Main Topic: How to Optimise MCMC Choices?

In theory, MCMC works with essentially <u>any</u> update rules, as long as they leave π stationary.

- <u>Any</u> symmetric proposal distribution Q. (Choices!)
- <u>Non</u>-symmetric proposals, with a suitably modified acceptance probability. ("Metropolis-Hastings") (e.g. Independent, Langevin)
- Update one coordinate at a time. ("Componentwise")
- Update from full conditional distributions. ("Gibbs Sampler")

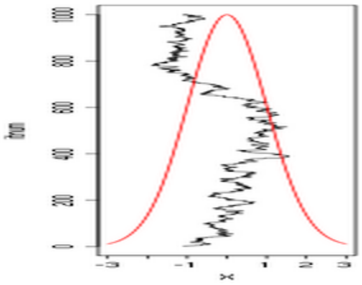
But what choice works <u>best</u>? e.g. What γ in [APPLET]?

- If γ too small (say, $\gamma = 1$), then usually accept, but move very slowly. (Bad.)
- If γ too large (say, $\gamma = 50$), then usually $\pi(Y_{n+1}) = 0$, i.e. hardly ever accept. (Bad.)
- Best γ is <u>between</u> the two extremes, i.e. acceptance rate should be far from 0 <u>and</u> far from 1. ("Goldilocks Principle")

Example: Metropolis for N(0,1)

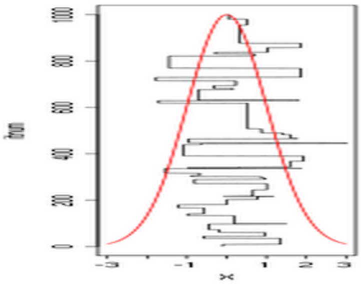
Target $\pi = N(0,1)$. Proposal $Q(x,\cdot) = N(x,\sigma^2)$.

How to choose σ? Big? Small? What <u>acceptance rate (A.R.)</u>?



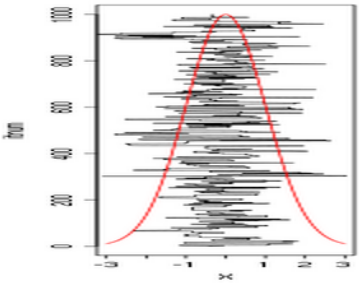
$\sigma = 0.1$?
too small!

A.R. $= 0.962$



$\sigma = 25$?
too big!

A.R. $= 0.052$



$\sigma = 2.38$?
just right!

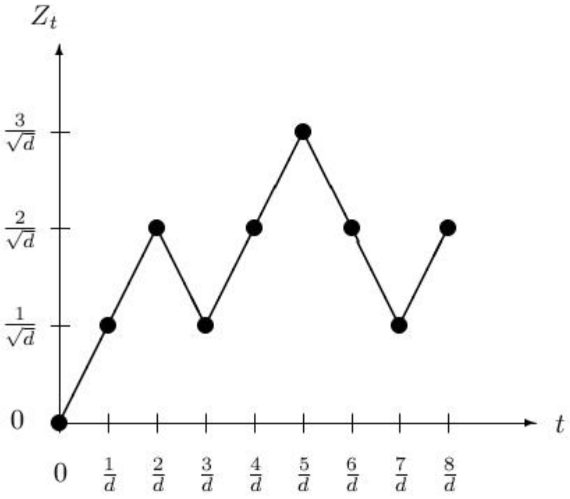
A.R. $= 0.441$

The Goldilocks Principle in action!

What about higher-dimensional examples? If d increases, then σ should: decrease. But how quickly? On what scale? Theory?

Theoretical Progress: Diffusion Limits

Recall: if $\{X_n\}$ is simple random walk, and $Z_t = d^{-1/2} X_{dt}$ (i.e., we speed up time, and shrink space), then as $d \to \infty$, the process $\{Z_t\}$ converges to Brownian motion (i.e., a diffusion). [GRAPHS]



Do similar limits hold for a Metropolis algorithm, in dimension d, as $d \to \infty$? Yes!

