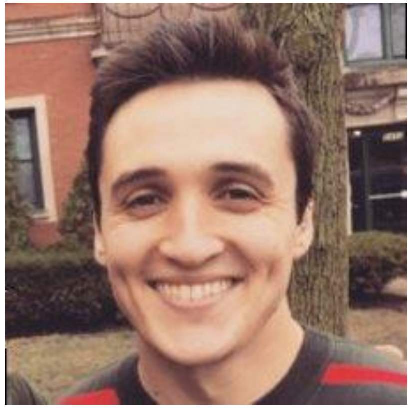
Truncated Marginal Neural Ratio Estimation

Empirically testable, simulation efficient & simulation-based posterior approximation.


Benjamin Kurt Miller


Alex Cole


Patrick Forré


Gilles Louppe


Christoph Weniger

5th Inter-experiment Machine Learning Workshop, CERN 2022

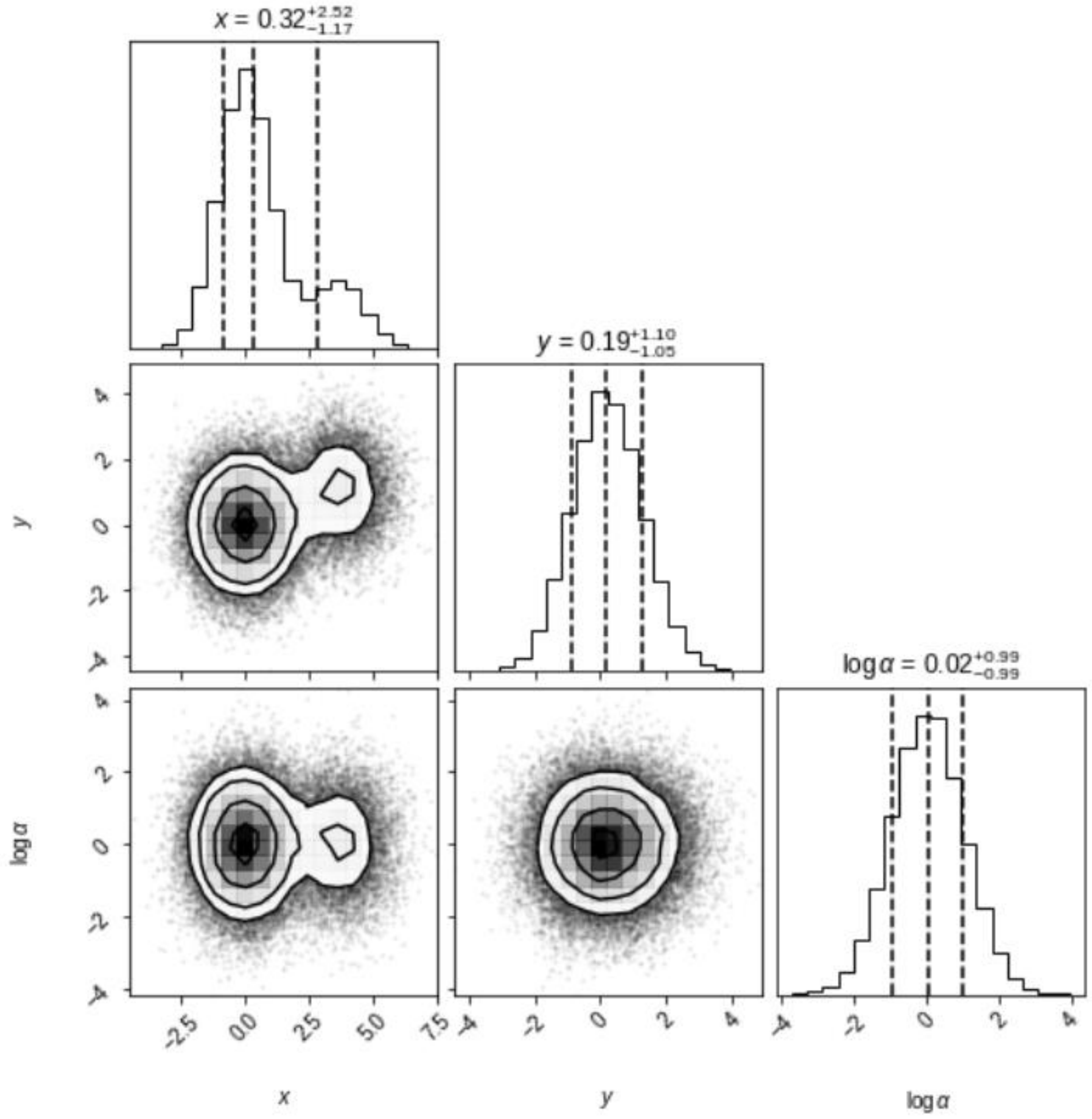
AMLAB Amsterdam Machine Learning Lab

GRAPPA GRavitation AstroParticle Physics Amsterdam

Marginal Inference

The posterior quantifies the uncertainty about parameters θ given data x.

$$p(\theta \mid x) = \frac{p(x \mid \theta)}{p(x)} p(\theta)$$



Foreman-Mackey, D. (2016). corner.py: Scatterplot matrices in Python. *The Journal of Open Source Software*, *1*(2), 24. https://doi.org/10.21105/joss.00024

Marginal Inference

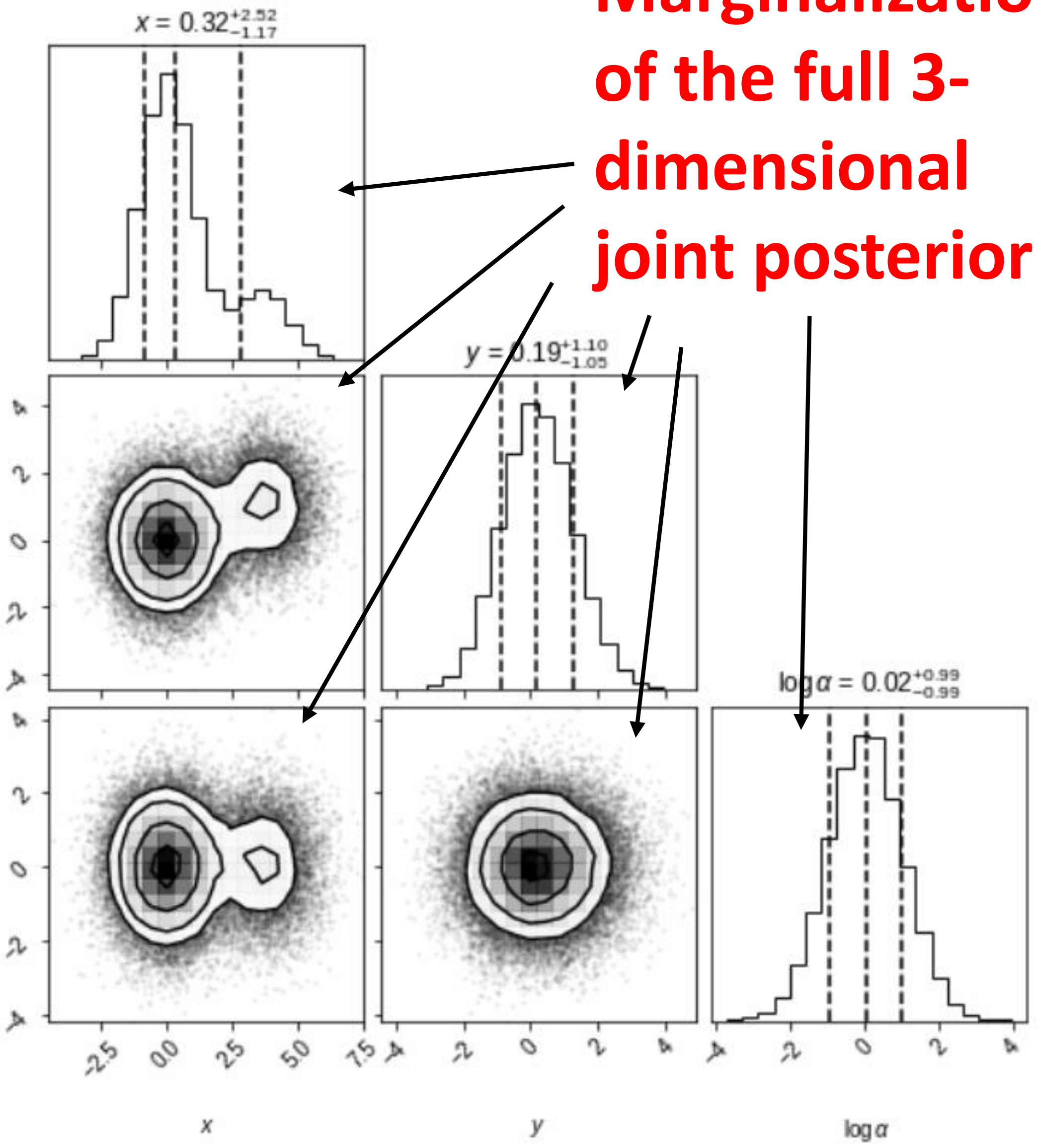
Marginal Inference: Estimate the marginal posteriors of interest directly.

