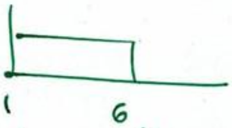
Inference

Dice → every time a different process → random

roll many times #

sufficiently many times →



uniform

is the dice fair? just repeat sufficiently many times

inference → many dice. Roll one, got 1, which one? (infere)

you can't observe the truth. Repeating don't help.

1 → 6 face → prior

12 face → best fitting model ≠ truth

remove dice + add more 6 faces → more likely

↓

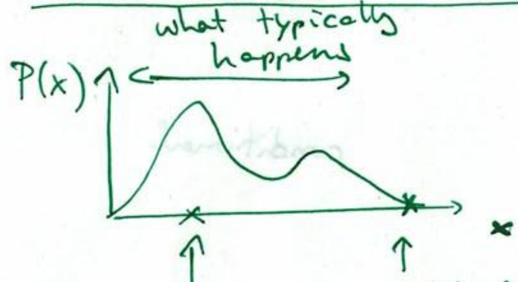
prior

everything you know before the datapoint is taken.

$\bar{x} \longrightarrow \theta$ unobserved (Ω)

priors are essential to learn from experiments.

Some distrib. & notation



draw random value → realisation

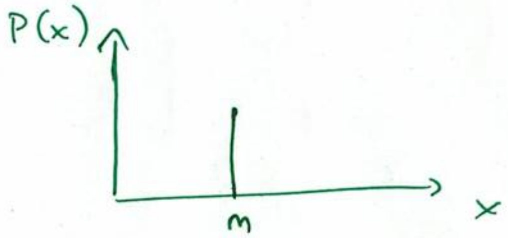
unlikely realisation

more likely.

what happens most of the time.

$P(x|\theta) \Rightarrow \partial_\theta P(x|\theta) = 0$ best fitting param.

maximum likelihood stimate (χ^2 method if G)



$\delta_D(x-m)$

classical statement (physical law) as prob. distribution

$E = mc^2 \longrightarrow \delta_D(E - mc^2)$

can be random var

only possible if $E = mc^2$.

A, B : $P(A,B)$ joint

statistically indep $P(A,B) = P(A)\,P(B)$

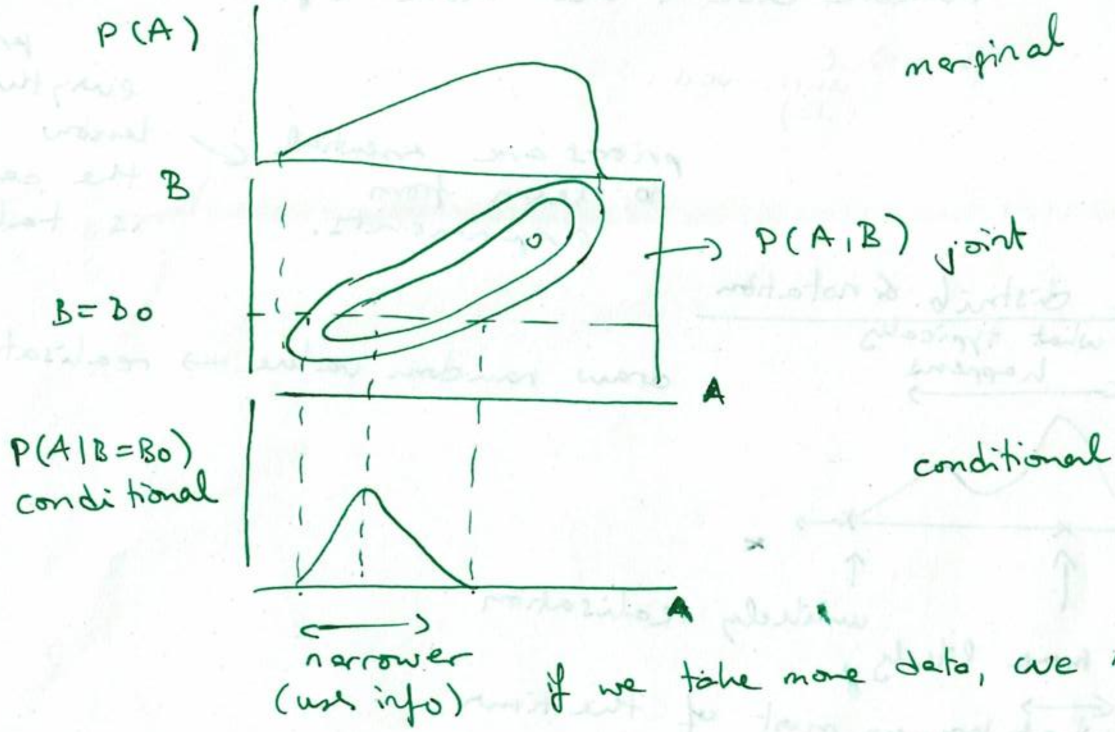
else : $P(A,B) = P(A)\,\underbrace{P(B|A)}_{\text{conditional}} = P(B)\,P(A|B)$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad \text{Bays Th.}$$

$P(A,B)$ joint

$P(A|B)$ conditional → uses information

$P(A) = \int P(A,B)\, dB$ marginal → drops info



if we take more data, we narrow dist.

