Statistical applications of contrastive learning

Michael U. Gutmann

michael.gutmann@ed.ac.uk

School of Informatics, University of Edinburgh

15th October 2021

Acknowledgements

Research presented here is the result of joint work with Aapo Hyvärinen, Jukka Corander, Jun-ichiro Hirayama, Chris Drovandi, and my PhD students Ben Rhodes and Steven Kleinegesse.

Main messages

1. The likelihood function is computationally intractable for energy-based and simulator-based models.
2. Contrastive learning is an intuitive and computationally feasible alternative to likelihood-based learning.
3. We used it in a broad range of tasks: (1) parameter estimation, (2) Bayesian inference, and (3) Bayesian experimental design.

Contents

Computational difficulties in likelihood-based learning

Contrastive learning

Applications in statistical inference and experimental design

Contents

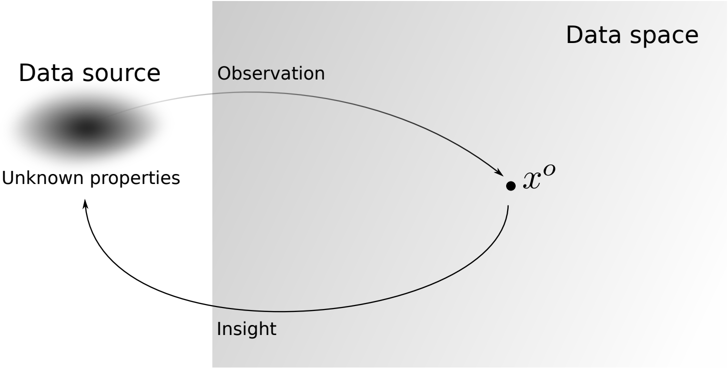
Computational difficulties in likelihood-based learning

Contrastive learning

Applications in statistical inference and experimental design

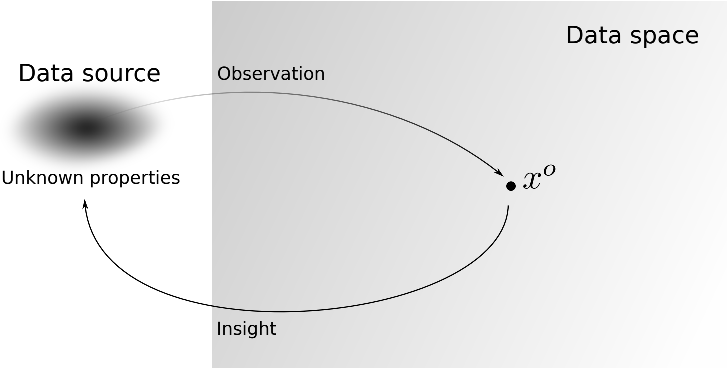
Overall goal

- Goal: Understanding properties of some data source
- Enables predictions, decision making under uncertainty, ...



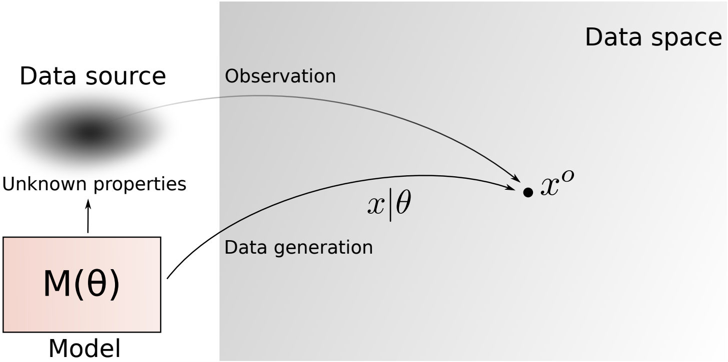
Two fundamental tasks

- Inference task : Given $\mathbf{x}^o$, what can we robustly say about the properties of the source?
- Experimental design task : How to obtain a $\mathbf{x}^o$ that is maximally useful for learning about the properties?



The likelihood function $L(\boldsymbol{\theta})$

- Probability that the model generates data like $\mathbf{x}^o$ when using parameter value $\boldsymbol{\theta}$
- Classically, the main workhorse to solve the inference and design task.



The likelihood function $L(\boldsymbol{\theta})$

- For models expressed as a family of pdfs $\{p(\mathbf{x}|\boldsymbol{\theta})\}$ indexed by $\boldsymbol{\theta}$: $L(\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ where $\mathbf{x}$ is fixed.
- Inference:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{x}|\boldsymbol{\theta}) \qquad \text{or} \qquad p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} p(\boldsymbol{\theta}) \tag{1}$$

  with $\mathbf{x}$ fixed to $\mathbf{x}^o$.
- Experimental design via mutual information: expand model to include (deterministic) design variable $\mathbf{d}$, $\{p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})\}$

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\operatorname{argmax}}\, \mathrm{MI}_{\mathbf{d}}(\mathbf{x}, \boldsymbol{\theta}) \tag{2}$$

$$\mathrm{MI}_{\mathbf{d}}(\mathbf{x}, \boldsymbol{\theta}) = \mathrm{KL}\left(p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{d}) || p(\boldsymbol{\theta}|\mathbf{d}) p(\mathbf{x}|\mathbf{d})\right) \tag{3}$$

$$= \mathbb{E}_{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})}{p(\mathbf{x}|\mathbf{d})}\right] \tag{4}$$

Energy and simulator-based models

- Not all models are specified as family of pdfs.
- Two important classes considered here
  1. Energy-based (unnormalised) models
  2. Simulator-based (implicit) models
- The models are rather different, common point:

