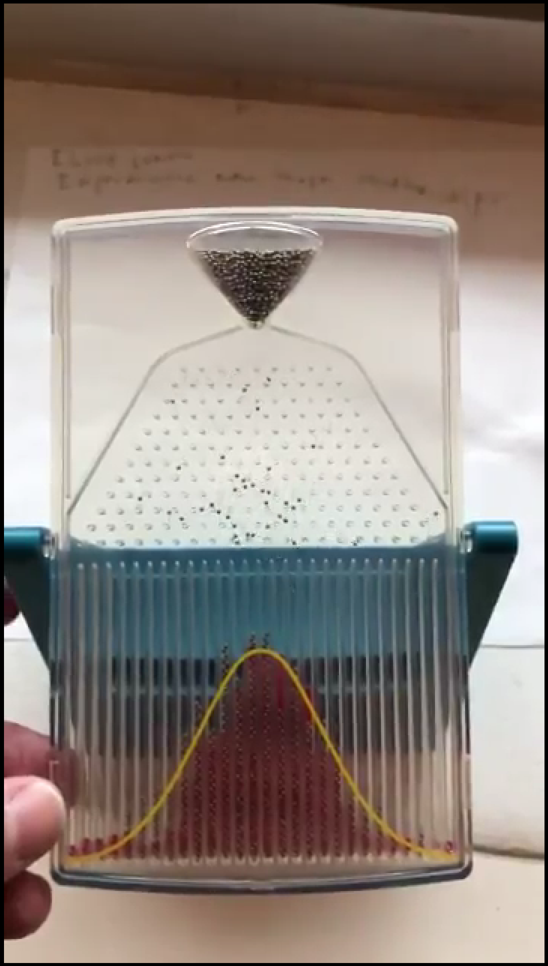
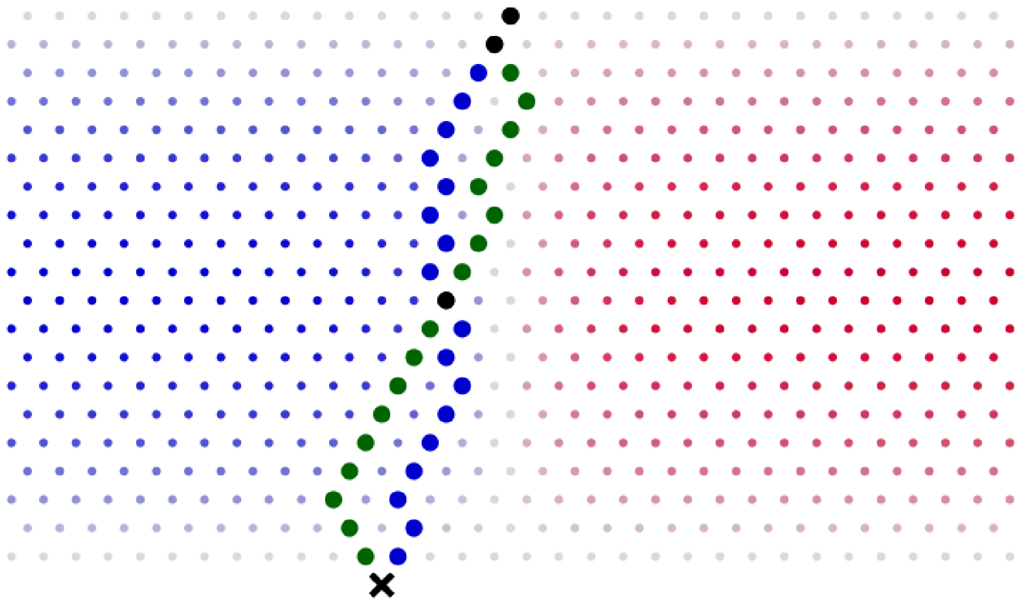
Likelihood ratio estimation for statistical inference in physical sciences

NTT
April 19, 2019

Gilles Louppe
@glouppe

LIÈGE université





The probability of ending in bin x corresponds to the total probability of all the paths z from start to x.

