Optimal Metropolis Scaling: Previous Proofs and Possible Generalizations

L'émir Omar Chéhab
under the supervision of Prof. Jeff Rosenthal
at the University of Toronto

July 30, 2017

Abstract

In the world of multidimensional random walk Metropolis algorithms, the seminal paper of Roberts, Gelman and Gilks (1997) [Roberts et al.(1997)Roberts, Gelman, Gilks, et al.] proposed a scaling of the proposal distribution that maximizes the algorithm's efficiency, in the case of an iid target distribution. Subsequent generalizations have been made (cf. Rosenthal, Bédard, Roberts, Stuart, etc.), where some hypothesis on the target distribution has been somewhat relaxed, though keeping some simplifying property (product form, symmetry, etc.). The next step would seek a stronger generalization, for example to target distributions satisfying certain properties and where no product form is assumed. Discussion follows.

Contents

1 The Random Walk Metropolis Algorithm and Variants

1.1 The Random Walk Metropolis Algorithm

The Random Walk Metropolis Algorithm, or RWA, is motivated by a simple fact of life: sampling for a complicated distribution can be hard.
The idea of MCMC is this: in order to sample from the *target distribution*, we are going to 'approach' it with a sequence of approximations. Because it is easier to work with a sequence of random variables, we associate each approximating distribution to a random variable. We start somewhere in the state space, jump randomly in a certain perimeter defined by a *proposal distribution*, look at the new state we obtain, and decide to move to it with a probability describing how closer this new position is to a 'high-density zone' of the target distribution. After a while, this stochastic process, which happens to be a Markov Chain, tends towards positions of 'higher target distribution density', and therefore the distributions associated with every new step better emulate the target distribution.
In conclusion, we start somewhere and randomly explore the state space in directions of higher density of target distribution we wish to emulate.

MATHEMATICAL NOTATION

$X^n_{s,i}$ where i is the vector component, s is the time-index or step of process, and n is the vector dimension. We will sometimes write one subscript index, which will refer to either time or vector component, according to context.

The Markov process at hand, or 'sequence of approximations', is $X^n = (X^n_0, X^n_1, X^n_2, ..., X^n_\infty)$

On the Target distribution:
It is the distribution we want to sample from. We write it $\Pi(x)$, where x is a n-dimensional vector.

On the Proposal distribution:
At X_i, we jump with an iid Gaussian centered where we are and of variance σ^2. It is the easiest and less expensive choice. The expectancy of the Gaussian is well-defined, sigma is the only 'loose' parameter: for the scaling, we might choose it to be inversely proportional to n, the vector dimension of the process X^n: $\sigma^2 = \frac{l^2}{n-1}$.

On the variance parameter:
Let us interpret intuitively how the proposal variance affects the Markov process. σ is sort of the average jump distance at each step of the process: if it is too small, the state space takes too long to be explored, and the convergence is significantly too slow; if it is too big, it is more probably that the possible next step is of lower stationary density than the one we are at currently, such a move would be refused by the transition probability and stationarity would result. Finding an optimal value for the variance is key.

On the probability of jumping to a new state:
At state X^n_i, we obtain a candidate Y^n_i for X^n_{i+1} by jumping following the *proposal distribution*: $Y^n_i = X^n_i + J^n_i$ where $J^n_i \sim \mathcal{N}(0, \sigma^2 I_n))$, is the 'jump' following the RWM proposal distribution. Now that we have a candidate Y^n_i, we must choose whether to move to it or remain in the current state X^n_i by *weighting out choice in favor of higher target density*: this 'weight' is the acceptance function $\alpha(X^n, Y^n) = min(1, \frac{\Pi(Y^n_i)}{\Pi(X^n_i)})$. More specifically, $\alpha(x^n, Y^n))$ is the probability that we choose Y^n_i over X^n_i. It follows that the probability that we remain at X^n_i is $1 - \alpha(X^n, Y^n)$.
On a side note, we often only have access to a distribution *proportional* to the true target distribution. Thus, the quotient form $\frac{\Pi(Y^n_i)}{\Pi(X^n_i)})$ that compares the target density of states X^n_i and Y^n_i is relevant, as it is independent of that the coefficient of proportionality.

Efficiency criteria:
1/ a priori, to maximize the estimated squared jumping distance: $\max_l \mathbf{E}[((X^d_{n+1,i} - X^d_{n,i})\alpha(x^n, Y^n))^2]$
[Atchadé et al.(2011)Atchadé, Roberts, and Rosenthal]
2/ after scaling and some work, maximize the velocity of the asymptotic continuous process, a Langevin diffusion.

