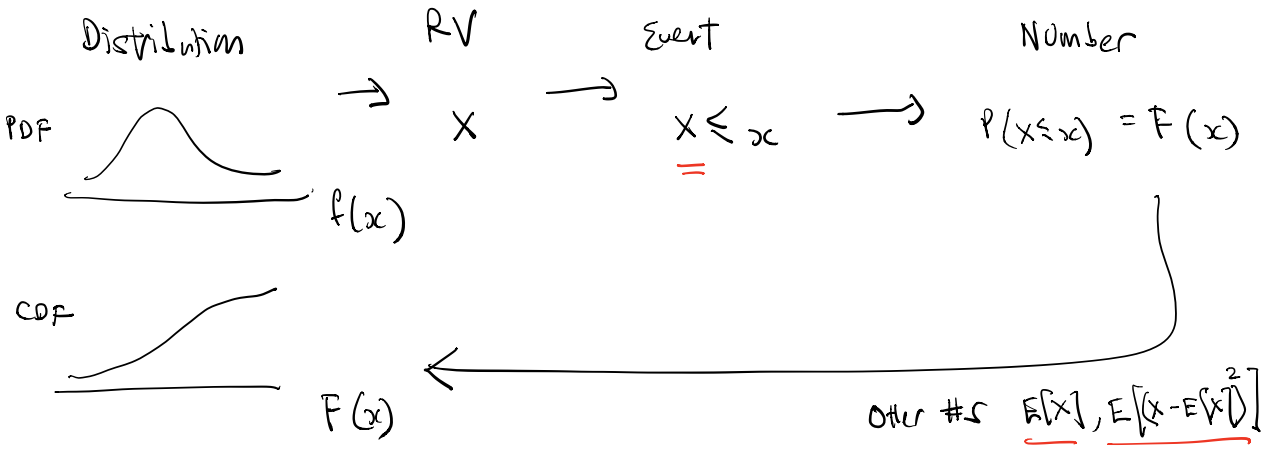


A RV is a function that maps an event in sample space to a number

$$X : \Omega \to \mathbb{R}$$

2 coin tosses → X is # heads

HH	HT
TH	TT

$$X = \begin{cases} HH \to 2 \\ HT \to 1 \\ TH \to 1 \\ TT \to 1 \end{cases}$$

Family of distributions

$N(\mu, \sigma^2)$

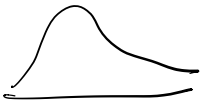
↙

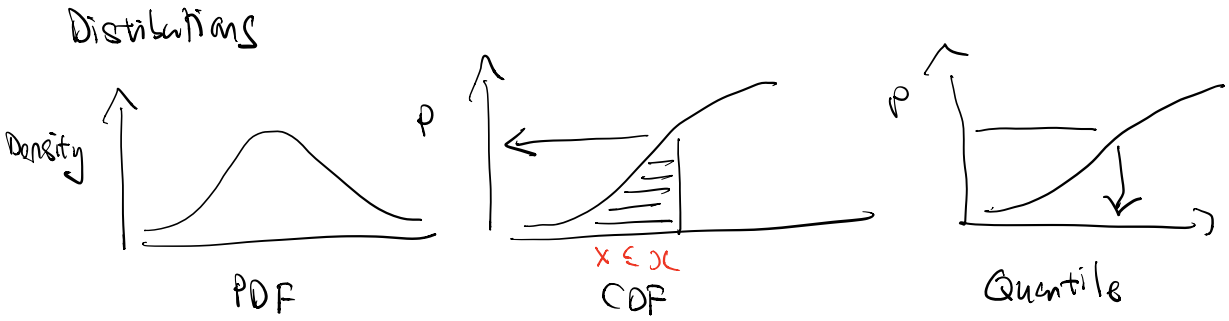
Every realization of $\underline{\theta} = (\mu, \sigma)$ defines a member of the normal family → $\theta = (\mu, \sigma)$ are parameters of the distribution

Types of distribution

Discrete $X \in \mathbb{Z}$

Continuous $X \in \mathbb{R}$





E.g. Suppose IQ is normally distributed with

$\mu = 100$ & $\sigma = 15$

(1) Fraction of $IQ \leq 120$, $80 \leq IQ \leq 110$ (cdf)

