Neural Approximate Sufficient Statistics

Michael U. Gutmann

michael.gutmann@ed.ac.uk

School of Informatics, University of Edinburgh

28 October 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

Background on sufficient statistics

Proposed method to learn approximate sufficient statistics

Application to Bayesian inference with implicit models

Contents

Sufficient statistics

- Consider a parametric statistical model $p(\mathbf{x}|\boldsymbol{\theta})$ for data $\mathbf{x}$.

Sufficient statistics

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Sufficient statistics

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

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

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

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

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:

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 $T(\mathbf{x})$

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

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

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

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 $\mathbf{x}$ but only the value of $T(\mathbf{x})$.

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

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 $\mathbf{x}$ 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.

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

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 $I(\mathbf{y}_1; \mathbf{y}_2)$,

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

Characterisation in terms of mutual information

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

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

- (Data-processing inequality) For a Markov chain $\boldsymbol{\theta} \rightarrow \mathbf{x} \rightarrow \mathbf{z}$,

$$I(\boldsymbol{\theta}; \mathbf{z}) \leq I(\boldsymbol{\theta}; \mathbf{x}) \tag{8}$$

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

Characterisation in terms of mutual information

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

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

- (Data-processing inequality) For a Markov chain $\boldsymbol{\theta} \rightarrow \mathbf{x} \rightarrow \mathbf{z}$,

$$I(\boldsymbol{\theta}; \mathbf{z}) \leq I(\boldsymbol{\theta}; \mathbf{x}) \tag{8}$$

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

- No information loss for sufficient statistics:

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

Contents

Sufficient statistics are infomax representations

- MI-based characterisation of sufficient statistics T

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

Sufficient statistics are infomax representations

- MI-based characterisation of sufficient statistics T

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

- Since for deterministic transformations g

$$I(\boldsymbol{\theta}; g(\mathbf{x})) \leq I(\boldsymbol{\theta}; \mathbf{x}) \tag{11}$$

we have a variational characterisation of sufficient statistics

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

Sufficient statistics are infomax representations

- MI-based characterisation of sufficient statistics T

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

- Since for deterministic transformations g

$$I(\boldsymbol{\theta}; g(\mathbf{x})) \leq I(\boldsymbol{\theta}; \mathbf{x}) \tag{11}$$

we have a variational characterisation of sufficient statistics

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

- 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.

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.

Combining the idea with MI estimators

- Estimating MI is hard.
- We are interested in finding an $\operatorname{argmax}_g I(\boldsymbol{\theta}; g(\mathbf{x}))$ rather than knowing the precise value of the MI.
- Broadens the set of applicable MI estimators to include surrogate quantities.

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

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)

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.

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:

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

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:

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

- 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; Hermans et al, 2020)

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 $A_{\mathbf{x}}$ is a double-centred (random) distance function

$$A_{\mathbf{x}} = \|\mathbf{x} - \mathbf{x}'\| - \mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - \mathbf{x}'\|] - \mathbb{E}_{\mathbf{x}'}[\|\mathbf{x}' - \mathbf{x}\|] + \mathbb{E}_{\mathbf{x}'}\mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - \mathbf{x}'\|]$$

(equivalent definition for $A_{\boldsymbol{\theta}}$)

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 $A_{\mathbf{x}}$ is a double-centred (random) distance function

$$A_{\mathbf{x}} = \|\mathbf{x} - \mathbf{x}'\| - \mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - \mathbf{x}'\|] - \mathbb{E}_{\mathbf{x}'}[\|\mathbf{x}' - \mathbf{x}\|] + \mathbb{E}_{\mathbf{x}'}\mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - \mathbf{x}'\|]$$

(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).

  Links to maximum mean discrepancy (MMD, Sejdinovic et al, 2013)

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).

  Links to maximum mean discrepancy (MMD, Sejdinovic et al, 2013)
- Key properties:

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).

  Links to maximum mean discrepancy (MMD, Sejdinovic et al, 2013)
- Key properties:
  - $0 \leq R(\boldsymbol{\theta}, \mathbf{x}) \leq 1$

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).

