Calibrated Uncertainty Quantification in Simulator-Based Inference

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

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

Bridging Classical Statistics and Machine Learning in Scientific Inference

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

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

The Interplay Between Theory/Models and Data

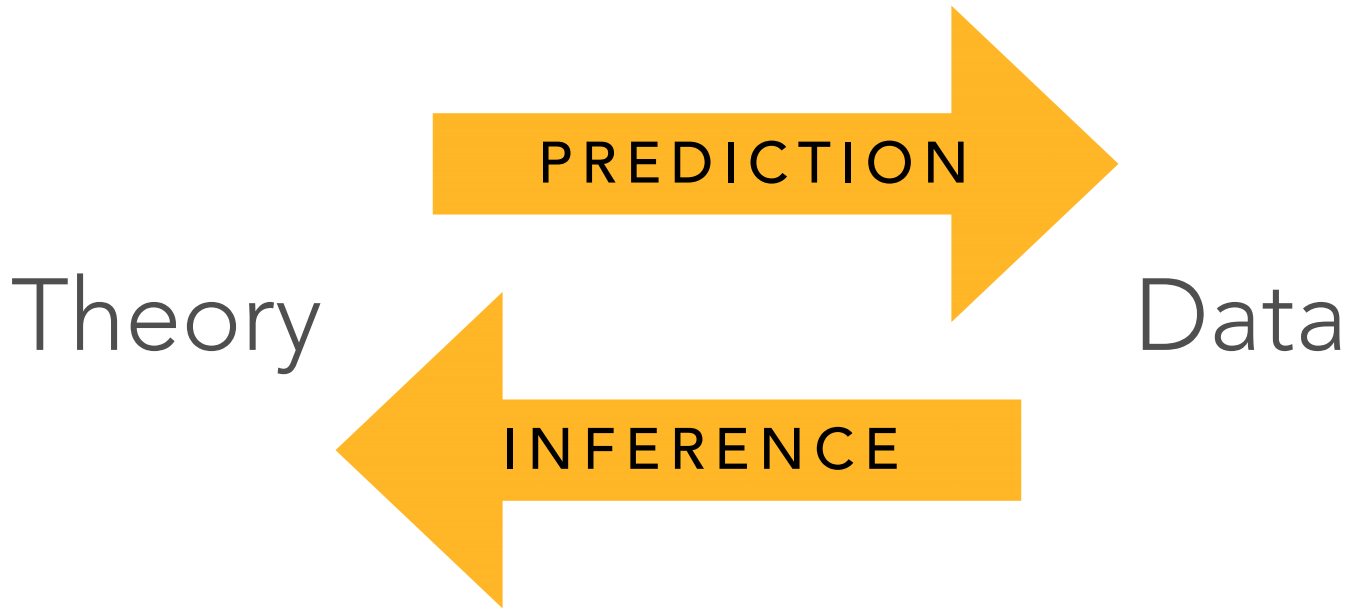
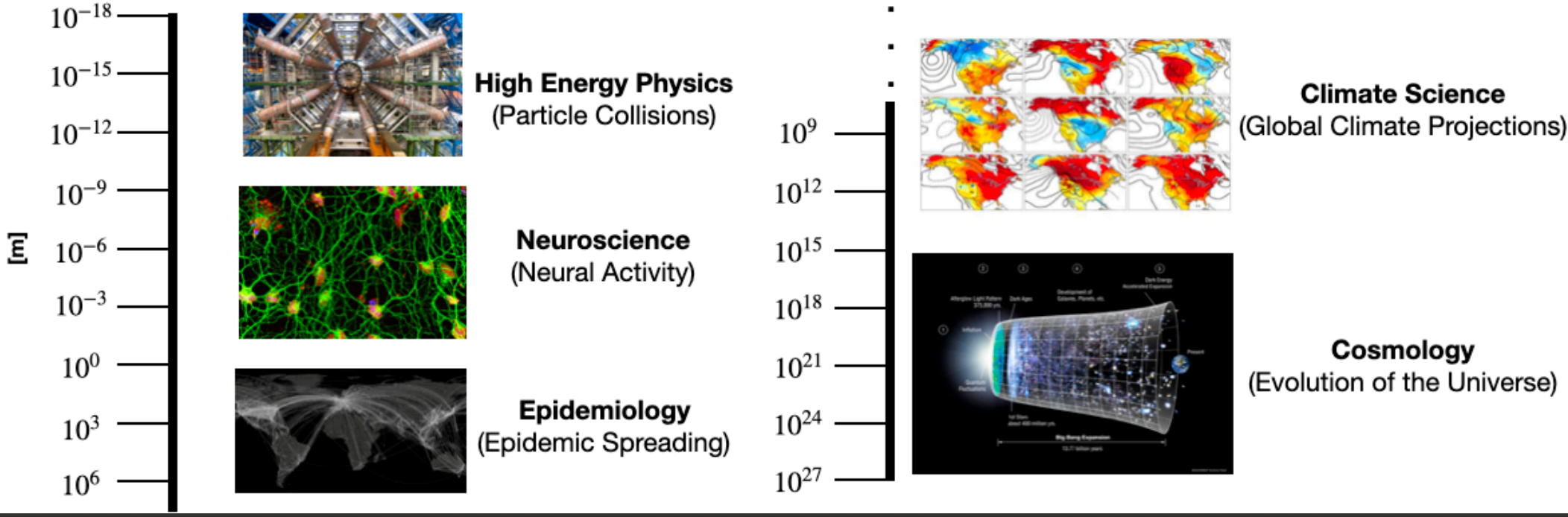


Figure credit: Kyle Cranmer

Simulators are Ubiquitous in Science



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

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

Taxonomy of Different Types of Simulators

Image credit: Kyle Cranmer

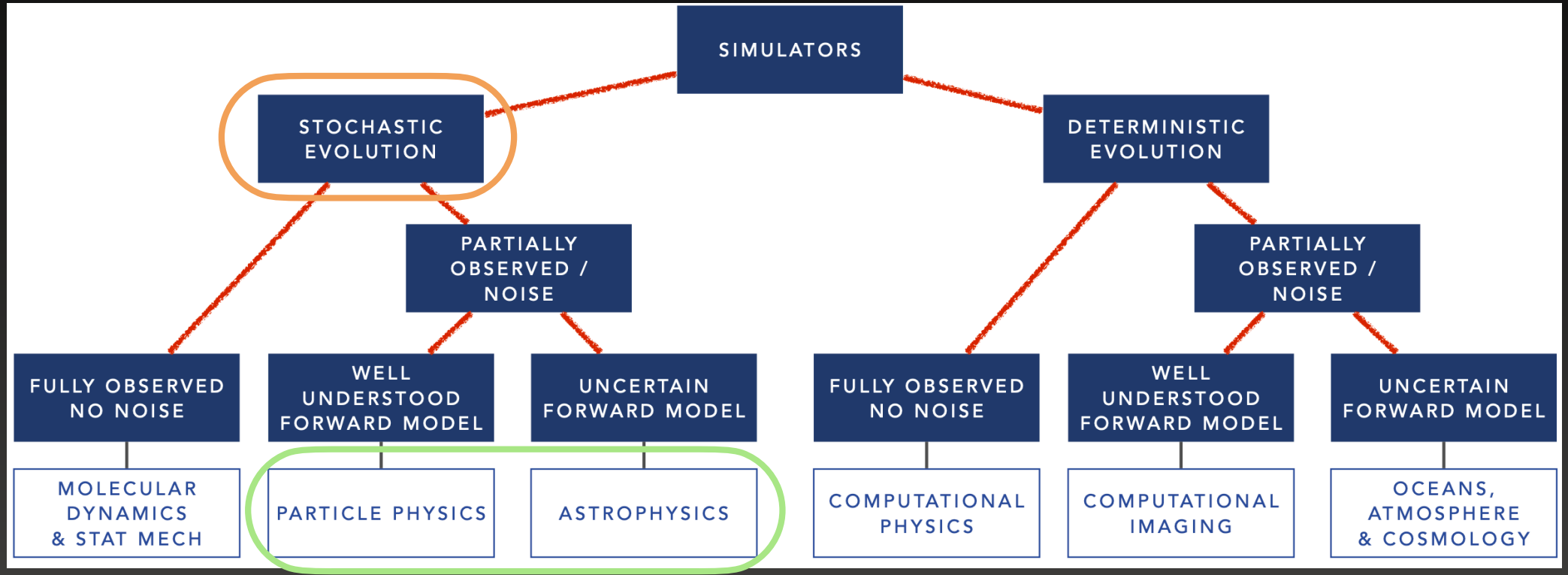
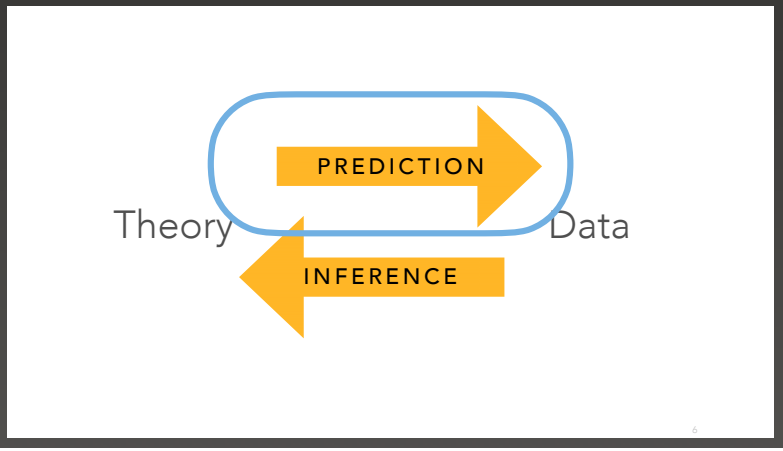


Figure credit: Kyle Cranmer

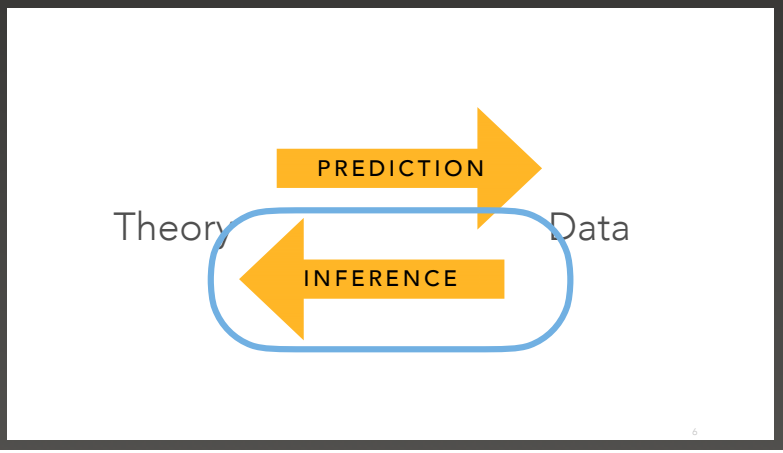
Simulation/Prediction to Scientific Inference

- Revolution in simulators and AI deep generative models (GANs, transformers, diffusion models etc) & high-performance prediction algorithms.
- But what about scientific inference?
  - Simulators are often poorly suited for the "inverse problem" of inferring the causes behind observed phenomena.



Simulation/Prediction to Scientific Inference

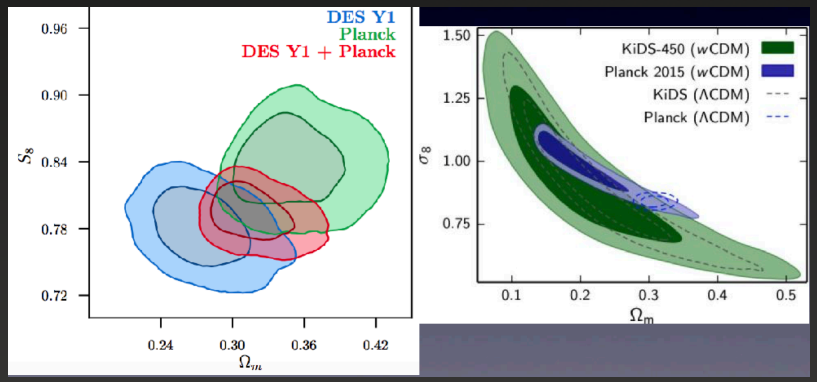
- Revolution in simulators and AI deep generative models (GANs, transformers, diffusion models etc) & high-performance prediction algorithms.
- But what about scientific inference?
  - Simulators are often poorly suited for the "inverse problem" of inferring the causes behind observed phenomena.



Scientific Inference and Causation

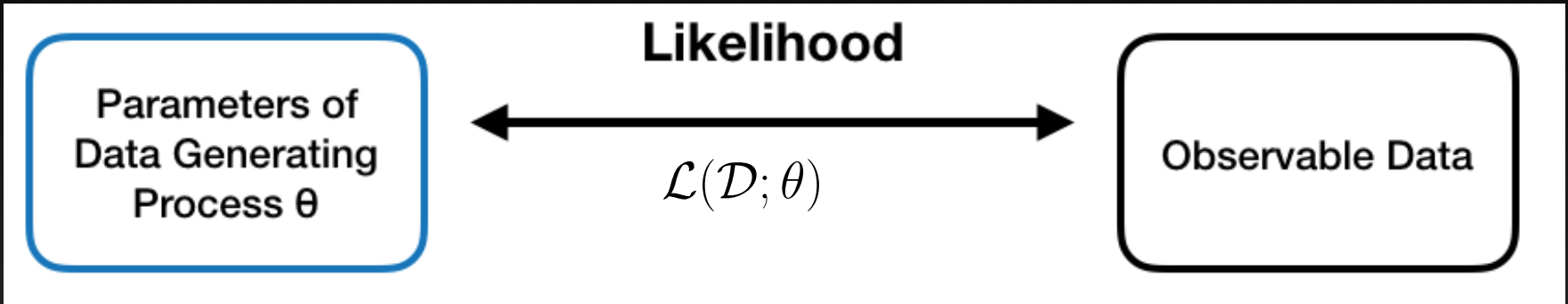
- Much of ML targets "forward" problems and generative models.

In many science applications, however, the quantities of interest are internal parameters θ, i.e. the ``causes'' of observations in $\theta \rightarrow D$ problems.



Given observed data, constrain internal parameters of interest using assumed theoretical/simulation model. Valid measures of uncertainty.

Likelihood-Based Inference



Likelihood-Free Inference (LFI)

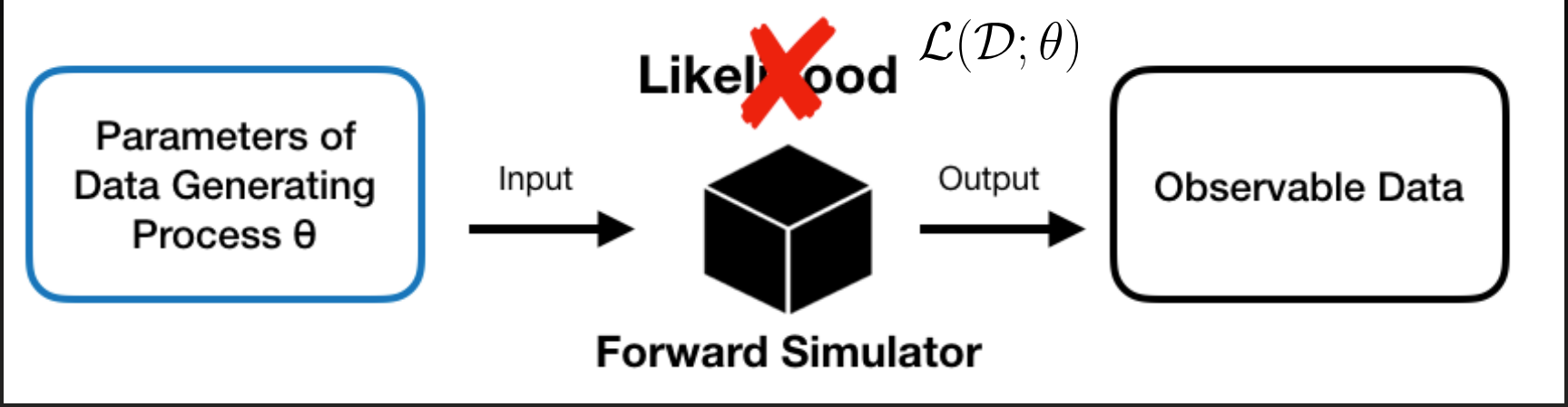


Image credit: Nic Dalmasso

- The likelihood cannot be evaluated. But it is implicitly encoded by the simulator…
- Inference on parameters in this setting is called likelihood-free inference (LFI)

$$\text{Simulate } \theta_i \sim r(\cdot),\ \mathcal{D}_i|\theta_i \sim \mathcal{L}(\cdot;\theta_i) \text{ where } \mathcal{D}_i = (\mathbf{X}_{i,1},\ldots,\mathbf{X}_{i,n})$$

$$\implies \mathcal{T}_B = \{(\theta_1,\mathcal{D}_1),(\theta_2,\mathcal{D}_2),\ldots,(\theta_B,\mathcal{D}_B)\},$$

Likelihood-Free Inference (LFI)

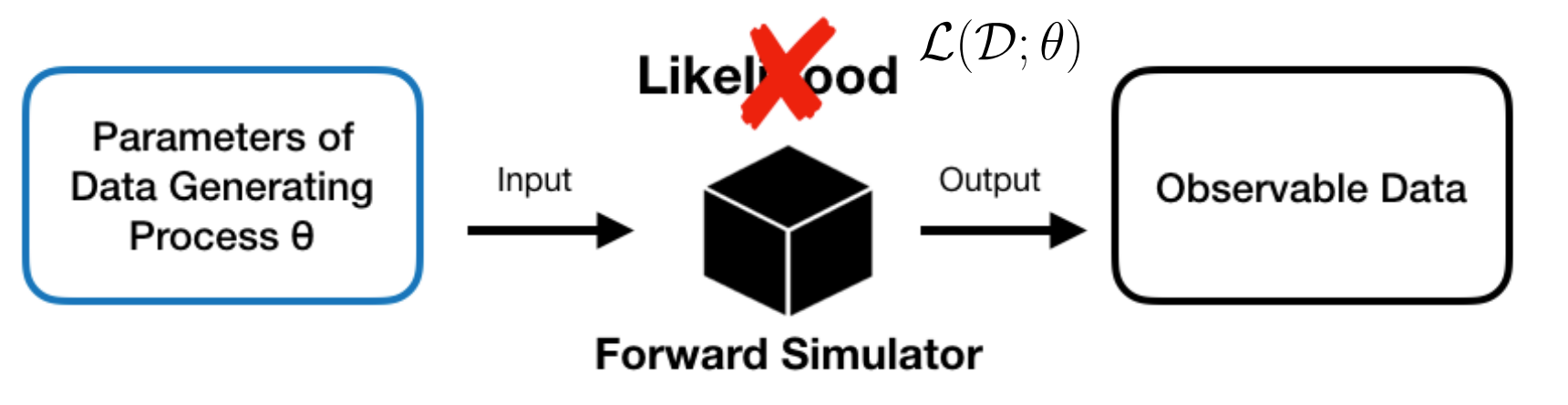


Image credit: Nic Dalmasso

- The likelihood cannot be evaluated. But it is implicitly encoded by the simulator…
- Inference on parameters in this setting is called likelihood-free inference (LFI), a.k.a. SBI

$$\text{Simulate } \theta_i \sim r(\cdot),\ \ \mathcal{D}_i|\theta_i \sim \mathcal{L}(\cdot;\theta_i) \text{ where } \mathcal{D}_i = (\mathbf{X}_{i,1},\ldots,\mathbf{X}_{i,n})$$

$$\implies \mathcal{T}_B = \{(\theta_1,\mathcal{D}_1),(\theta_2,\mathcal{D}_2),\ldots,(\theta_B,\mathcal{D}_B)\},$$

Classical LFI: Approximate Bayesian Computation (ABC) [Rubin 1984; Overview of ABC: Sisson et al, 2018]

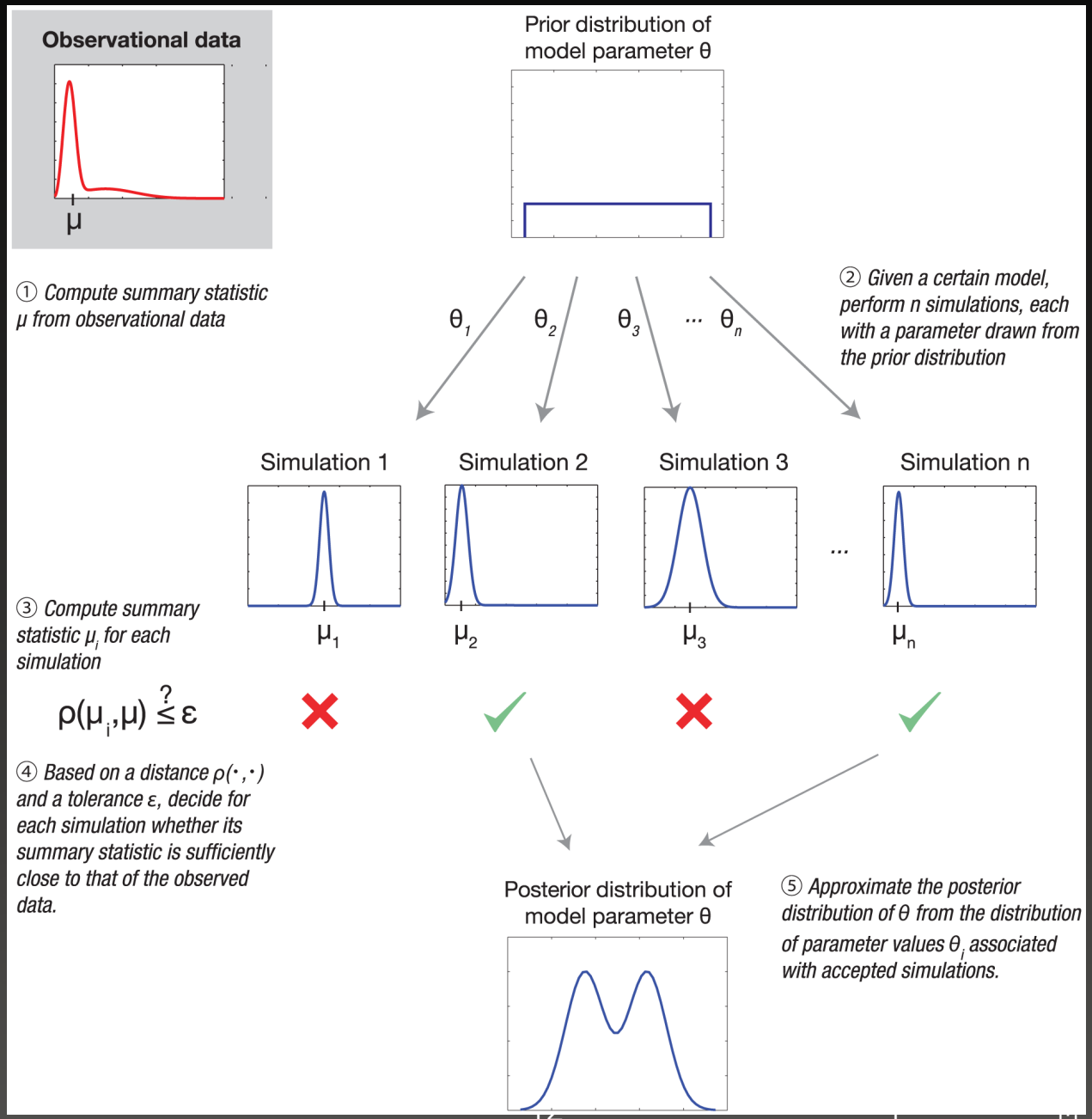


Image credit: Sunnaker et al. 2013

Modern LFI Landscape [review: Cranmer et al, PNAS 2019]

- Use ML-based algorithms to directly estimate key inferential quantities from simulated "train" set

$$\mathcal{T}_B = \{(\theta_1, \mathcal{D}_1), (\theta_2, \mathcal{D}_2), \ldots, (\theta_B, \mathcal{D}_B)\}, \text{ where } \theta \sim r(\cdot), \;\; \mathcal{D}|\theta \sim \mathcal{L}(\cdot; \theta)$$

Ex: Learning the Likelihood Ratio $f(D|\theta_1)/f(D|\theta_2)$

[e.g, Cranmer et al, 2015; Thomas et al, 2016; Hermans et al, 2020; Durkan et al, 2020; Brehmer et al, 2020]

Approximating Likelihood Ratios with Calibrated Discriminative Classifiers

Kyle Cranmer[1], Juan Pavez[2], and Gilles Louppe[1]
[1]New York University
[2]Federico Santa María University

March 21, 2016 arXiv:1506.02169

$$r(x;\theta_0,\theta_1) = \frac{p(x \mid \theta_0)}{p(x \mid \theta_1)} = 1 - \frac{1}{s(x;\theta_0,\theta_1)}$$

In light of this result, the likelihood ratio estimation problem can now be recast as a (probabilistic) classification problem, by noticing that the decision function

$$s^*(\mathbf{x}) = \frac{p_{\mathbf{X}}(\mathbf{x}|\theta_1)}{p_{\mathbf{X}}(\mathbf{x}|\theta_0) + p_{\mathbf{X}}(\mathbf{x}|\theta_1)}. \tag{2.10}$$

modeled by a classifier trained to distinguish samples $\mathbf{x} \sim p_{\theta_0}$ from samples $\mathbf{x} \sim p_{\theta_1}$

Learning Likelihood Ratios with Neural Network Classifiers

Shahzar Rizvi,[1,*] Mariel Pettee,[2,†] and Benjamin Nachman[2,3,‡]
[1]*Department of Statistics, University of California, Berkeley, CA 94720, USA*
[2]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[3]*Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*

arXiv:2303.10300

Ex: Learning the Likelihood Function f(D|θ)

[e.g, Izbicki et al, 2014; Thomas et al, 2016; Durkan et al, 2020; Brehmer et al., 2020]

High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation

AISTATS 2014; arXiv:1506.02169

Rafael Izbicki **Ann B. Lee** **Chad M. Schafer**

Department of Statistics – Carnegie Mellon University

mate the ratio

$$\mathcal{L}(\mathbf{x};\theta) \equiv \frac{f(\mathbf{x}|\theta)}{g(\mathbf{x})}, \qquad (5)$$

where $f(\mathbf{x}|\theta)$ is the conditional density of $\mathbf{x}$ given θ, and $g(\mathbf{x})$ is the marginal density for $\mathbf{x}$. This is, up to a multiplicative factor that is not a function of θ, the standard definition of the likelihood function.