Diffusion Limits for the Metropolis Algorithm

[Roberts, Gelman, Gilks, AAP 1997]

- Consider a d-dimensional Metropolis algorithm $\{X_t^d\}_{t \geq 0}$, with proposal distribution $N(x, (\ell^2/d) I_d)$ for some fixed $\ell > 0$ (i.e., with proposal size shrinking as $1/\sqrt{d}$).

- Assume it starts in stationarity, i.e. $X_0^d \sim \pi$.

- Let $U_t^d = X^d_{PP(td),1}$ be the first component of the algorithm, at time $t \times d$ (i.e., U^d is sped up by a factor of d, and is converted to continuous-time via a Poisson Process).

- Assume (for now) that the target density π^d takes on a very special/unrealistic form, namely $\pi^d(x) = \prod_{i=1}^d f(x_i)$ where f is a fixed positive one-dimensional well-behaved (i.e., f'/f Lipschitz, $\mathbf{E}_f[(f'/f)^8] < \infty$, $\mathbf{E}_f[(f''/f)^4] < \infty$) density function.

- Then as $d \to \infty$, the process U^d converges (weakly, in the Skorokhod topology) to a fixed one-dimensional diffusion process U, defined by . . .

Diffusion Limits for Metropolis (cont'd)

- This limiting process U has dynamics

$$dU_t \;=\; \sqrt{h(\ell)}\; dB_t \,+\, h(\ell)\, \frac{f'(U_t)}{2\, f(U_t)}\, dt\,,$$

where $h(\ell) = 2\,\ell^2\,\Phi(-\ell\,\sqrt{\mathcal{I}}\,/2)$ with $\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$ and $\mathcal{I} = \mathbf{E}_f[(f'/f)^2]$.

- The process U is thus a Langevin diffusion, with stationary density f, and "speed" $h(\ell)$.

- Indeed, equivalently, $U_t = V_{h(\ell)\,t}$ is a speeded up (by a factor of $h(\ell)$) version of a Langevin diffusion V of unit speed, satisfying

$$dV_t \;=\; dB_t \,+\, \frac{f'(V_t)}{2\, f(V_t)}\, dt\,.$$

- So, to optimise the algorithm, we should maximise $h(\ell)$.
- The maximisation gives: $\ell_{opt} \doteq 2.38/\sqrt{\mathcal{I}}$.
- Then we compute that: $AR(\ell_{opt}) \doteq 0.234$. (constant!)

Diffusion Limits for Metropolis (cont'd)

- So, for a Metropolis algorithm in d dimensions, with $Q(x,\cdot) = N(x, \sigma^2 I_d)$, it is optimal to choose $\sigma^2 = \ell_{opt}^2\,/\,d$ $\doteq (2.38)^2\,/\,\mathcal{I}d$, corresponding to an (optimal) acceptance rate of 0.234. Clear, simple "0.234" rule. Good! Useful! (Used in BUGS!)

- The unrealistic form of π^d was later generalised to: inhomogeneous product form (Bédard & R., CJS 2008), infinite-dimensional absolutely continuous distributions (Stuart et al.), discrete hypercubes (Roberts, Stoch Rep 1998), spherical targets (Neal and Roberts, MCAP 2008), elliptically symmetric targets (Sherlock and Roberts, Bernoulli 2009), and discontinuous targets (Neal et al., AAP 2012).

- Numerical studies (e.g. Roberts and R., Stat Sci 2001): same optimality appears to "approximately" hold for more general π^d.

- Different optimal AR of 0.574 for Langevin diffusion algorithms (Roberts & R., JRSSB 1998).

New Generalisations?

(Yang, Negrea, Roberts, R., work in progress)

In the original RGG result, the unrealistic i.i.d. nature of π^d was used to apply Laws of Large Numbers when taking limits of the generators of the processes U^d.

Can the same proof techniques be used under weaker conditions?

It appears that if, as $d \to \infty$:

- in π^d, the dependence of x_1 on $x_2, \ldots, x_d$ goes to zero, and
- π^d and its derivatives satisfy strong moment order bounds,

then diffusion limits similar to the i.i.d. case still hold.

In particular, 0.234 is still the optimal acceptance rate.

