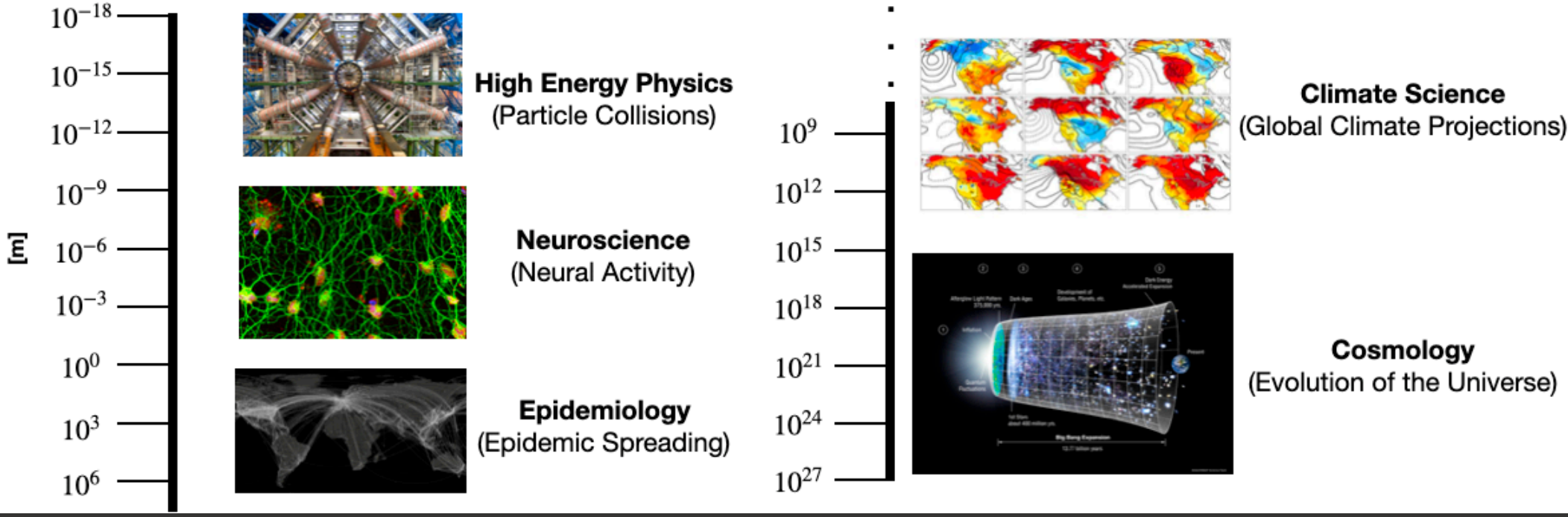
Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage

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

Collaborators: Nic Dalmasso (JP Morgan); Luca Masserano (CMU); Tommaso Dorigo (Padova); Rafael Izbicki (UFSCar); Mikael Kuusela (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

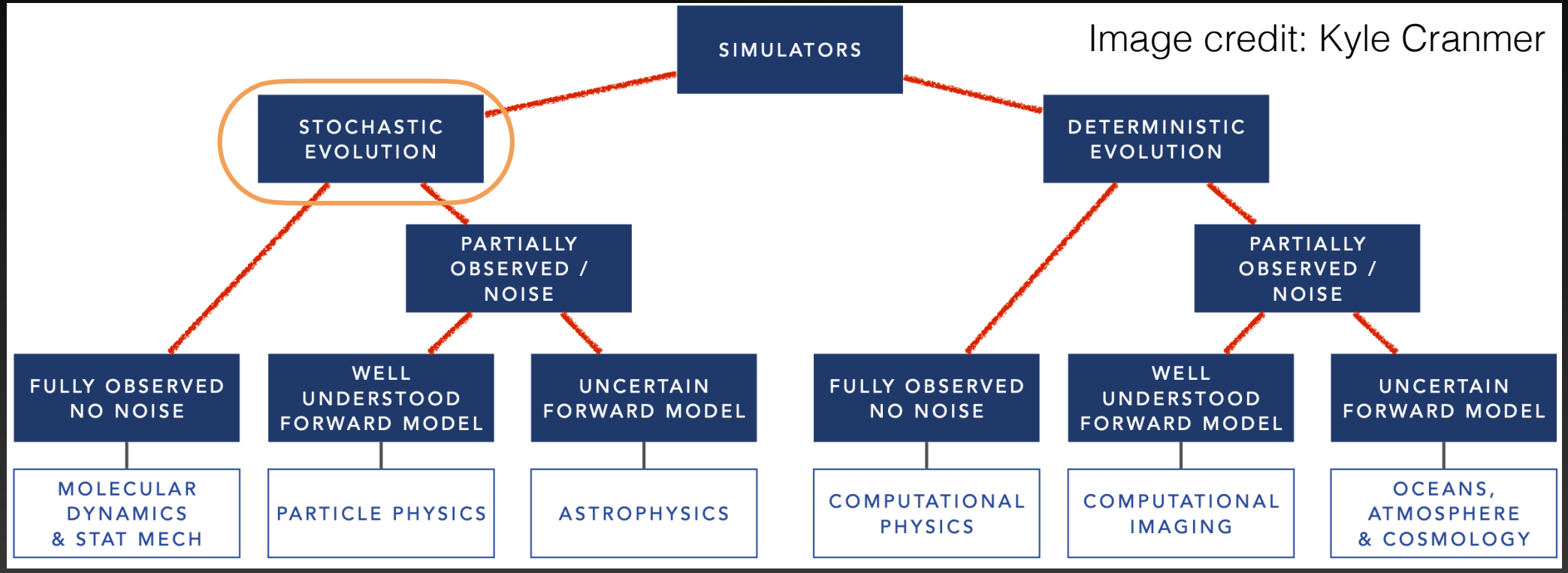


Image credit: Kyle Cranmer

- 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.

Likelihood-Based Inference

Likelihood-Free Inference (LFI)

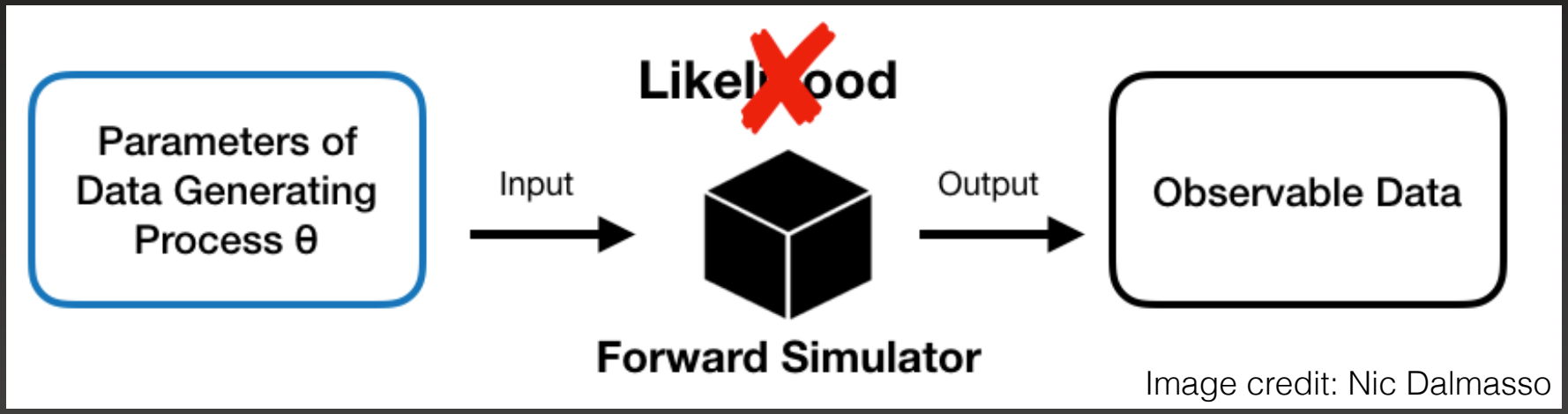


Image credit: Nic Dalmasso

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

Classical LFI: Approximate Bayesian Computation (ABC)

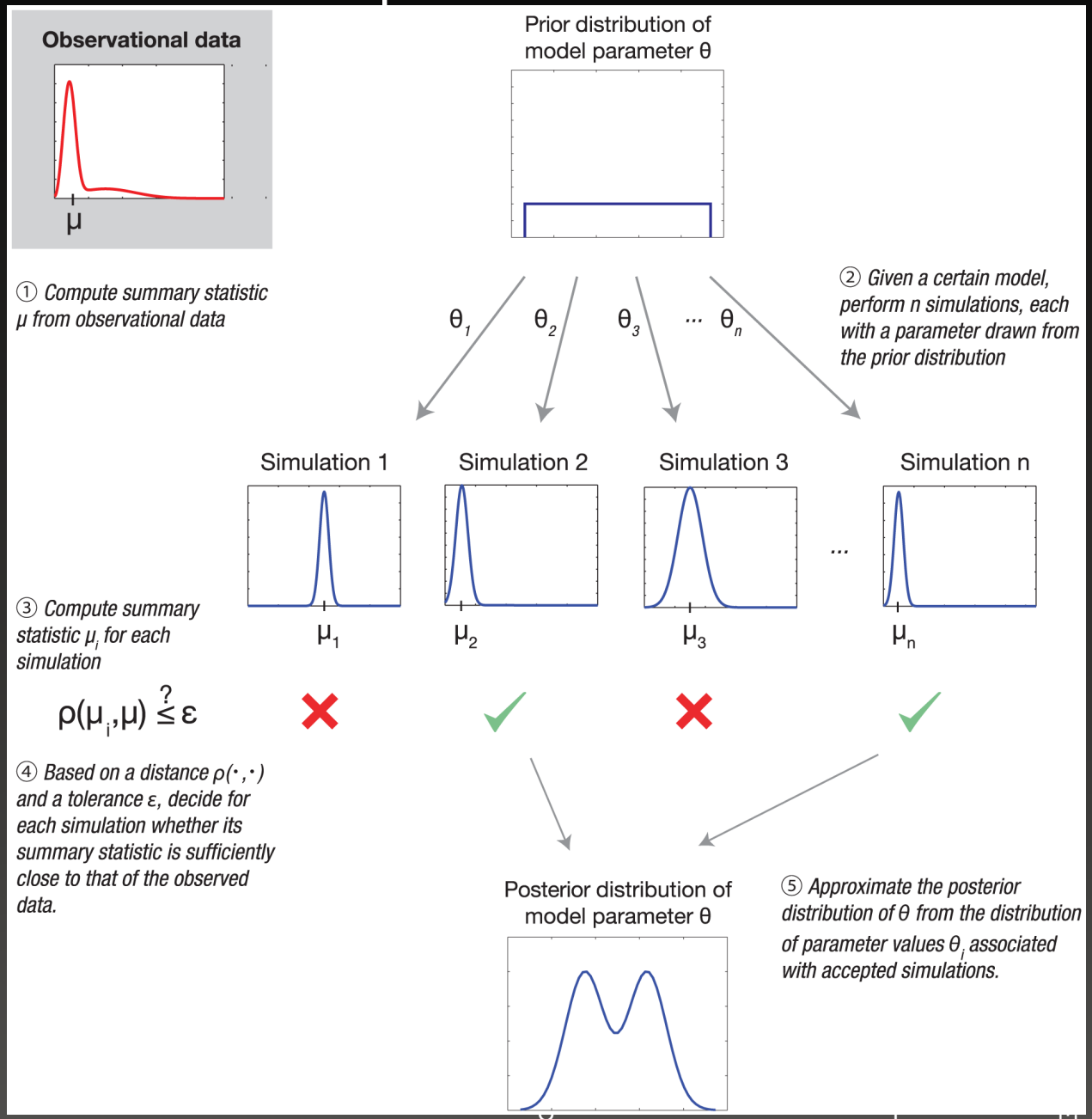


Image credit: Sunnaker et al. 2013

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]

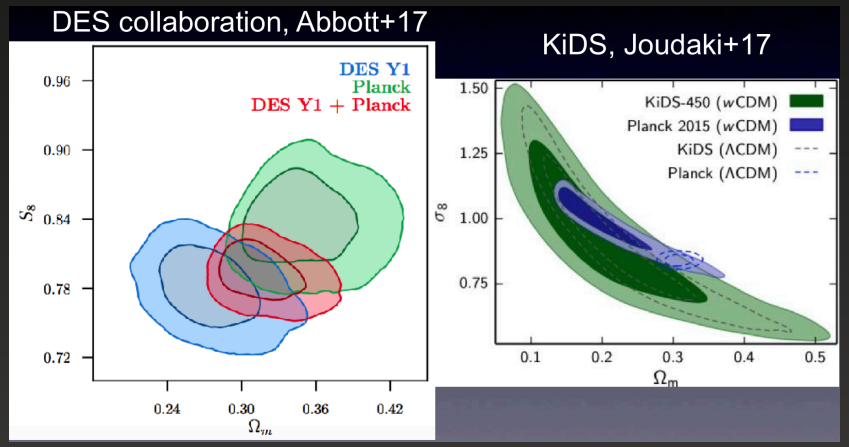
- These new training-based approaches provide "amortized" inference. Can handle complex high-dimensional data without a prior dimension reduction.

So What's Missing in the LFI-ML Literature?

- Shortage of practical inferential and diagnostic tools with finite-sample guarantees of freq. coverage.

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

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

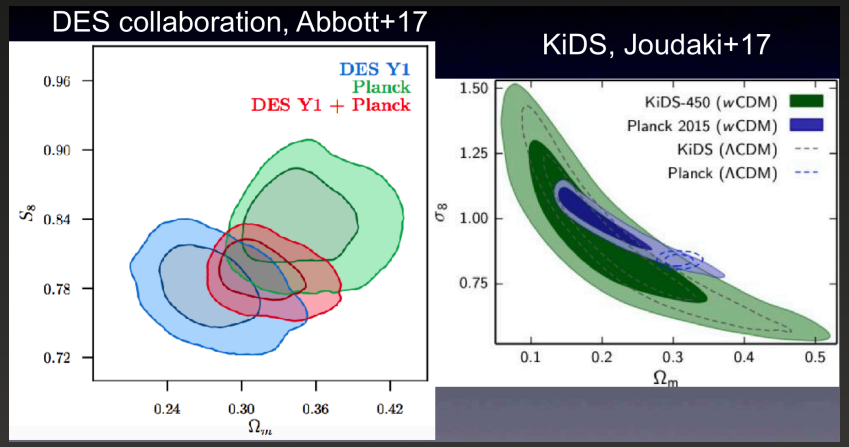


So What's Missing in the LFI-ML Literature?

