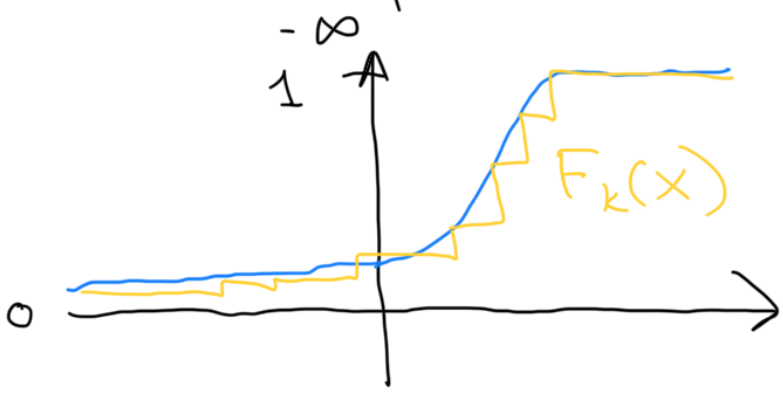
Sampling algorithms. Monte Carlo

$x \in \mathbb{R}$, $p(x)$ – a distribution of x

$F(t) = \int_{-\infty}^{t} p(x)\,dx$ – CDF of x



1. Empirical distribution function

Sum of delta functions

$$F_k(x) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}[x > x_i]$$

2. Dense steps for higher $\frac{\partial F(t)}{\partial t} = p(t)$

3. Kolmogorov – Smirnov test:

$$D_k = \sup_{x \in \mathbb{R}} |F_k(x) - F(x)|$$

$\sqrt{k} D_k \to$ Kolmogorov distribution

convergence in distribution

Ex. 1 $p(y|\vec{x}) = \int p(y|\vec{x},\vec{\theta})\, p(\vec{\theta}|D)\, d\vec{\theta} \approx$

(y – output, $\vec{x}$ – input, $p(\vec{\theta}|D)$ – posterior)

$$\approx \frac{1}{k} \sum_{i=1}^{k} p(y|\vec{x},\vec{\theta}_i),\ \vec{\theta}_i \sim p(\vec{\theta}|D)$$

- unbiased estimate
- variance decreases $O\left(\frac{1}{k}\right)$

$$\mathbb{E} f(x) = \int f(x) p(x)\,dx$$

$$\hat{F}_k f(x) = \int f(x)\,\hat{p}_k(x)\,dx = \frac{1}{k}\sum_{i=1}^{k} f(x_i)$$

$$\hat{p}_k(x) = \frac{1}{k}\sum_{i=1}^{k} \delta(x - x_i)$$

$$\mathbb{E}\left(\hat{F} f(x) - \hat{F}_k f(x)\right)^2 = \mathbb{E}\left(\hat{F} f(x) - \frac{1}{k}\sum_{i=1}^{k} f(x_i)\right)^2 =$$

$$= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left(\hat{F} f(x) - f(x_i)\right)\right]^2 = \frac{1}{k^2}\, k\,\mathbb{E}\left(\hat{F} f(x) - f(x_i)\right)^2 +$$

$$+ c\cdot\sum_{i \neq j}\mathbb{E}\left(\hat{F} f(x) - f(x_i)\right)\left(\hat{F} f(x) - f(x_j)\right) \to 0 =$$

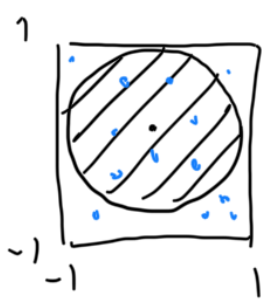
$$= \frac{1}{k}\,\mathrm{Var}\left(f(x)\right)$$

▲

- we also need estimates for variational inference

Ex.2 Find π, toy problem

$\Theta_1, \Theta_2 \sim \mathcal{U}([-1,1])$, ind.



$$V_i = \left[\left\| \underset{m}{\Theta}^{(i)} \right\|_2^2 \leq 1\right]$$

$$S_m = \frac{1}{m}\sum_{i=1} V_i \approx \frac{\pi}{4}$$

How to sample from $p(\vec{\Theta}|D)$?