  Multiple integrals needed to be solved to represent the models in terms of pdfs.
- Solving the integrals exactly is computationally impossible (curse of dimensionality)

  $\Rightarrow$ No model pdfs

  $\Rightarrow$ No standard likelihood-based inference or experimental design

Energy-based models

- Widely used:
  - computer vision and modelling of images
  - natural language processing and machine translation
  - modelling social or biological networks
  - ...
- Specified via an energy function $E(\mathbf{x};\boldsymbol{\theta})$ so that $\phi(\mathbf{x}|\boldsymbol{\theta}) = \exp(-E(\mathbf{x};\boldsymbol{\theta})) \propto p(\mathbf{x}|\boldsymbol{\theta})$,

$$\int \cdots \int \phi(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = Z(\boldsymbol{\theta}) \neq 1 \qquad p(\mathbf{x}|\boldsymbol{\theta}) = \frac{\phi(\mathbf{x}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

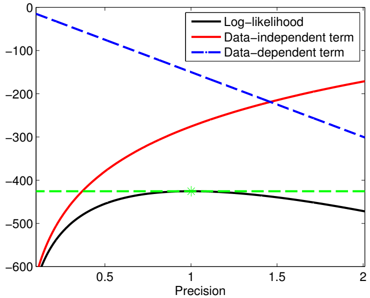
- Advantage: Specifying an energy $E(\mathbf{x};\boldsymbol{\theta})$ is often easier than specifying normalised models
- Disadvantage: Integral defining the partition function $Z(\boldsymbol{\theta})$ can generally not be computed. Model pdf and likelihood function are intractable.

We cannot just ignore the partition function

- Consider $p(x;\theta) = \frac{\phi(x;\theta)}{Z(\theta)} = \frac{\exp\left(-\theta \frac{x^2}{2}\right)}{\sqrt{2\pi/\theta}}$
- Log-likelihood function for precision $\theta \geq 0$

$$\ell(\theta) = -n \log \sqrt{\frac{2\pi}{\theta}} - \theta \sum_{i=1}^{n} \frac{x_i^2}{2} \tag{5}$$

- Data-dependent (blue) and independent part (red) balance each other.
- Ignoring $Z(\boldsymbol{\theta})$ leads to meaningless estimates.



Simulator-based models

- Widely used:
  - computer models/simulators in the natural sciences
  - evolutionary biology to model evolution
  - epidemiology to model the spread of an infectious disease
  - ...
- Specified via a measurable function g, typically not known in closed form but implemented as a computer programme.

$$\mathbf{x} = g(\boldsymbol{\theta}, \boldsymbol{\omega}), \quad \boldsymbol{\omega} \sim p(\boldsymbol{\omega}) \tag{6}$$

  Maps parameters $\boldsymbol{\theta}$ and "noise" $\boldsymbol{\omega}$ to data $\mathbf{x}$
- Advantage: connects statistics to the natural sciences
- Disadvantage: Model pdf and lik function are intractable.

$$\Pr(\mathbf{x} \in \mathcal{A} | \boldsymbol{\theta}) = \Pr(\{\omega : g(\boldsymbol{\theta}, \boldsymbol{\omega}) \in \mathcal{A}\})$$

Contents

Computational difficulties in likelihood-based learning

Contrastive learning

Applications in statistical inference and experimental design

Basic idea

- The basic idea in contrastive learning is to learn the difference between the data of interest and some reference data.

Basic idea

- The basic idea in contrastive learning is to learn the difference between the data of interest and some reference data.
- Properties of the reference are typically known or not of interest; by learning the difference we focus the (computational) resources on learning what matters.

Basic idea

- The basic idea in contrastive learning is to learn the difference between the data of interest and some reference data.
- Properties of the reference are typically known or not of interest; by learning the difference we focus the (computational) resources on learning what matters.
- As straightforward as

$$\underbrace{b}_{\text{reference}} + \underbrace{a - b}_{\text{difference}} \Rightarrow \underbrace{a}_{\text{interest}} \tag{7}$$

Basic idea

- The basic idea in contrastive learning is to learn the difference between the data of interest and some reference data.
- Properties of the reference are typically known or not of interest; by learning the difference we focus the (computational) resources on learning what matters.
- As straightforward as

$$\underbrace{b}_{\text{reference}} + \underbrace{a - b}_{\text{difference}} \Rightarrow \underbrace{a}_{\text{interest}} \tag{7}$$

- Link to (log) ratio estimation (see e.g. Sugiyama et al's textbook)

$$\underbrace{\log p_b}_{\text{reference}} + \underbrace{\log p_a - \log p_b}_{\text{difference}} \Rightarrow \underbrace{\log p_a}_{\text{interest}} \tag{8}$$

Basic idea

- The basic idea in contrastive learning is to learn the difference between the data of interest and some reference data.
- Properties of the reference are typically known or not of interest; by learning the difference we focus the (computational) resources on learning what matters.
- As straightforward as

$$\underbrace{b}_{\text{reference}} + \underbrace{a - b}_{\text{difference}} \Rightarrow \underbrace{a}_{\text{interest}} \tag{7}$$

- Link to (log) ratio estimation (see e.g. Sugiyama et al's textbook)

$$\underbrace{\log p_b}_{\text{reference}} + \underbrace{\log p_a - \log p_b}_{\text{difference}} \Rightarrow \underbrace{\log p_a}_{\text{interest}} \tag{8}$$

- Link to Bayes' rule

$$\underbrace{\log p(\boldsymbol{\theta})}_{\text{reference}} + \underbrace{\log p(\mathbf{x}|\boldsymbol{\theta}) - \log p(\mathbf{x})}_{\text{difference}} \Rightarrow \underbrace{\log p(\boldsymbol{\theta}|\mathbf{x})}_{\text{interest}} \tag{9}$$

Logistic loss

- Link to classification: learning differences between data sets can be seen as a classification problem.