1.2 The Metropolis-Adjusted Langevin Algorithm

Abbreviated as MALA, this novelty in this algorithm is that it adds *intentionality* in going from a step of the Markov process to the next :

1. An 'intentional' or 'damped' proposed move.

   The proposed 'next move' is generated not by a symmetric Gaussian, a good working choice for sure. Rather, knowing that asymptotically the probability distribution of a state behaves like the invariant distribution $\Pi(x)$, we damp in the 'direction of high probability concentration of Π'. Simply put, the proposed 'next move' is generated according to an SDE dynamic relying on the direction of $\nabla\Pi$ and such that the stationary law of that dynamic matches the invariant law of the Markov process.

2. An acceptance function evaluating the proposed move.

   Here we use the traditional Metropolis acceptance function, in its entire - not symmetrical as in the previous pages - form. Indeed, the transition distribution used will be an exponential decreasing with the L^2 distance between the current and proposed states: it is a harder journey when the distance is long.

The result for optimality in the case of an iid target distribution is to fine tune the AOAR to 0.524 (vs. the 0.234 for the RMW) [Roberts and Rosenthal(1998)].

1.3 Hamiltonian Monte-Carlo

The statistics-physics analogy is this:
the physical *position* is the *state space variables*
the *potential energy* is minus log of the probability density of those variables
position space. The minus log expression is reminiscent of information theory and mathematical entropy...
the *momentum* variables are artificially introduced with respect to the position variables

For more information, we refer the reader to [Neal et al.(2011)].

2 The iid target distribution model

[Roberts et al.(1997)Roberts, Gelman, Gilks, et al.]

2.1 Preliminary notions

Generators A stochastic process (Z_t) is characterized by an infinitesimal *generator*.
For a continuous-time process, the generator is the differential space-operator in the Feynman-Kac equivalent PDE.
For a discrete-time process, the generator writes as $< G, V > ((Z_t)) = lim_{t\to 0} \frac{\mathbb{E}[V(Z_t)] - V(z_0)}{t} = lim_{t\to 0} \frac{\mathbb{E}[V(Z_t) - V(z_0)]}{t}$ where $Z_0 = z_0$, V is functional. The expectancy kills the stochasticity and brings the computation to $\mathbb{R}$, an easily totally ordered set. If the process is time-homogenous Markov, which loosely states that the 'starting point plays no big role', we can write the generator as: $lim_{t\to 0} \frac{\mathbb{E}[V(Z_s) - V(Z_t)|Z_t = z]}{t}$. Like its continuous-time counterpart, it is clear that **the generator somewhat emulates the logic of differentiating** at time 0 (or t for a time-homogenous process), **but for a stochastic process**.

Skorokhod topology Many central functions encountered in probability, for example the cumulative distribution function, are càdlàg: continuous on the right side, with limits on the left side. It therefore makes sense to group them in a space, which we call the Skorokhod space. The most commonly used metric on that space assigns to it a topology, which we call the Skorokhod topology. Among its properties, are a certain time-elasticity and space-elasticity: **a function that is epsilon-perturbed in time (input space) or in space (output-space) remains the same through the lens of the Skorokhod topology.**
In that sense, it generalizes the uniform convergence topology, which bears space-elasticity but not time-elasticity.

Convergence of stochastic processes Much of the theory is to be found in Ethier and Kurtz's *Characterization and Convergence*, 1986. We here cite the lemma, theorems, and corollaries that will support our proof:

Chapter 4, Theorem 8.2
In substance, it states that a sufficient condition for the finite-dimensional *distributions* of a sequence of time-continuous processes to convergence weakly *in the Skorokhod topology* to those of a Markov process, is L1 convergence of their generators.

Chapter 3, Theorem 7.8
To go beyond convergence of *distributions* to convergence of the time-continuous stochastic processes *themselves*, a sufficient condition is relative compactness of the sequence of stochastic processes.

Chapter 4, Corollary 8.7
The relative compactness can be verified by making the generators uniformly converge on a set or limiting probability 1.

Lemma
Since, in our proof, the function g=log'f is lipschitz, a core function V for the generator G will be $C^{\infty}_{compact}$.

Precision
Most of the Ethier and Kurtz results stated above require that the functional space of the test function V be complete and separable. That is the case of $C^{\infty}_{compact}$.

2.2 Intuition behind the proof

[Bédard and Rosenthal(2008)]

To 'force' the convergence of the Markov chain toward a continuous stochastic process, we rescale in time and space.

Time-wise, we accelerate the Markov process by the dimension number n. As such, the time between two consecutive jumps becomes $\frac{1}{n}$, which goes to zero as n grows to ∞.

Space-wise, we temper the proposal standard deviation, the typical 'jump' if you will, by a $O(1/n)$ magnitude. More precisely, we choose $\sigma^2 = \frac{l^2}{n-1}$. We want l to be a free parameter than we can optimize for convergence speed. One may think: the jumps decrease, what if we have premature convergence? This is compensated for by the time-wise rescaling: mobility = number of jumps x length of jumps $\sim n\frac{l^2}{n-1} \sim l^2$.

