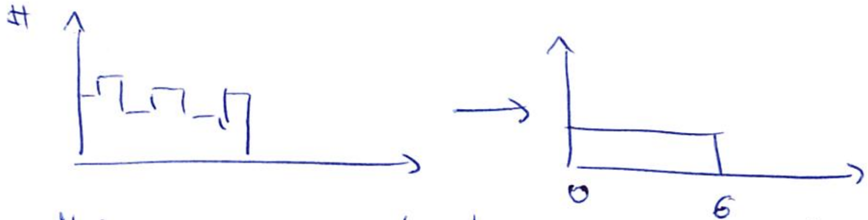
Field-level inference & SBI

- How many working on FLI?
- " " " " SBI?

stats are the intersting part → mostly stats & theory
exercices of flow charts + pseudo code

Dice example

typically: roll dice: Random process with different outcome due to noise
Repeat many times



inference: know something we cannot observe from data we observe

many dice → roll one: it gives a 1
what dice did I roll?
repeating doesn't help you.
you will not observe the truth

add more 6-face dice → more likely that I roll a 6-face? Prior knowledge

Statistics

$P(A,B)$ joint prob.

if A, B indep → $P(A,B) = P(A) \cdot P(B)$

if A, B dep → $P(A,B) = P(A)\, P(B|A)$

$= P(B) \underbrace{P(A|B)}_{\text{conditional}}$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ Bayes Theorem

Bayes vs freq: what they mean by prob (all agree on B. Th.)

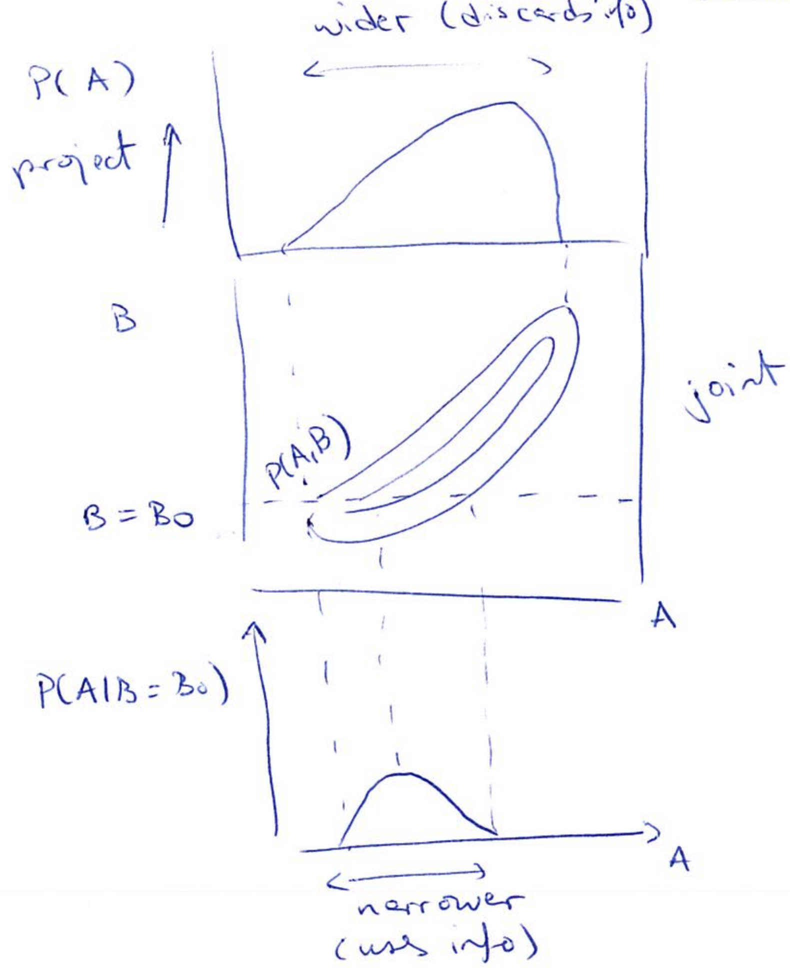
$P(A, B)$ joint

$P(A \mid B)$ conditional $\rightarrow$ uses info

$P(A)$ marginal $\rightarrow$ discards info

$$P(A) = \int P(A, B) \, dB$$

integrate over all possible values of B.



$P\left(\begin{array}{c}\text{what we want}\\ \text{to know}\end{array} \middle| \begin{array}{c}\text{what we}\\ \text{know}\end{array}\right) = \ ?$

Inference

$$P(\vec{x}) = \int P(\vec{x}, \vec{\theta})\, d^n\theta = \int P(\vec{x}|\vec{\theta})\, P(\vec{\theta})\, d^n\theta = \mathcal{E}$$

$P(\vec{x})$: sampling distribution (nature)

outliers — how likely data is

$\vec{\theta}$: param (human)

$P(\vec{x}|\vec{\theta})$: Theory (we can compute) Likelihood

$P(\vec{\theta})$: what are reasonable values of param. Prior

$\mathcal{E}$: evidence prob of data at all.