RKHS basis (orthogonal wrt P)

Hence, we define our likelihood function estimator by

$$\widehat{\mathcal{L}}(\mathbf{x};\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} \widehat{\beta}_{i,j}\widehat{\Psi}_{i,j}(\mathbf{x},\theta),$$

[Submitted on 30 Nov 2016 (v1), last revised 11 Sep 2020 (this version, v6)]

Likelihood-free inference by ratio estimation

Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, Michael U. Gutmann

arXiv:1611.10242

LR trick and logistic regression

Ex: Learning the Posterior f(θ|D)

[e.g, Papamakarios et al, 2016; Lueckmann et al, 2017; Alsing et al 2019; Izbicki et al, 2019; Greenberg et al, 2019]

Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation

NeurIPS 2016; arXiv:1605.06376

George Papamakarios
School of Informatics
University of Edinburgh
g.papamakarios@ed.ac.uk

Iain Murray
School of Informatics
University of Edinburgh
i.murray@ed.ac.uk

(S)NPE via MDNs
+ sequential updates
via proposal priors and reweighting

for $n = 1..N$ **do**
  sample $\boldsymbol{\theta}_n \sim \tilde{p}(\boldsymbol{\theta})$
  sample $\mathbf{x}_n \sim p(\mathbf{x} \mid \boldsymbol{\theta}_n)$
end
train $q_{\boldsymbol{\phi}}(\boldsymbol{\theta} \mid \mathbf{x})$ on $\{\boldsymbol{\theta}_n, \mathbf{x}_n\}$
$\hat{p}(\boldsymbol{\theta} \mid \mathbf{x} = \mathbf{x}_o) \leftarrow \frac{p(\boldsymbol{\theta})}{\tilde{p}(\boldsymbol{\theta})} q_{\boldsymbol{\phi}}(\boldsymbol{\theta} \mid \mathbf{x}_o)$

Flexible statistical inference for mechanistic models of neural dynamics

NeurIPS 2017; arXiv:1711.01861

Jan-Matthis Lueckmann[*1], **Pedro J. Gonçalves**[*1], **Giacomo Bassetto**[1], **Kaan Öcal**[1,2], **Marcel Nonnenmacher**[1], **Jakob H. Macke**[†1]
[1] research center caesar, an associate of the Max Planck Society, Bonn, Germany
[2] Mathematical Institute, University of Bonn, Bonn, Germany
{jan-matthis.lueckmann, pedro.goncalves, giacomo.bassetto, kaan.oecal, marcel.nonnenmacher, jakob.macke}@caesar.de

(S)NPE via MDNs
+ sequential updates via
importance-weighted loss

$$\mathcal{L}(\boldsymbol{\phi}) = -\frac{1}{N} \sum_n \frac{p(\boldsymbol{\theta}_n)}{\tilde{p}(\boldsymbol{\theta}_n)} K_\tau(\mathbf{x}_n, \mathbf{x}_o) \log q_{\boldsymbol{\phi}}(\boldsymbol{\theta}_n | \mathbf{x}_n),$$

Modern LFI Landscape [review: Cranmer et al, PNAS 2019]

- Use ML-based algorithms to directly estimate key inferential quantities from simulated "train" set

$$\mathcal{T}_B = \{(\theta_1, \mathcal{D}_1), (\theta_2, \mathcal{D}_2), \ldots, (\theta_B, \mathcal{D}_B)\}, \text{ where } \theta \sim r(\cdot), \;\; \mathcal{D}|\theta \sim \mathcal{L}(\cdot; \theta)$$

  - Likelihood ratios, $f(D|\boldsymbol{\theta}_1)/f(D|\boldsymbol{\theta}_2)$ [e.g, Cranmer et al, 2015; Thomas et al, 2016; Hermans et al, 2020; Durkan et al, 2020; Brehmer et al, 2020]
  - Likelihoods, $f(D|\boldsymbol{\theta})$ [e.g., Izbicki et al, 2014; Thomas et al, 2016; Durkan et al, 2020; Brehmer et al., 2020]
  - Posteriors, $f(\boldsymbol{\theta}|D)$ [e.g., Papamakarios et al, 2016; Lueckmann et al, 2017; Izbicki et al, 2018; Alsing et al 2019; Greenberg et al, 2019]

- These ML-based approaches can handle complex high-dimensional data without a prior dimension reduction. Base versions also provide "amortized" inference.

Two Open Problems in ML-Based Likelihood-Free Inference

Challenge 1: Valid UQ / Coverage

- Valid UQ / Coverage: How do we guarantee confidence sets R(D) to have nominal coverage —- for every θ?, any sample size n (including e.g. n=1), and any choice of reference distribution r?

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

Challenge 1: Valid UQ / Coverage

- Valid UQ / Coverage: How do we guarantee confidence sets R(D) to have nominal coverage —- for every θ, for any sample size n (including e.g. n=1)?, and any choice of reference distribution r?

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

Simulate $\theta_i \sim r(\cdot), \quad \mathcal{D}_i|\theta_i \sim \mathcal{L}(\cdot;\theta_i)$ where $\mathcal{D}_i = (\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,n})$

$$\Longrightarrow \mathcal{T}_B = \{(\theta_1, \mathcal{D}_1), (\theta_2, \mathcal{D}_2), \ldots, (\theta_B, \mathcal{D}_B)\},$$

Challenge 1: Valid UQ / Coverage

- Valid UQ / Coverage: How do we guarantee confidence sets R(D) to have nominal coverage —- for every θ, any sample size n (including e.g. n=1), and for any choice of reference or prior distribution r?

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

$$\text{Simulate } \theta_i \sim r(\cdot), \quad \mathcal{D}_i|\theta_i \sim \mathcal{L}(\cdot;\theta_i) \text{ where } \mathcal{D}_i = (\mathbf{X}_{i,1},\ldots,\mathbf{X}_{i,n})$$

$$\Longrightarrow \mathcal{T}_B = \{(\theta_1,\mathcal{D}_1),(\theta_2,\mathcal{D}_2),\ldots,(\theta_B,\mathcal{D}_B)\},$$

$$\mathcal{T}'_B = \{(\theta'_1,\mathcal{D}'_1),(\theta'_2,\mathcal{D}'_2),\ldots,(\theta'_B,\mathcal{D}'_B)\}, \text{ where } \theta' \sim r'(\cdot), \quad \mathcal{D}'|\theta' \sim \mathcal{L}(\cdot;\theta')$$

Challenge 2: Diagnostics/Validation of Coverage

- How do we check empirical coverage of the final constructed confidence sets across the entire parameter space? (Note: "Consistency checks" only check marginal coverage)
  - R. Cousins: "*Lectures on Statistics in Theory: Prelude to Statistics in Practice*", arXiv:1807.05996, 2018:

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.

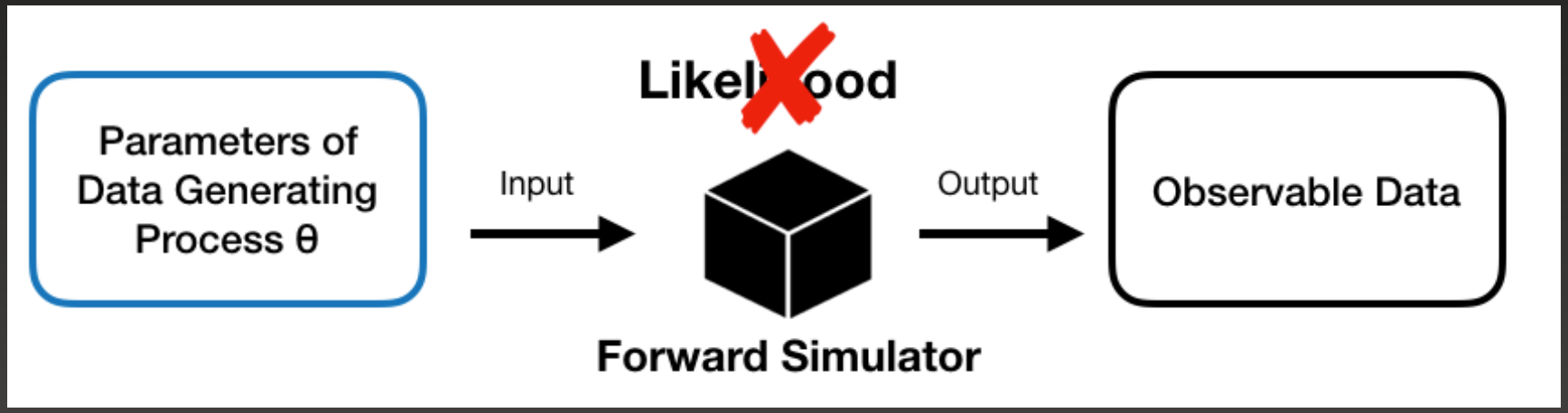
But... the ideal of a fine grid is usually impractical.

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

Recall: LFI Setting

Given observed data, constrain internal parameters of interest using assumed theoretical/simulation model. Valid measures of uncertainty.

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



$$\mathcal{T}_B = \{(\theta_1, \mathcal{D}_1), (\theta_2, \mathcal{D}_2), \ldots, (\theta_B, \mathcal{D}_B)\}, \text{ where } \theta \sim r(\cdot), \quad \mathcal{D}|\theta \sim \mathcal{L}(\cdot;\theta)$$

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

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

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

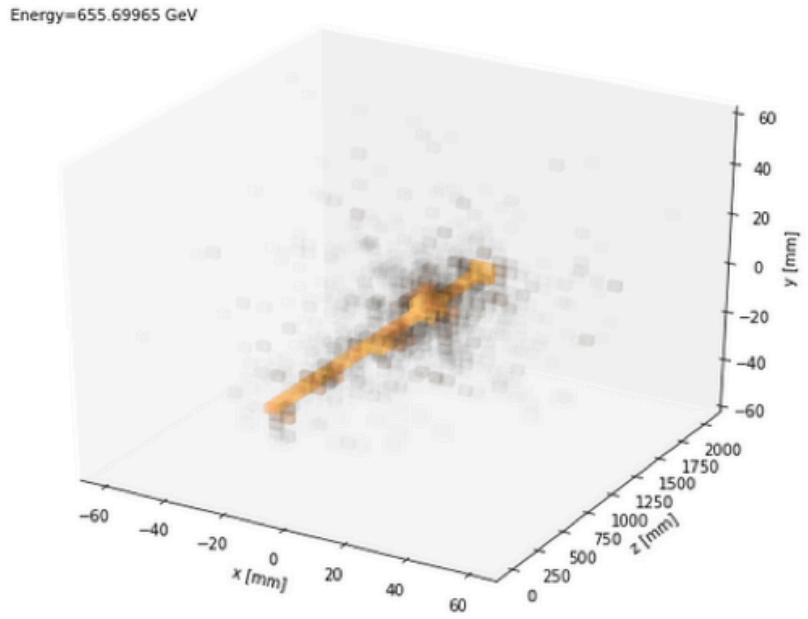


Figure 4: Muon entering the calorimeter in z direction.

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

1. Bias

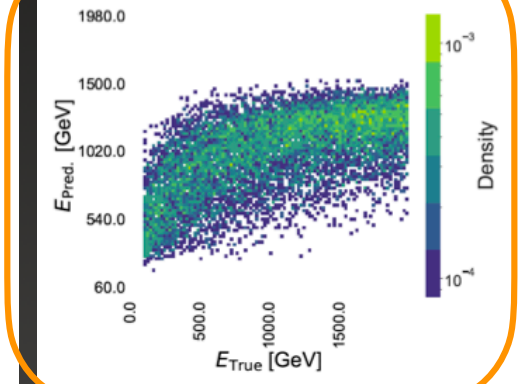


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

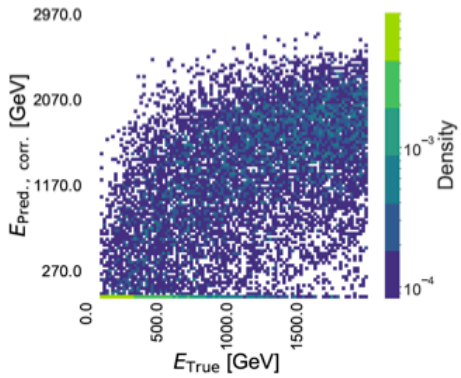


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

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

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

Similarly, neural posteriors via e.g. NFs do not guarantee coverage of internal parameters (often "over-confident")

Averting A Crisis In Simulation-Based Inference

https://arxiv.org/abs/2110.06581

Joeri Hermans*
University of Liège
joeri.hermans@doct.uliege.be

Arnaud Delaunoy*
University of Liège
a.delaunoy@uliege.be

François Rozet
University of Liège
francois.rozet@uliege.be

Antoine Wehenkel
University of Liège
antoine.wehenkel@uliege.be

Gilles Louppe
University of Liège
g.louppe@uliege.be

Abstract

We present extensive empirical evidence showing that current Bayesian simulation-based inference algorithms are inadequate for the falsificationist methodology of scientific inquiry. Our results collected through months of experimental computations show that all benchmarked algorithms – (S)NPE, (S)NRE, SNL and variants of ABC – may produce overconfident posterior approximations, which makes them demonstrably unreliable and dangerous if one's scientific goal is to constrain parameters of interest. We believe that failing to address this issue will lead to a well-founded trust crisis in simulation-based inference. For this reason, we argue that research efforts should now consider theoretical and method-

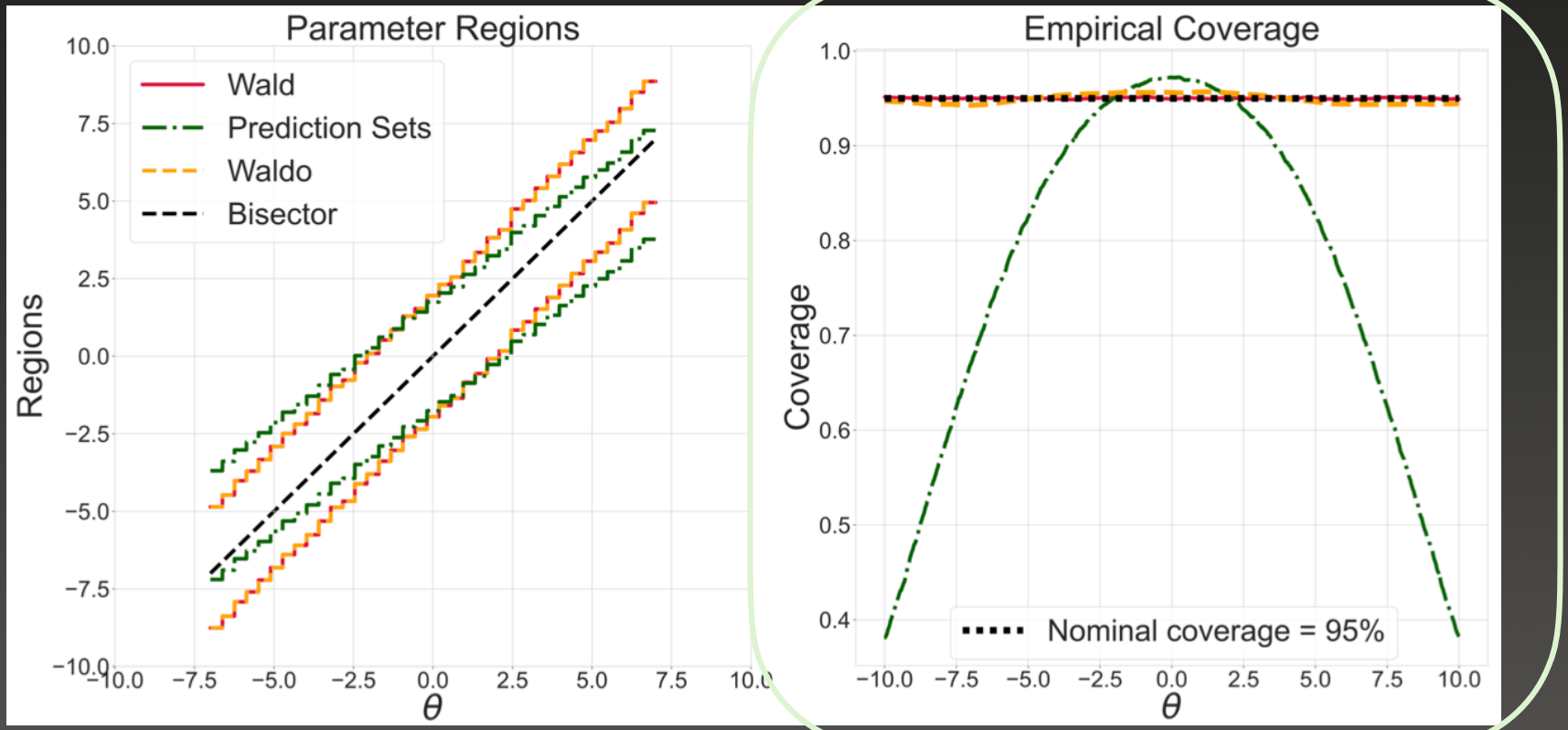
evaluation requires the often *intractable* integration of all stochastic execution paths. In this problem setting, statistical inference based on the likelihood becomes impractical. However, approximate inference remains possible by relying on likelihood-free *approximations* thanks to the increasingly accessible and effective suite of methods and software from the field of simulation-based inference (Cranmer et al., 2020).

While simulation-based inference targets domain sciences, advances in the field are mainly driven from a machine learning perspective. The field, therefore, inherits the quality assessments (Lueckmann et al., 2021) customary to the machine learning literature, such as the minimization of classical divergence criteria. Despite recent developments of post hoc diagnostics to inspect the quality of likelihood-free approximations (Cranmer et al., 2015; Brehmer et al., 2018, 2019; Hermans et al., 2021; Lueckmann et al., 2021; Talts et al.,

Toy Ex: Coverage of Prediction and Posterior Intervals Depends on the Choice of Prior

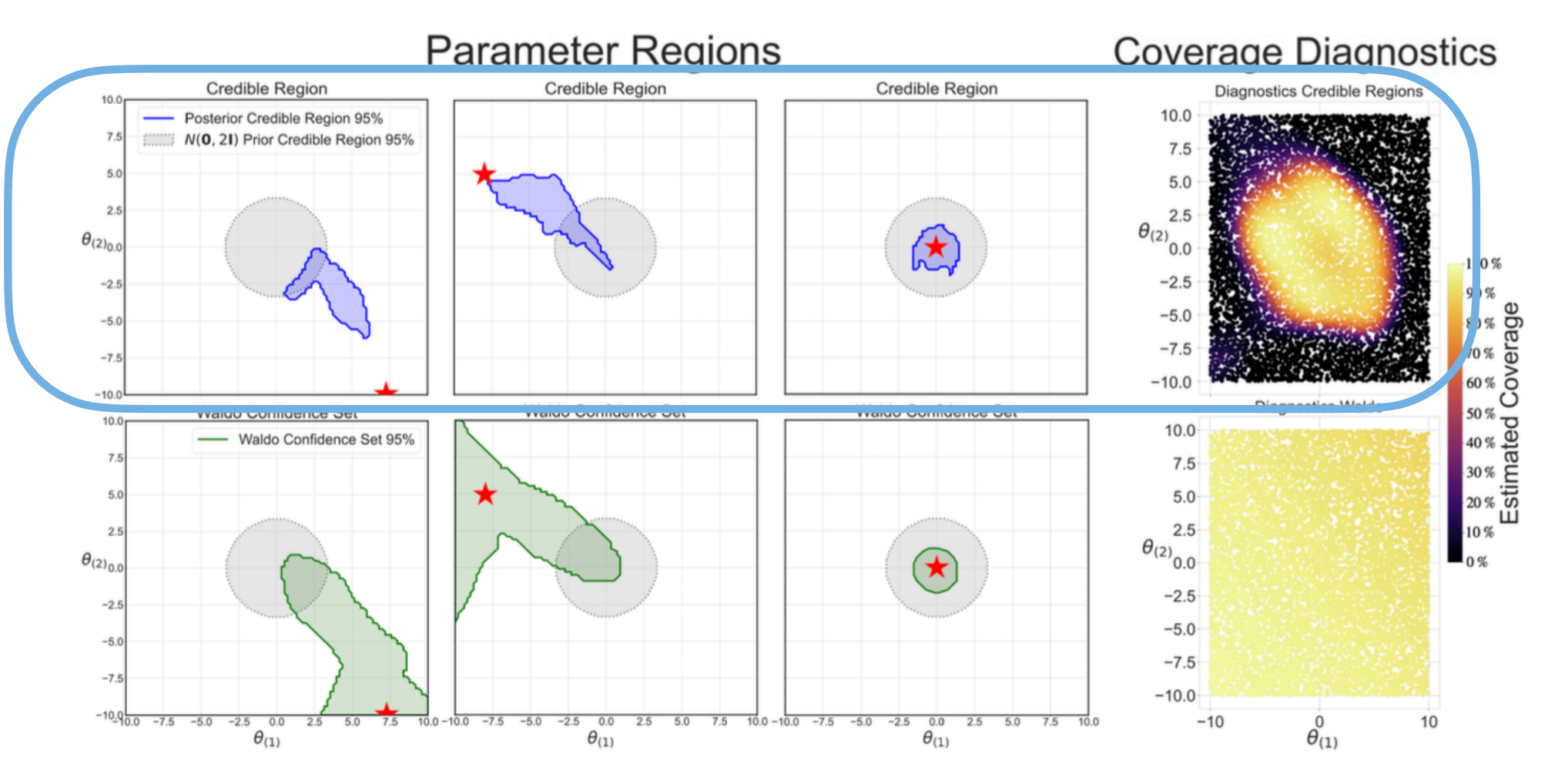
- Likelihood: $\mathscr{D} \mid \theta \sim \mathcal{N}(\theta,\ \sigma = 1)$
- Assume prior: $\theta \sim \mathcal{N}(\mu = 0, \sigma = 2)$

