Neural Approximate Sufficient Statistics for Implicit Models

Yanzhi Chen*[1] , Dinghuai Zhang*[2] , Michael U. Gutmann[1] , Aaron Courville[2] , Zhanxing Zhu[3]   [1]The University of Edinburgh, [2]MILA, [3]Beijing Institute of Big Data Research

Likelihood-free inference (LFI)

LFI considers the task of Bayesian inference when the likelihood function of the model is intractable but sampling data from the model is possible[1]:

$$\underbrace{\pi(\boldsymbol{\theta}|\mathbf{x}_o)}_{\text{posterior}} \propto \underbrace{\pi(\boldsymbol{\theta})}_{\text{prior}} \underbrace{p(\mathbf{x}_o|\boldsymbol{\theta})}_{?\ \text{likelihood}}$$

1. sample data: $\mathcal{D} = \{\mathbf{x}_i, \boldsymbol{\theta}_i\}_{i=1}^n, \quad \mathbf{x}_i \sim p(\mathbf{x}|\boldsymbol{\theta}_i), \boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta})$

2. learn $p(\boldsymbol{\theta}|\mathbf{x})$ with the data with e.g. ABC[2], NDE[3,4]

Curse of dimensionality

However, most existing methods suffer from the curse of dimensionality when modeling high-dimensional distributions. Our interest here is to find a low-dimensional statistic

$$\mathbf{s} = s(\mathbf{x})$$

that is near-sufficient, and could be applied to a wide range of LFI methods:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_o) \approx \pi(\boldsymbol{\theta}|\mathbf{s}_o) \propto \pi(\boldsymbol{\theta}) p(\mathbf{s}_o|\boldsymbol{\theta})$$

existing ways[7,8] for learning summary statistics cannot guarantee sufficiency

Neural sufficient statistics

learning sufficient statistics $\equiv$ learning infomax representation of data

Proposition 1. *Let $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$, and $s : \mathcal{X} \to \mathcal{S}$ be a deterministic function. Then $\mathbf{s} = s(\mathbf{x})$ is a sufficient statistic for $p(\mathbf{x}|\boldsymbol{\theta})$ if and only if*

$$s = \arg\max_{S:\mathcal{X}\to\mathcal{S}} I(\boldsymbol{\theta}; S(\mathbf{x})),$$

where S is deterministic mapping and $I(\cdot;\cdot)$ is the mutual information between random variables.

$$s = \arg\max_{S:\mathcal{X}\to\mathcal{S}} I(\boldsymbol{\theta}; S(X)),$$

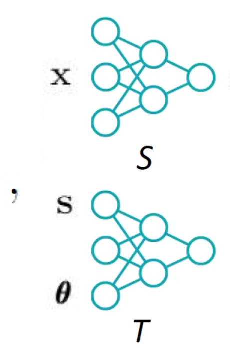
$$I(\boldsymbol{\theta}, \mathbf{s}) = KL[p(\boldsymbol{\theta}, \mathbf{s}) \| p(\boldsymbol{\theta})p(\mathbf{s})]$$

$\max_S \hat{I}^{\text{JSD}}(\boldsymbol{\theta}, S(X))$ other MI estimators $\max_S \hat{I}^{\text{DC}}(\boldsymbol{\theta}, S(X))$

We can use any proxy to KL (e.g. JSD, MMD, WD, DC) for sufficient statistics learning to achieve (a) better performance; and/or (b) faster execution time

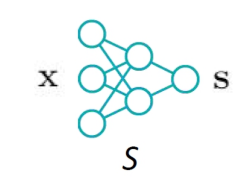
- Jensen-Shannon divergence estimator[5]:

$$\hat{I}^{\text{JSD}}(\boldsymbol{\theta}; \mathbf{s}) = \sup_{T:\Theta\times\mathcal{S}\to\mathbb{R}} \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{s})}\left[-\text{sp}(-T(\boldsymbol{\theta}, \mathbf{s}))\right] - \mathbb{E}_{p(\boldsymbol{\theta})p(\mathbf{s})}\left[\text{sp}(T(\boldsymbol{\theta}, \mathbf{s}))\right],$$