(2) IQ to be in top 5% of population (quantile)

(3) mode of IQ (pdf)

(4) Sample 100 IQ values in top 25 percent (rvs)

norm.cdf ≡ pnorm	norm.ppf ≡ qnorm
norm.pdf ≡ dnorm	norm.rvs ≡ rnorm

scipy.stats

Likelihood

Probability $x : x \mapsto f(x|\theta)$

Likelihood $\theta : x \mapsto f(x|\theta)$ } Equivalent

$L(\theta|x) = f(x|\theta)$ }

For independent events, likelihood of n events is the product of the individual likelihoods

$\downarrow$

$\log \prod \mapsto \sum$ (log likelihood)

$l(\theta|x)$

$\downarrow$

Function to come min, max as $L(\theta|x$

For optimization, often work with negative log likelihood

because we want a minimisation problem

Maximum likelihood estimation (MLE)

Find θ where $L(\theta|x)$ is max

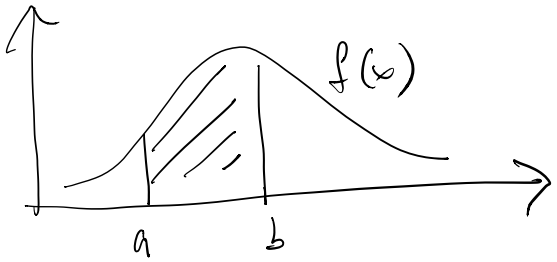
$\equiv$ Find θ where $-L(\theta|x)$ is min

} θ_{MLE}

$$\theta_{mle} = \arg\max L(\theta|x)$$

→ Notebook 8A

Probability & Integration



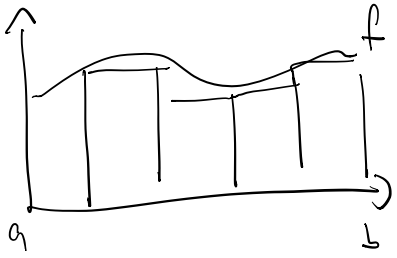
$$E[X] = \int x\, p(x)\, dx$$

$$E[f(X)] = \int f(x)\, p(x)\, dx$$

$$P(a \le X \le b) = \int_a^b f(x)\, dx$$

Finding probabilities involves integration

Numerical Quadrature



Some form of divide & conquer summation

Trapezoid, Simpsons etc.

Scales poorly to ↑ dimensions

if n is the # grid points → n^d

$n = 10$

$d=1$	10
$d=2$	100
$d=3$	1,000
$d=4$	10,000
$d=100$	...

Monte Carlo integration

Lotus

$$E[g(x)] = \int_x g(x)\, p(x)\, dx$$

$$\bar{g} = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \qquad x_i \sim p$$

Convergence is $O(n^{1/2})$ independent of dimensionality

Some intuition

$$I = \int f(x)\,dx$$

$$E\left[g(x)\right] = \int g(x)\,p(x)\,dx$$

Choose $g(x) = \frac{f(x)}{p(x)}$

$$E\left[g(x)\right] = \int \frac{f(x)}{p(x)}\,p(x)\,dx$$

$$= \int f(x)\,dx$$

$$= I$$

LLN $$I \approx \bar{g} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

Estimating π.

$$f(x) = \begin{cases} 1 & \text{if } x^2+y^2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$I = \int_0^1 \int_0^1 f(x)\, dx = \frac{\pi}{4}$$

$$\tilde{\pi} = 4 \frac{1}{N} \sum_{i=1}^{N} f(x)$$

$$x \sim U(0,1)$$

Bayes theorem

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

Parameter is RV (c.f. frequentist)

posterior

likelihood

prior

$$= \frac{p(x|\theta)\,p(\theta)}{\int p(x|\theta)\,p(\theta)\,d\theta}$$

marginal (normalizing constant)

often high-dimensional & intractable

MAP

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{\int p(x|\theta)\,p(\theta)\,d\theta}$$