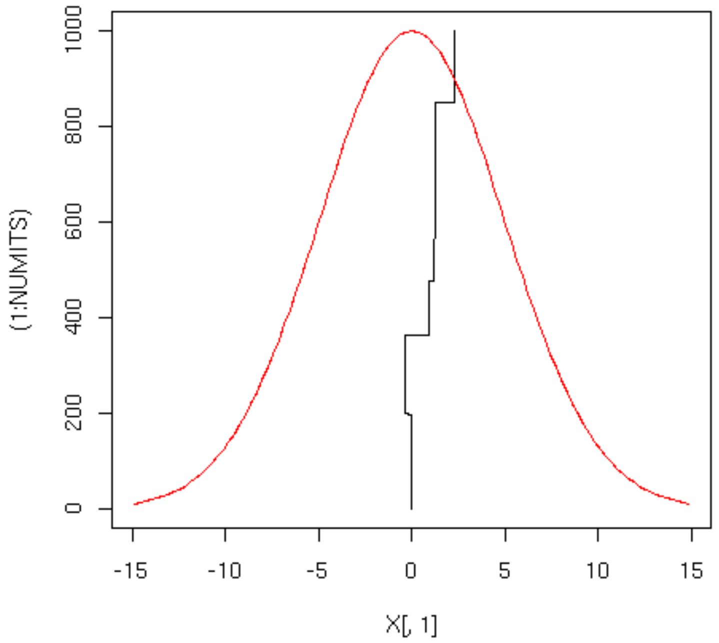
Anyway, 0.234 is a very useful rule of thumb.

But it is just a "one-dimensional" guideline.

What about further optimality, beyond "0.234"?

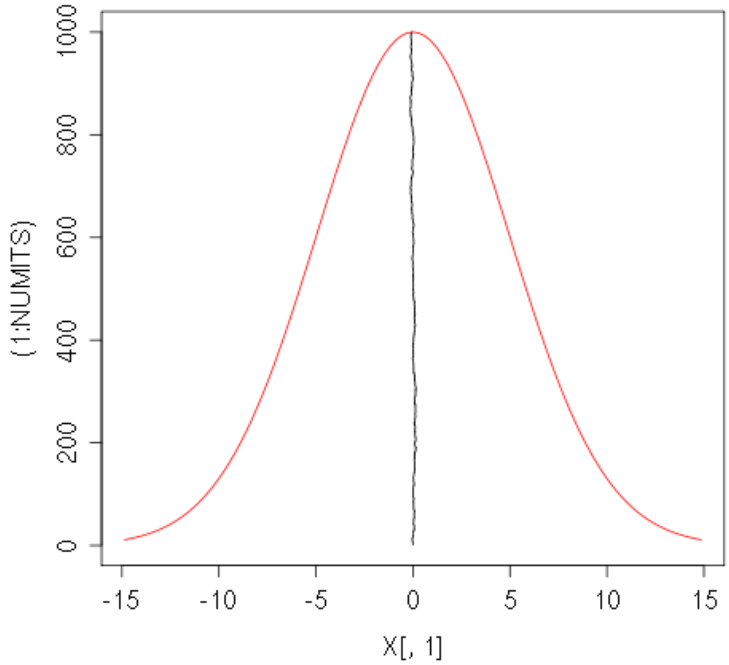
Example: $\pi = N(0, \Sigma)$ in dimension 20

First try: $Q(x, \cdot) = N(x, I_{20})$ (A.R. $= 0.006$)



Horrible: $\Sigma_{11} = 24.54$, $E(X_1^2) = 1.50$. Need smaller proposal!

