

4th COMCHA SCHOOL

Modern Bayesian Inference

Bryan Zaldivar

IFIC

14.04.2026

Outline

- * Basic concepts
(Bayesian approach, Uncertainty, Predictions)
- * Markov Chain Monte Carlo
(Metropolis-Hastings, HMC)
- * Variational Inference
(Kullback-Leibler, ELBO)
- * Simulation-based Inference
(NRE, NPE)

Basic Concepts

① Basic Concepts

Frequentist Approach

Probability $\Leftrightarrow$ Relative frequency

$$P(A) = \lim_{N \rightarrow \infty} \frac{\text{\# occurrences of } A}{N}$$

event $\rightarrow$

Examples:

- Getting "5" in a die
- Measuring 1 event in a neutrino experiment

Bayesian Approach

Probability $\Leftrightarrow$ Degree of believe in that a given hypothesis is TRUE

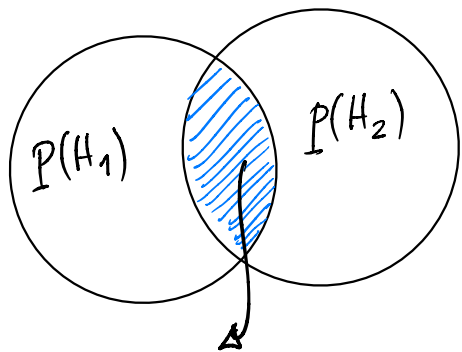
$$P(H) \rightarrow \text{hypothesis}$$

Note: No experiment repetition needed

Examples:

- What is the probability of for the Higgs mass to be inside a given interval?
- What is the most probable value of the Dark Matter mass, given some dataset?

① Basic Concepts



$$P(H_1 \cap H_2) \Leftrightarrow P(H_1, H_2)$$

Properties

- $P(H) \geq 0 \quad \forall H$
- $P(H_1) + P(H_2) = 1$
- etc

Joint
probability

fixed ←
random ← Prob. of H_1 GIVEN H_2

$$\begin{aligned} P(H_1, H_2) &= P(H_1 | H_2) P(H_2) \\ &= P(H_2 | H_1) P(H_1) \end{aligned}$$

↓ **BAYES
THEOREM**

$$P(H_1 | H_2) = \frac{P(H_2 | H_1) P(H_1)}{P(H_2)}$$

① Basic Concepts

Typical use of Bayes Theorem

- Dataset $\mathcal{D} = \{\vec{x}_i, y_i\}_{i=1}^N$
- Fitting function $f(\vec{x}; \vec{\theta})$

e.g. parameters of
our physics model

- Prior knowledge $P(\vec{\theta})$

↑ information about $\vec{\theta}$
before "seeing" $\mathcal{D}$

$$P(\vec{\theta} | \mathcal{D}) = \frac{P(\mathcal{D} | \vec{\theta}) P(\vec{\theta})}{P(\mathcal{D})}$$

"Posterior distrib." ↑
"Likelihood" ↑
"Prior" ↑
"Evidence" ↑

{ $p(\cdot)$: DENSITY DISTRIB.
 $\int p(x) dx = 1$

① Basic Concepts (Uncertainty)

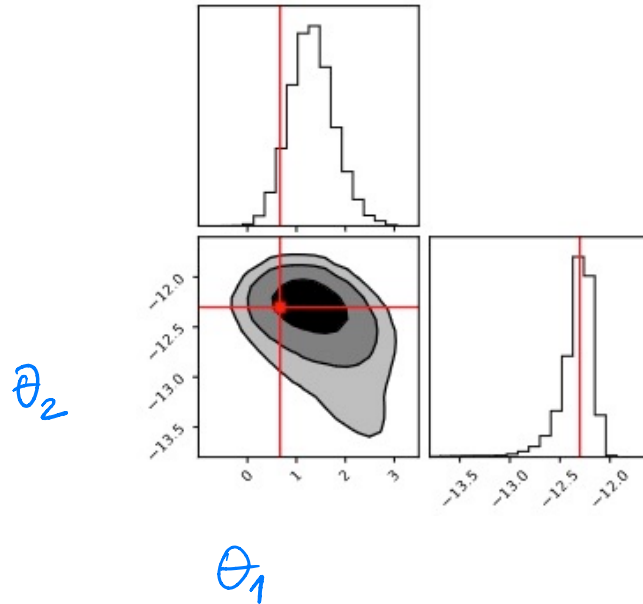
$$1 - \beta = \int_{\theta_{\text{Low}}}^{\theta_{\text{High}}} p(\theta | \text{data}) d\theta \quad \left\{ \begin{array}{l} [\theta_{\text{Low}}, \theta_{\text{High}}] \\ \text{Interval containing the true value} \\ \text{with "degree of belief"} = 1 - \beta \end{array} \right. \quad \xrightarrow{\text{"credible interval"}}$$