Logistic loss

- Link to classification: learning differences between data sets can be seen as a classification problem.
- Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the data of interest, $\mathbf{x}_i \sim p$ (iid), and $\{\mathbf{y}_1, \ldots \mathbf{y}_m\}$ be reference data, $\mathbf{y}_i \sim q$ (iid).

Logistic loss

- Link to classification: learning differences between data sets can be seen as a classification problem.
- Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the data of interest, $\mathbf{x}_i \sim p$ (iid), and $\{\mathbf{y}_1, \ldots \mathbf{y}_m\}$ be reference data, $\mathbf{y}_i \sim q$ (iid).
- Label the data: $(\mathbf{x}_i, 1)$, $(\mathbf{y}_i, 0)$ and minimise the (rescaled) logistic loss $J(h)$

$$\begin{aligned} J(h) =& \frac{1}{n} \sum_{i=1}^{n} \log\left[1 + \nu \exp(-h(\mathbf{x}_i))\right] + \\ & \frac{\nu}{m} \sum_{i=1}^{m} \log\left[1 + \frac{1}{\nu} \exp(h(\mathbf{y}_i))\right] \end{aligned} \tag{10}$$

where $\nu = m/n$

Logistic loss

- Link to classification: learning differences between data sets can be seen as a classification problem.
- Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the data of interest, $\mathbf{x}_i \sim p$ (iid), and $\{\mathbf{y}_1, \ldots \mathbf{y}_m\}$ be reference data, $\mathbf{y}_i \sim q$ (iid).
- Label the data: $(\mathbf{x}_i, 1)$, $(\mathbf{y}_i, 0)$ and minimise the (rescaled) logistic loss $J(h)$

$$J(h) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \nu \exp(-h(\mathbf{x}_i))\right] + \frac{\nu}{m}\sum_{i=1}^{m} \log\left[1 + \frac{1}{\nu}\exp(h(\mathbf{y}_i))\right] \tag{10}$$

  where $\nu = m/n$
- For large sample sizes n and m (and fixed ratio ν), the optimal h is

$$h^* = \log p - \log q \tag{11}$$

Logistic loss

Two key points:

1. The optimisation is done without any constraints (e.g. normalisation). The optimal h is automagically the ratio between two *densities*

$$h^* = \log p - \log q \tag{12}$$

2. We only need samples from p and q; we do not need their densities or model of them (but we do need an appropriate model for the ratio)

Logistic loss

- For large sample sizes n and m, $J(h) \to \bar{J}(h)$ and the corresponding minimal loss is

$$\bar{J}(h^*) = \mathbb{E}_{\mathbf{x}\sim p} \log \left[1 + \nu \frac{q(\mathbf{x})}{p(\mathbf{x})}\right] + \nu \mathbb{E}_{\mathbf{y}\sim q} \log \left[1 + \frac{p(\mathbf{y})}{\nu q(\mathbf{y})}\right] \tag{13}$$

$$= \dots$$

$$= -\mathrm{KL}(p||M_\nu) - \nu \mathrm{KL}(q||M_\nu) + 2\log 2 \tag{14}$$

with $M_\nu = (p + \nu q)/2$

Logistic loss

- For large sample sizes n and m, $J(h) \to \bar{J}(h)$ and the corresponding minimal loss is

$$\bar{J}(h^*) = \mathbb{E}_{\mathbf{x}\sim p} \log\left[1 + \nu\frac{q(\mathbf{x})}{p(\mathbf{x})}\right] + \nu\mathbb{E}_{\mathbf{y}\sim q}\log\left[1 + \frac{p(\mathbf{y})}{\nu q(\mathbf{y})}\right] \tag{13}$$

$$= \dots$$

$$= -\mathrm{KL}(p||M_\nu) - \nu\mathrm{KL}(q||M_\nu) + 2\log 2 \tag{14}$$

with $M_\nu = (p + \nu q)/2$

- For $\nu = 1$, $\bar{J}(h^*) = -2\mathrm{JSD}(p, q) + 2\log 2$, and hence

$$\bar{J}(h) \geq -2\mathrm{JSD}(p, q) + 2\log 2 \tag{15}$$

Logistic loss

- For large sample sizes n and m, $J(h) \to \bar{J}(h)$ and the corresponding minimal loss is

$$\bar{J}(h^*) = \mathbb{E}_{\mathbf{x}\sim p} \log \left[1 + \nu \frac{q(\mathbf{x})}{p(\mathbf{x})}\right] + \nu \mathbb{E}_{\mathbf{y}\sim q} \log \left[1 + \frac{p(\mathbf{y})}{\nu q(\mathbf{y})}\right] \tag{13}$$

$$= \dots$$

$$= -\mathrm{KL}(p||M_\nu) - \nu \mathrm{KL}(q||M_\nu) + 2\log 2 \tag{14}$$

  with $M_\nu = (p + \nu q)/2$
- For $\nu = 1$, $\bar{J}(h^*) = -2\mathrm{JSD}(p, q) + 2\log 2$, and hence

$$\bar{J}(h) \geq -2\mathrm{JSD}(p, q) + 2\log 2 \tag{15}$$

- Contrastive learning via classification with the logistic loss estimates the JSD.

Other loss functions

- In the following, I will focus on the logistic loss as done in our early work on contrastive learning for the estimation of unnormalised models, "Noise-contrastive estimation (NCE)" (Gutmann and Hyvärinen, AISTATS 2010).