We assume that we can sample from

Ex.0 $u \sim \mathcal{U}([0,1])$ - uniform distribution with finite support

But typically $p(\vec{\Theta}|D)$ is more complex

Ex.1 Exponential distribution

- inverse sampling

$$x = -\frac{\ln u}{\lambda}$$

/ 1 - e

$p(x) = \lambda \exp(-\lambda x)$

quantile

$x \in \mathbb{R}$

F_x - CDF

$u \in U([0,1])$

$F_x^{-1}(u) \sim p(x)$

Ex. Gaussian distribution,

Box-Muller transformation

$u_1 \sim U([0,1])$, $u_2 \sim U([0,1])$

polar coordinates: $z = \sqrt{-2\ln(u_1)}$

$\varphi = 2\pi u_2$

$\Rightarrow \quad z \in (0, +\infty)$, $\varphi \in (0, 2\pi)$

Cartesian coordinates: $x = z \sin\varphi$

$y = z \cos\varphi$

$\varphi \sim U([0, 2\pi])$

$z = g(u_1)$ - monotonic one to one transf.

$\ln(u_1) = -\frac{z^2}{2}$, $u_1 = \exp(-\frac{z^2}{2}) = g^{-1}(z)$

$$P(z_1 < z < z_2) = \int_{g^{-1}(z_1)}^{g^{-1}(z_2)} f_{u_1}(s)\,ds =$$

$$= -\int_{g^{-1}(z_1)}^{g^{-1}(z_2)} ds = \int_{z_1}^{z_2} -t\exp(-t^2/2)\,dt, \quad t = \sqrt{-2\ln s}$$

$\Rightarrow \quad f_R(z) = z\exp(-z^2/2)$

$x = g_x(z, \varphi)$, $y = g_y(z, \varphi)$

$$P(x_1 < x < x_2,\ y_1 < y < y_2) =$$

$$= \iint_A f_{z,\varphi}(u, v)\,du\,dv = \iint_A f_z(u)\,f_\varphi(v)\,du\,dv =$$

$$= \iint_A u \exp(-u^2/2) \frac{1}{2\pi} du\, dv \qquad (1)$$

$$p = u\cos(v), \quad q = u\sin(v)$$

$$u = \sqrt{p^2+q^2}, \quad v = \arctan(q/p)$$

$$du\,dv = \left|\det \frac{\partial(u,v)}{\partial(p,q)}\right| dp\,dq =$$

$$= \frac{1}{\sqrt{p^2+q^2}}\, dp\,dq$$

$$P(x_1 < x < x_2,\ y_1 < y < y_2) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} \underbrace{\frac{\sqrt{p^2+q^2}}{2\pi}}_{\text{from (1)}} \cdot \underbrace{\exp(-(p^2+q^2)/2)}_{x_2} \underbrace{\frac{1}{\sqrt{p^2+q^2}}}_{y_2} dp\,dq = \text{change of var.}$$

$$= \int_{x_1}^{} \frac{1}{\sqrt{2\pi}} \exp(-p^2/2)\, dp \int_{y_1}^{} \frac{1}{\sqrt{2\pi}} \exp(-q^2/2)\, dq =$$

$\Rightarrow$ x, y are independent and

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

$$f_y(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$$

For other distributions we can get something similar, if we know their analytical form, but in general for $p(\vec{\theta}\,|\,D)$ we know only numerator and it's hard to design something similar

---

Importance sampling

$$\mathbb{E}_p f(\vec{\theta}) \approx \sum_{k=1}^{K} p(\vec{\theta}_k)\, f(\vec{\theta}_k), \quad \vec{\theta}_k \sim \text{uniform}$$

$$\mathbb{E}_p f(\vec{\theta}) = \int f(\vec{\theta})\, p(\vec{\theta})\, d\vec{\theta} = \int f(\vec{\theta}) \frac{p(\vec{\theta})}{q(\vec{\theta})} q(\vec{\theta})\, d\vec{\theta} =$$

$$= \mathbb{E}_q \left[f(\vec{\theta}) \frac{p(\vec{\theta})}{q(\vec{\theta})} \right] \approx \frac{1}{K} \sum_{i=1}^{K} f(\vec{\theta}_i) \frac{p(\vec{\theta}_i)}{q(\vec{\theta}_i)},$$

we use the proposal distribution $q(\vec{\theta})$

$\vec{\theta}_i \sim q(\vec{\theta})$, $z_i = \frac{p(\vec{\theta}_i)}{q(\vec{\theta}_i)}$ - importance weights