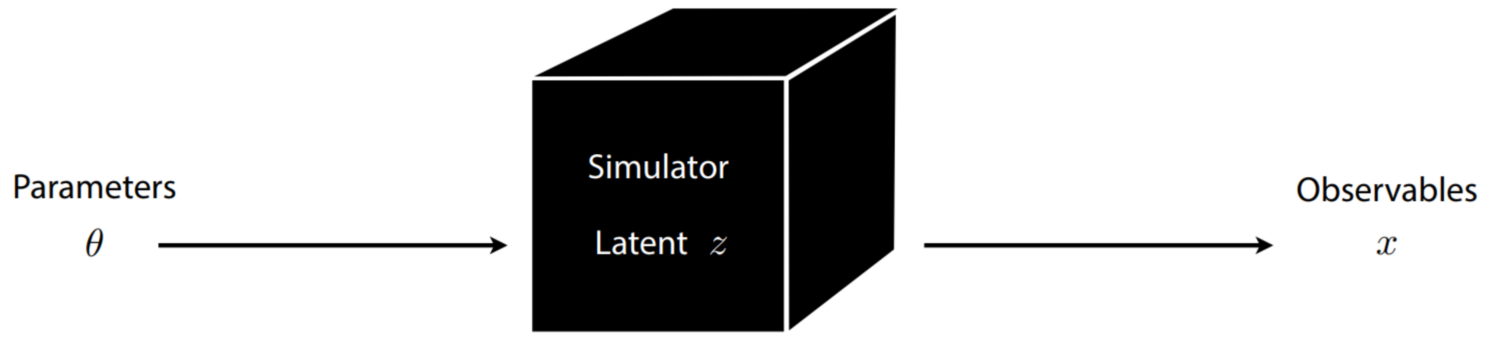
$$p(x|\theta) = \int p(x, z|\theta) dz = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

But what if we shift or remove some of the pins?

The Galton board is a metaphore of simulation-based science:

Galton board device	$\rightarrow$	Computer simulation
Parameters θ	$\rightarrow$	Model parameters θ
Buckets x	$\rightarrow$	Observables x
Random paths z	$\rightarrow$	Latent variables z (stochastic execution traces through simulator)

Inference in this context requires likelihood-free algorithms.

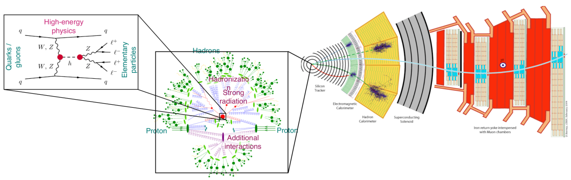


Prediction:
- Well-understood mechanistic model
- Simulator can generate samples

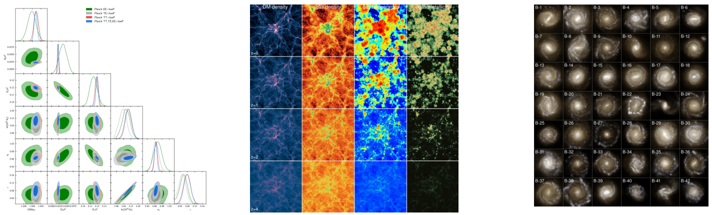
Inference:
- Likelihood function $p(x|\theta)$ is intractable
- Inference based on estimator $\hat{p}(x|\theta)$

Credits: Johann Brehmer.

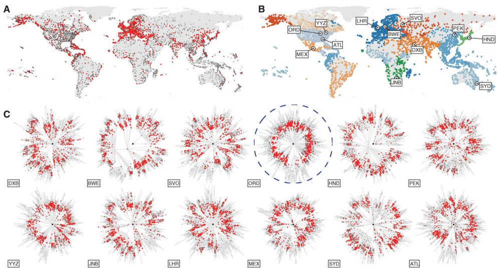
Applications (some)