density-free, more robust than KL-based estimator

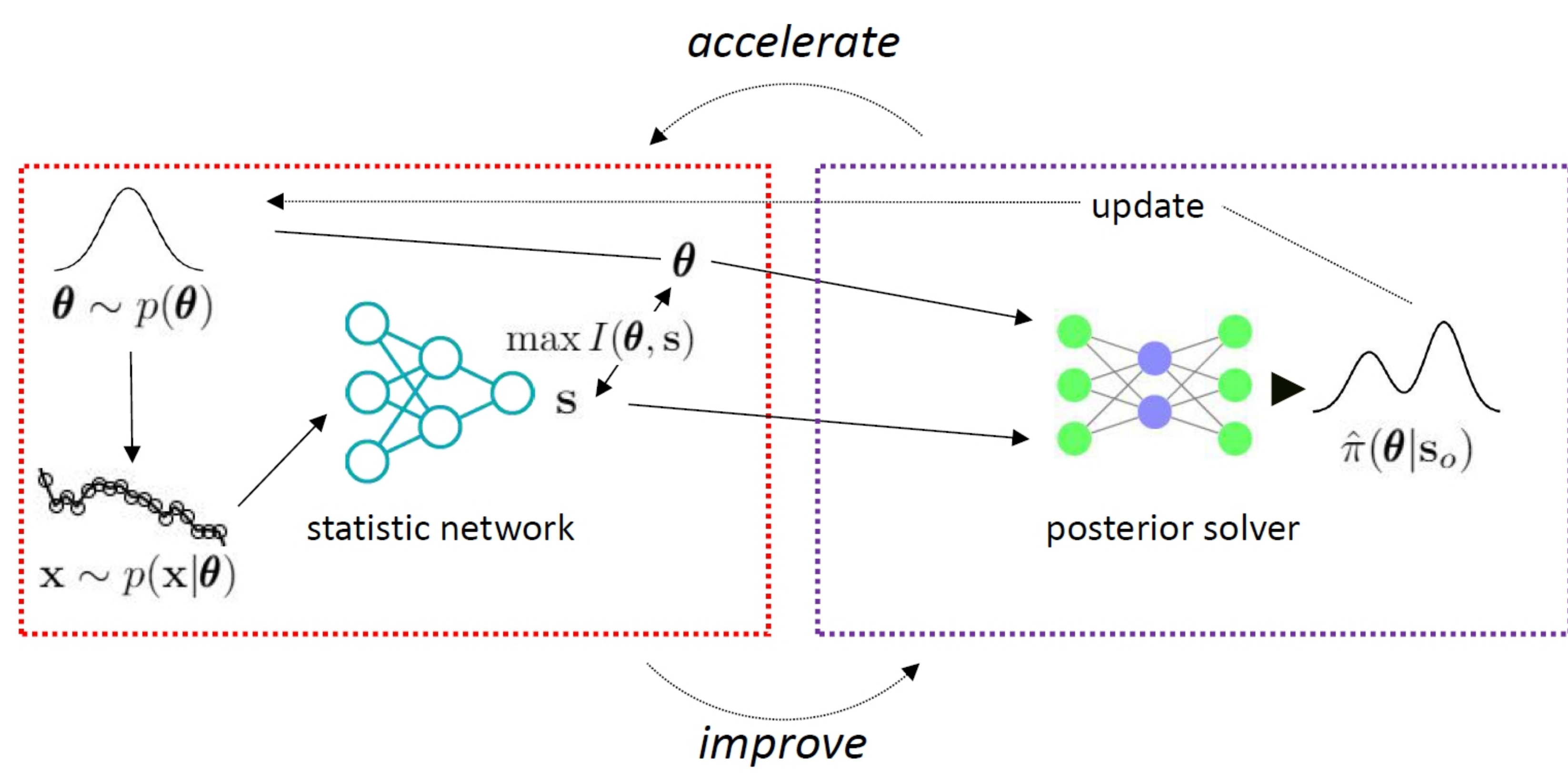
- Distance correlation estimator[6]:



$$\hat{I}^{\text{DC}}(\boldsymbol{\theta}; \mathbf{s}) = \frac{\mathbb{E}_{p(\boldsymbol{\theta},\mathbf{s})p(\boldsymbol{\theta}',\mathbf{s}')}[h(\boldsymbol{\theta}, \boldsymbol{\theta}')h(\mathbf{s}, \mathbf{s}')]}{\sqrt{\mathbb{E}_{p(\boldsymbol{\theta})p(\boldsymbol{\theta}')}[h^2(\boldsymbol{\theta}, \boldsymbol{\theta}')]} \cdot \sqrt{\mathbb{E}_{p(\mathbf{s})p(\mathbf{s}')}[h^2(\mathbf{s}, \mathbf{s}')]}},$$

