Examples of Metropolis-Hastings algorithms

① Metropolis algorithm

$$q(x'|x) = \mathcal{N}(x', \mathbf{C})$$

$$q(x'|x) \propto \exp\left(-\tfrac{1}{2}(x-x')^T C^{-1}(x-x')\right)$$

$$q(x|x') \propto \exp\left(-\tfrac{1}{2}(x-x')^T C^{-1}(x-x')\right) \propto q(x'|x)$$

$\Rightarrow$ acceptance ratio: $\alpha = \min\left\{1, \frac{p(x')}{p(x)}\right\}$

↳ Moves are likely towards regions of higher probability ($p(x') > p(x)$)

↳ If $p(x') < p(x)$, chain will likely stay put.

How to chose C?

A popular choice: set $C = I$

↳ Random Walk Metropolis.

This is great and very popular because it is

(i) easy to implement

(ii) problem independent (target only shows up in α)

(iii) Means also that it is slow for most problems.

An "optimal choice": set $C = \sigma I$

tune σ so that acceptance probability ~ 0.24

(Roberts et al 1997)

For high-dimensional problems, some paper suggests that $\sigma \sim \frac{1}{n}$

You can see why this becomes problematic in large n problems:

Proposed sample: $x' = x_k + \sqrt{\sigma}\, I$

$= x_k + \frac{1}{\sqrt{n}} I$

the proposed move is tiny if n is large

The small moves cause the samples to be correlated, which we will see will be a problem.

Other MH algorithms use different, often problem specific, proposals but this does not help with such high-dimensional issues.

Situation is similar as in importance sampling: "good" proposals are useful, but it does not "break" this issue. Localization does.

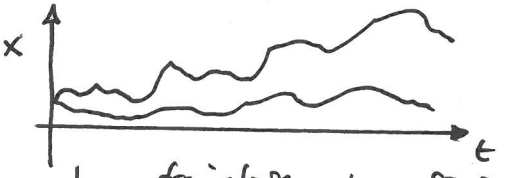
---

(2) Langevin MCMC

This is a stochastic differential eqn:

$$dx = f(x)\,dt + \sqrt{2}\,dW$$

↑ Brownian motion.



two trajectories of an SDE starting at the same initial condition

Def of BM:
1. $W(0) = 0$
2. $W(t+T) - W(t) \approx N(0, T)$
3. $W(t)$ is almost surely continuous (not differentiable)
4. $W(t)$ has independent increments.

$0 \le t_1 \le t_2 \le t_3 \le t_4$

then $W(t_1) - W(t_2)$ and $W(t_4) - W(t_3)$ are independent random variables.

Note: $X(t)$ is a random variable with distribution $P_t(x)$

Question: What is the distribution of x as $t \to \infty$

Answer: $P_\infty \propto \exp(U(x))$; $\nabla U(x) = f(x)$

↑ "potential"

Idea: Generate samples from a given target distribution $p(x)$ by solving the SDE

$$dx = \nabla \log p(x)\, dt + \sqrt{2}\, dW$$

This is good because it is problem dependent

Problems: solving SDE is hard.

- solving SDE numerically is also difficult to do accurately.

An "easy" solver is a variation of the Euler scheme

$$dW \approx \Delta W \approx N(0, \Delta t)$$
$$dt \approx \Delta t$$

Then: $\circledast$ $x_k = x_{k-1} + \nabla \log p(x_{k-1})\, \Delta t + \sqrt{2\Delta t}\, \xi_k$

$\xi_k \sim N(0, I)$ iid

How to use it?

"Simulate" SDE using $\circledast$ (or another numerical scheme) until you reach k^* which is large.

Repeat N_e times.

Samples are approximately distributed as $p(x)$.

The approximation is due to the discretization of the SDE and errors go to zero as $\Delta t \to 0$.

But a small Δt slows down the algorithm $\Rightarrow$ you need to simulate to a large $T = N\Delta t$ for each sample.

A great thing is that the algorithm is very easy to parallelize.

3) Metropolis-adjusted Langevin algorithm "MALA"

Idea: use Langevin algorithm in discrete time as a proposal, get rid of discretization error by an accept/reject step.

Proposal: $x' = x_k + \nabla \log p(x_k)\,\Delta t + \sqrt{2\Delta t}\,\xi \quad , \xi \sim N(0, I)$

$\hookrightarrow q(x' | x_k) = N(x_k + \nabla \log p(x_k)\,\Delta t,\ 2\Delta t)$

$$q(x'|x_k) \propto \exp\left(-\frac{1}{4\Delta t}\left\| x' - x_k - \nabla \log p(x_k)\,\Delta t \right\|^2\right)$$

$$\alpha = \min\left\{1, \frac{p(x')\,q(x_k|x')}{p(x_k)\,q(x'|x_k)}\right\}$$

How to choose Δt, especially in high-dimensions?

Roberts & Rosenthal 1998: optimal acceptance ratio is 0.574

and $\Delta t \sim n^{-1/3}$

↑ much better than RWM which was n^{-1}.

