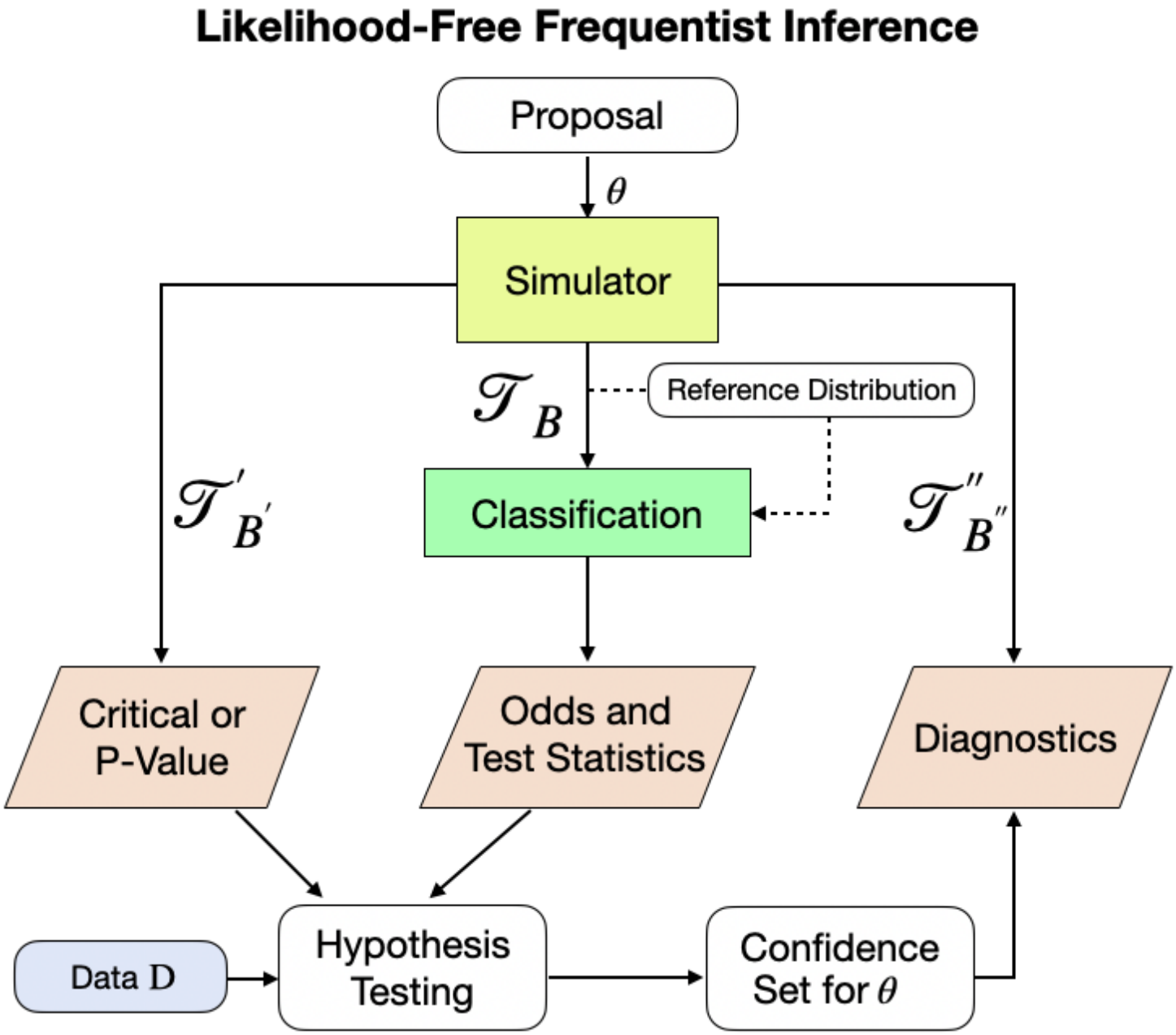
are conditional distribution functions which often vary smoothly as a function of the (unknown) parameters.

Efficient Construction of Finite-Sample Confidence Sets



Rather than running a batch of Monte Carlo simulations for every null hypothesis $\theta = \theta_0$ on, e.g., a fine enough grid in Θ, we can interpolate across the parameter space using training-based ML algorithms.

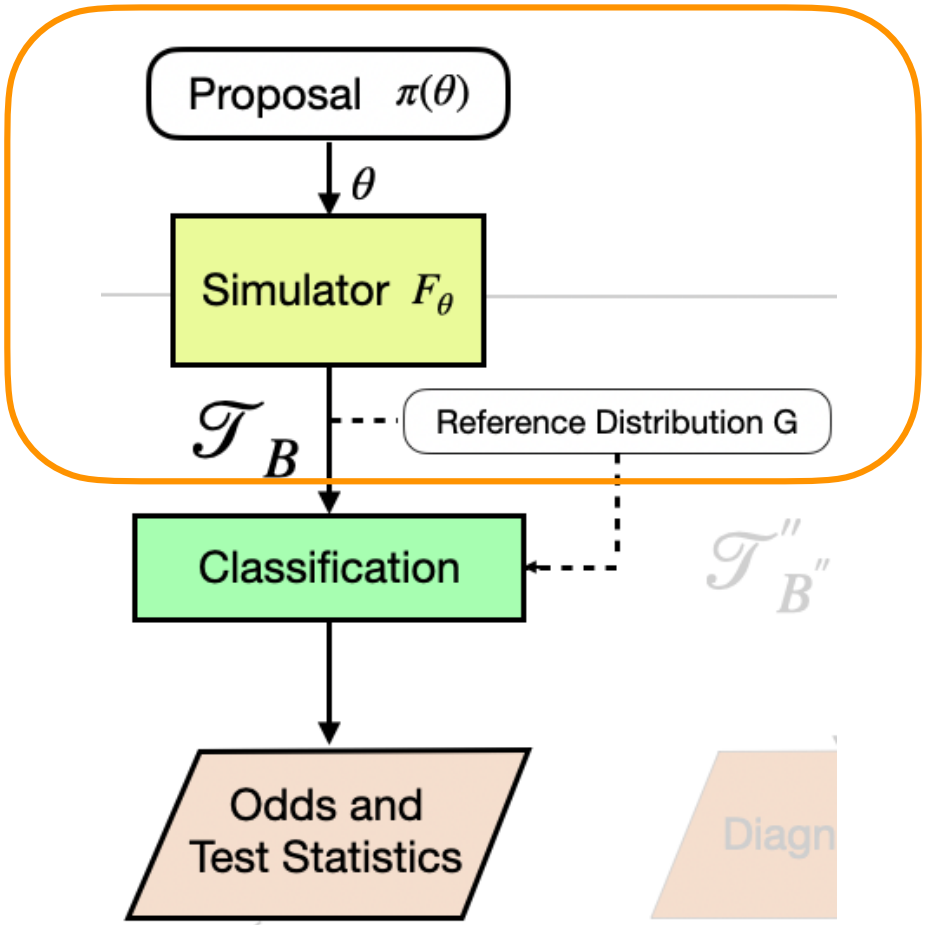
Our Inference Machinery



Center Branch: Estimating Odds and Test Statistic

Parameter: $\theta \in \Theta$

Simulated data: $\mathbf{X},\ \mathbf{x} \in \mathcal{X}$. *Observed data*: $\mathbf{X}^{\text{obs}}, \mathbf{x}^{\text{obs}} \in \mathcal{X}$.



1. Proposal distribution $\pi(\theta)$ over the parameter space Θ
2. Forward simulator F_θ
   - $F_{\theta_1} \neq F_{\theta_2}$ for $\theta_1 \neq \theta_2 \in \Theta$
3. Reference distribution G over the feature space $\mathcal{X}$
   - $F_\theta \ll G$ for all $\theta \in \Theta$
4. A simulated sample of size B to estimate odds and test statistic

Estimate Odds via Probabilistic Classification

Simulate two samples:

- $\{(\theta_k, \mathbf{X}_k, Y_k = 1)\}_{k=1}^{B/2}$, where $\theta \sim \pi(\theta)$, $\mathbf{X} \sim F_\theta$
- $\{(\theta_l, \mathbf{X}_l, Y_l = 0)\}_{l=1}^{B/2}$ where $\theta \sim \pi(\theta)$, $\mathbf{X} \sim G$

Probabilistic classifier r:

$$r : (\theta, \mathbf{X}) \longrightarrow \mathbb{P}(Y = 1 | \mathbf{X}, \theta)$$

Define the **odds** at $\theta \in \Theta$ and fixed $\mathbf{x} \in \mathcal{X}$ as

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

Interpretation: Chance that $\mathbf{x}$ was generated from F_θ rather than G.

Estimate Odds via Probabilistic Classification

Simulate two samples:

- $\{(\theta_k, \mathbf{X}_k, Y_k = 1)\}_{k=1}^{B/2}$, where $\theta \sim \pi(\theta)$, $\mathbf{X} \sim F_\theta$
- $\{(\theta_l, \mathbf{X}_l, Y_l = 0)\}_{l=1}^{B/2}$ where $\theta \sim \pi(\theta)$, $\mathbf{X} \sim G$

Probabilistic classifier r:

$$r : (\theta, \mathbf{X}) \longrightarrow \mathbb{P}(Y = 1 | \mathbf{X}, \theta)$$

Define the **odds** at $\theta \in \Theta$ and fixed $\mathbf{x} \in \mathcal{X}$ as

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

Interpretation: Chance that $\mathbf{x}$ was generated from F_θ rather than G.

Test Statistics Based on Odds: ACORE and BFF

Suppose we want to test:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1, \quad \text{where } \Theta_1 = \Theta_0^c.$$

For observed data $\mathrm{D} = \{\mathbf{X}_1^{\text{obs}}, ..., \mathbf{X}_n^{\text{obs}}\}$, we define

- ACORE (Approximate Computation via Odds Ratio Estimation):

$$\widehat{\Lambda}(\mathcal{D}; \Theta_0) := \log \frac{\sup_{\theta_0 \in \Theta_0} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta)}$$

- BFF (Bayesian Frequentist Factor):

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

where π_0 and π_1 are the restrictions of a proposal distribution π_τ over Θ to Θ_0 and Θ_0^c, respectively.

ACORE and BFF are Approximations of the LR Statistic and the Bayes Factor respectively!

Lemma (Fisher's Consistency)

If $\quad \widehat{\mathbb{P}}(Y = 1|\theta, \mathbf{X}) = \mathbb{P}(Y = 1|\theta, \mathbf{x}) \, \forall \theta, \mathbf{X}$

1. $\implies \widehat{\Lambda}(\mathcal{D}; \Theta_0) = \mathrm{LR}(\mathcal{D}; \Theta_0) \equiv \log \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\mathcal{D};\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\mathcal{D};\theta)},$

2. $\implies \widehat{\tau}(\mathcal{D}; \Theta_0) = \mathrm{BF}(\mathcal{D}; \Theta_0) \equiv \frac{\mathbb{P}(\mathcal{D}|H_0)}{\mathbb{P}(\mathcal{D}|H_1)} = \frac{\int_{\Theta_0} \mathcal{L}(\mathcal{D};\theta) d\pi_0(\theta)}{\int_{\Theta_1} \mathcal{L}(\mathcal{D};\theta) d\pi_1(\theta)}.$

Note: The Bayes factor is often used as a Bayesian alternative to significance testing but here we are treating it as a frequentist test statistic.