$$p(\vartheta \mid x) = \frac{\int p(x \mid \vartheta, \eta)\, p(\vartheta, \eta)\, d\eta}{p(x)} = \frac{p(x \mid \vartheta)}{p(x)} p(\vartheta)$$

Also see: Justin Alsing and Benjamin Wandelt. Nuisance hardened data compression for fast likelihood-free inference. [arXiv:1903.01473]()

Niall Jeffrey and Benjamin Wandelt. Solving high-dimensional parameter inference: . [arXiv: 2011.05991]()

Marginalizations of the full 3-dimensional joint posterior!



Foreman-Mackey, D. (2016). corner.py: Scatterplot matrices in Python. *The Journal of Open Source Software*, *1*(2), 24. https://doi.org/10.21105/joss.00024

Neural Ratio Estimation

We train a classifier and extract a likelihood-to-evidence ratio…

The classifier distinguishes between samples drawn jointly vs marginally

$$p(x,\theta \mid y) = \begin{cases} p(x,\theta) & if\ y = 1 \\ p(x)p(\theta) & if\ y = 0 \end{cases}.$$

The posterior for the "switching" variable y is

$$p(y = 1|x,\theta) = \frac{p(x,\theta \mid y = 1)}{p(x,\theta \mid y = 0)\ +\ p(x,\theta \mid y = 1)} = \frac{p(x,\theta)}{p(x)p(\theta) + p(x,\theta)}$$

Let $r(x,\theta) = \frac{p(x,\theta \mid y=1)}{p(x,\theta \mid y=0)} = \frac{p(x,\theta)}{p(x)p(\theta)} = \frac{p(x \mid \theta)}{p(x)}$ i.e., the likelihood to evidence ratio.

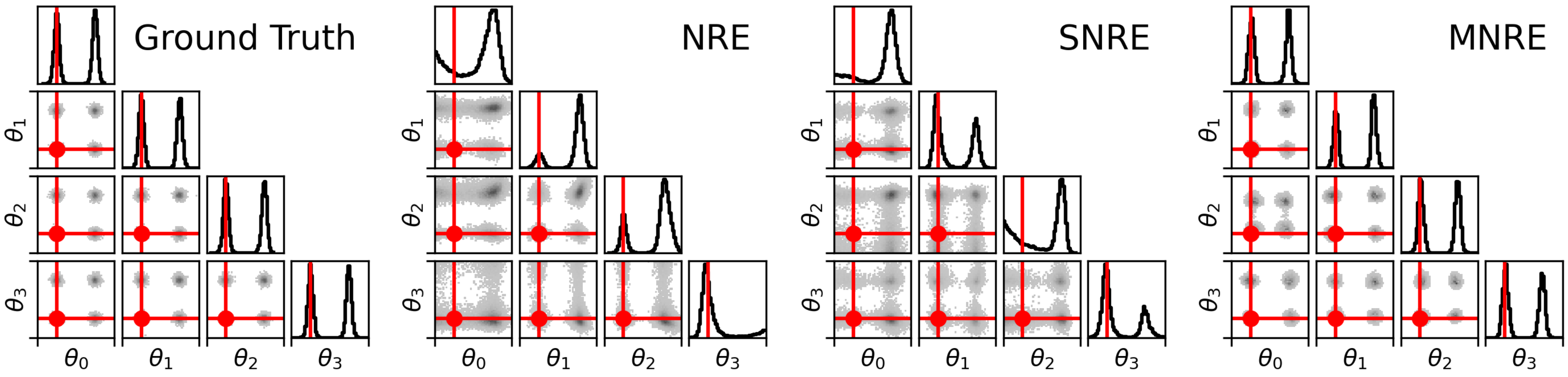
That means $p(y = 1 \mid x,\theta) = \frac{r(x,\theta)}{r(x,\theta) + 1} = \sigma\big(\log r(x,\theta)\big)$.

Eggbox: Is marginal ratio estimation simulation efficient?

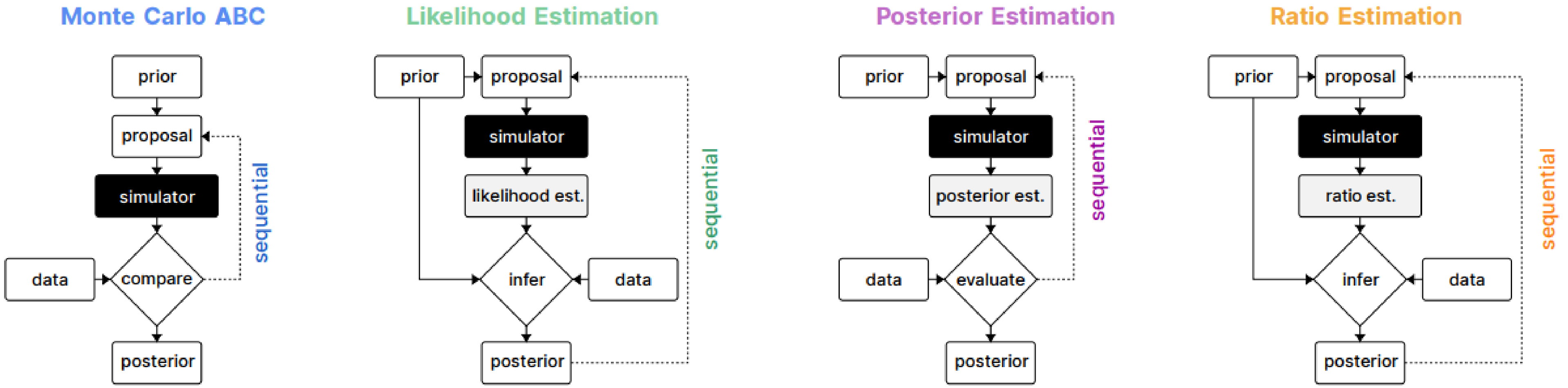
10-dimensional, $2^{10} = 1024$ modes, 10,000 simulations.

