Neural Approximate Sufficient Statistics

Michael U. Gutmann

michael.gutmann@ed.ac.uk

School of Informatics, University of Edinburgh

15 June 2021

References

Paper:
Yanzhi Chen*, Dinghuai Zhang*, Michael U. Gutmann, Aaron Courville, Zhanxing Zhu
Neural approximate sufficient statistics for implicit models
ICLR 2021
`https://openreview.net/pdf?id=SRDuJssQud`

Code:
`https://github.com/cyz-ai/neural-approx-ss-lfi`

*did the hard work, equal contribution

Key messages

1. Sufficient statistics are information maximising representations.
2. We can learn approximate sufficient statistics using estimators of mutual information or their proxies.
3. The learned statistics boost the performance of Bayesian inference methods for implicit models.

Contents

Contents

Sufficient statistics

- Consider a parametric statistical model $p(\mathbf{X}|\boldsymbol{\theta})$ for data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.
- A statistic T is a vector-valued function of the data. Basic example is the sample average:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad (1)$$

- Fisher–Neyman factorisation: A statistic T is sufficient for $\boldsymbol{\theta}$ if and only if the joint $p(\mathbf{X}|\boldsymbol{\theta})$ factorises as

$$p(\mathbf{X}|\boldsymbol{\theta}) = u(\mathbf{X})v(T(\mathbf{X}), \boldsymbol{\theta}) \qquad (2)$$

for all $\mathbf{X}$ and $\boldsymbol{\theta}$, where u and v are two non-negative functions.

Example

- Classic example: n iid observations of a Gaussian random variable with mean θ and known variance σ^2.

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right) \tag{3}$$

$$= \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right)}_{u(\mathbf{X})} \underbrace{\exp\left(\frac{2n\theta\bar{x} - n\theta^2}{2\sigma^2}\right)}_{v(T(\mathbf{X}),\theta)}$$

with $T(\mathbf{X}) = \bar{x}$.

- Intuition:
  (1) the model parameters $\boldsymbol{\theta}$ only interact with $\mathbf{X}$ through $T(\mathbf{X})$
  (2) $\boldsymbol{\theta} \perp\!\!\!\perp \mathbf{X} \mid T(\mathbf{X})$

Log likelihood function

- Assume $p(\mathbf{X}|\boldsymbol{\theta}) = u(\mathbf{X})v(T(\mathbf{X}), \boldsymbol{\theta})$
- Given some observed data $\mathbf{X}$, the log likelihood function is

$$\ell(\boldsymbol{\theta}) = \log v(T(\mathbf{X}), \boldsymbol{\theta}) + \text{const} \tag{4}$$

- To infer $\boldsymbol{\theta}$ from $\mathbf{X}$, we do not need to know $\mathbf{X} = (\mathbf{x}_1 \ldots, \mathbf{x}_n)$ but only the value of $T(\mathbf{X})$.
- Gaussian example

$$\ell(\theta) = \frac{n}{2\sigma^2}(2\theta\bar{x} - \theta^2) \tag{5}$$

so that $\hat{\theta}_{\text{MLE}} = \bar{x}$

- Sufficient statistics are important both for MLE and Bayesian inference

$$p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta}|T(\mathbf{X})) \propto v(T(\mathbf{X}), \boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{6}$$

where $\pi(\boldsymbol{\theta})$ is the prior.

Computational benefits of sufficient statistics

- Dimensionality reduction: Both the posterior and the (log)-likelihood only depend on $\mathbf{X}$ via $T(\mathbf{X}) \Rightarrow$ we don't need to store or work with the raw data but can work with $T(\mathbf{X})$, which is often much easier.
- Gaussian example: 1 number vs n numbers
- Many algorithms work by comparing data sets to each other. But comparing $\mathbf{X}$ with $\mathbf{X}'$ is very hard due to high dimensionality. Comparing $T(\mathbf{X})$ with $T(\mathbf{X}')$ is often simpler.