Inference

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

$P(\vec{x})$ sampling dist

$\mathcal{L}(\vec{x}\,|\,\vec{\theta})$ likelihood

$\pi(\vec{x})$ evidence (prob of getting the data at all) (funct model)

$\pi(\vec{\theta})$ prior (what are good values for param?)

$P(\vec{x})$ sampling distrib (nature) (not funct of model)

likely data | outliers
std

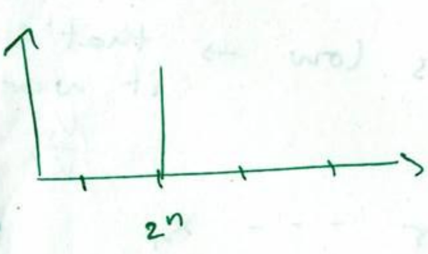
$$P(\vec{x}) = \int P(\vec{x},\vec{\theta})\,d^n\theta = \int \underbrace{p(\vec{x}\,|\,\vec{\theta})}_{\text{likelihood}}\;\underbrace{P(\vec{\theta})}_{\text{prior}}\,d^n\theta) = \underbrace{\mathcal{E}}_{\text{evidence}}$$

we want to introduce param (human)

likelihood (theory/sim)

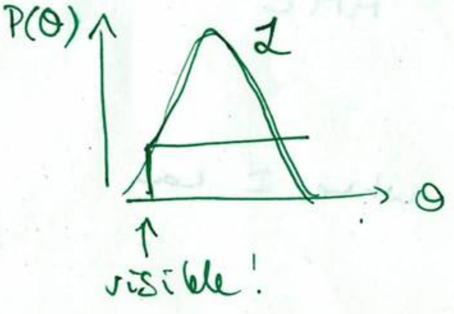
Priors

nature rule game : 2 4 8



you cannot overrule the zeros. You need a very good reason to set something to 0. Data cannot overrule this.

Prior : the more data you collect, the less the prior matters.



visible!

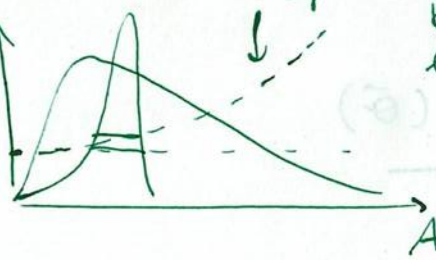
BHM

need to deal with variability

$L = \sigma_T A T^4$

invisible hard cut in the A.

hyperparameters



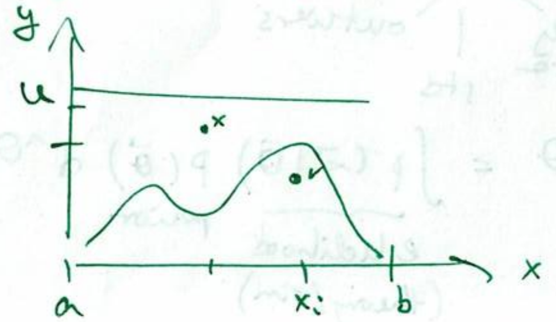
upweight α here
because I know
large values of A are more likely.

Sampling

fullfill $\pi(\vec{\theta}) \propto P(\vec{\theta} \mid \dots)$

$x \sim G(0, \sigma^2)$

① rejection sampler :



$x \sim \text{unif}\,(a, b)$ } 1 variable 2 dimension
$y \sim \text{unif}\,(0, u)$
prob > 0

(x_i, y_i) ~~check~~ keep if $P(x_i) \geq y_i$

result: $\{x_i\}$ $\propto P(x)$ (populate sim with s|x)
ensamble
... $n \leq 3$

more sample where
P(x) is high and
less where P(x) is low → that's why it works.