Compare marginal estimation (MNRE) to joint estimations (NRE, SNRE).

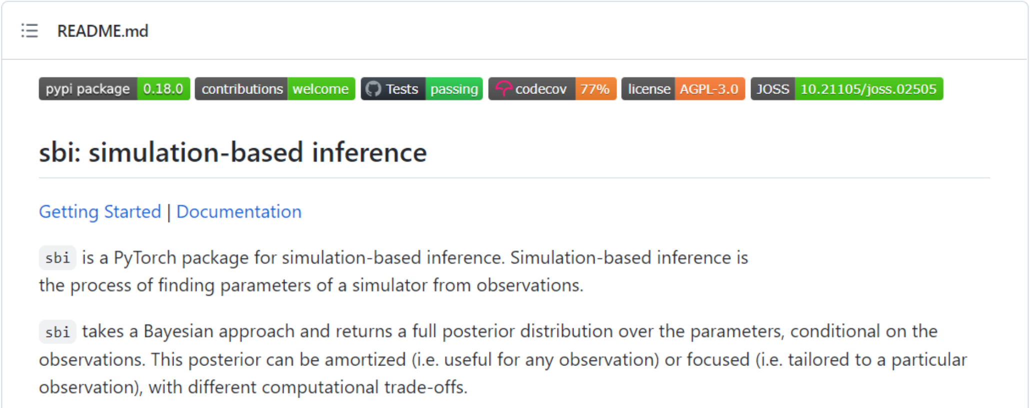
Ours



Sidenote: what are amortized vs. sequential methods?



Jan-Matthis Lueckmann, et. al. Benchmarking Simulation-Based Inference.
https://arxiv.org/abs/2101.04653



README.md

pypi package 0.18.0 | contributions welcome | Tests passing | codecov 77% | license AGPL-3.0 | JOSS 10.21105/joss.02505

sbi: simulation-based inference

Getting Started | Documentation

`sbi` is a PyTorch package for simulation-based inference. Simulation-based inference is the process of finding parameters of a simulator from observations.

`sbi` takes a Bayesian approach and returns a full posterior distribution over the parameters, conditional on the observations. This posterior can be amortized (i.e. useful for any observation) or focused (i.e. tailored to a particular observation), with different computational trade-offs.

https://github.com/mackelab/sbi

<u>Truncated Marginal Neural Ratio Estimation</u>

- **Estimates marginal posteriors directly…**

- Extends *Neural Ratio Estimation*, which estimates the likelihood-to-evidence ratio $\frac{p(x \mid \theta)}{p(x)}$ by training a classifier.

Hermans, et. al. 2019. Likelihood-free MCMC with Amortized Approximate Ratio Estimators. arXiv: 1903.04057.

- **Truncates uninformative regions** where $p(\theta \mid x_o) \approx 0$…

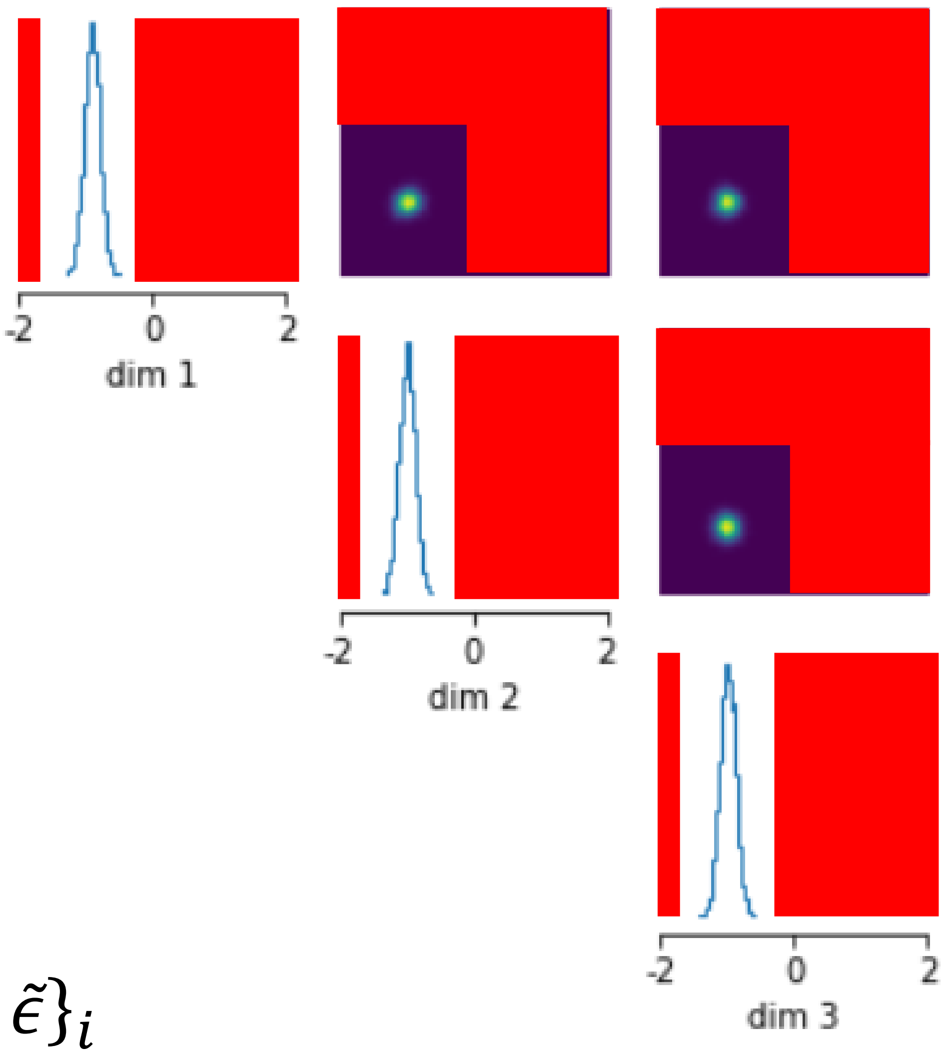
- **Enables consistency checks through local amortization…**

Hermans and Delaunoy, et. al. 2021. Averting A Crisis In Simulation-Based Inference. arXiv: 2110.06581.

Truncated Bayesian Inference

Posteriors can be quite narrow compared to priors.

Truncated Inference: Sample only regions near the posterior mass.