- Shortage of practical inferential and diagnostic tools with finite-sample guarantees of freq. coverage.

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

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



$$\text{Simulate} \quad (\theta_1, \mathcal{D}_1), (\theta_2, \mathcal{D}_2), \ldots, (\theta_B, \mathcal{D}_B),$$
$$\text{where} \quad \theta_i \sim \pi(\theta), \ \mathcal{D}_i = \{\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n}\} \sim F_{\theta_i}$$

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.

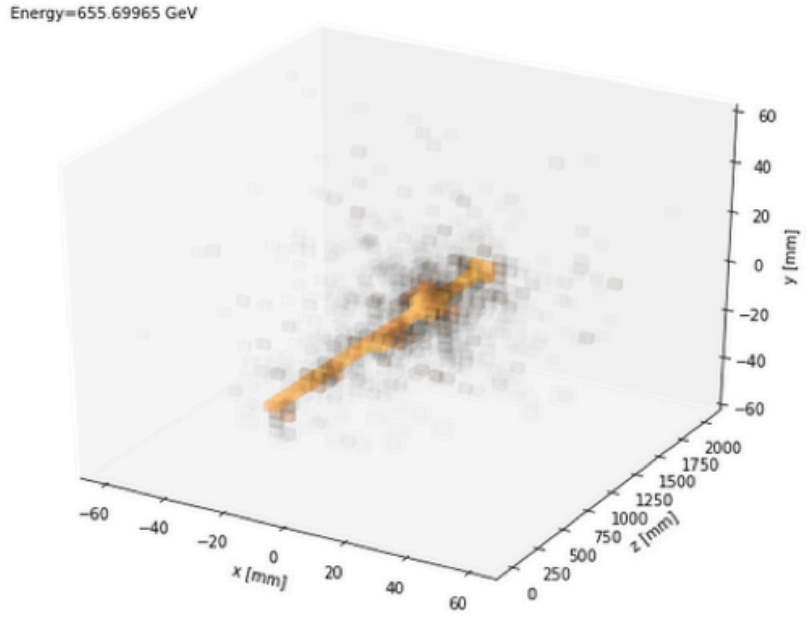


Figure 4: Muon entering the calorimeter in z direction.

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

1. Bias

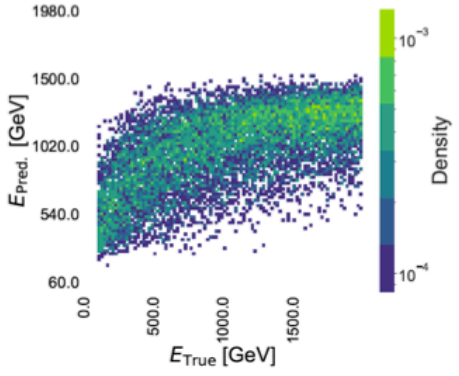


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

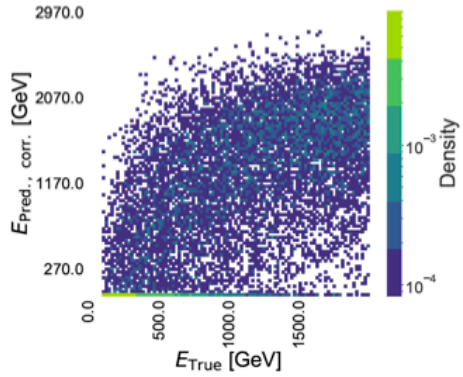
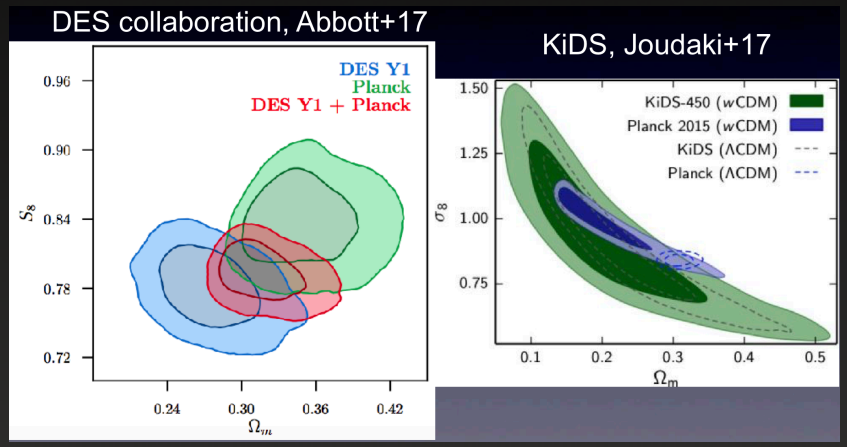


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

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

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

How about Frequentist LFI Approaches?



Confidence sets with correct conditional coverage (for small n)?

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

- Most approaches that estimate likelihoods or likelihood ratios
  - rely on asymptotic assumptions (Wilks 1938) for downstream inference
  - do not assess validity across entire parameter space, or
  - use costly MC simulations at fixed parameter settings on a grid

Unified Inference Machinery for Frequentist LFI

- Bridges ML with classical statistics to provide:

  (i) valid inference: confidence sets and 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 output from any prediction algorithm)

Preliminary work (Feb 2020) 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 (revised April 2022)
https://arxiv.org/abs/2107.03920

Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage

Niccolò Dalmasso[*†] NICCOLO.DALMASSO@GMAIL.COM

Luca Masserano[*‡] LMASSERA@ANDREW.CMU.EDU

David Zhao[†] DAVIDZHAO@CMU.EDU

Rafael Izbicki[§] RAFAELIZBICKI@GMAIL.COM

Ann B. Lee [†¶] ANNLEE@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, particularly outside 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 confidence sets with correct conditional coverage for small sample sizes. This paper unifies classical statistics with modern machine learning to present (i) a practical procedure for the Neyman construction of confidence sets with finite-sample guarantees of nominal coverage, and (ii) diagnostics that estimate conditional coverage over the entire parameter space. We refer to our framework as *likelihood-free frequentist inference* (LF2I). Any method that defines a test statistic, like the likelihood ratio, can leverage the LF2I machinery to create valid confidence sets and diagnostics without costly Monte Carlo samples at fixed parameter settings. We study the power of two test statistics (`ACORE` and `BFF`), which, respectively, maximize versus integrate an odds function over the parameter space. Our paper discusses the benefits and challenges of LF2I, with a

Equivalence of Tests and Confidence Sets

- Data $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_n\} \sim F_\theta$
- Test statistic $\lambda(\mathcal{D}; \theta)$
- Critical values

$$\text{Reject } H_0 : \theta = \theta_0 \iff \lambda(\mathcal{D}; \theta_0) < C_{\theta_0, \alpha}$$

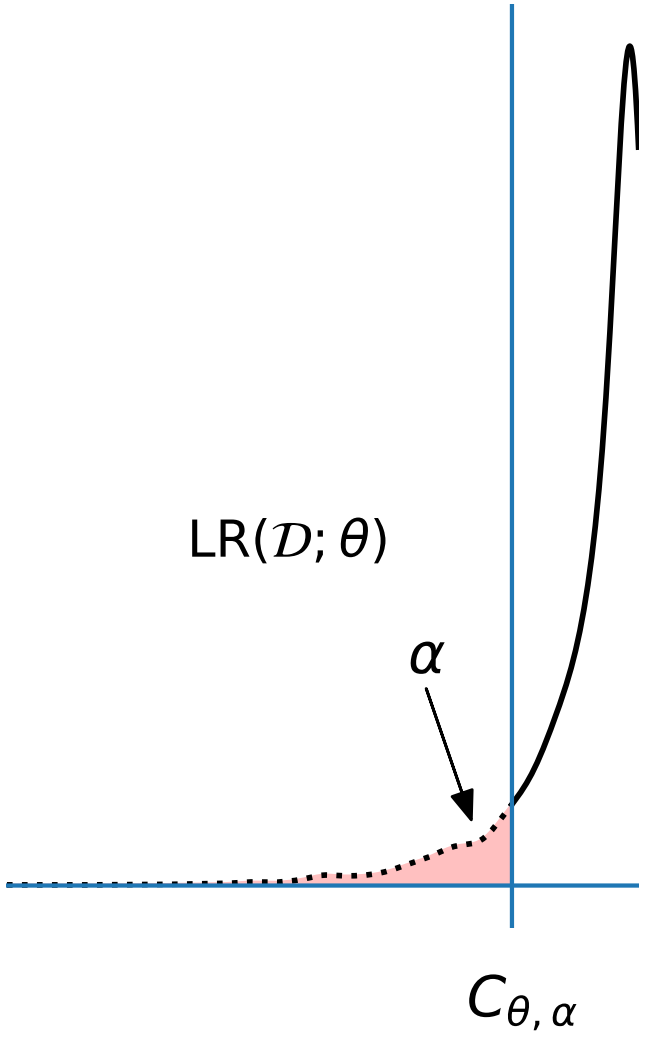
Theorem (Neyman 1937)

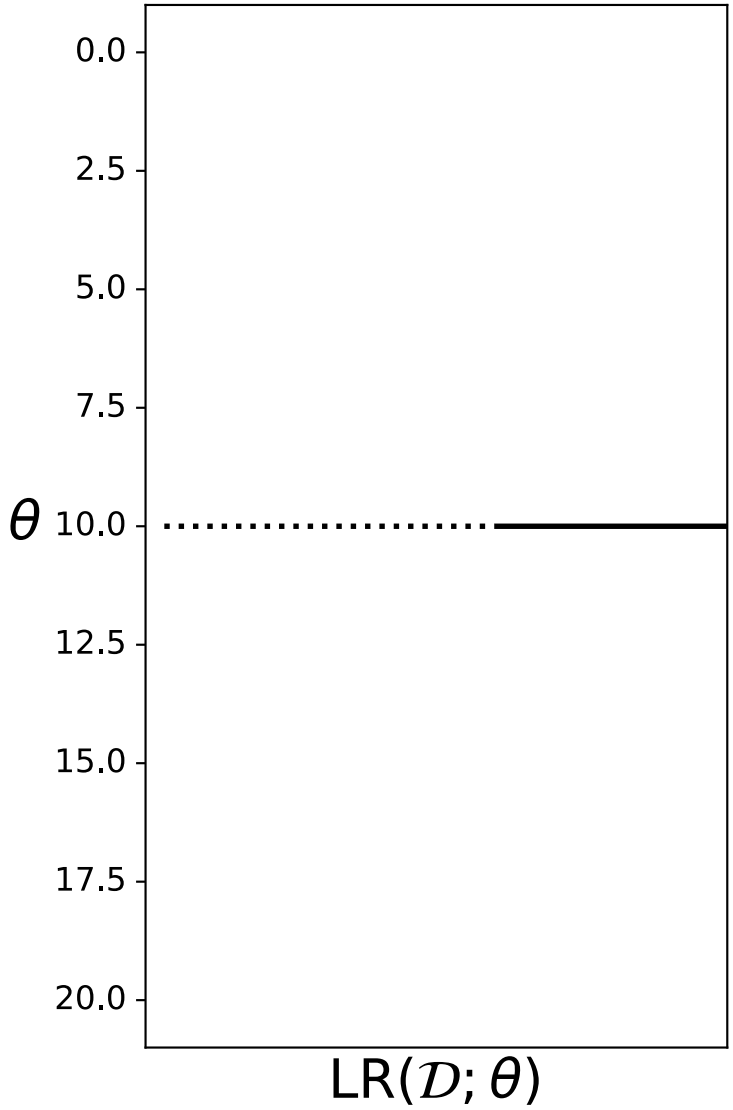
Constructing a $1 - \alpha$ confidence set for θ is equivalent to testing

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

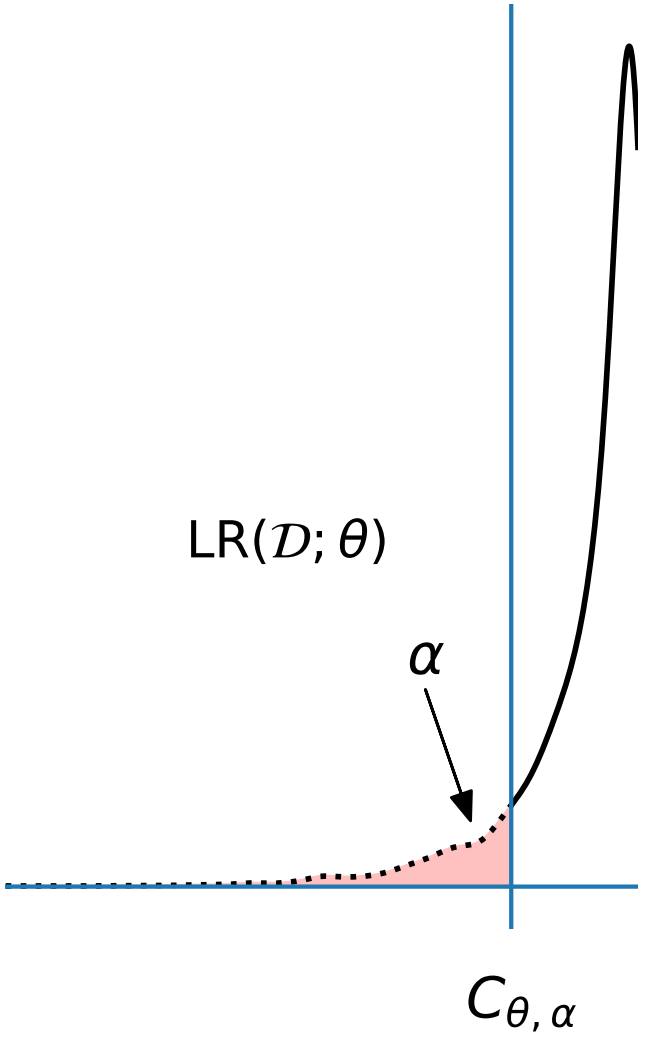
for every $\theta_0 \in \Theta$.

