Telescoping ratio estimation for Simulation-based Inference

Dan Leonte, Raphaël Huser, Almut Veraart

King Abdullah University of Science and Technology, Saudi Arabia.
Imperial College London, United Kingdom.

Abstract

We propose a fast, MCMC-free, sample-efficient Simulation-based Inference (SBI) approach for the amortized posterior inference of intractable stochastic processes.

Main ideas:

- learn the posterior density sequentially across parameter dimensions,
- efficiently generate independent posterior samples with Chebyshev polynomial approximations even if MCMC mixes poorly,
- introduce novel quality checks and further amortize via post-hoc calibration.

Neural ratio estimation (NRE)

Let $\mathbf{Z} \sim q(\mathbf{z})$ and $\tilde{\mathbf{Z}} \sim \tilde{q}(\mathbf{z})$ be supported on $\mathcal{Z}$. The ratio $r(\mathbf{z}) = q(\mathbf{z})/\tilde{q}(\mathbf{z})$ can be approximated by training a classifier to distinguish between samples from q and $\tilde{q}$

$$c^* := \underset{c:\mathcal{Z}\to[0,1]}{\arg\max}\ \mathbb{E}\left[\log c(\mathbf{Z}) + \log\left(1 - c(\tilde{\mathbf{Z}})\right)\right]. \qquad \text{(BCE)}$$

The optimal classifier satisfies $c^* = \frac{q(\mathbf{z})}{q(\mathbf{z})+\tilde{q}(\mathbf{z})}$ and yields $r(\mathbf{z}) = \frac{c^*(\mathbf{z})}{1-c^*(\mathbf{z})}$.

Applications: ▶ Covariate shift ▶ Direct density estimation with $\tilde{q} = \mathcal{N}(\mathbf{0}, I)$.

NRE in SBI: With $\mathbf{z} = (\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{Z} = \mathcal{X}\times\Theta \subset \mathbb{R}^k \times \mathbb{R}^m$, set $q = p(\boldsymbol{x}, \boldsymbol{\theta})$ and $\tilde{q} = p(\boldsymbol{x})p(\boldsymbol{\theta})$

$$r(\mathbf{z}) = \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x})p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta})}{p(\boldsymbol{x})} \propto p(\boldsymbol{x} \mid \boldsymbol{\theta}) \text{ for fixed } \boldsymbol{x}.$$

Telescoping ratio estimation (TRE) for SBI and post-training calibration

Let $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^m) \in \Theta \subset \mathbb{R}^m$, $\boldsymbol{\theta}^{i:j} = (\theta^i, \ldots, \theta^j)$ and

$$q_i(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{\theta}^{1:i})p(\boldsymbol{\theta}^{i+1:m}), \qquad i = 1, \ldots, m-1,$$

with $q_m(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{\theta})$, $q_0(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x})p(\boldsymbol{\theta})$ and

$$r_i(\boldsymbol{x}, \boldsymbol{\theta}) := \frac{q_i(\boldsymbol{x}, \boldsymbol{\theta})}{q_{i-1}(\boldsymbol{x}, \boldsymbol{\theta})} = \frac{p(\boldsymbol{x}, \boldsymbol{\theta}^{1:i})p(\boldsymbol{\theta}^{i+1:m})}{p(\boldsymbol{x}, \boldsymbol{\theta}^{1:i-1})p(\boldsymbol{\theta}^{i:m})} = \frac{p(\theta^i \mid \boldsymbol{x}, \boldsymbol{\theta}^{1:i-1})}{p(\theta^i \mid \boldsymbol{\theta}^{i+1:m})}.$$

Then $r(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^m r_i(\boldsymbol{x}, \boldsymbol{\theta})$ can be approximated sequentially.

Sample efficiency If $p(\boldsymbol{\theta}) = p(\theta^1)\cdots p(\theta^m)$, then

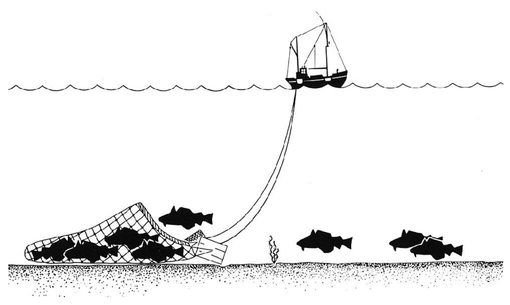
$$D_{\mathsf{KL}}(p(\boldsymbol{x}, \boldsymbol{\theta}) \parallel p(\boldsymbol{x})p(\boldsymbol{\theta})) = D_{\mathsf{KL}}(q_m \parallel q_0) = \sum_{i=1}^m D_{\mathsf{KL}}(q_i \parallel q_{i-1}).$$

Post-training calibration: Apply monotonic transformations $T\colon [0,1] \to [0,1]$ to each of the trained classifiers $\hat{c}_{\mathsf{cal}} = T \circ \hat{c}$:
▶ Platt scaling ▶ Beta-calibration ▶ Isotonic regression

Lévy bases and trawl processes

Lévy bases: A Lévy basis L is a collection of infinitely-divisible, real-valued random variables $\{L(A) : A \in \mathcal{B}_{\mathsf{Leb}}(\mathbb{R}^d)\}$ such that for any disjoint sets $A, B \in \mathcal{B}_{\mathsf{Leb}}(\mathbb{R}^d)$, the random variables $L(A)$ and $L(B)$ are independent and $L(A \cup B) = L(A) + L(B)$.