$\Rightarrow$ empirical coverage (green curve)



Ex: Credible Regions from Neural (NF) Posteriors

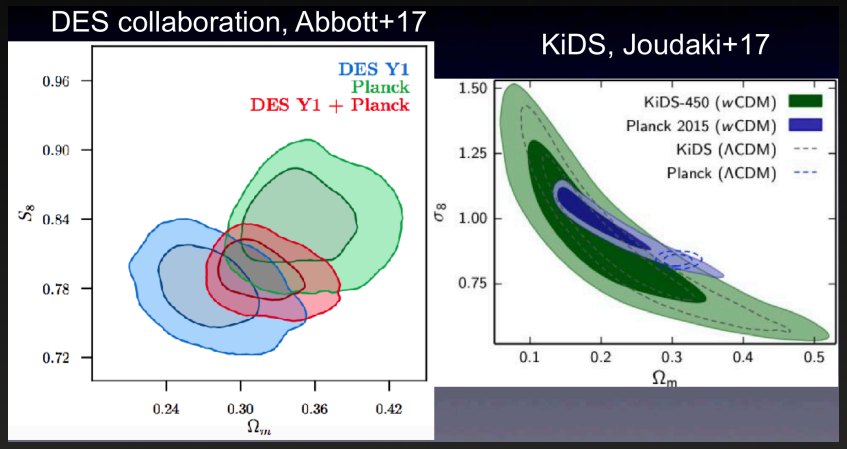
$$\mathcal{D}|\boldsymbol{\theta} \sim \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}) + \tfrac{1}{2}\mathcal{N}(\boldsymbol{\theta}, 0.01 \odot \mathbf{I}), \text{ where } \boldsymbol{\theta} \in \mathbb{R}^2 \text{ and } n = 1$$



Blue contours: 95% credible regions from Normalizing Flows

Overly confident when prior is poorly specified

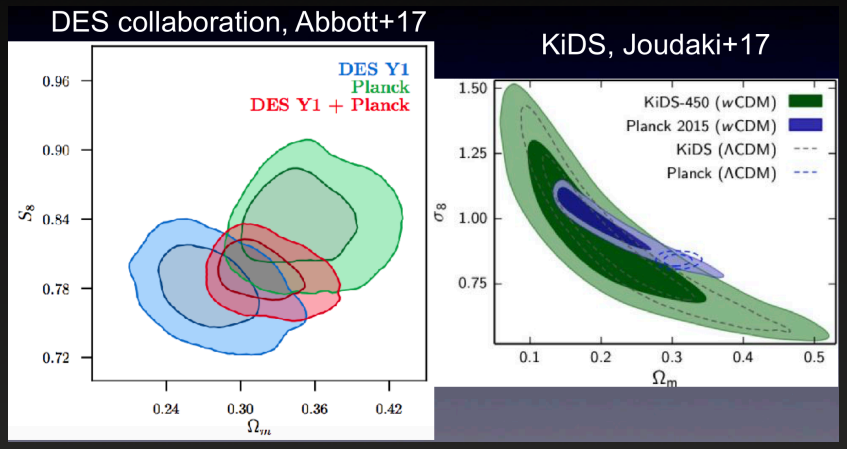
How about Frequentist LFI Approaches?



Guarantee nominal coverage at every θ (regardless of n and design prior)?

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

How about Frequentist LFI Approaches?



Guarantee nominal coverage at every θ (regardless of n and design prior)?

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

- Frequentist approaches (that estimate likelihoods or likelihood ratios) are *by construction* robust to prior prob shift
- However, most such approaches
  - rely on asymptotic assumptions (e.g. Wilks 1938) for downstream inference
  - do not assess validity across entire parameter space
  - do not take advantage of "good" prior information

Can we have it all?

Reliable inference regardless of observed sample size n, even for poorly specified prior.

But higher constraining power if well-specified prior.

Interpretable diagnostics.

* All done by leveraging the arsenal of ML/AI tools “as is” (same network architecture and same loss functions, etc)

Toward a General Inference Machinery for LFI

- Bridges classical statistics with ML to provide:

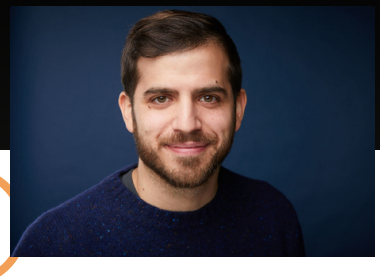
  (i) valid inference: confidence sets with finite-n (e.g. n=1) guarantees of nominal coverage for all parameters

  (ii) practical diagnostics: independent check of actual coverage across entire parameter space (separate from calibration step)

- Goal: Modular procedures with theoretical guarantees.
  - For any test statistic, and any reference or prior distribution
  - Ideally, plug in your favorite SBI algorithm for computing likelihoods, likelihood ratios, posteriors (NPE, NLE, NRE), …

https://arxiv.org/abs/2002.10399 (ICML 2020)

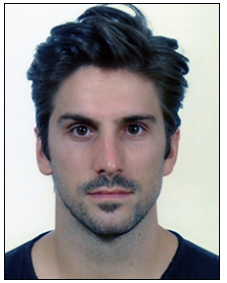
https://arxiv.org/abs/2205.15680 (AISTATS 2023)



LF2I

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

https://arxiv.org/abs/2107.03920



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

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

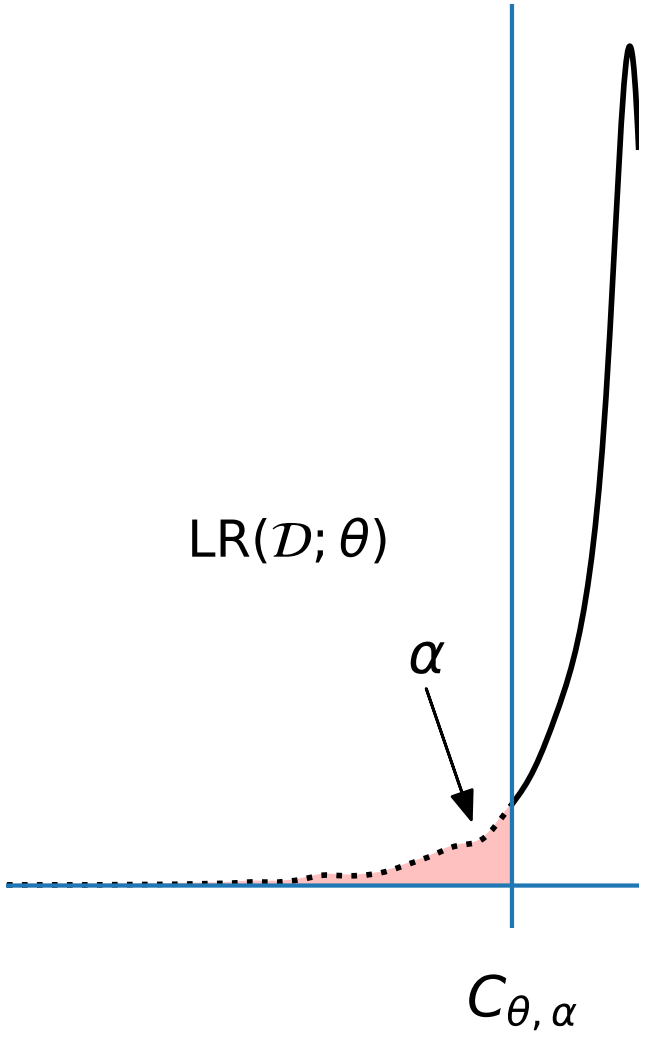
Theorem (Neyman 1937)

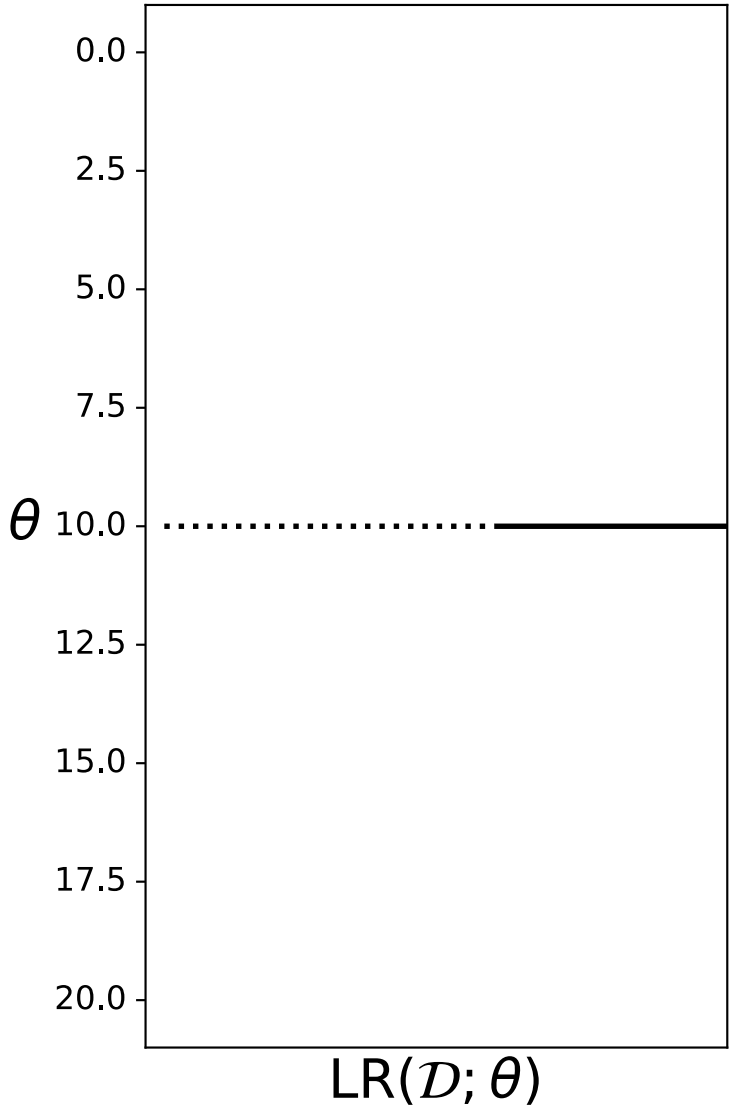
Constructing a $1 - \alpha$ *confidence set for* θ *is equivalent to testing*

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

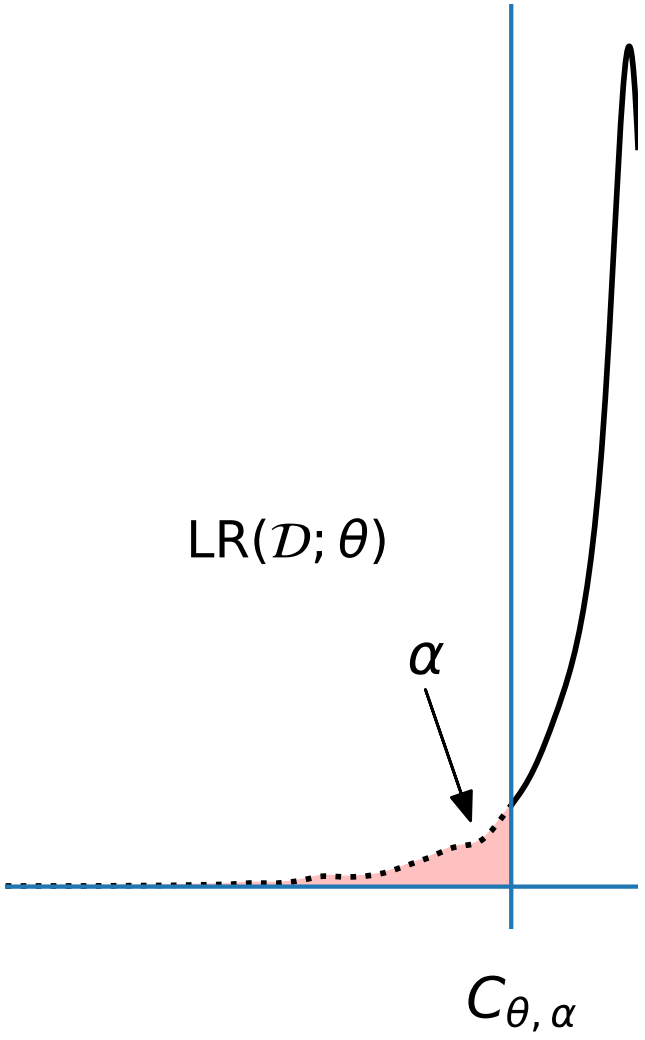
for every $\theta_0 \in \Theta$.

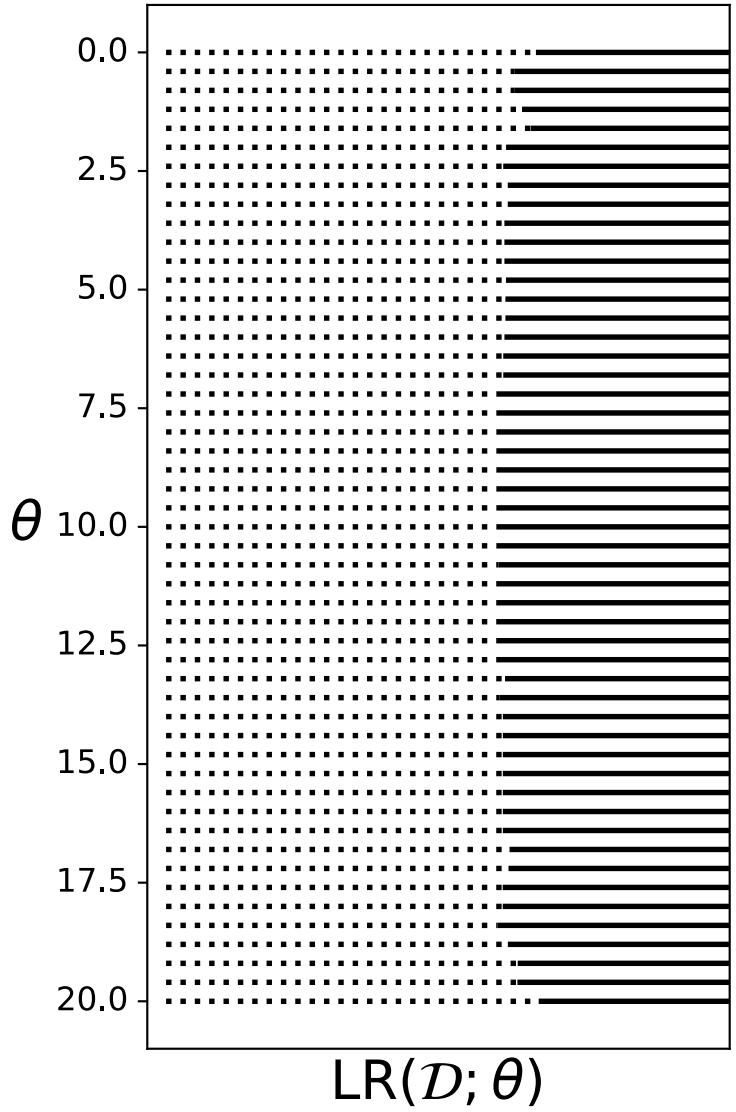
1. Fixed θ. Find the rejection region for test statistic λ.



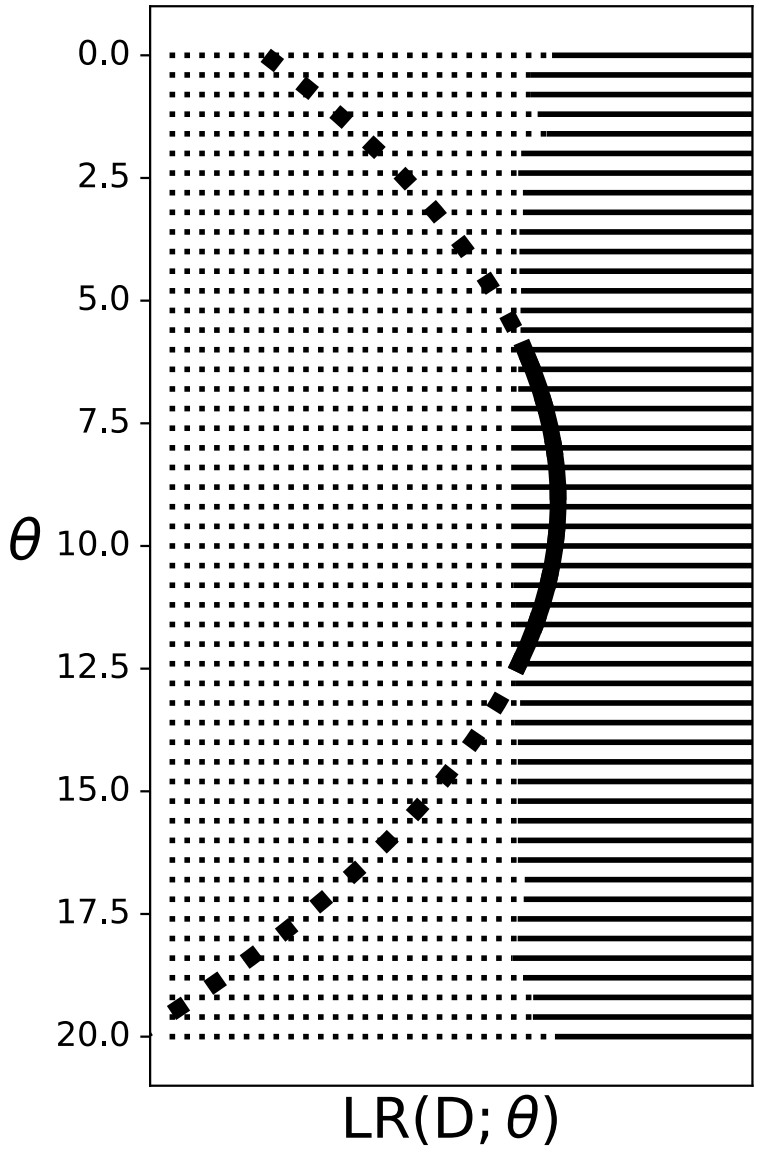


2. Repeat for every θ in parameter space.

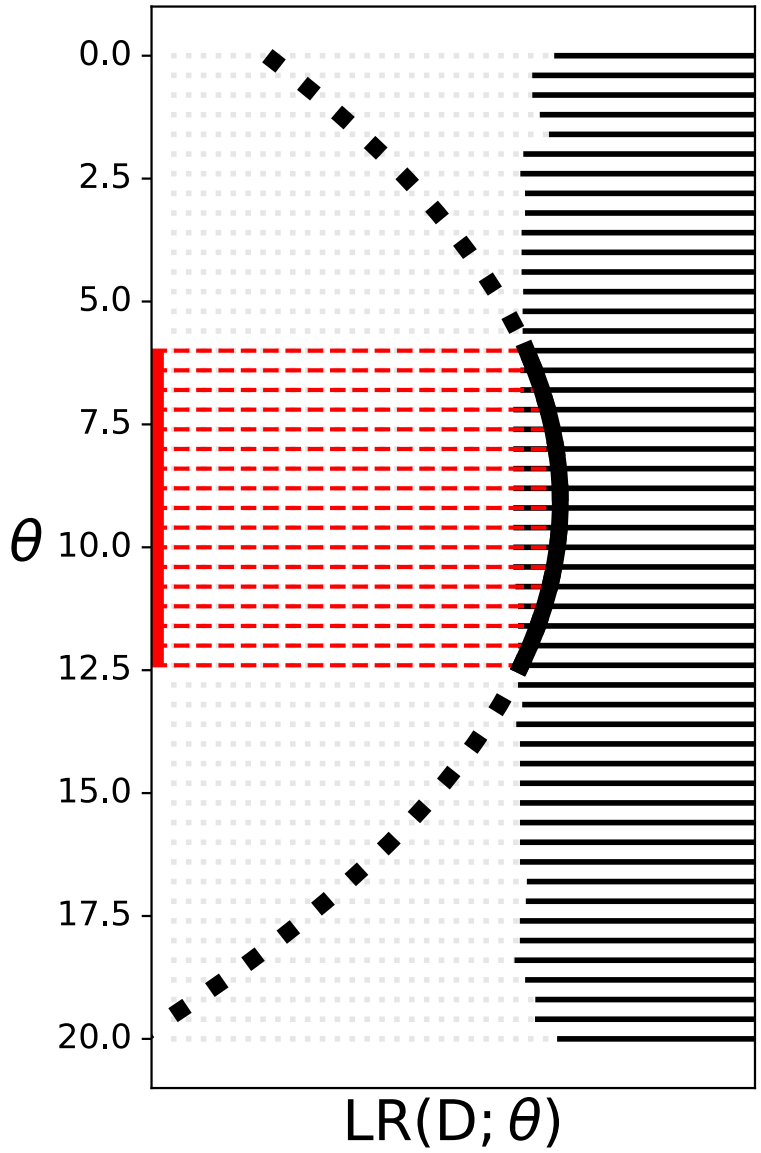




3. Observe data $\mathcal{D} = \mathrm{D}$. Evaluate $\lambda(\mathrm{D}; \theta)$.



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, 2018:

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

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

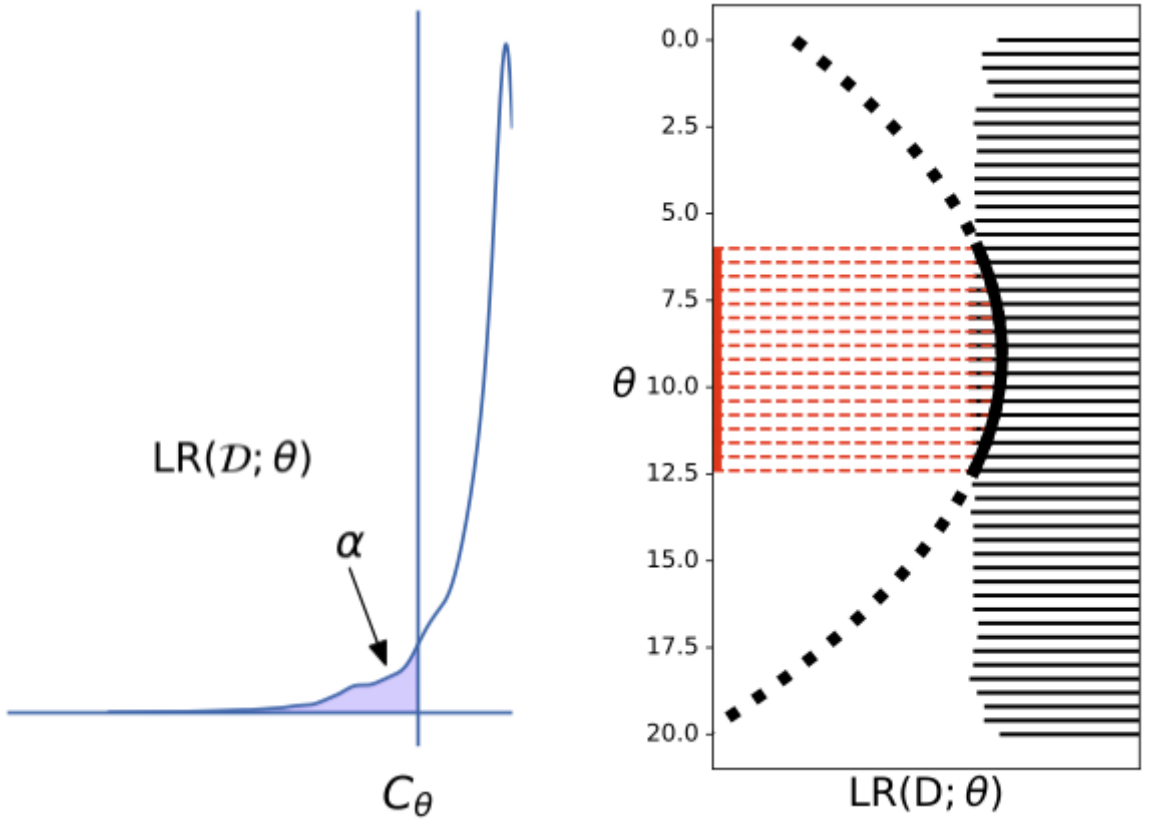
for **every** $\theta_0 \in \Theta$.