But typically we know $p(\vec{\theta})$ up to normalization constant (e.g. $p(\vec{\theta}|\mathcal{D})$)

$$p(\vec{\theta}) = \frac{\tilde{p}(\vec{\theta})}{Z_p} \leftarrow \text{normalization constant}$$

$$\mathbb{E}_p f(\vec{\theta}) = \int f(\vec{\theta})\, p(\vec{\theta})\, d\vec{\theta} =$$

$$= \frac{Z_q}{Z_p} \frac{1}{K} \sum_{i=1}^{K} \tilde{z}_i f(\vec{\theta}_i), \quad \vec{\theta}_i \sim q(\vec{\theta}) = \frac{\tilde{q}(\vec{\theta})}{Z_q},$$

$$\tilde{z}_i = \frac{\tilde{p}(\vec{\theta}_i)}{\tilde{q}(\vec{\theta}_i)}$$

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q} \int \tilde{p}(\vec{\theta})\, d\vec{\theta} = \int \frac{\tilde{p}(\vec{\theta})}{\tilde{q}(\vec{\theta})} q(\vec{\theta})\, d\vec{\theta}$$

$$\approx \frac{1}{K} \sum_{i=1}^{K} \tilde{z}_i$$

$$\Rightarrow$$

$$\mathbb{E}_p f(\vec{\theta}) \approx \sum_{i=1}^{K} w_i f(\vec{\theta}_i), \quad w_i = \frac{\tilde{z}_i}{\sum_{j=1}^{K} \tilde{z}_j}$$

- big variance
- no diagnostics
- $q(\vec{\theta})$ should be close to $p(\vec{\theta})$

Markov Chain Monte Carlo

we don't know

Metropolis algorithm : sampling from $p(\vec{\theta}) = \frac{\tilde{p}(\vec{\theta})}{Z_p}$

1. Initialize $\vec{\theta}^{(0)}$
2. Repeat k times

depends on $\vec{\theta}^{t}$

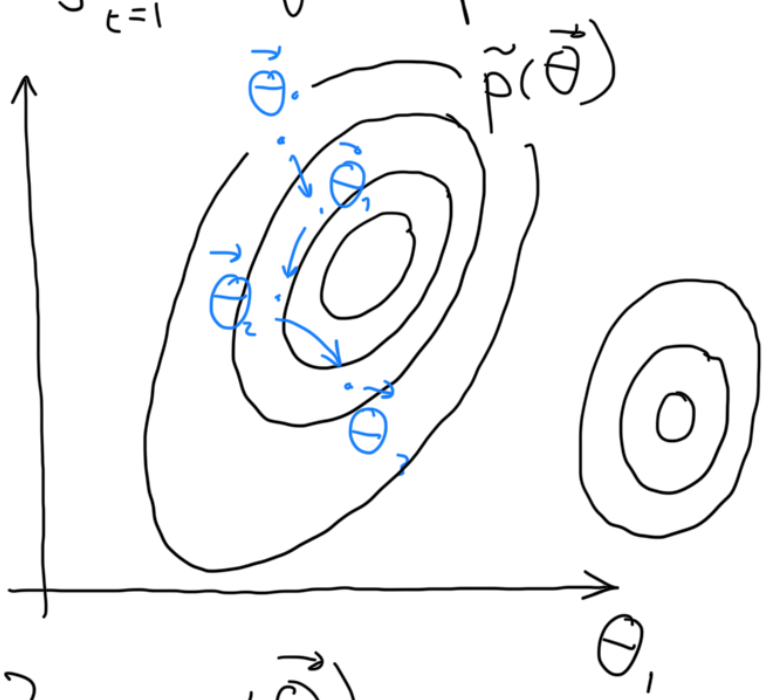
2.1. Sample $\vec{\theta}^{*}$ from the proposal distribution $q(\vec{\theta} \mid \vec{\theta}^{(t)})$

2.2. Accept $\vec{\theta}^{*}$ with probability

$$a = A(\vec{\theta}^{*}, \vec{\theta}^{(t)}) = \min\left(1, \frac{\tilde{p}(\vec{\theta}^{*})}{\tilde{p}(\vec{\theta}^{(t)})}\right):$$

$$\vec{\theta}^{(t+1)} = \begin{cases} \vec{\theta}^{(t)}, & u < a \\ \vec{\theta}^{*}, & u > a \end{cases}, \quad u \in \mathcal{U}([0,1])$$