Trawl process: Define $X_t = L(A_t)$, where $A_t = A + (t, 0)$ for some fixed set $A \subset \mathbb{R}^d = \mathbb{R}^2$.



Correlation structure:

$$\mathrm{Corr}(X_s, X_t) = \frac{\mathrm{Leb}(A_s \cap A_t)}{\mathrm{Leb}(A)}.$$

Marginal distribution: any infinitely divisible law.

Assessing and improving the quality of the learned approximation

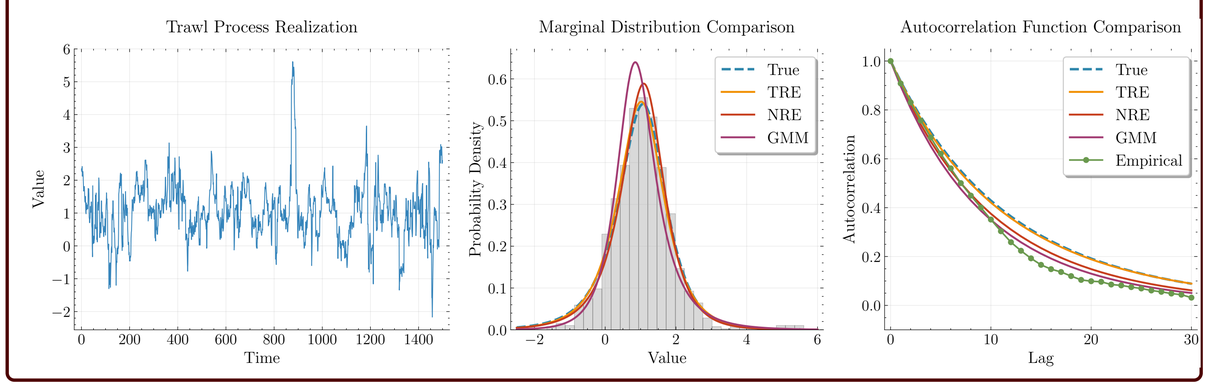
TRE decomposition for trawl processes

$$r(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x})p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{x}, \theta_{\mathsf{acf}}, \mu, \sigma, \beta)}{p(\boldsymbol{x})\,p(\theta_{\mathsf{acf}})\,p(\mu)\,p(\sigma)\,p(\beta)}$$

$$= \underbrace{\frac{p(\boldsymbol{x}, \theta_{\mathsf{acf}}, \mu, \sigma, \beta)}{p(\boldsymbol{x}, \theta_{\mathsf{acf}}, \mu, \sigma)\,p(\beta)}}_{\beta\text{ classifier}} \cdot \underbrace{\frac{p(\boldsymbol{x}, \theta_{\mathsf{acf}}, \mu, \sigma)}{p(\boldsymbol{x}, \theta_{\mathsf{acf}}, \mu)\,p(\sigma)}}_{\sigma\text{ classifier}} \cdot \underbrace{\frac{p(\boldsymbol{x}, \theta_{\mathsf{acf}}, \mu)}{p(\boldsymbol{x}, \theta_{\mathsf{acf}})p(\mu)}}_{\mu\text{ classifier}} \cdot \underbrace{\frac{p(\boldsymbol{x}, \theta_{\mathsf{acf}})}{p(\boldsymbol{x})\,p(\theta_{\mathsf{acf}})}}_{\text{ACF classifier}}$$

