Bayesian experimental design and inference for simulator models

Michael U. Gutmann

michael.gutmann@ed.ac.uk

School of Informatics, University of Edinburgh

27th October 2023

Contents

Research objective
- Two main goals: inference and experimental design
- Tasks are computationally intractable for simulator models

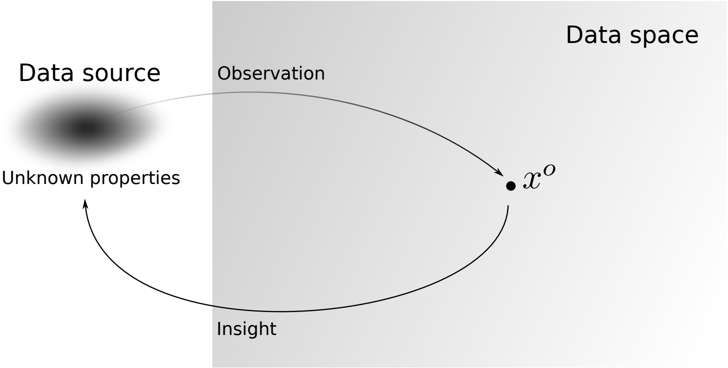
Contrastive learning to deal with intractability
- Link to logistic regression and Jensen-Shannon divergence
- Technical challenge: the density-chasm problem

Bayesian inference and experimental design
- Via contrastive learning of density ratios
- Exploiting bounds to increase computational efficiency

Contents

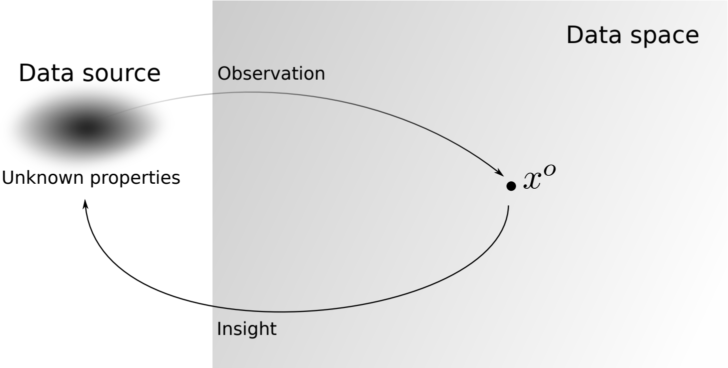
Overall goal

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



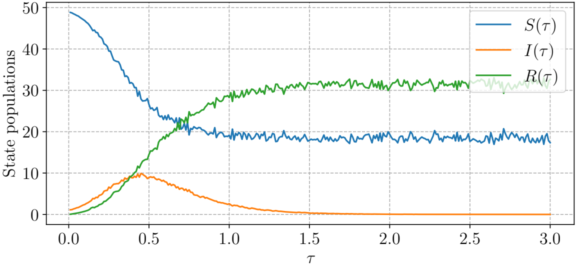
Two fundamental tasks

- 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?



Example: stochastic SIR model

- Stochastic model describing the population of susceptibles $S(\tau)$, infected $I(\tau)$ and recovered $R(\tau)$ as a function of time.
- Parameters $\boldsymbol{\theta}$: rate of infection β and the rate of recovery γ.
- Inference task : determine plausible values of β and γ given some measurements of the population sizes.
- Exp design task : find the optimal times at which to perform the measurements to most accurately estimate β and γ.



(Figure by Steven Kleinegesse)

Bayesian inference and design with tractable models

- Assume model is expressed as a family of pdfs $\{p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})\}$ indexed by parameter $\boldsymbol{\theta}$ and design variable $\mathbf{d}$.
- Bayesian inference of $\boldsymbol{\theta}$ for data $\mathbf{x}_o$ obtained with design $\mathbf{d}_o$:

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{d}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})}{p(\mathbf{x}|\mathbf{d})} p(\boldsymbol{\theta}|\mathbf{d}) \tag{1}$$

  with $\mathbf{x}$ fixed to $\mathbf{x}_o$ and $\mathbf{d}$ to $\mathbf{d}_o$.
- Experimental design by maximising mutual information (MI) between data $\mathbf{x}$ and parameters $\boldsymbol{\theta}$:

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

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

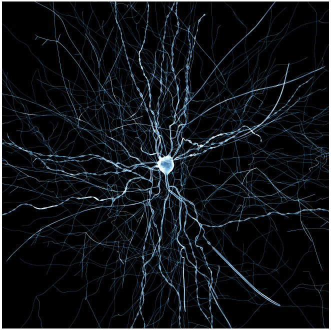
$$= \mathbb{E}_{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})}{p(\mathbf{x}|\mathbf{d})}\right] \tag{4}$$

Simulator models

- Not all models are specified as family of pdfs.
- We consider here the important class of simulator models: models which are specified via a (stochastic) mechanism for generating data
- Think of a latent continuous-time stochastic process described by a system of coupled SDEs followed by observation model.
- Key properties:
  - No closed form expression for $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})$ available
  - Models typically contain too many latent variables for standard MCMC inference to be feasible.
  - Sampling data $\mathbf{x}|\boldsymbol{\theta}, \mathbf{d} \sim p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})$ is possible

Simulator models are widely used in the natural sciences

- Evolutionary biology:
  to model evolution
- Biochemistry:
  to model gene expression
- Neuroscience:
  to model neural processing
- Cognitive sciences:
  to model human decision making
- Epidemiology:
  to model the spread of an infectious disease
- ...



Simulated neural activity in rat somatosensory cortex
(Figure from `https://bbp.epfl.ch/nmc-portal`)

Research objective

- Simulator models have great modelling power.
- However, we pay the price when attempting to perform inference and experimental design:
  evaluating $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})$ is computationally intractable
- Paradoxical situation: we have great models from the natural sciences but cannot fully use them because we lack principled tools for inference and experimental design.
- Research objective:
  Develop efficient tools for Bayesian inference and experimental design with simulator models.

Technical aside on the intractability of simulator models

- A simulator model is technically a measurable function g

$$\mathbf{x} = g(\boldsymbol{\theta}, \mathbf{d}, \boldsymbol{\omega}), \quad \boldsymbol{\omega} \sim p(\boldsymbol{\omega}) \tag{5}$$

  Maps parameters $\boldsymbol{\theta}$, design variables $\mathbf{d}$ and "noise" $\boldsymbol{\omega}$ to data $\mathbf{x}$

- Function g is not known in closed form but implemented as a computer programme.
- The probability distribution of $\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{d}$ is defined by the probability of the inverse image: $\Pr(\mathbf{x} \in \mathcal{A} | \boldsymbol{\theta}, \mathbf{d}) = \Pr(\{\omega : g(\boldsymbol{\theta}, \mathbf{d}, \boldsymbol{\omega}) \in \mathcal{A}\})$
- Computing inverse image and it's measure is generally not possible.

Contents

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

$$\underbrace{b}_{\text{reference}} + \underbrace{a-b}_{\text{difference}} \Rightarrow \underbrace{a}_{\text{interest}} \tag{6}$$

- Link to (log) ratio estimation

$$\underbrace{\log p_b}_{\text{reference}} + \underbrace{\log p_a - \log p_b}_{\text{difference}} \Rightarrow \underbrace{\log p_a}_{\text{interest}} \tag{7}$$

- Link to Bayes' rule

$$\underbrace{\log p(\boldsymbol{\theta})}_{\text{reference}} + \underbrace{\log p(\mathbf{x}|\boldsymbol{\theta}) - \log p(\mathbf{x})}_{\text{difference}} \Rightarrow \underbrace{\log p(\boldsymbol{\theta}|\mathbf{x})}_{\text{interest}} \tag{8}$$

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 $J(h)$

$$\begin{aligned} J(h) =& \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \nu \exp(-h(\mathbf{x}_i))\right] + \\ & \frac{\nu}{m}\sum_{i=1}^{m} \log\left[1 + \frac{1}{\nu}\exp(h(\mathbf{y}_i))\right] \end{aligned} \quad (9)$$

  where $\nu = n/m$ and h is a nonlinearity (e.g. neural network) that we learn.
- For large sample sizes n and m (and fixed ratio ν), the optimal h is

$$h^* = \log p - \log q \quad (10)$$

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*

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

2. We only need samples from p and q; we do need not 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

$$\bar{J}(h^*) = -\mathrm{KL}(p||M_\nu) - \nu \mathrm{KL}(q||M_\nu) + 2\log 2 \tag{12}$$