Posterior:

$$\boxed{P(\vec{\theta}|\vec{x}) = \frac{\mathcal{L}(\vec{x}|\vec{\theta})\,\pi(\vec{\theta})}{\pi(\vec{x})}}$$

$P(\vec{\theta}|\vec{x})$: what are good values of $\vec{\theta}$ given data

Bays Th.

likelihood : model of nature producing some data.

prior : price to introduce params to nature (what are ~~good~~ reasonable values?)

posterior : what are credible values of θ given $\vec{x}$ want to reverse the process & see how nature produced the data.

evidence : falsify ourselves.

Gaussian
many sources of randomness (CLT)

Poisson
counting

Gamma
only positive (areas)

f_D
physical law

Baysian hierarchical model

slide

noise detector
manufacturer

or

marbles : 3g 2.9g 3.1g 2g

size
material
wearing
impurities
manufacturer
noisy detector.

many sources of uncertainty!

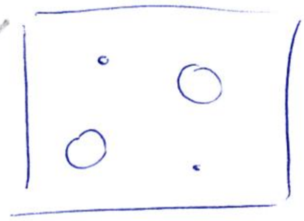
↓

Baysian hierarchical Model (BHM)

$\boxed{\text{BHM}}$ accounts for all sources of uncertainties and propagate the uncertainties from one step to the next
↳ submodel.

Analytical example

slide



colours → T
sizes → A
noisy detector

2 sources of variability: universe, detector

$L = \sigma_T A T^4$

↑ want to infere σ_T → $P(\sigma_T \mid \hat{L})$

observed estimator of L because of noise (every time different)

collect all random variabl:

A → population variability

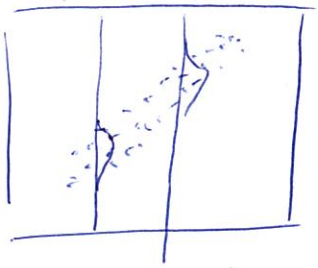
T → function of Rand. Var.

L → noise

$\hat{L}$ → uncertain (it has a prob)

σ_T →

T



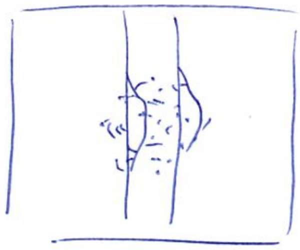
A

I learn T from A

$$P(A,T) = P(A)\,P(T\,|\,A)$$

does not factorise

T



A

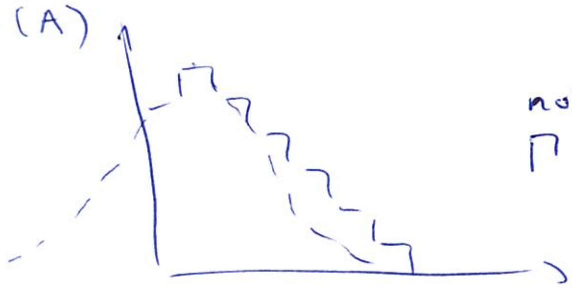
only looking at data, no stats yet

factorises

$$P(A,T) = P(A)\,P(T)$$

↑ let's assume this.

hist (A)



A

not G

$$\Gamma(A\,|\,\nu, A_0) = \frac{1}{\Gamma(A_0)} A_0^{\nu} A^{\nu-1} \exp(A_0 A)$$

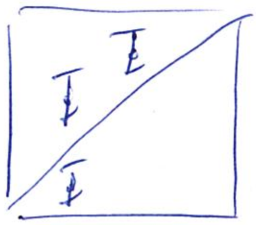
(powerlaw × exp)

if hist well resolved, ~~we~~ can fit A_0 and ν.

- $P(A) = \Gamma(A\,|\,A_0, \nu)$
- $P(A,T) = \Gamma(A\,|\,A_0, \nu)\; G(T)$

detector → counts photons but many photons → G

$\hat{L}$



$L = \sigma_T A T^4$

scatter is big because error bars are only the noise, but there are other sources of uncertainty (A, T)

- $P(\hat{L}\,|\,L) = G(L, \sigma_L^2)$

want $P(\sigma_T\,|\,\hat{L})$

$$P(\sigma_T | \hat{L}) = \int P(\underbrace{\sigma_T, A, T, L}_{\text{everything that affects the data}} | \hat{L}) \underbrace{dA\, dT\, dL}_{\text{I don't care}} =$$

go to joint

$$= \int \frac{P(\sigma_T, A, T, L, \hat{L})}{P(\hat{L})}\, dA\, dT\, dL =$$

factor out what you know

univ does not care if I look at L

$$= \int \frac{P(\sigma_T, A, T | L, \cancel{\hat{L}})\, P(L, \hat{L})}{P(\hat{L})}\, dA\, dT\, dL =$$

$$= \int \frac{P(\sigma_T, A, T | L)\, \overbrace{P(\hat{L} | L)}^{\text{detector } G}\, P(L)}{P(\hat{L})}\, dA\, dT\, dL$$

$$= \int \frac{P(\sigma_T, A, T, L)\, G(\hat{L} | L)}{P(\hat{L})}\, dA\, dT\, dL =$$

try stell boltz law

$$= \int \frac{P(L | A, T, \sigma_T)\, P(\sigma_T, A, T)\, G(\hat{L} | L)}{P(\hat{L})}\, dA\, dT\, dL$$

$$= \int \frac{\delta_D(L - \sigma_T A T^4)\, P(\sigma_T, A, T)\, G(\hat{L} | L)}{P(\hat{L})}\, dA\, dT\, dL$$