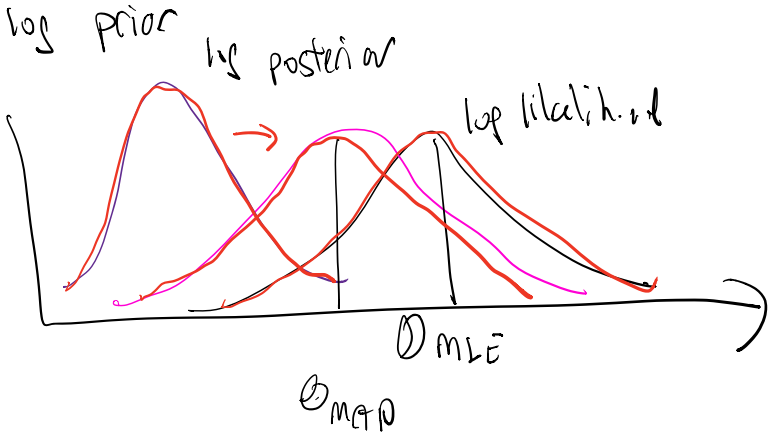
Normalizing constant

$$\propto p(x|\theta)\,p(\theta)$$

We can optimize → just like maximum likelihood

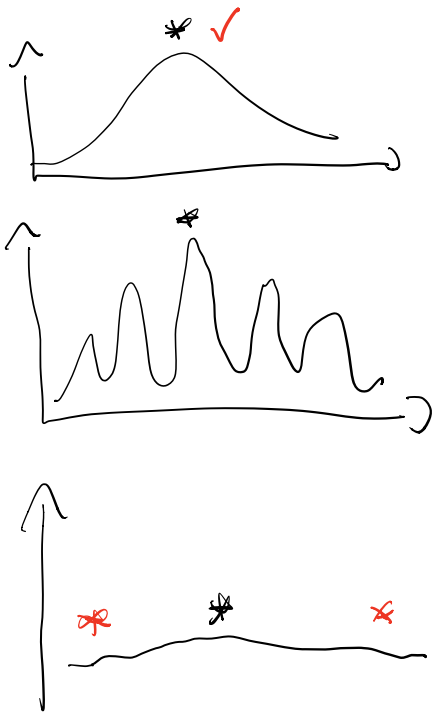
+ additional term for prior