Other loss functions

- In the following, I will focus on the logistic loss as done in our early work on contrastive learning for the estimation of unnormalised models, "Noise-contrastive estimation (NCE)" (Gutmann and Hyvärinen, AISTATS 2010).
- But other loss functions can be used:
  - multinomial logistic loss when we contrast more than two data points.
  - Bregman divergences
  - f-divergences
  - ...

Constructing reference data

Choice depends on the specific application of contrastive learning.

- Fit a preliminary model and keep it fixed (as often done in NCE)
- Iterative approach: fitted model becomes reference in the next iteration (as also done in our original work on NCE!)
- Use other segments for time series data
  (Hyvärinen and Morioka, NeurIPS 2016)
- For Bayesian inference, use prior predictive distribution
  (Thomas et al, 2016; Thomas et al, Bayesian Analysis, 2020)
- Generate it conditionally on observed data
  (Ceylan and Gutmann, ICML 2018)
- Iterative adaptive approach with implicit models: gives GANs
  (Goodfellow et al, NeurIPS 2014)
- Iterative adaptive approach with flexible density model such as flows ("Flow-contrastive estimation", Gao et al, NeurIPS 2019)
- ...

The density-chasm problem

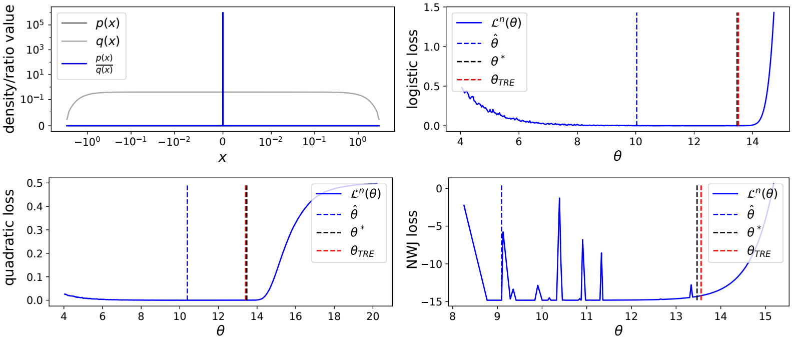
- Single ratio methods are sample inefficient if the two distributions are very different ("density chasm")

The density-chasm problem

- Single ratio methods are sample inefficient if the two distributions are very different ("density chasm")
- Consider ratio between two zero-mean Gaussians. 10'000 samples from each distribution. Ratio parametrised by $\theta \in \mathbb{R}$.

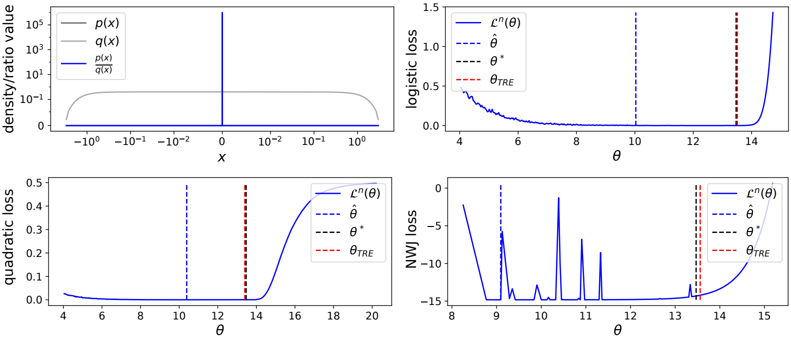
The density-chasm problem

- Single ratio methods are sample inefficient if the two distributions are very different ("density chasm")
- Consider ratio between two zero-mean Gaussians. 10'000 samples from each distribution. Ratio parametrised by $\theta \in \mathbb{R}$.



The density-chasm problem

- Single ratio methods are sample inefficient if the two distributions are very different ("density chasm")
- Consider ratio between two zero-mean Gaussians. 10'000 samples from each distribution. Ratio parametrised by $\theta \in \mathbb{R}$.
- Solution in red bridges the "gap" using telescopic ratio estimation (TRE) (Rhodes, Xu, and Gutmann, NeurIPS 2020)

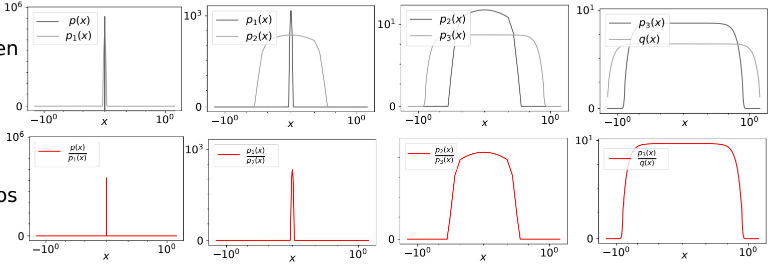


A single density-ratio fails to “bridge” the density-chasm.

Let us thus use multiple bridges.

Anneal between p(x) & q(x)

Estimate consecutive ratios

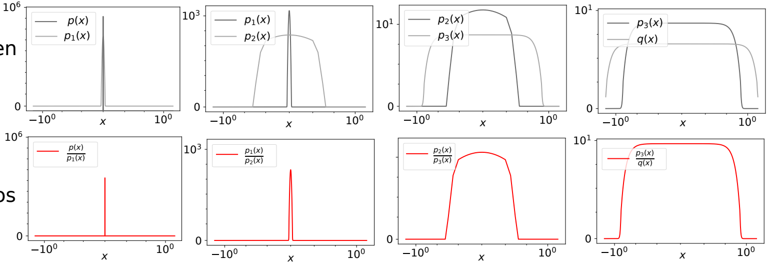


A single density-ratio fails to "bridge" the density-chasm.