Test Statistics Based on Odds: ACORE and BFF

Suppose we want to test:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

For observed data $\mathrm{D} = \{\mathbf{X}_1^{\text{obs}}, ..., \mathbf{X}_n^{\text{obs}}\}$, we define

- ACORE (Approximate Computation via Odds Ratio Estimation):

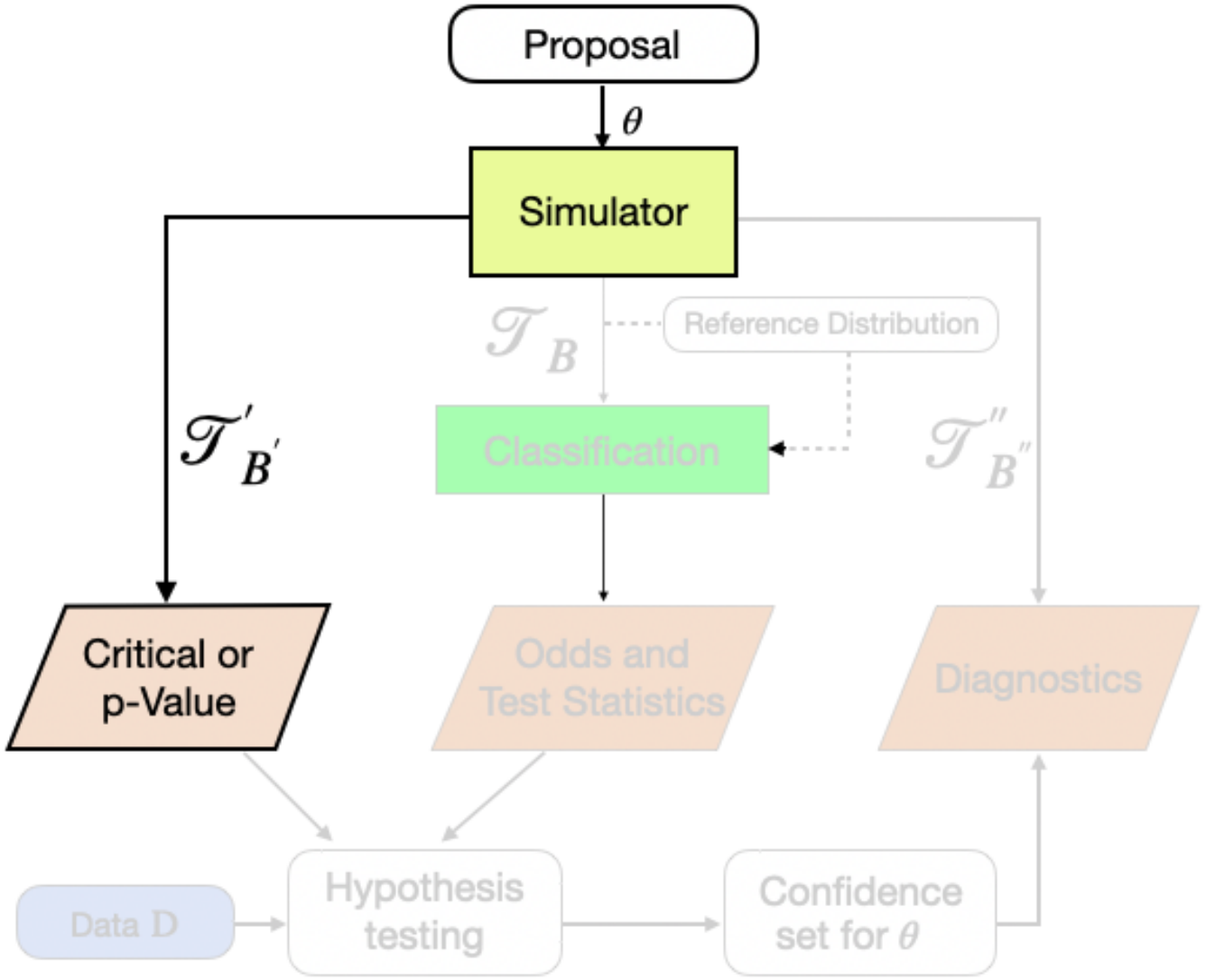
$$\widehat{\Lambda}(\mathcal{D}; \theta_0) := \log \frac{\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}$$

- BFF (Bayesian Frequentist Factor):

$$\widehat{\tau}(\mathcal{D}; \theta_0) := \frac{\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}{\int_\Theta \left(\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta) \right) d\pi_\tau(\theta)}.$$

where $\pi_\tau(\theta)$ is a probability distribution over the parameter space.

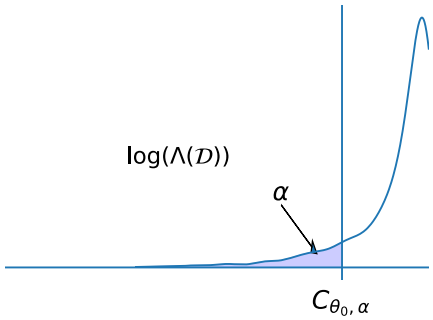
Left Branch: Estimate Critical Values or P-Values



We use $B^{'}$ simulations to estimate critical values.

Estimate Critical Values $C_{\theta_0,\alpha}$

To control Type I error at level α:
Reject $H_0 : \theta = \theta_0$ when $\tau(\mathcal{D}; \theta_0) < C_{\theta_0,\alpha}$, where

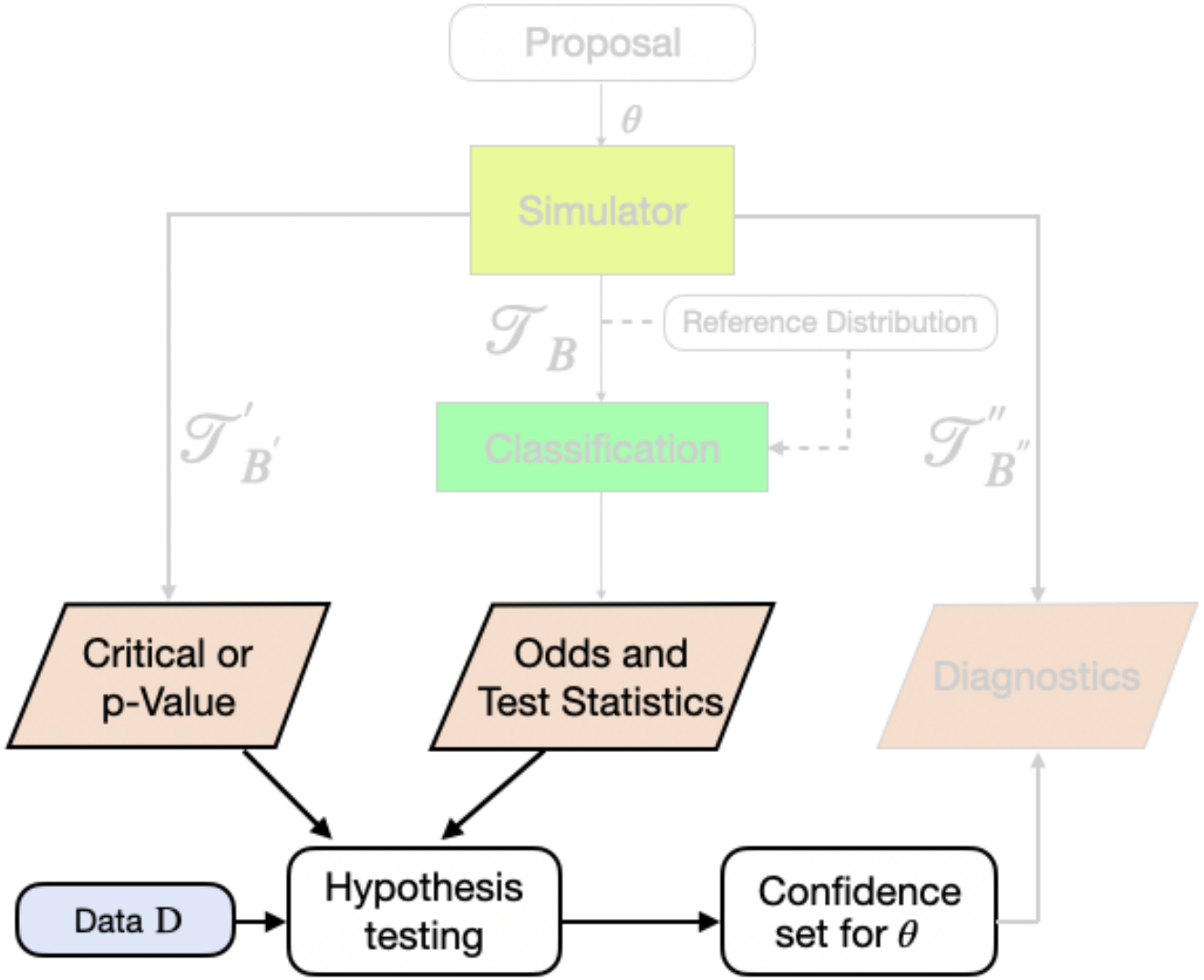


$$C_{\theta_0,\alpha} = \arg\sup_{C \in \mathbb{R}} \{C : \mathbb{P}(\tau(\mathcal{D}; \theta_0) < C \mid \theta = \theta_0) \leq \alpha\}.$$

Problem: Need to estimate $\mathbb{P}(\tau(\mathcal{D}; \theta) < C \mid \theta)$ for every $\theta \in \Theta$.

Solution: $F_{\tau|\theta}(C|\theta) \equiv \mathbb{P}(\tau(\mathcal{D}; \theta) < C|\theta)$ is a conditional CDF, so we can estimate its α-quantile via quantile regression $F^{-1}_{\tau|\theta}(\alpha|\theta)$.

Construct Confidence Set via Neyman Inversion



Are the constructed confidence sets valid?

Theorem (Validity for Any Test Statistic)

Let τ_B be an estimated test statistic and assume the quantile regression estimator is consistent. Then, for any fixed $\alpha \in (0,1)$ and $\theta_0 \in \Theta$:

$$\widehat{C}_{\theta_0,B} \xrightarrow[B' \longrightarrow \infty]{\mathbb{P}} C_{\theta_0,B},$$

where $C_{\theta_0,B}$ is such that $\mathbb{P}(\tau_B \leq C_{\theta_0,B}|\theta) = \alpha$.

If B' is large enough, we can construct a confidence set with guaranteed nominal coverage regardless of the observed sample size n

What Can We Say about Power?

Suppose we are testing

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0$$

and assume that the critical values are well estimated (that is, $B^{'}$ is large enough).

Consider

- $\phi_{\widehat{\tau}_B}(\mathcal{D}) = \mathbb{I}\left(\widehat{\tau}_B(\mathcal{D}; \theta_0) < C_{\theta_0, B}\right)$: decision of approximate test
- $\phi_{\tau}(\mathcal{D}) = \mathbb{I}\left(\tau(\mathcal{D}; \theta_0) < C_{\theta_0}\right)$: decision of exact test

Theorem

If the probabilistic classifier for learning the odds is consistent, and $C_{\theta,B} \xrightarrow[B \longrightarrow \infty]{\mathbb{P}} C_{\theta}$, *then, for every* $\theta \in \Theta$*:*

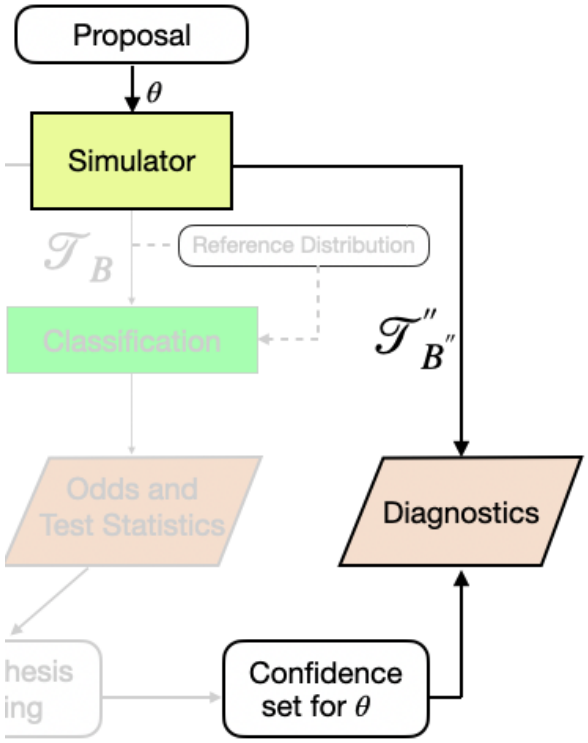
$$\mathbb{P}_{\mathcal{D}, \mathcal{T}_B | \theta}\left(\phi_{\widehat{\tau}_B}(\mathcal{D}) = 1\right) \xrightarrow[B \longrightarrow \infty]{} \mathbb{P}_{\mathcal{D}|\theta}\left(\phi_{\tau}(\mathcal{D}) = 1\right).$$

Right Branch: Checking Actual Coverage Across Θ

How do we check coverage $\mathbb{P}(\theta \in R(\mathcal{D}))$ as a function of $\theta \in \Theta$?

