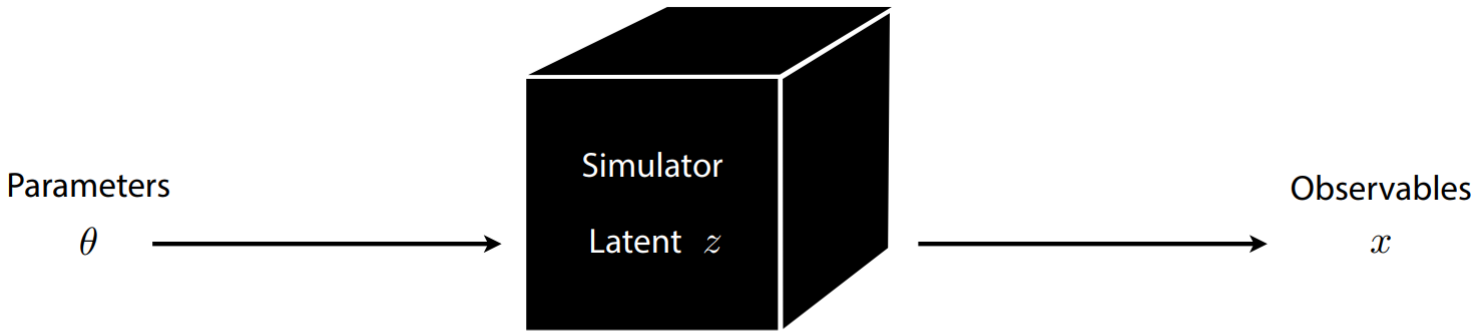
Neural Likelihood Free Inference For GW

Grappa ML Journal Club - Amsterdam

Antoine Wehenkel Gilles Louppe

A few days ago...



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.

$$p(x|\theta) = \underbrace{\iiint}_{\text{intractable}} p(z_p|\theta)p(z_s|z_p)p(z_d|z_s)p(x|z_d)dz_p dz_s dz_d$$

Likelihood ratio

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

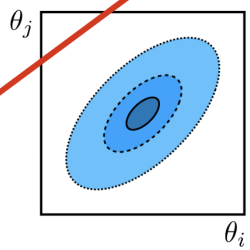
Gilles already explained that

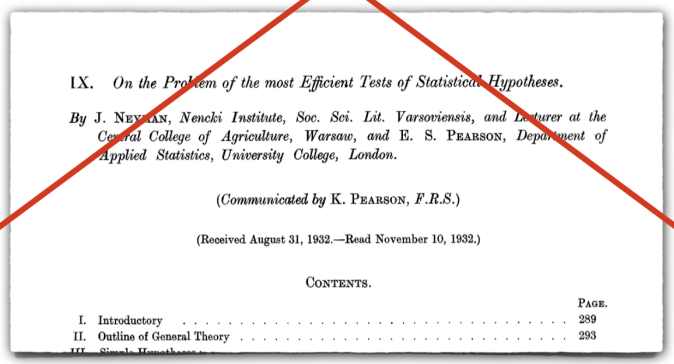
The frequentist (physicist's) way

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.