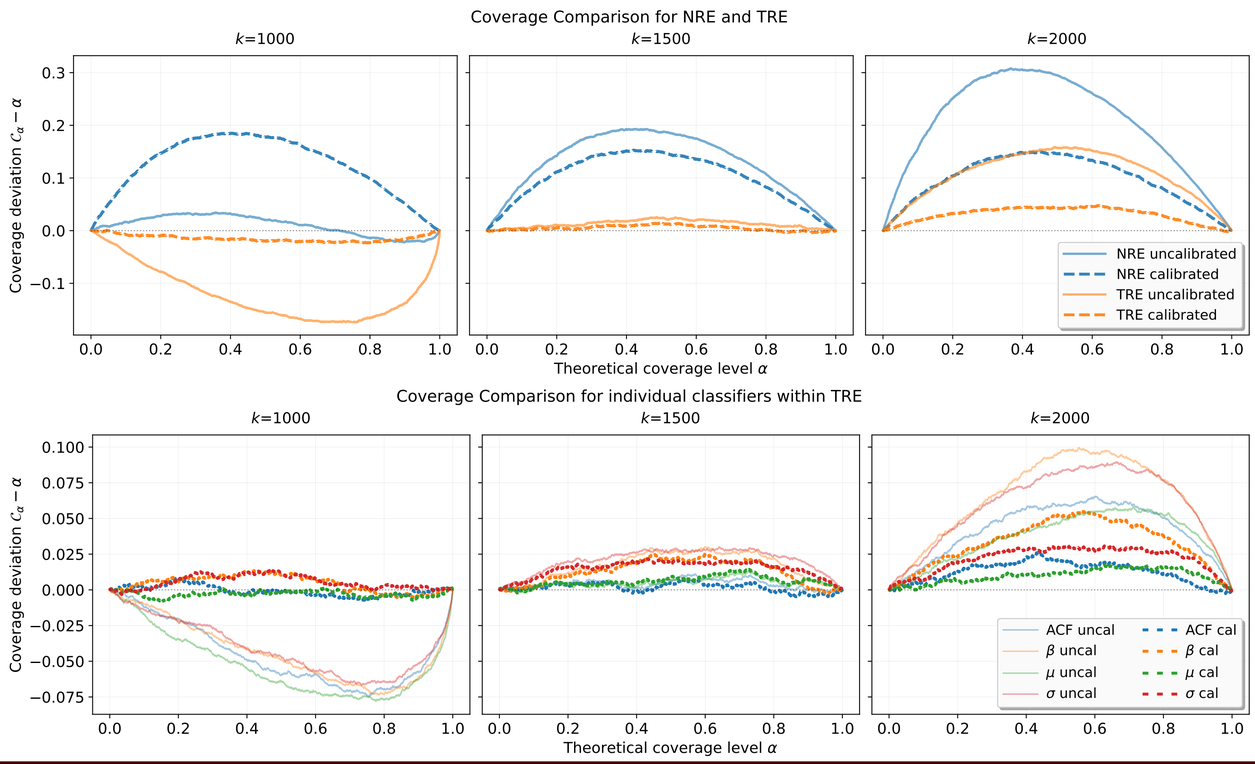
$$= \underbrace{\frac{p(\beta \mid \boldsymbol{x}, \theta_{\mathsf{acf}}, \mu, \sigma)}{p(\beta)}}_{\beta\text{ classifier}} \cdot \underbrace{\frac{p(\sigma \mid \boldsymbol{x}, \theta_{\mathsf{acf}}, \mu)}{p(\sigma)}}_{\sigma\text{ classifier}} \cdot \underbrace{\frac{p(\mu \mid \boldsymbol{x}, \theta_{\mathsf{acf}})}{p(\mu)}}_{\mu\text{ classifier}} \cdot \underbrace{\frac{p(\theta_{\mathsf{acf}} \mid \boldsymbol{x})}{p(\theta_{\mathsf{acf}})}}_{\text{ACF classifier}}.$$

Locating the posterior mode



Posterior coverage: Let $\Theta_{\hat{p}(\cdot\,|\boldsymbol{x})}(1-\alpha)$ be the $1-\alpha$ highest posterior density region of $\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$. The coverage at level $1-\alpha$ is