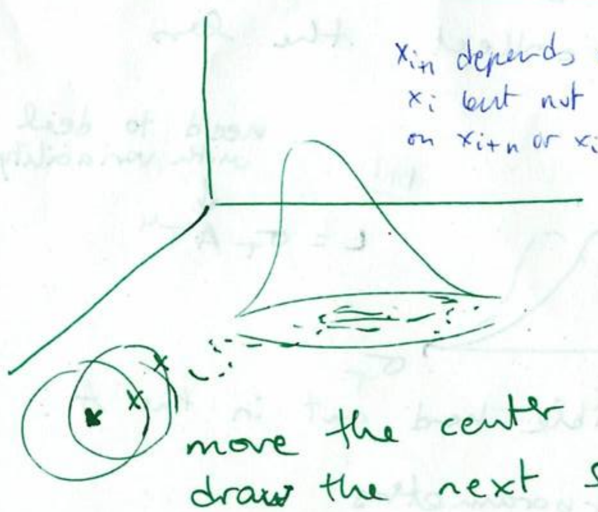
$\frac{V_0^n}{V_B^n} \to 0$

②

Sequence $x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad - - - \quad x_n$ → time

Markov chain

x_{i+1} depends on x_i but not on x_{i+n} or x_{i-n}

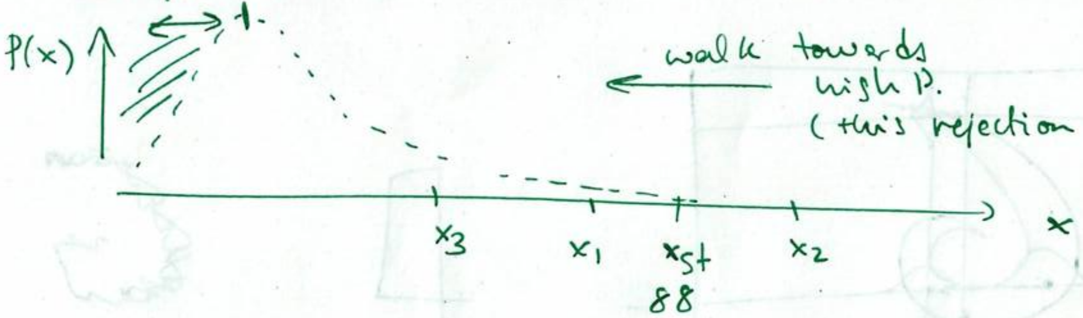


MCMC : Metropolis Hastings $n \sim 20$
integral depend.
Gibbs sampler $n \sim 10^6$
HMC $n \sim 10^6$

move the center of prob region where I can draw the next sample.

acceptance critria → to go towards high P region

Metropolis Hastings



i	x	P(x)
0	88	P(88)
1	82	P(82)
2	82	P(82)
3	70	P(70)

init at x_0

prop. distrib: → ~~evalua~~ draw Δ

$$P(x_1 | x_0) = P(x_0 | x_1)$$

(symmetric to undo step)

$x_1 = x_0 + \Delta$, accept-reject

keep if $P(x_1) > P(x_0)$

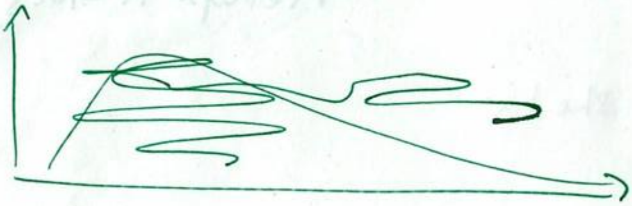
draw $x_2 = x_1 + \Delta$

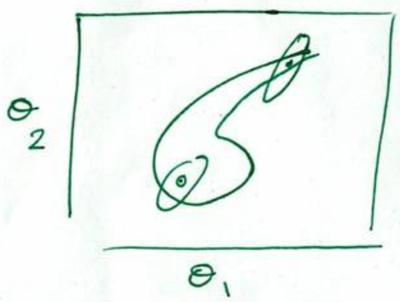
if $P(x_2) < P(x_1)$ → generate $\alpha \sim$ unif $[0, 1]$