Note: $\mathbb{P}(\theta \in R(\mathcal{D})|\theta) = \mathbb{E}[\mathbb{I}(\theta \in R(\mathcal{D}))|\theta]$

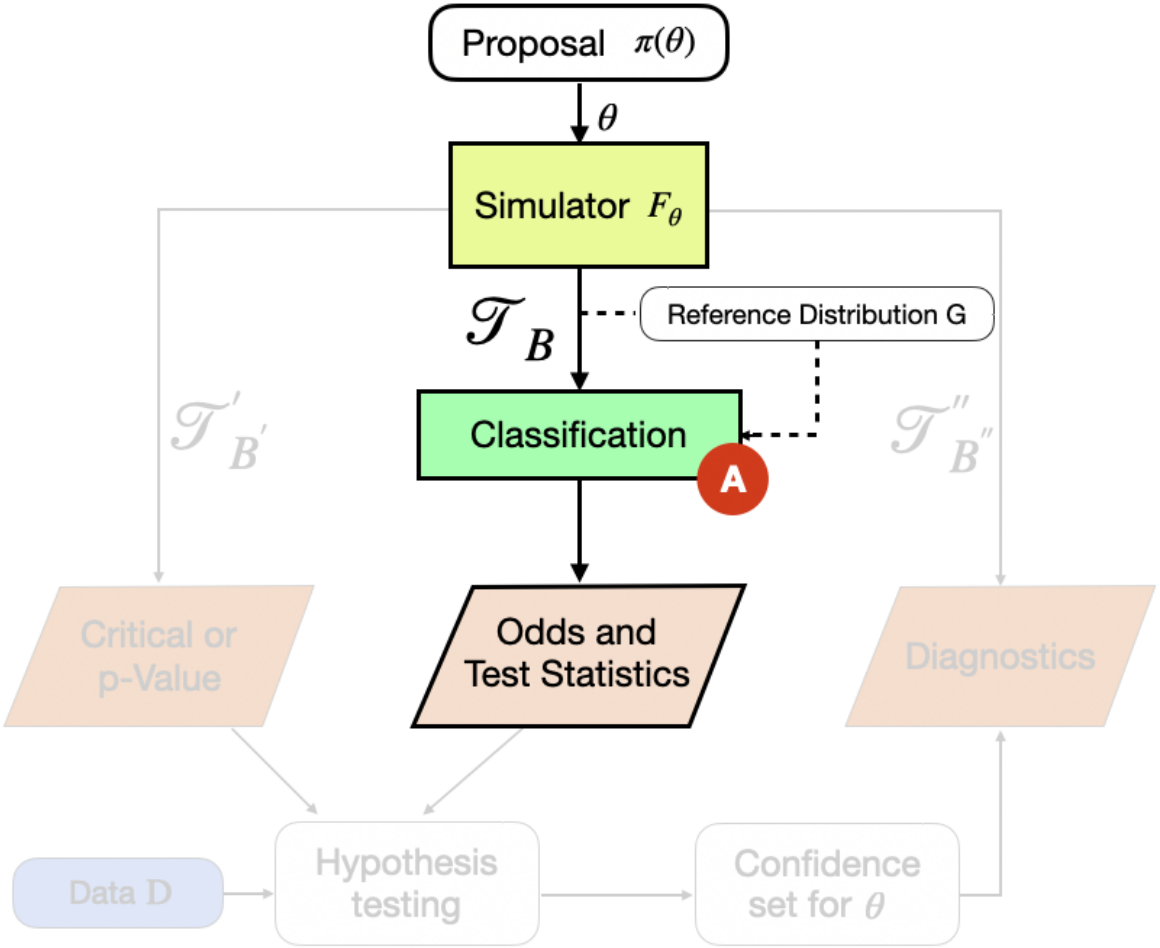
That is, we can estimate empirical coverage across the entire parameter space by regression (probabilistic classification):



1. Sample θ_i and data $\mathcal{D}_i \sim F_{\theta_i}$
2. Construct confidence set $R(\mathcal{D}_i)$
3. For $\{\theta_i, R(\mathcal{D}_i)\}_{i=1}^{B''}$, regress $Z_i = \mathbb{I}(\theta_i \in R(\mathcal{D}_i))$ on θ_i

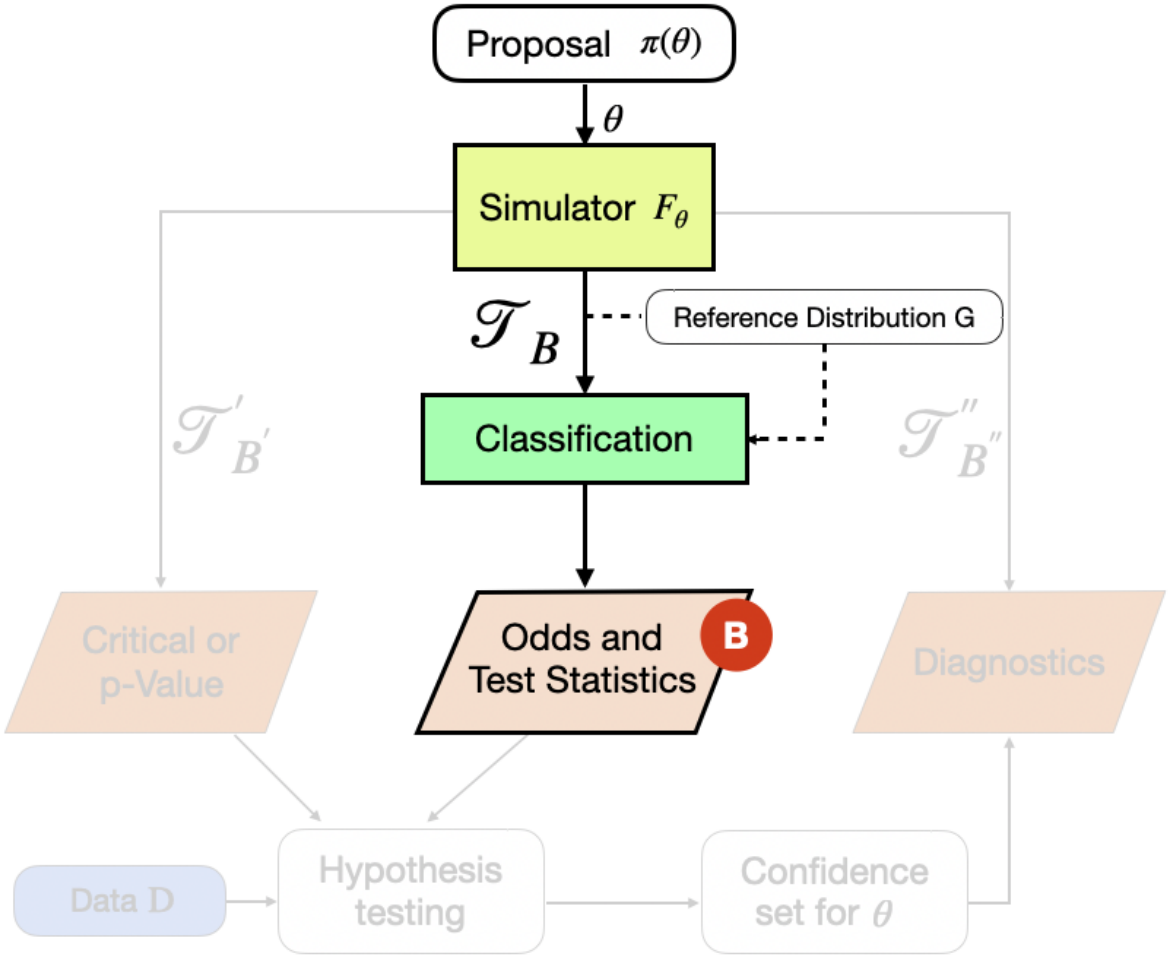
How close is the actual coverage to the nominal confidence level $1 - \alpha$?

A Practical Strategy



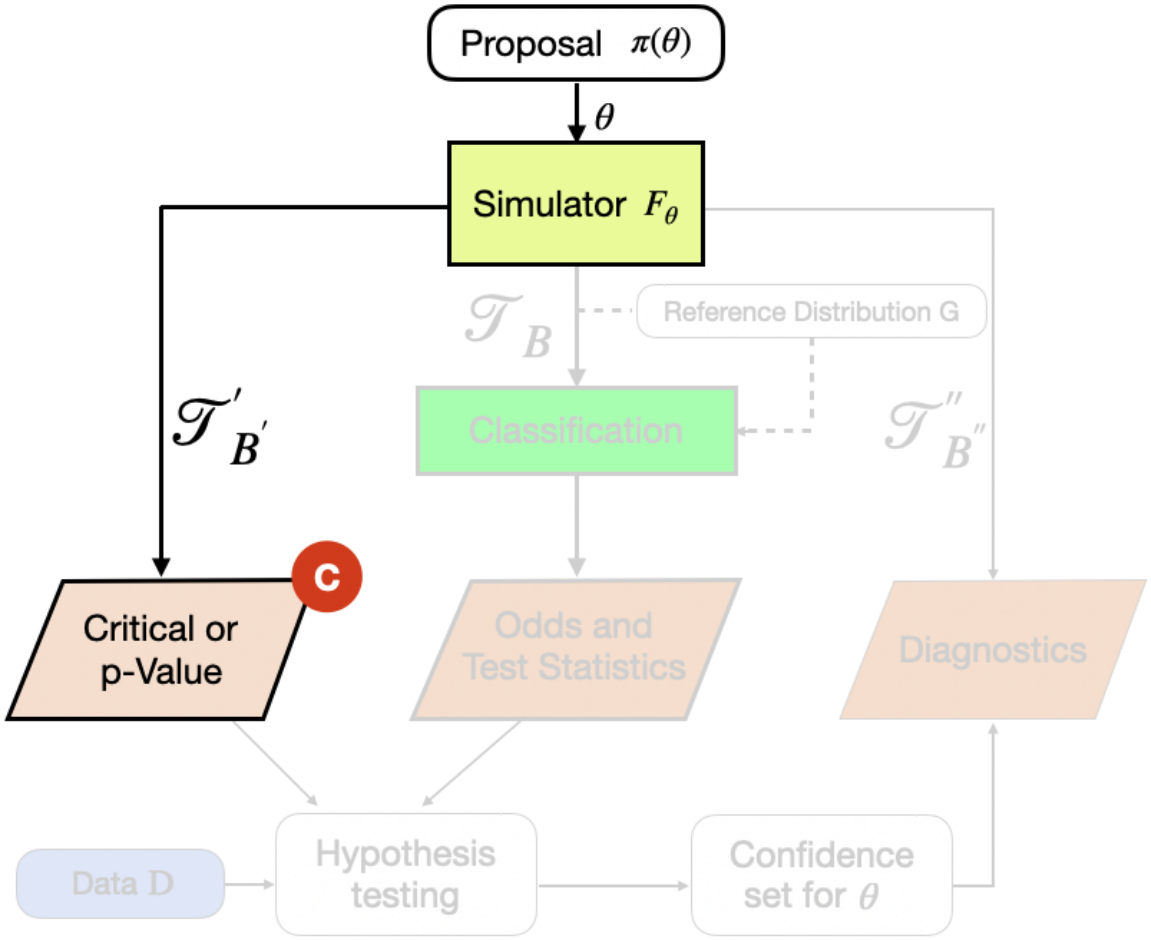
(A) Use the cross-entropy on a held-out set to select probabilistic classifier and sample size B for learning the odds

A Practical Strategy



(B) Compute the maximization (or integration) step in `ACORE` (or `BFF`) with all available computational budget.

A Practical Strategy



(C) Use our diagnostic tool to determine the quantile regression algorithm and sample size B' to achieve nominal coverage across Θ.