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

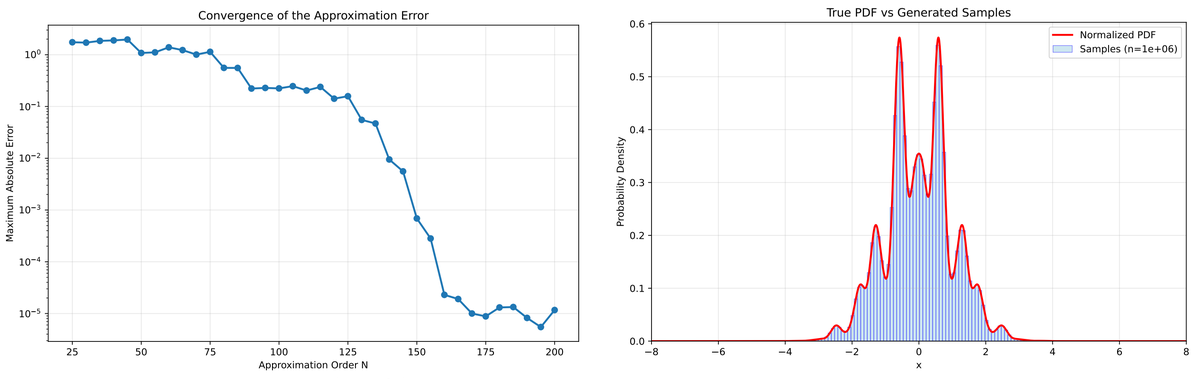


Efficient posterior inference

The Chebyshev polynomials $T_n : [-1, 1] \to \mathbb{R}$ are given by

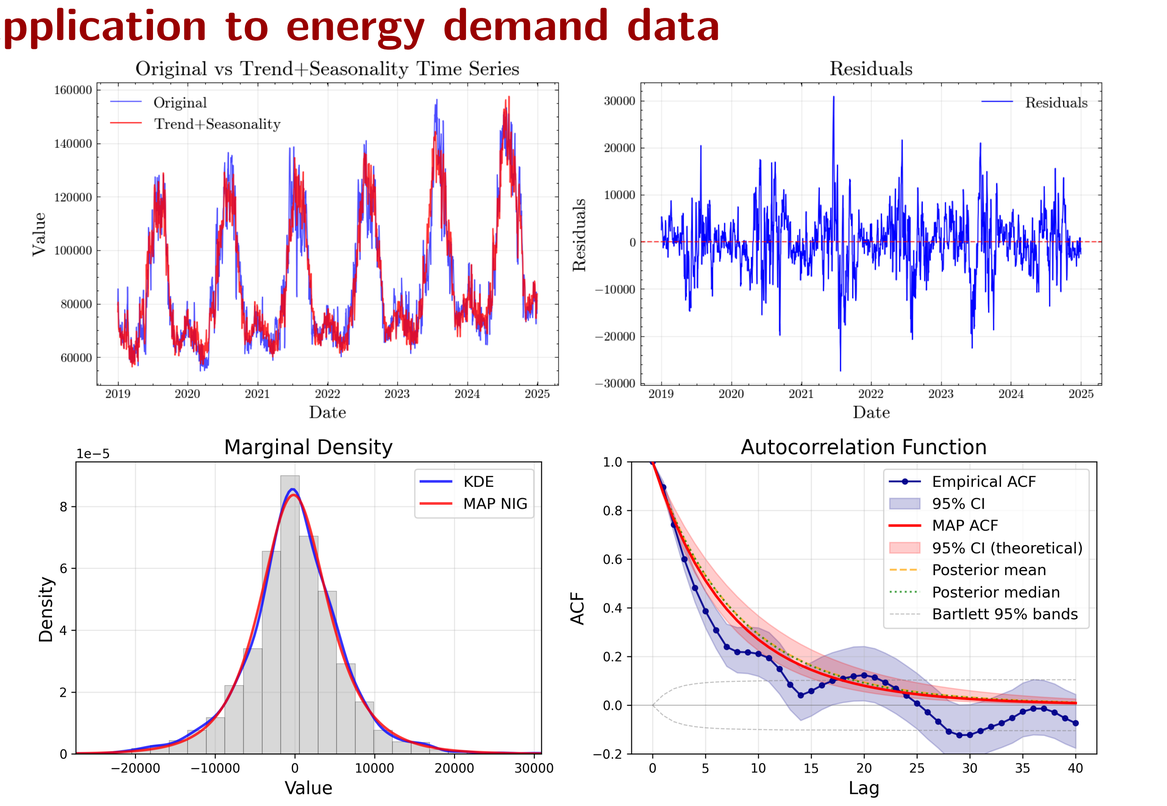
$$T_n(x) = \cos(n \arccos x), \quad n = 0, 1, 2, \ldots$$

- form an orthogonal basis with weight $w : [-1, 1] \to \mathbb{R}$, $w(x) = 1/\sqrt{1-x^2}$,
- can be efficiently evaluated and integrated by Clenshaw's algorithm.



$$f(x) = \exp\left(-\frac{x^2}{2}\right)\left(1 + \sin^2(3x)\right)\left(1 + \cos^2(5x)\right), \quad -8 \le x \le 8.$$

Application to energy demand data



References:

[1] Cranmer, K., Pavez, J. & Louppe, G. (2015), 'Approximating likelihood ratios with calibrated discriminative classifiers', arXiv: 1506.02169.
[2] Chatterjee, S. & Diaconis, P. (2018), 'The sample size required in importance sampling', Ann. Appl. Probab. 28, 1099–1135.
[3] Hermans, J., Begy, V. & Louppe, G. (2020), Likelihood-free MCMC with amortized approximate ratio estimators, in '37th International Conference on Machine Learning', PMLR, pp. 4239–4248.
[4] Rhodes, B., Xu, K. & Gutmann, M. U. (2020), Telescoping density-ratio estimation, in 'Advances in Neural Information Processing Systems', Curran Associates, pp. 4905–4916.