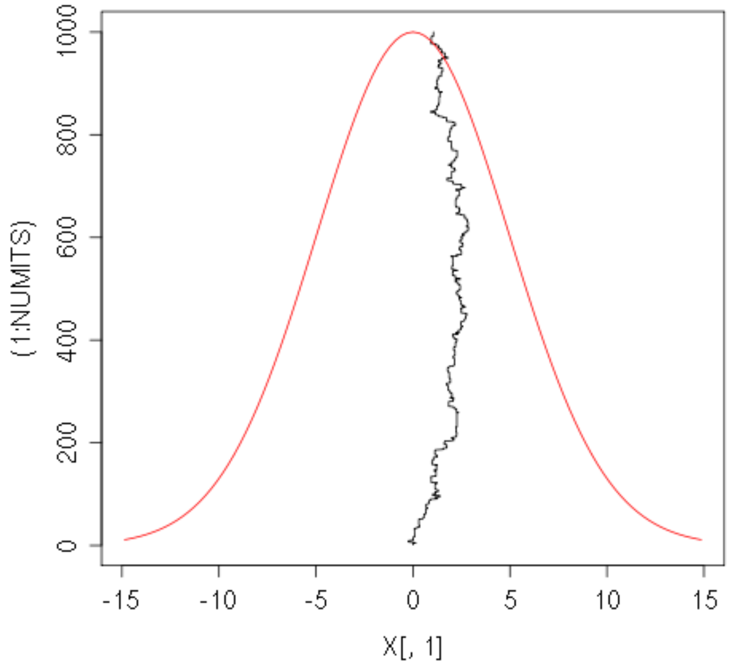
Second try: $Q(x,\cdot) = N\Big(x, (0.0001)^2 I_{20}\Big)$ (A.R.=0.9996)



Also horrible: $\Sigma_{11} = 24.54$, $E(X_1^2) = 0.0053$.

Need bigger proposal!

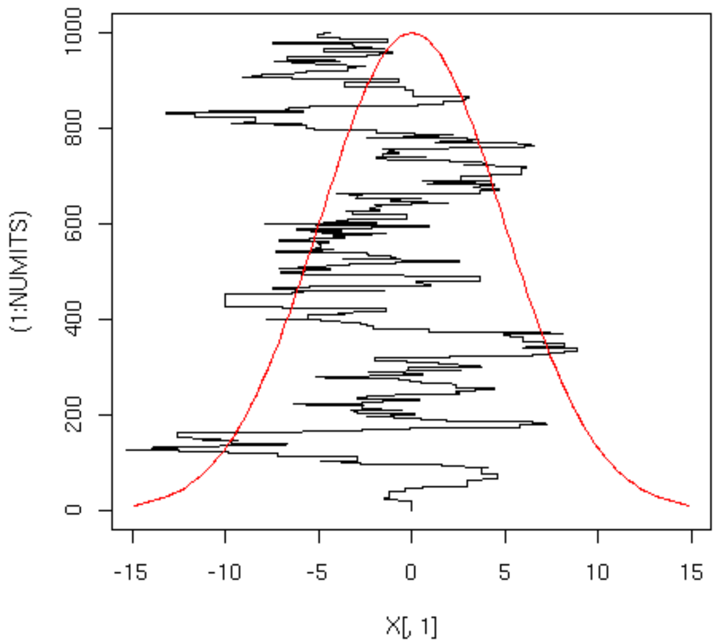
Third try: $Q(x,\cdot) = N\Big(x, (0.02)^2 I_{20}\Big)$ (A.R.=0.234)



Still terrible: $\Sigma_{11} = 24.54$, $E(X_1^2) = 3.63$.

But acceptance rate is “just right”. What gives?

Fourth try: $Q(x,\cdot) = N\Big(x, [(2.38)^2/20]\,\Sigma\Big)$ (A.R.=0.263)



Much better: $\Sigma_{11} = 24.54$, $E(X_1^2) = 25.82$.

Not perfect, but fairly good. Why?

Optimising the Proposal Covariance (Shape)

Theorem [Roberts and R., Stat Sci 2001]: If π is any orthogonal transform of any density satisfying the RGG conditions, then the optimal Gaussian proposal distribution as $d \to \infty$ is:

$$Q(x,\cdot) \;=\; N\Big(x,\; [(2.38)^2/d]\,\Sigma_t\Big)$$

where Σ_t is the target covariance. (Not $N(x,\, \sigma^2 I_d)$.)

So, want proposal covariance proportional to target covariance!

The corresponding asymptotic acceptance rate is again 0.234.