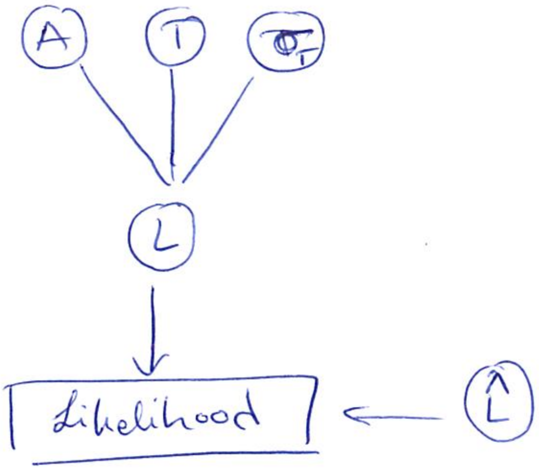
$$= \int_A \int_T \frac{P(\sigma_T, A, T)\, G(\hat{L} | \sigma_T A T^4)}{P(\hat{L})}\, dA\, dT\, \cancel{dL}$$

univ does not care if I put params.

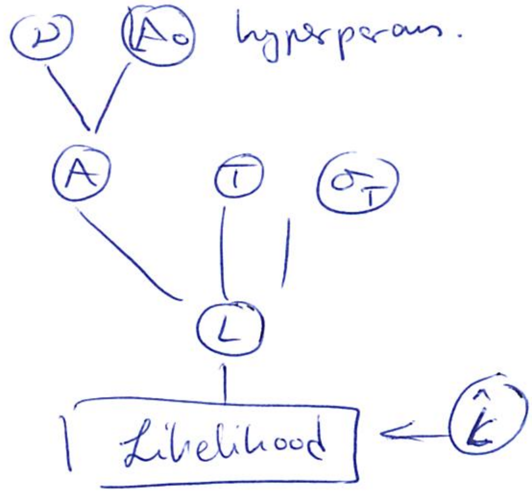
$$= \int_A \int_T \frac{P(A, T | \cancel{\sigma_T})\, P(\sigma_T)\, G(\hat{L} | \sigma_T A T^4)}{P(\hat{L})}\, dA\, dT$$

$$= \int_A \int_T \frac{P(A, T)\, \Pi(\sigma_T)\, G(\hat{L} | \sigma_T A T^4)}{P(\hat{L})}\, dA\, dT$$

$$= \int_A \int_T \frac{P(A)\, G(T)\, \pi(\sigma_T)\, G(\hat{L} | \sigma_T A T^4)}{P(\hat{L})}\, dA\, dT$$

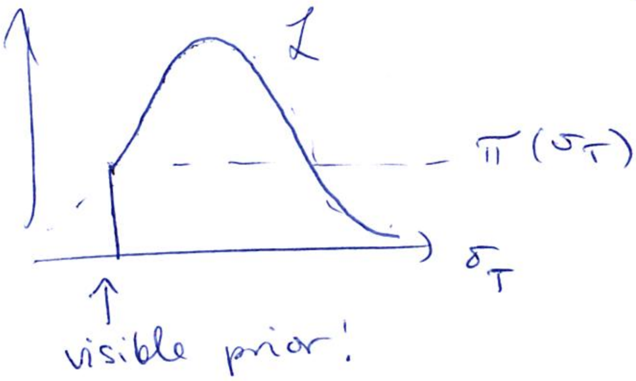


But imagine A_0, ν not well constraint by histogram:



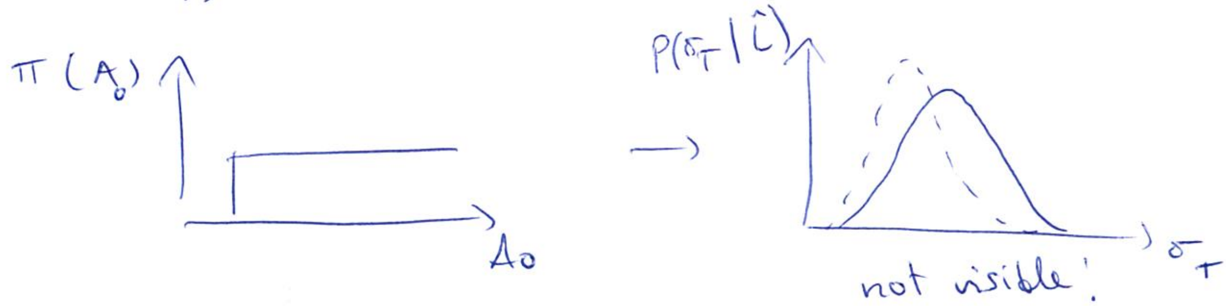
Priors

$P(\sigma_T | \hat{L})$

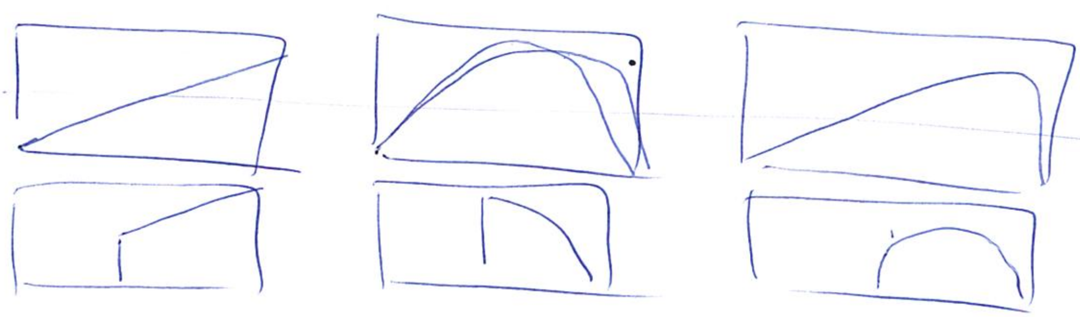


but ~~when~~ when prior is not on param of interest but on hyperparam:

$$\iint P(A)\, P(T)\, \pi(\sigma_T)\, G(\hat{L} | L)\, \underline{\pi(\nu)\, \pi(A_0)}\, dA\, dT\, d\nu\, d\underline{A_0}$$



Rule : 2, 4, 8 THTT



WL example

shear map (noisy)

cosmo params → observed map

want $P(\theta \mid \hat{map})$

collect random variables:

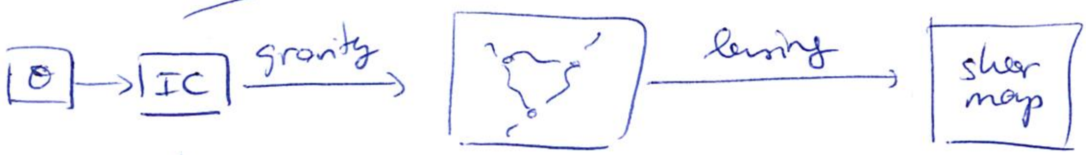
- θ uncertain
- map initial fluctuations δ
- $\hat{map}$ noisy

$$P(\theta \mid \hat{map}) = \int P(\theta, map \mid \hat{map})\, d(map)$$

~~10^6 params~~

$$P(\theta, map \mid \hat{map}) \propto \underbrace{P(\hat{map} \mid map)}_{\text{detector noise } G\text{, no cov.}}\, P(map, \theta)$$

$$P(map, \theta) = \underbrace{P(map \mid \theta)}_{\text{theory}}\, \underbrace{P(\cancel{map}\ \theta)}_{\pi(\theta)}$$



deterministic

$$\delta_D\left[IC - M(IC)\right]$$

whatever gravity model you want to put in simulator.

$$P(\theta \mid \hat{map}) = \int \underbrace{P(\hat{map} \mid map)}_{\substack{\text{detector}\\ G}}\ \underbrace{P(map \mid IC, \theta)}_{\substack{\text{simulator}\\ \delta_D}}\ \underbrace{P(IC \mid \theta)}_{\substack{\text{prior}\\ G}}\ \underbrace{P(\theta)}_{\substack{\text{prior}\\ \pi}}\ d\,map\ \underbrace{dIC}_{10^6\,\text{dim}}$$

$$\begin{tikzcd}
\theta \arrow[d] & \\
IC \arrow[d] & \gamma \arrow[dl] \\
map \arrow[d] & \\
\boxed{\text{Likelihood}} & \hat{map} \arrow[l]
\end{tikzcd}$$

baryon param (hyperparam?) γ

easy because we have model everything, ~~no cov~~.

BHM

splits problem into steps → sub-models
links sub-models and propagate uncertainties
we need conditional prob of each step
BHM is a complete model of the data
challenge: sub-models introduce latent variables
↓
many params.

~~Sum~~

$\boxed{\text{FLI}}$: BHM with a likelihood at the field-level

WL : $\log \mathcal{L} = \sum_b \sum_i \frac{(\gamma^1_{i} - \hat{\gamma}^1_{b,i})^2 + (\gamma^2_{b,i} - \hat{\gamma}^2_{b,i})^2}{2\sigma_{i,b}^2}$

$\sigma_i = \frac{\sigma_\epsilon}{\sqrt{N_{gal,i}}}$

GC: $\log \mathcal{L} =$

Ly α $\log \mathcal{L} = \sum_{skewer} \sum_i \frac{(F - FGPA(\delta))^2}{2\sigma^2}$