ratio-free, much faster execution time but comparable performance to JSD/KL ones

Iterative statistics-posterior learning



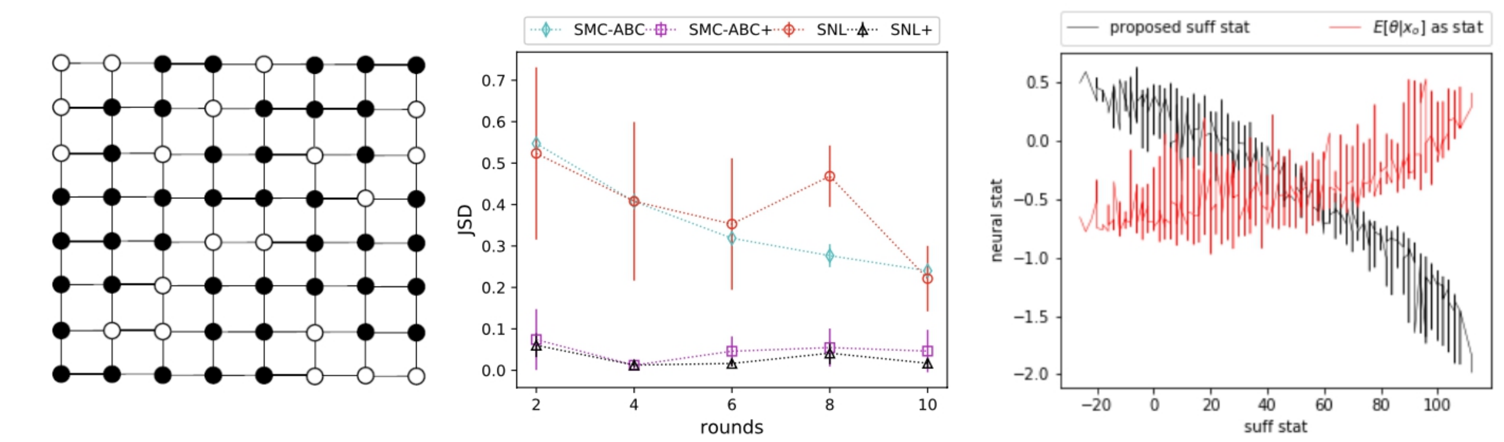
- The learned low-dimemsional statistics ***s*** can improve posterior estimate;
- The improved posterior as a better proposal accelerates the learning of ***s***

Experiments

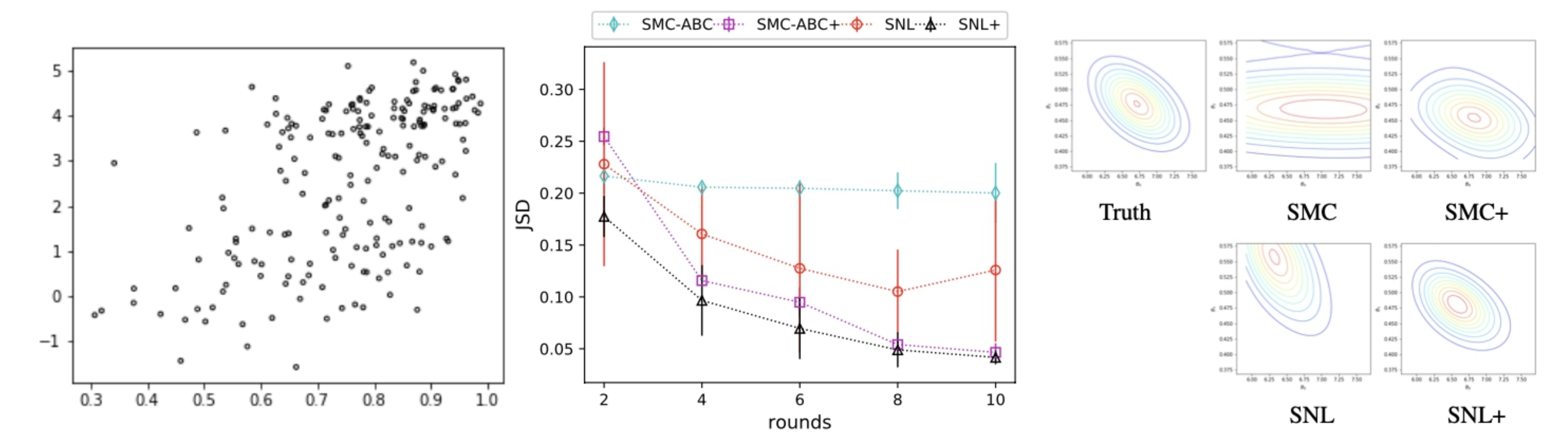
Algorithms

- SMC-ABC [5]: a traditional approximate Bayesian computation (ABC) approach
- SMC-ABC +: improved SMC-ABC with the proposed neural sufficient statistics
- SNL [4]: a recent neural density estimator (NDE) approach that learns likelihood
- SNL +: improved SNL with the proposed neural sufficient statistics