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

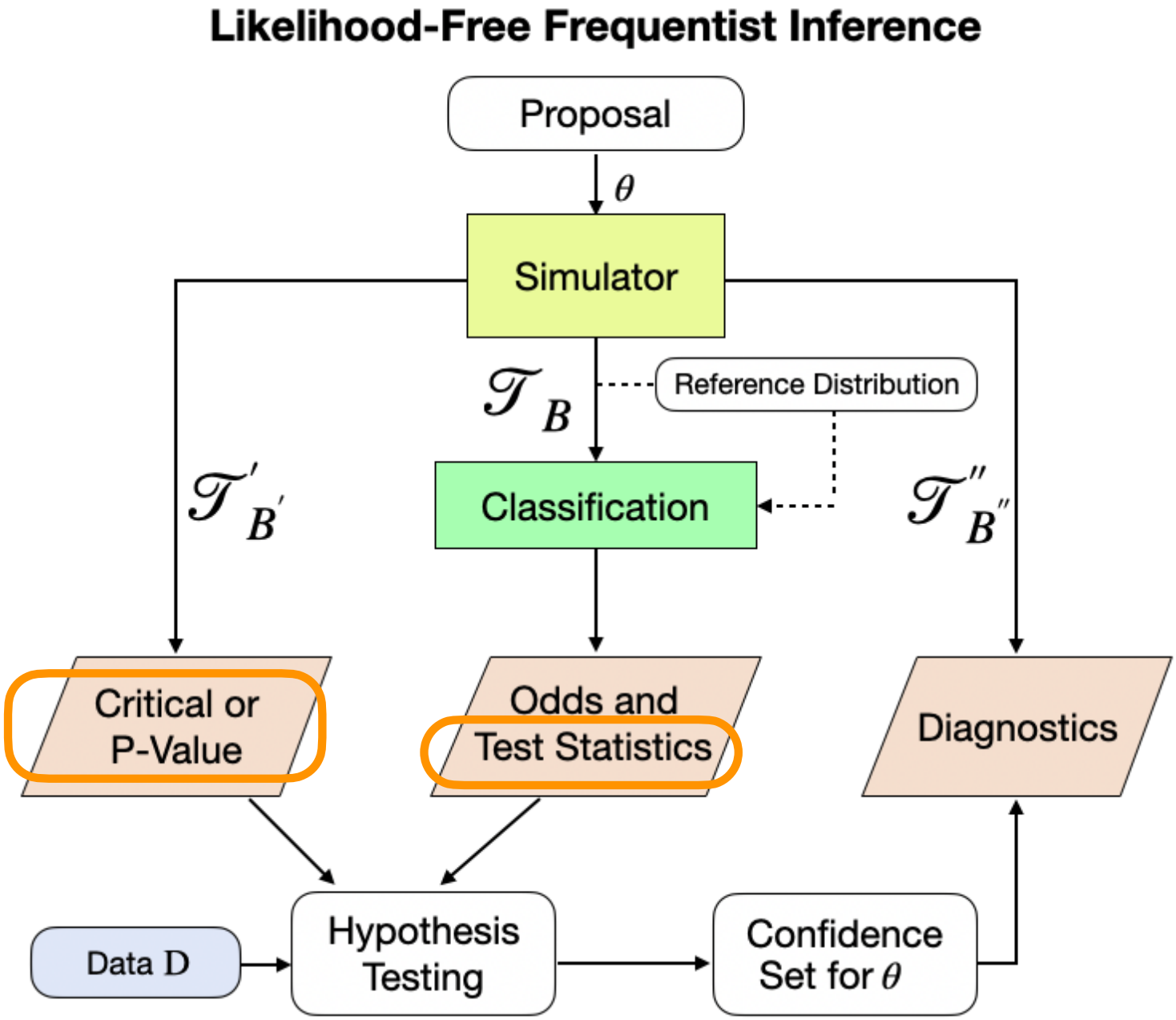
are **conditional distribution functions** of the (unknown) parameters, and often vary smoothly across the parameter space Θ.

Efficient Construction of Finite-Sample Confidence Sets

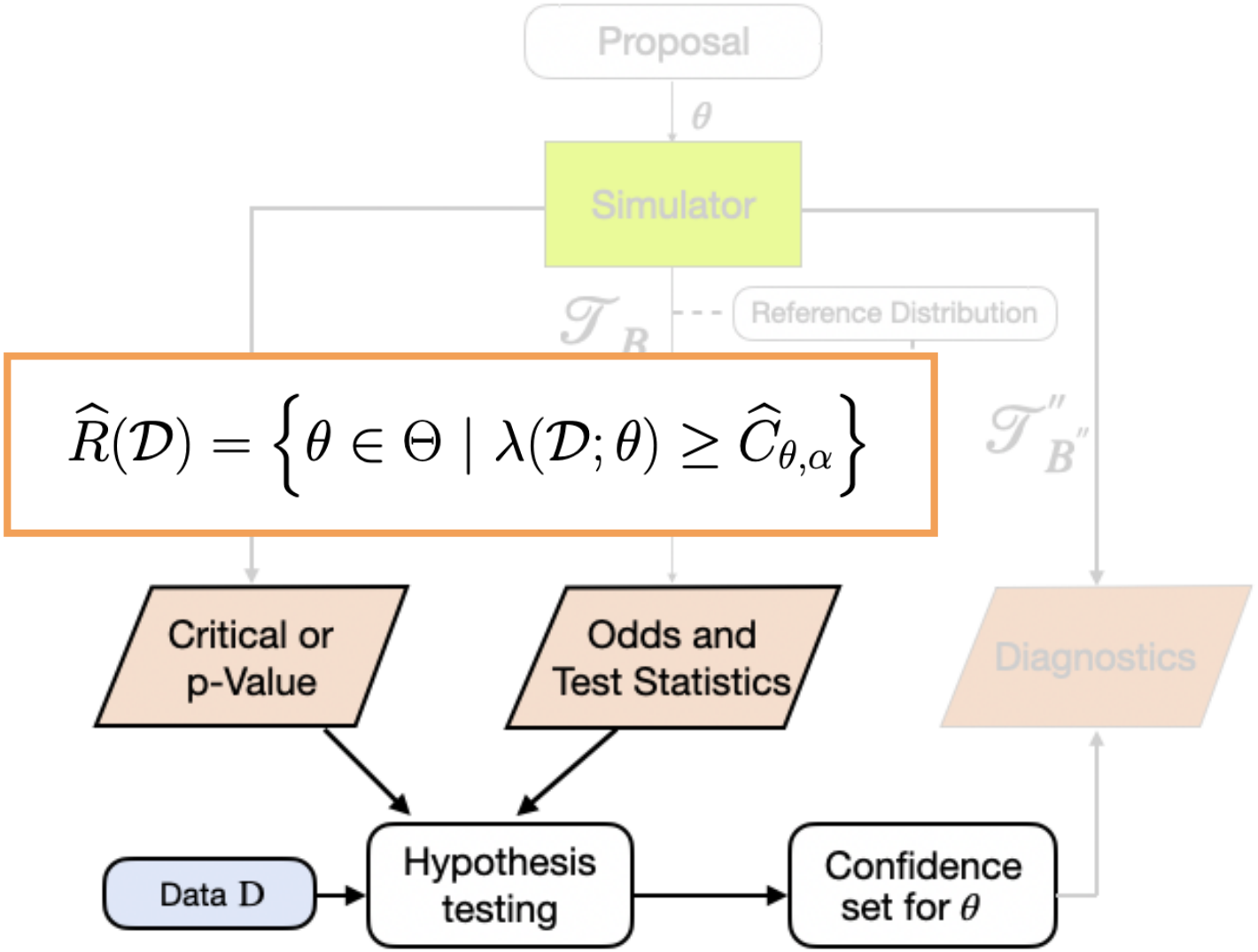


Rather than running a batch of Monte Carlo simulations for every null hypothesis $\theta = \theta_0$ on, e.g., a fine enough grid in Θ, we can interpolate across the parameter space using training-based ML algorithms.

Our Inference Machinery



Construct Confidence Set via Neyman Inversion



Test Statistics: Leverage ML Classification/ Prediction Algorithms

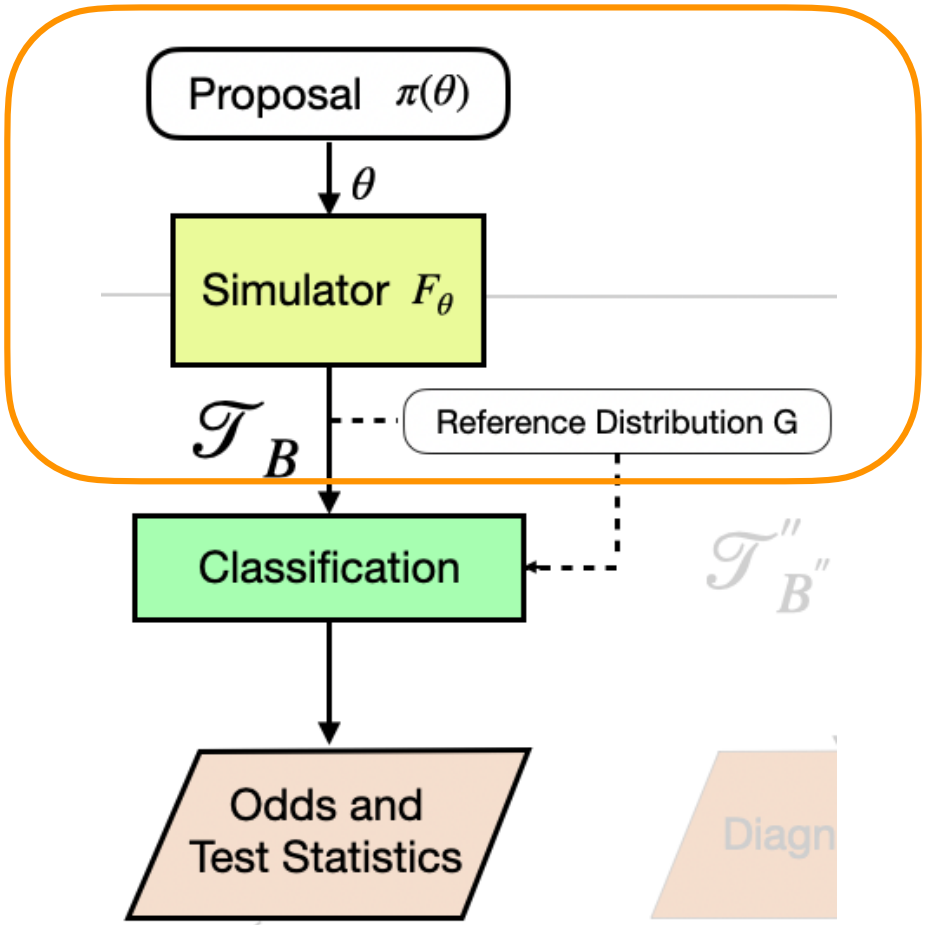
- Estimate "odds function" (from parameters to data)
  - → **ACORE** (approximate LRT) [Izbicki et al 2013; Cranmer et al 2015; Dalmasso et al 2020, arXiv:2002.10399]
  - → **BFF** (approximate Bayes Factor) [Dalmasso et al 2021, arXiv:2107.03920; Heinrich 2022, arXiv: 2203.13079]
- Obtain point predictions or posterior estimates (from data to parameters)
  - → **WALDO** (modified Wald test statistic) [Masserano et al 2022, arXiv:2205.15680]
  - → Frequentist Bayes sets [Masserano, Shen et al 2023-]

Center Branch: Estimating Odds and Test Statistic

[Izbicki et al 2013; Dalmasso et al 2020]

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):

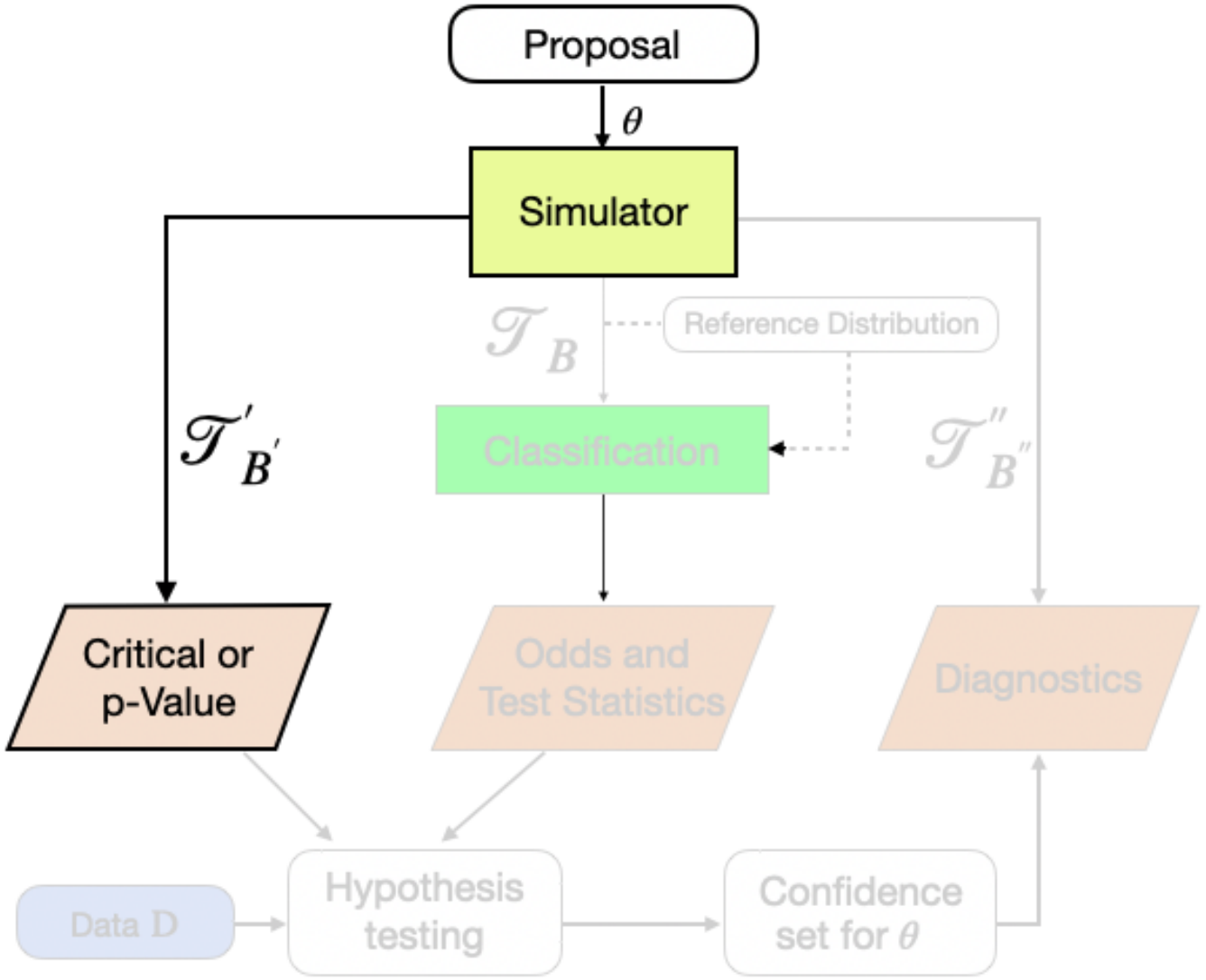
$$\widehat{\Lambda}(\mathcal{D}; \theta_0) := \log \frac{\prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta_0)}$$

- 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.

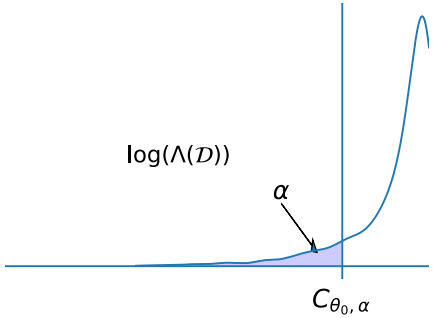
Left Branch: Estimate Critical Values or P-Values



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

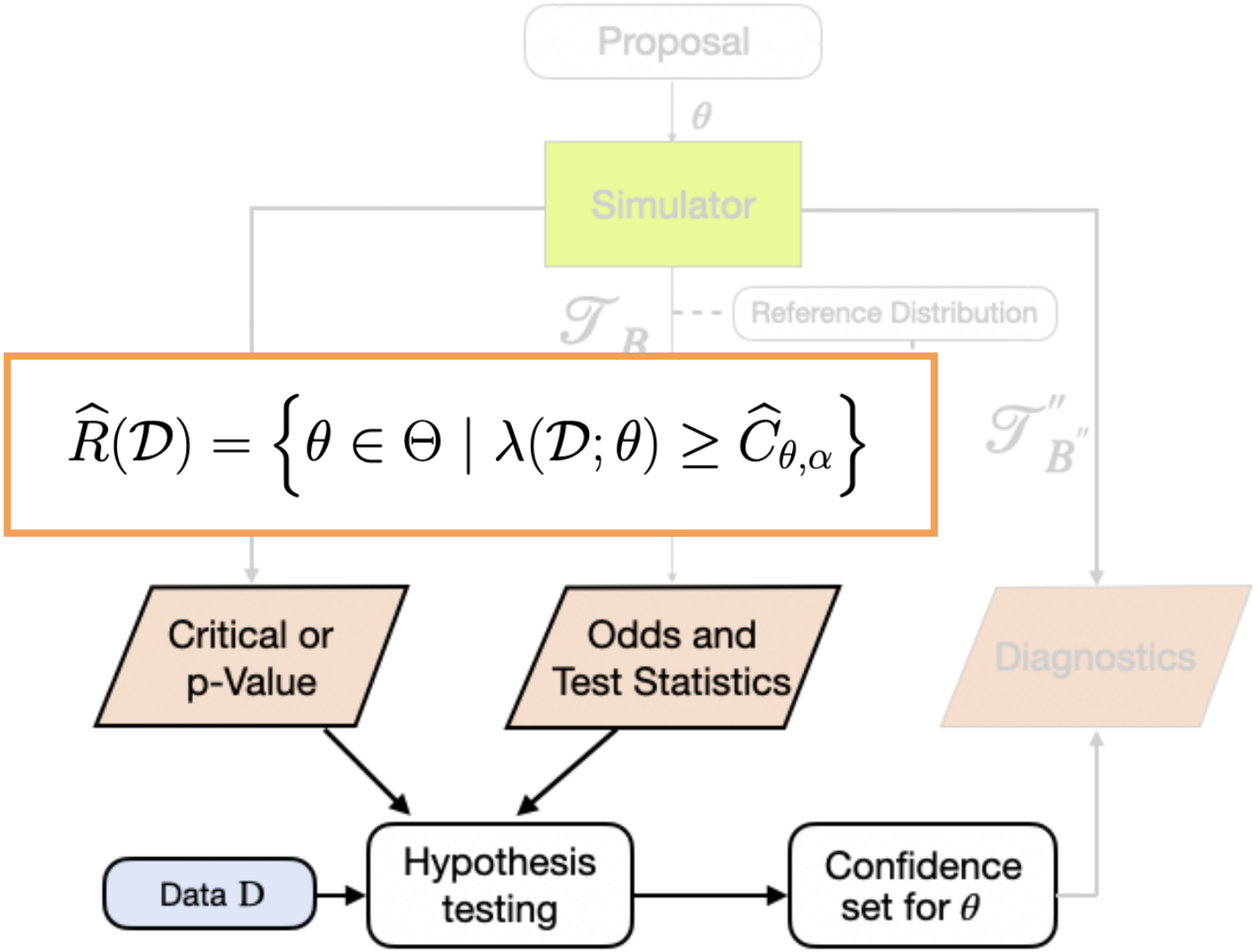


$$C_{\theta_0,\alpha} = \arg\sup_{C\in\mathbb{R}} \left\{ C : \mathbb{P}_{\mathcal{D}|\theta_0}\left(\lambda(\mathcal{D};\theta_0) < C\right) \leq \alpha \right\}.$$

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.

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

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

NOTE: Regardless of the number of observations n, how well we estimate the test statistic, and the proposal distribution r(θ)

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

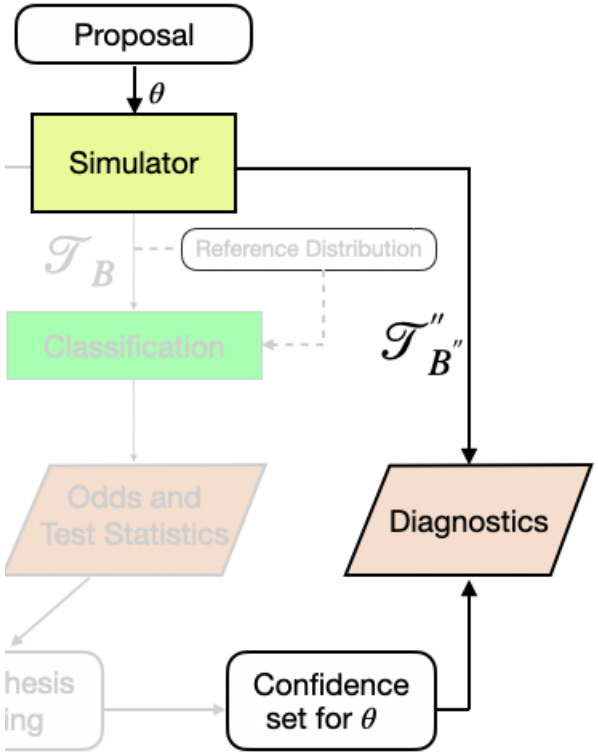
Right Branch: Assessing Conditional Coverage of $\widehat{R}(\mathcal{D})$

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 $1-\alpha$?