Note the difference with frequentist:

"If the experiment were repeated many times, this interval would cover the true value a fraction $1 - \beta$ of times"

① Basic Concepts (Predictions)

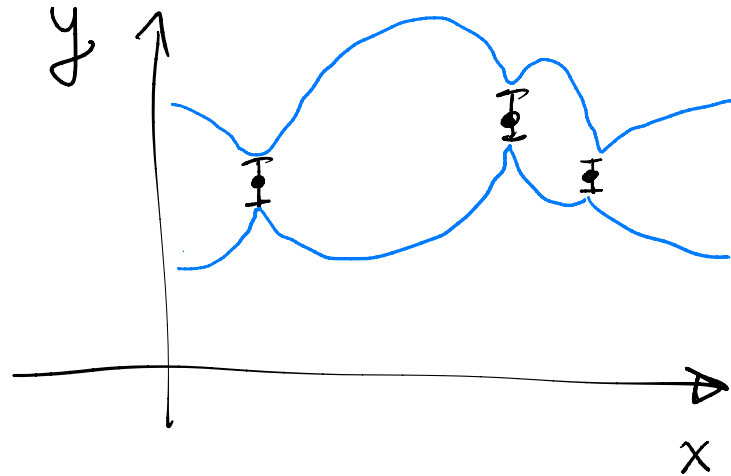
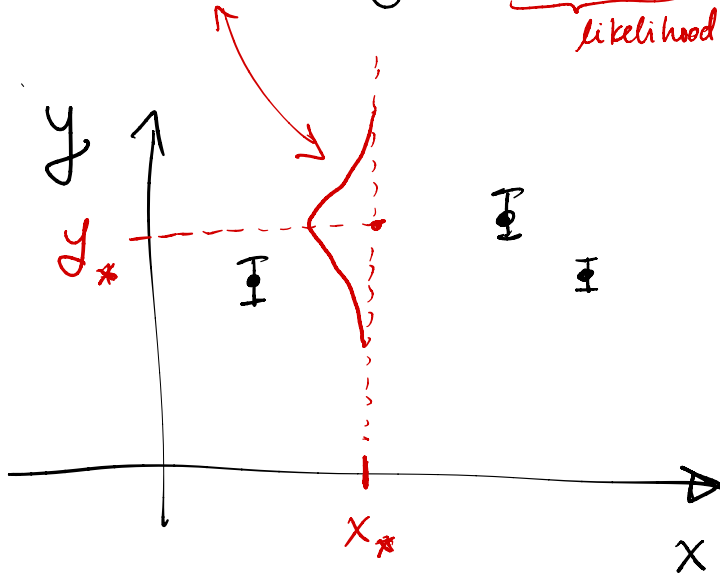
Ⓐ Report The posterior distribution
(typical in physics applications)



① Basic Concepts (Predictions)

② Report the "predictive distribution" *more common in the statistics community*

$$P(y_* | x_*, \mathcal{D}) = \int d\theta \underbrace{P(y_* | x_*, \theta)}_{\text{likelihood at a new point } (x_* \notin \mathcal{D})} P(\theta | \mathcal{D})$$



① Basic Concepts

* The primary goal of Bayesian Inference is to estimate the posterior.

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\underline{\underline{P(D)}}}$$

Typically this normalization is intractable!