⇒ MAP estimate



As $n \to \infty$ MAP → MLE

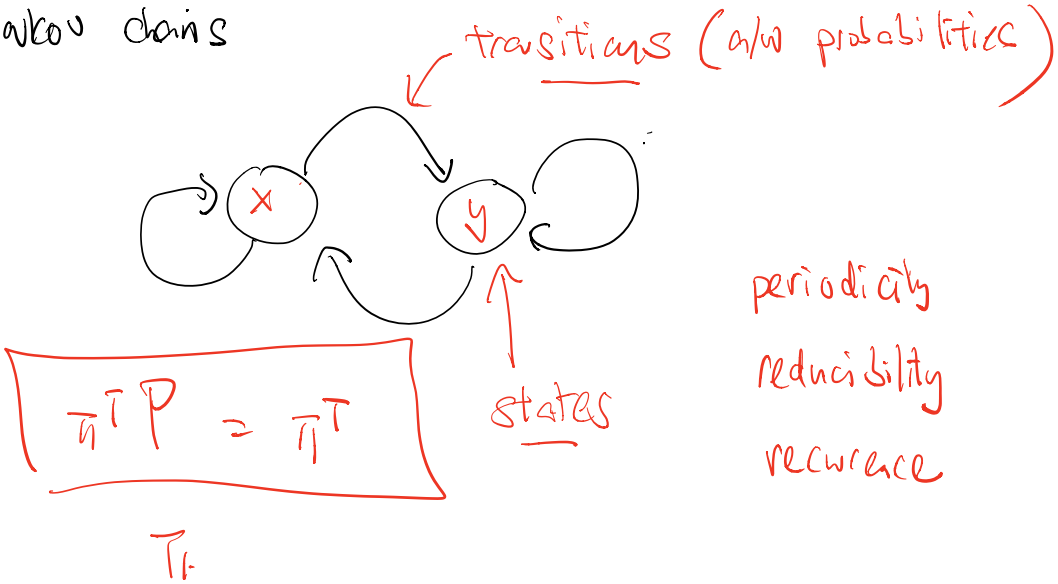
Limitation of MAP, MLE



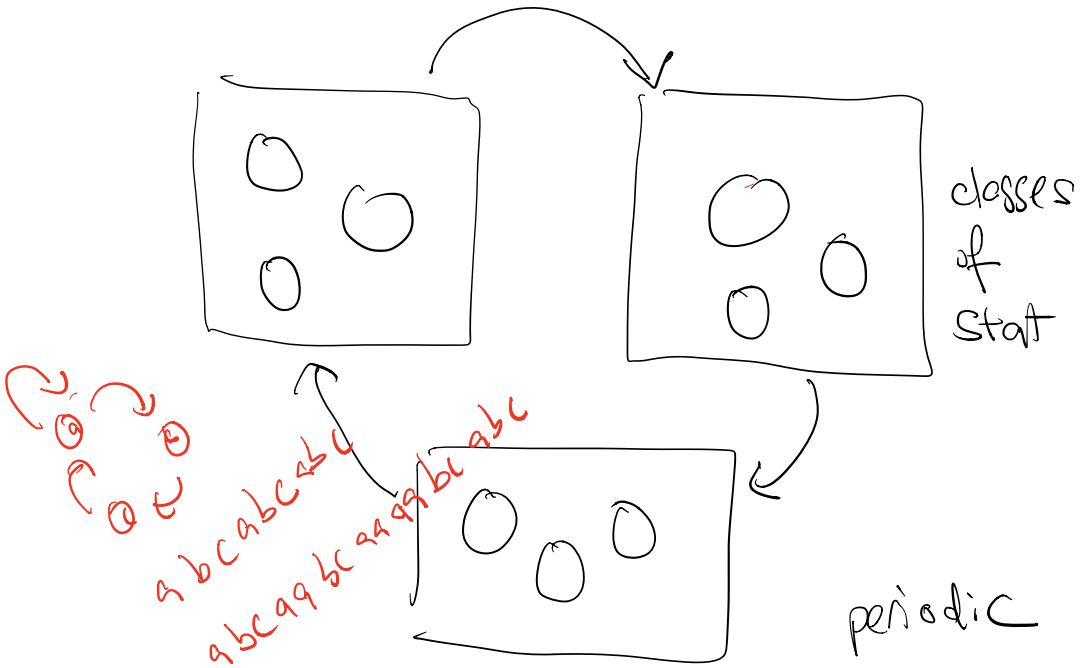
Same point estimate, very different posterior