Now we have $\{\vec{\theta}_t\}_{t=1}^{k}$ of dependent samples (we can just take every m-th sample)



1. $$\frac{\tilde{p}(\vec{\theta})}{\tilde{p}(\vec{\theta}')} = \frac{p(\vec{\theta}) \cdot Z_p}{p(\vec{\theta}') \cdot Z_p} = \frac{p(\vec{\theta})}{p(\vec{\theta}')}$$

2. $q(\vec{\theta} \mid \vec{\theta}') = q(\vec{\theta}' \mid \vec{\theta})$ - symmetric

If $q(\vec{\theta} \mid \vec{\theta}')$ is positive for all $\vec{\theta}, \vec{\theta}'$, then the distribution of $\vec{\theta}_t$ goes to $p(\vec{\theta})$ for $t \to \infty$ (sufficient cond.)

Ex. $q(\vec{\theta} \mid \vec{\theta}') = N(\vec{\theta}', \sigma^2 I_d) =$

$$= \frac{1}{(\sqrt{2\pi}\,\sigma)^d} \exp\left(-\frac{1}{2\sigma^2}\|\vec{\theta} - \vec{\theta}'\|_2^2\right)$$

We have a single parameter σ^2

σ^2 is small - high correlation between $\vec{\theta}^{(t)}, \vec{\theta}^{(t+1)}$

σ^2 is large - low acceptance prob.

Metropolis algorithm II (more general)

$\tilde{p}(\vec{\theta})$ - unnormalized density we can calculate

$q(\vec{\theta}' \mid \vec{\theta})$ - (unsymmetric) proposal distribution

1. Initial $\vec{\theta}^{(0)}$ from some distr.
2. k iterations
   - 2.1. $\vec{\theta}^* \sim q(\vec{\theta} \mid \vec{\theta}^{(t)})$
   - 2.2. Accept with probability $A(\vec{\theta}^*, \vec{\theta}^{(t)}) = \min\left(1, \frac{\tilde{p}(\vec{\theta}^*)\, q(\vec{\theta}^{(t)} \mid \vec{\theta}^*)}{\tilde{p}(\vec{\theta}^{(t)})\, q(\vec{\theta}^* \mid \vec{\theta}^{(t)})}\right)$

cancels if $q(\theta' \mid \theta)$ is symmetric

Let's prove, that this approach is correct

Markov chains

$\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(t)}, \vec{\theta}^{(t+1)}$ - a series of random variables

Def. A first order Markov chain is a series of random variables s.t.

$$p(\vec{\theta}^{(t+1)} \mid \vec{\theta}^{(1)} \ldots \vec{\theta}^{(t)}) = p(\vec{\theta}^{(t+1)} \mid \vec{\theta}^{(t)})$$

. the conditional density depends only on the previous state.

Ex. A. independent $\vec{\theta}^{(t)}$

$$p(\vec{\theta}^{(t+1)} \mid \vec{\theta}^{(1)}, \dots, \vec{\theta}^{(t)}) = p(\vec{\theta}^{(t+1)})$$

B. Gaussian Random Walk

$$p(\vec{\theta}^{(t+1)} \mid \vec{\theta}^{(1)}, \dots, \vec{\theta}^{(t)}) = \mathcal{N}(\vec{\theta} \mid \vec{\theta}^{(t)}, \delta^2 I)$$

How to fully specify Markov Chain

1. Initial distribution $p(\vec{\theta}^{(0)})$
2. Transition distribution

$$p(\vec{\theta}^{(m+1)} \mid \vec{\theta}^{(m)}) = T_m(\vec{\theta}^{(m)}, \vec{\theta}^{(m+1)})$$

If there's no dependence on m =>

homogeneous Markov Chains.

We'll consider only homogenous M.C.