"Ideal" truncated region?

$$\Gamma := \{\theta \in \text{supp } \mathrm{p}(\theta) \mid \mathrm{p}(\theta \mid x_o) > \epsilon\}$$

Our component-wise marginal estimate.

$$\Gamma_{rec,i} := \{\theta_i \in \text{supp } \mathrm{p}_i(\theta_i) \mid \hat{p}(\theta_i \mid x_o) > \tilde{\epsilon}\}_i$$

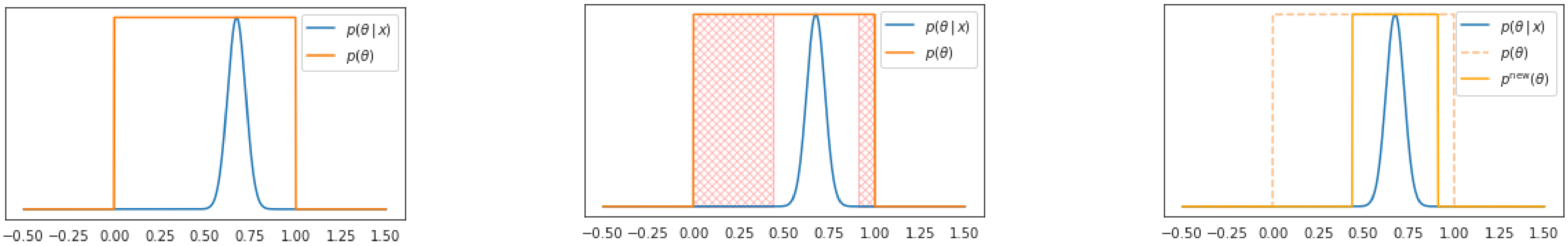
Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., & Macke, J. H. (2020). sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, *5*(52), 2505. https://doi.org/10.21105/joss.02505

Sketch of truncation scheme

1. Sample from the joint $(\boldsymbol{x}, \boldsymbol{\theta}) \sim p(x \mid \theta)\, p(\theta)$.
2. Learn all component-wise likelihood-to-evidence ratios $\dfrac{p(x \mid \vartheta_i)}{p(x)}$.
3. Truncate prior $p(\theta) \to p_{\Gamma_{rec}}(\theta)$ with $\Gamma_{rec,i} = \{\theta_i \in \mathrm{supp}\ \mathrm{p}_i(\theta_i) \mid \hat{p}(\theta_i \mid x_o) > \tilde{\epsilon}\}_i$.
4. Simulate more data $(\boldsymbol{x}, \boldsymbol{\theta}) \sim p(x \mid \theta)\, p_{\Gamma_{rec}}(\theta)$.
5. Repeat 2-4 until prior volume stabilizes.
6. Return truncated region Γ_{rec}, samples within, learn (marginal) posterior.

Truncation Visualization

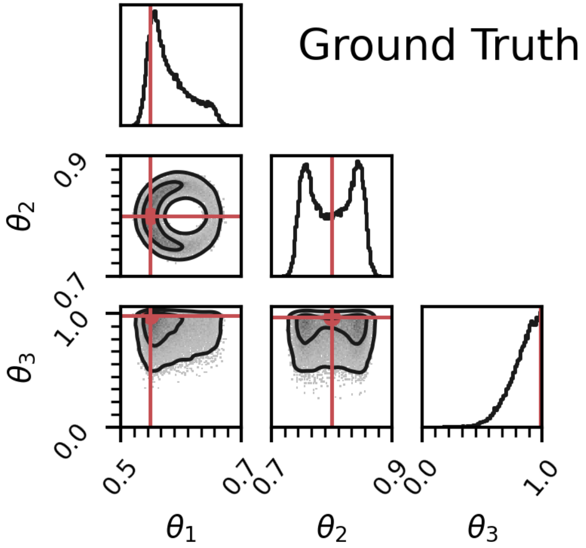
In the image, $p^{new}(\theta)$ is not normalized

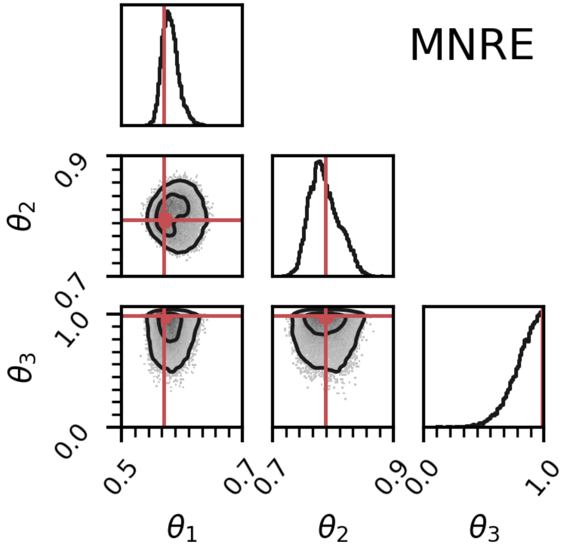


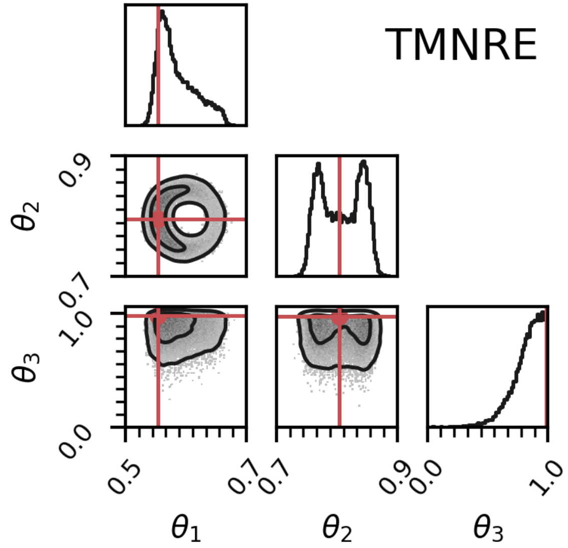
Torus: Is truncation with marginals efficient?