④ Hamiltonian MCMC (Hybrid Monte Carlo)

Hamiltonian dynamics.

$H(q, p)$ — position, momentum.

Eqns of motion

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

Hamiltonian is Conserved:

$$\frac{d}{dt}H = \sum_i \underbrace{\frac{\partial H}{\partial q_i}\frac{dq_i}{dt}}_{\frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i}} + \underbrace{\frac{\partial H}{\partial p_i}\frac{dp_i}{dt}}_{= -\frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i}} = 0$$

Distribution of q and p? $p(q,p) \propto \exp(-H(q,p))$

We often have: $H(q,p) = U(q) + K(p)$

U: potential energy; K: kinetic energy.

Example: $U(q) = \frac{1}{2}q^2$, $K(p) = \frac{p^2}{2m}$

$$\frac{\partial H}{\partial p} = \frac{\partial K}{\partial p} = p/m \quad (= v, \text{ velocity})$$

$$\frac{\partial H}{\partial q} = \frac{\partial U}{\partial q} = q$$

EOM: $\frac{dq}{dt} = p/m$; $\frac{dp}{dt} = -q$

In this case: $p(q,p) \propto \exp(-H(q,p)) \propto \exp(-U(q))\exp(-K(p))$

$\Rightarrow q, p$ are independent!

Idea of HMC pick $U(x) = -\log p(x)$, make up a kinetic energy $K(y)$

solve Hamiltonian eqns of motion

$$\frac{dx}{dt} = \frac{\partial K}{\partial y}$$

$$\frac{dy}{dt} = -\frac{\partial U}{\partial x}$$

$\exp(+\log p(x)) = p(x)$

marginal distribution of x: ~~Boltzmann~~ ✓

Similar to Langevin, same problems: discretizing Hamiltonian dynamics is hard because you need a solver that preserves H. The typical ones won't be good enough, you need symplectic integrators + you have errors due to discretization

HMC algorithm: use Leapfrog to discretize the dynamics.

$H(x,y) = \frac{1}{2} y^T y + F(x)$ where $F(x) = -\log p(x)$

$p(x,y) \propto \exp(-\frac{1}{2} y^T y) \exp(-F(x)) \Rightarrow y \sim N(0,I)$!

Leap-frog: given x sample $y \sim N(0,I)$

$$y_{h/2} = y - \frac{h}{2} \nabla F(x)$$

$$x_h = x + h\, y_{h/2}$$

$$y_h = y_{h/2} - \frac{h}{2} \nabla F(x_h)$$

x_h, y_h are proposed states.

accept with prob: $\alpha = \min\left\{1, \frac{\exp(-H(x_h, y_h))}{\exp(-H(x,y))}\right\}$.

How to pick step-size?

acceptance prob. ≈ 0.651

$h \approx n^{-1/4}$ even better than MALA!

Recap:

	Acceptance prob.	step size
RWM	0.24	n^{-1}
MALA	0.574	$n^{-1/3}$
HMC	0.651	$n^{-1/4}$

Beskos et al. 2011

---

⑤ pCN MCMC. pCN = preconditioned Crank Nicholson (pde solver)

Variation of RWM:

$$x' = \sqrt{1-\beta^2}\, x + \beta \xi, \quad \xi \sim N(0, I)$$

$$q(x|x') \propto \exp\left(-\frac{1}{2\beta^2}\left(x - \sqrt{1-\beta^2}\,x'\right)^T C^{-1}\left(x - \sqrt{1-\beta^2}\,x'\right)\right)$$

$$\propto \exp\left(-\frac{1}{2\beta^2}\left(x^T C^{-1} x - 2x^T C^{-1}\sqrt{1-\beta^2}\,x' + (1-\beta^2)\,x'^T C^{-1} x'\right)\right)$$

$$q(x'|x) \propto \exp\left(-\frac{1}{2\beta^2}\left(x'^T C^{-1} x' - 2x'^T C^{-1} x\sqrt{1-\beta^2} + (1-\beta^2)\,x^T C^{-1} x\right)\right)$$

$$\frac{q(x|x')}{q(x'|x)} = \exp\left(-\frac{1}{2\beta^2}\left(x^T C^{-1} x - x'^T C^{-1} x' + (1-\beta^2)\left(x'^T C^{-1} x - x^T C^{-1} x\right)\right)\right)$$

$$= \exp\left(-\frac{1}{2\beta^2}\left(-\beta^2\left(x'^T C^{-1} x' - x^T C^{-1} x\right)\right)\right)$$

$$= \exp\left(-\frac{1}{2}\left(x^T C^{-1} x - x'^T C^{-1} x'\right)\right)$$

Assume that: $p(x) \propto \exp\left(-\frac{1}{2} x^T C^{-1} x - \phi(x)\right)$

Then: $\frac{p(x')}{p(x)} \propto \exp\left(-\frac{1}{2}\left(x'^T C^{-1} x' - x^T C^{-1} x\right) - \phi(x') + \phi(x)\right)$.