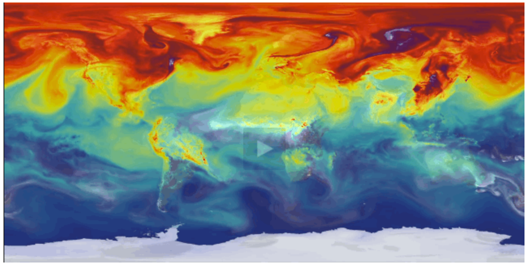
Particle physics



Cosmology

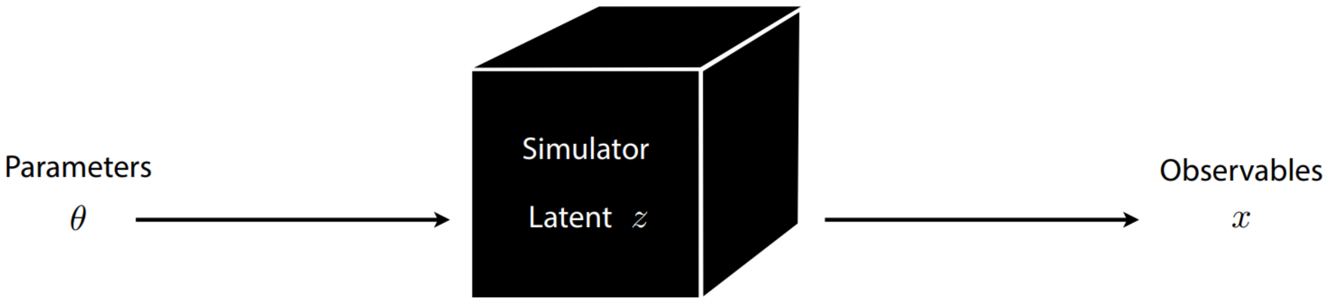


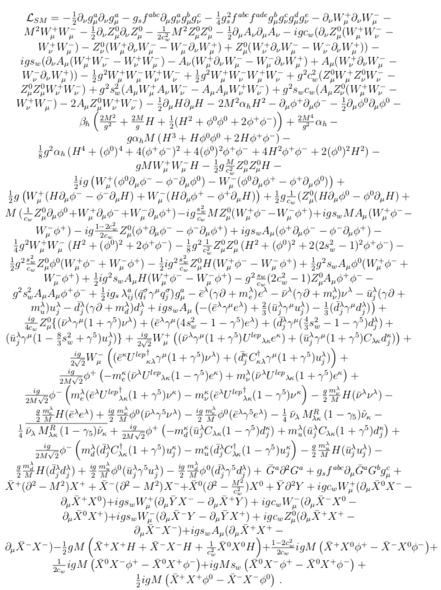
Epidemiology

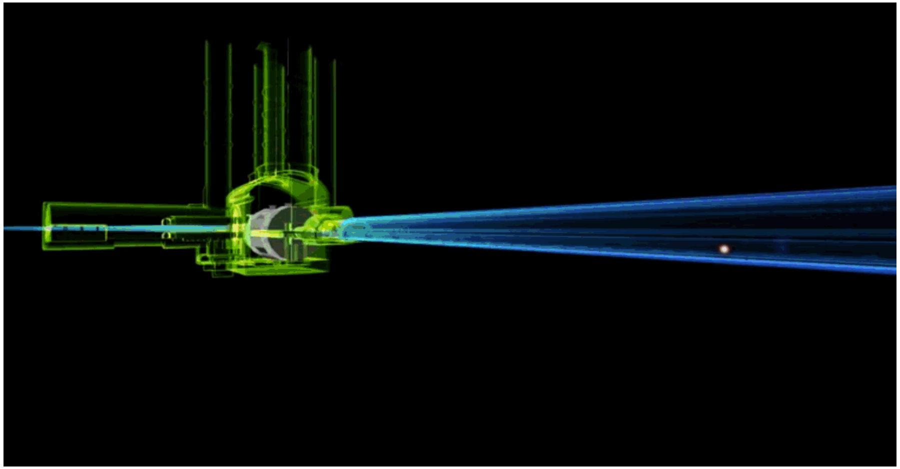


Climatology

Particle physics







Likelihood ratio estimation

The likelihood ratio

$$r(x|\theta_0, \theta_1) = \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

is the quantity that is central to many statistical inference procedures.

Examples

- Frequentist hypothesis testing
- Supervised learning
- Bayesian posterior sampling with MCMC
- Bayesian posterior inference through Variational Inference
- Generative adversarial networks
- Empirical Bayes with Adversarial Variational Optimization
- Optimal compression

When solving a problem of interest, do not solve a more general problem as an intermediate step. – Vladimir Vapnik

$$\frac{p_{\mathbf{x}}(\mathbf{x}|\theta_0)}{p_{\mathbf{x}}(\mathbf{x}|\theta_1)} = r(\mathbf{x};\theta_0,\theta_1)$$

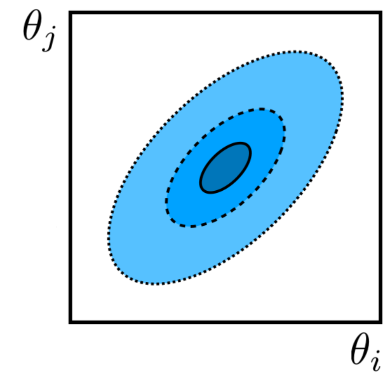
Direct likelihood ratio estimation is simpler than density estimation.

(This is fortunate, we are in the likelihood-free scenario!)

The frequentist physicist's way

The Neyman-Pearson lemma states that the likelihood ratio

$$r(x|\theta_0, \theta_1) = \frac{p(x|\theta_0)}{p(x|\theta_1)}$$



is the most powerful test statistic to discriminate between a null hypothesis θ_0 and an alternative θ_1.