  Links to maximum mean discrepancy (MMD, Sejdinovic et al, 2013)
- Key properties:
  - $0 \leq R(\boldsymbol{\theta}, \mathbf{x}) \leq 1$
  - $R(\boldsymbol{\theta}, \mathbf{x}) = 0 \iff \boldsymbol{\theta} \perp\!\!\!\perp \mathbf{x}$

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).

  Links to maximum mean discrepancy (MMD, Sejdinovic et al, 2013)
- 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.

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).

  Links to maximum mean discrepancy (MMD, Sejdinovic et al, 2013)
- 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:

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

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

Goal

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

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

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 $S(\cdot)$ and aim at $p(\boldsymbol{\theta}|\mathbf{s})$, with $\mathbf{s} = S(\mathbf{x})$, instead of $p(\boldsymbol{\theta}|\mathbf{x})$ to increase efficiency.

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 $S(\cdot)$ and aim at $p(\boldsymbol{\theta}|\mathbf{s})$, with $\mathbf{s} = S(\mathbf{x})$, instead of $p(\boldsymbol{\theta}|\mathbf{x})$ to increase efficiency.
- Focus on sequential inference methods:

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 $S(\cdot)$ and aim at $p(\boldsymbol{\theta}|\mathbf{s})$, with $\mathbf{s} = S(\mathbf{x})$, instead of $p(\boldsymbol{\theta}|\mathbf{x})$ to increase efficiency.
- Focus on sequential inference methods:
  - (variant of) sequential approximate Bayesian computation

    (SMC-ABC, Beaumont, 2009)

Goal

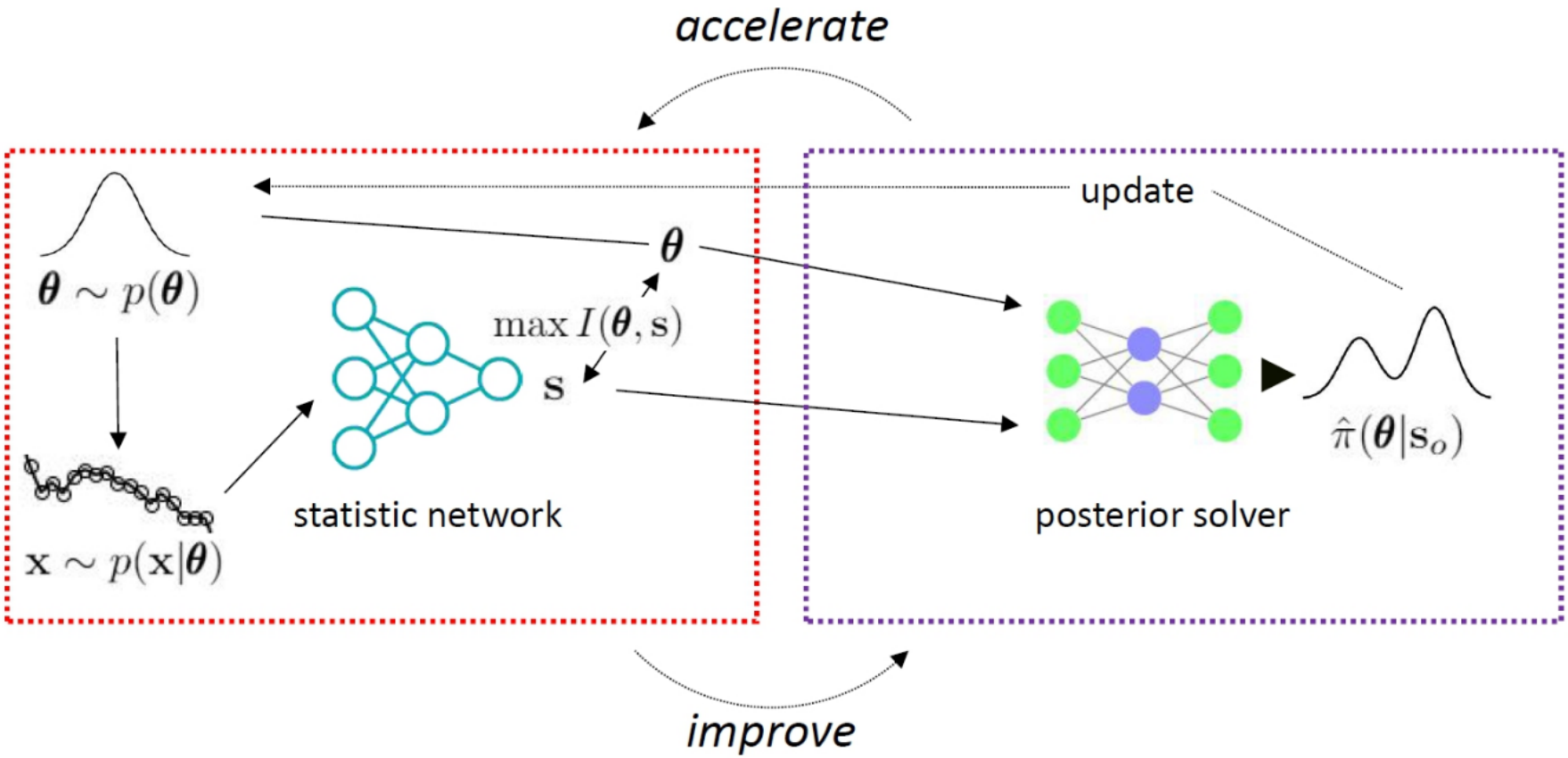
- Goal: approximate Bayesian parameter inference for implicit models
- Implicit models: models where sampling is possible but evaluating the likelihood function is computationally infeasible