Let us thus use multiple bridges.

Anneal between p(x) & q(x)

Estimate consecutive ratios

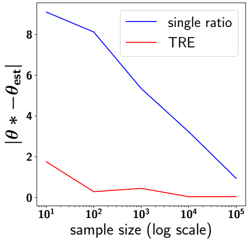


(relabel $p \equiv p_0$ and $q \equiv p_4$) and compute *telescoping* product

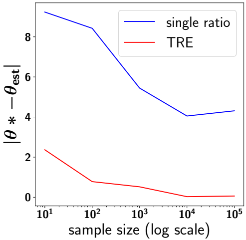
$$\frac{p_0(\mathbf{x})}{p_4(\mathbf{x})} = \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \frac{p_2(\mathbf{x})}{p_3(\mathbf{x})} \frac{p_3(\mathbf{x})}{p_4(\mathbf{x})}. \tag{16}$$

Sample efficiency curves for the 1d peaked ratio experiment.

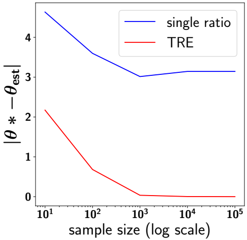
More results in the paper!



(a) Logistic loss



(b) NWJ loss



(c) Quadratic loss

Contents

Computational difficulties in likelihood-based learning

Contrastive learning

Applications in statistical inference and experimental design

Estimation of energy-based models

(Gutmann and Hyvärinen, AISTATS 2010; JMLR 2012)
(Pihlaja, Gutmann, and Hyvärinen, UAI2010; Gutmann and Hirayama, UAI 2011)
(Rhodes, Xu, and Gutmann, NeurIPS 2020)

- Data: random sample from $\mathbf{x} \sim p_{\mathbf{x}}$

Estimation of energy-based models $p(\mathbf{x}|\boldsymbol{\theta}) \propto \phi(\mathbf{x}|\boldsymbol{\theta})$

(Gutmann and Hyvärinen, AISTATS 2010; JMLR 2012)
(Pihlaja, Gutmann, and Hyvärinen, UAI2010; Gutmann and Hirayama, UAI 2011)
(Rhodes, Xu, and Gutmann, NeurIPS 2020)

- Data: random sample from $\mathbf{x} \sim p_{\mathbf{x}}$
- Introduce reference data $\mathbf{y} \sim q$

Estimation of energy-based models

(Gutmann and Hyvärinen, AISTATS 2010; JMLR 2012)
(Pihlaja, Gutmann, and Hyvärinen, UAI2010; Gutmann and Hirayama, UAI 2011)
(Rhodes, Xu, and Gutmann, NeurIPS 2020)

- Data: random sample from $\mathbf{x} \sim p_{\mathbf{x}}$
- Introduce reference data $\mathbf{y} \sim q$
- Estimate the log-ratio $h(\mathbf{x}; \boldsymbol{\theta}) \approx \log p_{\mathbf{x}}(\mathbf{x}) - \log q(\mathbf{x})$

Estimation of energy-based models

(Gutmann and Hyvärinen, AISTATS 2010; JMLR 2012)
(Pihlaja, Gutmann, and Hyvärinen, UAI2010; Gutmann and Hirayama, UAI 2011)
(Rhodes, Xu, and Gutmann, NeurIPS 2020)

- Data: random sample from $\mathbf{x} \sim p_{\mathbf{x}}$
- Introduce reference data $\mathbf{y} \sim q$
- Estimate the log-ratio $h(\mathbf{x}; \boldsymbol{\theta}) \approx \log p_{\mathbf{x}}(\mathbf{x}) - \log q(\mathbf{x})$
  - Either parametrise $h(\mathbf{x}; \boldsymbol{\theta})$ in terms of an energy-based model if provided, i.e. $h(\mathbf{x}; \boldsymbol{\theta}) = \log \phi(\mathbf{x}|\boldsymbol{\theta}) - \log q(\mathbf{x}) + \text{const}$

Estimation of energy-based models

(Gutmann and Hyvärinen, AISTATS 2010; JMLR 2012)
(Pihlaja, Gutmann, and Hyvärinen, UAI2010; Gutmann and Hirayama, UAI 2011)
(Rhodes, Xu, and Gutmann, NeurIPS 2020)

- Data: random sample from $\mathbf{x} \sim p_{\mathbf{x}}$
- Introduce reference data $\mathbf{y} \sim q$
- Estimate the log-ratio $h(\mathbf{x};\boldsymbol{\theta}) \approx \log p_{\mathbf{x}}(\mathbf{x}) - \log q(\mathbf{x})$
  - Either parametrise $h(\mathbf{x};\boldsymbol{\theta})$ in terms of an energy-based model if provided, i.e. $h(\mathbf{x};\boldsymbol{\theta}) = \log \phi(\mathbf{x}|\boldsymbol{\theta}) - \log q(\mathbf{x}) + \text{const}$
  - Or parametrise the log-ratio $h(\mathbf{x};\boldsymbol{\theta})$ directly, e.g. for deep unsupervised learning.