Methods

② Markov Chain Monte Carlo

② MCMC (General Principles)

* It's a sampling method $\rightarrow$ typically it suffices to have samples of the posterior

$$\mathbb{E}[f] = \int d\theta p(\theta|D) f(\theta) \approx \frac{1}{N} \sum_{i=1}^N f(\theta_i) \quad \text{with } \theta_i \sim p(\theta|D)$$

* It builds a "chain" of samples $\{\bar{\theta}_1, \bar{\theta}_2, \dots, \bar{\theta}_N\}$

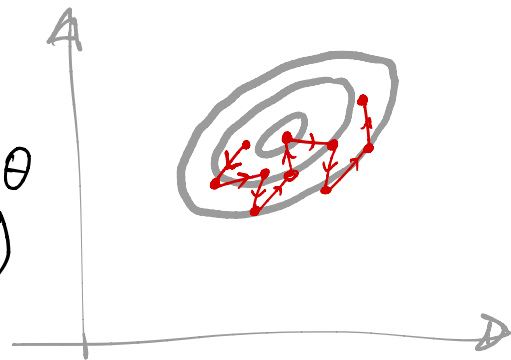
* It uses the Markov property $\rightarrow$ "proposal distribution"

$$q(\theta_i | \theta_{i-1}, \theta_{i-2}, \dots, \theta_1) = q(\theta_i | \theta_{i-1})$$

② MCMC (General Principles)

* The chain should be Ergodic

- It should be able to reach any state θ
- Aperiodic (Avoid $\theta_1 \leftrightarrow \theta_2$ oscillations)



* The chain should converge to the target $\pi(\theta)$ as its stationary distribution

$$P(\theta \rightarrow \theta') = q(\theta' | \theta) \alpha(\theta, \theta')$$

→ acceptance probability for this move

$$\pi(\theta) P(\theta \rightarrow \theta') = \pi(\theta') P(\theta' \rightarrow \theta) \Rightarrow \text{"Detailed balance"}$$

② MCMC (Metropolis-Hastings)

* Input: $\tilde{\pi}(\theta)$, $q(\theta'|\theta)$, θ_0 , N { $\pi(\theta) \equiv \frac{1}{Z} \tilde{\pi}(\theta)$ }

* For $t=0, \dots, N-1$

- $\theta' \sim q(\theta'|\theta_t)$

- $\alpha(\theta_t, \theta') = \min \left(1, \frac{\tilde{\pi}(\theta') q(\theta_t|\theta')}{\tilde{\pi}(\theta_t) q(\theta'|\theta_t)} \right)$

- $u \sim \text{unif}[0, 1]$

If $u < \alpha(\theta_t, \theta') : \theta_{t+1} = \theta' \text{ (accept)}$

Else: $\theta_{t+1} = \theta_t \text{ (reject)}$

Return $\{\theta_1, \theta_2, \dots, \theta_N\}$

② MCMC (Hamiltonian Monte Carlo)

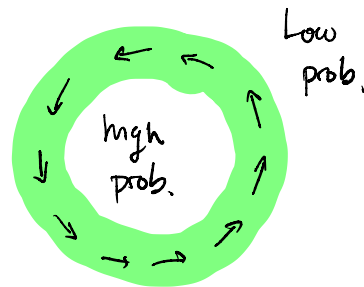
Motivation and key features

* Better exploration than MH

random walk behavior

* Analogy with a satellite in orbit

$\vec{\theta} \rightarrow \{\vec{\theta}, \vec{p}\}$ auxiliary "momentum" variables



* Hamiltonian dynamics

target $\tilde{\pi}(\vec{\theta}) \rightarrow \pi(\vec{\theta}, \vec{p}) = \pi(\vec{p}) \cdot \tilde{\pi}(\vec{\theta})$

Hamiltonian $H \equiv -\log \pi(\vec{\theta}, \vec{p}) = K(\vec{p}) + V(\vec{\theta})$

"kinetic energy" "potential energy"

$$\left. \begin{aligned} \frac{d\vec{\theta}}{dt} &= \frac{\partial H}{\partial \vec{p}} \\ \frac{d\vec{p}}{dt} &= -\frac{\partial H}{\partial \vec{\theta}} \end{aligned} \right\}$$

- Hamilton eqs
- Contain gradients of target density
- Deterministic proposal for θ_{t+1} [BEFORE ACCEPT/REJECT]

② MCMC (Hamiltonian Monte Carlo)

* $\vec{p} \sim \mathcal{N}(0, M) \Rightarrow K(\vec{p}) = \frac{1}{2} \vec{p}^T \cdot \vec{M}^{-1} \vec{p}$

* Ideally: Since energy is conserved $\vec{\Theta}, \vec{p} \rightarrow \vec{\Theta}', \vec{p}'$
 $\Rightarrow$ HMC would give 100% acceptance!!