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

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

Overview of the approach

We jointly learn the statistics and the posterior in multiple rounds.



Variant of SMC-ABC with neural suff stats (SMC-ABC+)

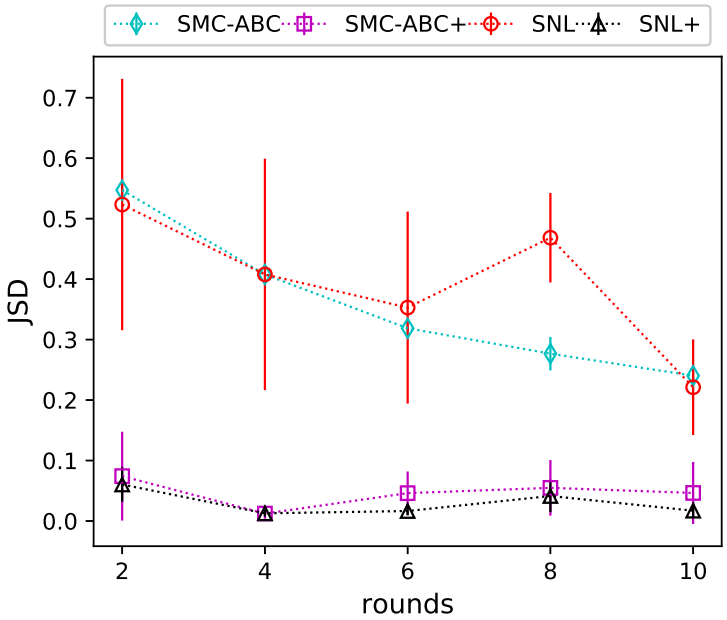
Input: prior $\pi(\boldsymbol{\theta})$, observed data $\mathbf{x}_o$
Output: estimated posterior $\hat{\pi}(\boldsymbol{\theta}|\mathbf{x}_o)$
Initialisation: $\mathcal{D} = \varnothing, p_1(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$
for j in 1 to r **do**
  repeat
    sample $\boldsymbol{\theta}_i \sim p_j(\boldsymbol{\theta})$; *sample from proposal distr*
    simulate $\mathbf{x}_i \sim p(\mathbf{x}|\boldsymbol{\theta}_i)$; *sample data from the model*
  until n samples
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{\boldsymbol{\theta}_i, \mathbf{x}_i\}_{i=1}^n$
  learn approx suff stats $S(\cdot)$ using $\mathcal{D}$; $\longleftarrow$ *the main modification*
  compute $\mathbf{s}_o = S(\mathbf{x}_o)$ and $\mathbf{s}_i = S(\mathbf{x}_i)$ for $(\boldsymbol{\theta}_i, \mathbf{x}_i) \in \mathcal{D}$;
  sort $\mathcal{D}$ according to $\|\mathbf{s}_i - \mathbf{s}_o\|$;
  fit $p(\boldsymbol{\theta}|\mathbf{s}_o)$ with the top m $\boldsymbol{\theta}$s in $\mathcal{D}$; *e.g. MoG or copula model*
  $q_j(\boldsymbol{\theta}|\mathbf{s}_o) \propto p(\boldsymbol{\theta}|\mathbf{s}_o)\pi(\boldsymbol{\theta}) / \sum_l^j p_l(\boldsymbol{\theta})$; $\pi / \sum_l^j p_l$: *importance weights*
  $p_{j+1}(\boldsymbol{\theta}) \leftarrow q_j(\boldsymbol{\theta}|\mathbf{s}_o)$; *update proposal distribution*
end for
return $\hat{\pi}(\boldsymbol{\theta}|\mathbf{x}_o) = q_r(\boldsymbol{\theta}|\mathbf{s}_o)$ *alternatively: return the weighted samples*