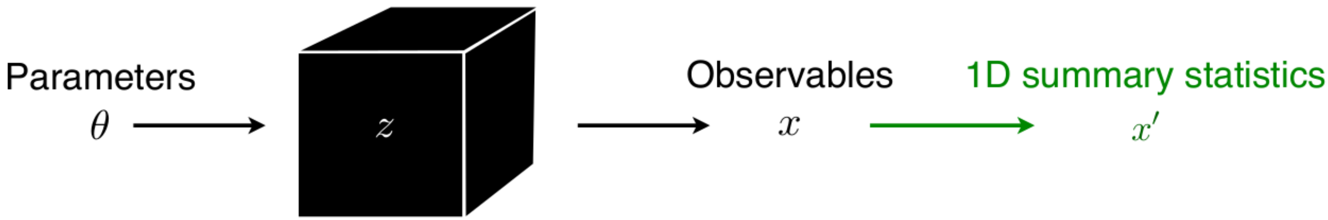
IX. *On the Problem of the most Efficient Tests of Statistical Hypotheses.*

By J. NEYMAN, *Nencki Institute, Soc. Sci. Lit. Varsoviensis, and Lecturer at the Central College of Agriculture, Warsaw, and* E. S. PEARSON, *Department of Applied Statistics, University College, London.*

(*Communicated by* K. PEARSON, *F.R.S.*)

(Received August 31, 1932.—Read November 10, 1932.)

CONTENTS.



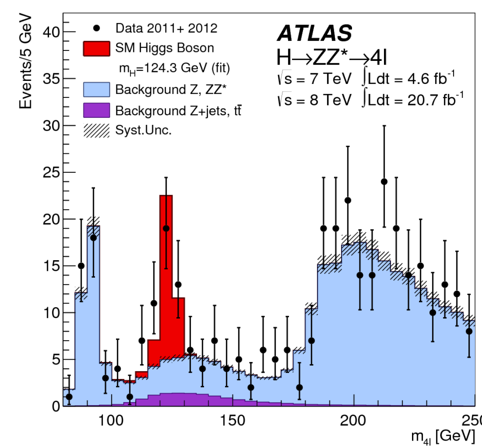
Define a projection function $s : \mathcal{X} \to \mathbb{R}$ mapping observables x to a summary statistics $x' = s(x)$.

Then, approximate the likelihood $p(x|\theta)$ as

$$p(x|\theta) \approx \hat{p}(x|\theta) = p(x'|\theta).$$

From this it comes

$$\frac{p(x|\theta_0)}{p(x|\theta_1)} \approx \frac{\hat{p}(x|\theta_0)}{\hat{p}(x|\theta_1)} = \hat{r}(x|\theta_0, \theta_1).$$

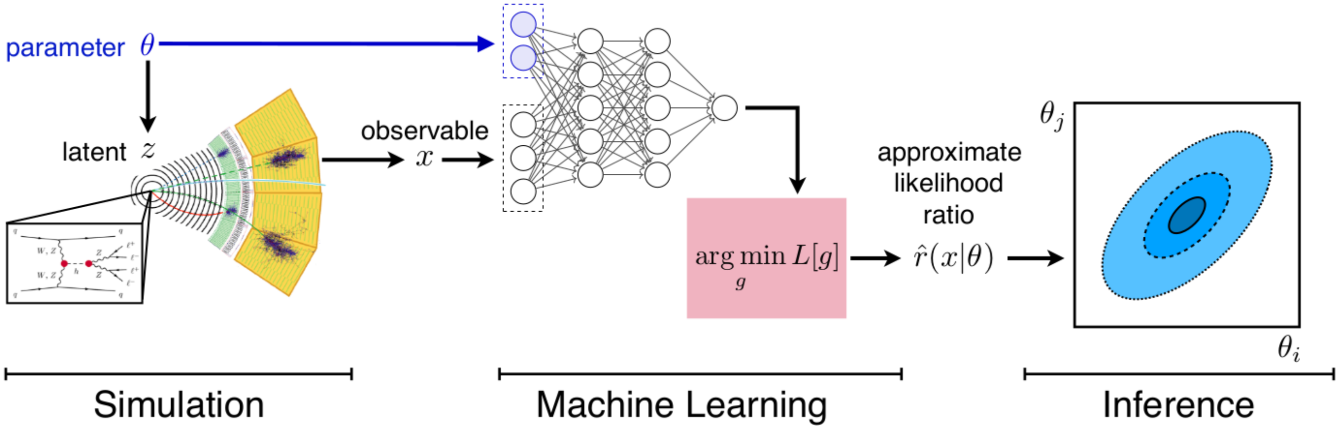


CARL

Supervised learning provides a way to automatically construct s:

- Let us consider a binary classifier $\hat{s}$ (e.g., a neural network) trained to distinguish $x \sim p(x|\theta_0)$ from $x \sim p(x|\theta_1)$.
- $\hat{s}$ is trained by minimizing the cross-entropy loss

$$L[\hat{s}] = -\mathbb{E}_{p(x|\theta)\pi(\theta)}[1(\theta = \theta_0) \log \hat{s}(x) + 1(\theta = \theta_1) \log(1 - \hat{s}(x))]$$



The solution $\hat{s}$ found after training approximates the optimal classifier

$$\hat{s}(x) \approx s^*(x) = \frac{p(x|\theta_1)}{p(x|\theta_0) + p(x|\theta_1)}.$$

Therefore,

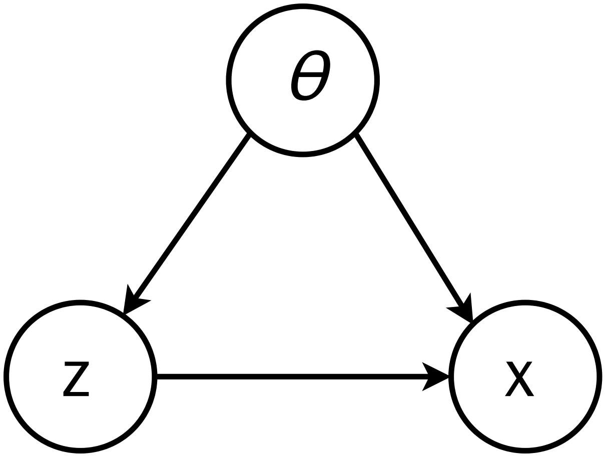
$$r(x|\theta_0, \theta_1) \approx \hat{r}(x|\theta_0, \theta_1) = \frac{1 - \hat{s}(x)}{\hat{s}(x)}.$$

That is, supervised classification is equivalent to likelihood ratio estimation.

Bayesian inference

Bayesian inference usually consists in computing the posterior

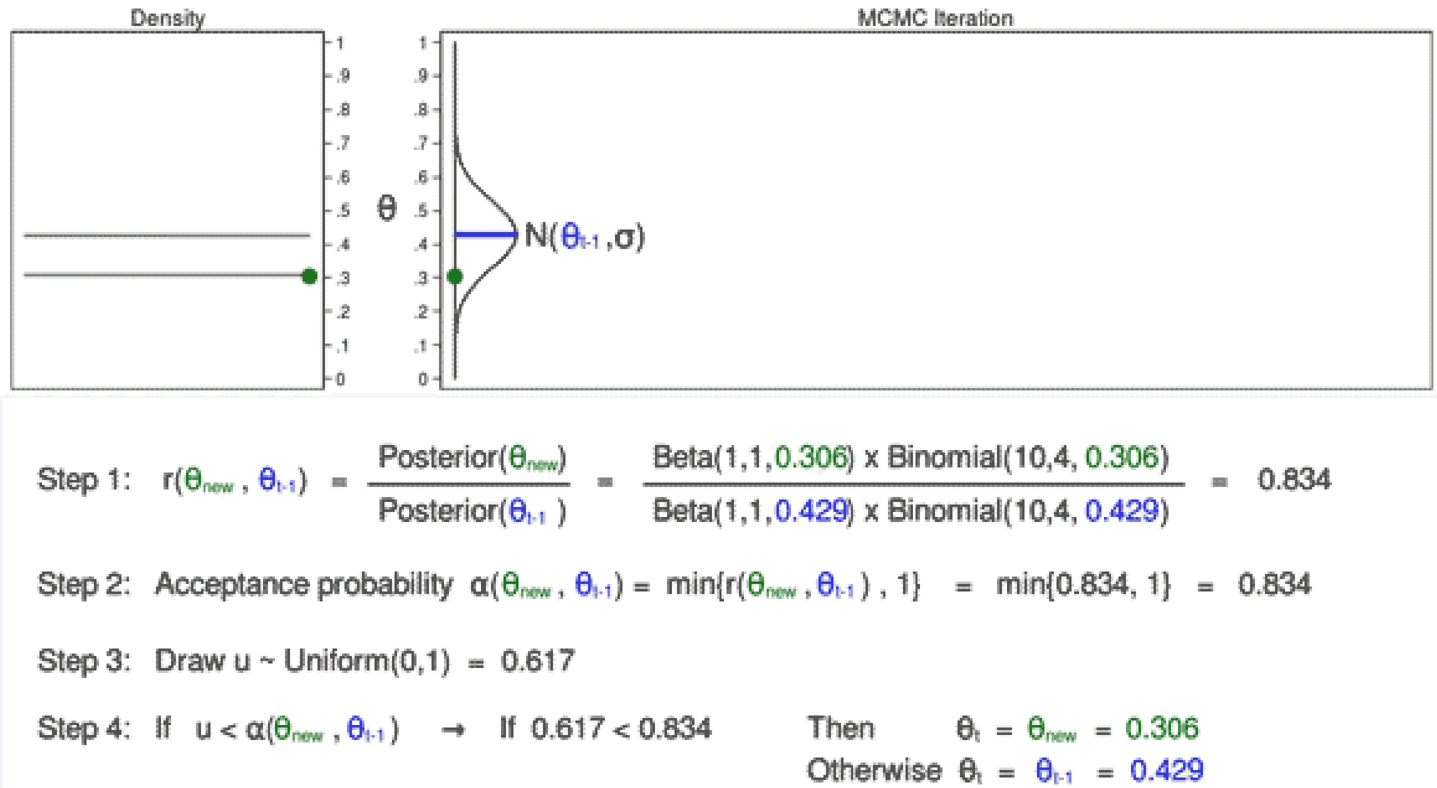
$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$