Marginal

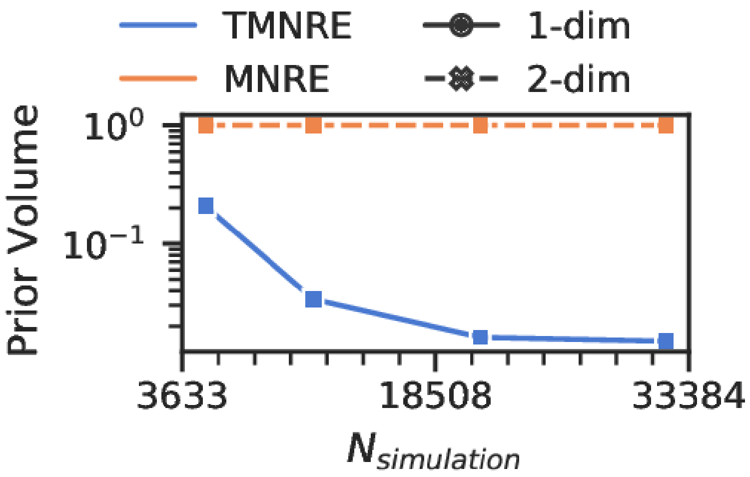
Truncated Marginal

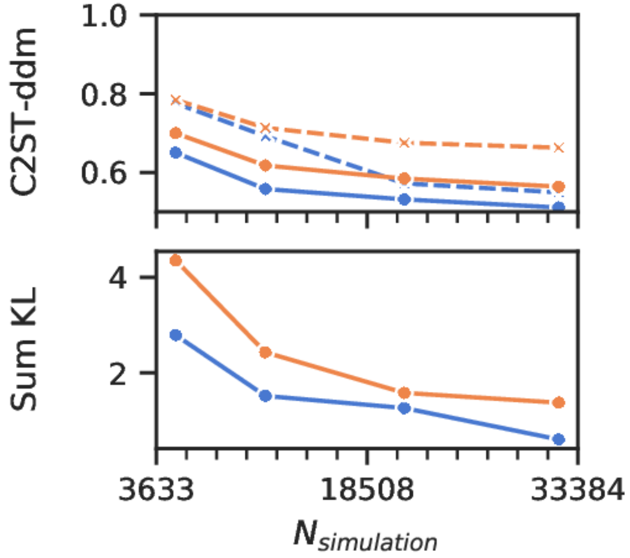






Prior volume reduction

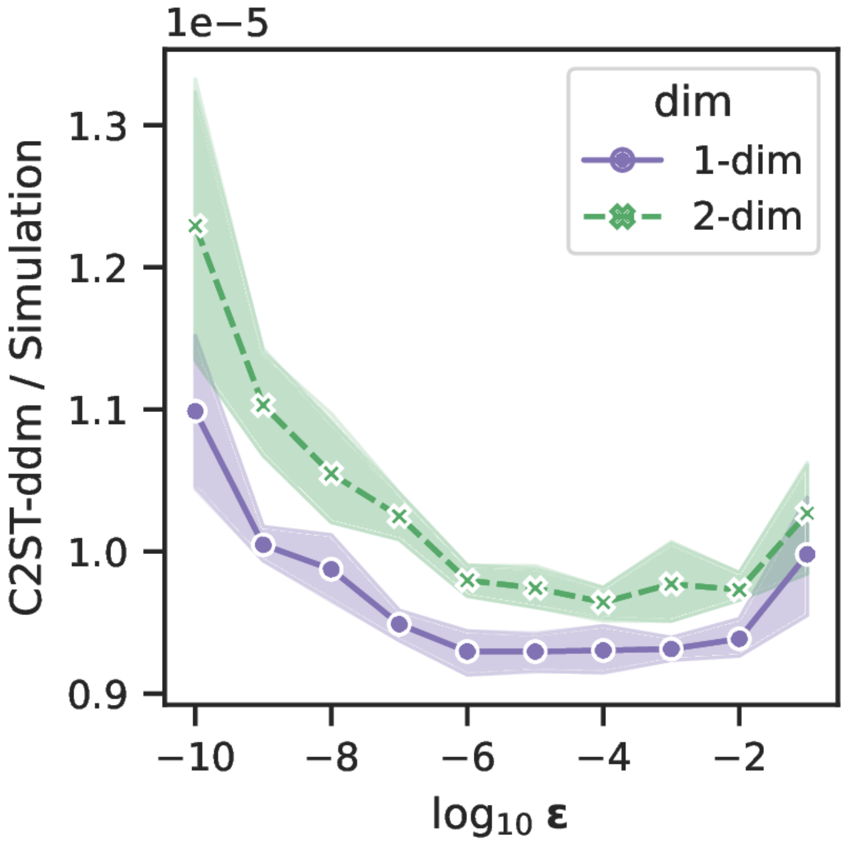




Classifier 2-Sample Test (C2ST)

Kullback-Leibler (KL) divergence.

Torus: How did we select our cutoff?

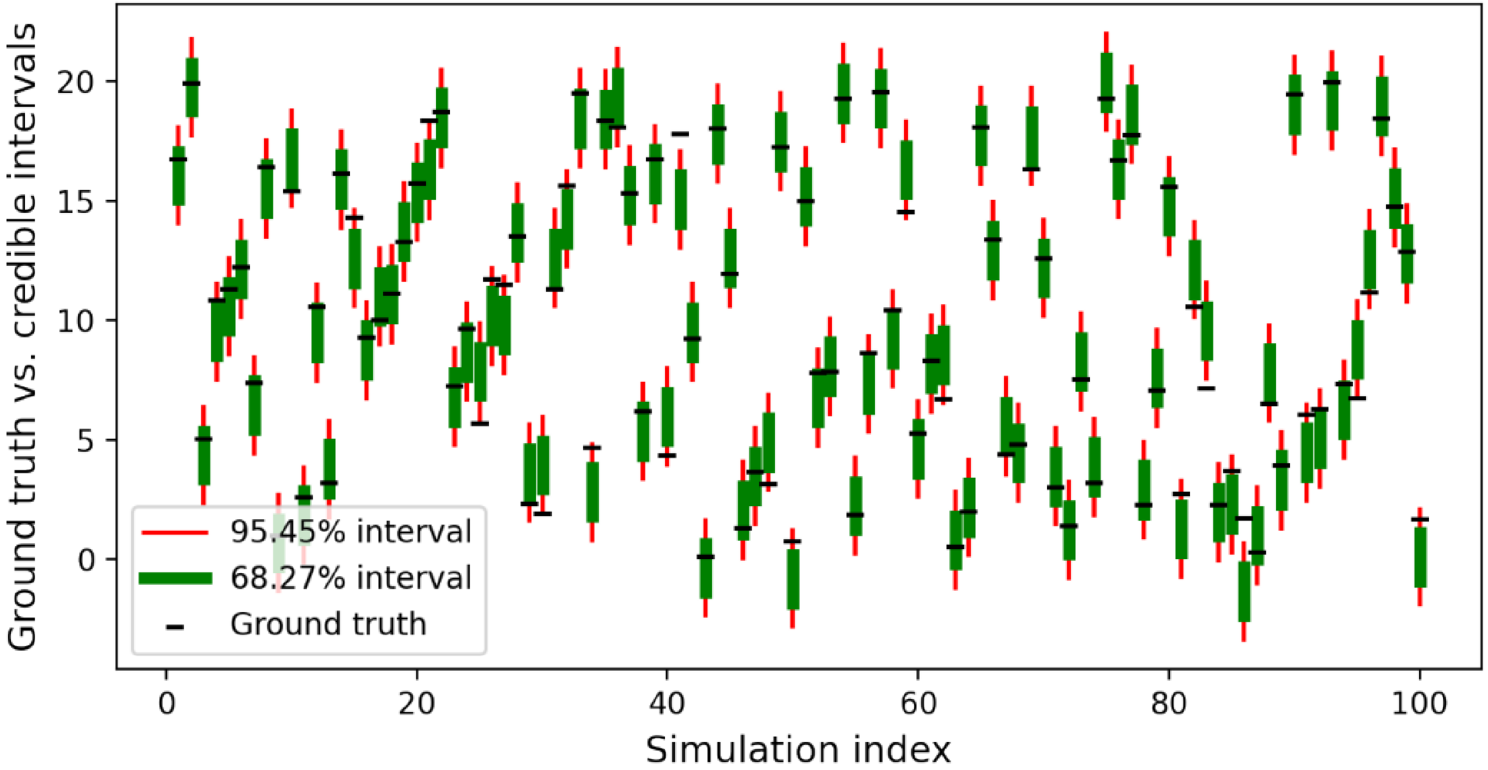


- Grid search on 10 values of the truncation cutoff ϵ.
- $\epsilon_0 = 10^{-6}$ conservatively minimized C2ST / simulation.
- Truncates a gaussian posterior at $\pm\sqrt{-2\ln\epsilon_0}\,\sigma \approx 5.26\,\sigma$. **Truncation affects only very-low probability credibility contours!**

Empirical tests with local amortization

How do we know we got the inference right
(when we don't have the truth)?