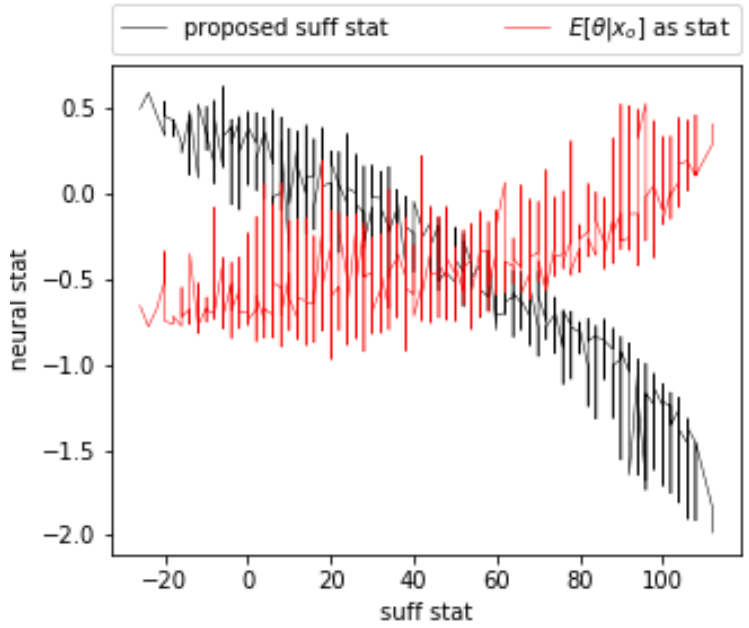
SNL with neural suff stats (SNL+)

Input: prior $\pi(\boldsymbol{\theta})$, observed data $\mathbf{x}_o$
Output: estimated posterior $\hat{\pi}(\boldsymbol{\theta}|\mathbf{x}_o)$
Initialisation: $\mathcal{D} = \varnothing, p_1(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$
for j in 1 to r **do**
  repeat
    sample $\boldsymbol{\theta}_i \sim p_j(\boldsymbol{\theta})$; *sample from proposal distr*
    simulate $\mathbf{x}_i \sim p(\mathbf{x}|\boldsymbol{\theta}_i)$; *sample data from the model*
  until n samples
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{\boldsymbol{\theta}_i, \mathbf{x}_i\}_{i=1}^n$
  learn approx suff stats $S(\cdot)$ using $\mathcal{D}$; *⟵ the modification*
  compute $\mathbf{s}_o = S(\mathbf{x}_o)$ and $\mathbf{s}_i = S(\mathbf{x}_i)$ for $(\boldsymbol{\theta}_i, \mathbf{x}_i) \in \mathcal{D}$;
  fit $q(\mathbf{s}|\boldsymbol{\theta})$ as in original SNL; *neural density estimator, e.g. MAF*
  $q_j(\boldsymbol{\theta}|\mathbf{s}_o) \propto \pi(\boldsymbol{\theta}) \cdot q(\mathbf{s}_o|\boldsymbol{\theta})$; *prior times estimated likelihood*
  $p_{j+1}(\boldsymbol{\theta}) \leftarrow q_j(\boldsymbol{\theta}|\mathbf{s}_o)$; *update proposal distribution*
end for
return $\hat{\pi}(\boldsymbol{\theta}|\mathbf{x}_o) = q_r(\boldsymbol{\theta}|\mathbf{s}_o)$