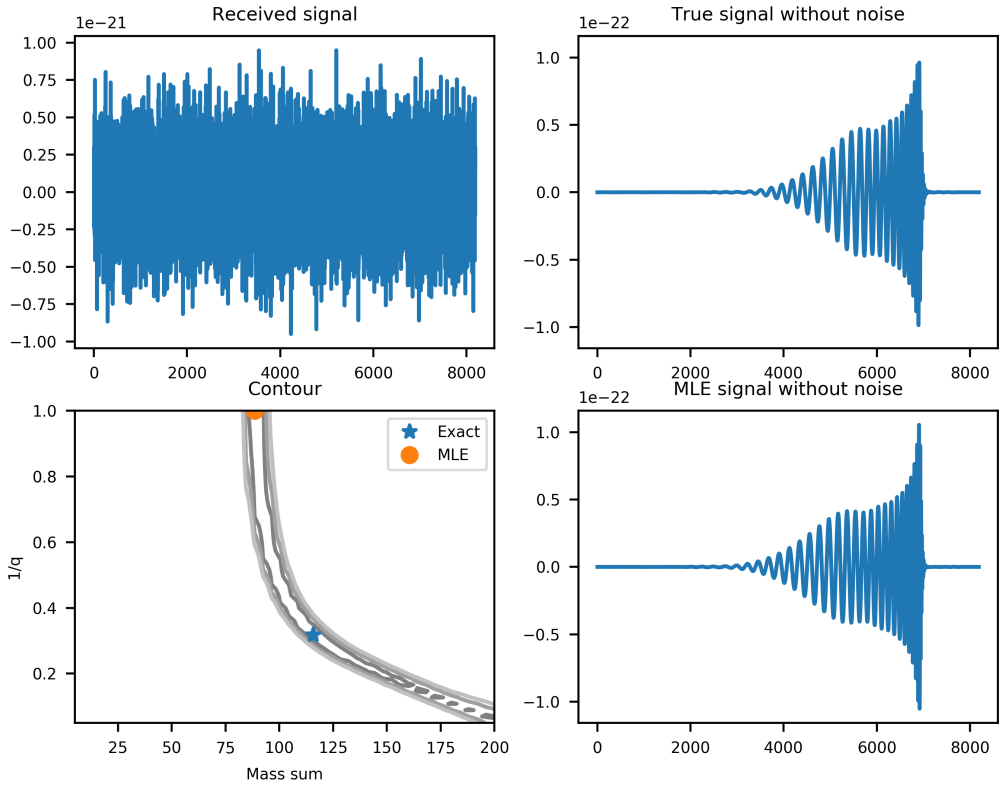
Page.

I. Introductory 289
II. Outline of General Theory 293

14 / 47

Bayesian inference

On-going project

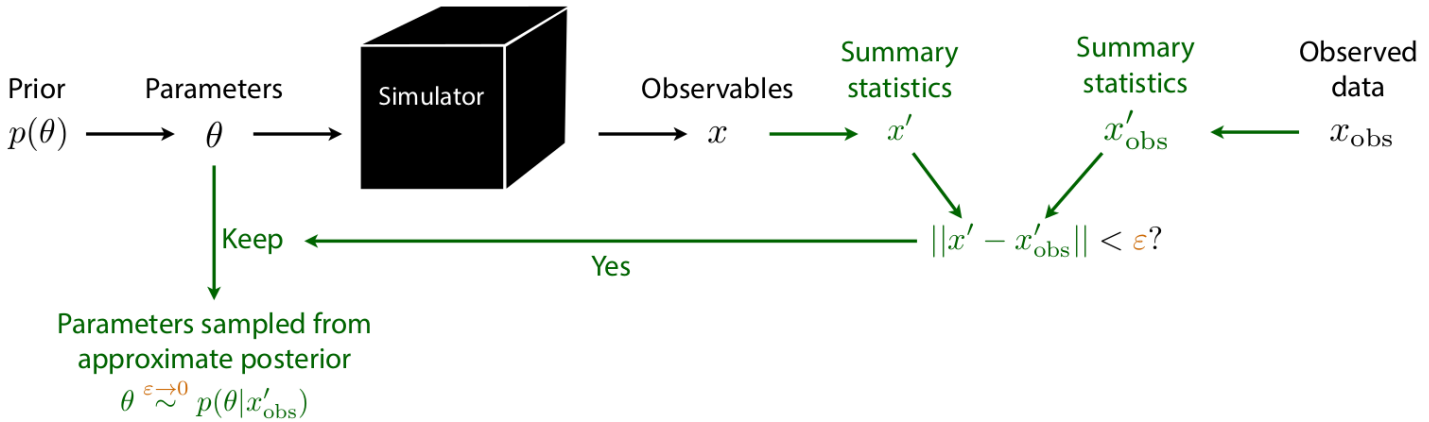


Bayesian inference

We want to evaluate $p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}$, where:

- $p(\theta)$ is your chosen prior.
- $p(x|\theta)$ the unknown/intractable likelihood function:
- $p(x)$ the marginal distribution of the data.