Characterisation in terms of mutual information

- Denote the mutual information between by two random variables $\mathbf{y}_1$ and $\mathbf{y}_2$ by $I(\mathbf{y}_1; \mathbf{y}_2)$,

$$I(\mathbf{y}_1; \mathbf{y}_2) = \mathbb{E}_{\mathbf{y}_1,\mathbf{y}_2} \log \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_1)p(\mathbf{y}_2)} \tag{7}$$

- (Data-processing inequality) For a Markov chain $\boldsymbol{\theta} \to \mathbf{X} \to \mathbf{Z}$,

$$I(\boldsymbol{\theta}; \mathbf{Z}) \leq I(\boldsymbol{\theta}; \mathbf{X}) \tag{8}$$

  We can't gain MI but only lose it by processing data. Inequality also holds for deterministic functions $\mathbf{Z} = g(\mathbf{X})$.

- No information loss if T is a sufficient statistic:

$$T \text{ is a sufficient statistic} \iff I(\boldsymbol{\theta}; T(\mathbf{X})) = I(\boldsymbol{\theta}; \mathbf{X}) \tag{9}$$

Contents

Sufficient statistics are infomax representations

- MI-based characterisation of sufficient statistics T

$$T \text{ is a sufficient statistic} \iff I(\boldsymbol{\theta}; T(\mathbf{X})) = I(\boldsymbol{\theta}; \mathbf{X}) \qquad (10)$$

- Since for deterministic transformations g

$$I(\boldsymbol{\theta}; g(\mathbf{X})) \leq I(\boldsymbol{\theta}; \mathbf{X}) \qquad (11)$$

  we have a variational characterisation of sufficient statistics

$$T \text{ is a sufficient statistic} \iff I(\boldsymbol{\theta}; T(\mathbf{X})) = \max_{g \in \mathcal{G}} I(\boldsymbol{\theta}; g(\mathbf{X})) \qquad (12)$$

- Choosing a function family $\mathcal{G}$ can introduce an approximation. We work with neural networks with a fixed number of outputs ($2 \dim(\boldsymbol{\theta})$).

Combining the idea with MI estimators

- Estimating MI is hard.
- We are interested in finding an $\text{argmax}_g I(\boldsymbol{\theta}; g(\mathbf{X}))$ rather than knowing the precise value of the MI.
- Broadens the set of applicable MI estimators to include surrogate quantities.
  - MI as KL divergence between joint and marginals $\rightarrow$ Use the more robust Jensen-Shannon divergence (JSD) instead of the KL divergence
  - MI as a nonlinear dependency measure $\rightarrow$ Use the ratio-free distance correlation (Székely and Rizzo, 2009, 2014)

Székely and Rizzo, Brownian distance covariance, The Annals of Applied Statistics, 2009

Székely and Rizzo, Partial distance correlation with methods for dissimilarities, The Annals of Statistics, 2014

Learning sufficient statistics with the JSD

- A density-free variational formulation of the JSD between $p(\boldsymbol{\theta}, \mathbf{X})$ and $p(\boldsymbol{\theta})p(\mathbf{X})$ is

$$\sup_F \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{X})}\left[-\operatorname{sp}(-F(\boldsymbol{\theta},\mathbf{X}))\right] - \mathbb{E}_{p(\boldsymbol{\theta})p(\mathbf{X})}\left[\operatorname{sp}(F(\boldsymbol{\theta},\mathbf{X}))\right] \quad (13)$$

  where $\operatorname{sp}(t) = \log(1 + \exp(t))$ is the softplus function.
- Objective for learning sufficient statistics:

$$\sup_{S,F} \mathbb{E}_{p(\boldsymbol{\theta},\mathbf{X})}\left[-\operatorname{sp}\left(-F(\boldsymbol{\theta},S(\mathbf{X}))\right)\right] - \mathbb{E}_{p(\boldsymbol{\theta})p(\mathbf{X})}\left[\operatorname{sp}\left(F(\boldsymbol{\theta},S(\mathbf{X})\right)\right] \quad (14)$$

- Same as learning the ratio $p(\boldsymbol{\theta}, \mathbf{X})/p(\boldsymbol{\theta})p(\mathbf{X}) = p(\boldsymbol{\theta}|\mathbf{X})$ by logistic regression (classification) with a particular constraint on the processing of $\mathbf{X}$. (see"LFI by ratio estimation" by Thomas et al, 2016, 2021)

Learning sufficient statistics via distance correlation

(Székely and Rizzo, 2014)