Intuitively, as the dimension increases, the proposed moves become smaller and closer in time, giving an asymptotically continuous process. This limiting process is actually driven by a Langevin SDE: asymptotically optimizing convergence speed is now equivalent to maximizing the easily recognizable coefficient of diffusion-speed.

In a sense, this is a more complicated case of the simple symmetric random walk converging to a Brownian motion.

2.3 Walkthrough RGG's proof of weak convergence with generators

---

PROCESSES

$X^n = (X_t^n)_{t\in\mathbb{Z}}$ is **the original Markov process**. It is reversible with respect to π_n, and is π_n-irreducible, aperiodic and hence ergodic [Roberts and Smith(1994)] or [Mengersen et al.(1996)Mengersen, Tweedie, et a

$Z^n = (Z_t^n)_t$ is **the accelerated Markov process**. *How do we choose the acceleration-type?* The first guideline is this: Ethier and Kurtz's results on convergence are stated for continuous-time processes: *we must therefore transform our original discrete Markov process into a continuous one.*

The most obvious acceleration-type is:

- a *deterministic* n-linear acceleration: $(Z_t^n)_t = (X_{[nt]}^n)_t$
  It is the easiest starting point. We *know* that this accelerated process will jump every other $\frac{1}{n}$ and stay put between them. Let us note that throughout this document we use expectancies, and the jumps being at discrete times, they are of measure 0 and do not interfere in a an expectancy. What is more, this accelerated version *does not preserve the time-homogeneous property* of the discrete-time version *(cf. Appendix D) 11* which, if not absolutely necessary to maintain Ethier and Kurtz's results, remains a desirable property.

How do we preserve the time-homogeneous Markov property of the discrete-time and transpose it into a continuous-time version? This property denotes a 'memoryless' stochastic process. In order to affect time with this 'memorylessness', what better option than to link time transitions with the 'memoryless' property of the exponential distribution! The time index with thus follow a Poisson process. Which brings us to our second acceleration-type...

- a *stochastic* acceleration
  ...allowing the process to fall an epsilon away from the deterministic jumping points. The time between two consecutive steps is distributed according to an exponential with mean the time-scaling term $\frac{1}{n}$.

The distinction is a *theoretical* one: while the jumping times of the deterministic process are every other $\frac{1}{n}$, for the stochastic process the jumping times can fall an epsilon before or afterhand. The time-elasticity of the Skorokhod topology views them as equivalent: indeed, in *practical* terms, both have the same generator. *(cf. Appendix C) 10.*

Onwards, we write the *accelerated process* $(Z_t^n)_t$, knowing it is *stochastically accelerated* but can be conceptually understood as *deterministically accelerated*, thanks to the time-elasticity of the Skorokhod topology and the sharing of a same generator.

$U^n = (Z_{t,1}^n)_t$ is **the 1st component of the accelerated process** Z^n. Thought it is embedded in Z^n which is Markov, U^n itself is not Markov as it is *one-dimensional* and the acceptance depends on *all* dimensions. We anticipate the results of our proof, yet let us state here that U^n will convergence to a one-dimensional Langevin, independent from the other components, and satisfying the Markov property as solution of a regular SDE. In short, U^n *is Markov-embedded, asymptotically Markov, but not Markov itself.*

GENERATORS

$< G_n, V > (x^n) = n\mathbf{E}_Y[(V(Y^n) - V(x^n))\alpha(x^n, Y^n)]$ where V acts on $\mathbb{R}^{\ltimes}$, is the **generator of the accelerated process** Z^n *(cf. Appendix B) 9.*
Notation reminder: Y in subscript means that the expectancy acts on Y and is conditional to $X^n(s) = x^n(s)$, the lowercase indicating a constant.

$< G_n, V > (x^n) = n\mathbf{E}_Y[(V(Y^n) - V(x^n))\alpha(x^n, Y^n)]$ where V is restricted to $\mathbb{R}$, is the **generator of the one-dimensional accelerated process** U^n.
To be rigorous, restricting the core function of the generator to one dimension shouldn't necessarily give us the generator of the process restricted to one dimension: in other words, restricting to one dimension and taking the generator don't commute. It so happens that this is justified in our case. We spare the details and content ourselves with saying that the *supposed* generator verifies a martingale condition and is therefore the *correct* generator [Bédard(2006)].

$< G, V > (x) = h(l)[\frac{1}{2}V''(x) + \frac{1}{2}\frac{d}{dx}(logf)(x)V'(x)]$ is the anticipated **limiting generator**. It acts on functions of $\mathbb{R}$.

TARGET DISTRIBUTION

In the iid case, the target distribution is written $\Pi(x^n) = f(x_1^n)...f(x_n^n)$. That is the whole point: to break up the complexity of the problem by studying (decorrelated) one-dimensional components instead of an n-dimensional vector.