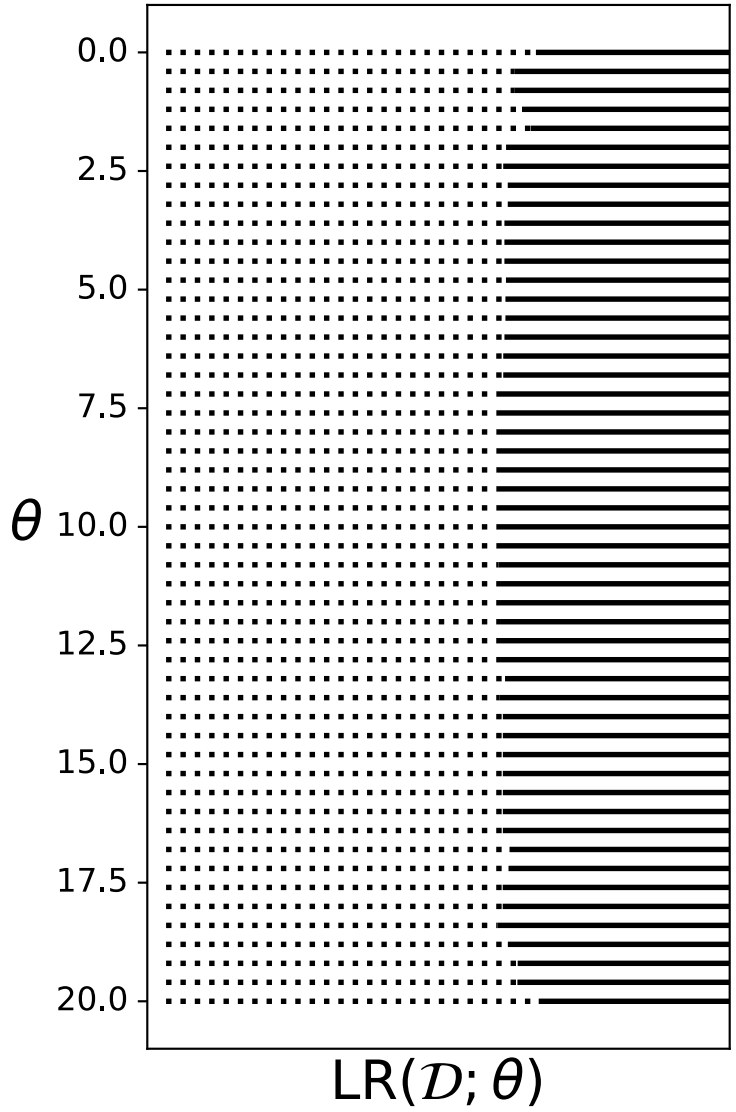
1. Fixed θ. Find the rejection region for test statistic λ.



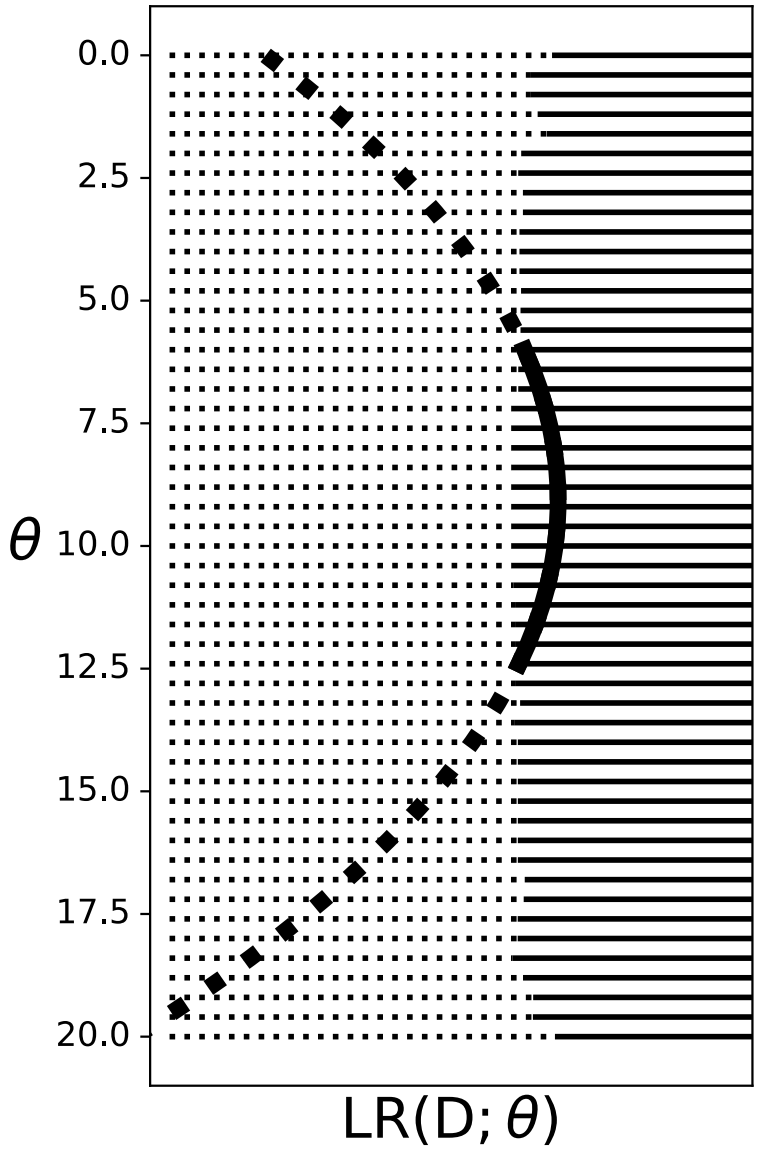


2. Repeat for every θ in parameter space.

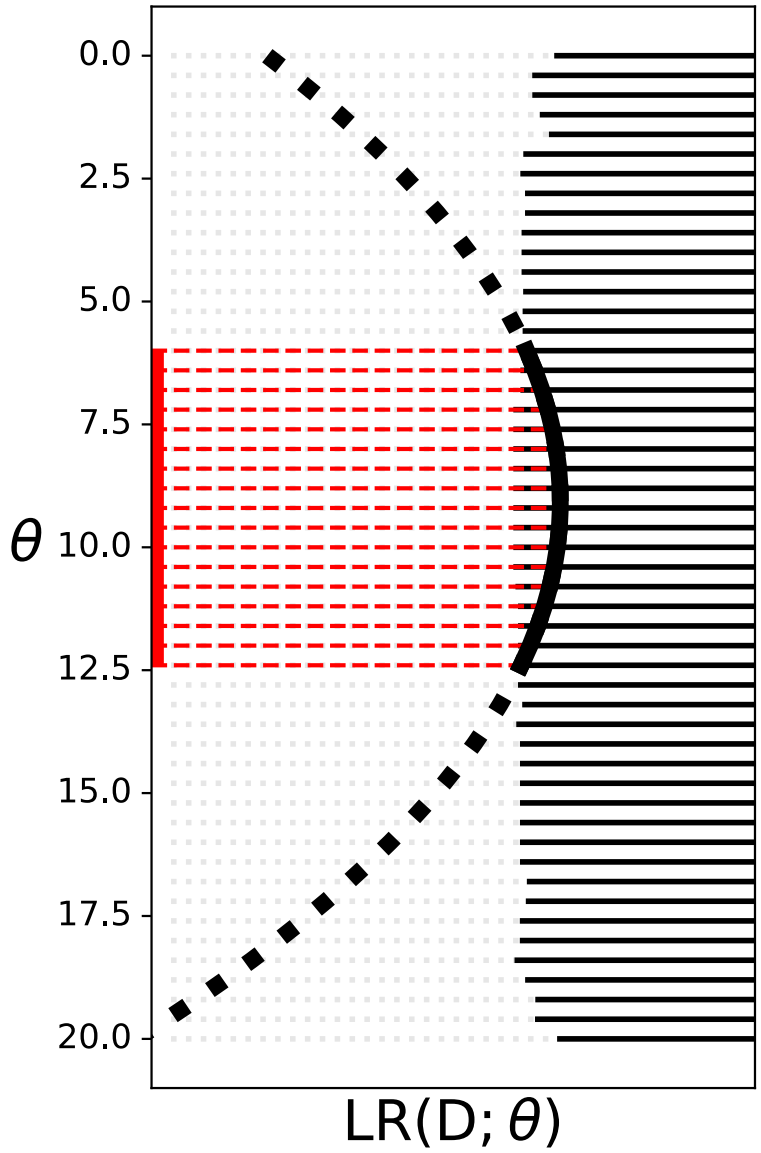




3. Observe data $\mathcal{D} = \mathrm{D}$. Evaluate $\lambda(\mathrm{D}; \theta)$.



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

- Validation 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 Neyman Construction and Validation into Practical Procedures?

The Neyman construction requires one to test

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

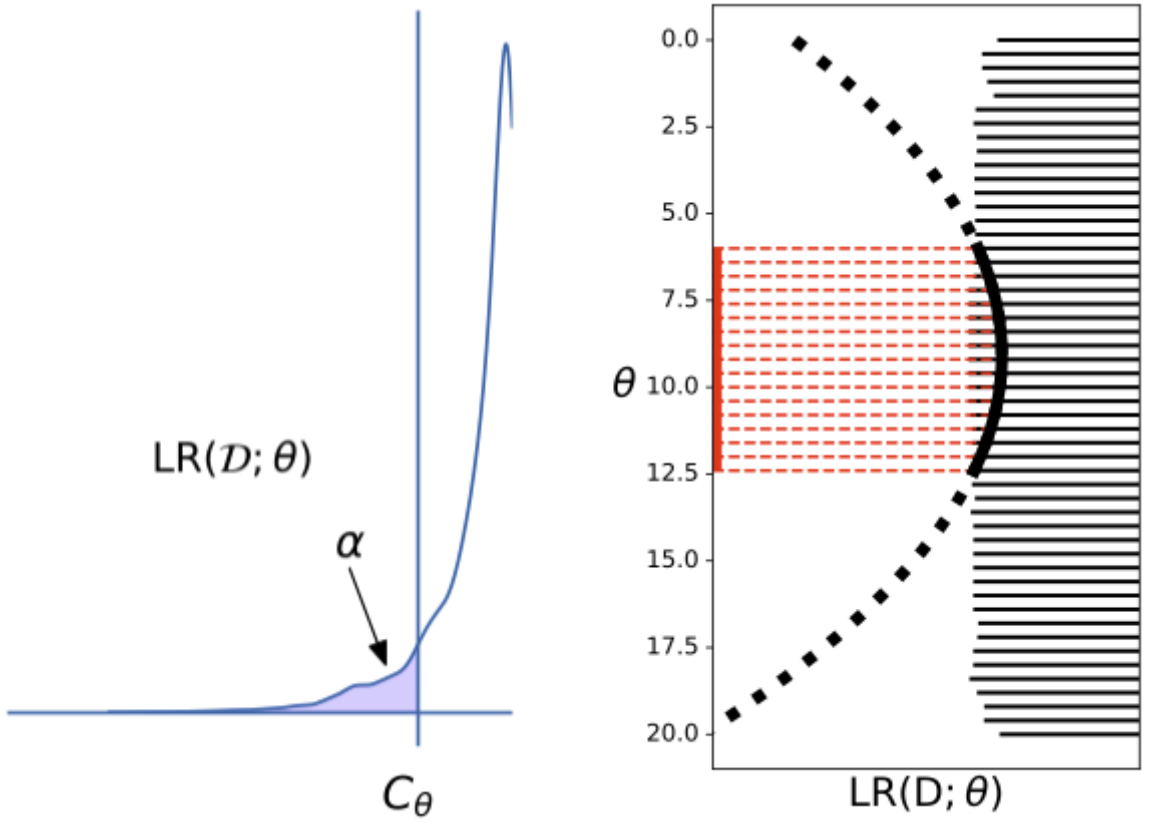
for **every** $\theta_0 \in \Theta$.