Marginal distribution for $\vec{\theta}^{(m+1)}$:

$$p(\vec{\theta}^{(m+1)}) = \sum_{\vec{\theta}^{(m)}} p(\vec{\theta}^{(m+1)} \mid \vec{\theta}^{(m)}) p(\vec{\theta}^{(m)})$$

Def. A distribution is stationary (invariant), if each step in the chain doesn't change it

$$p^*(\vec{\theta}) = \sum_{\vec{\theta}'} T(\vec{\theta}', \vec{\theta}) p^*(\vec{\theta})$$

Ex. Identity transitions => all are invariant

A sufficient condition is called the detailed balance property for $p^*(\vec{\theta})$

$$p^*(\vec{\theta}) p(\vec{\theta} \mid \vec{\theta}') = p^*(\vec{\theta}') p(\vec{\theta} \mid \vec{\theta}')$$

Stat. The invariance property follows from the detailed balance property

$$\sum_{\vec{\Theta}'} p(\vec{\Theta}|\vec{\Theta}')\, p^*(\vec{\Theta}') = \sum_{\vec{\Theta}'} p(\vec{\Theta}'|\vec{\Theta})\, p^*(\vec{\Theta}) =$$

$$= p^*(\vec{\Theta}) \sum_{\vec{\Theta}'} p(\vec{\Theta}'|\vec{\Theta}) = p^*(\vec{\Theta})$$

A Markov chain that respects the detailed balance is said to be reversible.

We design a Markov chain for sampling:

1. the desired distribution is <u>invariant</u>
2. Irrespective of the choise of the initial distribution

$$p(\vec{\Theta}^{(t)}) \to p^*(\vec{\Theta}), \text{ as } t \to \infty$$

<u>Ergodicity</u> holds.

M.C. will be ergodic under weak restrictions on the invariant distr. and transition probabilities.

M.-H. invariant distribution

k possible options

accept. probability $A_k(\vec{\Theta}^*, \vec{\Theta}^{(t)}) = \min\left(1, \frac{\tilde{p}(\vec{\Theta}^*)\, q_k(\vec{\Theta}^{(t)}|\vec{\Theta}^*)}{\tilde{p}(\vec{\Theta}^{(t)})\, q_k(\vec{\Theta}^*|\vec{\Theta}^{(t)})}\right)$

<u>Stat.</u> $p(\vec{\Theta})$ is an invariant distribution of M.C.

We'll check the detailed balance

$$p(\vec{\Theta})\, q_k(\vec{\Theta}'|\vec{\Theta})\, A_k(\vec{\Theta}', \vec{\Theta}) =$$

$$= \min\left(p(\vec{\theta})\, q_k(\vec{\theta}' \mid \vec{\theta}),\ p(\vec{\theta}')\, q_k(\vec{\theta} \mid \vec{\theta}') \right) =$$

$$= \min\left(p(\vec{\theta}')\, q_k(\vec{\theta} \mid \vec{\theta}'),\ p(\vec{\theta})\, q_k(\vec{\theta}' \mid \vec{\theta}) \right) =$$

$$= p(\vec{\theta}')\, q_k(\vec{\theta} \mid \vec{\theta}')\, A_k(\vec{\theta}, \vec{\theta}') \quad \blacktriangle$$

Ergodicity - also holds

Ex. Non-symmetric Gaussian

number of steps to have "independent" samples: $\left(\varepsilon_{max} / \varepsilon_{min}\right)^2$

Gibbs sampling

$$p(\vec{\theta}) = p(\theta_1, \dots, \theta_d)$$

1. initialize $\vec{\theta}^{(0)} = \left(\theta_1^{(0)}, \dots, \theta_d^{(0)}\right)$
2. $\tau = 1, \dots, T$
   - sample $\theta_1^{(\tau+1)} \sim p\left(\theta_1 \mid \theta_2^{(\tau)}, \dots, \theta_d^{(\tau)}\right)$
   - sample $\theta_2^{(\tau+1)} \sim p\left(\theta_2 \mid \theta_1^{(\tau+1)}, \theta_3^{(\tau)}, \dots, \theta_d^{(\tau)}\right)$
   - $\vdots$
   - sample $\theta_d^{(\tau+1)} \sim p\left(\theta_d \mid \theta_1^{(\tau+1)}, \theta_2^{(\tau+1)}, \dots, \theta_{d-1}^{(\tau+1)}\right)$

- Markov chain type sampling

Stat. Is it invariant for $p(\vec{\theta})$

$p(\vec{\theta}')$ - we change 1st component

We'll prove, that marginal and conditional are the same

$$p(\vec{\theta}') = p(\theta_1' \mid \vec{\theta}')\, p(\vec{\theta}') = p(\theta_1' \mid \vec{\theta}')\, p(\theta_1', \vec{\theta}_{-1}')$$

$$p(\vec{\theta}) = p(\theta_1 \mid \vec{\theta}_{/1})\, p(\vec{\theta}_{/1}) = p(\theta_1 \mid \vec{\theta}_{/1}) \cdot p(\vec{\theta}_{/1}) = p(\vec{\theta})$$ ▲