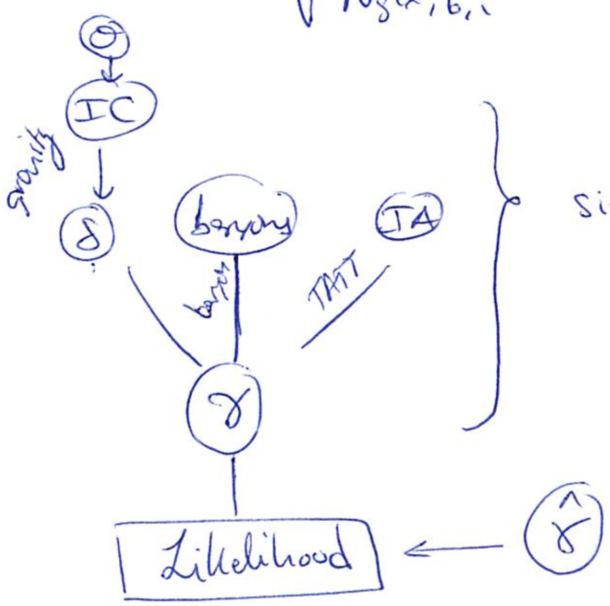
FLGPA: $F = e^{-A(1+\delta)^3}$

Field-level inference

BHM of the data at field level + $\mathcal{L}$ pixel by pixel

⊛ WL: $\log \mathcal{L} = \sum_b \sum_i \frac{(\gamma^1_{b,i} - \hat{\gamma}^1_{b,i})^2 + (\gamma^2_{b,i} - \gamma^2_{b,i})^2}{2\sigma^2_{b,i}}$

$\sigma_{b,i} = \frac{\sigma_\epsilon}{\sqrt{N_{glx,b,i}}}$ no cov. matrix



simulator within the stats framework.

↓

all sims are informed by the data (not random)

↓

result: Θ and IC

⊛

GC: Poisson.

~~$P(N_k^g | \lambda_k)$~~

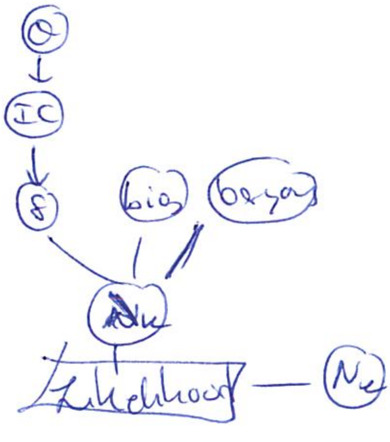
$\log \mathcal{L} = \prod_k \frac{(\lambda_k)^{N_k^g} e^{-\lambda_k}}{N_k^g!}$

observed

expected

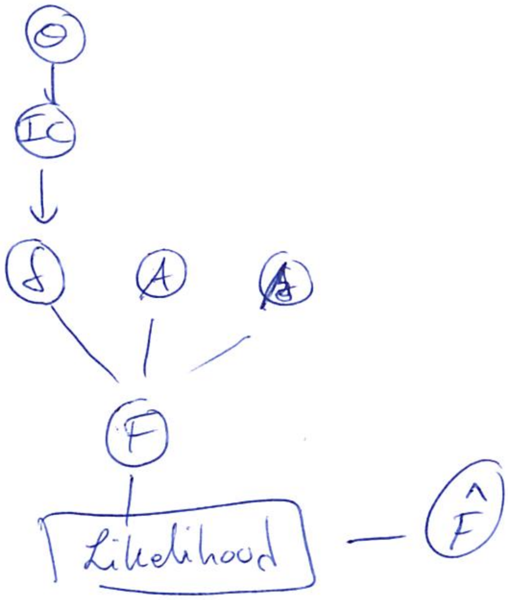
survey map

$\lambda_k = R_N \bar{N} (1 + B(\delta^L)_k)$



④ Ly-α forest:

$$\log \mathcal{L} = \sum_{\text{skewers}} \sum_{i} \frac{\left(\hat{F} - e^{-A(1+\delta)^{\beta}}\right)^2}{2\sigma^2}$$



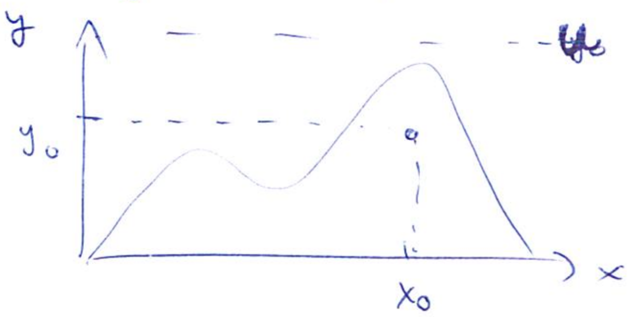
<u>Sampling</u> : 10^6 params of IC + bunch of Ω

Aim: $n(\vec{\theta}) \propto P(\vec{\theta} \mid \hat{F})$

density of samples is prop to prob.
then all integrals are replaced by Monte Carlo integration.