Key insight:

1. Test statistic $\lambda(\mathcal{D}; \theta)$
2. Critical values $C_{\theta_0,\alpha}$ or p-values $p(D; \theta_0)$ of the test
3. Coverage $\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D})\right)$ of the constructed confidence set

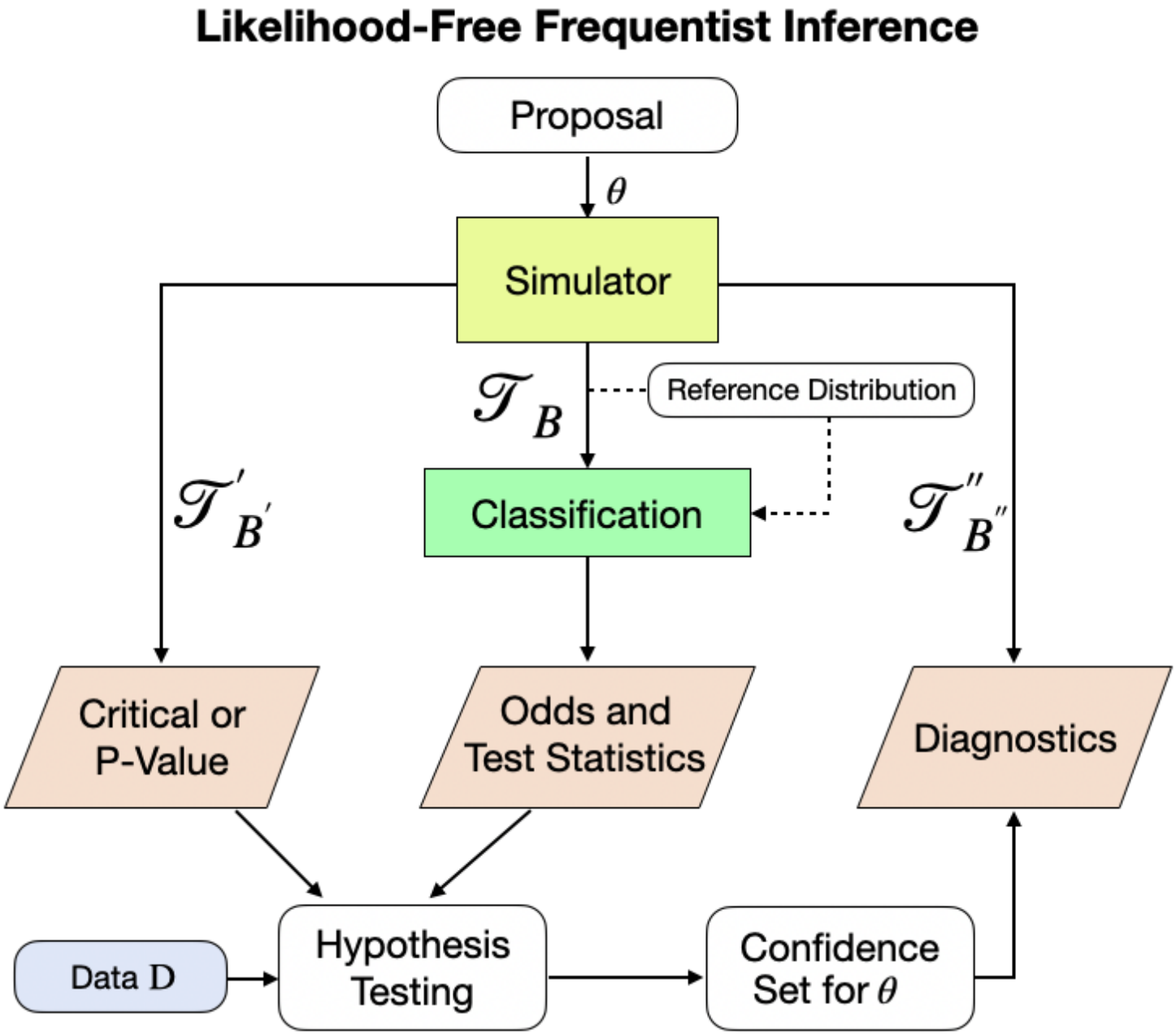
are **conditional distribution functions** of the (unknown) parameters, and often vary smoothly across the parameter space Θ.

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.

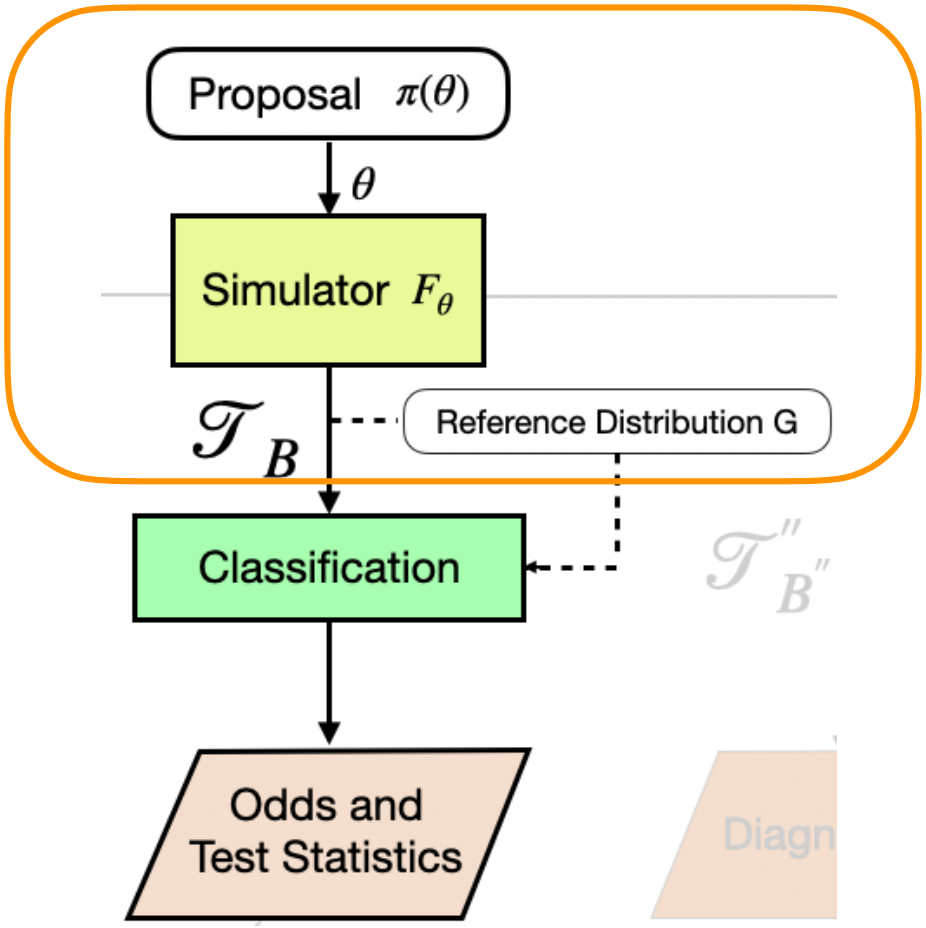
Our Inference Machinery



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.

Test Statistics Based on Odds: ACORE and BFF

Suppose we want to test:

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

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

- ACORE (Approximate Computation via Odds Ratio Estimation):

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

- BFF (Bayesian Frequentist Factor):

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

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):

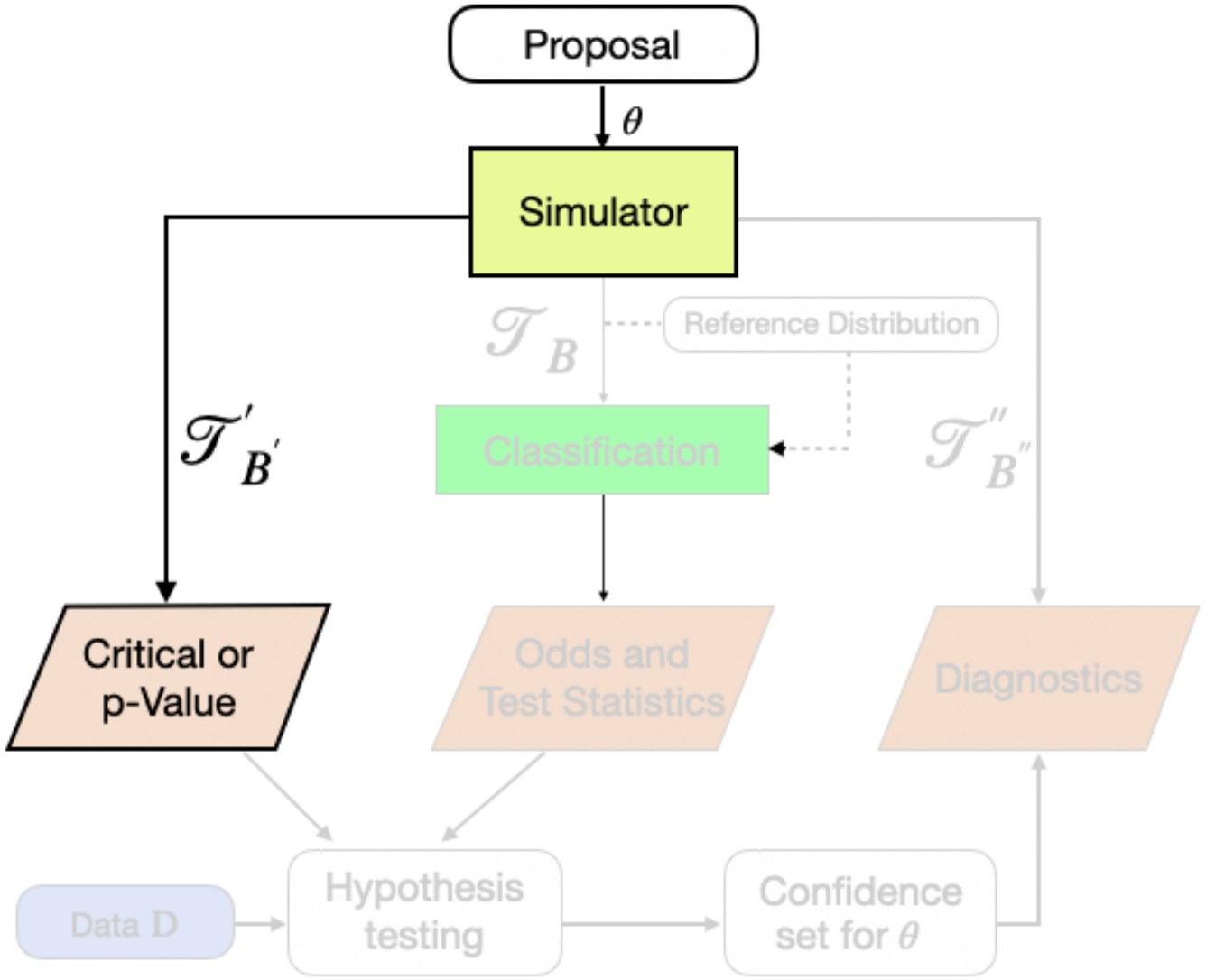
$$\widehat{\Lambda}(\mathcal{D}; \theta_0) := \log \frac{\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}$$

- 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.

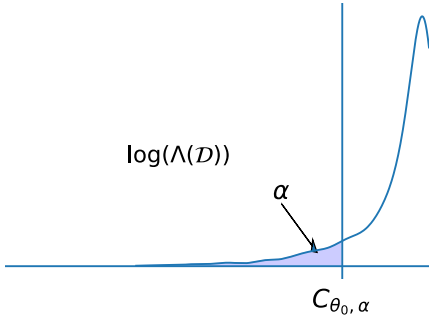
Left Branch: Estimate Critical Values or P-Values



We use B' simulations to estimate critical values.

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

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

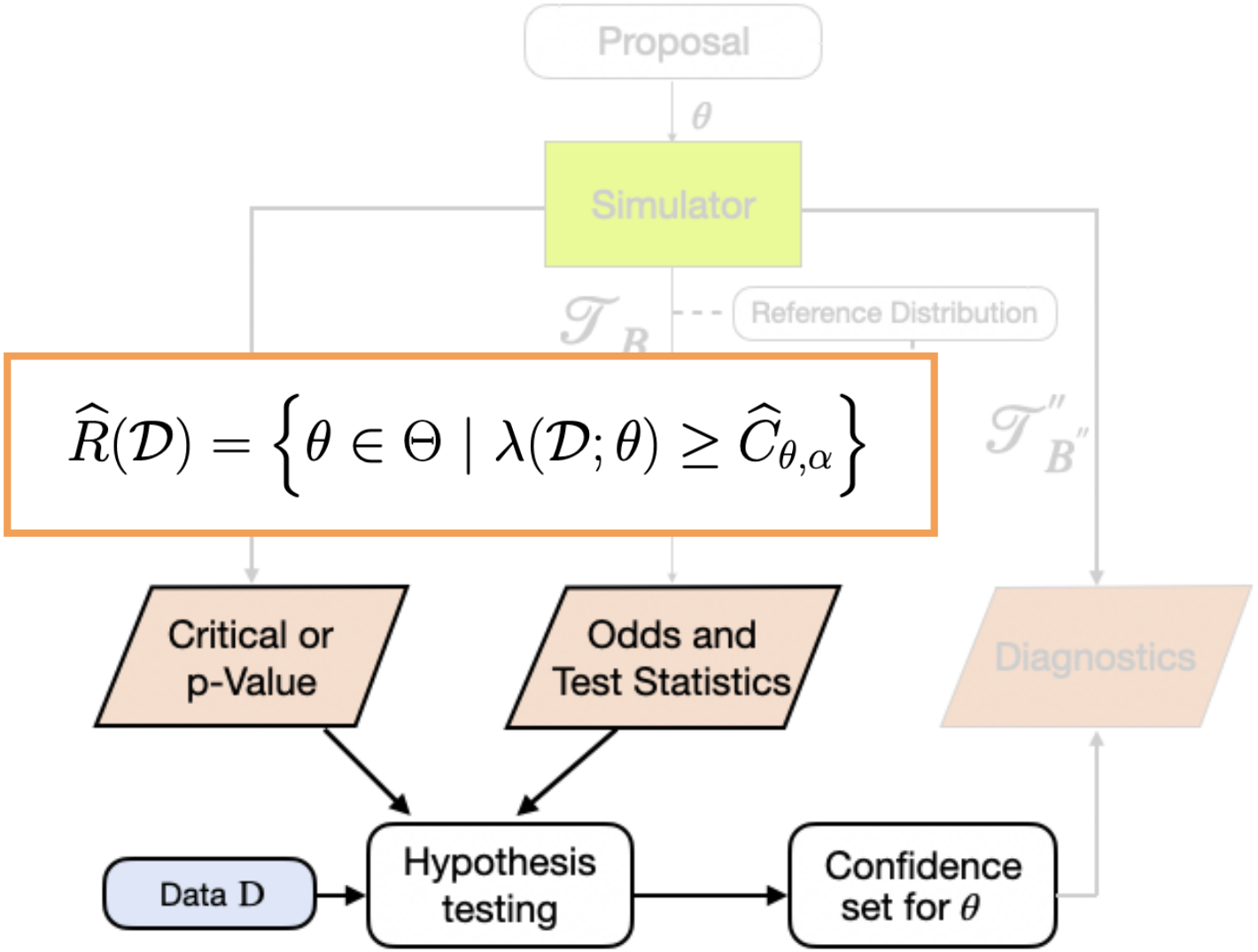


$$C_{\theta_0,\alpha} = \arg\sup_{C\in\mathbb{R}} \left\{ C : \mathbb{P}_{\mathcal{D}|\theta_0}\left(\lambda(\mathcal{D};\theta_0) < C\right) \leq \alpha \right\}.$$

Problem: Need to compute $\mathbb{P}_{\mathcal{D}|\theta}\left(\lambda(\mathcal{D};\theta) < C\right)$ for every $\theta \in \Theta$.

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

Construct Confidence Set via Neyman Inversion



Are the Constructed Confidence Sets Valid?