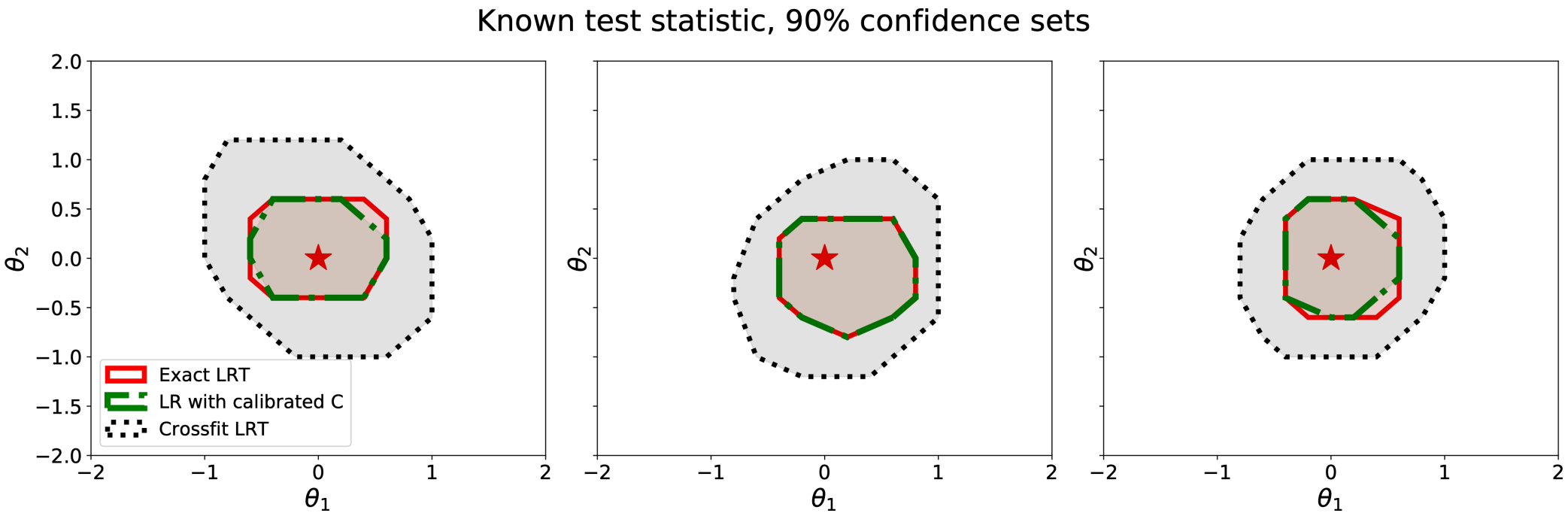
How Does Our Machinery Scale?

Consider an example where the forward model is just a MVG distribution $N(\boldsymbol{\theta}, \mathbf{I}_d)$. Construct a confidence set for the unknown mean $\boldsymbol{\theta} \in \mathbb{R}^d$.

- Suppose the observed data are $\mathbf{X}_1, \ldots, \mathbf{X}_{10} \sim N(\mathbf{0}, \mathbf{I}_d)$ so $n = 10$ and $\boldsymbol{\theta} = \mathbf{0}$ (unknown parameter).
- Both parameter and feature space have dimension d
- Compare our results to Exact LRT (known baseline, UMPU)
  1. Setting 1: Test statistic is known, but not its null distribution and the critical values. [Compare our results to Crossfit LRT — a "universal inference"[1] method for constructing valid finite-sample confidence sets without regularity conditions or calibration.]
  2. Setting 2: "LFI setting". Need to estimate both test statistic and the critical values.

[1] Wasserman, Ramdas and Balakrishnan; PNAS 2020

Confidence Sets for "Known Test Statistic" and 2D Gaussian Data.

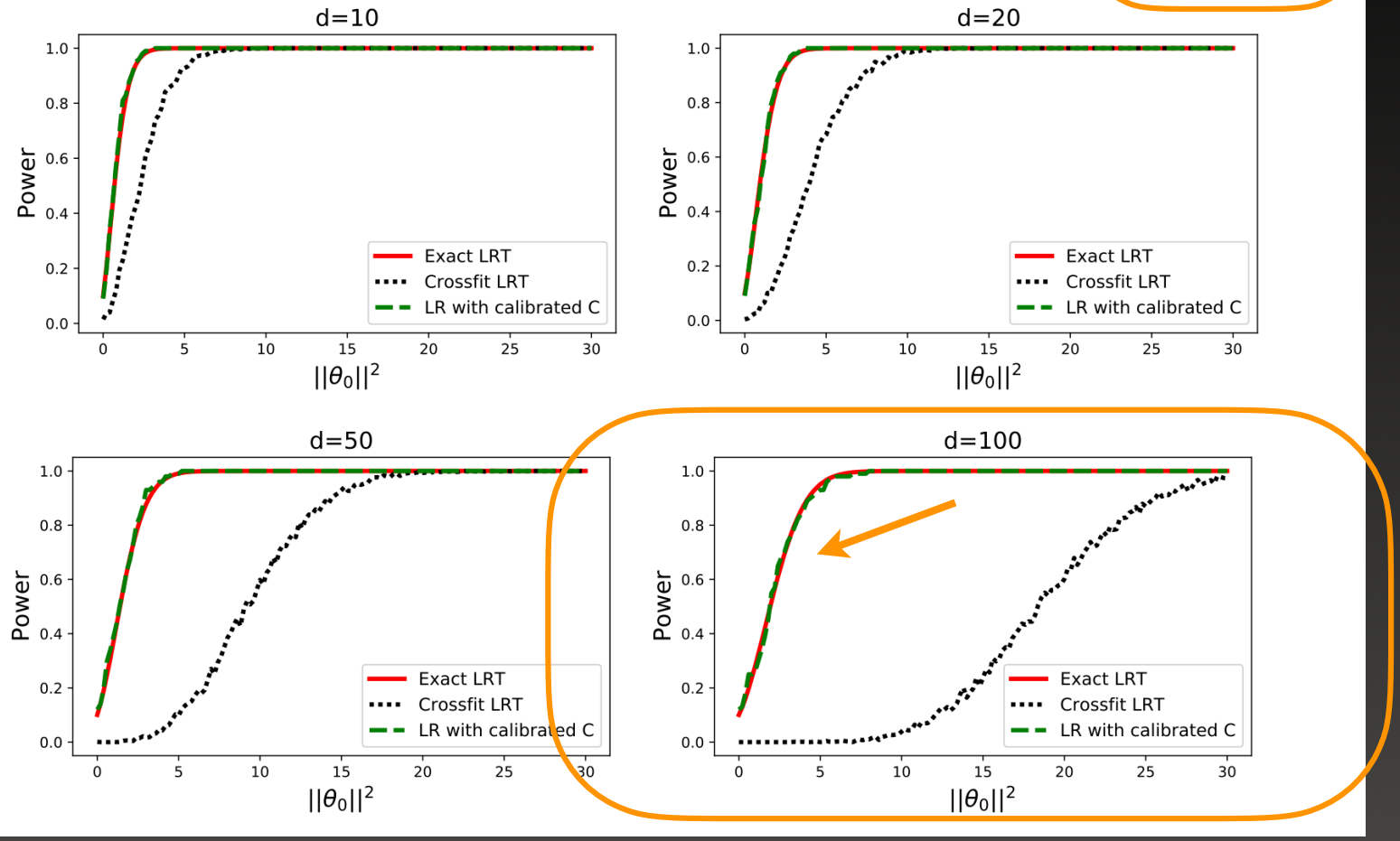


When d=2, our method for estimating critical values (GREEN) returns confidence sets that are close to "Exact LRT" (RED), but smaller than the more conservative universal inference approach with "Crossfit LRT" (GRAY dotted)

Coverage and Power for Known Test Statistic

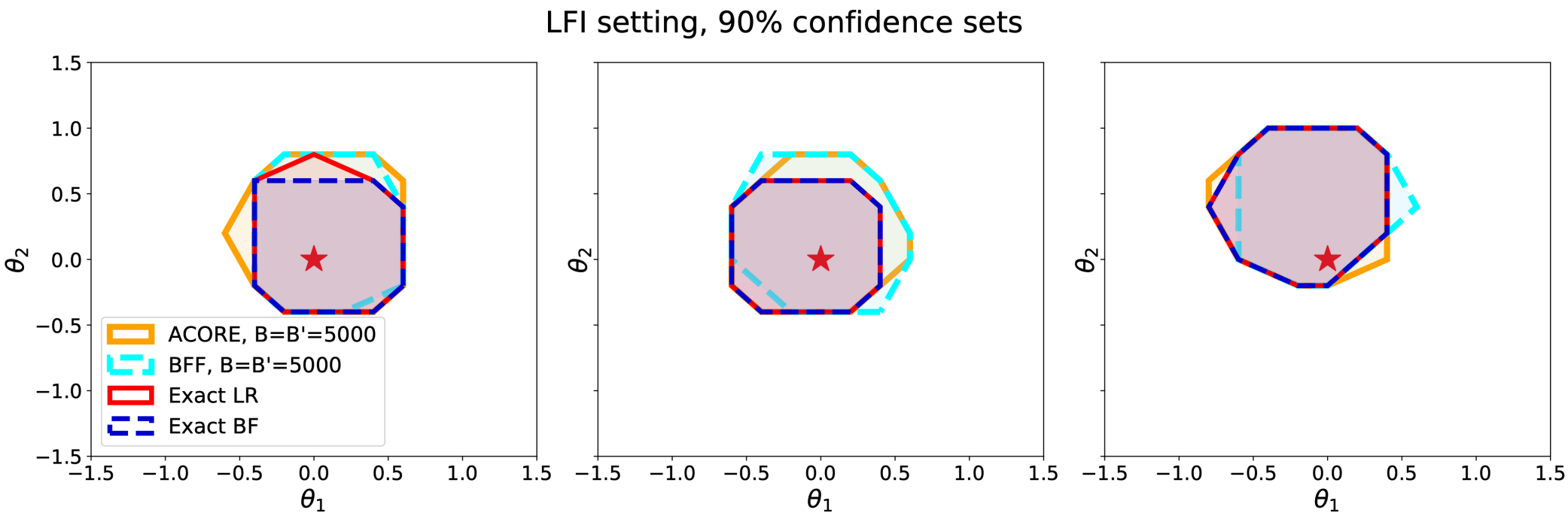
Finite-sample confidence sets for known test statistic

	d=10	d=20	d=50	d=100
Coverage of LR with calibrated C	0.91 ± 0.03	0.91 ± 0.03	0.88 ± 0.03	0.88 ± 0.03
Coverage of crossfit LRT	0.993 ± 0.008	0.997 ± 0.005	1.000 ± 0.000	1.000 ± 0.000



Our approach for estimating critical values yields the same power as the exact tests even in high dimensions, with a modest sample size of B'=5000.

Confidence Sets for "LFI Setting" and 2D Gaussian Data (all confidence sets valid)

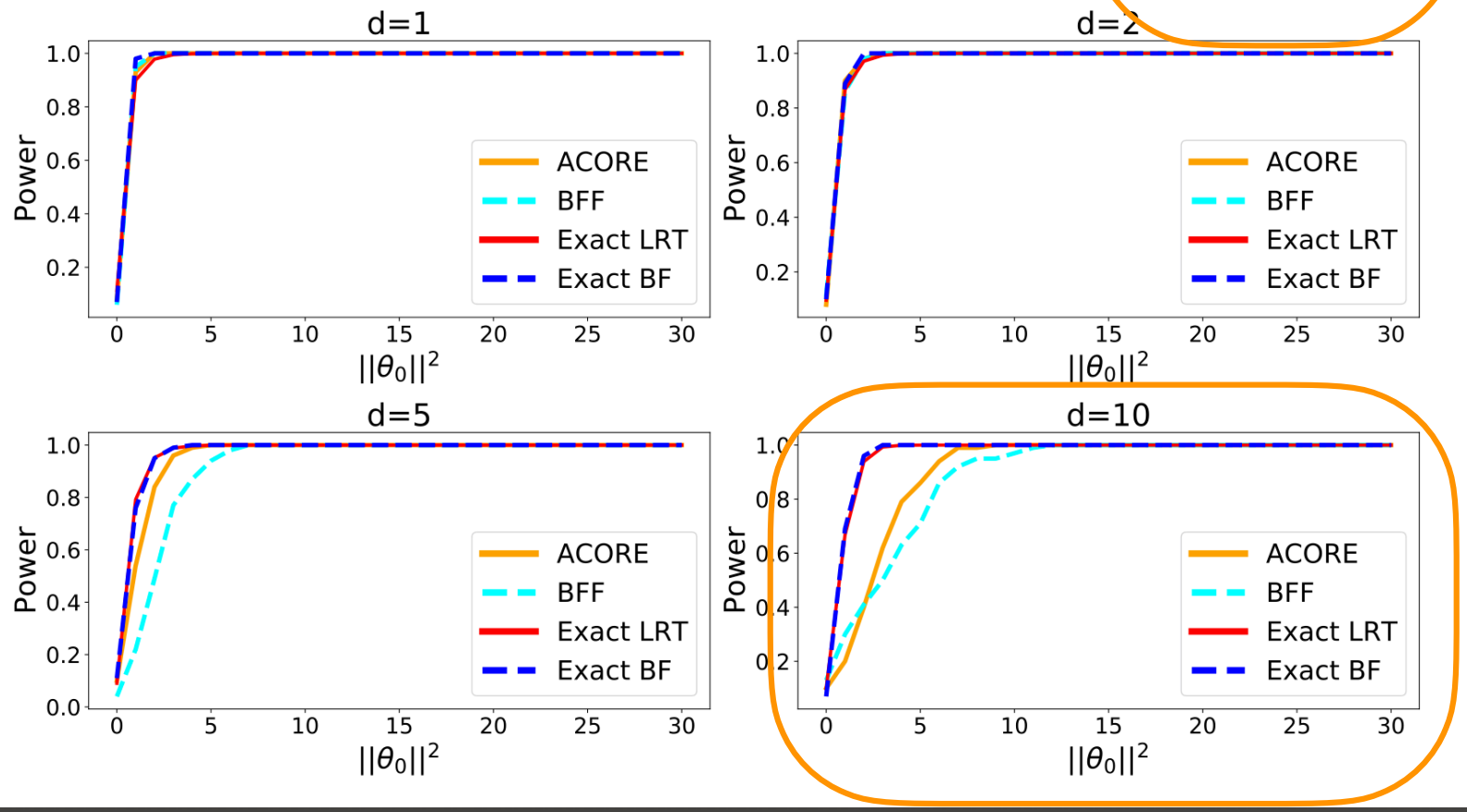


When d=2, ACORE and BFF confidence sets (for B=B'=5000) are similar in size to the Exact LR confidence sets.

