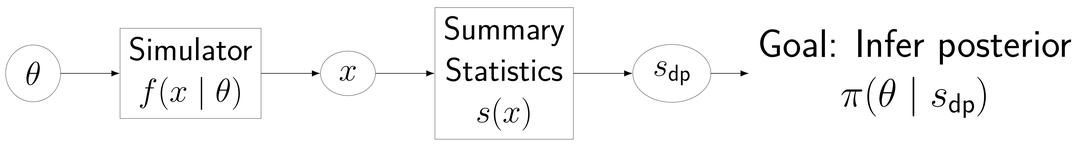
# Simulation-based Bayesian Inference from Privacy Protected Data

Yifei Xiong[1], Nianqiao Phyllis Ju[1], Sanguo Zhang[2]

[1]Department of Statistics, Purdue University [2]University of Chinese Academy of Sciences

Published in Transactions on Machine Learning Research (TMLR) arXiv:2310.12781




## Introduction

Many modern statistical analysis and machine learning applications require training models on sensitive user data. Differentially private algorithms inject calibrated noise into the confidential data or during the data analysis process to produce privacy-protected datasets or queries. This work proposes simulation-based inference methods from privacy-protected datasets.

## Notations

- $\theta \in \Theta$: model parameter
- $\pi(\theta)$: prior of the model parameter
- $x = (x_1, \cdots, x_n) \in \mathbb{X}^n$ : confidential database
- $f(x \mid \theta)$: intractable likelihood function as a 'simulator'
- $s_{\text{dp}}$: differentially private queries for the confidential database $x$

## Likelihood-free Inference

Given observed dataset $x^o$, how to learn the parameter posterior $\pi(\theta \mid x^o) \propto \pi(\theta) f(x^o \mid \theta)$?

- **Approximate Bayesian Computation (ABC)** [1]: Circumvent likelihood evaluations with simulations.
  - Accept $\theta^{(i)}$ if $d(x^{(i)}, x^o) < \epsilon$, where $x^{(i)} \sim f(x \mid \theta^{(i)})$
  - Depends on distance function $d(\cdot, \cdot)$; Requires multiple simulations
  - SMC-ABC improves ABC by specifying a sequence of intermediate target distributions
- **Neural density estimation (NDE)** [2]: approximate $\pi(\theta \mid x^o)$ with neural density $\{q_\phi(\theta \mid x)\}_\phi$ by an architecture $q$ and weight parameters $\phi$.

$$\hat{\phi} = \arg\min_{\phi} \mathbb{E}_{p(\theta,x)} \left[ -\log q_\phi(\theta \mid x) \right]$$

  - Training relies on Monte Carlo approximations

## Differential Privacy

Let $\eta(s_{\text{dp}} \mid x)$ be the conditional density of the private output $s_{\text{dp}}$ given the confidential data $x$. We say $\eta$ satisfies $\epsilon$-DP if, for all possible values of $s_{\text{dp}}$ and for neighboring datasets $x, x'$,

$$\frac{\int_A \eta(s_{\text{dp}} \mid x)\, \mathrm{d}s_{\text{dp}}}{\int_A \eta(s_{\text{dp}} \mid x')\, \mathrm{d}s_{\text{dp}}} \leq \exp(\epsilon), \quad \forall A \subseteq \text{Range}(\eta).$$

## Inference with Privatized Queries

Our goal is to approximate the posterior distribution of $\theta$ given privatized queries $s_{\text{dp}}$.

$$\pi(\theta \mid s_{\text{dp}}) \propto \pi(\theta) f(s_{\text{dp}} \mid \theta) \propto \pi(\theta) \int_{\mathbb{X}^n} f(x \mid \theta) \eta(s_{\text{dp}} \mid x) \mathrm{d}x$$

- Intractable $f(x \mid \theta)$ and integration over $\mathbb{X}^n$
- Data-augmentation MCMC methods [3] work when $f(x \mid \theta)$ is tractable

## Our Methods

We present two complementary approaches:

**Private data likelihood estimation (PLE)**: approximate the private data marginal likelihood $f(s_{\text{dp}} \mid \theta)$. Minimizing

$$\mathbb{E}_{\pi(\theta)} \left[ \mathcal{D}_{\text{KL}} \left( f(s_{\text{dp}} \mid \theta) \| q_\phi(s_{\text{dp}} \mid \theta) \right) \right]$$

is equivalent to minimizing

$$\ell_{\text{PLE}}(\phi) = \mathbb{E}_{p(\theta,x)} \left[ -\int_{\mathbb{S}} \eta(s_{\text{dp}} \mid x) \log q_\phi(s_{\text{dp}} \mid \theta) \mathrm{d}s_{\text{dp}} \right]$$

up to a constant independent of $\phi$. The posterior approximation can be $\hat{\pi}_{\text{PLE}}(\theta) \propto \pi(\theta) q_{\hat{\phi}}(s_{\text{dp}} \mid \theta)$.