Theorem (Validity for any test statistic)

Let $C_{B'}$ be the critical value of a level-α test based on the statistic $\lambda(\mathcal{D}; \theta_0)$. Then, if the quantile regression estimator is consistent,

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

where C^* is such that

$$\mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \theta_0)) \leq C^*) = \alpha.$$

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* $\theta \in \Theta$*:*

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

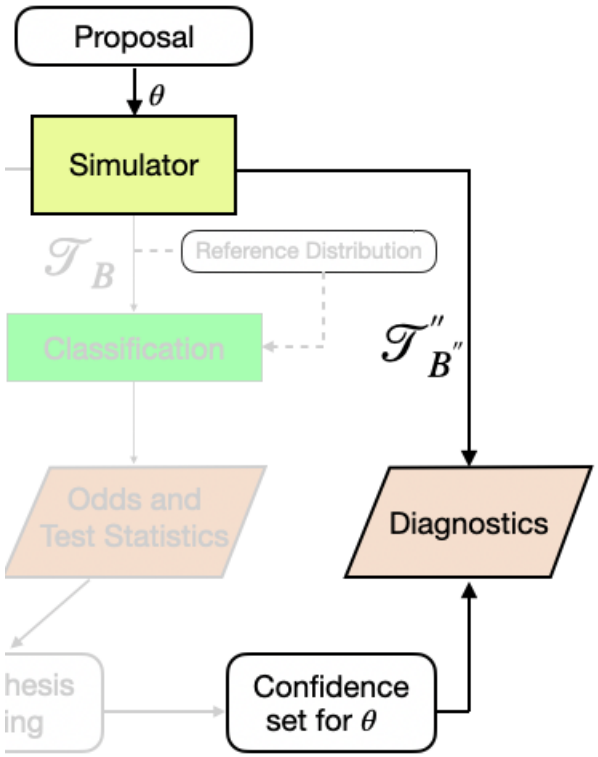
Right Branch: Assessing Conditional Coverage of $\widehat{R}(\mathcal{D})$

How do we check coverage of constructed confidence sets across Θ?

Note:

$$\widehat{R}(\mathcal{D}) = \left\{\theta \in \Theta \mid \lambda(\mathcal{D};\theta) \geq \widehat{C}_{\theta,\alpha}\right\}$$

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \widehat{R}(\mathcal{D}) \mid \theta\right) = \mathbb{E}_{\mathcal{D}|\theta}\left[\mathbb{I}\left(\theta \in \widehat{R}(\mathcal{D})\right) \mid \theta\right]$$

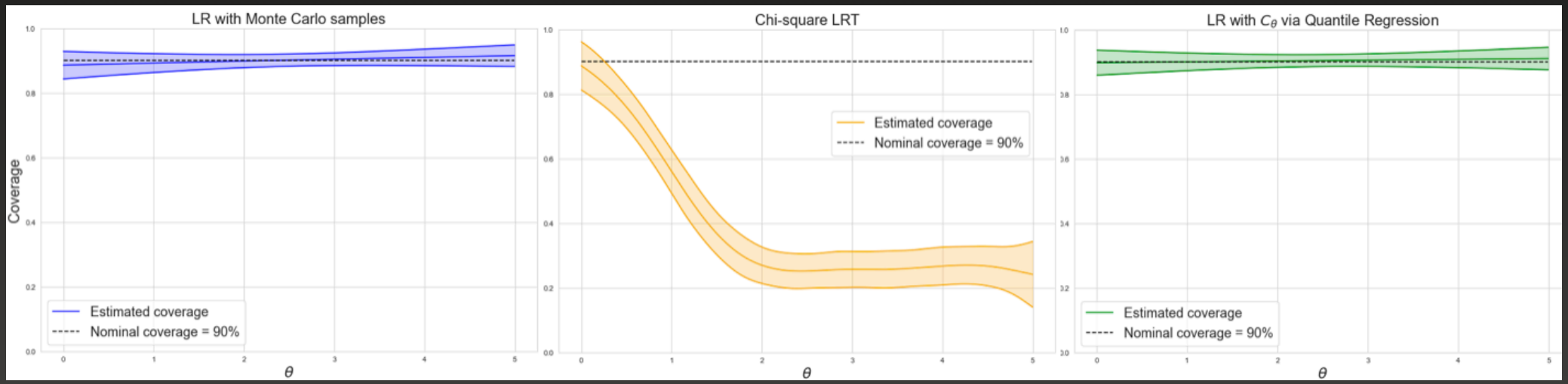


1. Sample θ_i and data $\mathcal{D}_i \sim F_{\theta_i}$
2. Construct confidence set $\widehat{R}(\mathcal{D}_i)$
3. For $\{\theta_i, \widehat{R}(\mathcal{D}_i)\}_{i=1}^{B''}$, regress $Z_i := \mathbb{I}(\theta_i \in \widehat{R}(\mathcal{D}_i))$ on θ_i.

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

Ex: Estimate Critical Values (GMM) & Run Diagnostics Across the Parameter Space

$$X_1, \ldots, X_n \sim 0.5N(\theta, 1) + 0.5N(-\theta, 1)$$



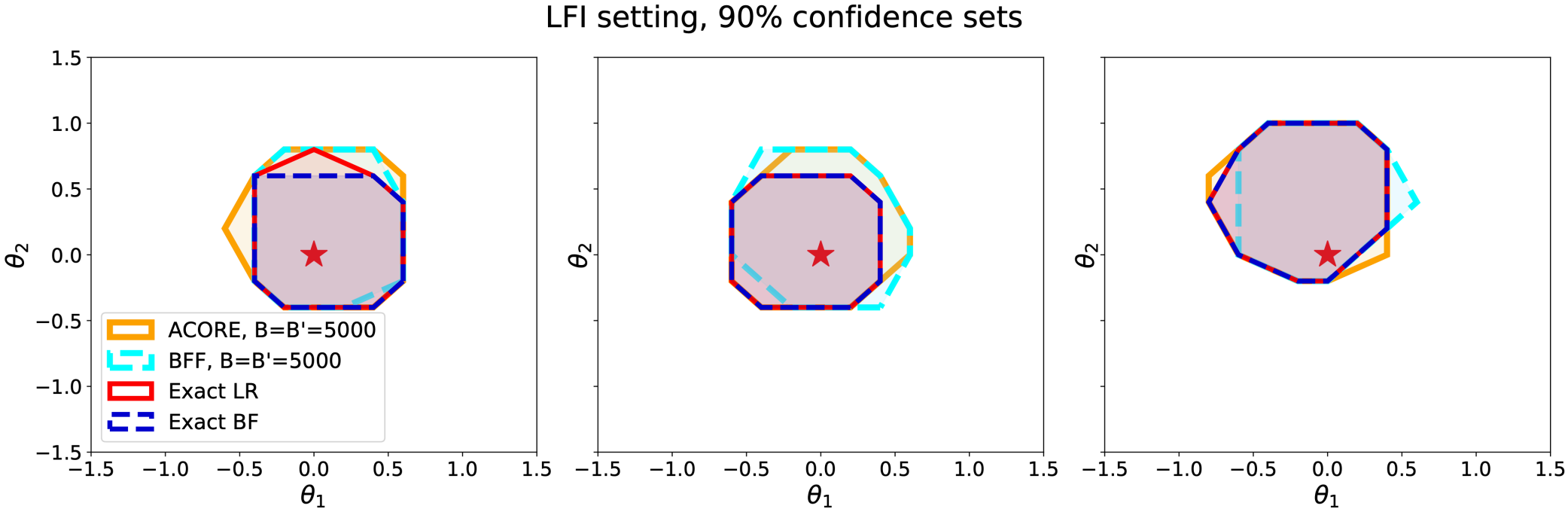
(Left) LR with 1000 MC simulations at each θ on a fine grid

(Center) Assume chi-squared distribution of LR statistic

(Right) LR with quantile regression with B'=1000 simulations total

Ex: Construct Confidence Sets (MVG data)

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N(\boldsymbol{\theta}, \mathbf{I}_d), \text{ where } n = 10, \ \boldsymbol{\theta} = \mathbf{0}$$

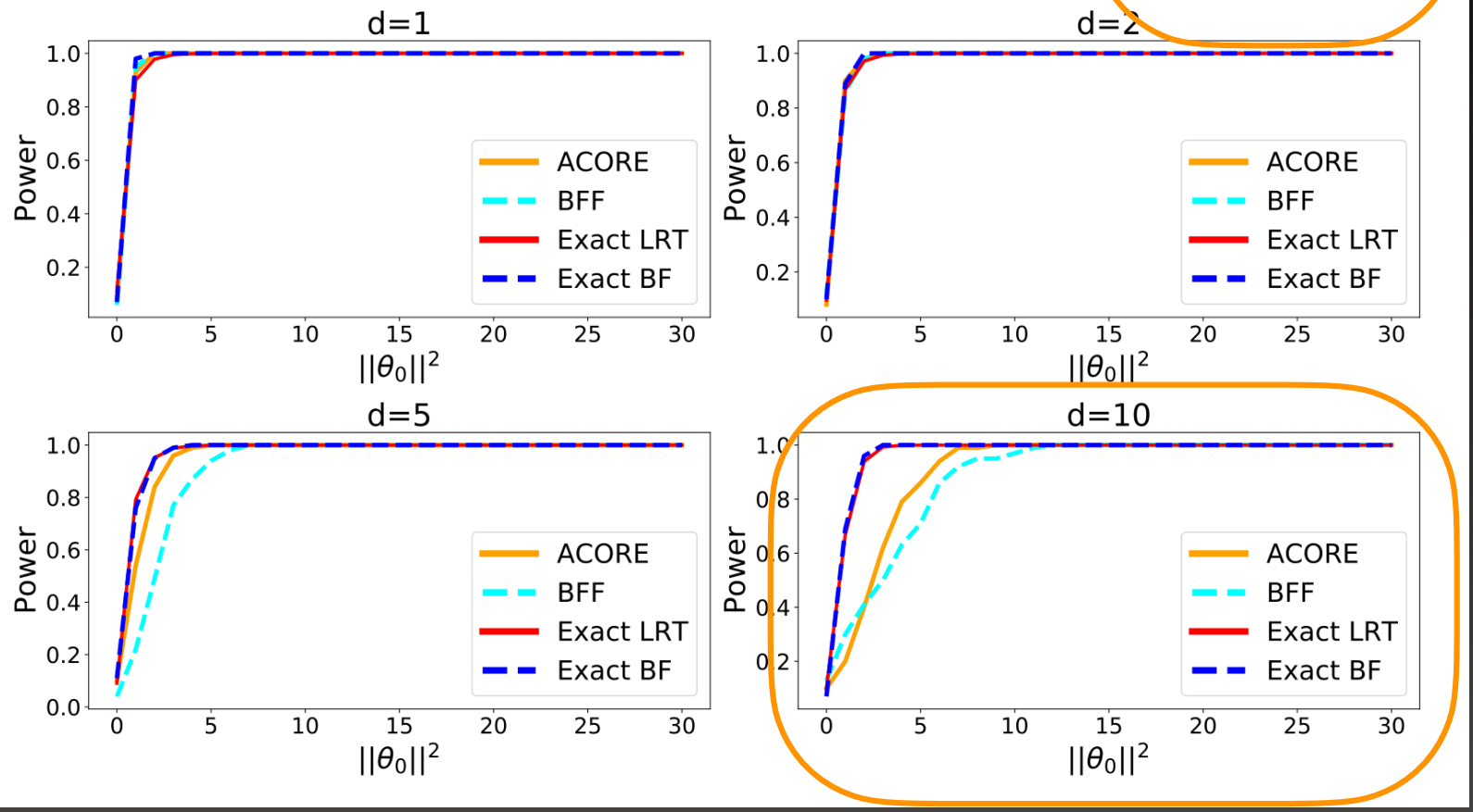


When d=2, ACORE and BFF confidence sets (for B=B'=5000) are similar in size to the Exact LR confidence sets.

Coverage and Power of LF2I Confidence Sets

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 LF2I

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

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

- `BFF` (Bayesian Frequentist Factor):

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

- 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 e3 (if B′ large enough, then e3≈0)

Break-Down of Sources of Errors in LF2I

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

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

- `BFF` (Bayesian Frequentist Factor):

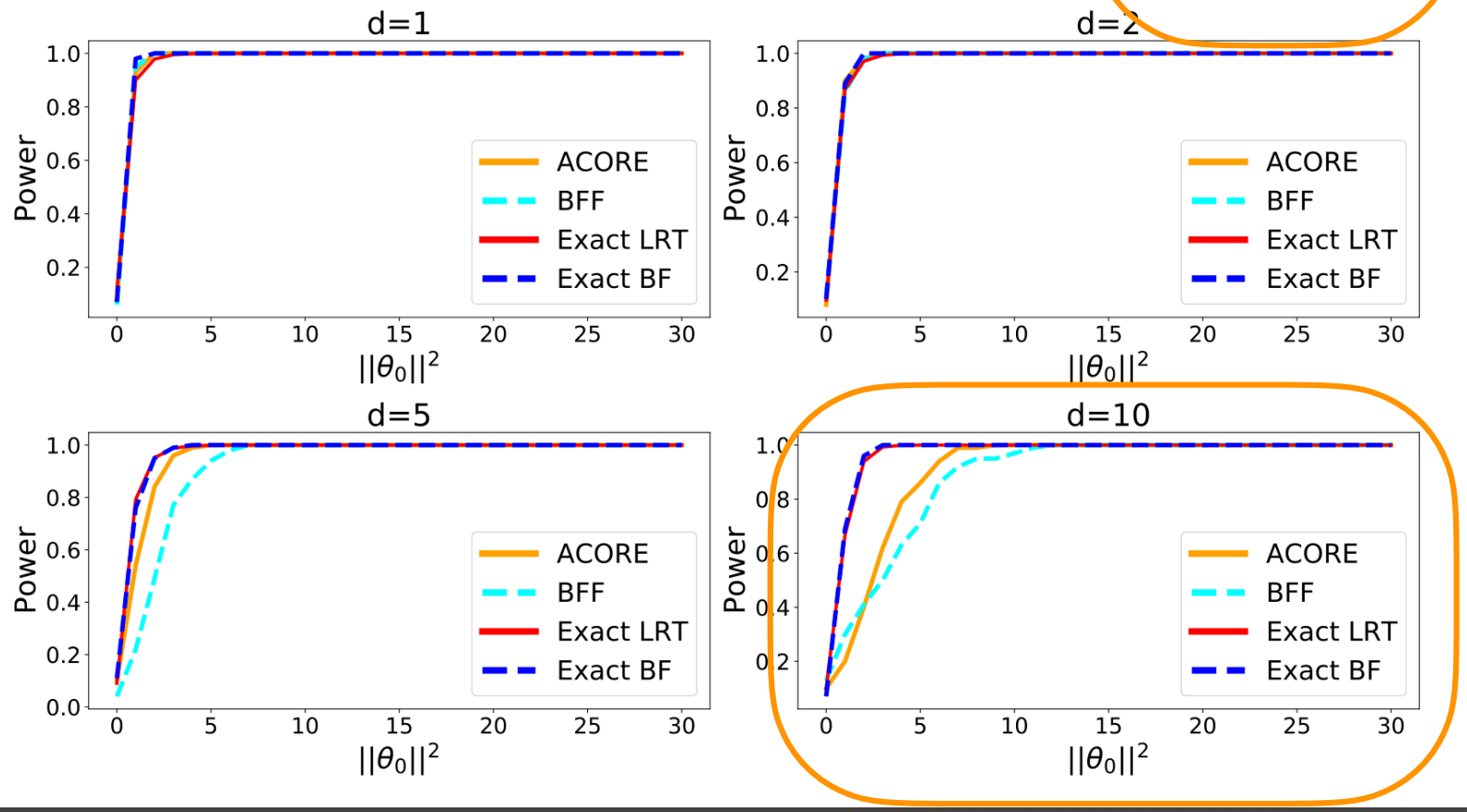
$$\widehat{\tau}(\mathcal{D};\theta_0) := \frac{\prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta_0)}{\int_{\Theta} \prod_{i=1}^{n} \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}};\theta)\ d\pi_{\tau}(\theta)}.$$

- 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 of LF2I Confidence Sets

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.