Coverage and Power in an LFI Setting

Finite-sample confidence sets in a likelihood-free inference setting

	d=1	d=2	d=5	d=10
Coverage of ACORE	0.92 ± 0.03	0.92 ± 0.03	0.90 ± 0.03	0.90 ± 0.03
Coverage of BFF	0.94 ± 0.02	0.89 ± 0.03	0.96 ± 0.02	0.87 ± 0.03



In higher dimensions, ACORE and BFF confidence sets are still valid but lose some power with respect to their exact counterparts.

Break-Down of Sources of Errors in Freq LFI

- `ACORE` (Approximate Computation via Odds Ratio Estimation):

$$\widehat{\Lambda}(\mathcal{D};\theta_0) := \log \frac{\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\sup_{\theta\in\Theta} \prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}$$

- `BFF` (Bayesian Frequentist Factor):

$$\widehat{\tau}(\mathcal{D};\theta_0) := \frac{\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\int_{\Theta} \left(\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta)\right) d\pi_{\tau}(\theta)}.$$

- e_1: error in estimating the odds function
- e_2: numerical error when computing test statistics
  - power depends on both e_1 and e_2
- e_3: error in estimating the critical values
  - validity determined by e_3 (if B' large enough, then $e_3 \approx 0$)

Break-Down of Sources of Errors in Freq LFI

- `ACORE` (Approximate Computation via Odds Ratio Estimation):

$$\widehat{\Lambda}(\mathcal{D};\theta_0) := \log \frac{\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\sup_{\theta\in\Theta} \prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}$$

- `BFF` (Bayesian Frequentist Factor):

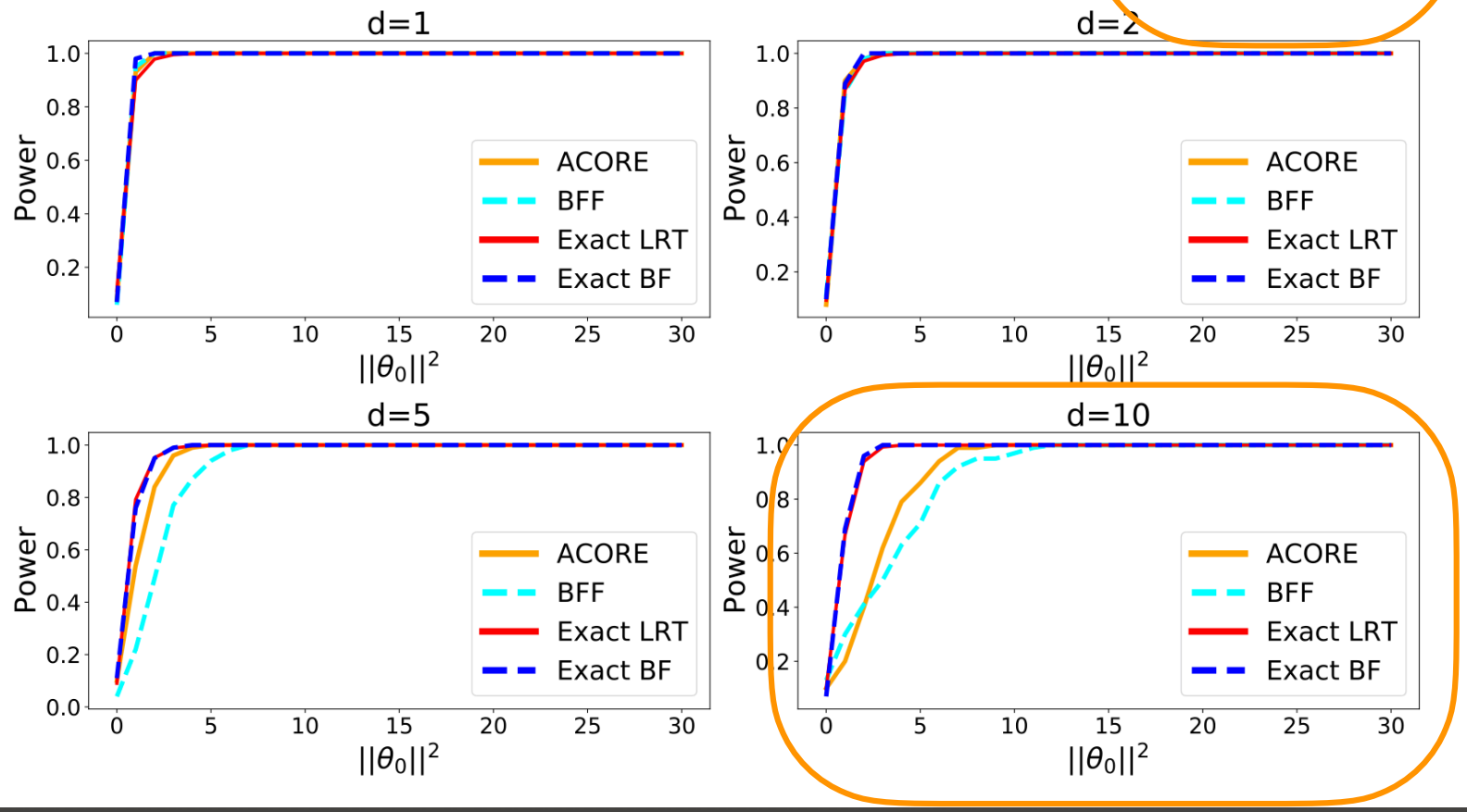
$$\widehat{\tau}(\mathcal{D};\theta_0) := \frac{\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\int_{\Theta} \left(\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta)\right) d\pi_{\tau}(\theta)}.$$

- e_1: error in estimating the odds function
- e_2: numerical error when computing test statistics
  - power depends on both e_1 and e_2
- e_3: error in estimating the critical values
  - validity determined by e_3 (if B' large enough, then $e_3 \approx 0$)

Coverage and Power in an LFI Setting

Finite-sample confidence sets in a likelihood-free inference setting

	d=1	d=2	d=5	d=10
Coverage of ACORE	0.92 ± 0.03	0.92 ± 0.03	0.90 ± 0.03	0.90 ± 0.03
Coverage of BFF	0.94 ± 0.02	0.89 ± 0.03	0.96 ± 0.02	0.87 ± 0.03



In higher dimensions, ACORE and BFF confidence sets are still valid but lose some power with respect to their exact counterparts.

How Do we Handle Nuisance Parameters?

In many applications, the parameter space can be decomposed as $\Theta = \Phi \times \Psi$, where Φ are the parameters of interest, and Ψ are nuisance parameters not of immediate interest.

To guarantee frequentist coverage with Neyman's inversion technique, we need to test null hypotheses

$$H_{0,\phi_0} : \phi = \phi_0 \quad \text{versus} \quad H_{1,\phi_0} : \phi \neq \phi_0 \quad \text{for } \phi_0 \in \Phi$$

by comparing test statistics to the cutoffs $\widehat{C}_{\phi_0} := \inf_{\psi \in \Psi} \widehat{C}_{(\phi_0,\psi)}$.

Can lead to numerically unwieldy and costly computations.

ACORE: Handling Nuisance Parameters by Maximization

For ACORE, we use a hybrid or "likelihood profiling" method.[2]

For each ϕ, we compute an approximation of the MLE of ψ for observed data $\mathcal{D}$:

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

Rather than comparing the ACORE test statistic $\widehat{\Lambda}(\mathcal{D}; \phi_0) = \widehat{\Lambda}(\mathcal{D}; (\phi_0, \widehat{\psi}_{\phi}))$ to $\widehat{C}_{\phi_0} := \inf_{\psi \in \Psi} \widehat{C}_{(\phi_0, \psi)}$, we use the hybrid cutoffs:

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

where the quantile regression is based on a training sample $\mathcal{T}'$ generated at *fixed* $\hat{\psi}_{\phi_0}$.

[2] Van der Vaart, 2000; Chuang & Lai, 2000; Feldman, 2000; Sen et al. 2009

BFF: Handling Nuisance Parameters by Integration

For BFF, we eliminate the nuisance parameters via integration.

By definition,

$$\widehat{\tau}(\mathcal{D};\phi_0) := \frac{\int_\Psi \prod_{i=1}^n \widehat{\mathbb{O}}\left(\mathbf{X}_i^{\mathsf{obs}};(\phi_0,\psi)\right) d\pi(\psi)}{\int_\Theta \left(\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\mathsf{obs}};\theta)\right) d\pi(\theta)},$$

where $\pi(\psi)$ is a distribution over Ψ, the nuisance parameter space.

Instead of using hybrid resampling, we approximate the cutoffs at parameter of interest ϕ_0 according to

$$\widehat{C}_{\phi_0} := \widehat{F}^{-1}_{\widehat{\tau}(\mathcal{D};\phi_0)\big|(\phi_0)}(\alpha)$$

Hybrid Methods and Confidence Sets

- Hybrid methods (which maximize or average 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 will give better frequentist coverage for the parameter of interest*" (Cousins 2018)
- Can our diagnostic tools provide guidance as to which method to choose for the problem at hand?

Toy Example: Poisson Counting Experiment

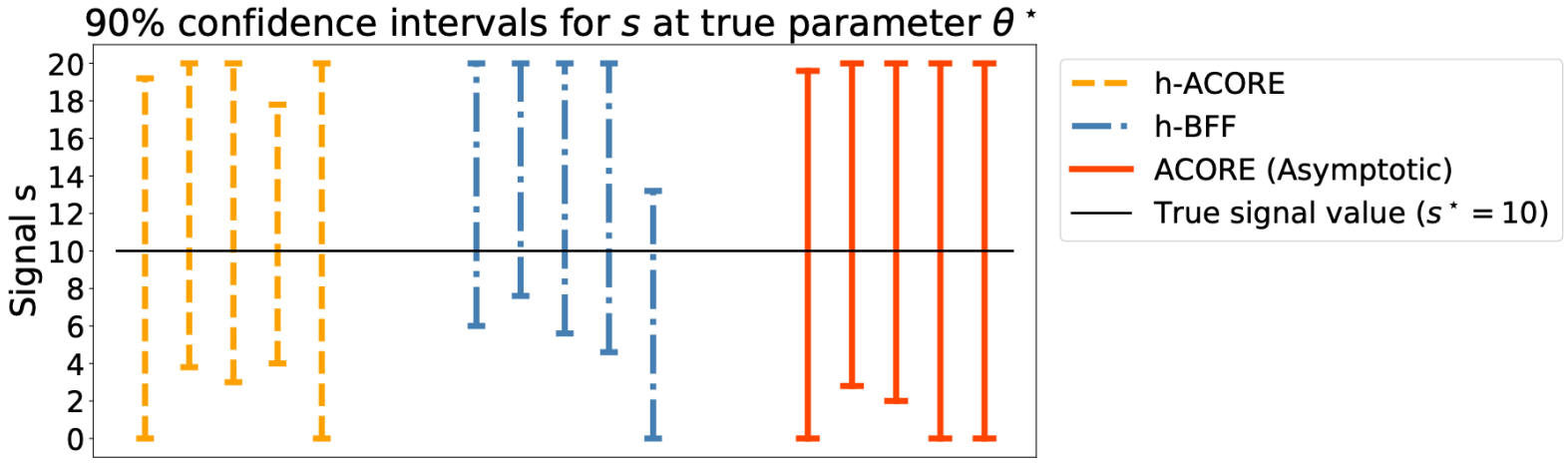
- Particle collision events counted under the presence of a background process.