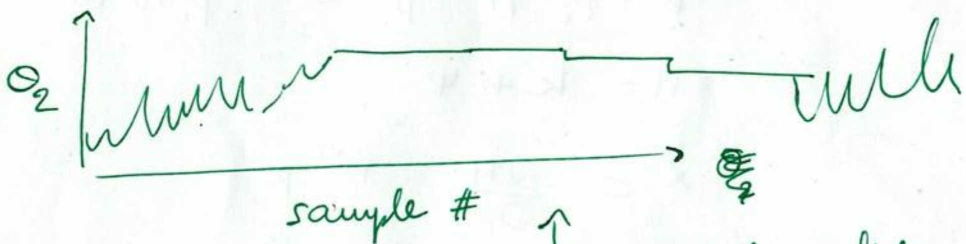
~~if α <~~ $r = \dfrac{P(x_2)}{P(x_1)}$

if $\alpha < r$, keep x_2

problem if prob can be 0. or tails



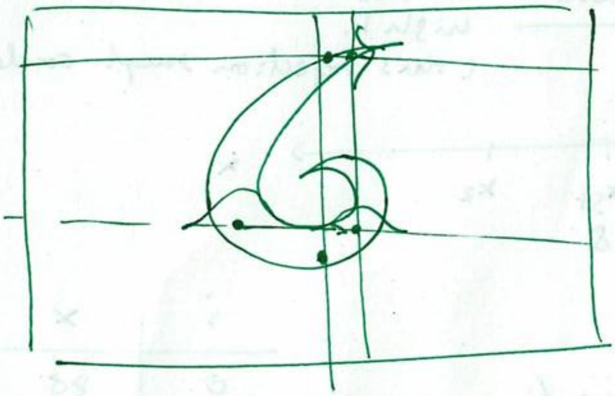


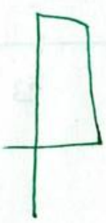


↑ very likely value not true → bias

solution → Gibbs

Gibbs sampling

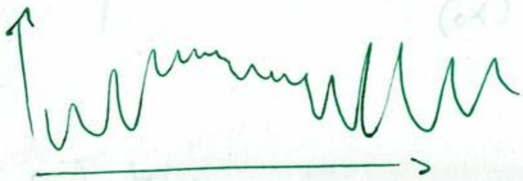




$$\vec{\Theta} = \begin{pmatrix} \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_n \end{pmatrix} \begin{matrix} \} \text{ fixed} \\ \\ \} \text{ sample (random)} \end{matrix}$$

we need

$\mathcal{P}(\Theta_r | \Theta_f)$



$T \sim P(T)$

$A \sim P(A)$

$\sigma_T \sim P(\sigma_T)$

$\sigma_T A T^4 \to L$

$G(\hat{L} | L) \to$ number

Metrop. Hasting accept / reject

HMC

use physics to solve stats

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

$$k = p_i^T M^{-1} p_i \qquad p_i \sim G$$

$$H = k + \Psi$$

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

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

$$\min \left[1, \exp(-H^i + H^{i+1})\right] \leftarrow \text{conserved} / \text{good}$$

BHM : stellar field

$L = \sigma_T A T^4$ → want $P(\sigma_T \mid \hat{L})$ ← stimation (noise, different answer every measurement)

random variables:

$\left.\begin{array}{l} A \\ T \end{array}\right\}$ pop. variability $\left\{\begin{array}{l} P(A \mid \alpha, A_0) = P(A) \\ G(T) = P(T) \end{array}\right.$

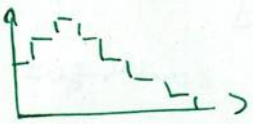
L funct of R.V.

$\hat{L}$ noise $\qquad P(\hat{L} \mid L) = G(\hat{L}, L)$

σ_T prob associated to it. → $\cancel{P(\sigma_T \mid \hat{L})}$ $P(\sigma_T)$ prior

Assume $P(A, T) = P(A)\, P(T)$ indep.

hist A



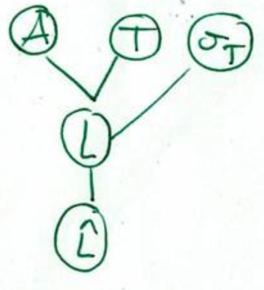
→ Gamma dist. (only positive) $P(x \mid \alpha, s)$

detector / noise: counting photons in CCD pixels

↳ Poisson → G
very bright, many photons.

$$P(\sigma_T \mid \hat{L}) = \int \frac{P(\underbrace{\sigma_T, A, T, L}_{\text{everything affecting data.}} \mid \hat{L})}{P(\hat{L})} \underbrace{dA\,dT\,dL}_{\text{don't care about these RV}} =$$

$$= \int P(A)\, P(T)\, \pi(\sigma_T)\, \underbrace{P(L \mid \sigma_T, A, T)}_{\delta_D(L - \sigma_T A T^4)}\, \underbrace{P(\hat{L} \mid L)}_{G(\hat{L} \mid L)}\, dA\,dT\,dL$$



Test & diagnostics

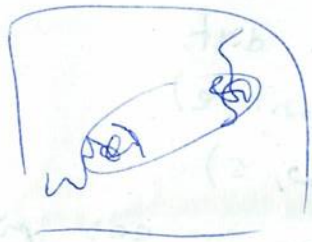
burn-in

stepsize too large

stepsize too small

↓ thinning

→ corr. length!



not combined
except if converged!
Gellman - Rubin test