(high prob → more samples)

easy way: rejection sample:



$x \sim \text{unif}\,[a, b]$
$y \sim \text{unif}\,[0, u]$
(x_i, y_i) check if $P(x_i) > y_i$
That gives $\{x_i\} \propto P(x)$
(introduces additional dimension)

works for $n < 3$ → $\dfrac{V_0^n}{V_D^n} \to 0$

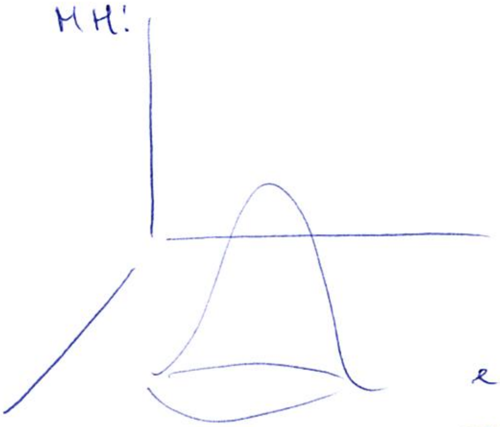
keep info from previous step

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_n$

Markov chain $\rightarrow$ x_{i+1} depends on x_i
but not on x_{i-n} or x_{i+n}

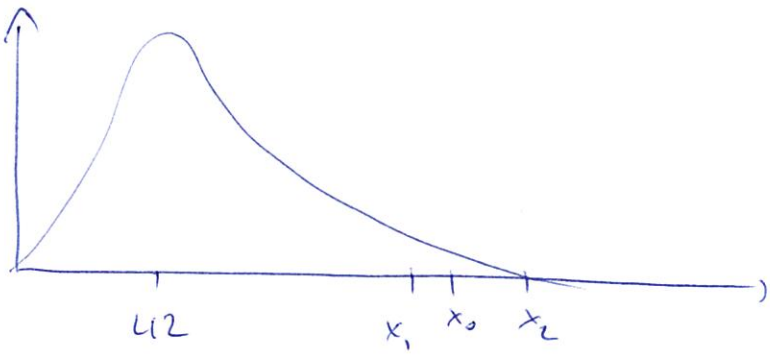
MCMC
- Metropolis Hastings $n \sim 20-30$
- Gibbs
- HMC ~~[illegible]~~ } $n \sim 10^6$

MH:



explore vicinity and move center. Accept / reject step to walk towards high prob

$\rightarrow$ prop distrib symmetric $P(x_0 | x_1) = P(x_1 | x_0)$



i	x
0	88
1	82
2	82
3	70

① init at x_0
evaluate $P(x_0)$
store

② draw Δ (usually G)
$x_1 = x_0 + \Delta$
evaluate $P(x_1)$
test if $P(x_1) > P(x_0)$ keep (towards high prob)
↳ new! not in rejection!
~~otherwise introduce α~~ ~~$\rightarrow x_2$~~ $= x_1$

(3) introduce α

$x_2 = x_1 + \Delta$

evaluate $P(x_2)$

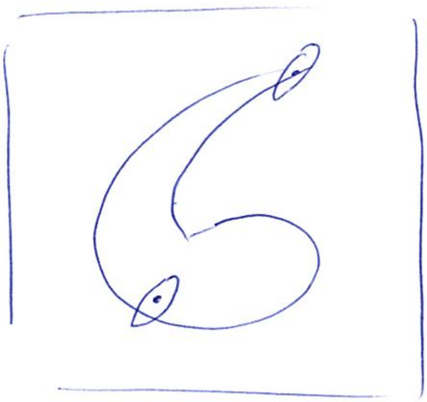
if $P(x_2) > P(x_1)$ accept

otherwise, draw $\alpha \sim \text{unif}\,[0,1]$

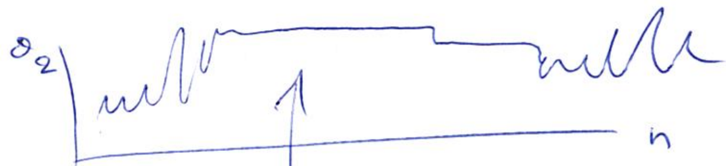
compute $r = \dfrac{P(x_2)}{P(x_1)}$

if $(\alpha < r)$ accept x_2

Problem:



same prop. distrib everywhere



highly likely

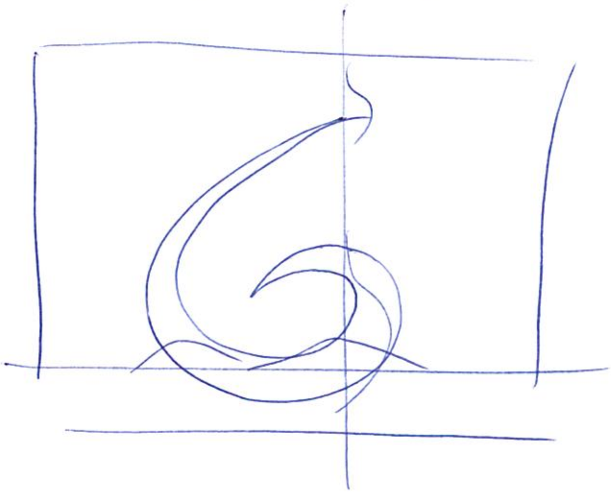
$n(\theta) \not\propto P(\theta)$

Fix: Gibbs sampling

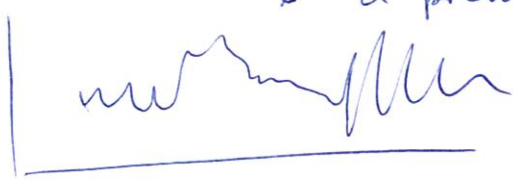
$$\vec{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_f \end{pmatrix} \begin{matrix} \} \text{ fixed} \\ \\ \} \text{ random} \end{matrix}$$

$$\vec{\theta} = \begin{pmatrix} \vec{\theta}_d = \vec{f} \\ \theta_{f-d} = \vec{r} \end{pmatrix}$$

need $P(\vec{r} \mid \vec{f})$



less of a problem



① $T \sim G(T)$

② $A \sim \Gamma(A)$

③ $\sigma_T \sim \pi(\sigma_T)$

④ $\sigma_T A T^4 \rightarrow$ intro $\delta_D \rightarrow$ gives L

⑤ input L to Gaussian $\rightarrow$ prob (number)