$$\text{Observed data } D = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{10})$$
$$\mathbf{X} = (M, N), \text{ where } M \sim \text{Pois}(\gamma b),\ N \sim \text{Pois}(b + \epsilon s)$$

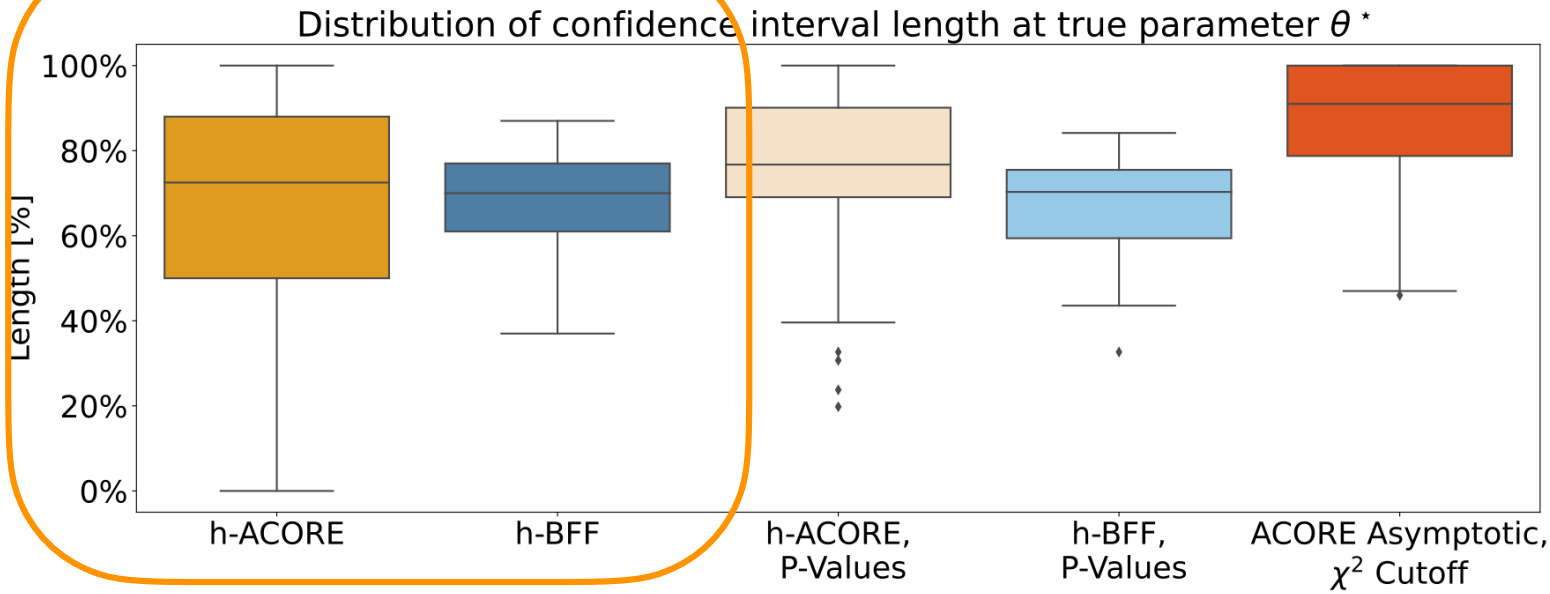
- The observed data D consist of n=10 realizations of X=(M,N), where
  - M is the number of events in the control region (assume γ=1)
  - N is the number of events in the signal region
- Unknown parameters:
  - signal strength (s); two nuisance parameters (b and ϵ)

Confidence sets at a fiducial point

HEP example with nuisance parameters



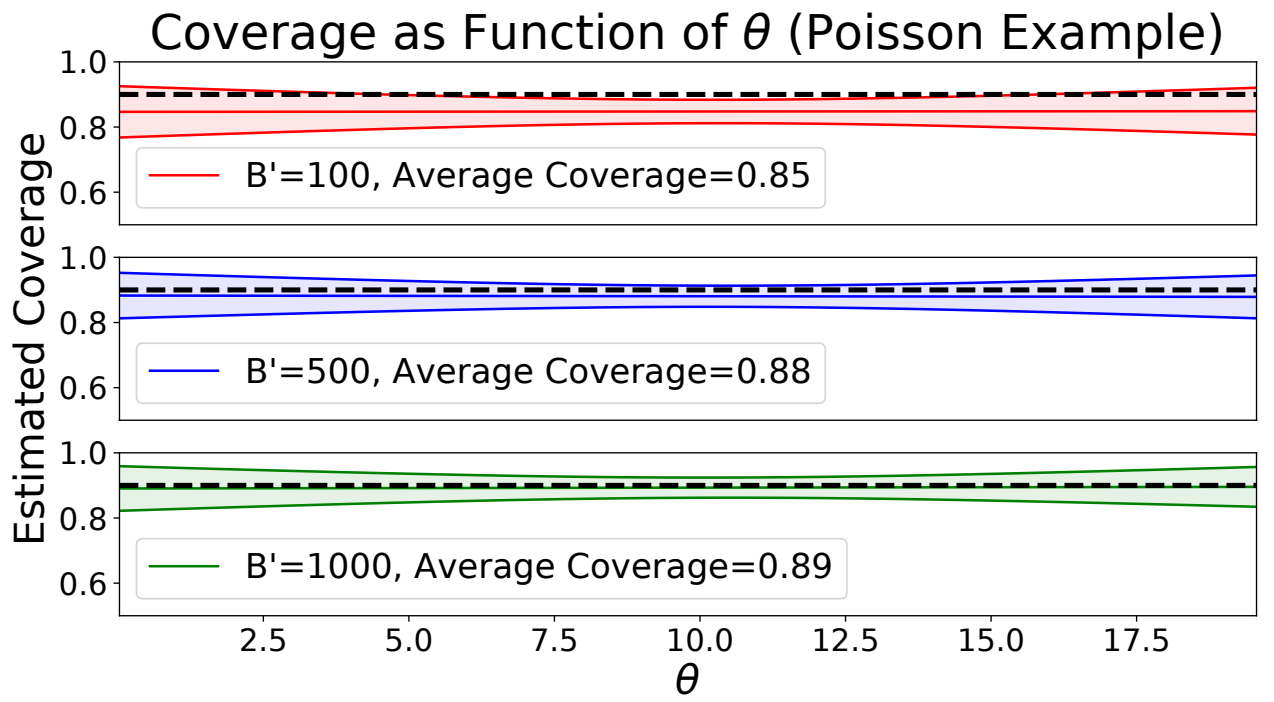
	h-ACORE	h-BFF	h-ACORE (p-values)	h-BFF (p-values)	ACORE (Asymptotic)
Coverage	0.87 ± 0.03	0.91 ± 0.03	0.92 ± 0.03	0.94 ± 0.02	0.97 ± 0.02



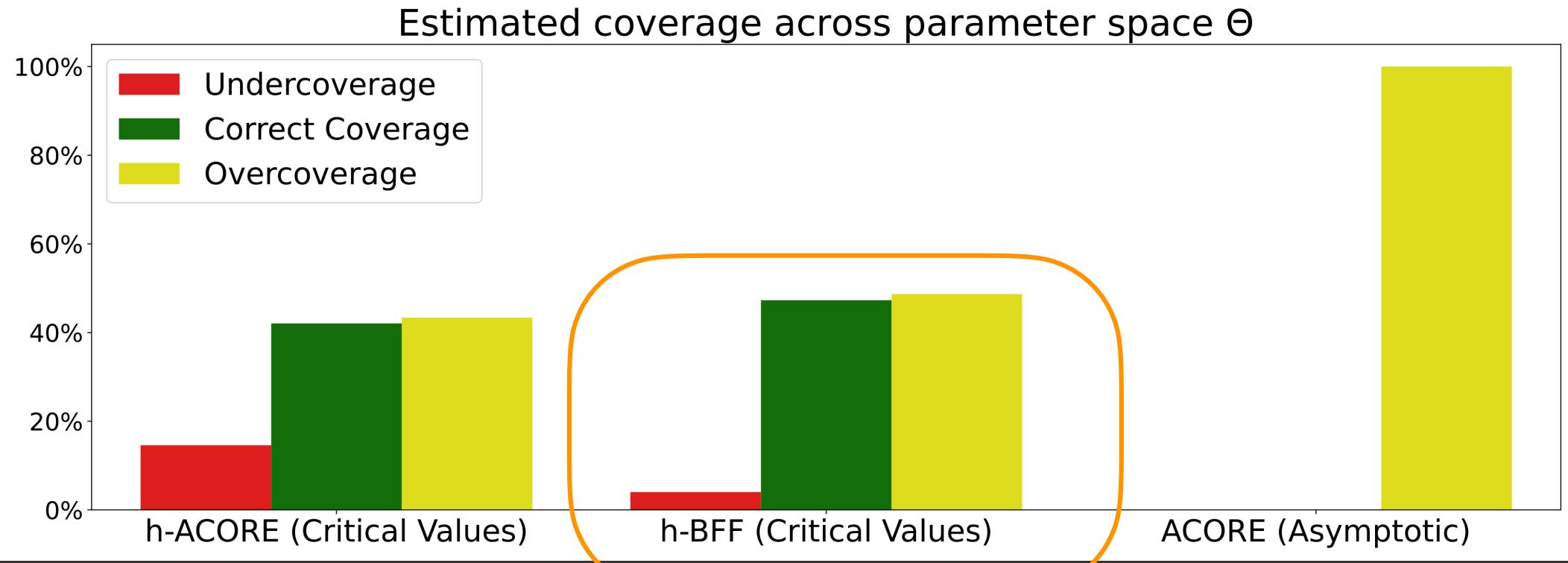
Diagnostic Tool to Check Coverage

- With logistic regression and B''=500 simulations, we estimate the coverage with a 2**σ** prediction band for all (s, b, **ϵ**) across the entire parameter space s∈[0,20] b∈[90.110], **ϵ**∈[0.5,1.0]

- If the nominal coverage of 1-α=0.9 falls within the prediction band => correct coverage. Upper/lower 2**σ** limit falls below/ above 0.9 => under/over coverage.



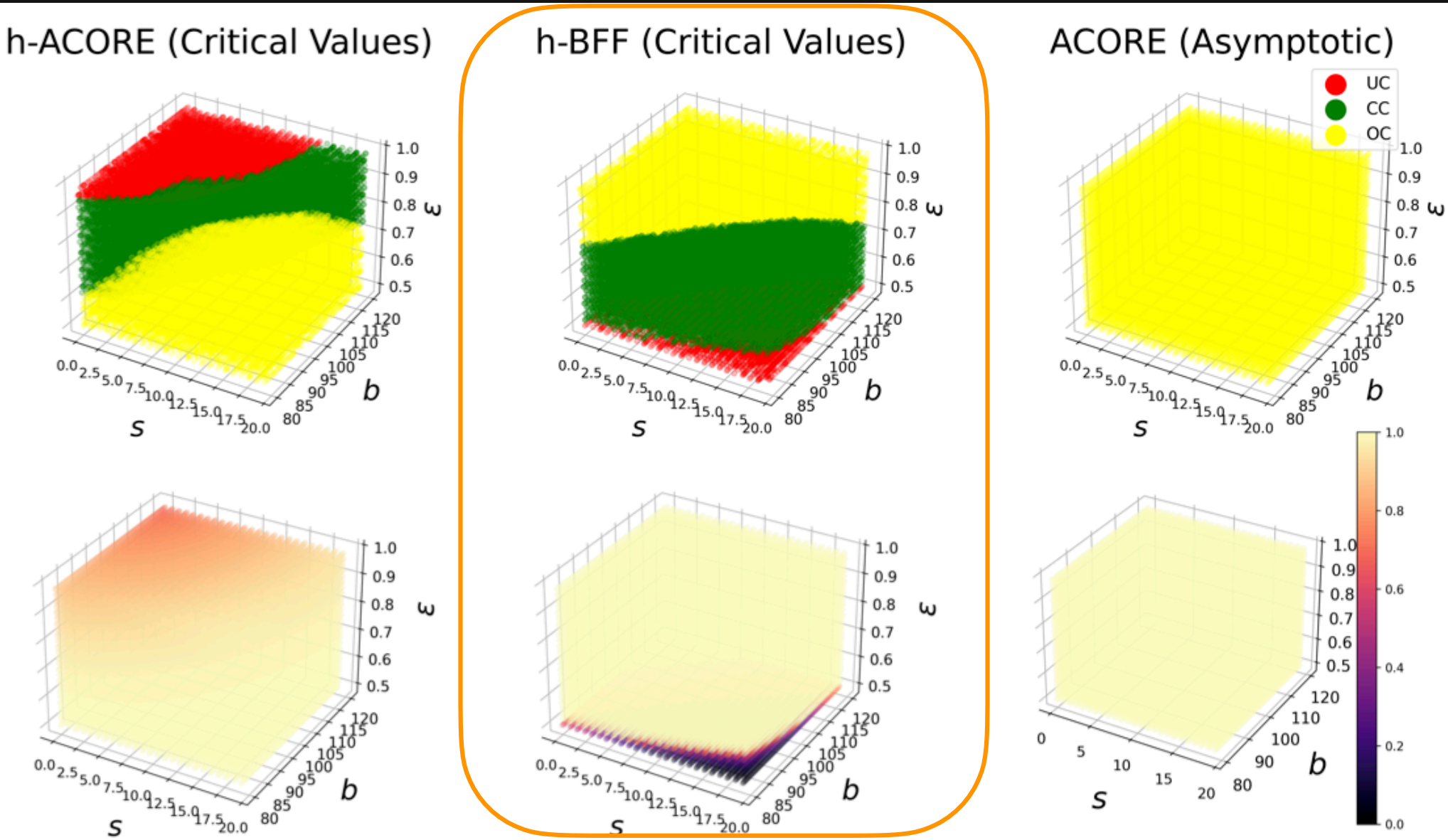
Diagnostic Tool to Check Coverage (UC, CC or OC)