Doubly intractable in the likelihood-free scenario:

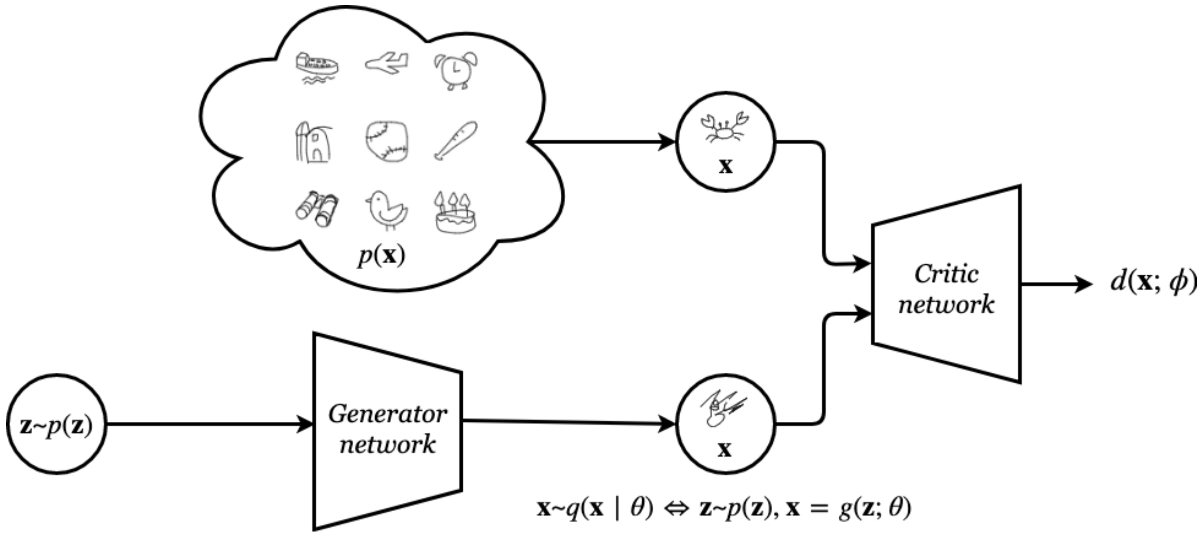
- Cannot evaluate the evidence $p(x) = \int p(x|\theta)p(\theta)d\theta$.
- Cannot evaluate the likelihood $p(x|\theta) = \int p(x,z|\theta)dz$.

Posterior sampling



MCMC algorithms can be made likelihood-free by plugging in the likelihood ratio.

GANs



$$
\begin{aligned}
&\min_\theta \max_\phi \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log(d(\mathbf{x};\phi))\right] + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x}|\theta)}\left[\log(1-d(\mathbf{x};\phi))\right] \\
&= \min_\theta \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x};\theta)+p(\mathbf{x})}\right] + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x};\theta)}\left[\log \frac{q(\mathbf{x};\theta)}{q(\mathbf{x};\theta)+p(\mathbf{x})}\right] \\
&= \min_\theta \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log \frac{1}{1+r(\mathbf{x};\theta)^{-1}}\right] + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x};\theta)}\left[\log \frac{1}{1+r(\mathbf{x};\theta)}\right]
\end{aligned}
$$

Goodfellow et al, 2014 [arXiv:1406.2661].