* In Practice: Numerical errors in solving eqs.
 $\Rightarrow$ we need good integrators ["symplectic"]

✗ • Runge-Kutta

✗ • Euler

✓ • Leap frog

✓ • 4th-order (Yoshida, etc)

② MCMC (Hamiltonian Monte Carlo)

* Input: $\tilde{\pi}(\theta)$, M , θ_0 , N , ϵ , L
Leapfrog

* For $t=0, 1, \dots, N-1$

- $\vec{p}_t \sim N(\theta, M)$

- $\{\vec{\theta}', \vec{p}'\} = \text{Leapfrog}(\vec{\theta}_t, \vec{p}_t, \epsilon, L)$

- $\alpha(\vec{\theta}_t, \vec{\theta}') = \min\left\{1, \exp\left[-H(\vec{\theta}', \vec{p}') + H(\vec{\theta}_t, \vec{p}_t)\right]\right\}$

- Accept/Reject # idem to MH

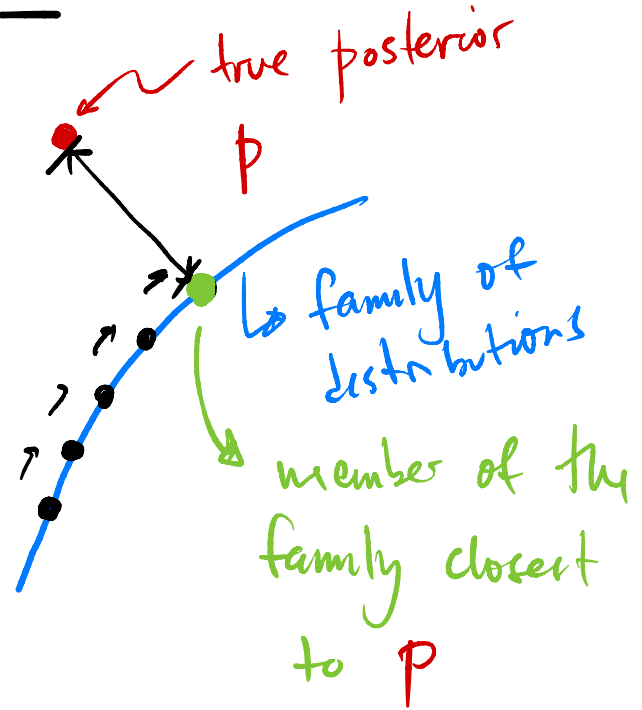
Return $\{\theta_1, \theta_2, \dots, \theta_N\}$

Methods

④ Variational Inference

③ Variational Inference

- * Optimisation method
- * Approximate inference
- * Existing research beyond parametric approach
- * Arguably The most active research field in Stats nowadays



③ Variational Inference

How $q(\theta) \rightarrow \pi(\theta)$?

* Information-theoretic guideline

$$KL[\pi \| q] = \int d\theta \pi(\theta) \log \frac{\pi(\theta)}{q(\theta)}$$

Relative entropy
↓
Excess information-content when using q instead of π

↪ "Kullback-Leibler Divergence" [FORWARD]

$$KL[q \| \pi] = \int d\theta q(\theta) \log q(\theta) / \pi(\theta)$$

↪ "Reverse KL"