Current work in progress…

- Handling of nuisance parameters/systematics and more efficient methods for optimization or integration ⟹ next time?
- Alternative test statistics
  - "WALDO" ⟹ rest of the talk!

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{\left(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0\right)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

https://arxiv.org/abs/2205.15680

Simulation-Based Inference with WALDO: Perfectly Calibrated Confidence Regions Using Any Prediction or Posterior Estimation Algorithm

Luca Masserano
Department of Statistics and Data Science
Carnegie Mellon University
lmassera@andrew.cmu.edu

Tommaso Dorigo
INFN
Sezione di Padova
tommaso.dorigo@cern.ch

Rafael Izbicki
Department of Statistics
Federal University of São Carlos
rafaelizbicki@gmail.com

Mikael Kuusela
Department of Statistics and Data Science
Carnegie Mellon University
mkuusela@andrew.cmu.edu

Ann B. Lee
Department of Statistics and Data Science
Carnegie Mellon University
annlee@stat.cmu.edu

Abstract

The vast majority of modern machine learning targets prediction problems, with algorithms such as Deep Neural Networks revolutionizing the accuracy of point predictions for high-dimensional complex data. Predictive approaches are now used in many domain sciences to directly estimate internal parameters of interest in theoretical simulator-based models. In parallel, common alternatives focus

LF2I-Waldo for Calorimetric Muon Energy Measurement

[Luca Masserano, Rafael Izbicki, Tommaso Dorigo, Mikael Kuusela]

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

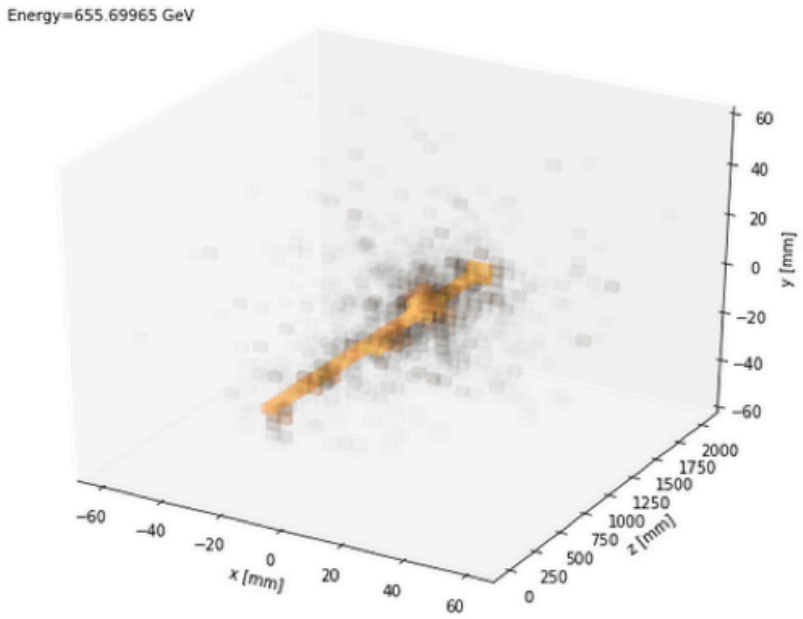


Figure 4: Muon entering the calorimeter in z direction.

1. Bias

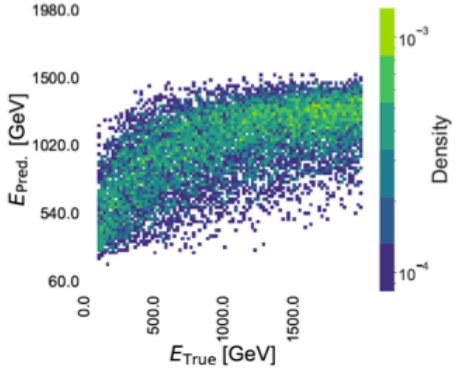


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

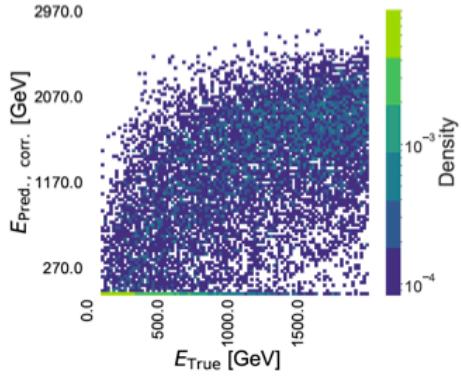


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

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

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

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

Can we do frequentist inference for muon energy?

We are mainly interested in **two questions**:

1. Infer, from the pattern of the energy deposits in the calorimeter, how much energy the incoming muon had *and* construct a **confidence set for it with proper coverage**

   → **goal**: Reconstruct muon properties with rigorous uncertainties for downstream analyses

2. How much added value does a **high granularity of the calorimeter** cells offer over the 1D and 28D representations?

   → **goal**: devise better and more cost-effective calorimeters for future particle colliders

Slide credit: Luca Masserano

Inputs: 1D energy-sum, 28 features and full calorimeter

Prediction algorithms used

Three "nested" datasets:

1. One-dimensional energy sum: minimizer of Cross-Validation MSE loss (XGBoost)
2. 27 features + 1D energy sum: minimizer of Cross-Validation MSE loss (XGBoost)
3. Full calorimeter (51200-D) + 28 features: custom CNN (with MSE loss) from Kieseler et al. (2022)

→ We estimate $\mathbb{E}[\theta \,|\, \mathscr{D}]$ and $\mathbb{V}[\theta \,|\, \mathscr{D}]$ for each of these. Muon energy is θ

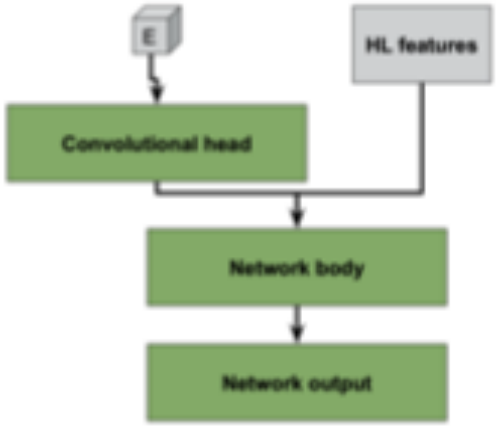


Image credit: Kieseler et al. (2022)

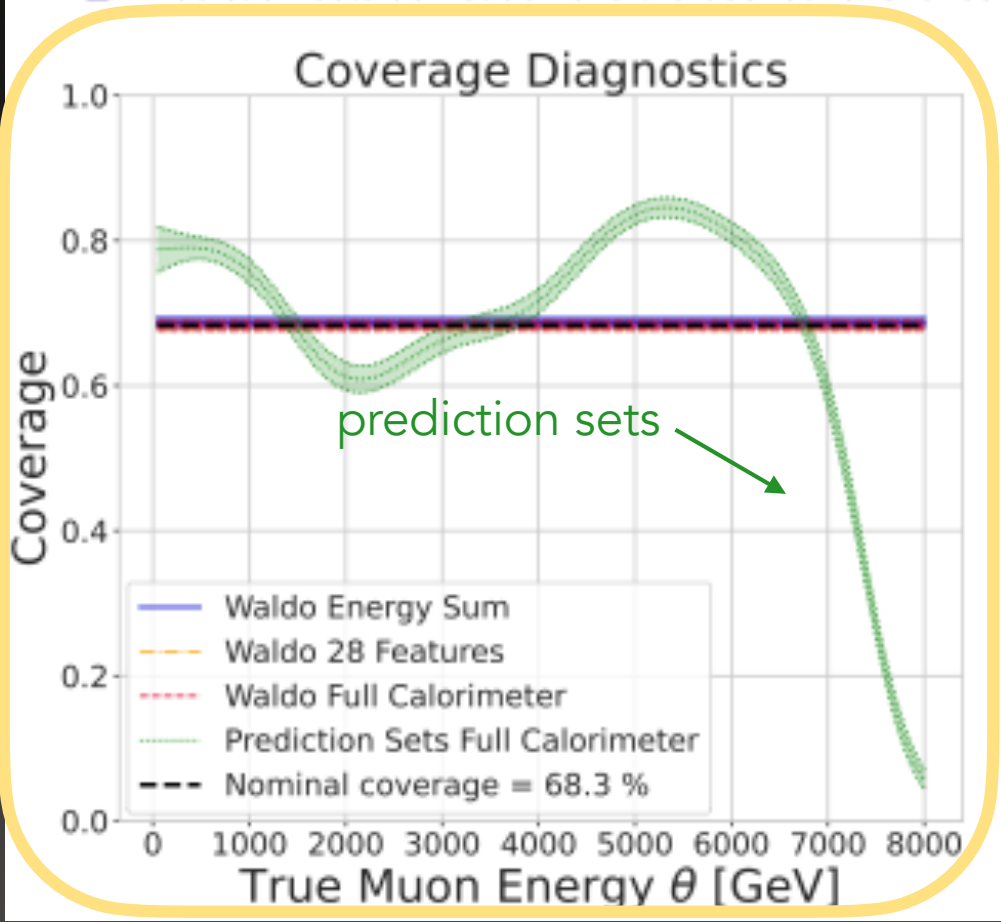
$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

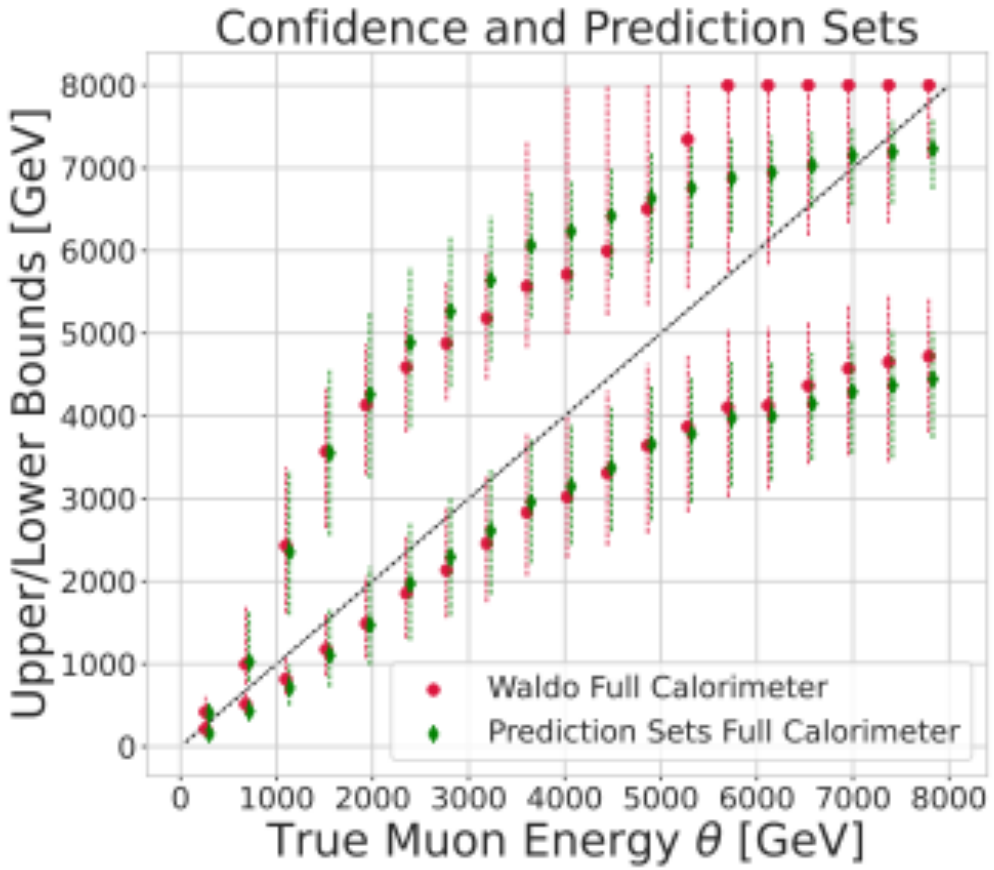
Slide credit: Luca Masserano

Valid confidence sets?