MCMC → Markov chain Monte Carlo

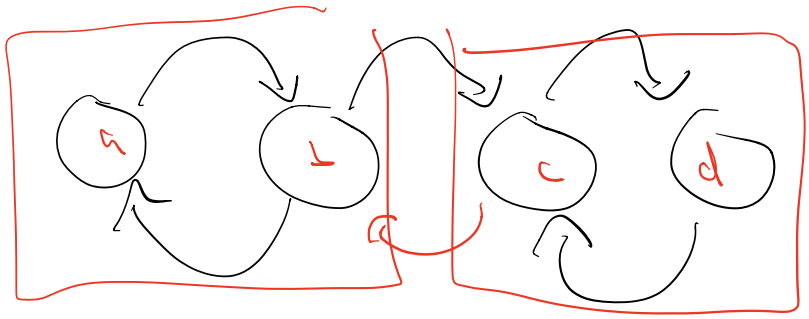
Markov chains



periodicity
reducibility
recurrence



aperiodic if not periodic



i, j communicate if $i \leftrightarrow j$

Put states into communication classes

reducible $\Rightarrow \geq 2$ communication classes

irreducible $\Rightarrow$ 1 communication class

recurrent if $P(i \to i) > 0$

For continuous Markov chains

positive recurrent if expected return time $< \infty$

Then as $t \to \infty$

i will be visited an ∞ # times

If a Markov chain is aperiodic, irreducible

& positive recurrent

$\Rightarrow$ unique steady state $\bar{\pi}$

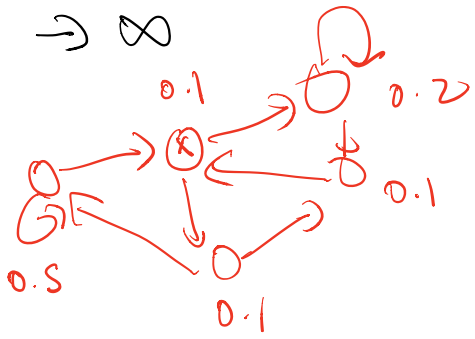
$$\bar{\pi}^T P = \bar{\pi}^T$$

Intuition

$\pi_i \simeq$ Average # of times in state i

as $t \to \infty$

Ergodic

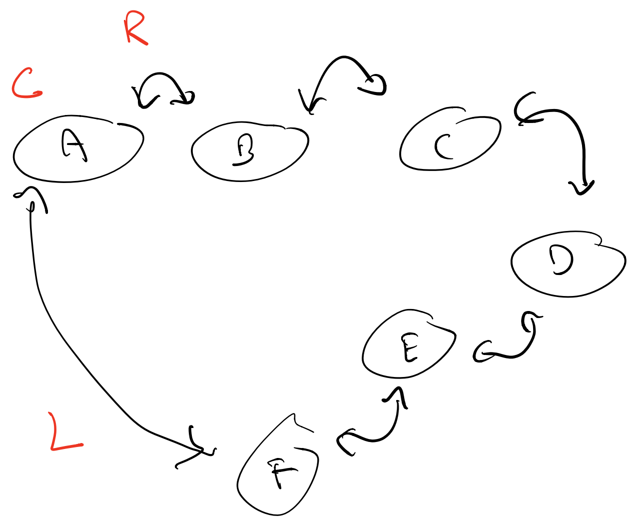


MCMC (1)

Steady state probability = posterior probability

(2) Why?

Kruschke's island census example



Can travel L, R or stay on current island.

Rules to move between islands

① Toss a coin

H → Ask for population (R) n_R

T → Ask for population (L) n_L

② Suppose you get H

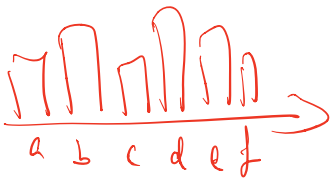
if $n_R > n_C$ → go (R)

else generate $p \in U(0,1)$

if $p < \frac{n_R}{n_C}$ → go (R)

else → stay

③ Repeat



post distribution

④ Distribution of population on islands
& Regency of visits (over long time)

MH → Metropolis Hastings

① Toss a coin ← symmetric proposal distribution

H → Ask for population (R) n_R

T → Ask for population (L) n_L

② Suppose you get H

if $n_R > n_C$ → go (R)

else generate $p \in U(0,1)$ ← standard uniform

if $p < \frac{n_R}{n_C}$ → go (R)

else → stay

③ Repeat

target function

MH.

(symmetric) proposal distribution

θ (start)

$\theta_p = \theta + \Delta\theta$ $\quad \Delta\theta \sim N(0, \sigma)$

$$\alpha = \frac{f(\theta_p|x)}{f(\theta|x)}$$ } acceptance probability; Ratios of portions

$\Rightarrow$ marginal cancels out!

$$\alpha = \frac{f(x|\theta_p)\,p(\theta_p)}{f(x|\theta)\,p(\theta)}$$

if $\alpha \geq 1$ $\quad \theta \leftarrow$

if $\alpha < 1$ $\quad \theta \leftarrow \theta_p$ = probability α

$\theta \leftarrow \theta$ = probability $1-\alpha$

$\theta_0, \theta_1, \theta_2, \theta_2, \theta_2, \theta_3, \theta_4, \ldots$

$$\tilde{\theta} = \frac{1}{N} \sum_{i=k}^{k+N} \theta_i$$

θ_4

Burn-in

Why does MH work?