2.3.1 Starting point and Strategy

Looking at the generator, we have n times the expectancy of a product of terms (the integrand), each clouding with functions the known behavior of $Y_t^n - x_t^n \sim \mathcal{N}(0, \sigma^2 I_n)$), the 'gaussian jump'. We would like to 'extract' that known behavior from the functions: that can be done by using a Taylor formula on each of the product terms: we will be left with a polynomial in $(Y_t^n - x_t^n)$ plus a remainder. The expectancy kills the randomness, and replaces the polynomial in and spits out a polynomial in $(Y_t^n - x_t^n)$ with a polynomial in $\sigma = O(\frac{1}{\sqrt{n}})$, plus the expected remainder. Multiplied with n, the lower-order terms of the $\frac{1}{\sqrt{n}}$ polynomial are saved whilst the higher-order ones including the remainder disappear, as $n \to \infty$. We will be left with the generator of the limiting process.

Is there any smart economy of effort that can be done to spare from multivariate Taylor formulas and convergence theorems? The limiting process is distributed according to an iid target distribution, so we can sensibly expect X^n to asymptotically act iid. As such, we choose to anticipate that behavior by considering the generator not of (U_t), but of its first component $(U_{t,1})$. We are entitled to do so *is* insofar as $(U_{t,1})$ is indeed stochastic process. It may not be independent from the other components $(U_{t,i})$, it may not be Markov, we do not *know* that its limiting stochastic process will be independent from that of the other components $(U_{t,i})$ (that is our hope, not a certainty, at this point). Our new generator writes as:

$$< G_n, V > (x^n) = n\mathbf{E}_{Y_1}[(V(Y^n) - V(x^n))\mathbf{E}_{Y_{2,\dots,n}}\alpha(x^n, Y^n)] = n\mathbf{E}_{Y_1}[(V(Y_1^n) - V(x_1^n))\mathbf{E}_{Y_{2,\dots,n}}\alpha(x^n, Y^n)]$$

2.3.2 Taylor for the second product term

The generator is the product of two inner terms. We take one of them, E, simplify it by looking at its asymptotic behavior E_{lim}, and compute a new, asymptotic generator G_1, hopefully malleable enough to optimize the rescaled MCMC algorithm's efficiency.

We smartly rewrite the second product term

The second product term contains itself a product which we classically write in exp o ln form to manipulate a sum, in which we isolate the term in y_1 as it relates to the outer expectancy and is, in the eyes of our inner expectancy, deterministic.

$$E = \mathbf{E}_{Y_{2,\dots,n}} \alpha(x^n, Y^n)] = \mathbf{E}\left[1 \wedge exp\left\{log(\tfrac{f(Y_1)}{f(x_1)}) + \textstyle\sum_{i=2}^{n}(log(Y_i) - log(x_i))\right\}\right]$$

We can conveniently work the Taylor-Lagrange formula on $\mathbb{R}^1$ (this results from the product form of the target distribution), for each term of the long sum.

$$E = \mathbf{E}\left[1 \wedge exp\left\{log(\tfrac{f(Y_1)}{f(x_1)}) + \textstyle\sum_{i=2}^{n}\left[(logf(x_i))'(Y_i - x_i) + \frac{1}{2}(logf(x_i))''(Y_i - x_i)^2 + \frac{1}{6}(logf(Z_i))'''(Y_i - x_i)^3\right]\right\}\right]$$

To simplify notation, as we are dealing with the function logf, let us write it g. As such,

$g(x_i) = logf(x_i)$
$g'(x_i) = (logf(x_i))' = \frac{f'(x_i)}{f(x_i)}$ measures discontinuities of f
$g''(x_i) = (logf(x_i))'' = \frac{f''(x_i)}{f(x_i)} - g'(x_i)^2$

We look at our rewritten term and notice that it most likely simplifies asymptotically, making it more easily computable. We conjecture the simplified limit form

Can our E be simplified? Given the three sums, each of iid random variables, there is something in hoping that asymptotically there occurs some simplification, by the likes of a law of large numbers.

Let us look at each term more closely:

- 1^{st} order term
  It follows a $\mathcal{N}(0, \sum_{i=2}^{n} g'(x_i)^2)$ and we are comfortable working with a Gaussian.
- 2^{nd} order term
  Each term follows a $\chi^2(1)$ law multiplied by a constant, which would add complexity to our computation by hand. Does the sum simplify asymptotically? In spirit of a law of large numbers, we might expect:
  $\sum_{i=2}^{n} \frac{1}{2} g''(x_i)(Y_i - x_i)^2 \sim \sum_{i=2}^{n} \frac{1}{2} g''(x_i)\sigma^2$ as $\mathbb{E}[(Y_i - x_i)^2] = \sigma^2$.
  Also $\sim \sum_{i=2}^{n} -\frac{1}{2} g'(x_i)^2\sigma^2$ as $g''(x_i) = (logf(x_i))'' = \frac{f''(x_i)}{f(x_i)} - g'(x_i)^2$
  and it is reasonable to hope that $\frac{f''(x_i)}{f(x_i)}$ is of higher order than $g'(x_i)^2$, and therefore asymptotically less consequential.
  The asymptotically constant 2^{nd} order term would add itself to the mean of the Gaussian.
- 3^{rd} order term
  In the same spirit, heuristically:
  $(Y_i - x_i)^3 = O(1/n^{\frac{3}{2}})$, so $\sum_{i=2}^{n} \frac{1}{6} g'''(Z_i)(Y_i - x_i)^3 = O(1/n^{\frac{1}{2}})$.
  We therefore hope to be able to asymptotically neglect the 3^{rd} order term.

We therefore conjecture this asymptotic simplification:

$$\begin{cases} E \to E_{lim} = \mathbb{E}[1 \wedge \gamma] \\ \gamma = exp(log(\frac{f(Y_1)}{f(x_1)}) + A) \\ A \sim \mathcal{N}(\mu_n, \Sigma_n^2), \mu_n = \sum_{i=2}^{n} -\frac{\sigma^2}{2} g'(x_i)^2, \Sigma_n^2 = \sum_{i=2}^{n} g'(x_i)^2 \end{cases}$$

Discussion We could have kept g'' instead of $-g'$ and have found a different, more obvious, set a limiting probability one, perhaps the set of points verifying $|\sum_{i=2}^{n} g''(x_i)(\frac{(Y_i - x_i)^2}{2} - \frac{\sigma^2}{2})| \leq f(n)$ where f decreases as n increases. In that case, we might have made the following conjecture.
As it happens, a property often found optimizing MCMC algorithms is: $\mathbb{E}[\mu_n] = -\frac{\sigma_n^2}{2}\mathbb{E}[\Sigma_n^2]$, which says that in an infinite sum, g" behaves like $-g'^2$. The original RGG proof anticipates this and that is why, following its format, we use the above conjecture.

$$\begin{cases} E \to E_{lim} = \mathbb{E}[1 \wedge \gamma] \\ \gamma = exp(log(\frac{f(Y_1)}{f(x_1)}) + A) \\ A \sim \mathcal{N}(\mu_n, \Sigma_n^2), \mu_n = \sum_{i=2}^{n} \frac{\sigma^2}{2} g''(x_i), \Sigma_n^2 = \sum_{i=2}^{n} g'(x_i)^2 \end{cases}$$

Let us put our conjecture to the test!

We prove our conjecture of the simplified limit form of the inner term

Let us show that E converges to E_{lim}, where:

$$E = \mathbf{E}\left[1 \wedge exp\left\{log(\tfrac{f(Y_1)}{f(x_1)}) + \sum_{i=2}^{n}\left[g'(x_i)(Y_i - x_i) + \tfrac{1}{2}g''(x_i)(Y_i - x_i)^2 + \tfrac{1}{6}g'''(Z_i)(Y_i - x_i)^3\right]\right\}\right]$$

$$E_{lim} = \mathbf{E}\left[1 \wedge exp\left\{log(\tfrac{f(Y_1)}{f(x_1)}) + \sum_{i=2}^{n}\left[g'(x_i)(Y_i - x_i) - \tfrac{l^2}{2(n-1)}g'(x_i)^2\right]\right\}\right]$$

Consider $\left|E - E_{lim}\right|$. The arguments of the exp function are so similar, it would be nice to compare them instead of $\left|E - E_{lim}\right|$. In other words, it would be nice that difference of argument in the exp function play a role in dominating $\left|E - E_{lim}\right|$: that is a Lipschitz property, and it bodes well that the $x \to 1 \wedge exp(x)$ function is 1-Lipschitz! Thus,

$$\left|E - E_{lim}\right| \leq \mathbf{E}\left[\left|\sum_{i=2}^{n}\tfrac{1}{2}g''(x_i)(Y_i - x_i)^2 - \tfrac{l^2}{2(n-1)}g'(x_i)^2\right|\right] + \mathbf{E}\left[\sum_{i=2}^{n}\left|\tfrac{1}{6}(logf(Z_i))'''(Y_i - x_i)^3\right|\right]$$

- The second term

  The second term, using the Gaussian distribution of $Y_i - x_i$, the linearity of expectancy, and the iid hypothesis, is bounded by: $\sup_z |logf(z)'''|\frac{1}{6(n-1)^{\frac{1}{2}}}\frac{4l^3}{(2\pi)^{\frac{1}{2}}} = 0(\frac{1}{n^{\frac{1}{2}}})$ as expected.

  Point of observation: this justifies which sigma was chosen to be in $\frac{1}{n-1}$.