Approximate Bayesian Computation (ABC)



Issues

- How to choose x'? ϵ? $||\cdot||$?
- No tractable posterior.
- Need to run new simulations for new data or new prior.

———

Credits: Johann Brehmer.

Amortizing Bayes

The Bayes rule can be rewritten as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = r(x|\theta)p(\theta) \approx \hat{r}(x|\theta)p(\theta),$$

where $r(x|\theta) = \frac{p(x|\theta)}{p(x)}$ is the likelihood-to-evidence ratio.

Amortizing Bayes

The Bayes rule can be rewritten as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = r(x|\theta)p(\theta) \approx \hat{r}(x|\theta)p(\theta),$$

where $r(x|\theta) = \frac{p(x|\theta)}{p(x)}$ is the likelihood-to-evidence ratio.

The likelihood-to-evidence ratio can be approximated a classifier trained to distinguish $x \sim p(x|\theta)$ from $x \sim p(x)$, hence enabling direct and **amortized** posterior evaluation.

Proof (1)

> Proposition:
>
> An optimal Bayesian classifier approximates the likelihood-to-evidence ratio $\hat{r}(x|\theta) = \frac{p(x|\theta)}{p(x)}$.

Suppose the following classification problem:

H_0	H_1
• $\theta \sim p(\theta)$ (We use the prior)	• $\theta \sim p(\theta)$ (We use the prior)
• $x \sim p(x\|\theta)$ (We use the simulator)	• $x \sim p(x)$ (We sample at random from the simulated x)
• This means that $X = (\theta, x) \sim p(\theta, x)$	• This means that $X = (\theta, x) \sim p(\theta)p(x)$
• $y = 0$	• $y = 1$
• $p(H_0) = 0.5$	• $p(H_1) = 1 - p(H_0) = 0.5$

Proof (2)

> Reminder:
>
> - $d_\phi(X) : \mathbb{R}^n \to [0, 1]$, a discriminator.
> - $\phi = \arg_\phi \max \mathbb{E}_{p(X,y)} \left[\log(d_\phi(X)) 1_{y=1} + \log(1 - d_\phi(X)) 1_{y=0} \right]$ (BCE).

Let's rewrite this:

$$
\begin{aligned}
& \mathbb{E}_{p(X,y)} \left[\log(d_\phi(X)) 1_{y=0} + \log(1 - d_\phi(X)) 1_{y=1} \right] \\
= & \int p(X, y) \left[\log(d_\phi(X)) 1_{y=0} + \log(1 - d_\phi(X)) 1_{y=1} \right] dX dy \\
= & \int \left[p(X|y=0) p(y=0) \log(d_\phi(X)) + p(X|y=1) p(y=1) \log(1 - d_\phi(X)) \right] dX dy \\
= & \frac{1}{2} \int \left[p(X|y=0) \log(d_\phi(X)) + p(X|y=1) \log(1 - d_\phi(X)) \right] dX dy \\
= & F(d_\phi).
\end{aligned}
$$

Proof (3)

When is $F(d)$ maximized?

> F(d) is strictly concave because the (non-null) sum and integral of strictly concave functions is strictly concave. So we have a unique maximum.

Let's derive the quantity $d^\star$ which maximizes F:

$$
\begin{aligned}
0 &= \frac{\partial F(d)}{\partial d}\Big|_{d=d^\star} \\
&= \frac{\partial}{\partial d}\frac{1}{2}\int \left[p(X|y=0)\log(d(X)) + p(X|y=1)\log(1-d(X))\right] dX\,dy\Big|_{d=d^\star} \\
&= \frac{1}{2}\int \left[\frac{p(X|y=0)}{d(X)} - \frac{p(X|y=1)}{1-d(X)}\right] dX\,dy\Big|_{d=d^\star}.
\end{aligned}
$$

Proof(4)

When is $F(d)$ maximized?