$$KL \geq 0$$

$$KL = 0 \text{ iff } q = \pi$$

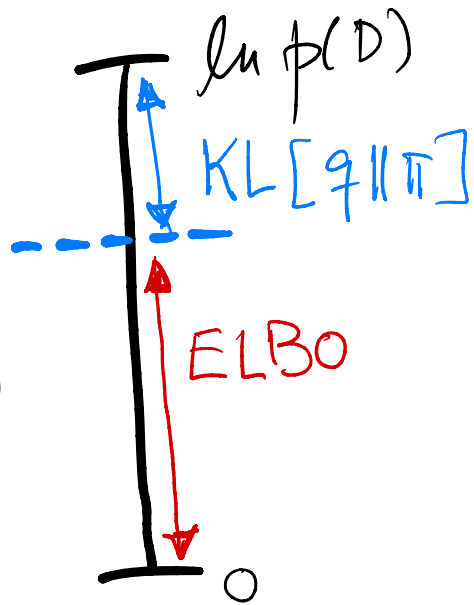
③ Variational Inference

$$KL[q \parallel \pi]$$

* $\pi(\theta) = p(\theta | D)$ $\leadsto$ target is the posterior

* $q = q_{\alpha}(\theta)$ $\leadsto \alpha$: optimisable parameters

$$\begin{aligned} KL[q \parallel \pi] &= \int d\theta q_{\alpha}(\theta) \log \frac{q_{\alpha}(\theta) p(D)}{p(D, \theta)} \\ &= \underbrace{- \int d\theta q_{\alpha}(\theta) \log \frac{p(D, \theta)}{q_{\alpha}(\theta)}}_{\text{"ELBO": Tractable}} + \log p(D) \end{aligned}$$



③ Variational Inference

$$KL[q \parallel \pi]$$

* Input : $\tilde{\pi}(\theta) = p(D|\theta)p(\theta)$

• $q_2(\theta)$ • α_0 # initial parameters' values

• N # number of iterations

* For $t = 0, \dots, N-1$

• M # number of samples

• $\theta_i \sim q_{\alpha_t}(\theta)$

• $\nabla_{\alpha_t} KL \approx \frac{1}{M} \sum_{i=1}^M \nabla_{\alpha_t} [\log q_{\alpha_t}(\theta_i) - \log \tilde{\pi}(\theta_i)]$

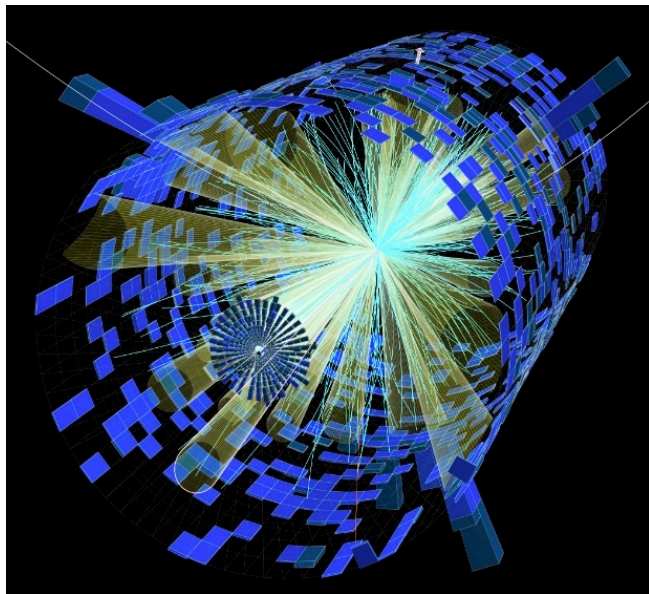
• $\alpha_{t+1} = \alpha_t - \nabla_{\alpha_t} KL$ # gradient descent

Methods

④ Simulation-based Inference (SBI)

④ SBI (Motivation)