⑥ MH accept - reject

HMC : uses physics to solve stats

① build potential from $P(x)$

$$\psi = -\ln P(x)$$

② define kinetic energy $\quad p_i \sim G$

$$k = p_i^T M^{-1} p_i$$

③ define Hamiltonian $\quad H = k + \psi$

④ evolve dynamical syst

$$\dot{x} = \frac{\partial H}{\partial p} = M^{-1} p$$

$$\dot{p} = -\frac{\partial H}{\partial x} = -\frac{\partial \psi}{\partial x}$$

⑤ integ. eq. of motion (leapfrog)

$$p_i\left(t + \frac{\varepsilon}{2}\right) = p_i(t) - \frac{\varepsilon}{2}\left(\frac{\partial p}{\partial x}\right)_{x(t)}$$

$$x_i(t+\varepsilon) = x_i(t) + \varepsilon\, p_i\left(t + \frac{\varepsilon}{2}\right)$$

$$p_i\left(t + \frac{\varepsilon}{2}\right) = p_i\left(t + \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2}\left(\frac{\partial p}{\partial x}\right)_{x(t+\varepsilon)}$$

$$(x_i, p_i) \rightarrow (x_{i+1}, p_{i+1})$$

why it works: H conserved $\rightarrow$ 100% acceptance in MH

$\min\left[1, \exp(-H^i + H^{i+1})\right]$

gradients

COLOURS (8)

Pseudo-code:

for $i = 0$ to N_{MCMC}:

if $i = 0$ → choose starting point x_i

else → use current point

draw random velocity $p_i \sim G(0, \mathbb{I})$

leapfrog loop : use x_i p_i as inial point

for $j = 0$ to N_{Leap}:

$(x_j, p_j) \to (x_{j+1}, p_{i+1})$

$(x_{NL}, p_{NL}) = (x_{j+1}, p_{i+1})$

draw $\alpha \sim U(0, 1)$

$R = \exp[-H_i + H_{NL}]$

if $\alpha > R$ x_{NL} rejected, $x_{i+1} = x_i$

if $\alpha < R$ x_{NL} accepted $x_{i+1} = x_{NL}$

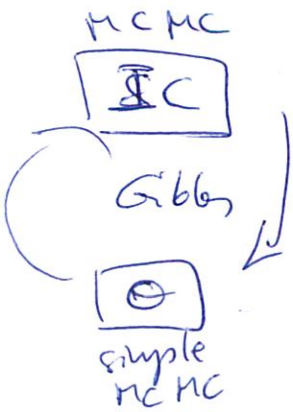
Price: tunning param: ε = timestepr } accept rate

N_L : number of timesteps

M : mass matrix → efficienty

$\frac{1}{\mathrm{cov}(x)}$

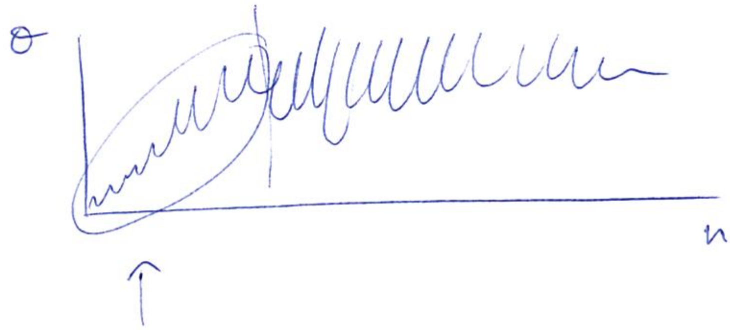
(update)



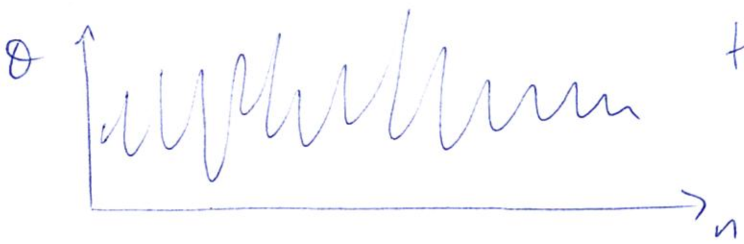
slide

Tests & diagnostics

slide



burn-in phase

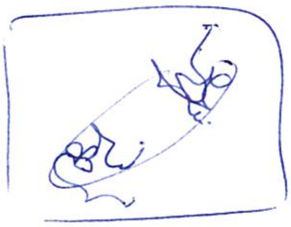


thinning $\longrightarrow$

corr. length!