Estimation of energy-based models

$p(\mathbf{x}|\boldsymbol{\theta}) \propto \phi(\mathbf{x}|\boldsymbol{\theta})$

(Gutmann and Hyvärinen, AISTATS 2010; JMLR 2012)
(Pihlaja, Gutmann, and Hyvärinen, UAI2010; Gutmann and Hirayama, UAI 2011)
(Rhodes, Xu, and Gutmann, NeurIPS 2020)

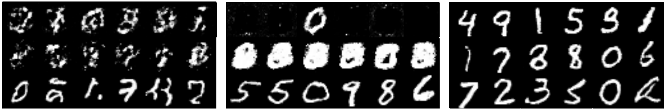
- Data: random sample from $\mathbf{x} \sim p_{\mathbf{x}}$
- Introduce reference data $\mathbf{y} \sim q$
- Estimate the log-ratio $h(\mathbf{x}; \boldsymbol{\theta}) \approx \log p_{\mathbf{x}}(\mathbf{x}) - \log q(\mathbf{x})$
  - Either parametrise $h(\mathbf{x}; \boldsymbol{\theta})$ in terms of an energy-based model if provided, i.e. $h(\mathbf{x}; \boldsymbol{\theta}) = \log \phi(\mathbf{x}|\boldsymbol{\theta}) - \log q(\mathbf{x}) + \text{const}$
  - Or parametrise the log-ratio $h(\mathbf{x}; \boldsymbol{\theta})$ directly, e.g. for deep unsupervised learning.
- Set $\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}) = \underbrace{\log q(\mathbf{x})}_{\text{reference}} + \underbrace{h(\mathbf{x}; \hat{\boldsymbol{\theta}})}_{\text{difference}}$

(Gutmann and Hyvärinen, AISTATS 2010; JMLR 2012)
(Pihlaja, Gutmann, and Hyvärinen, UAI2010; Gutmann and Hirayama, UAI 2011)
(Rhodes, Xu, and Gutmann, NeurIPS 2020)

- Data: random sample from $\mathbf{x} \sim p_{\mathbf{x}}$
- Introduce reference data $\mathbf{y} \sim q$
- Estimate the log-ratio $h(\mathbf{x};\boldsymbol{\theta}) \approx \log p_{\mathbf{x}}(\mathbf{x}) - \log q(\mathbf{x})$
  - Either parametrise $h(\mathbf{x};\boldsymbol{\theta})$ in terms of an energy-based model if provided, i.e. $h(\mathbf{x};\boldsymbol{\theta}) = \log \phi(\mathbf{x}|\boldsymbol{\theta}) - \log q(\mathbf{x}) + \text{const}$
  - Or parametrise the log-ratio $h(\mathbf{x};\boldsymbol{\theta})$ directly, e.g. for deep unsupervised learning.
- Set $\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}) = \underbrace{\log q(\mathbf{x})}_{\text{reference}} + \underbrace{h(\mathbf{x};\hat{\boldsymbol{\theta}})}_{\text{difference}}$

Noise distribution **Single ratio (NCE)** **TRE**

Gaussian
Copula
Flow



(Figure from Rhodes, Xu, and Gutmann, NeurIPS 2020)

Bayesian inference for simulator-based models

(Likelihood-Free Inference by Ratio Estimation, Thomas et al, 2016; 2020)
(Dinev and Gutmann, arXiv:1810.09899, 2018)

- Consider simulator-based model $\mathbf{x} = g(\boldsymbol{\theta}, \boldsymbol{\omega}), \ \boldsymbol{\omega} \sim p(\boldsymbol{\omega})$
- Task: estimate the posterior $p(\boldsymbol{\theta}|\mathbf{x}^o)$
- Contrastive interpretation of Bayes' rule:

$$\underbrace{\log p(\boldsymbol{\theta})}_{\text{reference}} + \underbrace{\log p(\mathbf{x}|\boldsymbol{\theta}) - \log p(\mathbf{x})}_{\text{difference}} \Rightarrow \underbrace{\log p(\boldsymbol{\theta}|\mathbf{x})}_{\text{interest}} \qquad (17)$$

- Use simulator to generate data from $p(\mathbf{x}|\boldsymbol{\theta})$ and from $p(\mathbf{x})$.
- Learning the difference provides an estimate of the desired $h(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta}) - \log p(\mathbf{x})$ and hence an estimate of the posterior.

Experimental design for simulator-based models

(Kleinegesse and Gutmann, AISTATS 2019; ICML 2020; arXiv:2105.04379)
(Kleinegesse, Drovandi and Gutmann, Bayesian Analysis 2020)
(Ivanova, Foster, Kleinegesse, Gutmann and Rainforth, NeurIPS 2021)

- Example: Stochastic SIR model with noisy observations
  Latent process: Susceptibles $\rightarrow$ Infected $I(t) \rightarrow$ Recovered
  Observation model: $y(t)|\boldsymbol{\theta} \sim \text{Poisson}(y; \phi I(t))$

Experimental design for simulator-based models

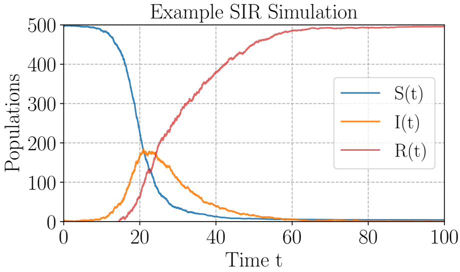
(Kleinegesse and Gutmann, AISTATS 2019; ICML 2020; arXiv:2105.04379)
(Kleinegesse, Drovandi and Gutmann, Bayesian Analysis 2020)
(Ivanova, Foster, Kleinegesse, Gutmann and Rainforth, NeurIPS 2021)

- Example: Stochastic SIR model with noisy observations
  Latent process: Susceptibles $\rightarrow$ Infected $I(t)$ $\rightarrow$ Recovered
  Observation model: $y(t)|\boldsymbol{\theta} \sim \text{Poisson}(y; \phi I(t))$
- Parameters $\boldsymbol{\theta} = (\beta, \gamma)$: infection rate and recovery rate