Acceptance ratio:

$$\alpha = \min\left\{1, \underbrace{\frac{p(x')\, q(x|x')}{p(x)\, q(x'|x)}}\right\}$$

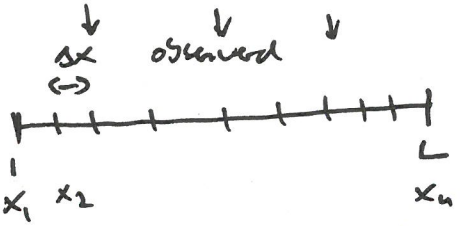
$$= \exp\left(-\frac{1}{2}\left(x'^T C^{-1} x' - x^T C^{-1} x\right) - \phi(x') + \phi(x)\right)$$

$$\cdot \exp\left(-\frac{1}{2}\left(x^T C^{-1} x - x'^T C^{-1} x'\right)\right)$$

$$= \exp\left(\phi(x) - \phi(x')\right)$$

$$\Rightarrow \alpha = \min\left\{1, \exp\left(\phi(x) - \phi(x')\right)\right\}$$

It's a small modification of RWM, but it can perform better if random variable to be estimated is as follows:



x is continuous fctn on $[0, L]$ (or higher D version of th.3)

$n = L/\Delta x$

Fixed observations at S locations ($S < n$)

As $\Delta x \to 0$ we approach the continuous function / infinite dimensional random variable.

Then we keep S fixed, then pCN performs well as $\Delta x \to 0$ and $n \to \infty$.

Th.3 does not mean that pCN is a good algorithm. When L is large and S is large pCN might not perform so well.

⑥ Heat bath / partial resampling / Gibbs sampling

Single Component Metropolis-Hastings

↳ Don't update all of x "at once", update step-by-step

We discuss updating one component of x at a time, but this can also be done in "blocks" (chunks of x).

Note: Single component proposes *local* moves, so there are probably connections w.r.t. localization.

Notation: $x_k = (x_k^1, x_k^2, \dots, x_k^n)$ ← location; ↑ time; state of Markov chain at time k

x_k^i is a scalar

$x_k^{\setminus i} = \{x_k^1, x_k^2, \dots, x_k^{i-1}, x_k^{i+1}, \dots, x_k^n\}$ — everything except its component.

At time k: iterate over blocks.

1st iterate: $\hat{x}^1 \sim q_1(x^1 \mid x_k^{\setminus 1})$

accept with prob:

$$\alpha = \min\left\{1, \frac{p(\hat{x}^1 \mid x_k^{\setminus 1})\, q_1(x_k^1 \mid \hat{x}_k^1, x_k^{\setminus 1})}{p(x^1 \mid x_k^{\setminus 1})\, q_1(\hat{x}_k^1 \mid x_k^{\setminus 1}, x_k^1)}\right\}$$

↳ update x_k^1 to x_{k+1}^1, if accepted, by $\hat{x}^1$, otherwise set $x_{k+1}^1 = x_k^1$.

2nd iterate: $\hat{x}^2 \sim q_2(x^2 \mid x_{k+1}^1, x_k^{\setminus 2})$ ↑ updated state!

$\Big\downarrow$

we need more notation.

Define: $x_k^{\setminus i} = \{ \underbrace{x_{k+1}^1, x_{k+1}^2, \ldots, x_{k+1}^{i-1}}_{\text{updated}}, \underbrace{x_k^{i+1}, \ldots x_k^n}_{\text{not updated.}} \}$

At ith iteration:

$$\hat{x}^i \sim q_i(\cdot \mid x_k^{\setminus i})$$

$$\alpha = \min\left\{1, \frac{p(\hat{x}^i \mid x_k^{\setminus i})}{p(x^i \mid x_k^{\setminus i})} \frac{q_i(x^i \mid x_k^{\setminus i})}{q_i(\hat{x}^i \mid x_k^{\setminus i})}\right\}.$$

In pseudo code:

for samples : k

for blocks/coordinates : i

$q_i(\hat{x}^i \mid x_k^{\setminus i})$

$$\alpha = \min\left\{1, \frac{p(\hat{x}^i \mid x_k^{\setminus i})}{p(x^i \mid x_k^{\setminus i})} \frac{q_i(x^i \mid x_k^{\setminus i})}{q_i(\hat{x}^i \mid x_k^{\setminus i})}\right\}$$

end

end.

This can work well, but requires that we know the conditional distributions $p(x^i \mid x^{\setminus i})$.

This is sometimes the case, but more often than not it is <u>not</u> the case.

The heat bath / partial resampling / Gibbs sampler is a special case of a single-component MH sampler:

We chose: $q_i(\cdot \mid x_{\sim i}) = p(\cdot \mid x_{\sim i})$

↳ proposal is full conditional of the target distribution.

↳ We accept every move (check anyway!)

This has even stricter assumptions:

(i) we need to know how to <u>sample</u> the conditional distributions.

↳ This is easy for Gaussians, but often it is quite difficult.

<u>Note:</u> For the "Bissel" problem of sampling a high-dimensional isotropic Gaussian this sample is great!

↳ This is another hint that there may be deep connections between localization and Gibbs sampling / single component MH.