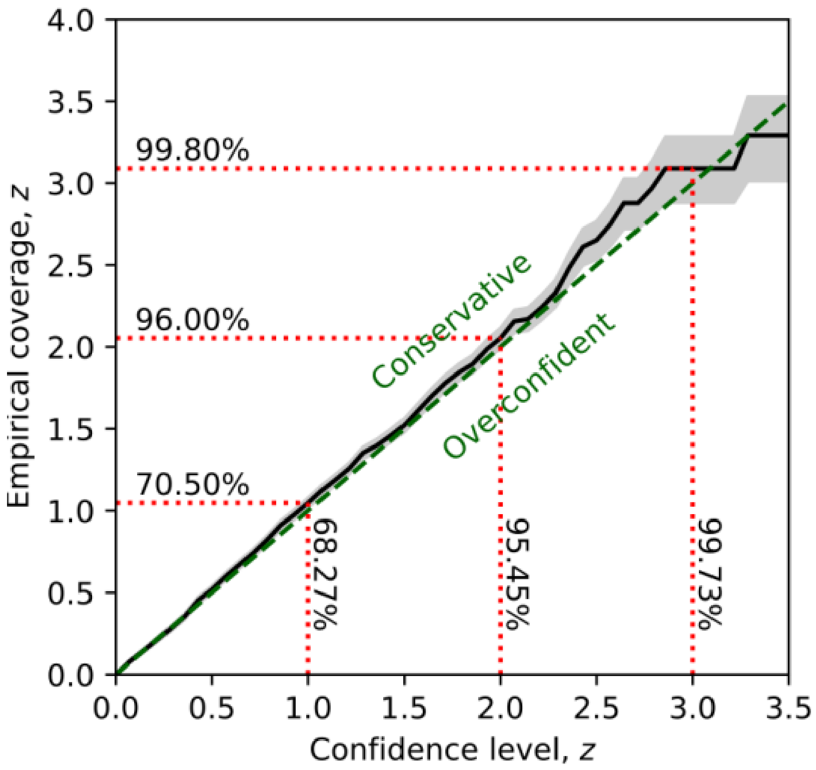
Check $p(\theta) = E_{x \sim p(x)}[\, p(\theta \mid x)\,]$!



Empirical tests with local amortization

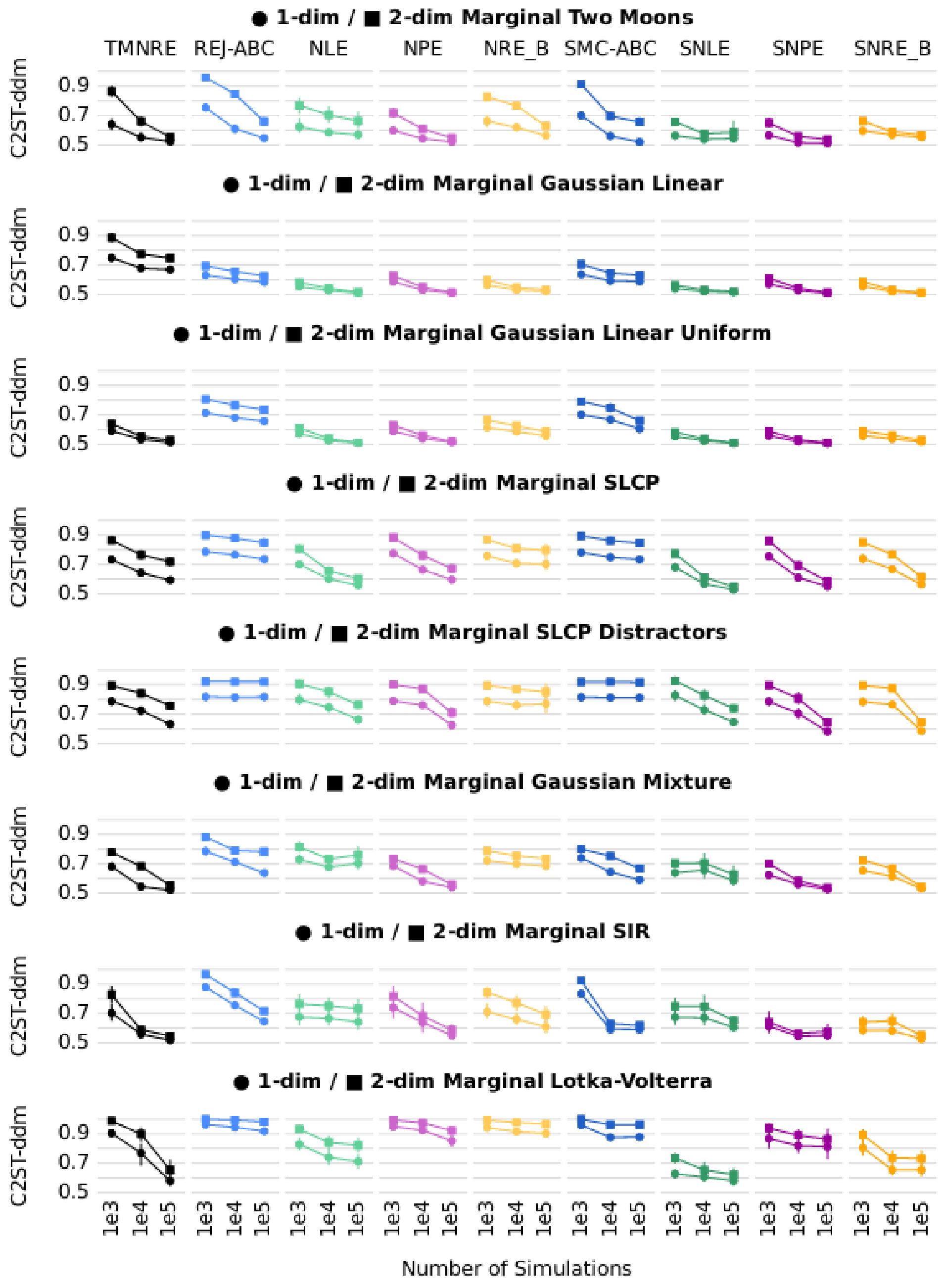
Simulation-based inference *constrains* parameters within compact regions using ***credible intervals***.

(Local) amortization enables us to check whether nominal credibility is calibrated.



$$1 - \hat{\alpha} = \mathbb{E}_{p(\boldsymbol{\vartheta}, \boldsymbol{x})} \left[\mathbb{1} \left[\boldsymbol{\vartheta} \in \Theta_{\hat{p}(\boldsymbol{\vartheta}|\boldsymbol{x})}(1 - \alpha) \right] \right]$$