This turns out to be nearly optimal for many other high-dimensional densities, too. Very useful advice . . . if Σ_t is known!

But what if the target covariance Σ_t is unknown?

Can we make use of this optimality result anyway?

Perhaps . . . if we “adapt” . . . (coming soon!).

Implications for Computational Complexity

- Above results say, if we speed up the Metropolis algorithm by a factor of $O(d)$, then it converges to a dimension-free diffusion, and hence must converge in time $O(1)$.

- So, this seems to imply that Metropolis converges in $O(d)$. Right?

- Problem #1: Result is only for very special forms of the target π. (But we're working to generalise this!)

- Problem #2: Result just gives weak convergence, not total variation distance. (But we can work with that!)

- Problem #3: How to define computational complexity on continuous unbounded state spaces? What initial distribution should be used? (Can't use "worst case".)

What to do?

Weak Convergence Complexity Result

- Use the Kantorovich-Rubinstein (KR) distance measure,

$$\|\mathcal{L}_x(X_t) - \pi\|_{KR} \ := \ \sup_{f \in \mathrm{Lip}_1^1} \Big| \mathbf{E}_x[f(X_t) - \pi(f) \Big|$$

where $\mathrm{Lip}_1^1 = \{f : \mathcal{X} \to \mathbf{R},\ |f(x)| \le 1,\ |f(x) - f(y)| \le dist(x,y)\}$, which metricises weak convergence.

- And average over starting values $X_0 \sim \pi$, i.e. use

$$\mathbf{E}_{X_0 \sim \pi} \|\mathcal{L}_{X_0}(X_t) - \pi\|_{KR} \ := \ \int_{x \in \mathcal{X}} \pi(dx)\ \|\mathcal{L}_x(X_t) - \pi\|_{KR}\,.$$

- Theorem [Roberts and Rosenthal, JAP 2016]: If $X^{(d)} \to X^{(\infty)}$ weakly, for any choice of $X_0^{(d)}$, and $X^{(\infty)}$ is càdlàg (or continuous), and $X^{(\infty)} \to \pi$, then $\mathbf{E}_{X_0^{(d)} \sim \pi} \|\mathcal{L}_{X_0^{(d)}}(X_t^{(d)}) - \pi\|_{KR} \to 0$ in $O(1)$ time, i.e. for any $\epsilon > 0$, there are $D < \infty$ and $T < \infty$ such that

$$\mathbf{E}_{X_0^{(d)} \sim \pi} \|\mathcal{L}_{X_0^{(d)}}(X_t^{(d)}) - \pi\|_{KR} < \epsilon\,, \quad \forall\ t \ge T,\ d \ge D\,.$$

Computational Complexity of Metropolis

- Combining this complexity result with the Metropolis weak convergence results immediately shows that:
  - The speeded-up processes U_t^d converge to π in $O(1)$ time.
- But U_t^d equals the original Metropolis algorithm's first coordinate process $X_{n,1}^d$, sped up by a factor of d.
- Hence, the original Metropolis algorithm's first coordinate process $X_{n,1}^d$ must converge to π in $O(d)$ time.
- Hence, the Metropolis algorithm converges (coordinatewise at least) in time $O(d)$. Right?
- One technicality: we need weak convergence from any starting point X_0, not from stationarity $X_0 \sim \pi$ ... but that also holds if the powers of the target density f in the moment assumptions are increased slightly (from 8 and 4, to 12 and 6). Phew!
- Also, still requires unrealistic conditions on π ... but we're working on that. Then have: convergence in $O(d)$ iterations!

How to Use the Optimality Information?

Recall: We have guidance about optimising MCMC in terms of acceptance rate, target covariance matrix Σ_t, etc.

In particular:

1. Want acceptance rate around 0.234.

2. Optimal Gaussian RWM proposal is $N\left(x, (2.38)^2 \, d^{-1} \, \Sigma_t\right)$, where Σ_t is the covariance matrix of the target π.

Great, except ... we don't know what proposal will lead to a desired acceptance rate. And, we don't know how to compute Σ_t.

So, what to do?

Trial and error? (difficult, especially in high dimension)

Or ... let the computer decide, on the fly!