- The first term

  As for the first term, let us rewrite it with easier notation to better strategize: $\mathbf{E}\big[|W_n|\big]$. We are hoping that it converges to zero as n gets bigger. Now, we know how to deal with $\mathbf{E}\big[W_n\big]$: through linearity of expectancy, we would have had to deal with a χ^2 and constants. But with $|W_n|$, getting the expectancy is more complicated. So it is best we bound $\mathbf{E}\big[W_n\big]$ by some other calculable term which can be shown to converge to 0. A classical proceeding is squaring the term and using Jensen's inequality: we rid ourselves of the absolute value and are able to compute!

  $\mathbf{E}\big[|W_n|\big]^2 \leq \mathbf{E}\big[W_n^2\big]$ where W_n^2 is an expression of well-known laws. The independence of component Gaussian-jumps simplifies the computation as well. It is just a matter of writing it out *(cf. Appendix A) 8.*

  At this point, let us mention that in Appendix A, we obtain, in addition to the computation of Roberts, Gelman, and Gilks, a cross-term which doesn't seem to trivially disappear as the dimension grows. We perhaps missed one step in their reasoning...

  We obtain: $\frac{1}{4(n-1)^2}\left[\sum_{i=2}^{n} g''(x_i) + g'(x_i)^2\right]^2 + \frac{2}{4(n-1)^2}\sum_{i=2}^{n} g''(x_i)^2$

  Before moving to any formal proof, let us informally discuss what we have.
  For the second part: g" is bounded so its sup will be finite and the term decreases therefore in $O(1/n)$.
  The first part, hopefully, must then converge to 0, which would be equivalent to: $\frac{1}{n-1}\sum_{i=2}^{n} g''(x_i) \sim_{n\to\infty} \frac{-1}{n-1}\sum_{i=2}^{n} g'(x_i)^2$, for example if both sides convergence to a common limit. If such is the case, that limit I would most likely be $\mathbf{E}_f(g'^2)$, by the weak law of large numbers, given that asymptotically behave iid.

Another way to write this conjecture is:

$$\begin{cases} P(Z_n^t \in F_n) \to_{n\to\infty} 1 \\ F_n = \{|R_n(x_2,...,x_n) - I| < n^{-\frac{1}{8}}\} \cap \{|S_n(x_2,...,x_n) - I| < n^{-\frac{1}{8}}\} \\ R_n(x_2,...,x_n) = \frac{1}{n-1}\sum_{i=2}^n g'(x_i)^2, S_n(x_2,...,x_n) = -\frac{1}{n-1}\sum_{i=2}^n g''(x_i), I = \mathbb{E}[g'(T)^2] \end{cases}$$

Parenthetically, we write the condition with Z^n because it is the accelerated process we are truly considering. We have worked with X^n until then because, as shown in *(cf. Appendix B) 9*, the generator of Z^n can be expressed in terms of X^n only.

Why is this useful at all? Because the formalism of a set of limiting probability 1 intervenes in the 'relative compactness' part of the convergence theorems in the Skorokhod topology (cf. preliminary notions).

Let us verify our conjecture:

At stationarity, so for 'n sufficiently big':

$P(Z_s^t \notin F_n, \text{ for some s} \in [0,t]) = P(\cup_{s\in[0,t]}\{X_{[ns]}^t \notin F_n\})$
$\leq tnP(T \notin F_n)$ by sub-additivity
$\leq tn\Big(P[|R_n(x_2,...,x_n) - I| \geq n^{-\frac{1}{8}}] + P[|S_n(x_2,...,x_n) - I| \geq n^{-\frac{1}{8}}]\Big)$ by sub-additivity

Regarding the part in R_n

- the part with R_n
  $= tnP[(R_n(x_2,...,x_n) - I)^4 \geq n^{-\frac{1}{2}}]$ in order to have a non-negative random variable for Markov's inequality (we could have only squared but the powers of n have been chosen all along for final convergence)
  $\leq tn\mathbb{E}[(R_n(x_2,...,x_n) - I)^4]n^{\frac{1}{2}} = tn^{\frac{3}{2}}\mathbb{E}[(\frac{1}{n-1}\sum_{i=2}^n g'(x_i)^2 - I)^4]$ by Markov's inequality

  Having a mental idea of how this develops, and knowing that I is a number not a random variable, we know we will have a an algebra-type combination of moments of $g'(x_i)^2$. The Cauchy-Schwarz inequality for product terms will justify looking at them individually for boundedness. The highest order moment will be 4, and it would guarantee all the lower-order moments. Therefore, **we impose the condition**: $\mathbb{E}_f[g'(x_i)^8] < \infty$. The Central Limit Theorem tells us that $R_n(Z) - I \sim \frac{1}{\sqrt{n}}\mathcal{N}(0, Var_\Pi)$ and therefore $\mathbb{E}[(R_n(x_2,...,x_n) - I)^4] = O(\frac{1}{n^2})$.
  In conclusion, we have $tnP[|R_n(x_2,...,x_n) - I| \geq n^{-\frac{1}{8}}] = O(tn^{-1/2})$ and therefore $\to 1$ as $n \to \infty$.