- h-BFF (averages over nuisance parameters) performs the best in terms of having the largest proportion of the parameter space with CC and only a small fraction of the parameter space with UC

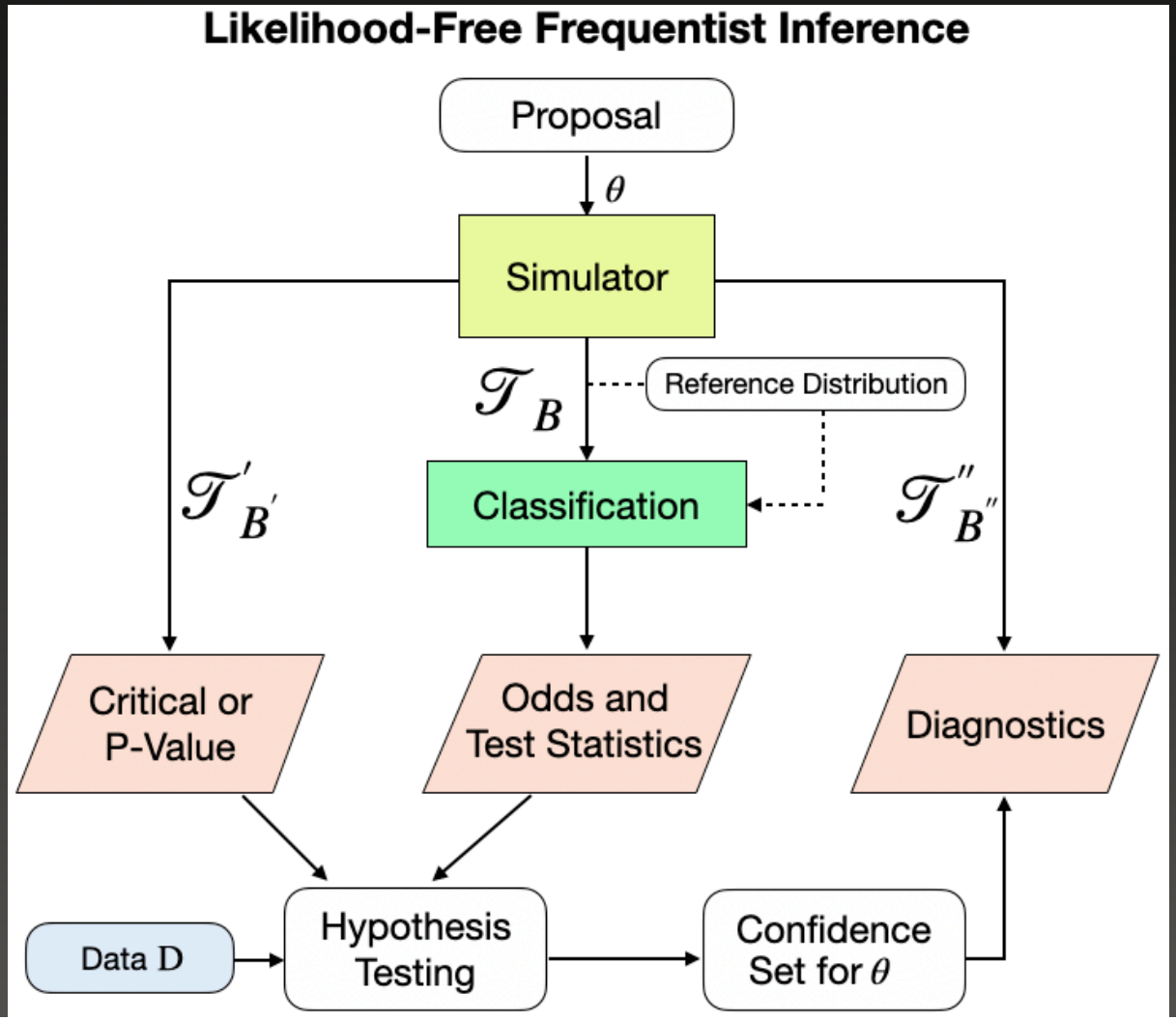
Our diagnostic tool can identify regions in parameter space with UC, CC and OC

(Bottom: heat maps of upper limit of 2**σ** prediction band)



Take-Away: Frequentist LFI (inverse problem)

- We can construct finite-sample confidence sets with nominal coverage, and provide diagnostics, even without a tractable likelihood. (Do not rely on asymptotic or costly MC samples)



Take-Away: Frequentist LFI (inverse problem)

- Validity: Any existing or new test statistic — that is, not only estimates of the LR statistic — can be used in our framework to create tests that control type I error.
  - Implicit assumption is that the null distribution of the test statistic varies smoothly in parameter space.
- Nuisance parameters and diagnostics: No guarantee that hybrid methods are valid. However, we have a practical tool for assessing empirical coverage across the entire parameter space.
- Power: Hardest to achieve in practice for LFI. Area where most statistical and computational advances will take place.

Collaborators

- Nic Dalmasso (JP Morgan AI)
- Rafael Izbicki (UFSCar)
- David Zhao (CMU)
- Luca Masserano (CMU)
- Mikael Kuusela (CMU)
- Tommaso Dorigo (INFN/Padova)

EXTRA SLIDES START HERE

For BFF confidence sets, we can analyze the power further for a special case...

Suppose

- Simple null hypotheses, $\Theta_0 = \{\theta_0\}$
- $\mathbf{X}^{\text{obs}} = \mathcal{D}$; i.e., $\mathbf{X}^{\text{obs}}$ contains all observations
- $G(\mathbf{x})$ is the marginal distribution of $F_\theta(\mathbf{x})$ w.r.t. $\pi(\theta)$

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

$$= \mathbb{O}(\mathbf{X}_i^{\text{obs}}; \theta_0)$$

We can then relate the power of BFF to an integrated odds loss:

$$\mathcal{L}(\widehat{\mathbb{O}}, \mathbb{O}) := \int \left(\widehat{\mathbb{O}}(\mathbf{X}; \theta) - \mathbb{O}(\mathbf{X}; \theta)\right)^2 dg(\mathbf{X}) d\pi(\theta).$$

Power of BFF (cont'd)

Theorem 3 *Let* $\phi_\tau(\mathcal{D}) = \mathbb{I}(\tau(\mathcal{D};\theta_0) < c)$ *and* $\phi_{\widehat{\tau}_B}(\mathcal{D}) = \mathbb{I}(\widehat{\tau}_B(\mathcal{D};\theta_0) < c)$ *be the testing procedures for testing* $H_0 : \theta = \theta_0$ *obtained using* τ *and* $\widehat{\tau}_B$*. Under Assumptions 3-5, there exists* $K' > 0$ *such that, for every* $0 < \epsilon < 1$*,*

$$\mathbb{P}_{\mathcal{D}|\theta,T_B}(\phi_\tau(\mathcal{D}) \neq \phi_{\widehat{\tau}_B}(\mathcal{D})) \leq \frac{K' \cdot \sqrt{L(\widehat{\mathbb{O}},\mathbb{O})}}{\epsilon} + \epsilon$$

- The probability that hypothesis tests based on the Bayes factor versus the BFF statistic lead to different conclusions is bounded by the integrated odds (which is easy to estimate in practice and also depends on the choice of probabilistic classifier)

Power of BFF (cont'd)

Theorem 3 *Let* $\phi_\tau(\mathcal{D}) = \mathbb{I}(\tau(\mathcal{D};\theta_0) < c)$ *and* $\phi_{\widehat{\tau}_B}(\mathcal{D}) = \mathbb{I}(\widehat{\tau}_B(\mathcal{D};\theta_0) < c)$ *be the testing procedures for testing* $H_0 : \theta = \theta_0$ *obtained using* τ *and* $\widehat{\tau}_B$*. Under Assumptions 3-5, there exists* $K' > 0$ *such that, for every* $0 < \epsilon < 1$,

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

Assumption 6 (Convergence rate of the probabilistic classifier) *The probabilistic classifier trained with* $\mathcal{T}_B$, $\widehat{\mathbb{P}}(Y = 1|\mathbf{x}, \theta)$ *is such that*

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

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

Theorem 4 *Under Assumptions 3-7, there exists* $K'' > 0$ *such that*

$$\mathbb{P}_{\mathcal{D},\mathcal{T}_B|\theta}(\phi_\tau(\mathcal{D}) \neq \phi_{\widehat{\tau}_B}(\mathcal{D})) \leq 2\sqrt{K''}B^{-\alpha/(4(\alpha+d))}.$$

Toy Example: Signal Detection in Counting Experiment

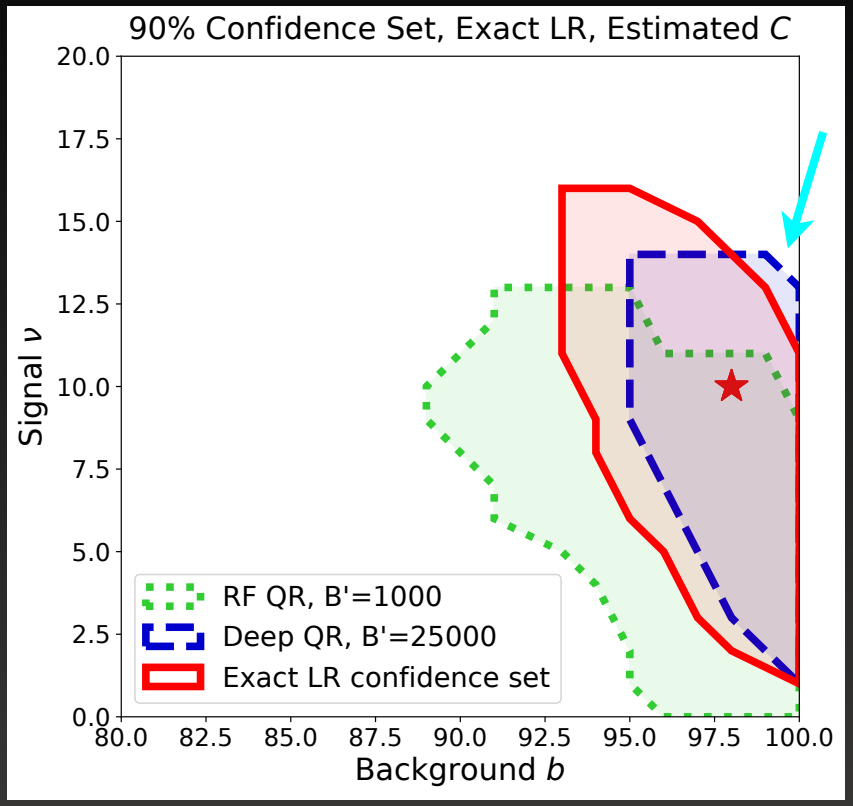
- Particle collision events counted under the presence of a background process.