Finding a $d^\star$ cancelling $\left[\frac{p(X|y=0)}{d(X)} - \frac{p(X|y=1)}{1-d(X)}\right]$ is sufficient:

$$\left[\frac{p(X|y=0)}{d^\star} - \frac{p(X|y=1)}{1-d\star}\right] = 0 \rightarrow d^\star = \frac{p(X|y=0)}{p(X|y=0)+p(X|y=1)}.$$

We can find the likelihood ratio between the two classes as:

$$r(X) = \frac{d^\star(X)}{1-d^\star(X)} = \frac{p(X|y=0)}{p(X|y=1)}$$

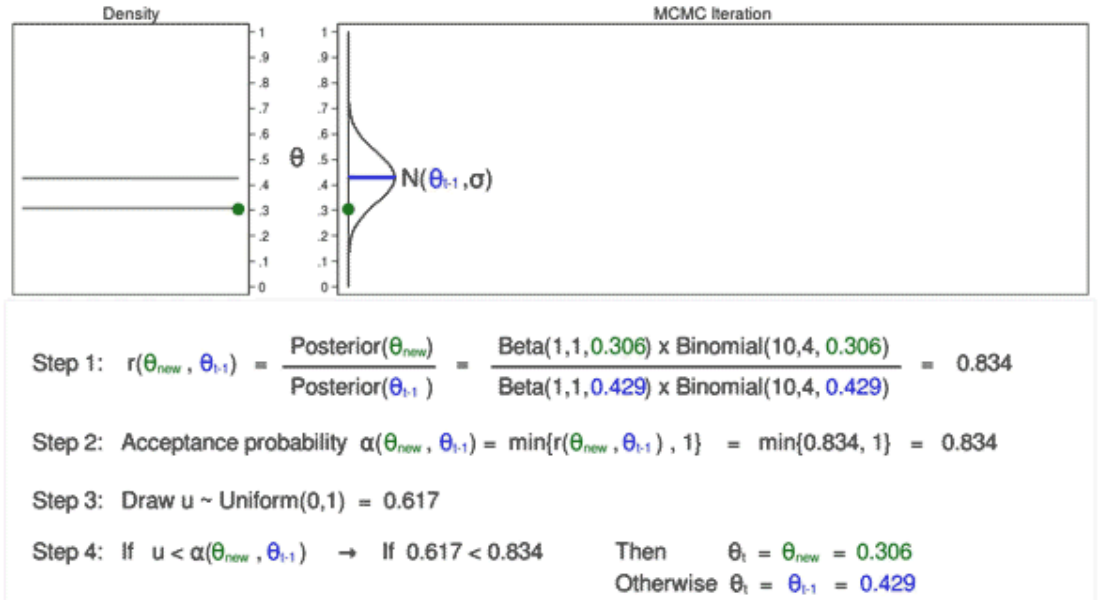
Plugging back our definition of H_0 and H_1 we obtain the likelihood-to-evidence ratio:

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

Bayesian inference

We now have access to the likelihood-to-evidence ratio $r(x, \theta) = \frac{p(x|\theta)}{p(x)}$, which can be weighted by the prior value to obtain the posterior.

- $r(x, \theta)p(\theta) = \frac{p(x|\theta)p(\theta)}{p(x)} = p(\theta|x)$
- This quantity can be used to sample from the posterior (MCMC) and draw credible intervals.



Back to frequentist analysis

We now have acccess to the likelihood ratio between any pair of parameters:

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

.