- The distance correlation between two random variables is a multivariate dependence coefficient defined as

$$R^2(\boldsymbol{\theta}, \mathbf{X}) = \frac{\mathbb{E}[A_{\boldsymbol{\theta}} A_{\mathbf{X}}]}{\sqrt{\mathbb{E}[A_{\boldsymbol{\theta}}^2]\mathbb{E}[A_{\mathbf{X}}^2]}} \tag{15}$$

  where $\mathbf{A}_{\mathbf{X}}$ is a double-centred (random) distance function

$$\mathbf{A}_{\mathbf{X}} = \|\mathbf{X}-\mathbf{X}'\| - \mathbb{E}_{\mathbf{X}}[\|\mathbf{X}-\mathbf{X}'\|] - \mathbb{E}_{\mathbf{X}'}[\|\mathbf{X}'-\mathbf{X}\|] + \mathbb{E}_{\mathbf{X}'}\mathbb{E}_{\mathbf{X}}[\|\mathbf{X}-\mathbf{X}'\|]$$

  (equivalent definition for $A_{\boldsymbol{\theta}}$)
- Expectation in the numerator is taken with respect to $(\mathbf{X}, \boldsymbol{\theta})$ and the independent and identically distributed tuple $(\mathbf{X}', \boldsymbol{\theta}')$.
  (The expectations in the denominator are taken with respect to the corresponding marginals.)

Learning sufficient statistics via distance correlation

- There are equivalent definitions in terms of characteristic functions and the so-called Brownian distance covariance

  (Székely and Rizzo, 2009, 2014)

- Key properties:
  - $0 \leq R(\boldsymbol{\theta}, \mathbf{X}) \leq 1$
  - $R(\boldsymbol{\theta}, \mathbf{X}) = 0 \iff \boldsymbol{\theta} \perp\!\!\!\perp \mathbf{X}$
  - $R(\boldsymbol{\theta}, \mathbf{X}) = 1$ means $\boldsymbol{\theta}$ and $\mathbf{X}$ are a linear transformation of each other.
- Objective for learning sufficient statistics:

$$\max_{S} R^2(\boldsymbol{\theta}, S(\mathbf{X})) \tag{16}$$

Note: we only need to train one network and not two as in the JSD (and other variational MI estimators), which makes this approach faster.

Contents

Goal

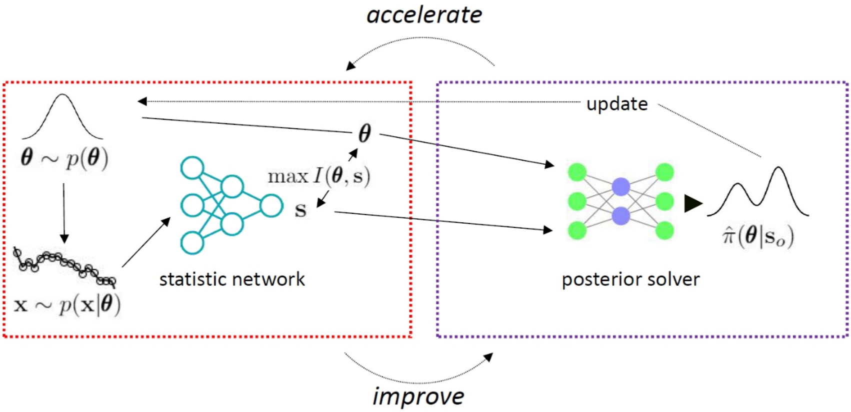
- Goal: approximate Bayesian parameter inference for implicit models
- Implicit models: models where sampling is possible but evaluating the likelihood function is computationally infeasible

$$\mathbf{X} \sim p(\mathbf{X}|\boldsymbol{\theta}) \tag{17}$$

- Approach: learn approximate sufficient statistics and aim at $p(\boldsymbol{\theta}|T(\mathbf{X}))$ instead of $p(\boldsymbol{\theta}|\mathbf{X})$ to increase efficiency.
- Focus on sequential inference methods:
  - (variant of) sequential approximate Bayesian computation (Beaumont, 2009)
  - sequential neural likelihood (Papamakarios et al., 2019)