Confidence sets for muon energy have proper coverage

- Nominal coverage is achieved regardless of the dataset used
- Prediction sets do not achieve the desired level of coverage

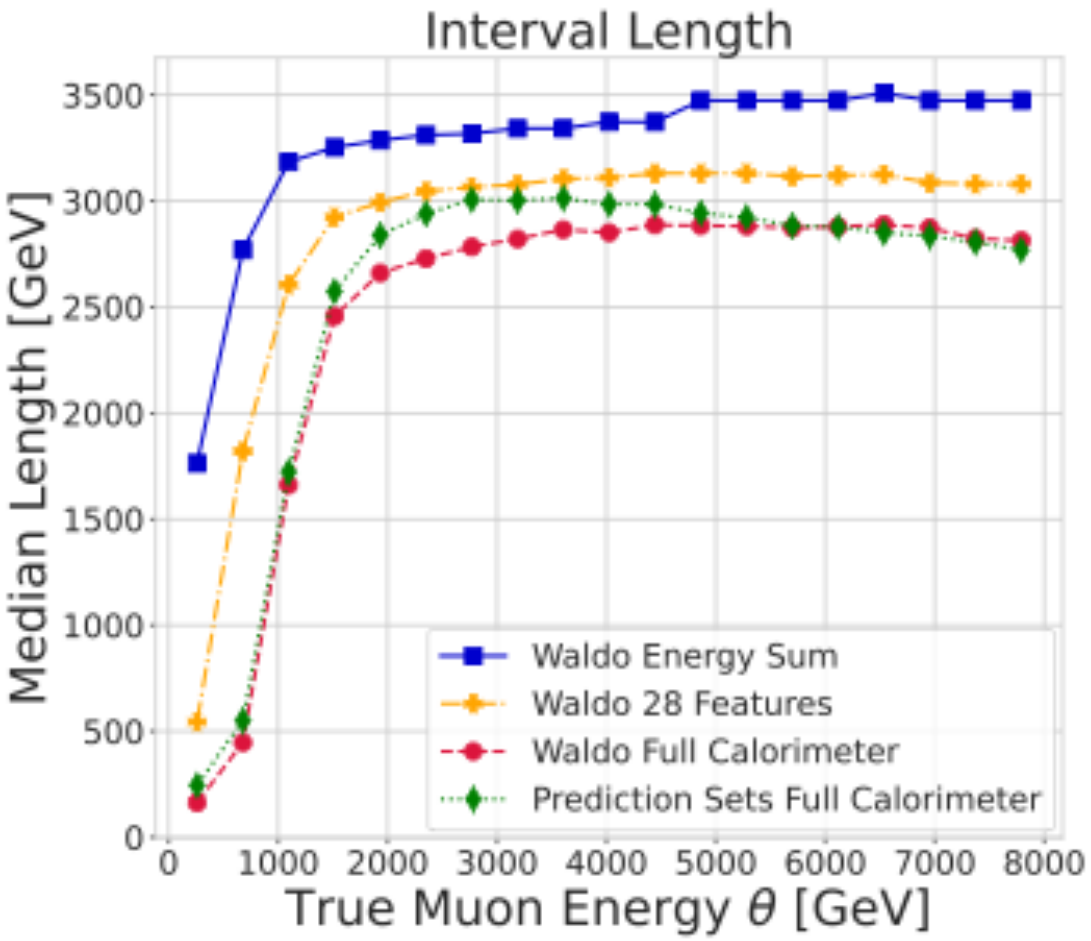




Slide credit: Luca Masserano

Constraining power?

Valuable information in high-granularity calorimeter

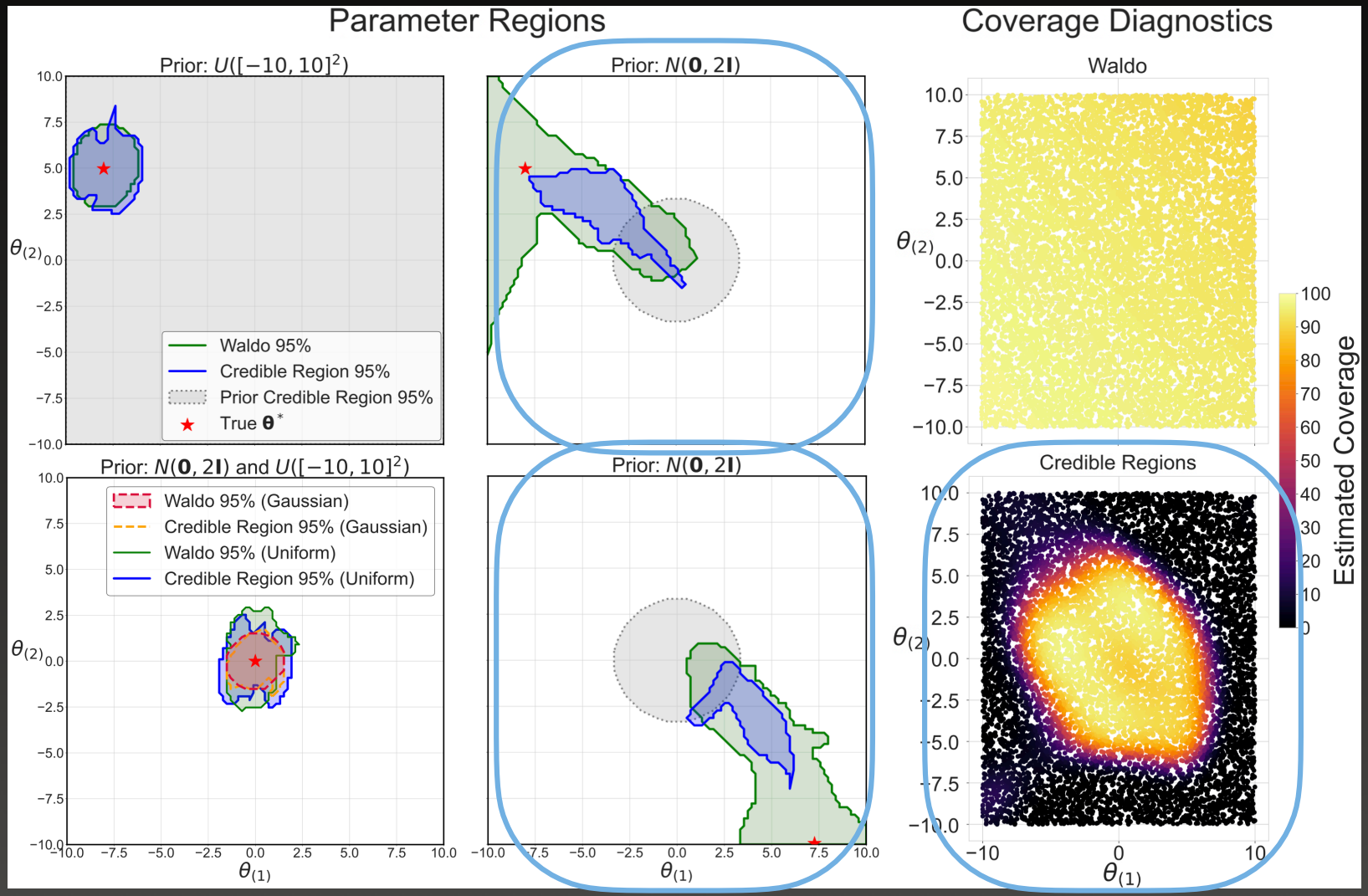


- Intervals are shorter as the data becomes higher-dimensional
- Prediction sets can even be larger than Waldo confidence sets (while also not guaranteeing coverage)

Slide credit: Luca Masserano

Ex: Recalibrating a Posterior Estimated via NFs

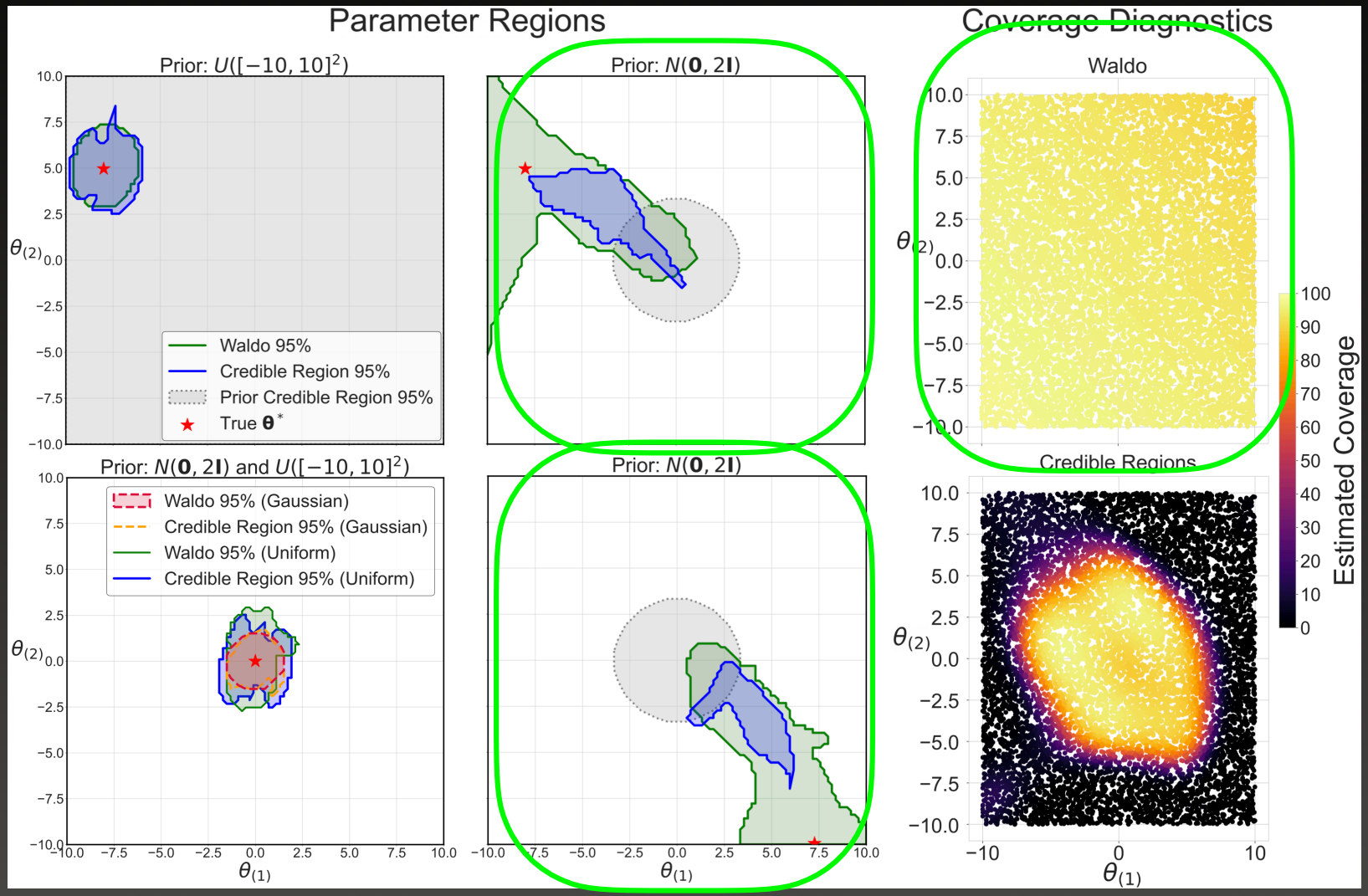
$$\mathcal{D}|\boldsymbol{\theta} \sim \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}) + \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, 0.01 \odot \mathbf{I}), \text{ where } \boldsymbol{\theta} \in \mathbb{R}^2 \text{ and } n = 1$$



Blue contours: 95% credible regions via Normalizing Flows (overly confident when prior is mis-specified)

Ex: Recalibrating a Posterior Estimated via NFs

$$\mathcal{D}|\boldsymbol{\theta} \sim \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}) + \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, 0.01 \odot \mathbf{I}), \text{ where } \boldsymbol{\theta} \in \mathbb{R}^2 \text{ and } n = 1$$

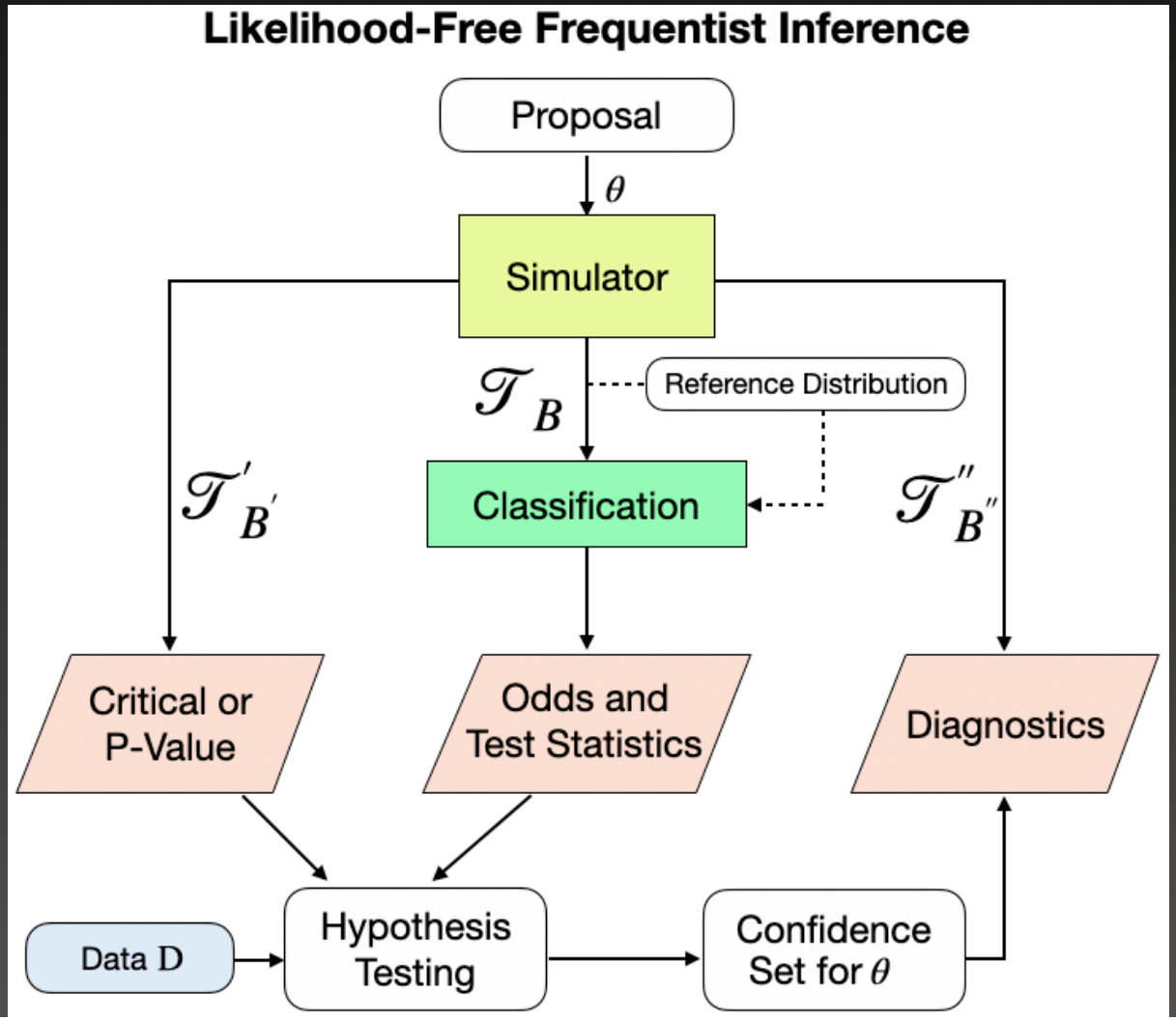


Waldo guarantees coverage everywhere, even if the prior is misspecified.