Wilks theorem

Consider the test statistic

$$q(\theta) = -2 \sum_x \log \frac{p(x|\theta)}{p(x|\hat{\theta})} = -2 \sum_x \log r(x|\theta, \hat{\theta})$$

for a fixed number N of observations $\{x\}$ and where $\hat{\theta}$ is the maximum likelihood estimator.

When $N \to \infty$, $q(\theta) \sim \chi_2$.

GW: In practice

Data

- $x \in \mathbb{R}^{8192}$ are timeseries (2 seconds of signal) and $\theta \in \mathbb{R}^2$ the parameters of interest (the two merger's masses).
- 100k generated data (H_0).

Classifier

A deep neural network made of the combination between $1D$ convolutions, residual connections and hypernetworks.

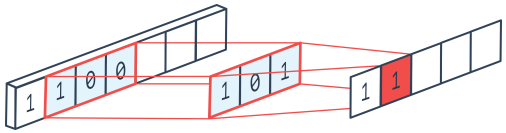
Why?

The marginalization with respect to a subset of parameters is amortized along the NN training. Thus once trained, the network is very fast to evaluate the marginal $p(m_1, m_2|x) = \int p(m_1, m_2, \chi_{eff}, d, \phi_t, ...) d\chi_{eff} dd d\phi_t ...$

GW: Neural Architecture

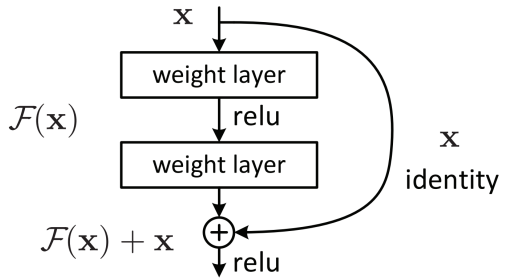
1D convolutions

Commonly used to process timeseries with a-priori known size. They are faster and easier to train than recurrent neural networks.



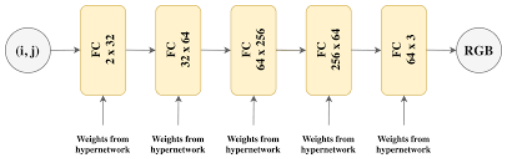
Residual connections

Commonly used for large neural networks in order to resolve the vanishing gradient problem.



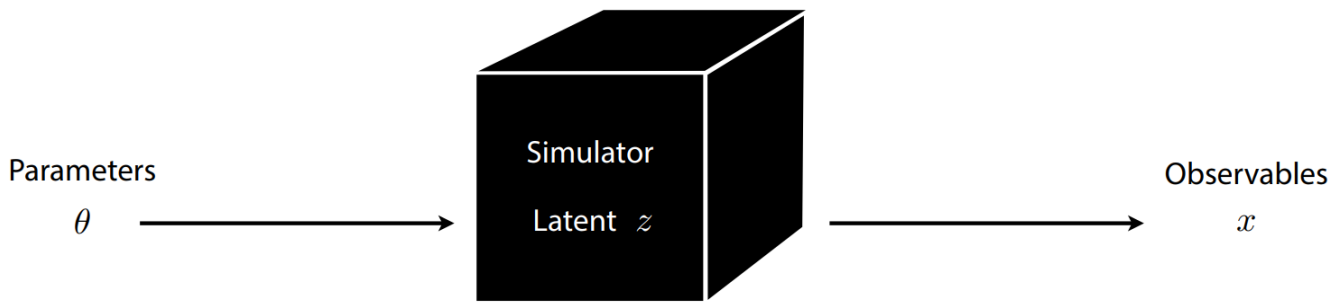
Hypernetworks

A "hypernetwork" takes as input the parameters θ and outputs the weights of the convolution filters. The purpose is to modulate the way the waveform is processed depending on the parameters value.



Summary

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



Thank you for the invitation!

References

- Hermans, J., Begy, V., & Louppe, G. (2019). Likelihood-free MCMC with Approximate Likelihood Ratios. arXiv preprint arXiv:1903.04057.

The end.