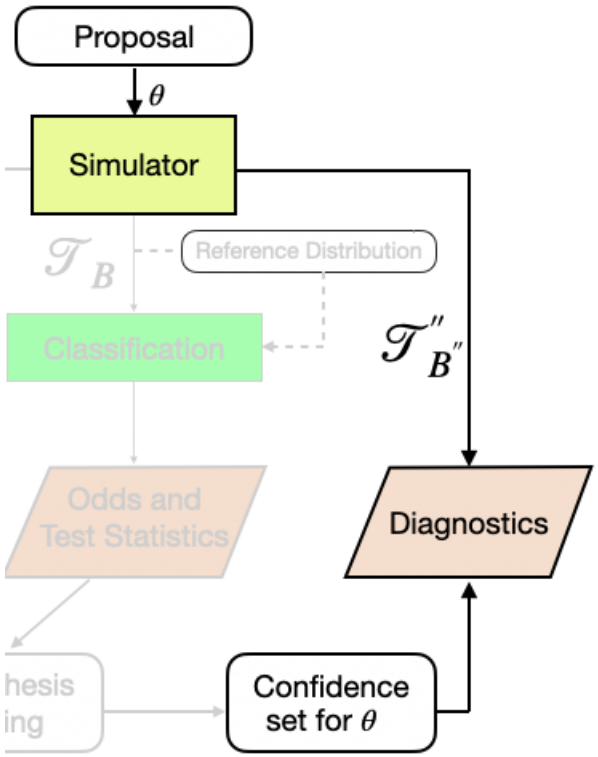
Right Branch: Assessing Conditional Coverage of $\widehat{R}(\mathcal{D})$

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

Note:

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

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



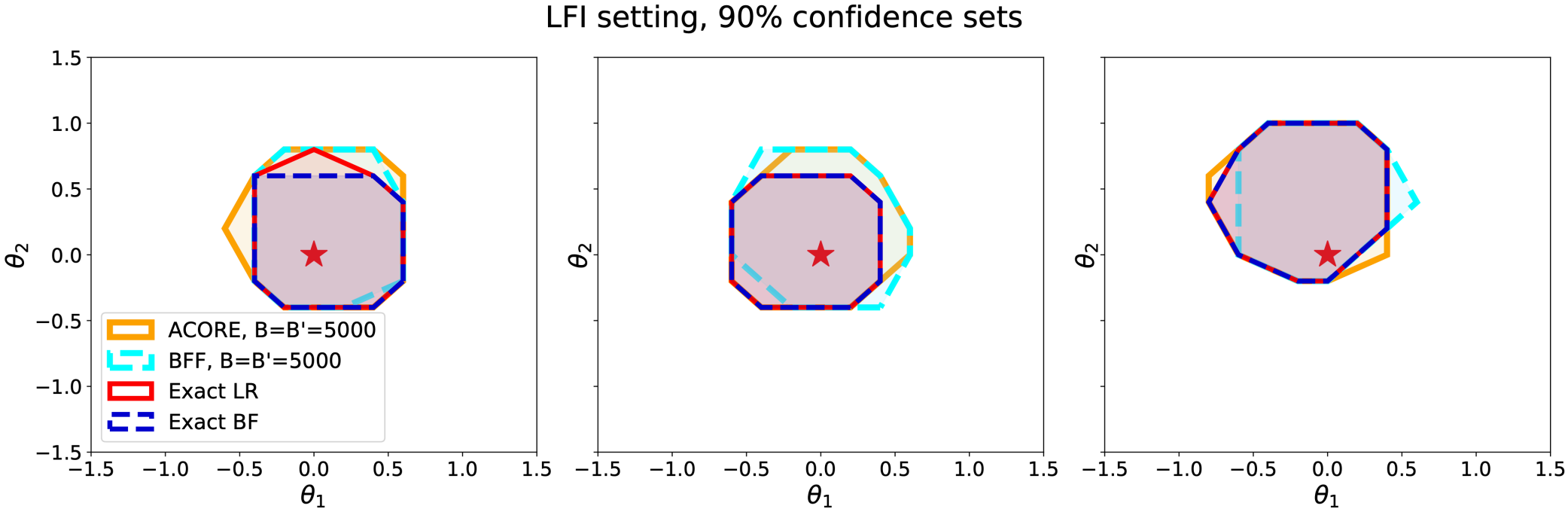
1. Sample θ_i and data $\mathcal{D}_i \sim F_{\theta_i}$
2. Construct confidence set $\widehat{R}(\mathcal{D}_i)$
3. For $\{\theta_i, \widehat{R}(\mathcal{D}_i)\}_{i=1}^{B''}$, regress $Z_i := \mathbb{I}(\theta_i \in \widehat{R}(\mathcal{D}_i))$ on θ_i.

Independent check of coverage across parameter space

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

Ex 1 (one MVG): Construct LF2I Confidence Sets

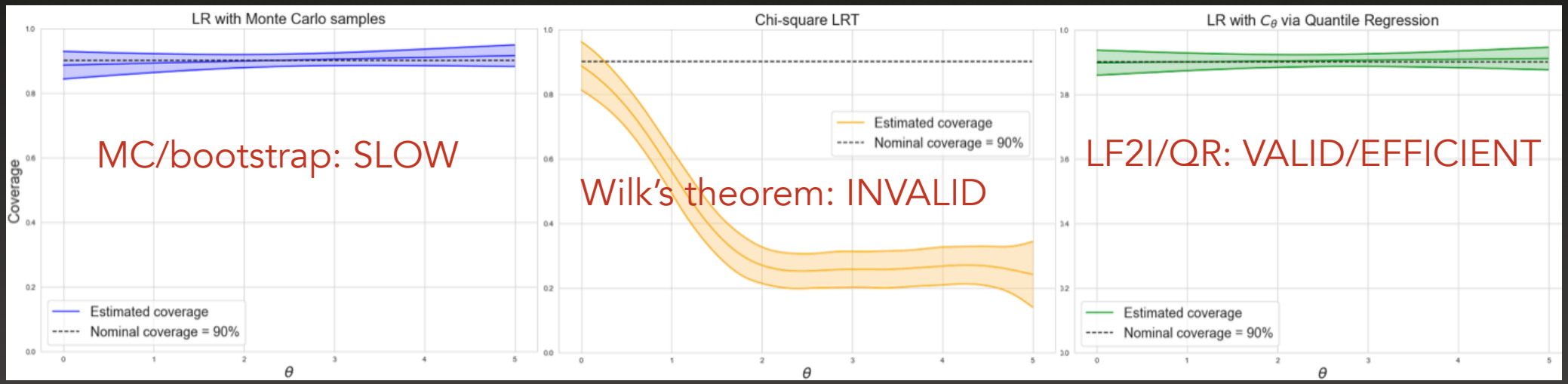
$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N(\boldsymbol{\theta}, \mathbf{I}_d), \text{ where } n = 10, \ \boldsymbol{\theta} = \mathbf{0}$$



When d=2, ACORE and BFF confidence sets (for B=B'=5000) are similar in size to the Exact LR confidence sets. (LF2I scales well if d<10)

Ex 2 (Gaussian Mixture): The distribution of the LR statistic is not known. Valid inference with nominal coverage (n=1000)?

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



(Left) LR with 1000 MC simulations at each θ on a fine grid in 1D

(Center) Assume chi-squared distribution of LR statistic

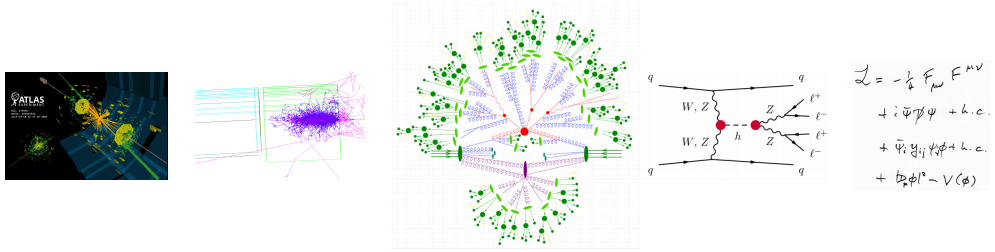
(Right) LR with quantile regression with B'=1000 simulations total

But what if we have >1,000 nuisance parameters?

The parameters θ

One more issue: the "theory" space is not the only thing effecting the data

- **every step of the forward process comes with its own parameters**
  (we understand the process generally but need additional knobs to model the data)



$$p(x|\theta) = \int \mathrm{d}z_d \mathrm{d}z_h \mathrm{d}z_p \quad p(x|z_d,\theta_x)\; p(z_d|z_h,\theta_d)\; p(z_h|z_p,\theta_h)\; p(z_p|\theta_p,\theta_{\mathrm{th}})$$

$p(x|z_d)$ $p(z_d|z_h)$ $p(z_h|z_p)$ $p(z_p|\theta)$

nuisance parameters

core "theory" parameters of inferest (e.g. "Higgs Mass"

12

Credit: Lukas Heinrich

How do we Handle Nuisance Parameters?

In many applications, the parameter space can be decomposed as $\Theta = \mathcal{M} \times \mathcal{N}$, where $\mathcal{M}$ contains the *main parameters* μ of interest, and $\mathcal{N}$ contains the *nuisance parameters* ν not of immediate interest.

Suppose we want to test

$$H_{0,\mu_0} : \mu = \mu_0 \quad \text{versus} \quad H_{1,\mu_0} : \mu \neq \mu_0 \quad \text{for } \mu_0 \in \mathcal{M}$$

How does one solve this problem within our inference machinery?

But Critical Value Estimation is Difficult with Many NPs

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

$$H_{0,\mu_0} : \mu = \mu_0 \quad \text{versus} \quad H_{1,\mu_0} : \mu \neq \mu_0 \quad \text{for } \mu_0 \in \mathcal{M}$$

by comparing test statistics to the cutoffs $\widehat{C}_{\mu_0} := \inf_{\nu \in \mathcal{N}} \widehat{C}_{(\mu_0,\nu)}$.

That is, one needs to control the type I error at each μ_0 for *all* possible values of the nuisance parameters.

Can lead to numerically unwieldy and costly computations if the number of nuisance parameters is large (>10 NPs).

Two Popular Approaches to Systematics

Hybrid Approaches to Critical Value Estimation

- h-ACORE: Hybrid Resampling or Profiling[1] of Nuisance Parameters
  - Compare ACORE test statistic with the *hybrid cut-off*

$$\widehat{C}'_{\mu_0} := \widehat{F}^{-1}_{\Lambda(\mathcal{D};\mu_0)\big|(\mu_0,\widehat{\nu}_{\mu_0})}\left(\alpha \,\big|\mu_0,\widehat{\nu}_{\mu_0}\right)$$

  where the quantile regression is based on a train sample $\mathcal{T}'$ generated at *fixed* $\widehat{\nu}_{\mu_0}$.

- h-BFF: Integration of Nuisance Parameters
  - Compare BFF test statistic with the *approximate cut-off*

$$\widehat{C}'_{\mu_0} := \widehat{F}^{-1}_{\tau(\mathcal{D};\mu_0)\big|\mu_0}\left(\alpha \,\big|\, \mu_0\right)$$

  where we draw the train sample $\mathcal{T}'$ from the entire parameter space $\Theta = \mathcal{M} \times \mathcal{N}$, but apply quantile regression *using* μ *only*

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

Assessing Confidence Sets

- *"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)
- Our LF2I diagnostic tool can
  - provide guidance as to which method to choose for the problem at hand, and
  - pinpoint regions of parameter space where inference may be unreliable, e.g., under/over-confident.
- The diagnostic branch works for any SBI method (including Bayesian credible regions)

Classical "On-Off" Problem

[Lyons 2008; Cowan et al 2011; Cowan 2012; L. Heinrich 2022]

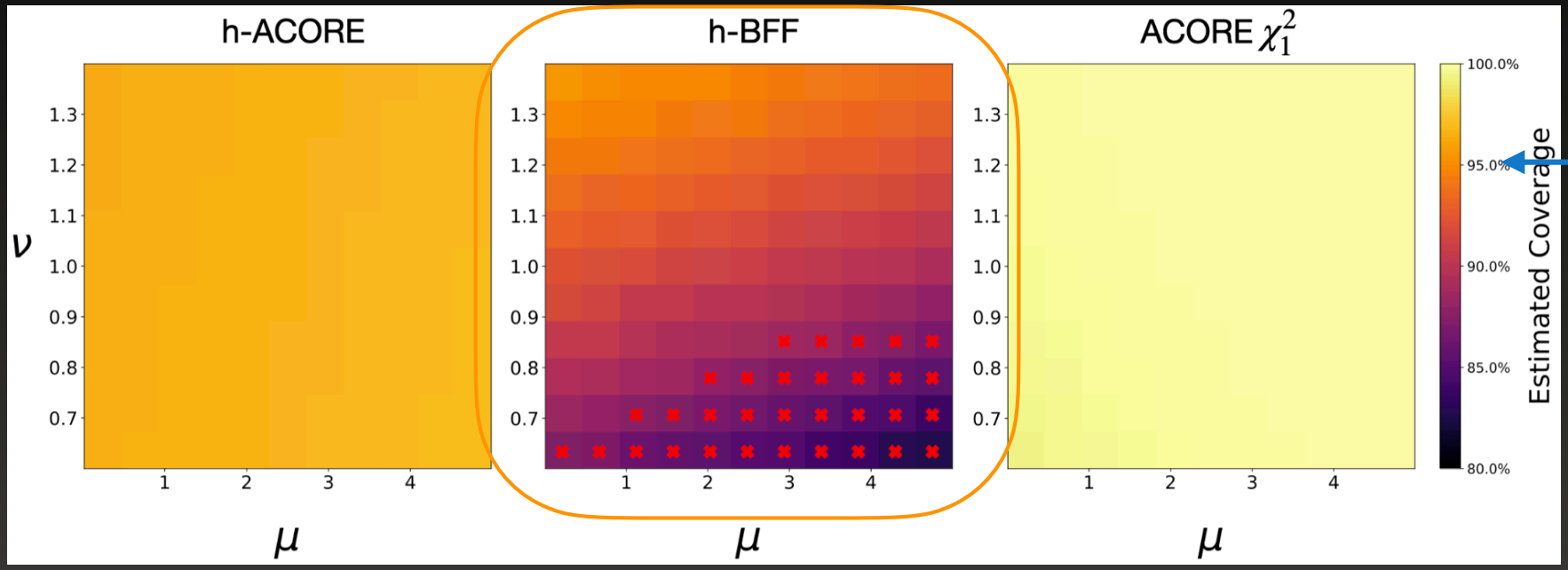
- Simultaneous measurements of two Poisson processes

$$\text{Observed data } \mathbf{X} = (N_b, N_s),$$
$$\text{where } N_b \sim \text{Pois}(\nu\tau b),\ N_s \sim \text{Pois}(\nu b + \mu s)$$

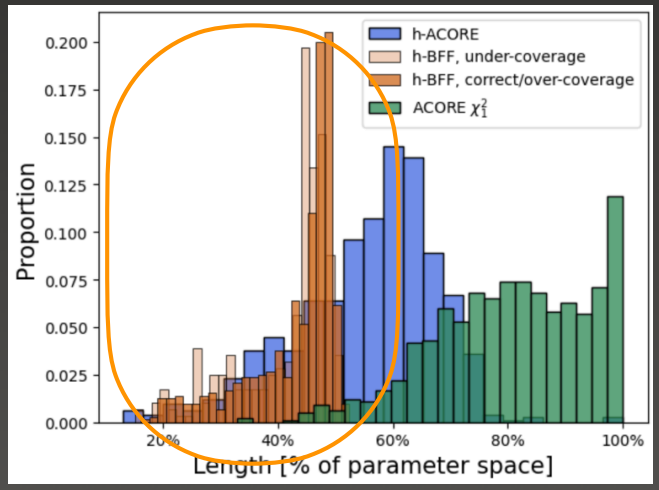
- N_B is the # of events in the background region (expected background count b)
- N_S is the # of events in the contaminated signal region (expected signal count s)
- Unknown parameters:
  - signal strength-POI (μ); scaling factor-NP (ν)
  - [L. Heinrich 2022] Set hyper-parameters at s=15, b=70, τ=1 ⇒ asymptotic regime with profiled values away from the MLE

Our diagnostic tool can identify regions in parameter space with under/over-coverage (95% nominal)

Left: profiling; Center: marginalization; Right: chi-square)



h-BFF (center top) has closest to nominal coverage with the highest constraining power (orange hist)



What's Next?

- Alternative test statistics based on direct predictions and posteriors
  - Reason 1: large arsenal of AI tools for prediction and NPEs. LR trick + maximization over many parameters sometimes hard to implement in practice (loss of power)
  - Reason 2: LR approaches do not benefit from good priors
  - "Freq Bayes sets" (in progress). Show some highlights of WALDO (1st version, 2022)

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

Ex 1: Back to the Problem of Calorimetric Muon Energy Measurement… [Masserano et al, AISTATS 2023]

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

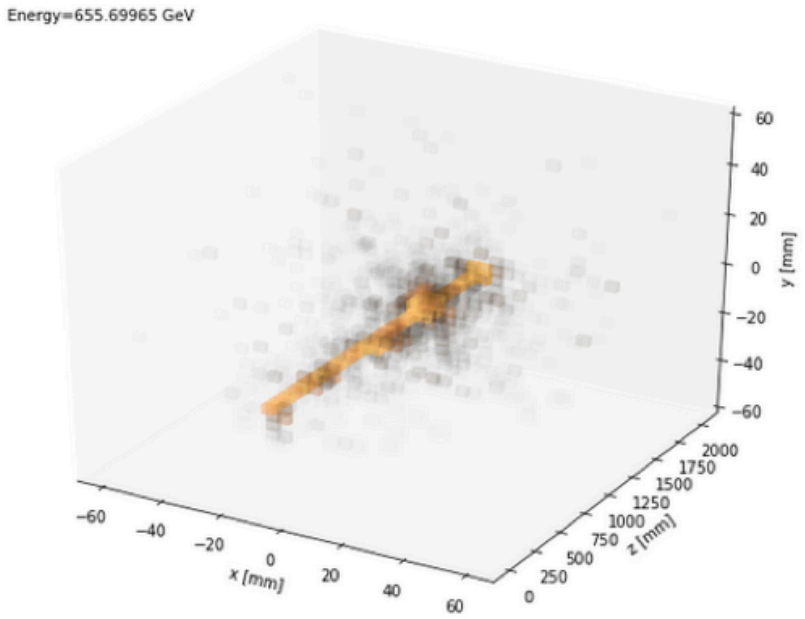


Figure 4: Muon entering the calorimeter in z direction.

1. Bias

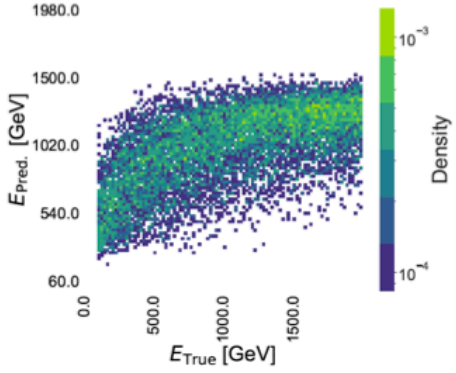


Figure 9: 2D histogram of uncorrected kNN prediction versus true energy for test data.

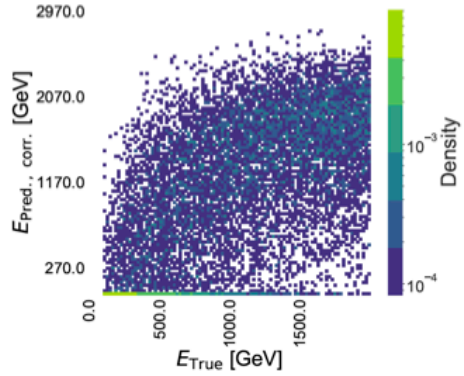


Figure 10: 2D histogram of corrected kNN prediction versus true energy for test data.

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

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

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

https://arxiv.org/abs/2205.15680

Simulation-Based Inference with WALDO: Confidence Regions by Leveraging Prediction Algorithms or Posterior Estimators for Inverse Problems

Luca Masserano[1] **Tommaso Dorigo**[2] **Rafael Izbicki**[3] **Mikael Kuusela**[1] **Ann B. Lee**[1]

[1]Department of Statistics & Data Science, Carnegie Mellon University

[2]INFN, Sezione di Padova [3]Department of Statistics, Federal University of São Carlos

Abstract

Predictive algorithms, such as deep neural net-

1 INTRODUCTION

The vast majority of modern machine learning targets pre-

of a data-generating process with reliable measures of

uncertainty. The parameters of interest, which we denote by

certainty quantification, especially when both pa-

sample theory. Many simulator-based inference

(SBI) methods are indeed known to produce bi-

denote observable data, where the "sample size" n refers

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 $\theta_0 \in \Theta$.

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

- **Wald** test statistic (1D case):

$$\tau^{Wald}(\mathscr{D}; \theta_0) := \frac{(\theta^{MLE} - \theta_0)^2}{\mathbb{V}[\theta^{MLE}]}$$

- **Waldo** test statistic (1D and p-D case):

$$\tau^{Waldo}(\mathscr{D}; \theta_0) := \frac{(\mathbb{E}[\theta \,|\, \mathscr{D}] - \theta_0)^2}{\mathbb{V}[\theta \,|\, \mathscr{D}]}$$

$$\tau^{Waldo}(\mathscr{D}; \theta_0) := (\mathbb{E}[\theta \,|\, \mathscr{D}] - \theta_0)^T \mathbb{V}[\theta \,|\, \mathscr{D}]^{-1} (\mathbb{E}[\theta \,|\, \mathscr{D}] - \theta_0)$$

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, or 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

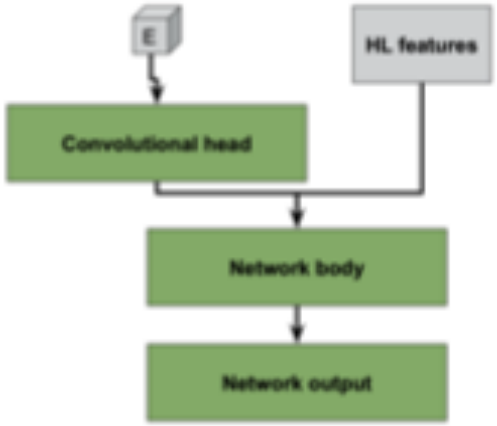


Image credit: Kieseler et al. (2022)

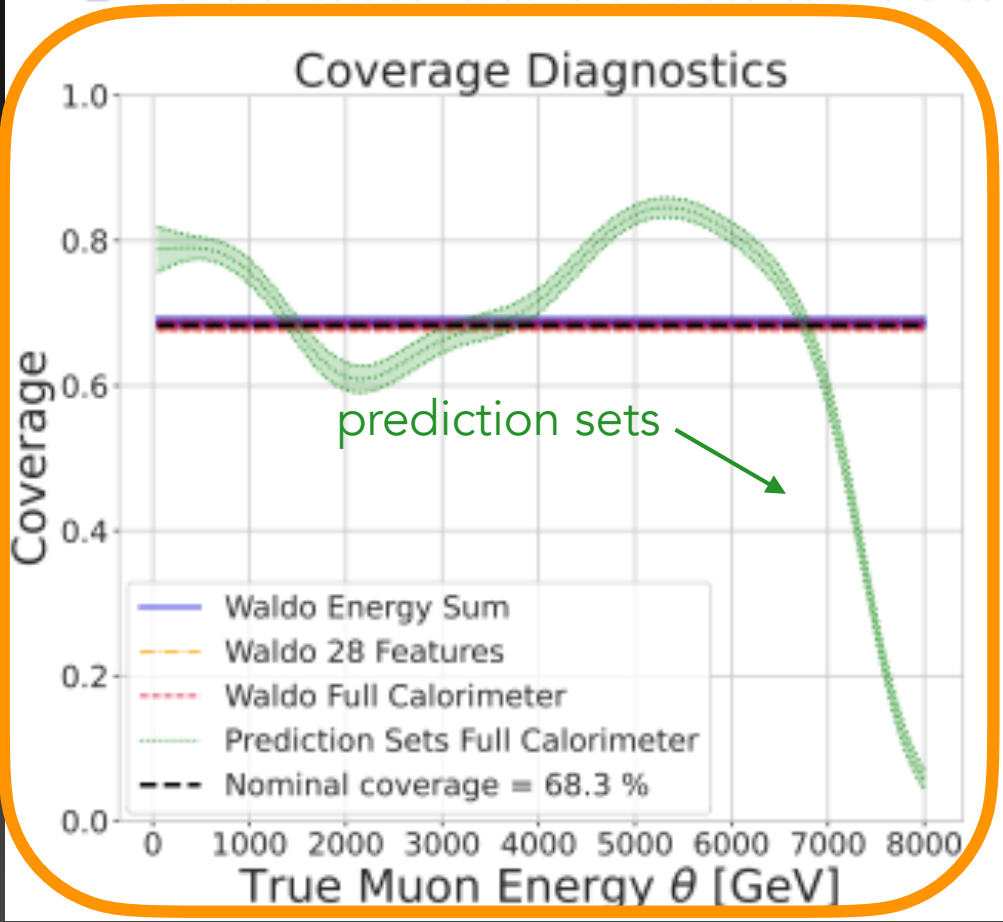
$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