$$c_n(\theta) = \frac{1}{N-n} \sum_{0}^{N-n} \frac{(\theta_i - \langle\theta\rangle)(\theta^{i+1} - \langle\theta\rangle)}{var(\theta)}$$

$$N_{eff} = \frac{N}{\sum \text{corr length}}$$



do not combine unless indep. converged

convergence test: Gelman - Rubin

run M chains of N samples each.

each chain $< \begin{matrix} \hat{\theta}_m \\ \sigma_m^2 \end{matrix}$

all chains $< \begin{matrix} \hat{\theta} = \frac{1}{M}\sum \hat{\theta}_m \\ B = \frac{N}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta})^2 \end{matrix}$

average variance $W = \frac{1}{M} \sum \sigma_m^2$

$$\hat{V} = \frac{N-1}{N} W + \frac{M+1}{MN} B$$

if $\sqrt{\frac{\hat{V}}{M}} \approx 1$ converged!

Posterior pred. test.

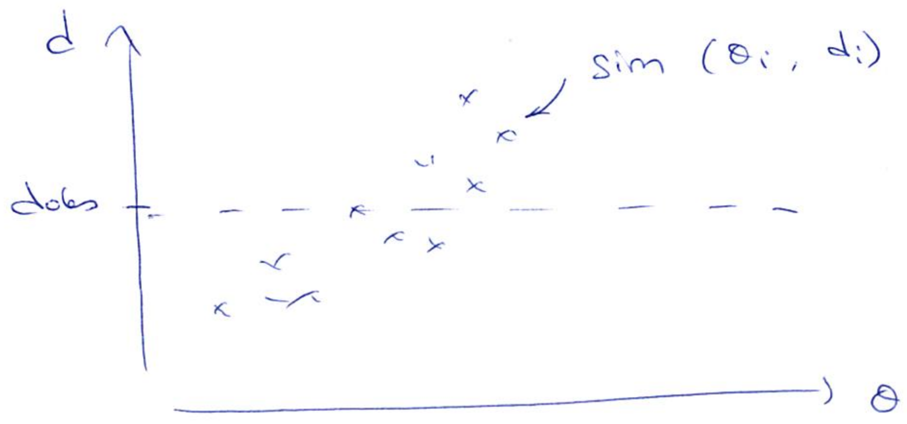
SBI

random variables: $\theta \in \mathbb{R}^D$, $d \in \mathbb{R}^N$

realisation:
- $\theta_{truth} \sim P(\theta)$ unknown
- $d_{obs} \sim P(d \mid \theta = \theta_{truth})$ observed

I can simulate $\theta \xrightarrow{BHM} d$ but I don't have / cannot compute Likelihood.

Goal: approx posterior distrib
$P(\theta \mid d = d_{obs})$ given a set of simulations
$\{(\theta_i, d_i)\}_{i=1 \dots N_{sim}}$



how do I get posterior without evaluating $P(d \mid \theta)$?

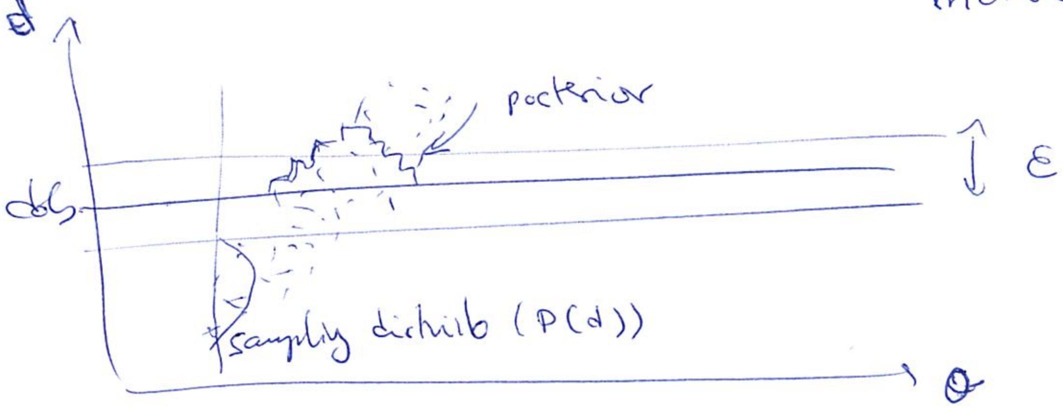
Density stimation:

1. run sims
2. stimate target distrib
3. extract posterior $P(d = d_{obs} \mid \theta)$ from modelled pdf

choices: target $p(\theta | d)$ or $p(d | \theta)$
type of density model (GMM, histogram...)
simulation schedule (predet? interactive?)
how to extract posterior from density model?

ABC

model $P(\theta | d)$ from "accepted" sims close to dobs
metric



hist of $\{\theta_i\}$ such $|d_i - d_{obs}| < \varepsilon$

✓ simple
✗ difficult to tune ε and metric

Neural density estim (kernel density - DELFI)

① fit NN model $g_\phi(\theta, d)$ to ~~$p(d|\theta)$~~ sample of $P(d|\theta)$ using NN techniques.

NN output, NN param, NN input

② run more sims (in regions of high density)
posterior $p(\theta | d = d_{obs}) \approx g(\theta_i, d_{obs}) \cdot p(\theta)$

iterative

✓ flexible
✗ evaluate prior