$$\text{Observed data } D = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{10})$$
$$\mathbf{X} = (N, M), \text{ where } N \sim \text{Poisson}(b + \nu),\ M \sim \text{Poisson}(b)$$

- The observed data D consist of n=10 realizations of X=(N,M), where
  - N is the number of events in the signal region,
  - M is the number of events in the background/control region
- Unknown parameters:
  - intensity of signal (ν); intensity of background (b)

Constructed Confidence Set for a Particular X^{obs}

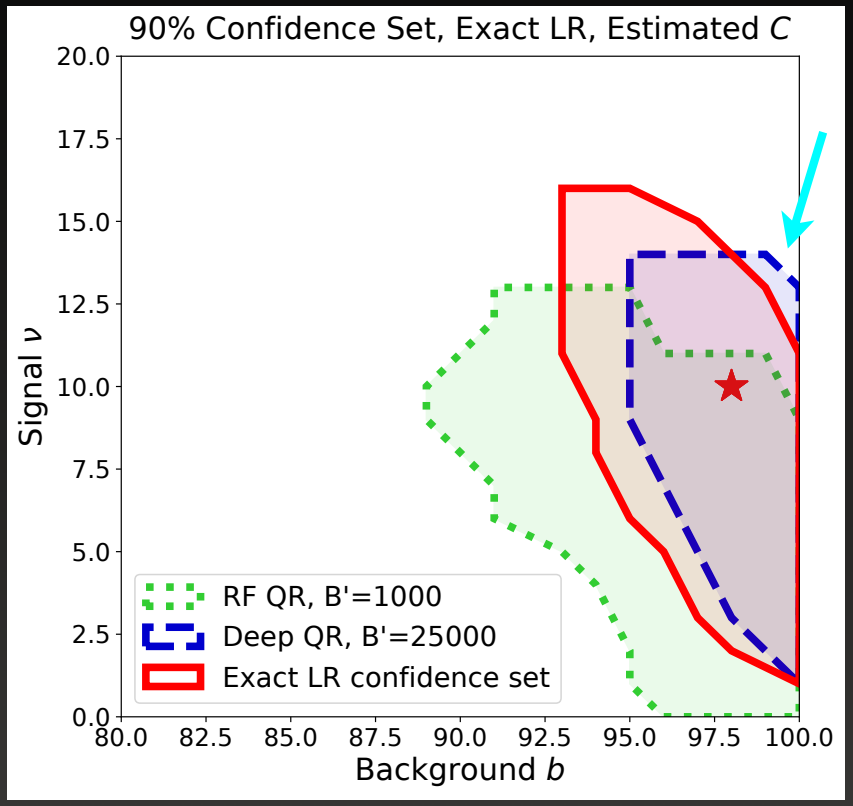
Our proposed strategy selects the BLUE confidence region

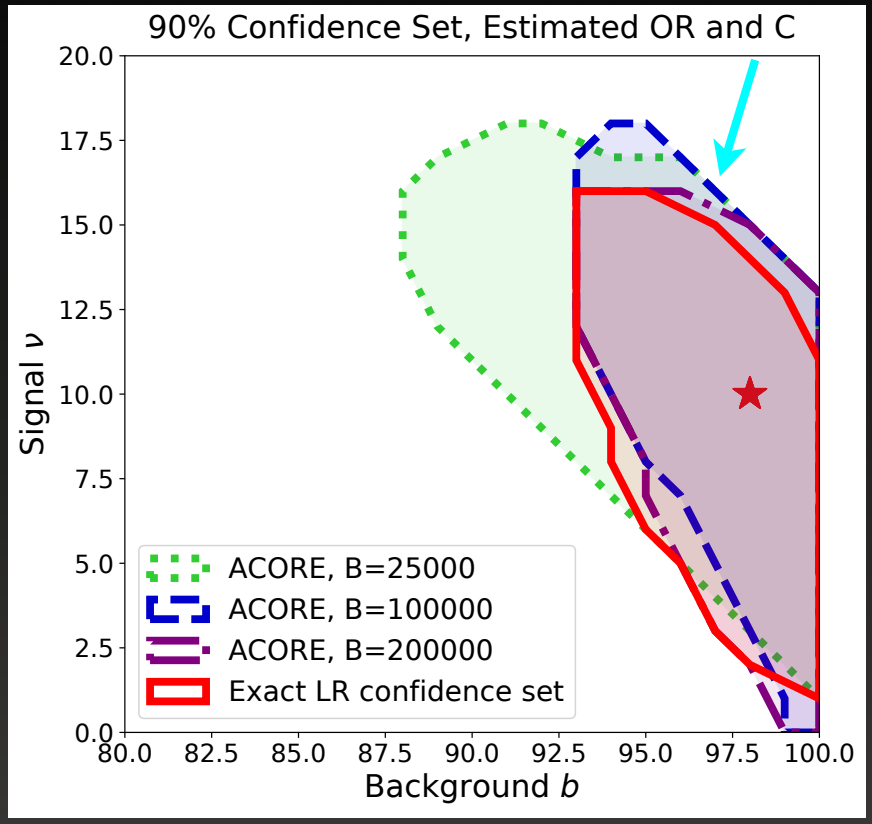


- Left: 90% confidence set computed with the exact LR statistic but estimated critical value. Estimating C can be challenging.
- Right: 90% confidence set with both estimated LR statistic and critical value. This is the true LFI setting.

Constructed Confidence Set for a Particular X^{obs}

Our proposed strategy selects the BLUE confidence region

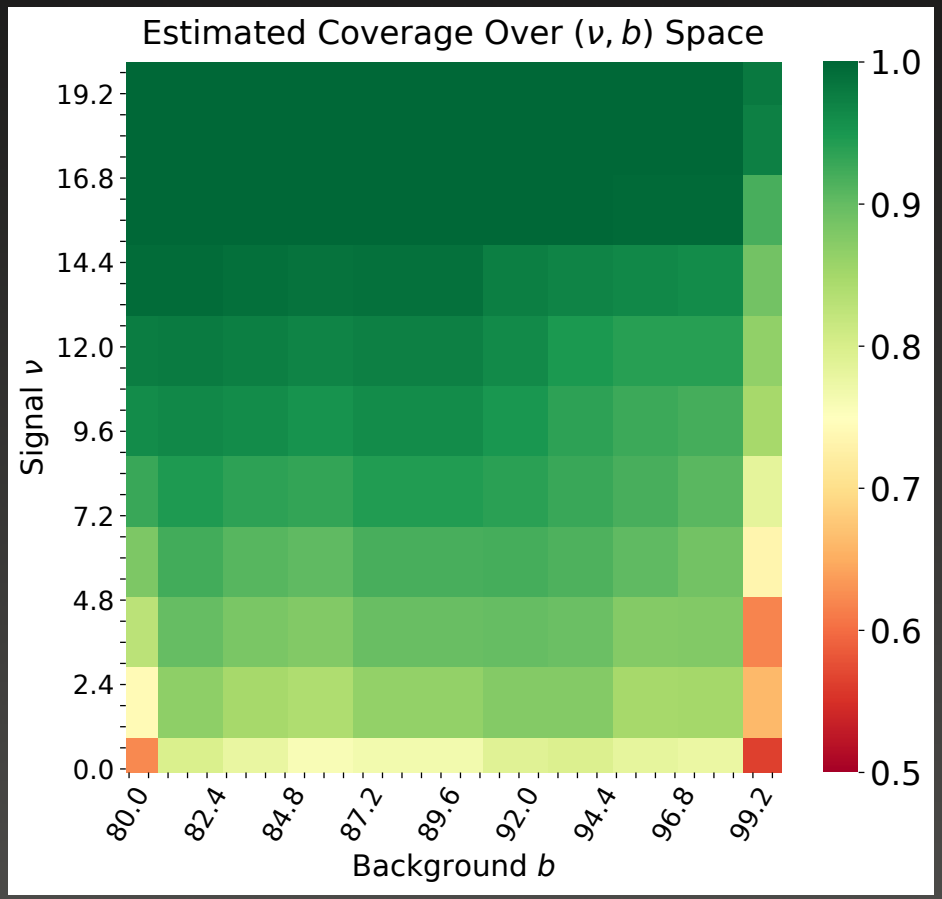




- Left: 90% confidence set computed with the exact LR statistic but estimated critical value. Estimating C can be challenging.
- Right: 90% confidence set with both estimated LR statistic and critical value. This is the true LFI setting.

Diagnostics: Do We Achieve Nominal Coverage (Type I Error Control) Across the Parameter Space?

$$\mathbb{P}[\theta_0 \in R(\mathcal{D}) \mid \theta = \theta_0] \geq 1 - \alpha \quad \text{for all } \theta_0 \in \Theta$$

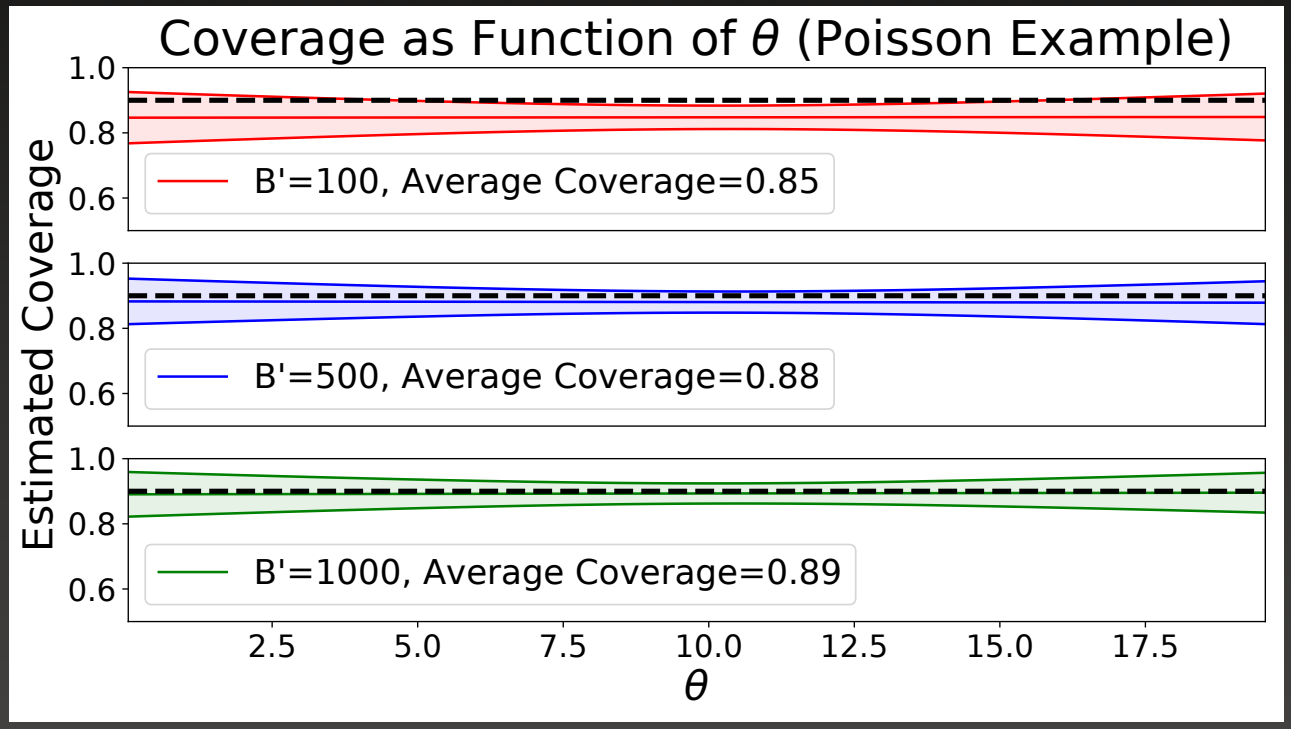


Heat map of estimated coverage for a confidence set that did not pass our goodness-of-fit diagnostic

- Overall coverage of confidence set is correct (92% vs the 90% nominal coverage)
- However, the set undercovers in low-signal and high-background regions.

Diagnostics: Do We Achieve Nominal Coverage (Type I Error Control) Across the Parameter Space?

$$\mathbb{P}[\theta_0 \in R(\mathcal{D}) \mid \theta = \theta_0] \geq 1 - \alpha \quad \text{for all } \theta_0 \in \Theta$$



Estimated coverage with prediction intervals for example with just one parameter. Forward model is a Poisson(100+θ), and n=10 obs.