* Highly complex experiments, highly dimensional data



* Binned analysis typically insufficient



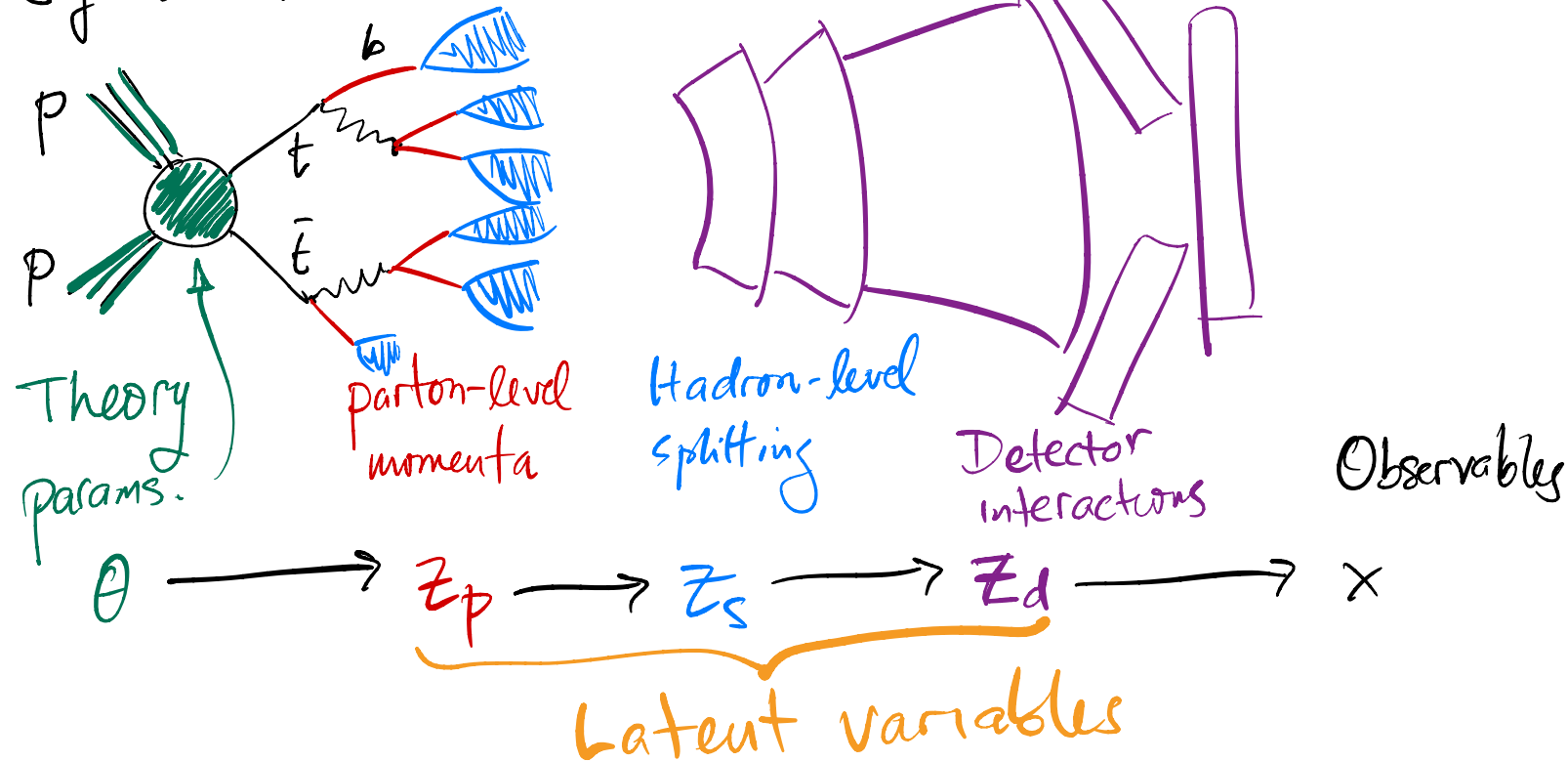
$$p(x|\theta)$$

Unbinned
Likelihood

* True likelihood is commonly intractable

④ SBI (Motivation)

e.g. at LHC:



④ SBI (Motivation)

$$p(x|\theta) = \int p(x|z_d) p(z_d|z_s) p(z_s|z_p) p(z_p|\theta) dz_p dz_s dz_d$$