- The part with S_n
  A similar reasoning with S_n giving rise to **imposed condition** $\mathbb{E}_f[\frac{f''}{f}(x_i)^4] < \infty$, gives: $tnP[|S_n(x_2,...,x_n) - I| \geq n^{-\frac{1}{8}}] \to 1$ as $n \to \infty$.

We have proved that $P(Z_s^t \in F_n, 0 \leq s \leq t) \to 1$ as $n \to \infty$ for fixed t. This concludes our proof for the overall result, that $\sup_{x^n \in F_n}\left|E - E_{lim}\right| = \phi(n) \to 0$ as $n \to \infty$.

We smartly rewrite the asymptotic inner term

Having established that E asymptotically behaves as E_{lim}, we are inclined to think that by plugging it into $< G_n, V > (x^n)$, we might obtain its limit for $n \to \infty$. But first: how does E_{lim} itself behave?

$$\begin{cases} E_{lim} = \mathbb{E}[1 \wedge exp(A)] \\ A \sim \mathcal{N}(\mu_n, \Sigma_n^2), \mu_n = log(\frac{f(Y_1)}{f(x_1)}) + \sum_{i=2}^n \frac{\sigma^2}{2} g''(x_i), \Sigma_n^2 = \sum_{i=2}^n g'(x_i)^2 \end{cases}$$

We might expect that the expectancy of an altogether not-too-complicated transformation of a Gaussian is computable. We here introduce this *useful lemma*:

$$\mathbb{E}[1 \wedge exp(A)] = \Phi(\tfrac{\mu}{\sigma}) + exp(\mu + \tfrac{\sigma^2}{2})\Phi(-\sigma - \tfrac{\mu}{\sigma})$$

where Φ is the cdf of a $\mathcal{N}(0,1)$ and $A \sim \mathcal{N}(\mu, \Sigma^2)$

We obtain:

$$E_{lim} = \Phi\left(R_n^{\frac{-1}{2}}\left(l^{-1}log(\tfrac{f(Y_1)}{f(x_1)}) - \tfrac{lR_n}{2}\right)\right) + exp\big(log(\tfrac{f(Y_1)}{f(x_1)})\big)\Phi\left(-\tfrac{-lR_n^{\frac{1}{2}}}{2} - log(\tfrac{f(Y_1)}{f(x_1)})R_n^{\frac{-1}{2}}l^{-1}\right)$$

that we can write as a function M of $\epsilon = log(\frac{f(Y_1)}{f(x_1)})$.

We show that looking at the generator asymptotically is nothing more than replacing the inner term by its asymptotic form

Now that we have a proper expression for E_{lim}, we are ready to look back at $< G_n, V > (x^n)$. Once again, it is reasonable to think that its limit for $n \to \infty$ is the same expression substituting E for E_{lim}. Let us verify that:

$< G_n, V > (x^n) = n\mathbf{E}_{Y_1}[(V(Y_1^n) - V(x_1^n))E]$
Our supposed limit: $< G_1, V > (x^n) = n\mathbf{E}_{Y_1}[(V(Y_1^n) - V(x_1^n))E_{lim}]$

$\sup_{x^n \in F_n} |(< G_n, V > - < G_1, V >)(x^n)| \leq \phi(n) n\mathbf{E}[|V(Y_1^n) - V(x_1^n)|] \to 0$

We have uniform convergence of G_n to G_1 on a set F_n of limiting probability 1. Therefore, by Ethier and Kurtz, we have weak convergence of X_n to the process described by the limiting generator, for the Skorokhod topology, on all finite sets of components.

2.3.3 Taylor for the first product term

Previously, our Taylor expansion of one inner term, E, has led us to a simplified asymptotic generator G1. By proceeding likewise with the other inner term, $V(Y_1^n) - V(x_1^n)$, we hope to obtain an even more malleable limiting generator G2.