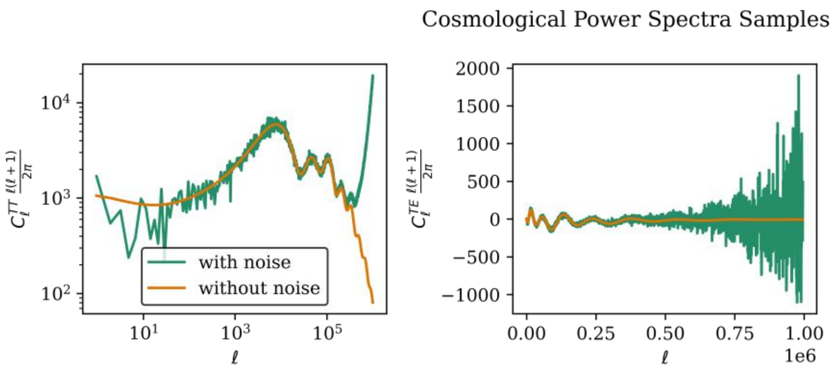
Simulation-based Inference Benchmark



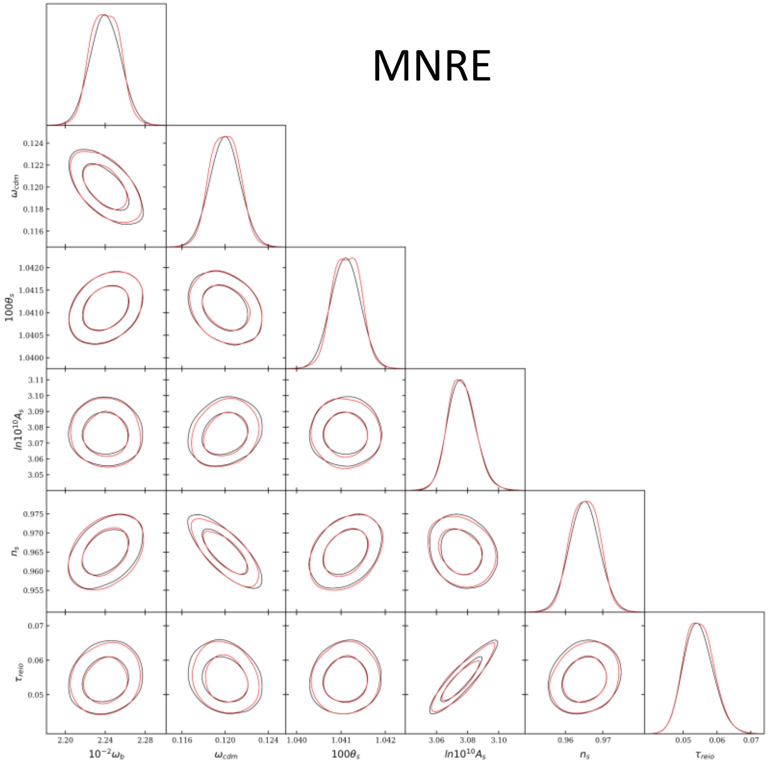
- Tested TMNRE on modified *sbibm*.
- Other methods (not TMNRE) trained to learn $p(\vartheta, \eta \mid x)$. Results on marginalized posterior samples.
- Mean & 95% CI of Classifier 2-Sample Test (C2ST) for 10 simulated x_0.
- **<u>TMNRE competitive results to sequential methods.</u>**

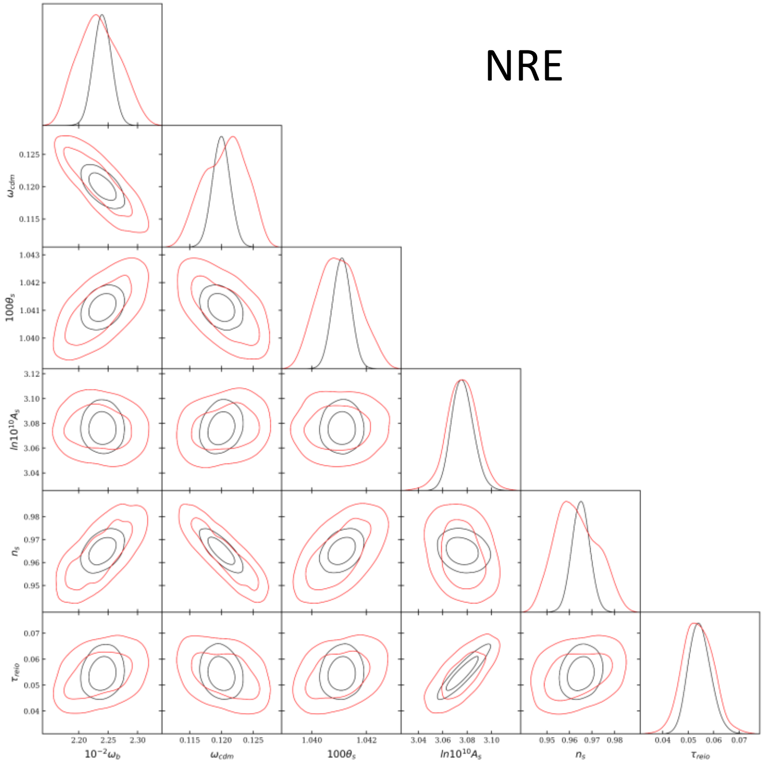
Plot style and benchmark from:
Jan-Matthis Lueckmann, et. al. 2021. Benchmarking Simulation-Based Inference.” arXiv: 2101.04653.

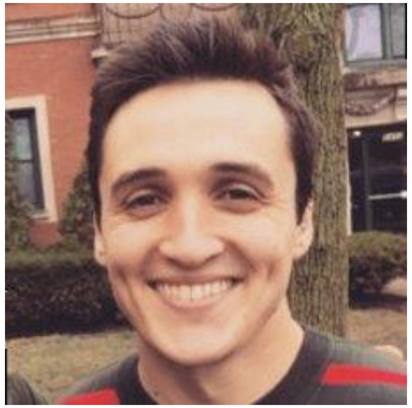
Cosmology



- Six parameters specify the lambda-CDM model.
- Data are power spectra (left) from the CMB.
- Simulator utilized to forecast the expected constraining power of future experiments.
- Budget of 5,000 simulations.







See Alex Cole's presentation at 16:30!