Optimal compression

The likelihood ratio r relates to the score

$$t(x|\theta_{\text{ref}}) = \nabla_\theta \log p(x|\theta)|_{\theta_{\text{ref}}} = \nabla_\theta r(x|\theta, \theta_{\text{ref}})|_{\theta_{\text{ref}}}.$$

- It quantifies the relative change of the likelihood under infinitesimal changes.
- It can be seen as a local equivalent of the likelihood ratio.

In a small patch around θ_{ref}, we have the approximation

$$p_{\text{local}}(x|\theta) = \frac{1}{Z(\theta)} p(t(x|\theta_{\text{ref}})|\theta_{\text{ref}}) \exp(t(x|\theta_{\text{ref}}) \cdot (\theta - \theta_{\text{ref}}))$$

where the score $t(x|\theta_{\text{ref}})$ are its sufficient statistics. Therefore,

- in the local model the likelihood ratio between θ_0 and θ_1 only depends on the product between the score and $\theta_0 - \theta_1$.
- That is, x can be compressed into a single scalar without loss of power.

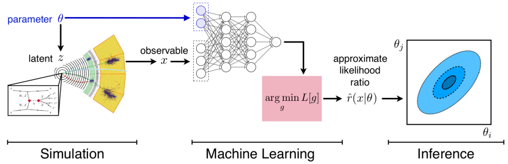
Brehmer et al, 2018 [arXiv:1805.12244].

How to estimate r or t?

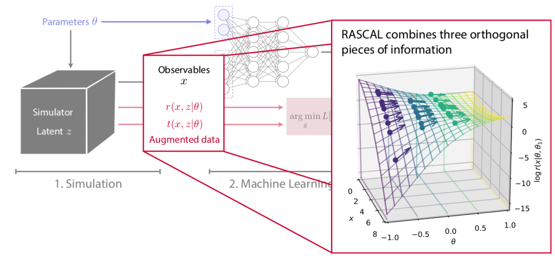
Treat the simulator as a black box

Make use of the inner structure

Learn a proxy for inference

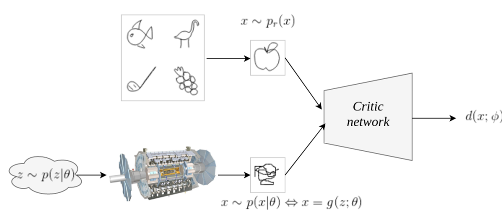


Histograms of observables
Supervised learning
Neural density (ratio) estimation

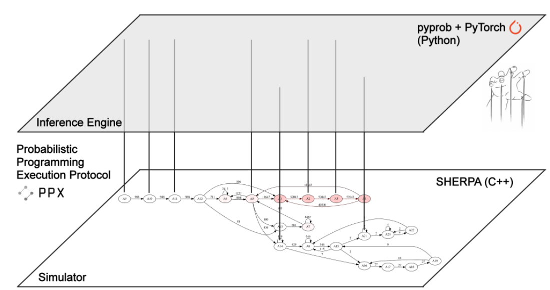


Mining gold from implicit models

Learn to control the simulator



Adversarial variational optimization



Probabilistic programming

Summary

- Much of modern science is based on "likelihood-free" simulations.
- The likelihood-ratio is central to many statistical inference procedures.
- Supervised learning enables likelihood-ratio estimation.
- (Better likelihood-ratio estimates can be achieved by mining simulators.)

Collaborators

Kyle Cranmer

Juan Pavez

Johann Brehmer

Joeri Hermans

The end.

References

- Hermans, J., Begy, V., & Louppe, G. (2019). Likelihood-free MCMC with Approximate Likelihood Ratios. arXiv preprint arXiv:1903.04057.
- Stoye, M., Brehmer, J., Louppe, G., Pavez, J., & Cranmer, K. (2018). Likelihood-free inference with an improved cross-entropy estimator. arXiv preprint arXiv:1808.00973.
- Baydin, A. G., Heinrich, L., Bhimji, W., Gram-Hansen, B., Louppe, G., Shao, L., ... & Wood, F. (2018). Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model. arXiv preprint arXiv:1807.07706.
- Brehmer, J., Louppe, G., Pavez, J., & Cranmer, K. (2018). Mining gold from implicit models to improve likelihood-free inference. arXiv preprint arXiv:1805.12244.
- Brehmer, J., Cranmer, K., Louppe, G., & Pavez, J. (2018). Constraining Effective Field Theories with Machine Learning. arXiv preprint arXiv:1805.00013.
- Brehmer, J., Cranmer, K., Louppe, G., & Pavez, J. (2018). A Guide to Constraining Effective Field Theories with Machine Learning. arXiv preprint arXiv:1805.00020.
- Casado, M. L., Baydin, A. G., Rubio, D. M., Le, T. A., Wood, F., Heinrich, L., ... & Bhimji, W. (2017). Improvements to Inference Compilation for Probabilistic Programming in Large-Scale Scientific Simulators. arXiv preprint arXiv:1712.07901.
- Louppe, G., Hermans, J., & Cranmer, K. (2017). Adversarial Variational Optimization of Non-Differentiable Simulators. arXiv preprint arXiv:1707.07113.
- Cranmer, K., Pavez, J., & Louppe, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. arXiv preprint arXiv:1506.02169.