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

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

$$\bar{J}(h) \geq -2\mathrm{JSD}(p, q) + 2\log 2 \tag{13}$$

- Contrastive learning via classification with the logistic loss corresponds to estimating the JSD.
- For a review paper on statistical applications of contrastive learning, see Gutmann, Kleinegesse, and Rhodes, Behaviormetrika, 2022.

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 (Srivastava, et al, TMLR 2023)
  - Bregman divergences (Gutmann and Hirayama, UAI 2011)
  - f-divergences (e.g. Rhodes and Gutmann, AISTATS, 2019)
  - ...

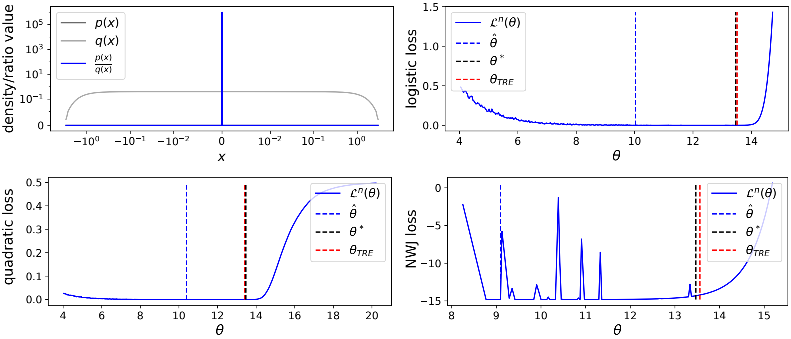
Constructing reference data

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 generative models: results into GANs (Goodfellow et al, NeurIPS 2014)
- Iterative adaptive approach with flexible density model such as flows ("Flow-contrastive estimation", Gao et al, NeurIPS 2019)
- ...

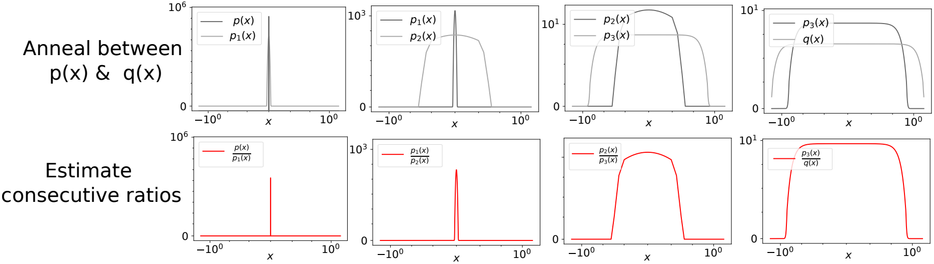
The density-chasm problem

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



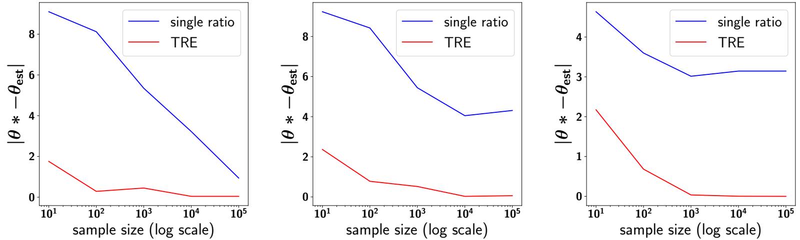
(relabel $p \equiv p_0$ and $q \equiv p_4$) and compute *telescoping* product

$$\frac{p_0(\mathbf{x})}{p_4(\mathbf{x})} = \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \frac{p_2(\mathbf{x})}{p_3(\mathbf{x})} \frac{p_3(\mathbf{x})}{p_4(\mathbf{x})}. \tag{14}$$

Telescoping density-ratio estimation (Rhodes, Xu, and Gutmann, NeurIPS 2020)

Sample efficiency curves for the 1d peaked ratio experiment.

