Neural Approximate Sufficient Statistics for Implicit Models

Yanzhi Chen*[1], Dinghuai Zhang*[2], Michael Gutmann[1], Aaron Courville[2], Zhanxing Zhu[3]

[1]The University of Edinburgh, [2]MILA, [3]Beijing Institute of Big Data Research

Overview

- **Background**
- **Method**
- **Related works**
- **Results**

Background

Implicit statistical models

defined by the *data generating process* rather than the *likelihood function*[1]

$$\mathbf{x} \sim \underbrace{p(\mathbf{x}|\boldsymbol{\theta})}_{?} \quad \Leftrightarrow \quad g_{\boldsymbol{\theta}}(\epsilon), \epsilon \sim p(\epsilon)$$

Examples

SIR model (**epidemiology**), Ricker's model (**ecology**), g-and-k model (**finance**)

Background

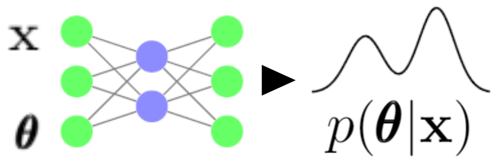
Likelihood-free inference

$$\pi(\boldsymbol{\theta}|\mathbf{x}_o) \propto \pi(\boldsymbol{\theta})\, \underbrace{p(\mathbf{x}_o|\boldsymbol{\theta})}_{?}$$

posterior prior likelihood

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

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



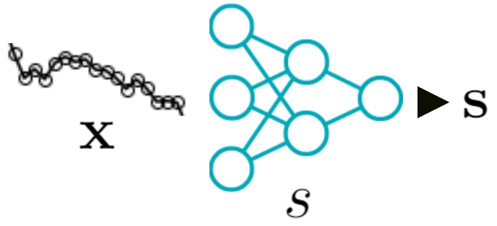
problem: high-dimensional density estimation is difficult

Method

Overview

1) first find a *low-dim, near-sufficient* statistics $s(\cdot)$

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



2) infer the posterior with **s**

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

learning $s(\cdot)$ may not require the estimation of density or density ratio

Method

Main idea

learning sufficient statistics <====> infomax representation learning

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

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

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

other MI estimators

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

we can maximize any non-KL proxy[5,6,7] of MI that has better properties

Method

Distance correlation (DC)[5] proxy:

$$\max_{S} \; \mathcal{L}(S) = \frac{\mathbb{E}^2_{p(\boldsymbol{\theta},\mathbf{x})p(\boldsymbol{\theta}',\mathbf{x}')}[h(\boldsymbol{\theta},\boldsymbol{\theta}')h(S(\mathbf{x}),S(\mathbf{x}'))]}{\mathbb{E}_{p(\boldsymbol{\theta})p(\boldsymbol{\theta}')}[h^2(\boldsymbol{\theta},\boldsymbol{\theta}')] \cdot \mathbb{E}_{p(\mathbf{x})p(\mathbf{x}')}[h^2(S(\mathbf{x}),S(\mathbf{x}'))]}$$

h is some 'centered' distance function

Jenson-Shannon divergence (JSD)[6] proxy:

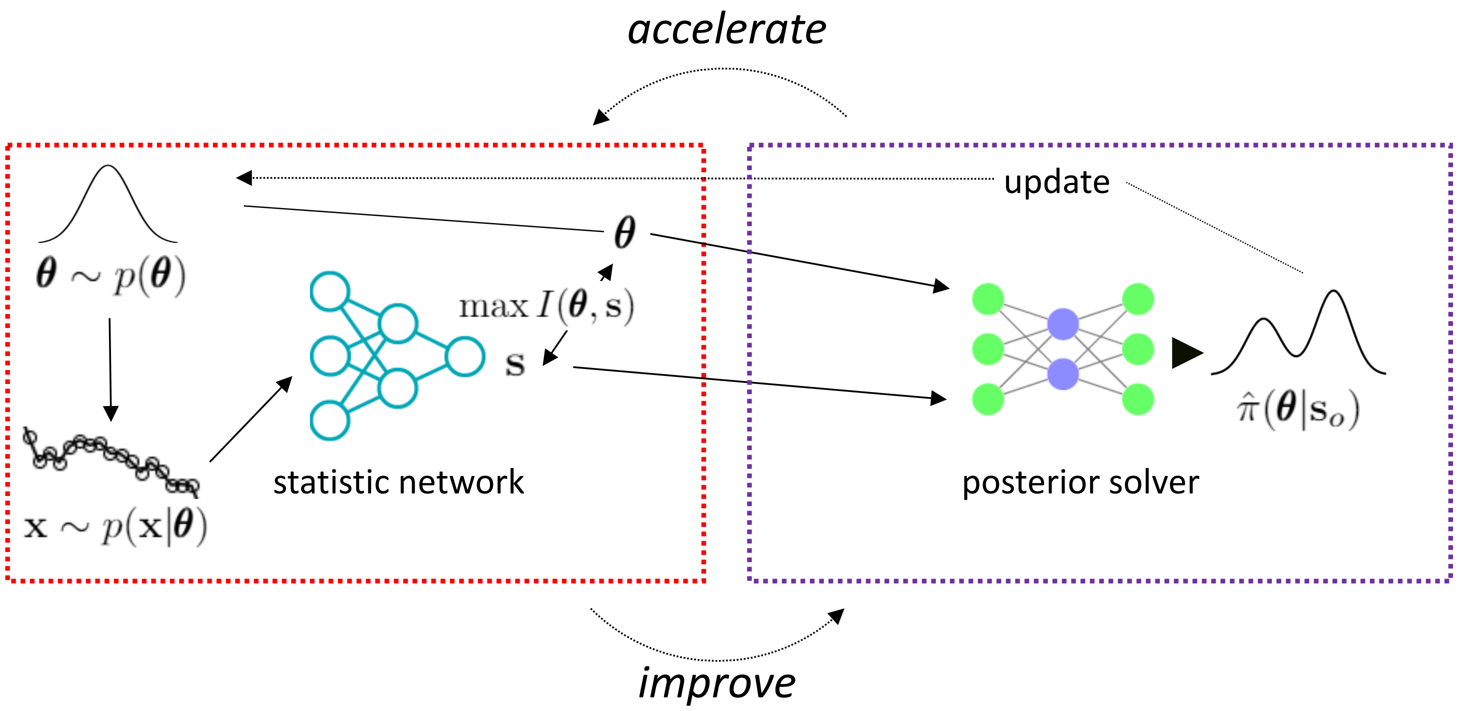
$$\max_{S,T} \; \mathcal{L}(S,T) = \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{x})}\left[-\,\mathrm{sp}\left(-T(\boldsymbol{\theta};S(\mathbf{x}))\right)\right] - \mathbb{E}_{p(\boldsymbol{\theta})p(\mathbf{x})}\left[\mathrm{sp}\left(T(\boldsymbol{\theta};S(\mathbf{x})\right)\right]$$

sp = softplus function

Method

Dynamic sufficient statistics learning

learn the statistics and posterior ***iteratively***



*posterior solver can be any sequential LFI algorithms e.g. SMC-ABC[2], SNL[4]

Related works

Related works

parameter-prediction-as-statistics[8]

$$s = \arg\min_{S:\mathcal{X}\to\mathcal{S}} \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{x})}[\|S(\mathbf{x}) - \boldsymbol{\theta}\|_2^2],$$

we prove it is (generally) *not sufficient*

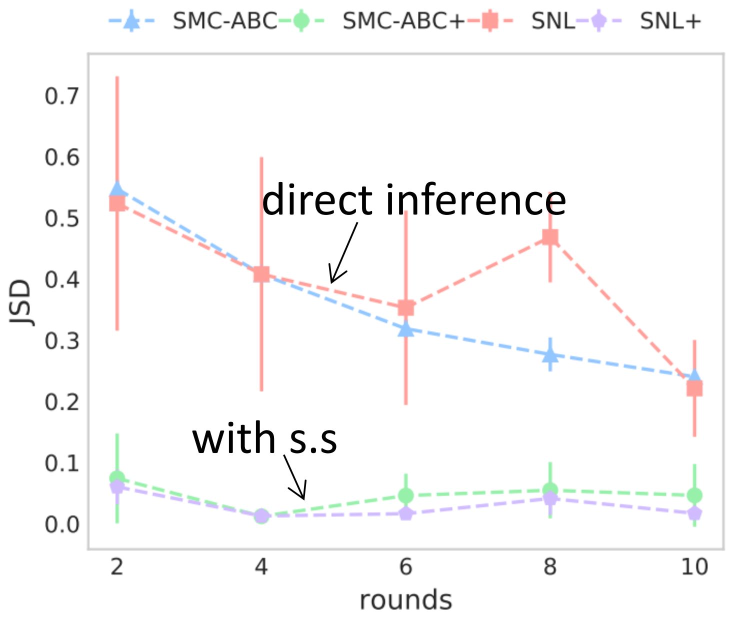
score-as-satistics[9]