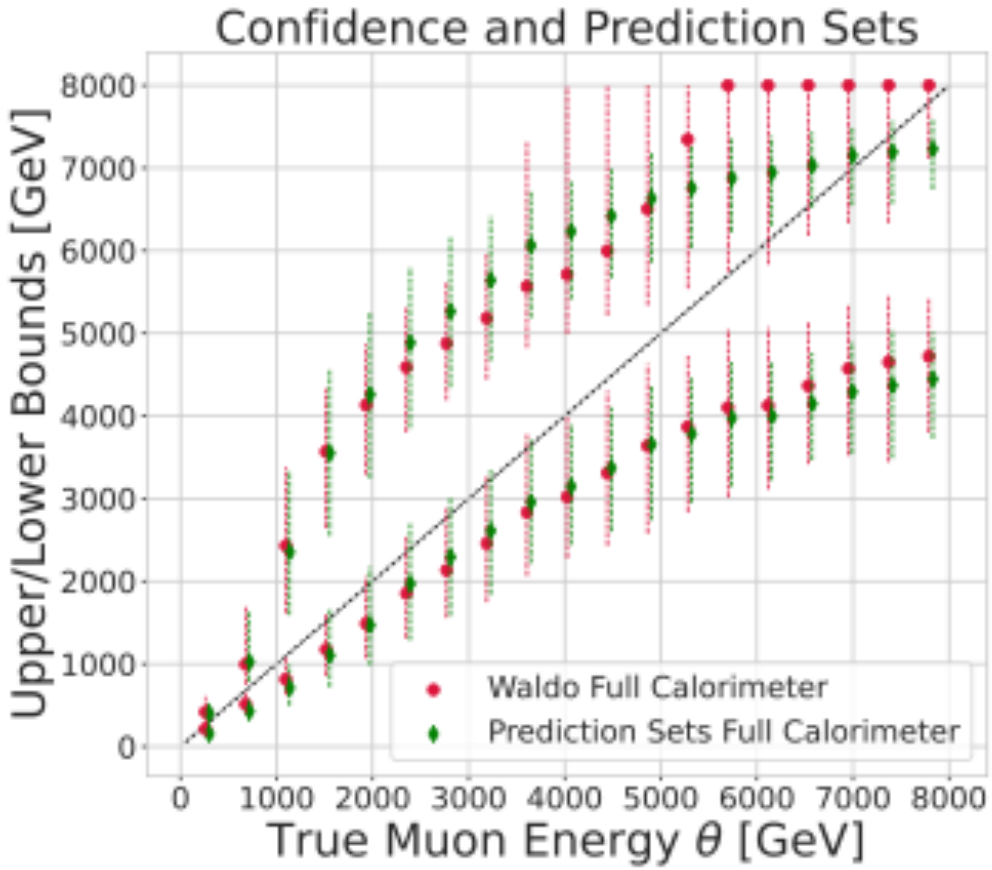
Slide credit: Luca Masserano

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

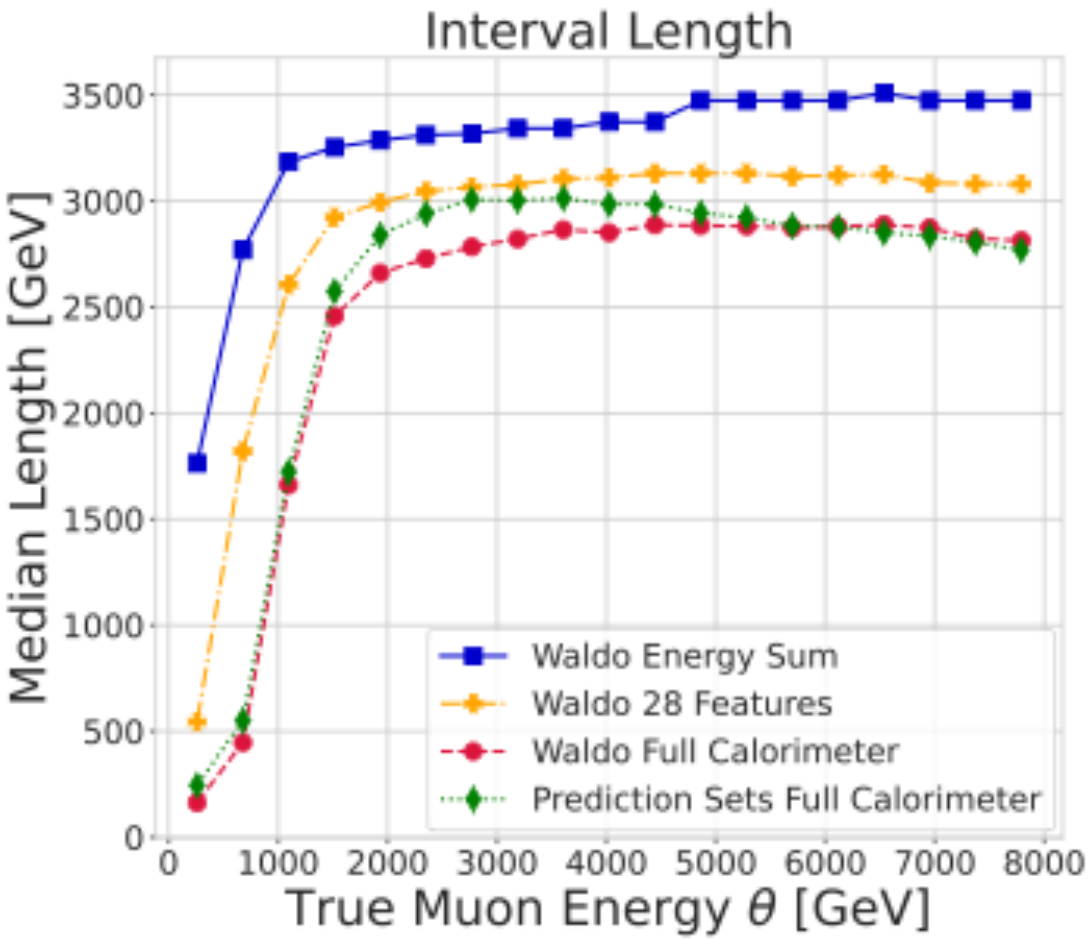




Slide credit: Luca Masserano

Constraining power?

Valuable information in high-granularity calorimeter

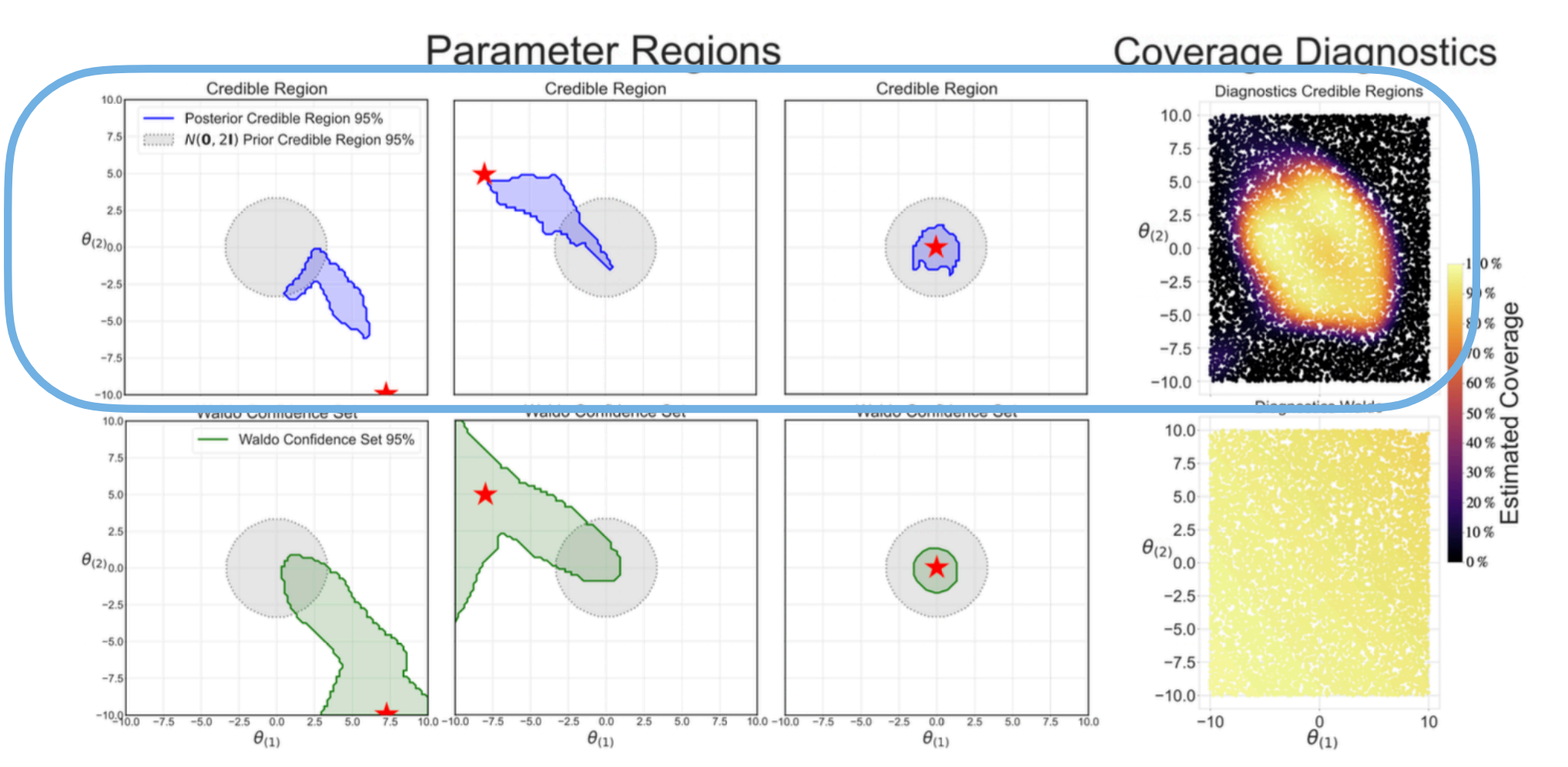


- Intervals are shorter as the data becomes higher-dimensional
- Prediction sets can even be larger than Waldo confidence sets (while also not guaranteeing coverage)

Slide credit: Luca Masserano

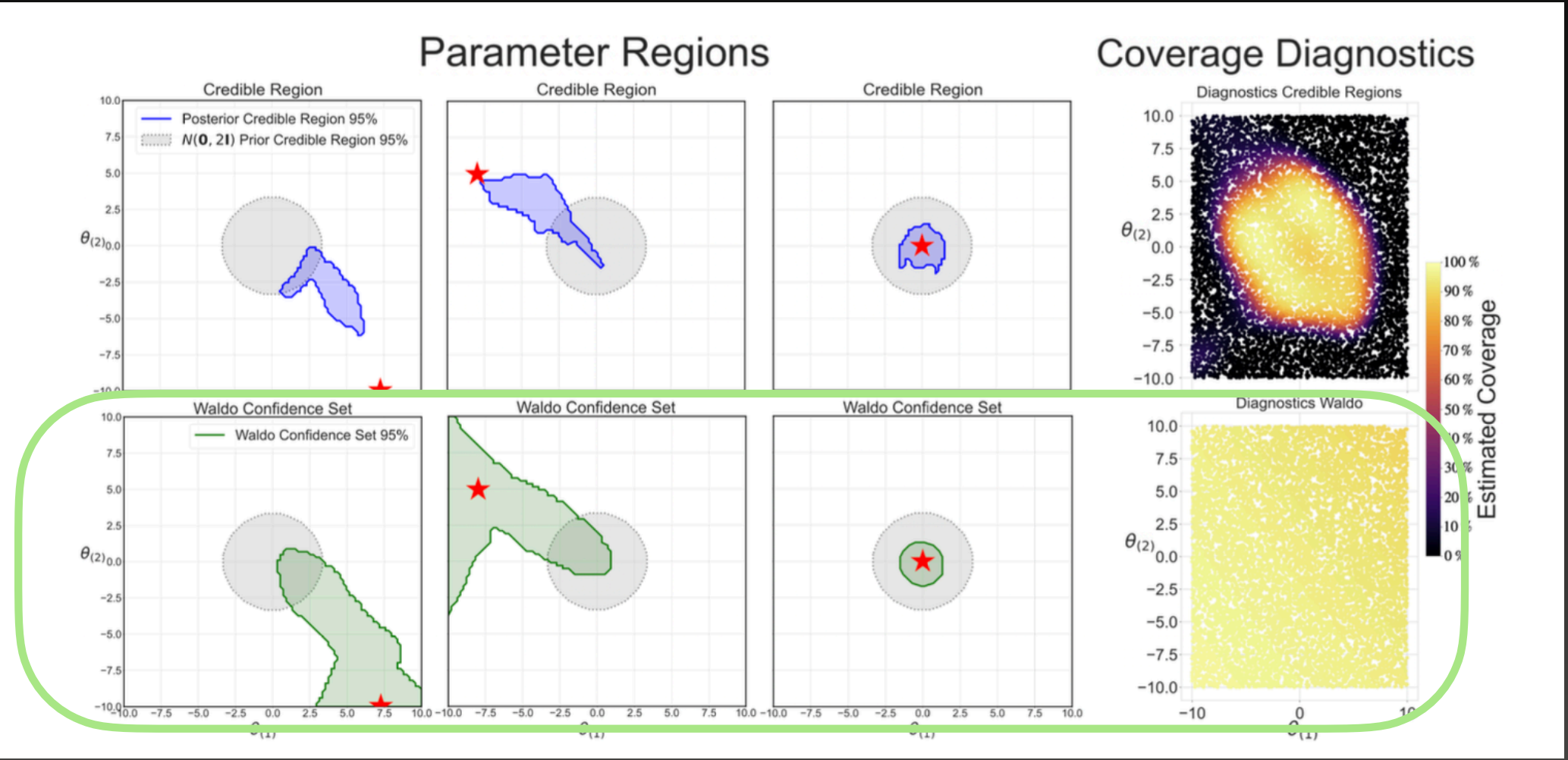
Ex 2: Credible Regions from Neural (NF) Posteriors

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



Blue contours: 95% credible regions from Normalizing Flows (overly confident when prior is poorly specified)

Ex: LF2I/Waldo Confidence Sets Derived from the Same Neural Posteriors ⇨ Correct Coverage

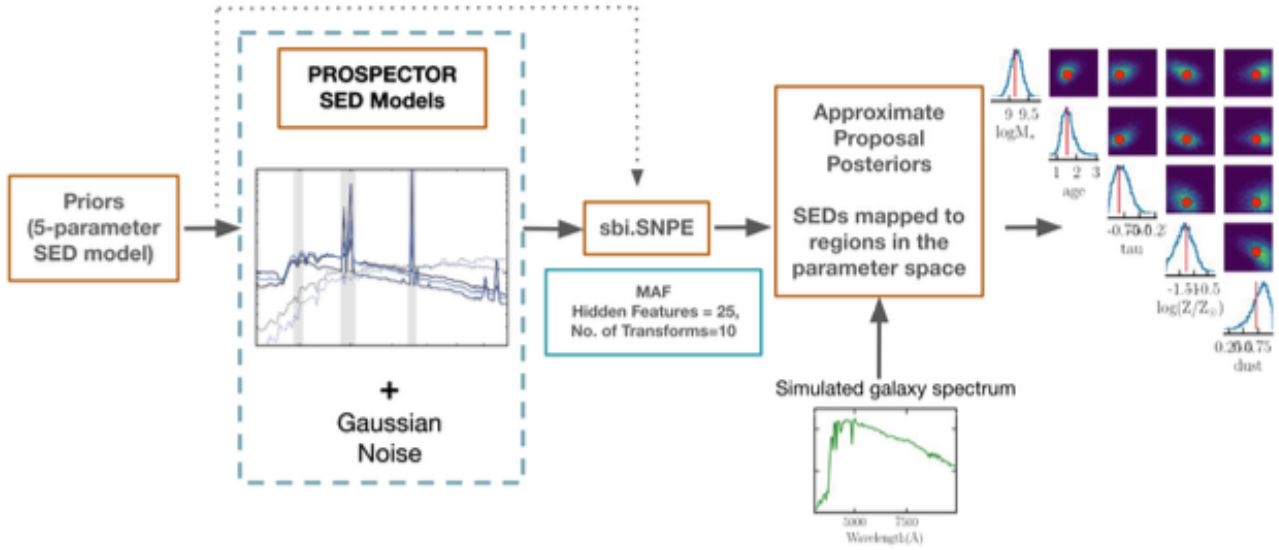


Waldo guarantees coverage everywhere, even if the prior poorly specified. Well-specified prior ⇨ power (tighter constraints)

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

Another application of Waldo to NPEs (5 POI) …

Astronomy: Infer galaxy parameters from SEDs via NPE



Why? Advent of billion-galaxy surveys with complex data needs efficient modeling of spectral energy distributions (SEDs) with robust uncertainty quantification

How? Combine SBI and NPE to infer galaxy parameters (5-parameter model)

Goal: use Waldo to obtain reliable constraints and check their validity against those obtained via NPE

Image taken from Khullar et al. (2022)

13

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

DIGS: Deep Inference of Galaxy Spectra with Neural Posterior Estimation

Gourav Khullar[1,2,3,4,5], **Brian Nord**[1,2,3], **Aleksandra Ćiprijanović**[1], **Jason Poh**[2,3], **Fei Xu**[2,3]

[1]Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
[2]Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA
[3]Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
[4]Kavli Institute for Astrophysics & Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA
[5]Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, Pittsburgh, PA 15260, USA

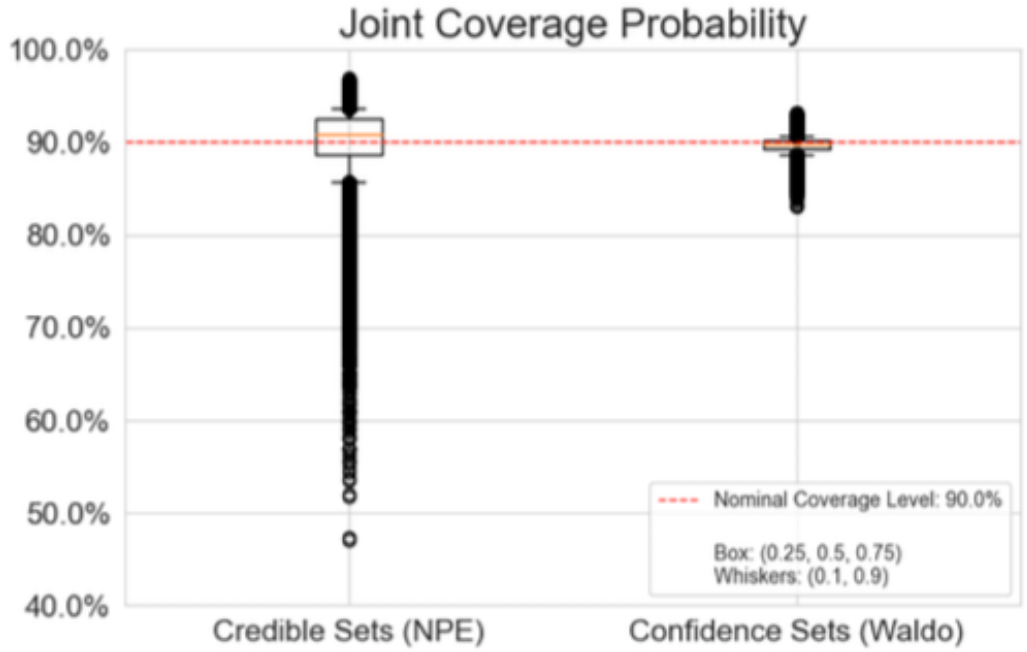
E-mail: gkhullar@uchicago.edu

Abstract.

With the advent of billion-galaxy surveys with complex data, the need of the hour is to efficiently model galaxy spectral energy distributions (SEDs) with robust uncertainty quantification. The combination of Simulation-Based inference (SBI) and amortized Neural Posterior Estimation (NPE) has been successfully used to analyse simulated and real galaxy photometry both precisely and efficiently. In this work, we utilise this combination and build on existing literature to analyse simulated noisy galaxy spectra. Here, we demonstrate a proof-of-concept study of spectra that is a) an efficient analysis of galaxy SEDs and inference of galaxy parameters with physically interpretable uncertainties; and b) amortized calculations of posterior distributions of said galaxy parameters

Coverage across the entire parameter space

$$r(\theta) := \mathbb{P}(\theta \in \mathscr{R}(\mathscr{D}) \mid \theta), \quad \theta \in \mathbb{R}^5$$

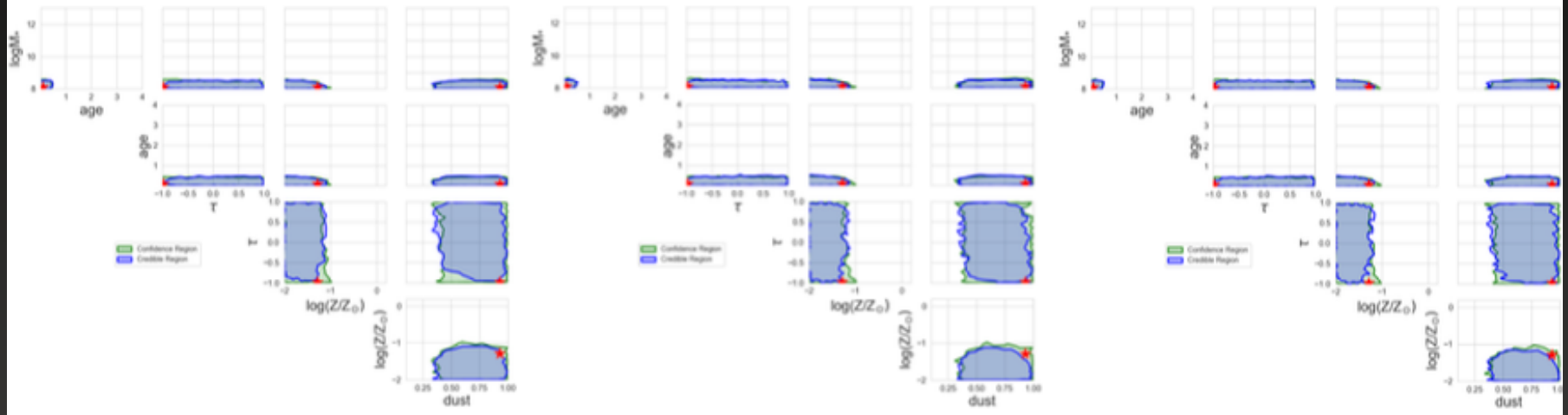


15

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

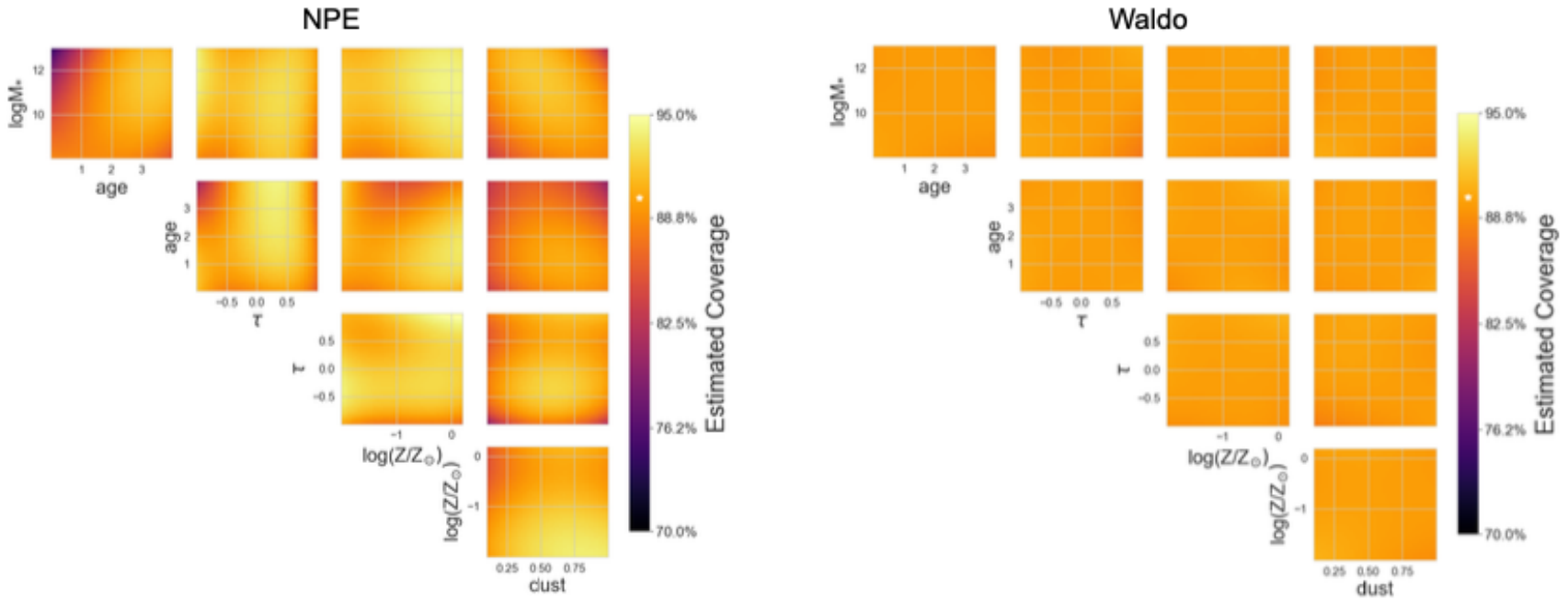
Example of parameter regions when NPE undercovers