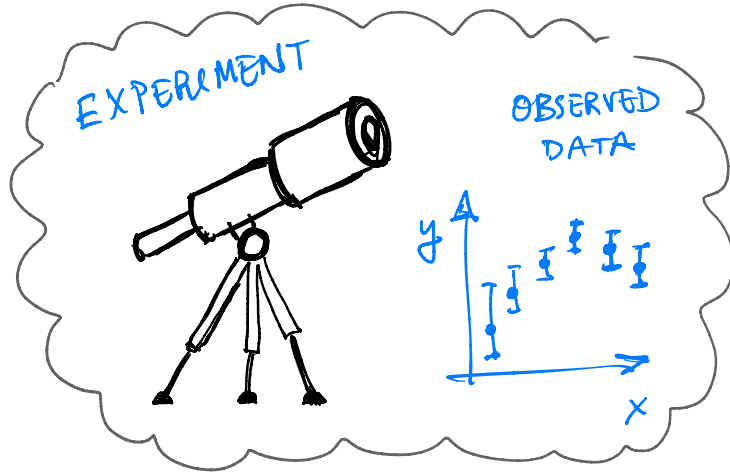
$\Rightarrow$ Intractable

* Analyses @ LHC don't compute this

$\Rightarrow$ Low dimensional modeling $\Rightarrow$ Typically good enough

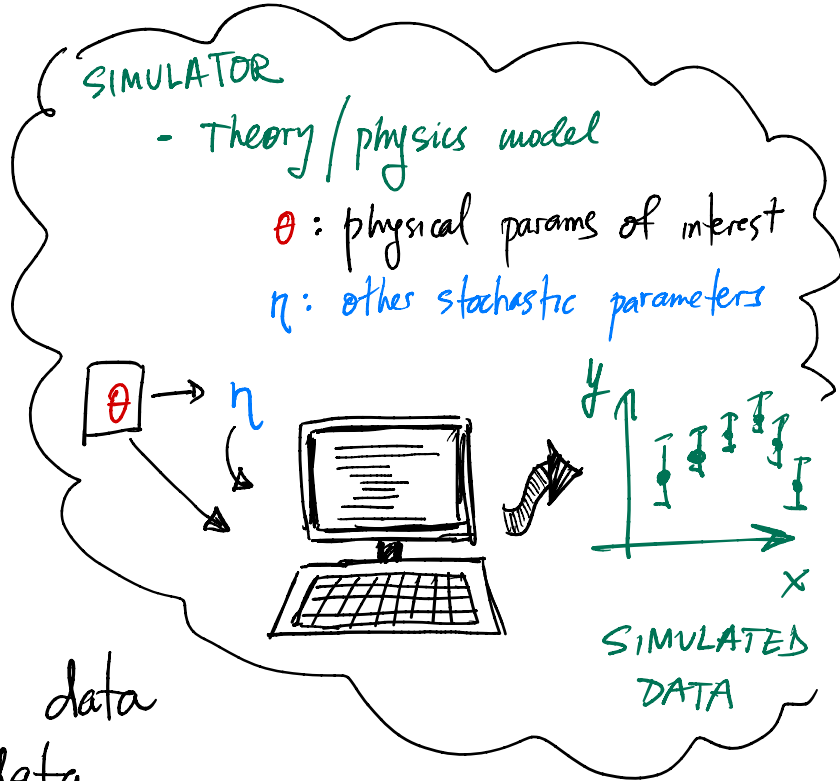
$\hookrightarrow$ Arguably insufficient for complicated signals

④ SBI (General Principles)



Aim: Solving the inverse problem

How? Selecting sets of **SIMULATED** data compatible with **OBSERVED** data



④ SBI (General Principles)

* Decades old idea (see "Approximate Bayesian Computation")
ABC ~ 80's)

with computational disadvantages

* Meteoric revival in the last decade Thanks to ML

$$\begin{matrix} p(\theta | \text{OBS. DATA}) \\ \text{or} \\ p(\text{OBS DATA} | \theta) \end{matrix} \approx f \left(\begin{array}{c} \bullet \quad \bullet \quad \bullet \\ \bullet \quad \bullet \quad \bullet \\ \bullet \quad \bullet \quad \bullet \end{array} \right)$$