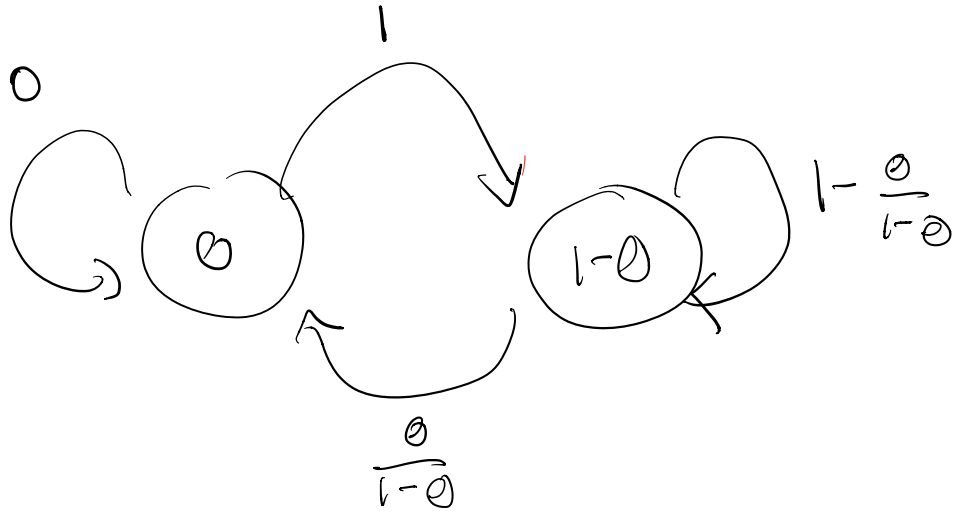
① There is a unique steady state

② By design, the steady state is the posterior distribution

System = just 2 states

positive probabilities

$[\theta, 1-\theta]$ $\theta < 0.5$



Solve stationary distribution

$$\begin{bmatrix} 0 & \theta/1-\theta \\ 1 & 1-\theta/1-\theta \end{bmatrix} \pi = \pi$$

$$\Rightarrow \quad \pi = \begin{bmatrix} \theta \\ 1-\theta \end{bmatrix} \leftarrow$$

Want $\pi(x)$ to be posterior $P(x)$

$0 \Leftrightarrow 0$

$$p(x)T(x \to y) = p(y)T(y \to x)$$

detailed balance $\Rightarrow$ steady state

$$\frac{T(x \to y)}{T(y \to x)} = \frac{p(y)}{p(x)}$$

$$\frac{q(x \to y)A(x \to y)}{q(y \to x)A(y \to x)} = \frac{p(y)}{p(x)}$$

$$\boxed{\frac{A(x \to y)}{A(y \to x)} = \frac{p(y)q(y \to x)}{p(x)q(x \to y)}}$$

Choose $A(x \to y) \to \min\left(1, \frac{p(y)q(y \to x)}{p(x)q(x \to y)}\right)$

$\Rightarrow$

$$\boxed{\frac{A(x \to y)}{A(y \to x)} = \frac{p(y)q(y \to x)}{p(x)q(x \to y)}} \checkmark$$

Metropolis

Metropolis-Hastings

Gibbs

$\theta_1, \theta_2, \theta_3$

$\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ $\quad p(\theta_1 | \theta_2, \theta_3, X) \rightarrow \theta_{-1}$

Suppose we can draw from full conditional distributions

Post N θ_i $\leftarrow$ $\boxed{p(\theta_i | \theta_{-i}, X)}$ $\leftarrow$ Need to be able to find this

we cycle through each parameter $\rightarrow$ at each step, the proposal is the posterior (for one parameter)

$\Rightarrow$ Always accept. ✓

$$\frac{P(y)\, g(y \to x)}{P(x)\, g(x \to y)} = 1$$

① $p(x_{-i}) = p(y_{-i})$ Apart from parameter i old & proposal state are identical

② $p(x_i | x_{-i})\, p(x_{-i}) = p(x_i, x_{-i}) = p(x)$ By definition

So

$$\frac{P(y)\, g(y \to x)}{P(x)\, g(x \to y)} = \frac{p(y_i | y_{-i})\, P(y_{-i})\, P(x_i | x_{-i})}{P(x_i | x_{-i})\, P(x_{-i})\, P(y_i | y_{-i})}$$