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)

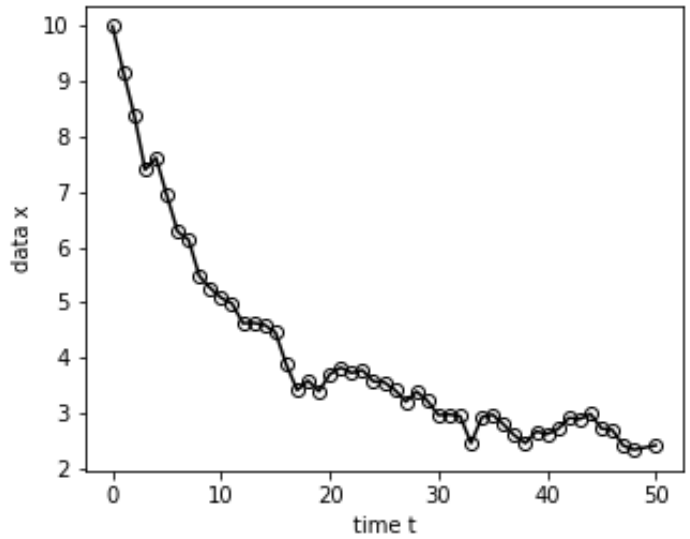
- Stochastic differential equation simulated with the Euler-Maruyama method

$$x_{t+1} = x_t + \Delta x_t \tag{18}$$

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

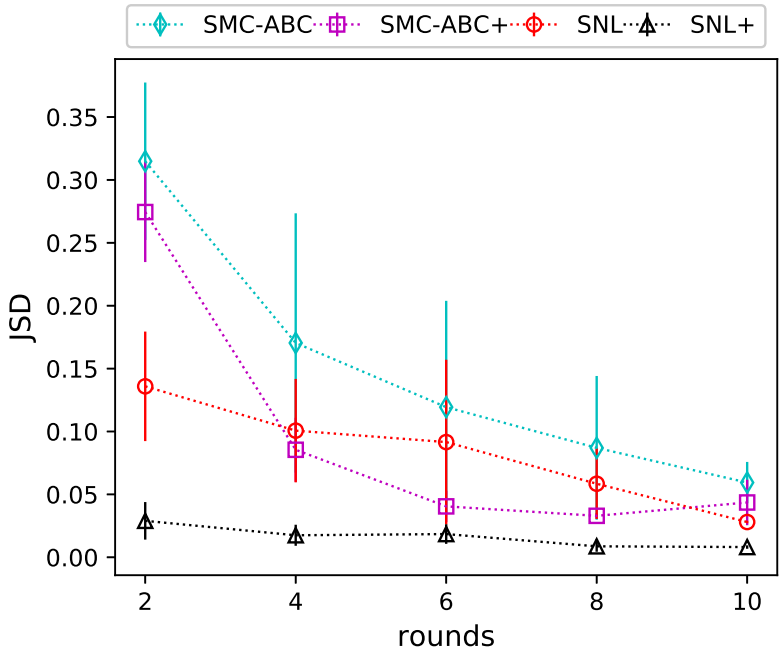
where $\Delta t = 0.2$ and $x_0 = 10$.