$$s = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

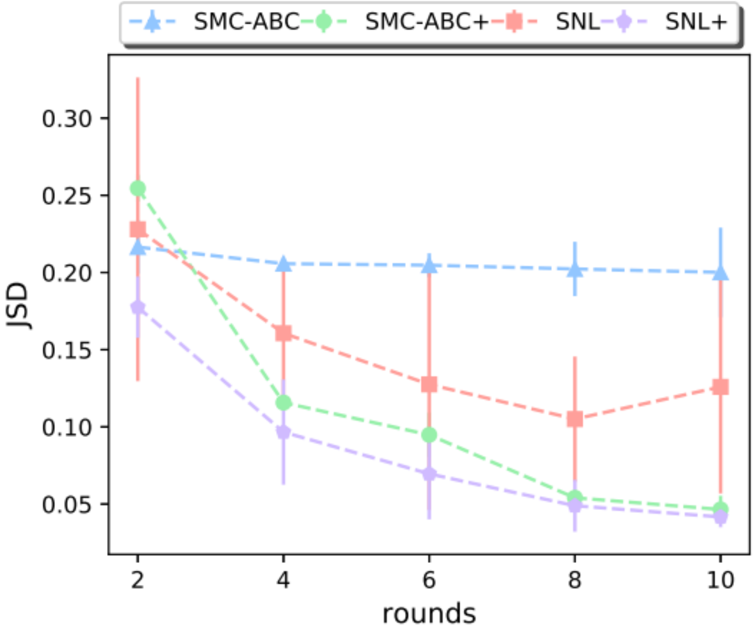
only *locally* sufficient around $\boldsymbol{\theta}^*$

Results

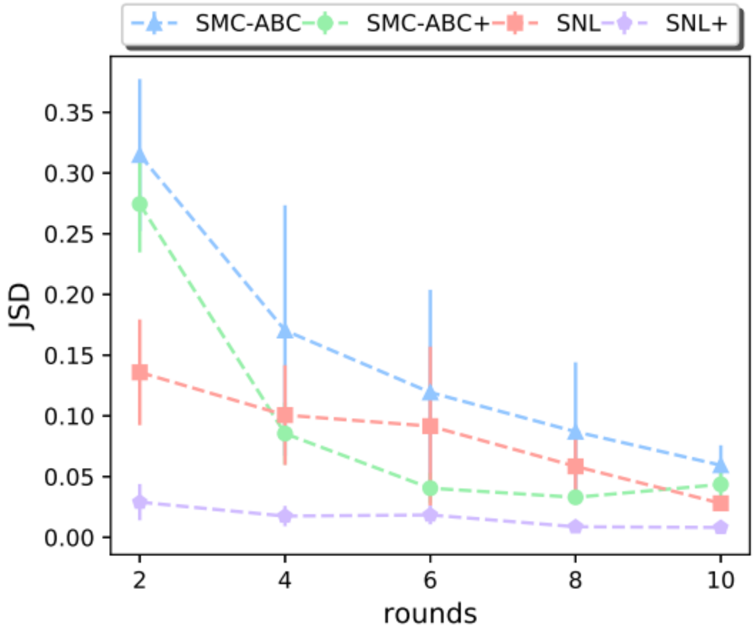
applying to existing LFI algorithms: SMC-ABC[2] ,SNL[4]



Ising model



Gaussian copula



OU Process

x-axis: learning rounds

y-axis: JSD(true P, learned P)

Contribution

- **For likelihood-free inference**

  new method for learning sufficient statistics based on infomax principle

- **For representation learning**

  establish a link between representation learning and Bayesian inference

Reference

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

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

[7]. Wasserstein dependency measure for representation learning, Neurips 19

[8]. Constructing summary statistics for ABC, JRSS B 09

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