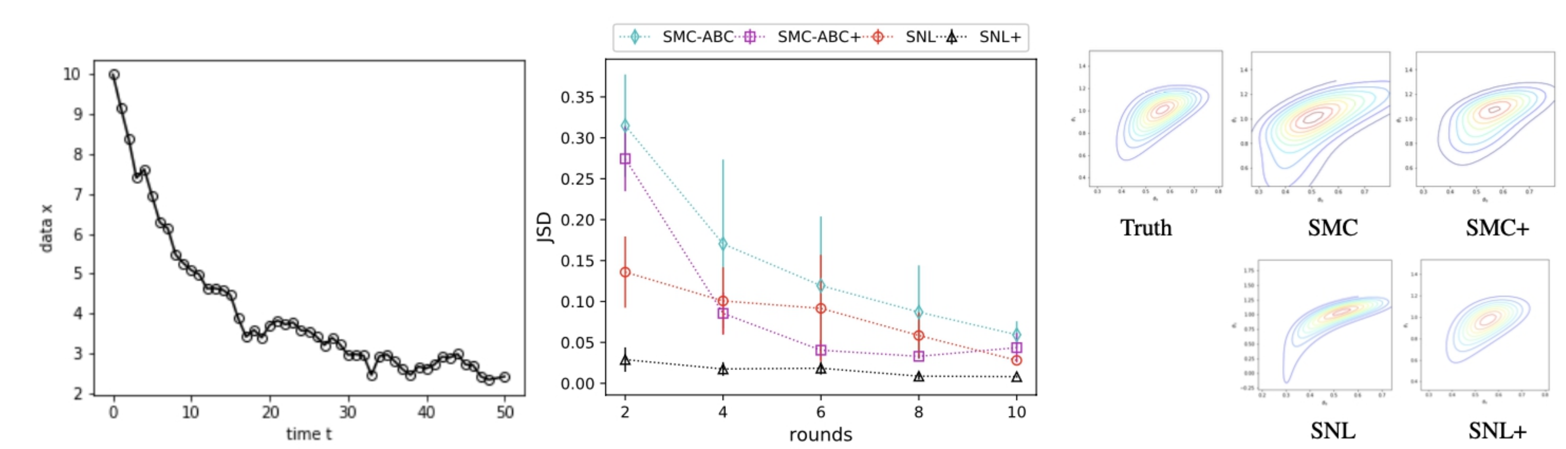
Inference problems: numerical experiments are performed on: (a) an Ising model; (b) a Gaussian copula model; (c) an Ornstein-Uhlenbeck process. The result here is for JSD estimator (see appendix for the results of other estmators).



Results on Ising model. Left: visualization of 64D observed data. Middle: the JSD between the true and the learned posteriors. Right: the relationship between the learned statistics and the sufficient statistic.



Results on Gaussian copula. Left: the observed data in this problem, which is comprised of a population of 200 i.i.d. samples. Middle: the JSD between the true/learned posteriors. Right: the contours of learned posterior.



Results on OU process. Left: the observed time-series data $\mathbf{x}_o = \{x_t\}_{t=1}^{50}$. Middle: the JSD between the true and the learned posteriors. Right: the contours of the true posterior and the learned posteriors.

References

[1] Monte Carlo methods of inference for implicit statistical models, JRSS B 1984

[2] Adaptive Approximate Bayesian Computation, Biometrika 09

[3] Fast epsilon-free Inference of Simulation Models with Bayesian Conditional Density Estimation, Neurips 16

[4] Sequential Neural Likelihood, AISTATS 19

[5] Learning deep representations by mutual information estimation and maximization, ICLR 19

[6] Partial distance correlation with methods for dissimilarities, Annals of Statistics 14

[7] Constructing summary statistics for ABC, JRSS B 09

[8] Mining gold from implicit models to improve likelihood-free inference, PNAS 20