Variational inference for hierarchical implicit models

Let $q(\mathbf{x}_n)$ be the empirical distribution on the observations $\mathbf{x}$ and consider using it in a "variational joint" $q(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta}) = q(\mathbf{x}_n)q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})$. Now subtract the log empirical $\log q(\mathbf{x}_n)$ from the ELBO above. The ELBO reduces to

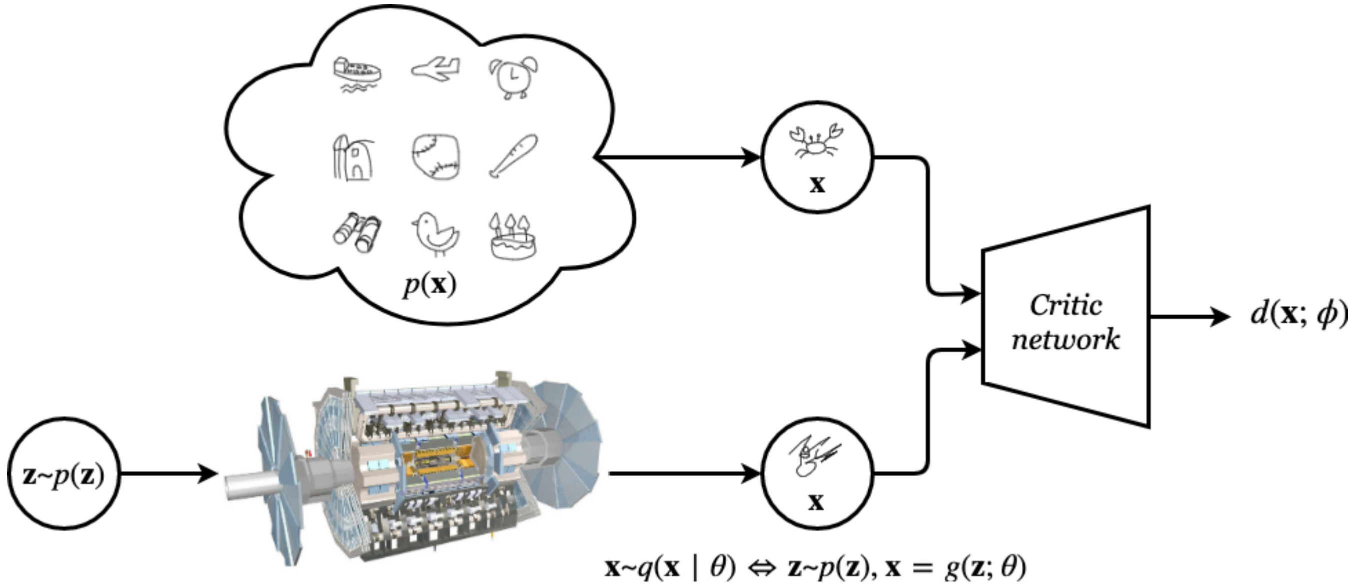
$$\mathcal{L} \propto \mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\boldsymbol{\beta}) - \log q(\boldsymbol{\beta})] + \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{\beta})q(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\beta})}\left[\log \frac{p(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}{q(\mathbf{x}_n, \mathbf{z}_n \mid \boldsymbol{\beta})}\right]. \tag{4}$$

(Here the proportionality symbol means equality up to additive constants.) Thus the ELBO is a function of the ratio of two intractable densities. If we can form an estimator of this ratio, we can proceed with optimizing the ELBO.

We apply techniques for ratio estimation [51]. It is a key idea in GANs [37, 56], and similar ideas have rearisen in statistics and physics [21, 8].

Tran et al, 2017 [arXiv:1702.08896].

Adversarial Variational Optimization



- Replace g with an actual scientific simulator.
- Bypass the non-differentiability with REINFORCE.

Louppe et al, 2017 [arXiv:1707.07113].