Confidence regions (Waldo, green) and credible regions (NPE, blue) obtained from three observations sampled from the same true parameter



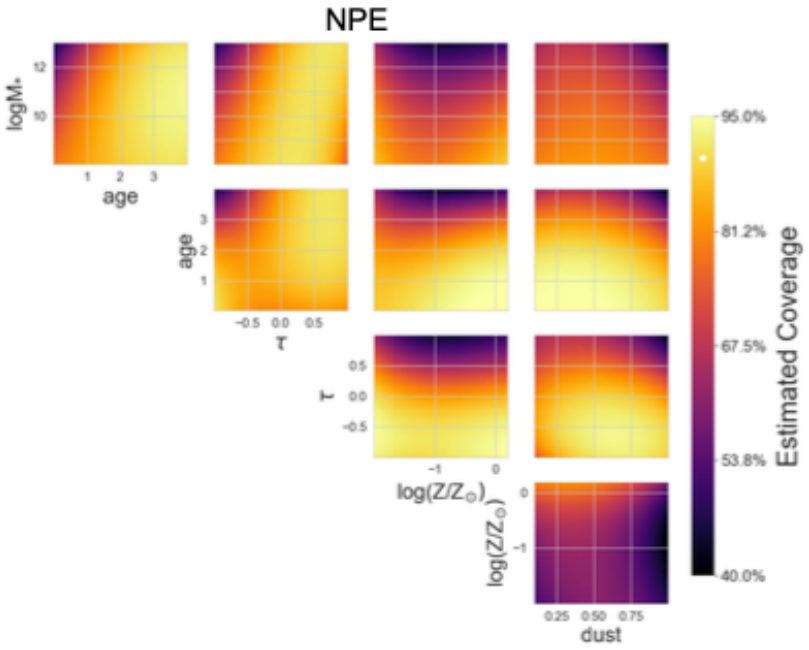
Partial dependence of coverage probability vs parameters

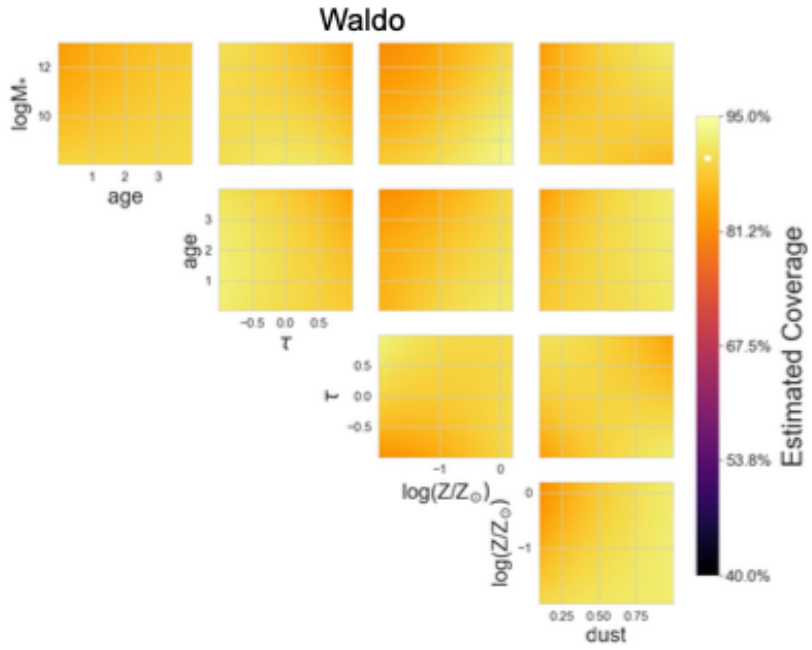
$$\int_{\theta_i\theta_j\theta_k} r(\theta)\mathrm{d}\theta_i\mathrm{d}\theta_j\mathrm{d}\theta_k, \quad \theta \in \mathbb{R}^5$$



Profiled dependence of coverage probability vs parameters

$$\min_{\theta_i \theta_j \theta_k} r(\theta), \quad \theta \in \mathbb{R}^5$$





17

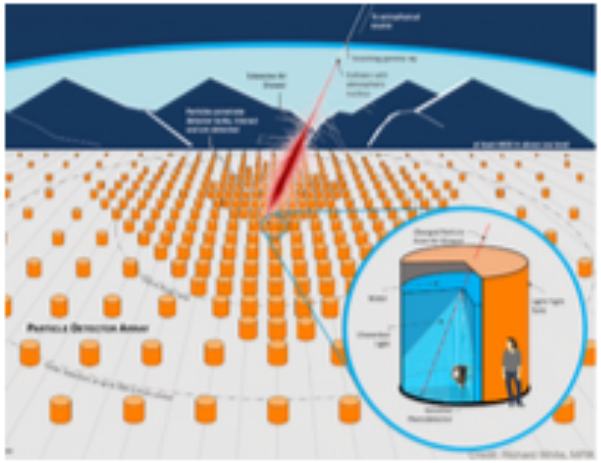
Classification Under Systematic Uncertainties:

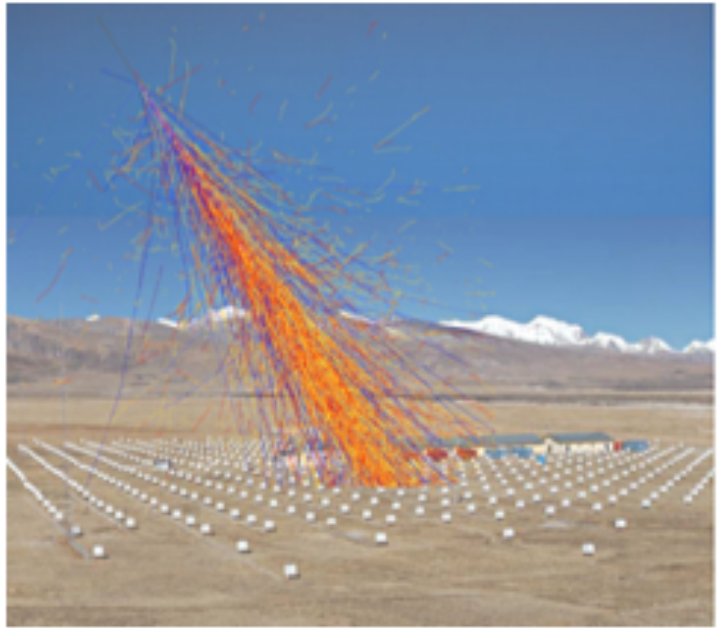
Application to Atmospheric Cosmic-Ray Showers

[IN PROGRESS 2023- with Alex Shen, Luca Masserano, Michele Doro, Tommaso Dorigo]

- SWGO plans to study ultra-energetic (above 1 TeV) photon showers ⇒ High-altitude array of Cherenkov tanks.
- Information on gamma flux requires separating gamma showers (very rare) from hadrons (common background). CORSIKA to generate G- and H-showers.

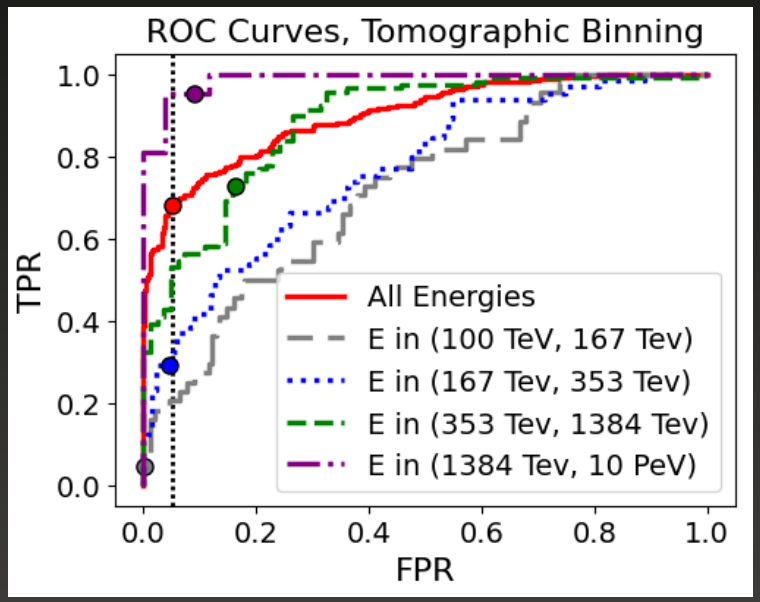
Right: a simulated photon-induced air shower over the LHAASO array in China. Below: a representation of the SWGO array.





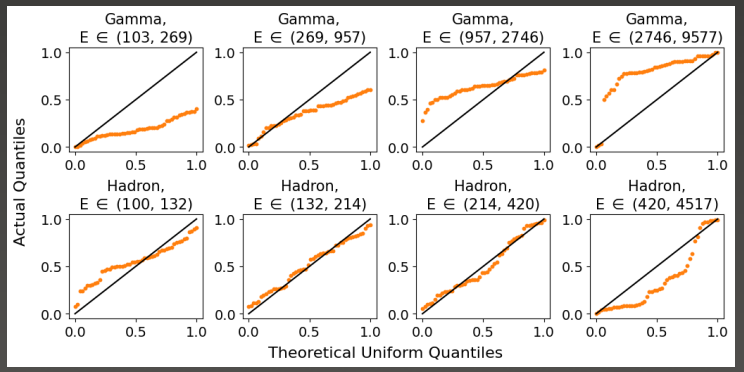
Classification Under Systematic Uncertainties

- Suppose we ignore shower parameters (e.g. energy E, direction) and various hyper-parameters, and directly classify showers (G or H) based on array measurements. That is, compare $\Pr(\mu=\mu_0 \mid x)$ to a constant C.



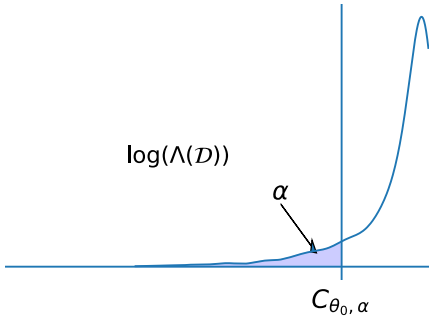
Left: *Direct classification ignoring NPs leads to a misleading ROC curve (red curve) and biased estimates (circles) when attempting to control e.g. FPR at 0.05*

Right: *Q-Q plots show that the estimated probabilities of rejection (TPR and FPR) are not calibrated across different energies.*



Recall:

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

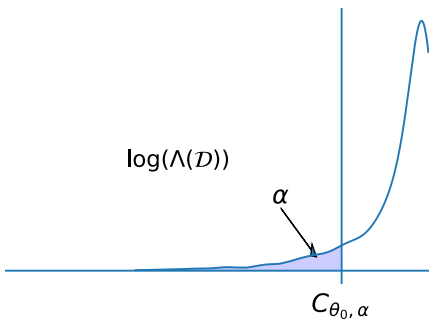
$$C_{\theta_0,\alpha} = \arg\sup_{C \in \mathbb{R}} \left\{ C : \mathbb{P}_{\mathcal{D}|\theta_0} \left(\lambda(\mathcal{D}; \theta_0) < C \right) \leq \alpha \right\}.$$

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 $F^{-1}_{\lambda|\theta}(\alpha|\theta)$.

Recall:

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 $F^{-1}_{\lambda|\theta}(\alpha|\theta)$.

1. Now use BF/posteriors instead of LRT as test statistic.
2. Learn the entire ROC as a function of POI and NPs

$\{(\mu_i, \mathbf{x}_i)\}_{i=1}^{B}$, we reformulate the gamma/hadron discrimination problem as a composite-versus-composite hypothesis test:

$$H_0 : \theta \in \Theta_0 \;\text{ versus }\; H_1 : \theta \in \Theta_1 \tag{1}$$

where $\Theta_0 = \{0\} \times \mathcal{N}$ and $\Theta_1 = \{1\} \times \mathcal{N}$. As test statistic, we exploit

$$\tau(\mathbf{x}) = \frac{\mathbb{P}'(\mu = 0|\mathbf{x})\ \mathbb{P}'(\mu = 0)}{\mathbb{P}'(\mu = 1|\mathbf{x})\ \mathbb{P}'(\mu = 1)}, \tag{2}$$

which is equivalent to the Bayes factor for the test 1, see Appendix B. This quantity can be estimated directly from the *pre-trained* classifier based on $\mathcal{T}_B$; there is no need for an extra step to, e.g., try to learn the likelihood function $\mathcal{L}(\mathbf{x}; \mu, \nu)$ or the associated likelihood ratio statistic from simulated data

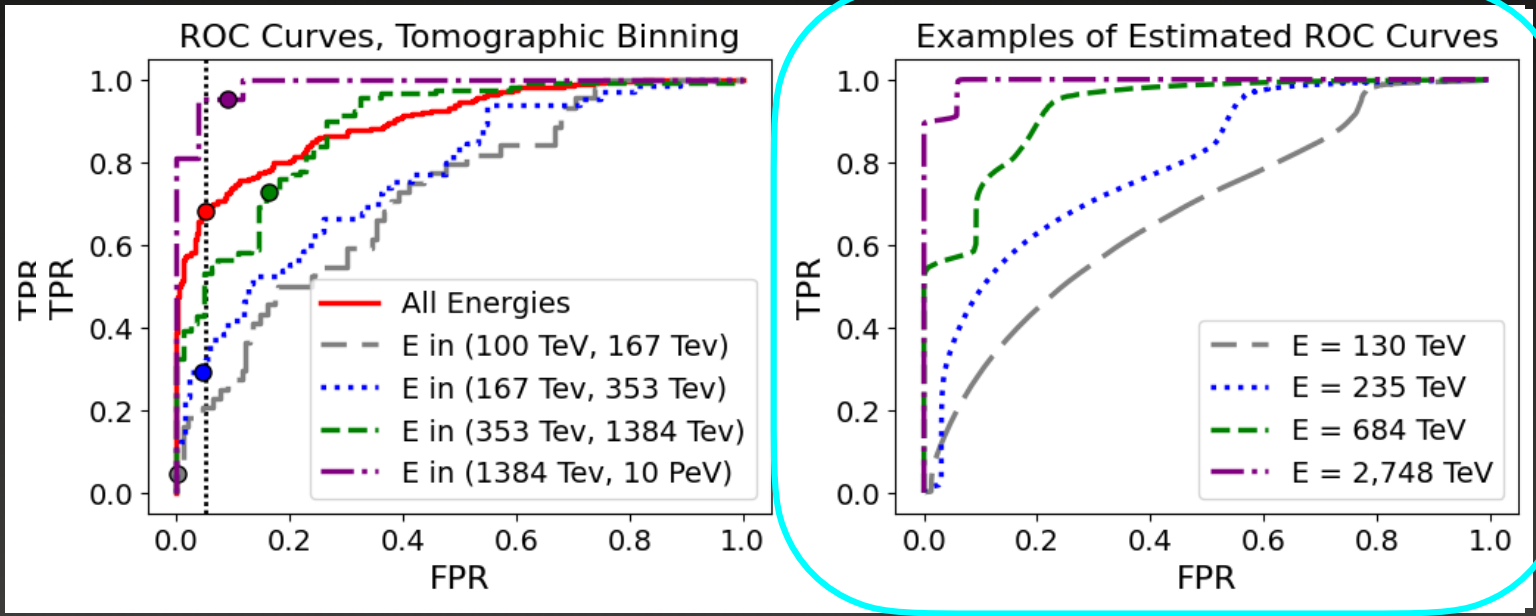
$$W_0(C; \nu) := \mathbb{P}\left(\hat{\tau}(\mathbf{x}) \leq C \mid C; \mu = 0, \nu\right) = \mathbb{E}_{\mathbf{x}|\mu=0,\nu}\left(I_C(\mathbf{x}) \mid C; \mu = 0, \nu\right) \tag{3}$$

$$W_1(C; \nu) := \mathbb{P}\left(\hat{\tau}(\mathbf{x}) \leq C \mid C; \mu = 1, \nu\right) = \mathbb{E}_{\mathbf{x}|\mu=1,\nu}\left(I_C(\mathbf{x}) \mid C; \mu = 1, \nu\right), \tag{4}$$

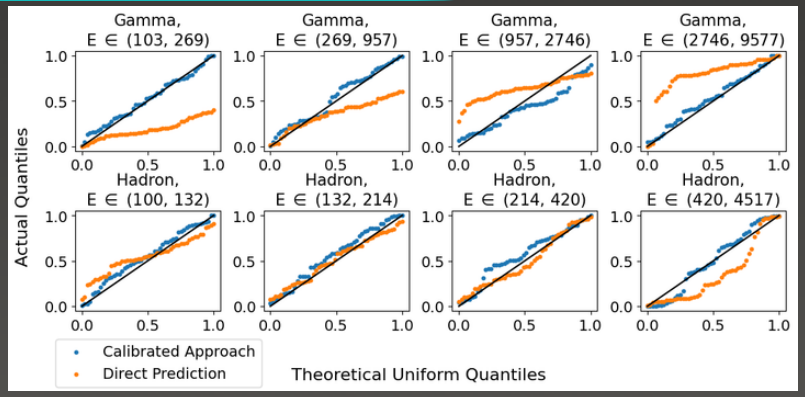
for all $C \in \mathbb{R}$ and all $\nu \in \mathcal{N}$. At fixed ν, the ROC curve is defined as the true positive rate (TPR; W_1) vs false positive rate (FPR; W_0) over the space of cutoffs C. Appendix C details our procedure for estimating $W_\mu(C; \nu)$ using calibration data $\mathcal{T}'_B = \{(\theta'_1, \mathbf{x}'_1), \ldots, (\theta'_B, \mathbf{x}'_{B'})\} \sim r(\theta)\mathcal{L}(\mathbf{x}; \theta)$.

Classification Under Systematic Uncertainties

- Using a new version of LF2I calibration, we can estimate the entire set of ROC's for all $\theta=(\mu,\eta)$, where μ is the class, and η are hyper-parameters.

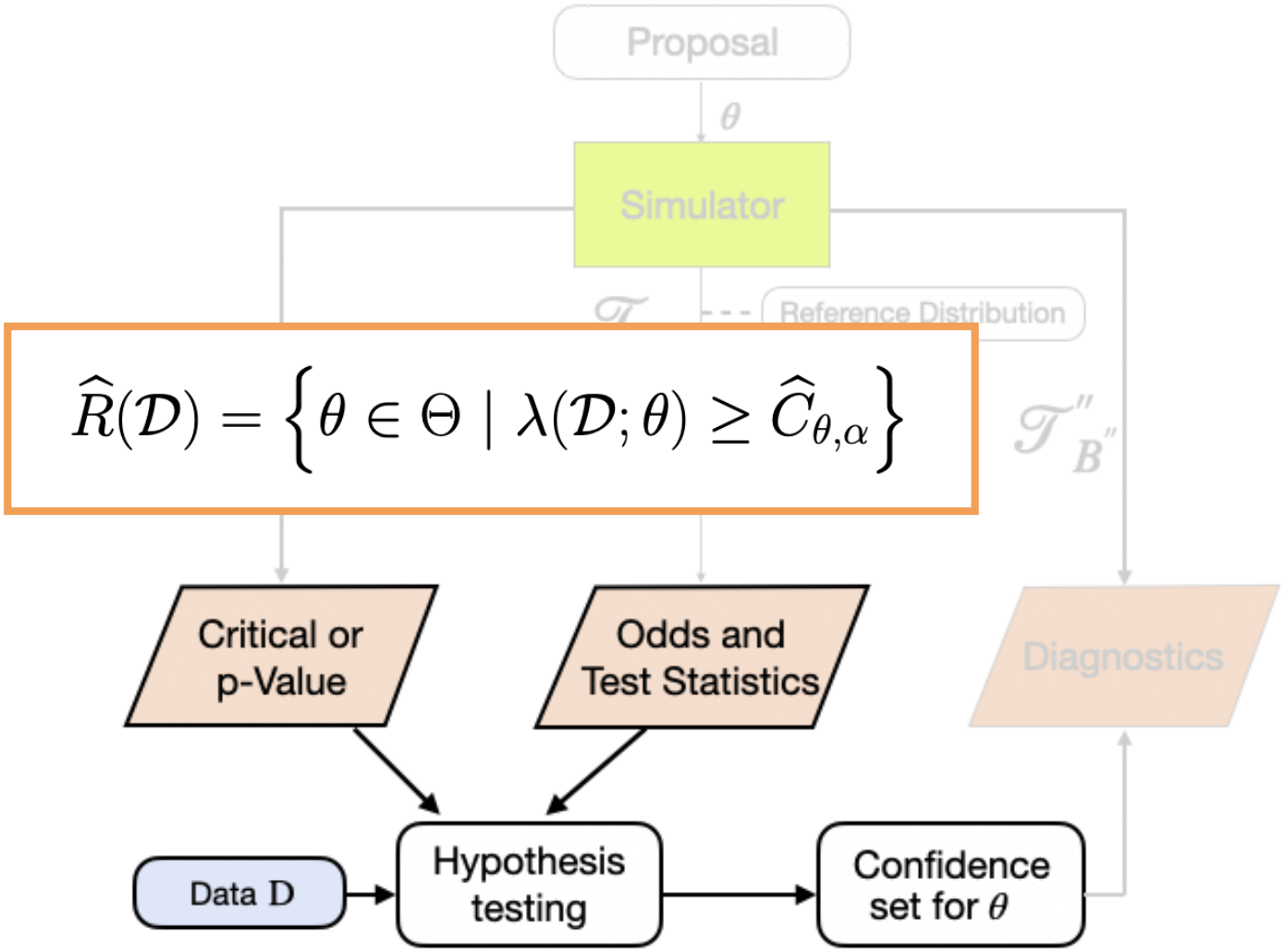


Right: Q-Q plots of prob of rejection curves for direct classification (orange) vs with LF2I calibration (blue)



Recall:

Construct Confidence Set via Neyman Inversion

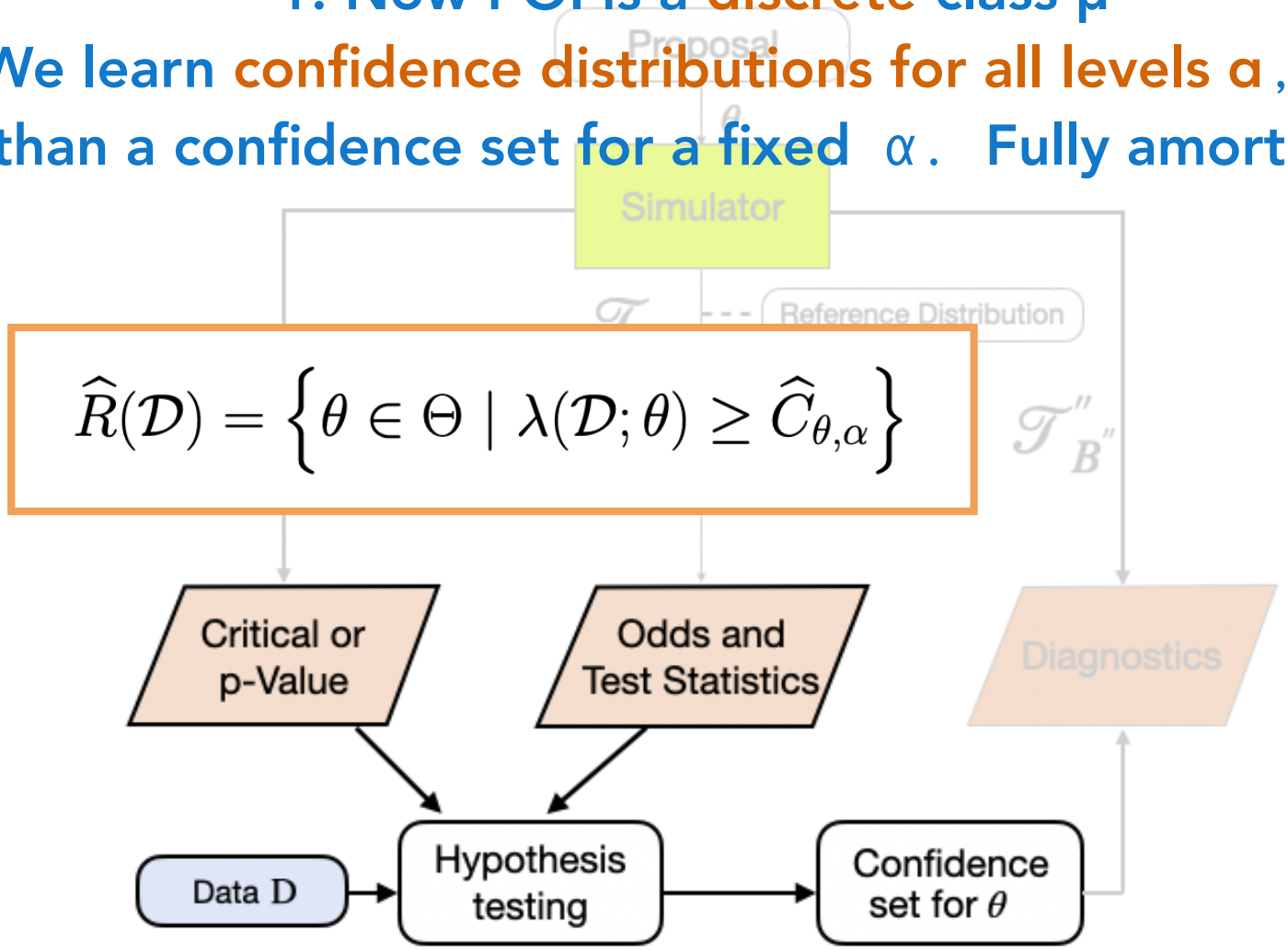


Recall:

Construct Confidence Set via Neyman Inversion

1. Now POI is a discrete class μ

2. We learn confidence distributions for all levels α, rather than a confidence set for a fixed α. Fully amortized



Construct Set-Valued Classifiers from Pr(μ=μ_0 | x)

- Some instances are ``ambiguous'' and hence difficult to label correctly. Set-valued classifiers output sets of plausible labels rather than a single label.

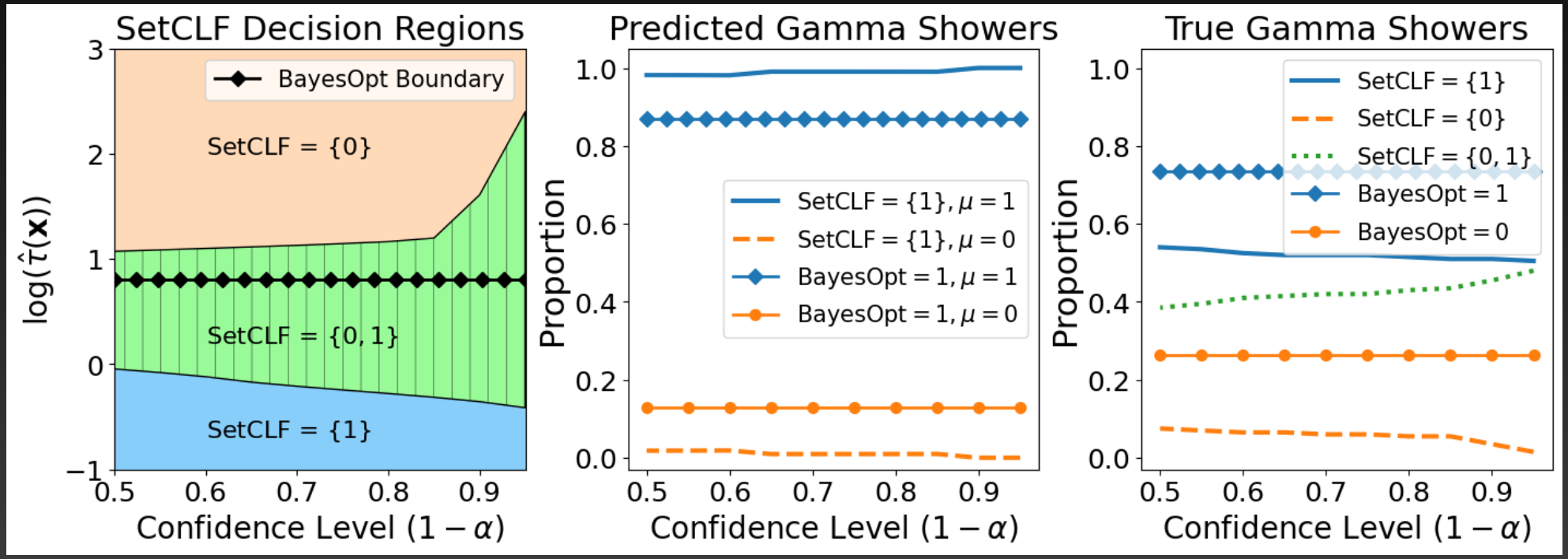
$$\mathbf{H}_\alpha : \mathbf{x} \mapsto \{0, 1, \{0, 1\}\}$$

- With LF2I calibration, the set-valued classifiers have nominal confidence (1-α), even under systematic uncertainties.

$$\mathbb{P}\left(\mu \in \mathbf{H}_\alpha(\mathbf{x}) \mid (\mu, \nu)\right) = 1 - \alpha, \quad \forall \mu \in \{0, 1\},\ \nu \in \mathcal{N}$$

- Amortized inference w.r.t. both observation x and level α

Set-Valued Classification Output Compared to ‘BayesOpt’ Under Systematic Uncertainties



- Left: Decision regions as a function of confidence level
- Center: Higher precision and lower FDR than BayesOpt
- Right: Lower miss rate but also lower recall than BayesOpt

Take-Away: LF2I is a practical procedure for Neyman construction of confidence sets

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

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

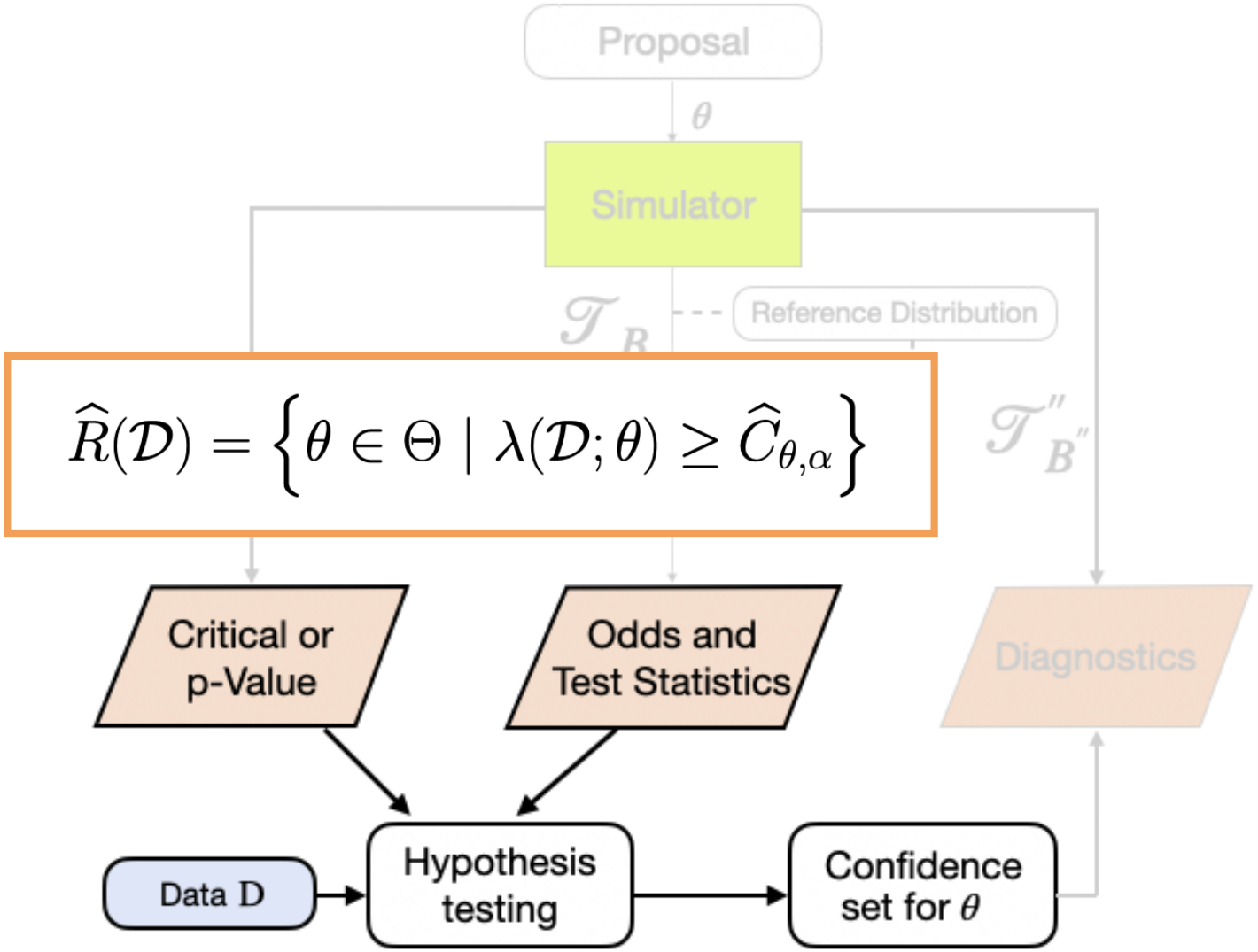
Theorem (Neyman 1937)

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 $\theta_0 \in \Theta$.

Construct Confidence Set via Neyman Inversion



Take-Away: LF2I

- Validity and Diagnostics: LF2I bridges ML and classical statistics to create valid confidence sets and run diagnostics.
- Prior Independence: LF2I guarantees (approximate) conditional coverage regardless of the prior and the sample size. Well-specified prior => higher power
- Power: Hardest to achieve in practice. Area where most statistical and computational advances will take place.

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

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

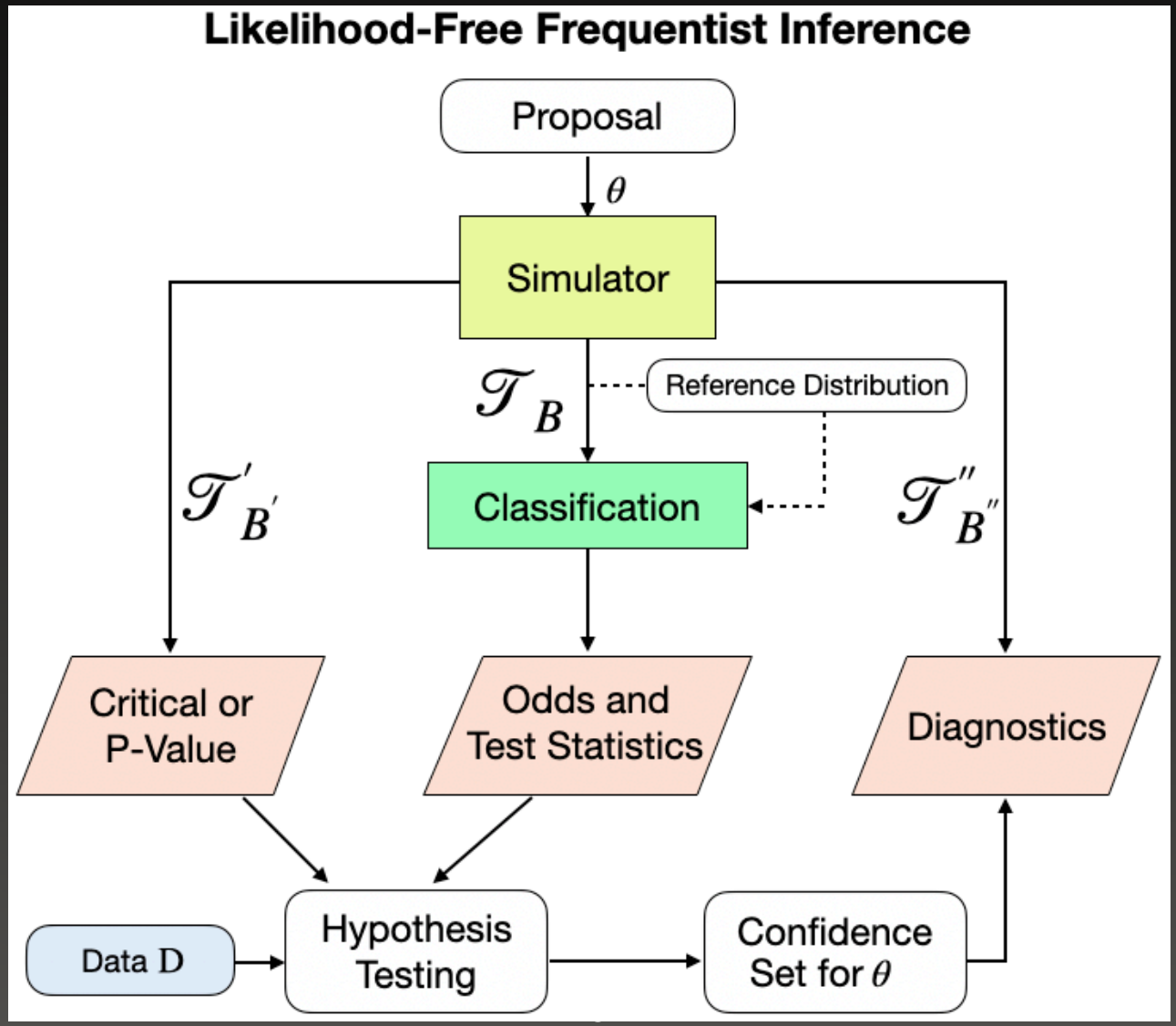
- `BFF` (Bayesian Frequentist Factor):

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

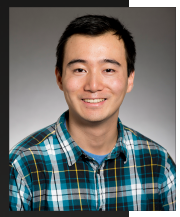
$$\tau^{\text{WALDO}}(\mathcal{D};\boldsymbol{\theta}_0) = \frac{(\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^2}{\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]}$$

- LF2I is a fully modular framework. We'd love to collaborate on code/science-problems! Please email annlee@andrew.cmu.edu

https://github.com/lee-group-cmu/lf2i



Finally, if you are instead interested in calibrated PDs and posteriors (consistent with a chosen prior)…



Diagnostics for Conditional Density Models and Bayesian Inference Algorithms

UAI, PMLR (161) 2021

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

[1]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]Department of Statistics, Federal University of São Carlos (UFSCar), S

Linhart et al 2022, ML4PS: Extension to multivariate response for NFs; Lemos et al, ICML 2023: TARP testing

Definition 1 (Global Consistency). *An estimate $\widehat{f}(y|\mathbf{x})$ is globally consistent with the density $f(y|\mathbf{x})$ if the following null hypothesis holds:*

$$H_0 : \widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x}) \text{ for every } \mathbf{x} \in \mathcal{X} \text{ and } y \in \mathcal{Y}. \quad (1)$$

arXiv:2205.14568

CONDITIONALLY CALIBRATED PREDICTIVE DISTRIBUTIONS BY PROBABILITY-PROBABILITY MAP: APPLICATION TO GALAXY REDSHIFT ESTIMATION AND PROBABILISTIC FORECASTING

BY BIPRATEEP DEY[1,a], DAVID ZHAO[2,d], JEFFREY A. NEWMAN[1,b], BRETT H. ANDREWS[1,c], RAFAEL IZBICKI[3,e], AND ANN B. LEE[4,f]

[1]*Department of Physics and Astronomy and PITT-PACC, University of Pittsburgh,* [a]*biprateep@pitt.edu;* [b]*janewman@pitt.edu;* [c]*andrewsh@pitt.edu*

[2]*Department of Statistics and Data Sci*

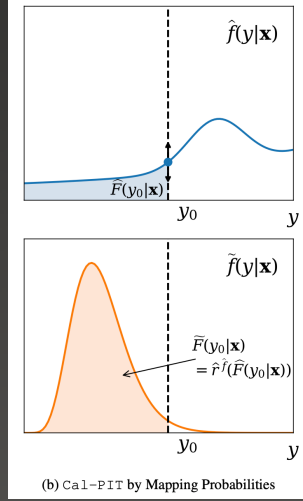
[3]*Department of Statistic*

[4]*Department of Statistics and Data S*

DEFINITION 3 (Recalibrated PD). *The recalibrated predictive distribution (PD) of Y given $\mathbf{x}$ is defined through a P-P map,*

$$\widetilde{F}(y|\mathbf{x}) := \widehat{r}^{\widehat{f}}\left(\widehat{F}(y|\mathbf{x}); \mathbf{x}\right), \quad (6)$$

where $\widehat{r}^{\widehat{f}}$ is the regression estimator of the PIT-CDF (Equation 4).



(b) `Cal-PIT` by Mapping Probabilities

Acknowledgments

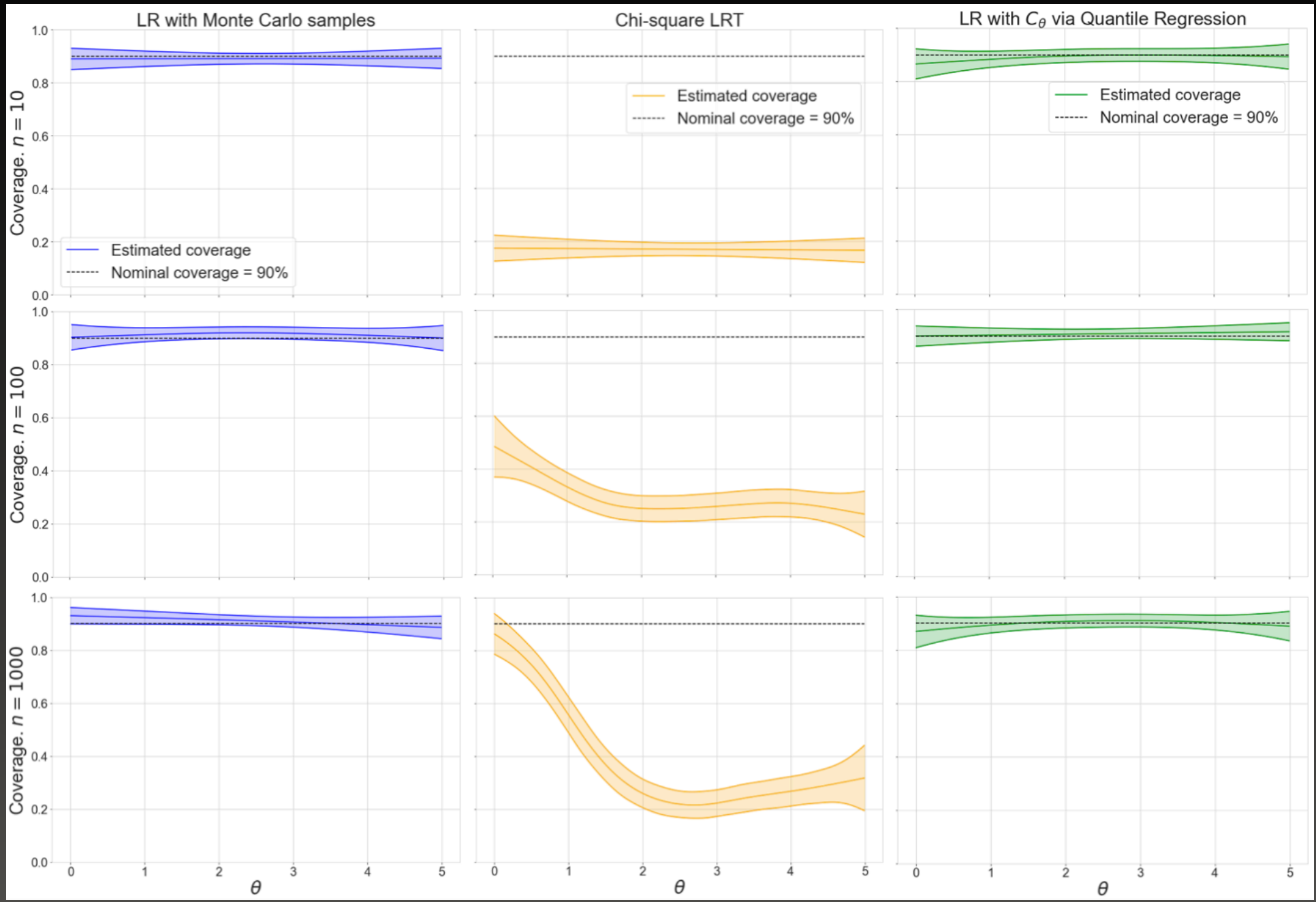
- Nic Dalmasso — original LF2I framework
- Rafael Izbicki (UFSCar)
- Luca Masserano
- Alex Shen, Mikael Kuusela, David Zhao
- Tommaso Dorigo, Michele Doro (INFN/Padova)
- Gourav Khullar (Univ of Pittsburgh)

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

This work is funded in part by NSF DMS-2053804 and NSF PHY-2020295.

Extra Slides Start Here

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



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

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

Ex: Power of BFF (n=1)

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*

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

Nuisance-Parameterized LF2I

Test composite vs composite hypotheses:

$$H_{0,\mu_0} : \theta \in \Theta_0 \quad \text{vs} \quad H_{1,\mu_0} : \theta \in \Theta_1,$$

where $\Theta_0 = \{(\mu_0, \nu) \mid \nu \in \mathcal{N}\}$, and $\Theta_1 = \Theta_0^c$.

- `ACORE` test statistic (by maximizing estimated odds)

$$\widehat{\Lambda}(\mathcal{D}; \mu_0) := \log \frac{\sup_{\nu \in \mathcal{N}} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; (\mu_0, \nu))}{\sup_{\theta \in \Theta} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta)}$$

- `BFF` test statistic (by integrating estimated odds)

$$\widehat{\tau}(\mathcal{D}; \mu_0) := \frac{\int_{\mathcal{N}} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; (\mu_0, \nu)) \; d\pi_0(\nu)}{\int_{\Theta_1} \prod_{i=1}^n \widehat{\mathbb{O}}(\mathbf{X}_i^{\text{obs}}; \theta) \; d\pi_1(\theta)}.$$

where $\pi_0(\nu)$ is a distribution over $\mathcal{N}$, the nuisance parameter space.

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*

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