Experimental design for simulator-based models

(Kleinegesse and Gutmann, AISTATS 2019; ICML 2020; arXiv:2105.04379)
(Kleinegesse, Drovandi and Gutmann, Bayesian Analysis 2020)
(Ivanova, Foster, Kleinegesse, Gutmann and Rainforth, NeurIPS 2021)

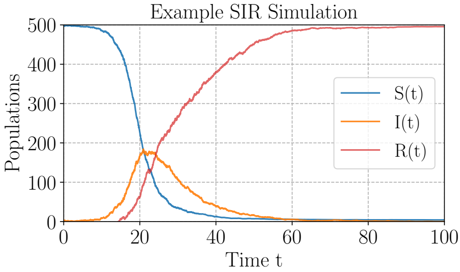
- Example: Stochastic SIR model with noisy observations
  Latent process: Susceptibles $\rightarrow$ Infected $I(t)$ $\rightarrow$ Recovered
  Observation model: $y(t)|\boldsymbol{\theta} \sim \text{Poisson}(y; \phi I(t))$
- Parameters $\boldsymbol{\theta} = (\beta, \gamma)$: infection rate and recovery rate



Experimental design for simulator-based models

(Kleinegesse and Gutmann, AISTATS 2019; ICML 2020; arXiv:2105.04379)
(Kleinegesse, Drovandi and Gutmann, Bayesian Analysis 2020)
(Ivanova, Foster, Kleinegesse, Gutmann and Rainforth, NeurIPS 2021)

- Example: Stochastic SIR model with noisy observations
  Latent process: Susceptibles $\rightarrow$ Infected $I(t)$ $\rightarrow$ Recovered
  Observation model: $y(t)|\boldsymbol{\theta} \sim \text{Poisson}(y; \phi I(t))$
- Parameters $\boldsymbol{\theta} = (\beta, \gamma)$: infection rate and recovery rate
- Task: find the optimal times at which to take measurements to most accurately estimate $\boldsymbol{\theta}$.



Experimental design for simulator-based models

- Experimental design by maximising mutual information (MI)

$$\hat{\mathbf{d}} = \operatorname*{argmax}_{\mathbf{d}} \mathbb{E}_{p(\mathbf{x},\boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta},\mathbf{d})}{p(\mathbf{x}|\mathbf{d})} \right] \tag{18}$$

- Use contrastive learning to estimate

$$h_{\mathbf{d}}(\mathbf{x},\boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta},\mathbf{d}) - \log p(\mathbf{x}|\mathbf{d}), \tag{19}$$

  and maximise sample average of $h_{\mathbf{d}}(\mathbf{x},\boldsymbol{\theta})$ with respect to $\mathbf{d}$
- Static setting: Kleinegesse and Gutmann, AISTATS 2019
- Sequential setting where we update our belief about $\boldsymbol{\theta}$ as we sequentially acquire the data: Kleinegesse, Drovandi and Gutmann, Bayesian Analysis 2020

Experimental design for simulator-based models

$$\hat{\mathbf{d}} = \operatorname{argmax}_{\mathbf{d}} \mathbb{E}_{p(\mathbf{x},\boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta},\mathbf{d})}{p(\mathbf{x}|\mathbf{d})} \right]$$

- Learning the ratio $h_{\mathbf{d}}(\mathbf{x}, \boldsymbol{\theta})$ and approximating the MI is computationally costly.

Experimental design for simulator-based models

$$\hat{\mathbf{d}} = \text{argmax}_{\mathbf{d}} \, \mathbb{E}_{p(\mathbf{x},\boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta},\mathbf{d})}{p(\mathbf{x}|\mathbf{d})} \right]$$

- Learning the ratio $h_{\mathbf{d}}(\mathbf{x}, \boldsymbol{\theta})$ and approximating the MI is computationally costly.
- But we do not need to estimate the MI accurately everywhere! Only around it's maximum.

Experimental design for simulator-based models

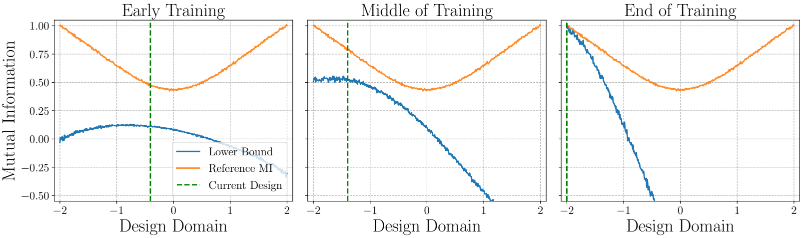
$$\hat{\mathbf{d}} = \operatorname{argmax}_{\mathbf{d}} \mathbb{E}_{p(\mathbf{x},\boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta},\mathbf{d})}{p(\mathbf{x}|\mathbf{d})} \right]$$

- Learning the ratio $h_{\mathbf{d}}(\mathbf{x}, \boldsymbol{\theta})$ and approximating the MI is computationally costly.
- But we do not need to estimate the MI accurately everywhere! Only around it's maximum.
- Suggests approach using lower bounds on the MI (or proxy quantities) where we concurrently tighten the bound and maximise the (proxy) MI.

Experimental design for simulator-based models

$$\hat{\mathbf{d}} = \text{argmax}_{\mathbf{d}}\, \mathbb{E}_{p(\mathbf{x},\boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta},\mathbf{d})}{p(\mathbf{x}|\mathbf{d})} \right]$$

- Learning the ratio $h_{\mathbf{d}}(\mathbf{x}, \boldsymbol{\theta})$ and approximating the MI is computationally costly.
- But we do not need to estimate the MI accurately everywhere! Only around it's maximum.
- Suggests approach using lower bounds on the MI (or proxy quantities) where we concurrently tighten the bound and maximise the (proxy) MI.