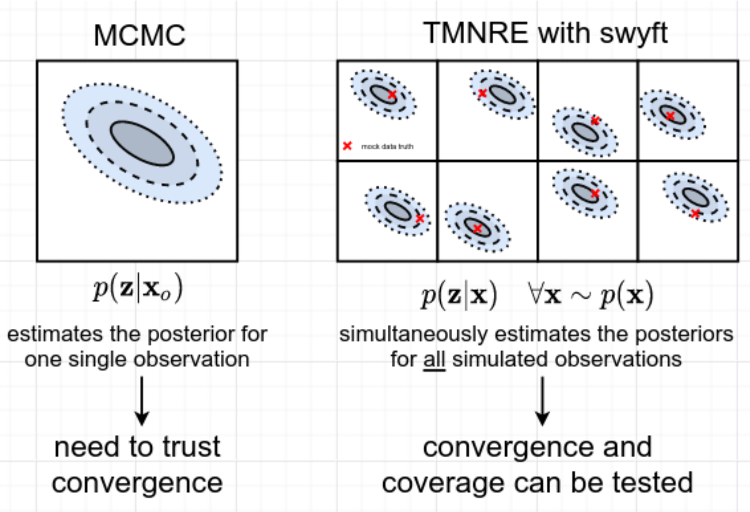
Conclusions

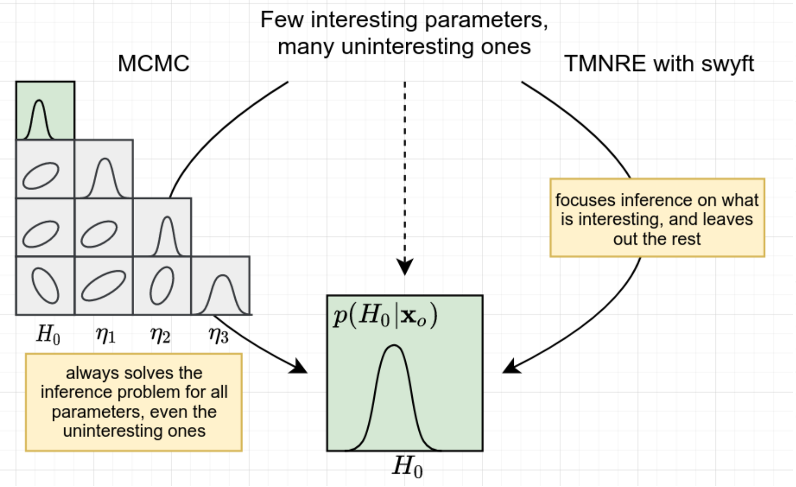
- Marginal Neural Ratio Estimation for *totally* amortized marginal inference.
- Proposed an iterative scheme (Truncated Marginal NRE) to focus on $p(\theta \mid x_0)$.
  - Increases simulation efficiency with truncation.
  - Enables empirical testing through *local* amortization.
  - **A method with both properties is unique.**

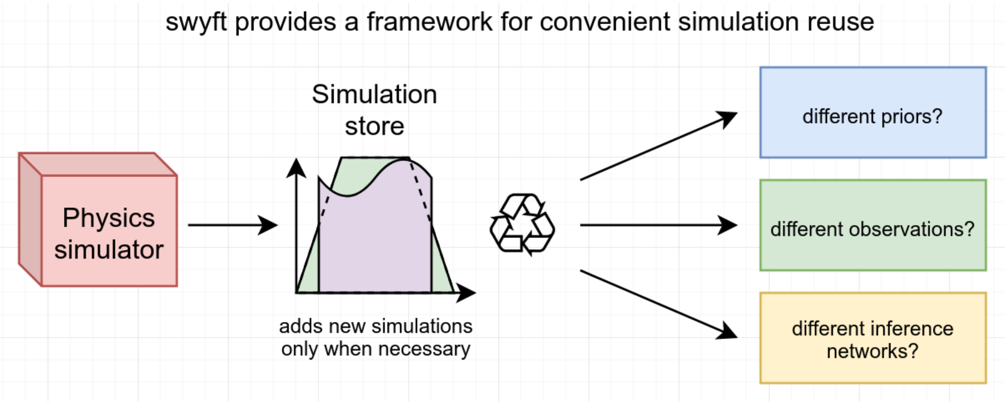
Packages:

swyft – implementation of method – https://github.com/undark-lab/swyft

tmnre – experimental results – https://github.com/bkmi/tmnre







Extra Slides

Hopefully, the answer to your question can be found in the next few slides…

Truncating the prior, based on the estimate posterior

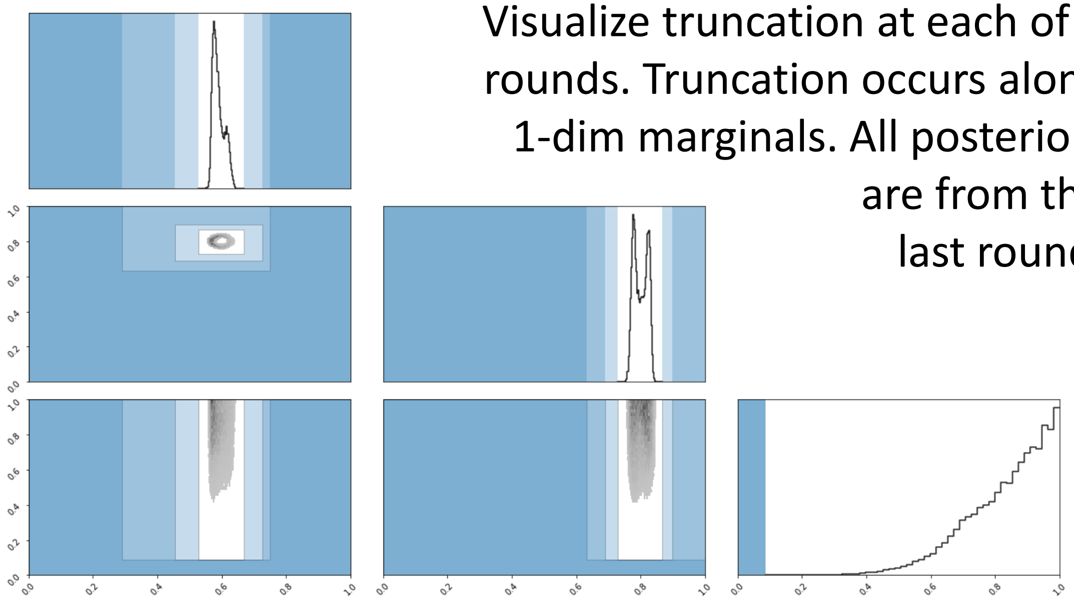
Why amortize the posterior when our focus is $p(\theta \mid x_o)$? --> Local amortization.

Determine region of interest: (*) $\Gamma = \{\theta \in \Omega \mid \forall d = 1, \dots, D: \frac{p(\theta_d \mid x_o)}{\max_{\theta_d} p(\theta_d \mid x_o)} > \delta\}$.

Discard parameters which lie outside this region (far tails).

Estimate Γ in a sequence of rounds:

1. Initialize $\Gamma^{(1)} = \Omega$.
2. Simulate data in $\Gamma^{(m)}$ & Train a ratio estimator on every 1-dim marginal.
3. Approximate (*) with the previous round's estimator and truncate.
4. Repeat until $\frac{\int 1_{\Gamma^{(m)}}(\theta) p(\theta) d\theta}{\int 1_{\Gamma^{(m-1)}}(\theta) p(\theta) d\theta} > \beta$.
5. Learn arbitrary marginal posteriors in the Γ estimate.



Visualize truncation at each of 3 rounds. Truncation occurs along 1-dim marginals. All posteriors are from the last round.

We use $\epsilon_0 = 10^{-6}$, for a gaussian joint posterior this truncates at $\pm\sqrt{-2 \ln \epsilon_0}\,\sigma$.
Truncation affects only very-low probability credibility contours.

C2ST-ddm

Classifier 2-Sample Test per d-Dimensional Marginal (C2ST-ddm) is a test statistic which reports the average *Classifier 2-Sample Test (C2ST)* across a set of d-dimensional marginals.

$X \sim P(X), Y \sim Q(Y)$ with $X, Y \in R^D$ and hyperparameter $1 \leq d \leq D$ that represents the marginal dimensionality of interest.

Let $(S_P, S_Q) \coloneqq \{(S_{P_k}, S_{Q_k}) : k \in \{1, 2, \dots, \binom{D}{d}\}\}$ where $S_{P_k} := \{x_k^{(1)}, \dots, x_k^{(n)} \sim P(X_k)$ and $S_{Q_k} := \{y_k^{(1)}, \dots, y_k^{(n)} \sim P(Y_k)$ are sets of n samples drawn from the kth d-dimensional marginal of P and Q respectively.

$$C2ST - ddm(S_P, S_Q) \coloneqq \frac{1}{K} \sum_{k=1}^{K} C2ST(S_{P_k}, S_{Q_k}), \text{ with } K = \binom{D}{d}.$$