Stat. It is a Metz.-Hast. algorithm

- We consider $q_k(\vec{\theta}^* \mid \vec{\theta}) = p(\theta_k^* \mid \vec{\theta}_{/k})$
  - conditional, $\vec{\theta}^*_{/k} = \vec{\theta}_{/k}$

$$A_k(\vec{\theta}^*, \vec{\theta}) = \frac{p(\vec{\theta}^*)\, q_k(\vec{\theta} \mid \vec{\theta}^*)}{p(\vec{\theta})\, q_k(\vec{\theta}^* \mid \vec{\theta})} =$$

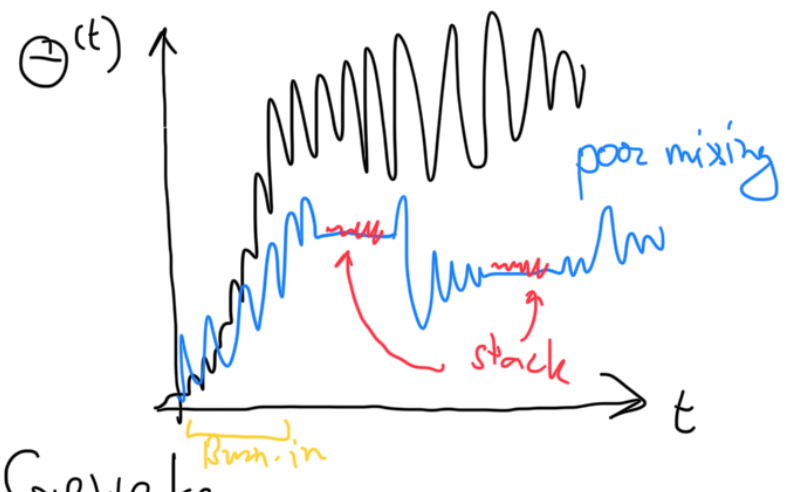
$$= \frac{p(\theta_k^* \mid \vec{\theta}_{/k})\, p(\vec{\theta}_{/k})\, p(\theta_k \mid \vec{\theta}_{/k})}{p(\theta_k \mid \vec{\theta}_{/k})\, p(\vec{\theta}_{/k})\, p(\theta_k^* \mid \vec{\theta}_{/k})} = 1$$ ▲

Design choices MCMC

Thinning: • skip steps, more independence

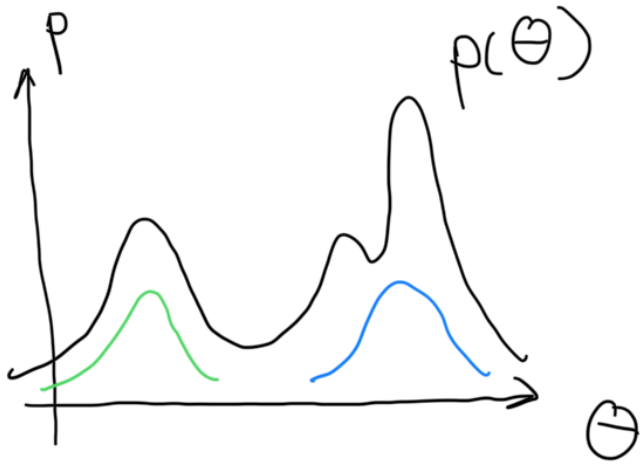
discarding: • burn-in period: don't depend on the initial distribution

adaptive proposals, multi-start:
- global / local maxima
- proposal distribution

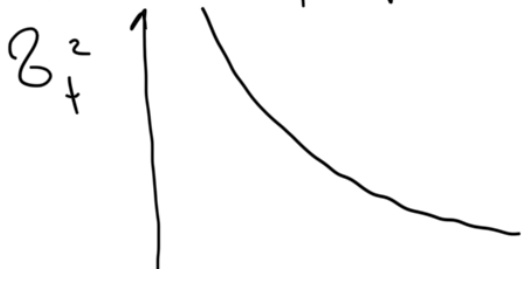


Diagnosis of MCMC

- eye test
- formal tests: - Geweke
  - Heidelberger - Welch
  - Raftery - Lewis



Adaptive proposal:



Next time:

- Hamiltonian Monte-Carlo
- NUT