Inputs $\bullet$
Output $\bullet$

depend on the
SBI flavor

$$f: \frac{p(\text{DATA} | \theta)}{p(\text{DATA})} \rightarrow \text{Neural Ratio Estimation}$$

$$f: q_w(\theta) \rightarrow \text{Neural Density Estimation}$$

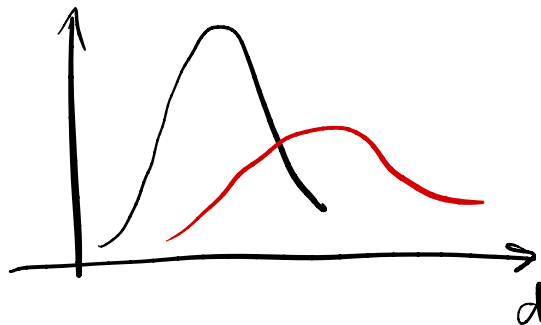
$\hookrightarrow$ VARIATIONAL DISTRIBUTION

④ SBI (NRE)

* Binary classifiers approximate Likelihood ratios

$$d_0 \sim P_0(d)$$

$$d_1 \sim P_1(d)$$



$$D = \{d_i, z_i\}$$
$$z = \{0, 1\}$$

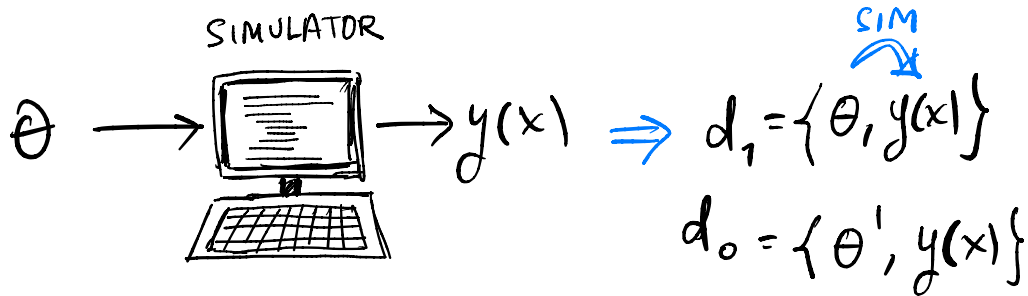
Classifier output:

$$p_{\alpha_*}(z=1|d) = \frac{p(d|z=1)p(z=1)}{\sum_k p(d|z=k)p(z=k)} = \frac{1}{1 + \frac{p(d|z=0)}{p(d|z=1)}}$$

(if equal priors)

e.g. a Neural network

④ SBI (NRE)



$$p(\theta, y) = p(y|\theta) p(\theta)$$

$$p(\theta', y) = p(y) p(\theta')$$

$$\frac{p(\theta, y)}{p(\theta', y)} = \frac{p(y|\theta)}{p(y)} = r_\alpha(y; \theta)$$

$$\boxed{p(\theta | y_{\text{obs}}) = r_{\alpha^*}(y_{\text{obs}}; \theta) p(\theta)}$$

Training data for the Neural network with parameters α

Class 1:

$$\{d_1^{(i)}, z_1^{(i)}\}$$

$$z_1^{(i)} = 1 \quad \forall i$$

Class 0:

$$\{d_0^{(i)}, z_0^{(i)}\}$$

$$z_0^{(i)} = 0 \quad \forall i$$

④ SBI (Neural Posterior Estimation)

* Data is only "class 1": $\mathcal{D} = \{ \theta_i, y_i(x) \}_{i=1}^N$

* Target:

$p(\theta | y_{\text{obs}}) \approx q_{\alpha_*}(\theta | y_{\text{obs}}) \Rightarrow$ variational distribution [Neural net-based]
(parametric form)

* Cost function

$$\mathbb{E}_{p(\mathcal{D})} [\text{KL}[p \parallel q]] \Rightarrow C(\alpha) = -\frac{1}{N} \sum_{i=1}^N \log q_{\alpha}(\theta_i | y_i)$$

④ SBI (cont...)

- * Neural Likelihood Estimation also possible (similar to NPE) and nowadays gaining popularity
- * NRE conceptually promising (recasting density estimation as a classification problem is a great idea) but in practice is less practical than NPE
- * The model inside SBI can be anything (modern implementations include Diffusion Models or Transformers)
- * Open issues $\Rightarrow$ very active area of research !

BAYESIAN INFERENCE

MCMC

MH

HMC

VARIATIONAL
INFERENCE

ELBO

Forward
KL

SBI

NPE

NRE

NLE