More results in the paper!



(a) Logistic loss (b) NWJ loss (c) Quadratic loss

For further improvements, see "Estimating the Density Ratio between Distributions with High Discrepancy using Multinomial Logistic Regression", Srivastava et al, TMLR 2023.

Contents

Bayesian inference for simulator models

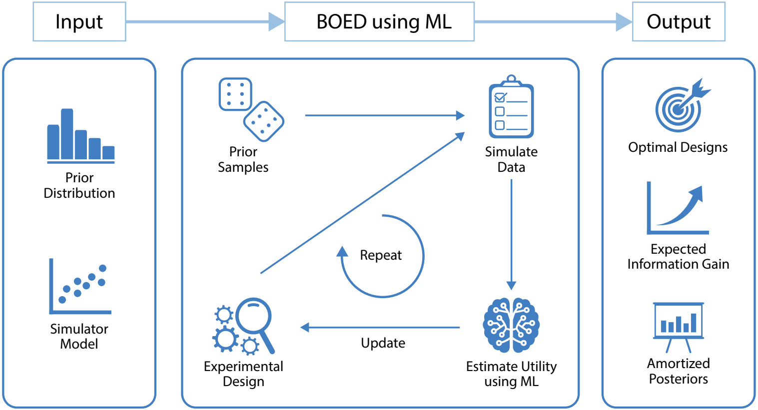
(Likelihood-Free Inference by Ratio Estimation, Thomas et al, 2016; 2020)
(Dinev and Gutmann, arXiv:1810.09899, 2018)

- Consider simulator 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:

$$\underbrace{\log p(\boldsymbol{\theta})}_{\text{reference}} + \underbrace{\log p(\mathbf{x}|\boldsymbol{\theta}) - \log p(\mathbf{x})}_{\text{difference}} \Rightarrow \underbrace{\log p(\boldsymbol{\theta}|\mathbf{x})}_{\text{interest}} \qquad (15)$$

- 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.
- It may make sense to work with lower-dim representations of the data (Chen, Gutmann, and Weller, ICML 2023; Chen et al, ICLR 2021).

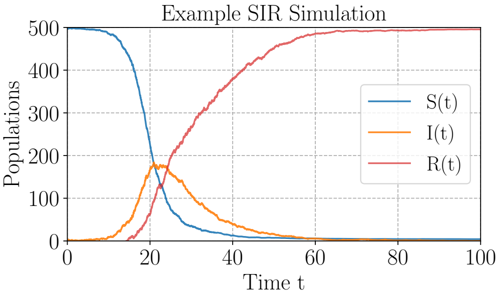
Experimental design for simulator models



(Valentin et al, arXiv:2305.07721, 2023)

(Kleinegesse and Gutmann, AISTATS 2019; ICML 2020; arXiv:2105.04379)
(Kleinegesse, Drovandi and Gutmann, Bayesian Analysis 2020)

- Example: Stochastic SIR model with noisy observations
  Latent process: Susceptibles $\to$ Infected $I(t)$ $\to$ 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 models

- Experimental design by maximising mutual information (MI)

$$\hat{\mathbf{d}} = \underset{\mathbf{d}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x},\boldsymbol{\theta}|\mathbf{d})} \log \left[\frac{p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d})}{p(\mathbf{x}|\mathbf{d})} \right] \tag{16}$$

- Use contrastive learning to estimate

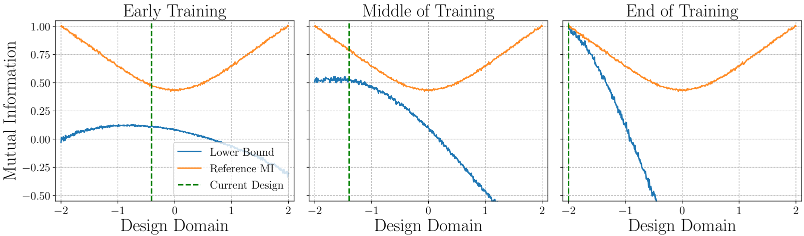
$$h_{\mathbf{d}}(\mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{d}) - \log p(\mathbf{x}|\mathbf{d}), \tag{17}$$

  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 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.
- Let us use lower bounds on the MI (or proxy) where we concurrently tighten the bound and maximise the (proxy) MI!