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{\left(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0\right)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

Take-Away: LF2I (inverse problem)

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



Take-Away: LF2I (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 frequentist confidence sets.
- Nuisance parameters and diagnostics: No guarantee that hybrid methods are valid. However, we have a practical tool for assessing coverage across the entire parameter space.
- Power: Hardest to achieve in practice. Area where most statistical and computational advances will take place.

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

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

- `BFF` (Bayesian Frequentist Factor):

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

$$\tau^{\text{WALDO}}(\mathcal{D};\boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

Collaborators

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

LF2I framework

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

EXTRA SLIDES START HERE

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?

HEP Example: "Poisson Counting Experiment"

[Lyons, 2008; Cowan et al, 2011; Cowan, 2012]

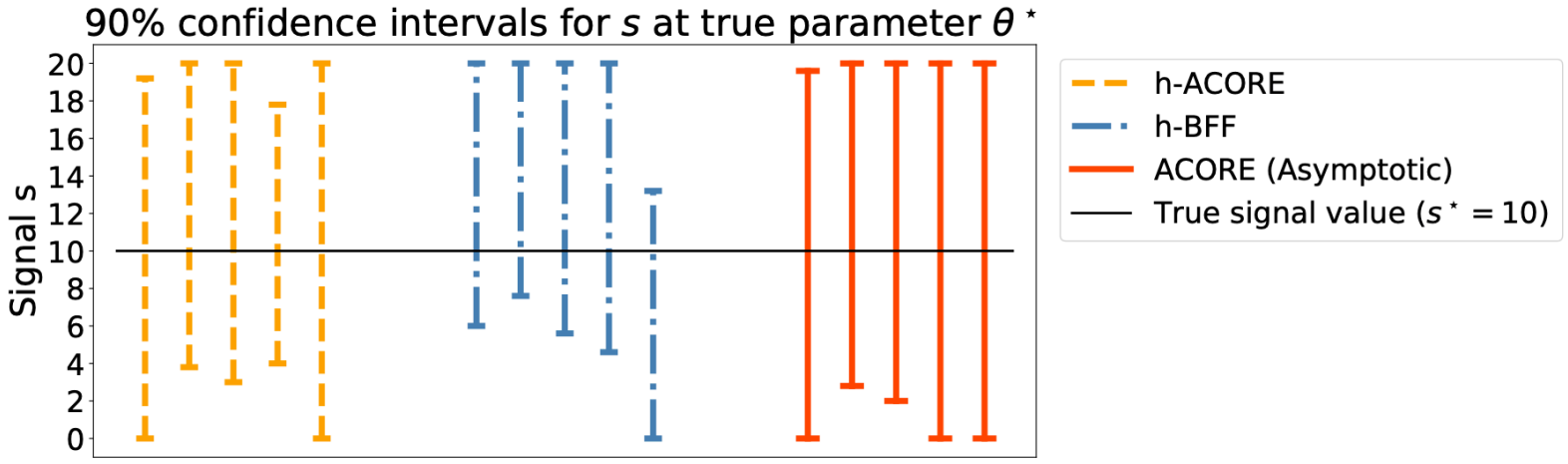
- Particle collision events counted under the presence of a background process.

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

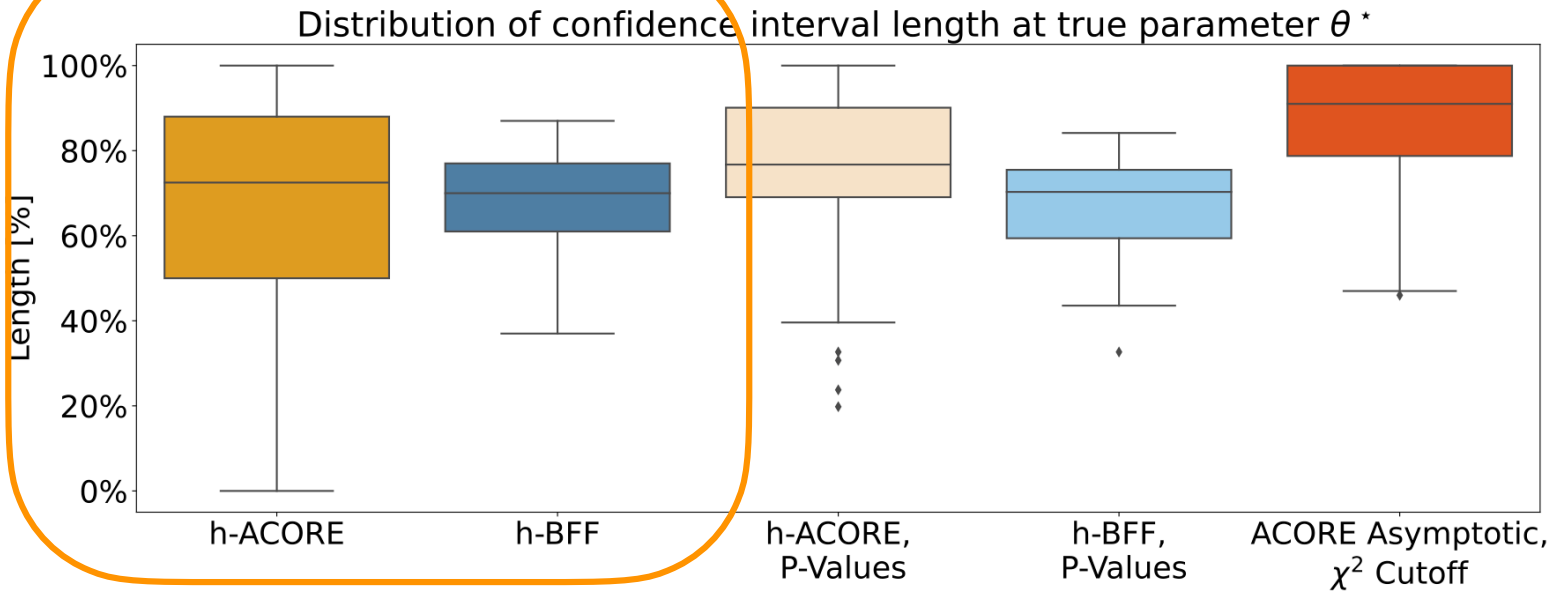
- The observed data D consist of n=10 realizations of X=(M,N), where
  - M is the number of events in the control region (assume γ=1)
  - N is the number of events in the signal region
- Unknown parameters:
  - signal strength (s); two nuisance parameters (b and ϵ)

Confidence sets at a fiducial point

HEP example with nuisance parameters



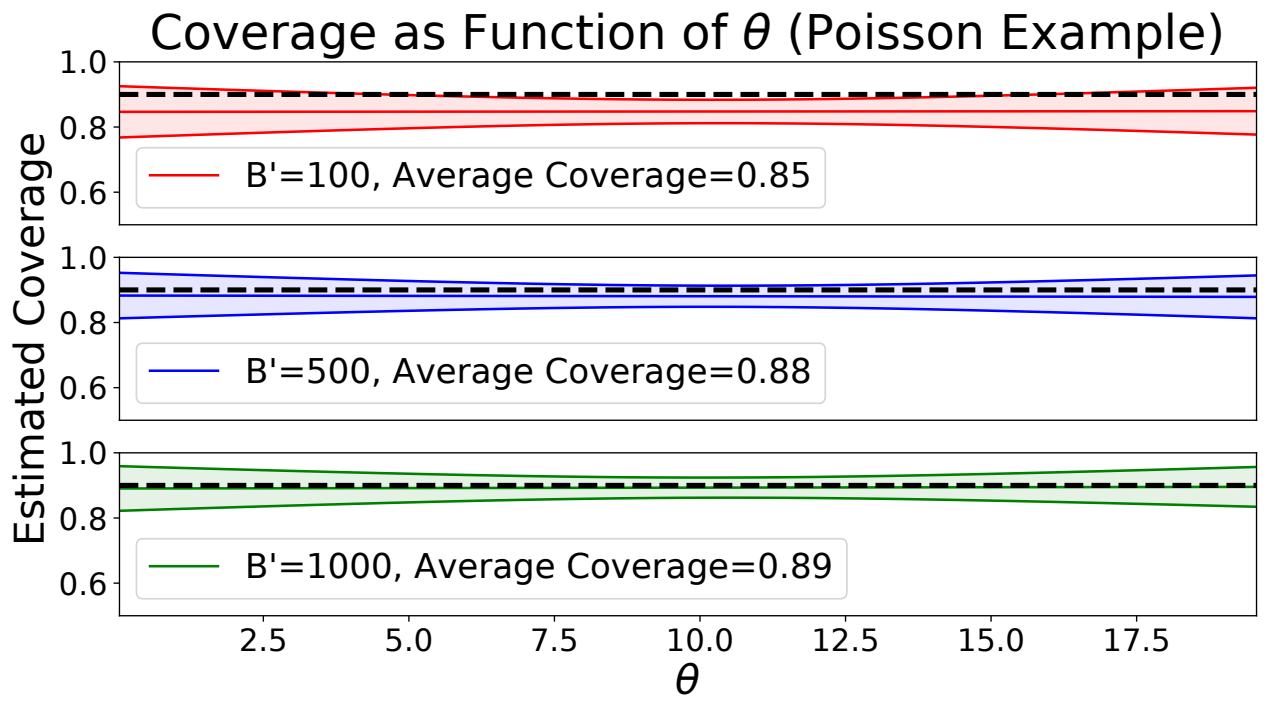
	h-ACORE	h-BFF	h-ACORE (p-values)	h-BFF (p-values)	ACORE (Asymptotic)
Coverage	0.87 ± 0.03	0.91 ± 0.03	0.92 ± 0.03	0.94 ± 0.02	0.97 ± 0.02



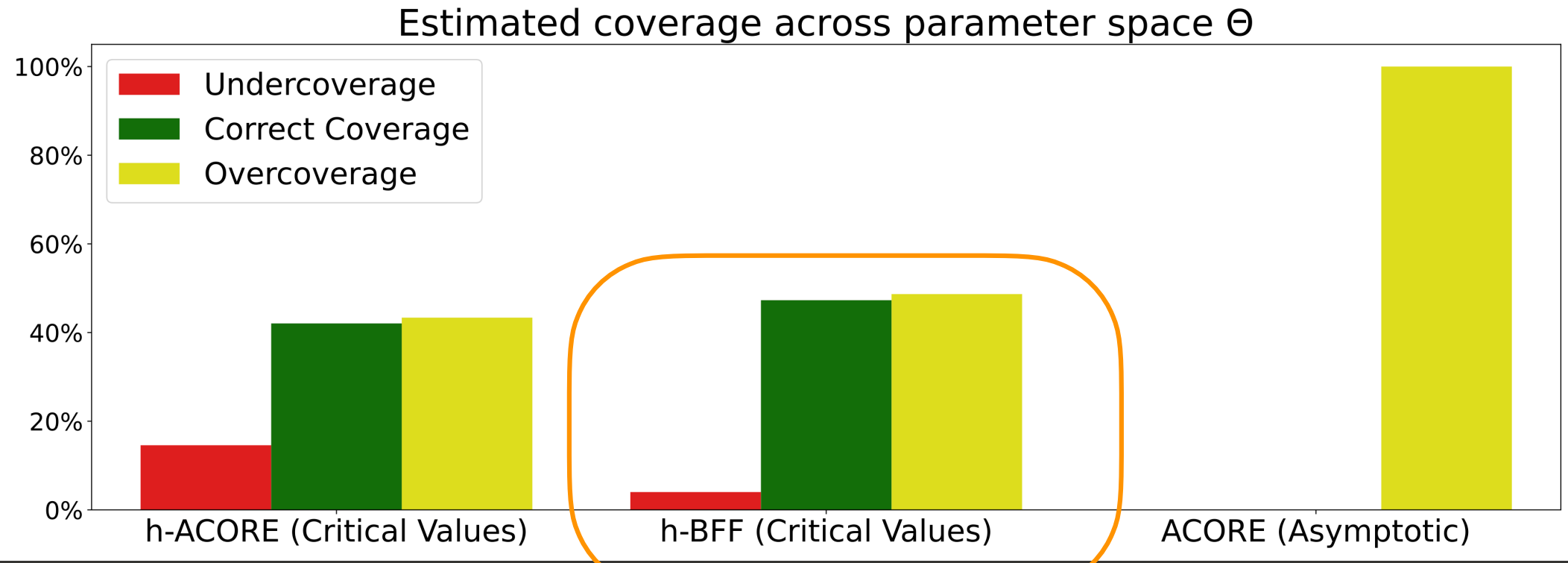
Assessing Conditional Coverage

- 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.



Diagnostics for HEP Example (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)