**Private data posterior estimation (PPE)**: approximate $\pi(\theta \mid s_{\text{dp}})$ directly. Minimizing

$$\mathbb{E}_{\pi(x)} \left[ \mathcal{D}_{\text{KL}} \left( \pi(\theta \mid s_{\text{dp}}) \| q_\phi(\theta \mid s_{\text{dp}}) \right) \right]$$

is equivalent to minimizing

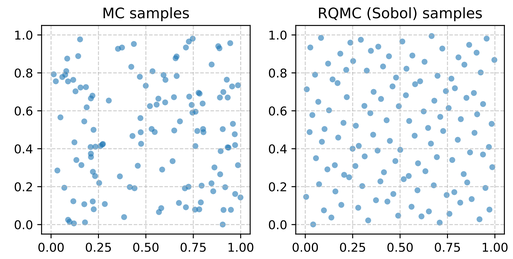
$$\ell_{\text{PPE}}(\phi) = \mathbb{E}_{p(\theta,x)} \left[ -\int_{\mathbb{S}} \eta(s_{\text{dp}} \mid x) \log q_\phi(\theta \mid s_{\text{dp}}) \mathrm{d}s_{\text{dp}} \right]$$

## Nested RQMC Estimators

- The integral in the loss function

$$I(\theta, x) = \int \eta(s_{\text{dp}} \mid x) g(s_{\text{dp}}, \theta) \mathrm{d}s_{\text{dp}}$$

can be approximated with standard Monte Carlo integration with the root-mean-squared-error (RMSE) on the order of $\mathcal{O}(M^{-1/2})$.




- Randomized quasi-Monte Carlo (RQMC) can generate correlated, low-discrepancy sequences to reduce the RMSE to $\mathcal{O}(M^{-1+\delta})$ under mild conditions.

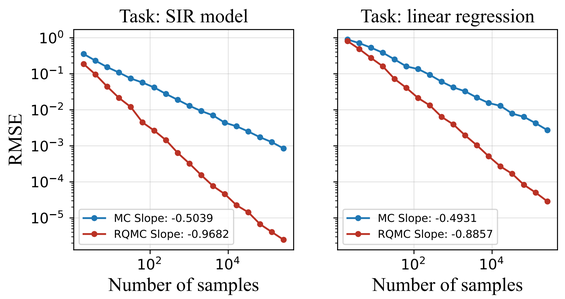



Figure: Rate of convergence of MC and RQMC. The RMSE of the RQMC method is approximately $\mathcal{O}(M^{-1})$.

## Sequential Neural Estimations

- We incorporate the current neural density estimator $q_\phi$ into the proposal distribution of the next training round
- Sequential training procedures can gradually move $q_\phi$ towards high-density regions of the private data posterior, and thus achieve good accuracy with fewer samples from the simulator

## Experiments

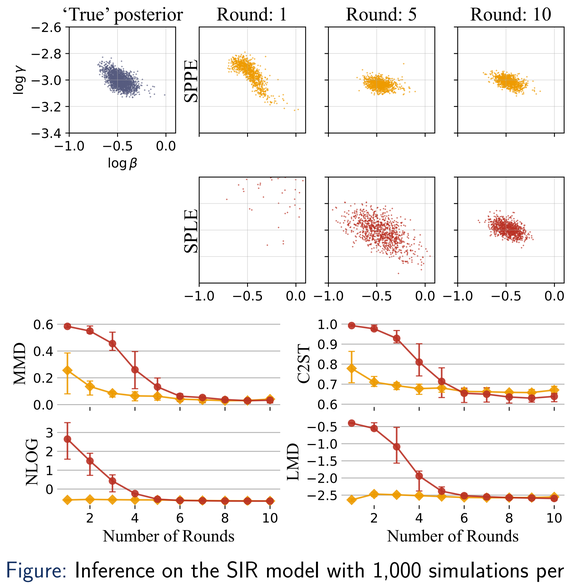



Figure: Inference on the SIR model with 1,000 simulations per round. Top: Convergence of posterior estimations; Bottom: Approximation accuracy by SPPE (orange) and SPLE (red) against the number of rounds.

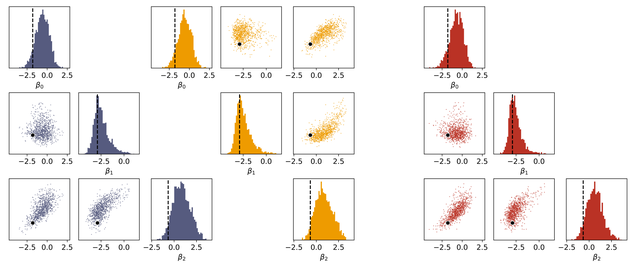


Figure: Posterior comparison on the Bayesian linear regression. Grey: Ground truth posterior; orange: SPPE; red: SPLE.

## References

[1] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[2] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

[3] Nianqiao Ju, Jordan Awan, Ruobin Gong, and Vinayak Rao. Data augmentation MCMC for Bayesian inference from privatized data. *Advances in Neural Information Processing Systems*, 35:12732–12743, 2022.