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

Experimental design for simulator models

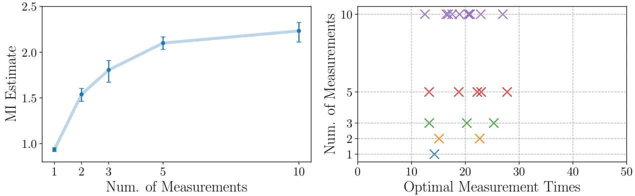
$$\hat{\mathbf{d}} = \text{argmax}_{\mathbf{d}} \, \text{KL}\left(p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{d}) || p(\boldsymbol{\theta}|\mathbf{d}) p(\mathbf{x}|\mathbf{d})\right)$$

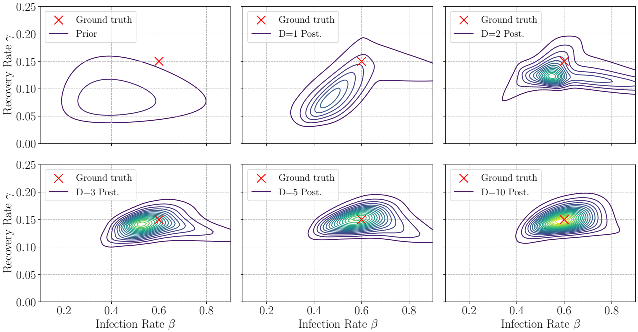
- We can (again!) leverage logistic regression.
- Logistic regression results in replacing the KL divergence with the JSD when measuring the MI.

$$\text{JSD}(p, q) \geq \log 2 - \frac{1}{2} \bar{J}(h) \tag{18}$$

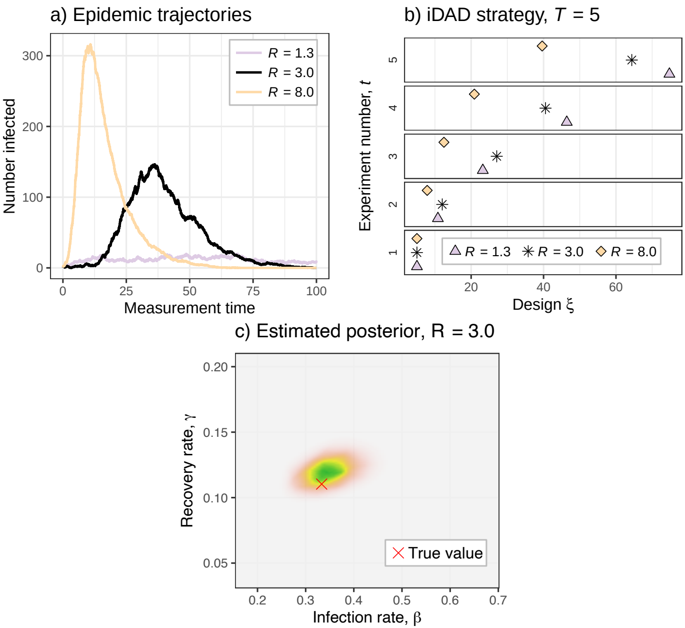
  where h is the regression function and $\bar{J}$ the logistic loss.
- Perform experimental design by maximising the negative logistic loss jointly with respect to h and $\mathbf{d}$.
- Learned h provides an estimate of the posterior (as before!)
- For more details and other loss functions: Kleinegesse and Gutmann, ICML 2020; arXiv:2105.04379
- For sequential setting: Ivanova et al, NeurIPS, 2021

SIR example: static case (Kleinegesse and Gutmann, ICML 2020)





SIR example: sequential case (Ivanova et al, NeurIPS, 2021)



Conclusions

Research objective

- Two main goals: inference and experimental design
- Tasks are computationally intractable for simulator models

Contrastive learning to deal with intractability

- Link to logistic regression and Jensen-Shannon divergence
- Technical challenge: the density-chasm problem

Bayesian inference and experimental design

- Via contrastive learning of density ratios
- Exploiting bounds to increase computational efficiency