(Kleinegesse and Gutmann, ICML 2020; arXiv:2105.04379)

Experimental design for simulator-based models

$$\hat{\mathbf{d}} = \text{argmax}_{\mathbf{d}} \, \text{KL}\left(p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{d}) || p(\boldsymbol{\theta}|\mathbf{d}) p(\mathbf{x}|\mathbf{d})\right)$$

- We can (again!) leverage logistic regression.

Experimental design for simulator-based models

$$\hat{\mathbf{d}} = \text{argmax}_{\mathbf{d}} \, \text{KL}\left(p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{d}) || p(\boldsymbol{\theta}|\mathbf{d}) p(\mathbf{x}|\mathbf{d})\right)$$

- We can (again!) leverage logistic regression.
- Logistic regression results in replacing the KL divergence with the JSD when measuring the MI.

$$\text{JSD}(p, q) \geq \log 2 - \frac{1}{2} \bar{J}(h) \tag{20}$$

where h is the regression function and $\bar{J}$ the logistic loss.

Experimental design for simulator-based models

$$\hat{\mathbf{d}} = \text{argmax}_{\mathbf{d}} \, \text{KL}\left(p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{d}) || p(\boldsymbol{\theta}|\mathbf{d}) p(\mathbf{x}|\mathbf{d})\right)$$

- We can (again!) leverage logistic regression.
- Logistic regression results in replacing the KL divergence with the JSD when measuring the MI.

$$\text{JSD}(p, q) \geq \log 2 - \frac{1}{2}\bar{J}(h) \tag{20}$$

  where h is the regression function and $\bar{J}$ the logistic loss.
- Perform experimental design by maximising the negative logistic loss jointly with respect to h and $\mathbf{d}$.

Experimental design for simulator-based models

$$\hat{\mathbf{d}} = \operatorname{argmax}_{\mathbf{d}} \mathrm{KL}\left(p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{d}) || p(\boldsymbol{\theta}|\mathbf{d}) p(\mathbf{x}|\mathbf{d})\right)$$

- We can (again!) leverage logistic regression.
- Logistic regression results in replacing the KL divergence with the JSD when measuring the MI.

$$\mathrm{JSD}(p, q) \geq \log 2 - \frac{1}{2}\bar{J}(h) \tag{20}$$

  where h is the regression function and $\bar{J}$ the logistic loss.
- Perform experimental design by maximising the negative logistic loss jointly with respect to h and $\mathbf{d}$.
- Learned h provides an estimate of the posterior (as before!)

Experimental design for simulator-based models

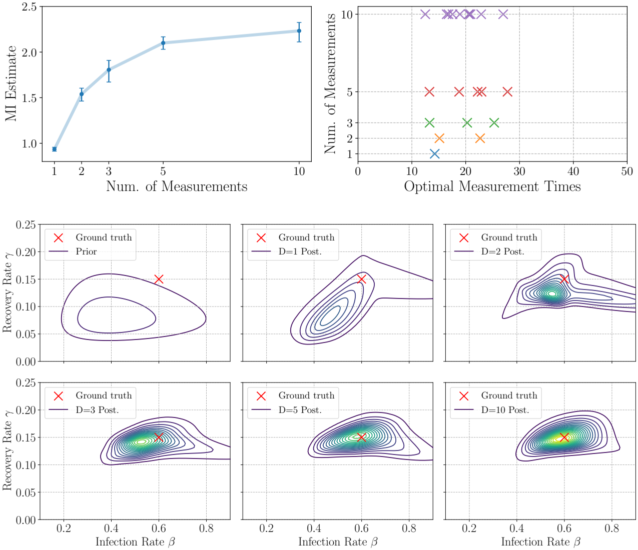
$$\hat{\mathbf{d}} = \text{argmax}_{\mathbf{d}} \, \text{KL}\left(p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{d}) || p(\boldsymbol{\theta}|\mathbf{d}) p(\mathbf{x}|\mathbf{d})\right)$$

- We can (again!) leverage logistic regression.
- Logistic regression results in replacing the KL divergence with the JSD when measuring the MI.

$$\text{JSD}(p, q) \geq \log 2 - \frac{1}{2}\bar{J}(h) \tag{20}$$

  where h is the regression function and $\bar{J}$ the logistic loss.
- Perform experimental design by maximising the negative logistic loss jointly with respect to h and $\mathbf{d}$.
- Learned h provides an estimate of the posterior (as before!)
- For more details and other loss functions:
  Kleinegesse and Gutmann, ICML 2020; arXiv:2105.04379

SIR example


MI Estimate
Num. of Measurements
Num. of Measurements
Optimal Measurement Times
Ground truth
Prior
D=1 Post.
D=2 Post.
D=3 Post.
D=5 Post.
D=10 Post.
Recovery Rate γ
Infection Rate β

Conclusions

- Introduced energy-based (unnormalised) and simulator-based (implicit) models.
- Pointed out that their likelihood function is computationally intractable.
- Introduced contrastive learning as an intuitive and computationally feasible alternative to likelihood-based learning.
- Contrastive learning is closely related to classification, logistic regression, and ratio estimation.
- We can use it to solve a range of difficult statistical problems:
  1. Parameter estimation for energy-based models
  2. Bayesian inference for simulator-based models
  3. Bayesian experimental design for simulator-based models
- For papers, see `https://michaelgutmann.github.io`