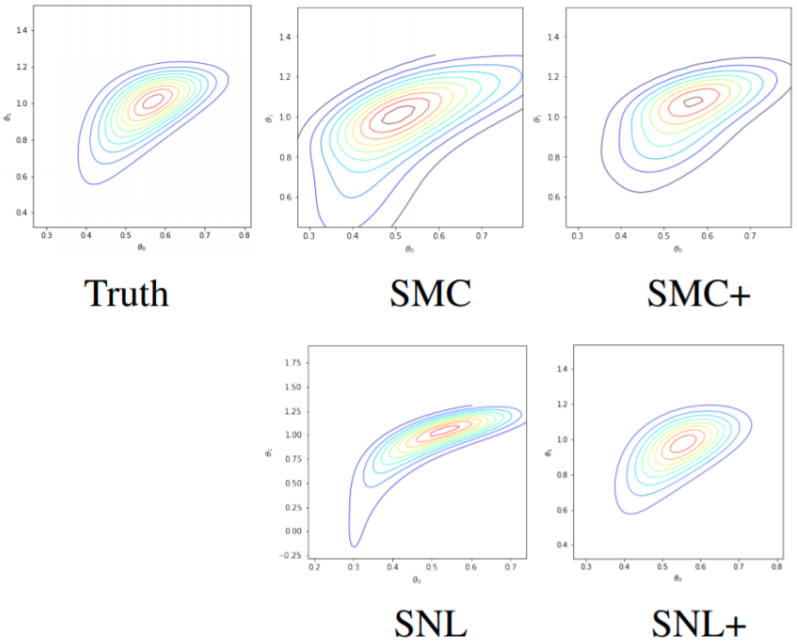
- Data: $x_1, \ldots, x_{50}$.
- Unknowns: θ_1 and θ_2 with priors $\mathcal{U}(0, 1)$ and $\mathcal{U}(-2.0, 2.0)$, respectively.



Example results: Ornstein-Uhlenbeck process (using JSD)

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

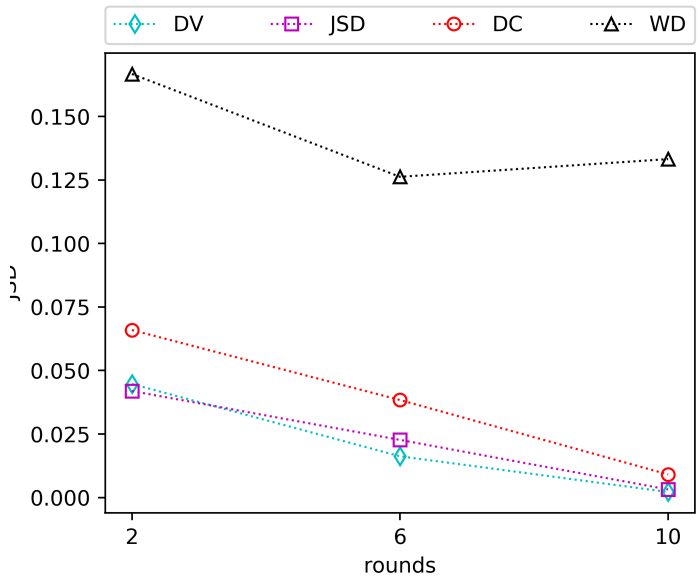




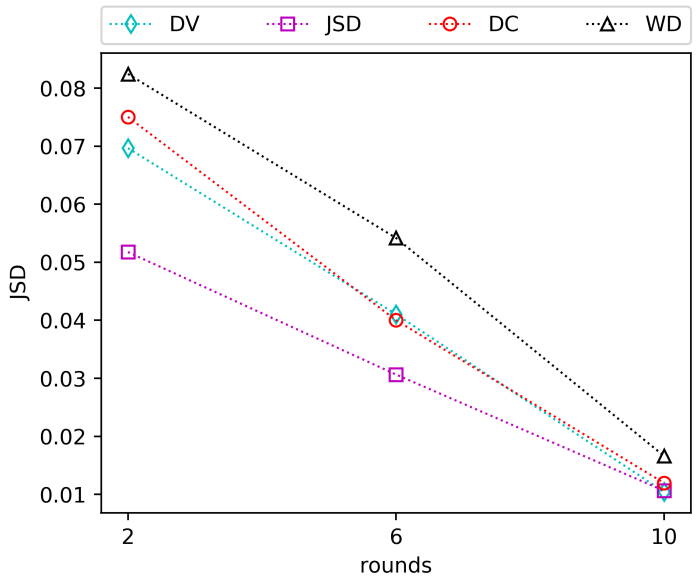
Other MI proxies

(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.
- "Approach plays well with others": can be used to enhance existing algorithms
- More results and comparisons in the paper
  Neural approximate sufficient statistics for implicit models
  `https://openreview.net/pdf?id=SRDuJssQud`