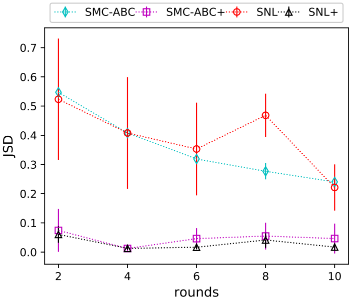
Overview of the sequential approach

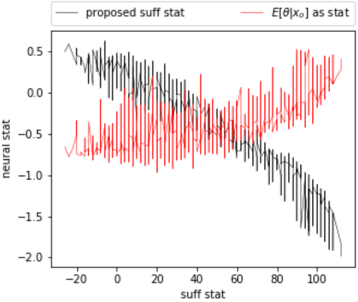
We jointly learn the statistics and the posterior in multiple rounds.
(see paper for details)



Example results: Ising model (using JSD)

- 64-dimensional Ising model, θ: coupling strength (prior: $\mathcal{U}(0, 1.5)$)
- Sufficient statistics are known. Reference posterior obtained by (expensive) rejection sampling.
- The learned statistics (algorithms with a $+$) improve the inference.
- Posterior mean as statistics is sub-optimal.





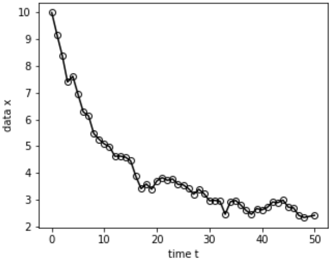
Example results: Ornstein-Uhlenbeck process (using JSD)

- Stochastic differential equation simulated with the Euler-Maruyama method

$$x_{t+1} = x_t + \Delta x_t \tag{18}$$
$$\Delta x_t = \theta_1(\exp(\theta_2) - x_t)\Delta t + 0.5\epsilon, \quad \epsilon \sim \mathcal{N}(\epsilon; 0, \Delta t) \tag{19}$$

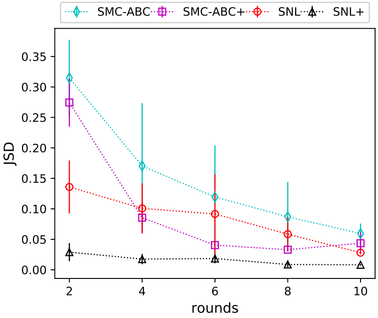
where $\Delta t = 0.2$ and $x_0 = 10$.

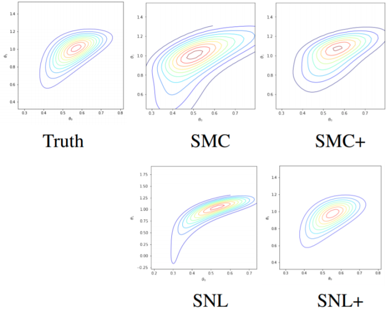
- Data: $x_1, \ldots, x_{50}$.
- Unknowns: θ_1 and θ_2 with priors $\mathcal{U}(0, 1)$ and $\mathcal{U}(-2.0, 2.0)$, respectively.



Example results: Ornstein-Uhlenbeck process (using JSD)

- Learning approximate sufficient statistics improves the inference.
- Learned statistics give better calibrated posteriors.

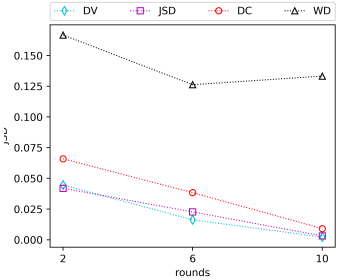




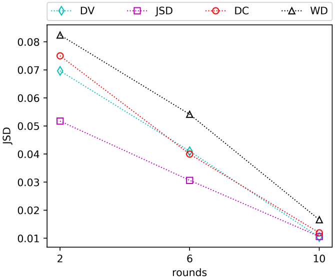
Other MI proxies

(Results for SNL+)

- We can use other MI proxies than JSD and DC. Results for Donsker-Varadhan (DV) and Wasserstein distance (WD).
- JSD performs here best but DC is about 15 times faster than the other methods.



Ising model



OU process

Conclusions

- Two characterisations of sufficient statistics:
  - Fisher-Neyman factorisation
  - Characterisation in terms of mutual information (MI)
- Variational characterisation: sufficient statistics are information maximising representations.
- Learn (approximate) sufficient statistics using (proxy) MI estimators.
- We used the learned statistics to boost the performance of Bayesian inference with implicit models.
- More results in the paper *Neural approximate sufficient statistics for implicit models* `https://openreview.net/pdf?id=SRDuJssQud`