Gibbs

$\Delta\theta \sim N(0, \sigma)$

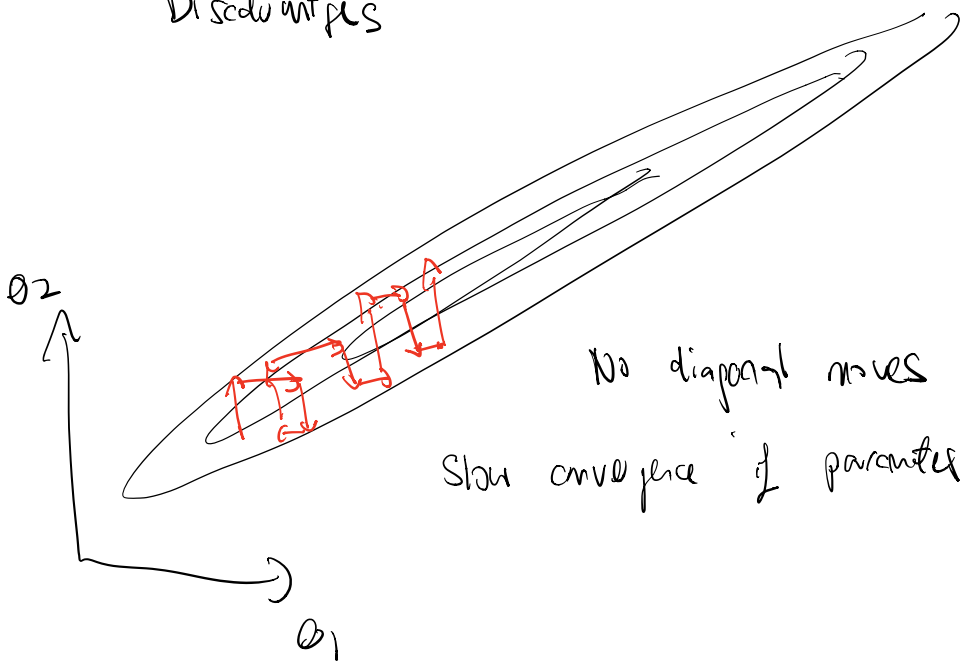
Advantage ① Always accept

② No tuning of proposal

$\sigma \ll 1$ 大√ 效X

$\sigma \gg 1$ 小↓

Disadvantages



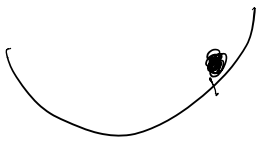
No diagonal moves

Slow convergence if parameters are correlated

Hamiltonian MC

From physics → energy conservation, path independence

E.g. Frictionless well



— KE + PE = constant

Particle ● → position } velocity } state
= unit mass x v

$$H(x,v) = K(v) + U(x)$$

Probability of state

$$p(x,v) \propto e^{-H(x,v)}$$

$$= e^{-U(x)-K(v)}$$

$$= e^{-U(x)} e^{-K(v)}$$

$$\propto p(x)\,p(v)$$

$$\Rightarrow \quad p(x,v) = p(x)\,p(v)$$

Contrast with

$p(x,v) = p(v|x)\,p(x)$

$= p(x|v)\,p(v)$

x, v independent

What happens in v does not affect x

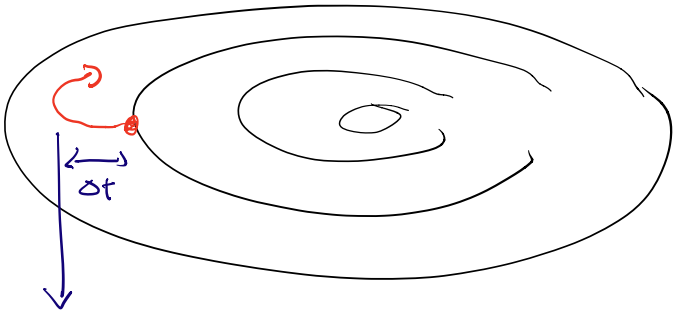
Proposal : Choose random v for particle, integrate forward in time, then use x^* as proposal

Acceptance ratio

$$\frac{e^{-U(x^*)-K(v^*)}}{e^{-U(x)-K(x)}} = e^{U(x)-U(x^*)+K(v)-K(v^*)}$$

acceptance $= 1$ if no numerical errors.

How long do we let particle travel?



Stop when there is a U-turn

⇓

No U-turn sampler (NUTS)