We Taylor-expand the first inner term

The asymptotic generator we're working with:
$< G_1, V > (x^n) = n\mathbf{E}_{Y_1}[(V(Y_1^n) - V(x_1^n))M(log(\frac{f(Y_1)}{f(x_1)}))]$

We apply the Taylor formula about x_1 for the first product term (as well as the simplified second term), working now with G_1, the Skorokhod limit of G_n. To no big surprise, we obtain a polynomial (product of polynomials) in $(Y_1 - x_1)$ with determined coefficients:

$$(V(Y_1^n) - V(x_1^n))M(log(\tfrac{f(Y_1)}{f(x_1)})) = \left(V'(x_1)(Y_1 - x_1) + \tfrac{1}{2}V''(x_1)(Y_1 - x_1)^2 + \tfrac{1}{6}V'''(Z_1)(Y_1 - x_1)^3\right)\left[M(0) + (Y_1 - x_1)M'(0)(g(x_i))' + \tfrac{1}{2}(Y_1 - x_1)^2 T(x_1, W_1)\right]$$

where

$$\begin{cases} T(x_1, W_1) = M''\left(log(\frac{f(W_1)}{f(x_1)})\right)g'(W_1)^2 + g''(W_1)M'\left(log\frac{f(W_1)}{f(x_1)}\right) \\ Z_1, W_1 \in [min(x_1, Y_1), max(x_1, Y_1)] \\ M(0) = 2M'(0) = 2\phi(\frac{-l\sqrt{R_n}}{2}) \end{cases}$$

This is what we wanted: to extract $(Y_i - X_i)$ whose behavior we know!

We simplify the Taylor expansion, the n in front multiplies a polynomial heuristically in $\frac{1}{n}$ and thus decides which terms remain and which will go to zero as the dimension grows

Now, let us not get muddled by all this fluff, but rather think clearly and what is going to happen: we are to multiply by n and apply the expectancy. There is no 0-order term. The 1st order term in $(Y_i - x_i)$ is evaluated by the expectancy. Now for the rest: heuristically, $Y_i - x_i$ is of magnitude $\sigma = O(\frac{1}{\sqrt{n}})$ so for the higher order terms, $n(Y_i - x_i)^p$ is of magnitude $n^{1-p/2}$. For order 2, that gives a magnitude of 1, but starting order 3, we have magnitudes of n to negative powers. What is more, coefficients formed of all-order derivatives of M or V are bounded, as smooth functions on the compact domain of the test function. In conclusion, as $n \to \infty$, terms of order 3 and higher will disappear.

All that remains is to write this out and clearly separate the terms of order 3 and higher. That gives:

$$< G_1, V > (x) = 2n\phi(\tfrac{-l\sqrt{R_n}}{2})\Big[\big(\tfrac{1}{2}V''(x_1) + \tfrac{1}{2}g(x_i)\big)V'(x_1)\Big]\mathbf{E}[(Y_i - x_i)^2] + \mathbf{E}[B(Y_1, x_1, n)]$$

The part with B corresponds to the disappearing terms. The arguments of bounded coefficients on the compact domain writes as:

$$\mathbf{E}[B(Y_1, x_1, n)] \leq n\Big(a_1(K)\mathbf{E}[|Y_i - x_i|^3] + a_2(K)\mathbf{E}[|Y_i - x_i|^4] + a_3(K)\mathbf{E}[|Y_i - x_i|^5]\Big)$$

We can show that this last term is uniformly $O(n^{\frac{1}{2}})$ and therefore $\sup_{x^n \in F_n} |(< G_1, V > - < G_2, V >)(x^n)| \to 0$. In other words, we have uniform convergence of G_1 to G_2 on a set F_n of limiting probability 1. Therefore, by Ethier and Kurtz, we have weak convergence of X_n to the process described by the limiting generator, for the Skorokhod topology, on all finite sets of components.

$$< G_2, V > (x) = h(l)[\tfrac{1}{2}V''(x) + \tfrac{1}{2}\tfrac{d}{dx}(logf)(x)V'(x)], \text{ with } h(l) = 2l^2\phi(\tfrac{-l\sqrt{I}}{2}).$$

2.3.4 The Limiting Process

We have our final, hopefully simpler, asymptotic generator that we can work with. As a generator, it encodes the dynamics of a stochastic process. Let us examine that process - an asymptotic version of our rescaled Markov chain. If all goes well, we will reap the benefits of asymptotic simplification by finding an easy way to optimize process speed, and therefore algorithm efficiency.

Now, from the generator of the limiting process, let us go to... the limiting process itself! The limiting generator corresponds to a Langevin diffusion, driven by the SDE:

$$dW_s = \sqrt{h(l)}dB_s + \frac{h(l)}{2}g'(W_s)$$

A word: the functional V has, out of convenience, been chosen to take the 1st vector component as input. However, in all generality, it could have taken any finite component. Therefore, on every finite subset of $\mathbb{N}$, every component asymptotically follows a Langevin dynamic that is independent from the other: intuitively, the SDE shows that every component can be expressed as an integral form of itself; it is a function of itself and of no other component, and having an iid property for the asymptotic target distribution...

h(l) identifies as the 'speed measure' and its expression is $2l^2\phi(\frac{-l\sqrt{I}}{2})$, where ϕ is the normal standard cdf.

Asymptotically optimizing convergence speed means optimizing h(l) with respect to l, giving $\hat{l} \approx \frac{2.38}{\sqrt{I}}$.

This makes sense! I measures in fact the roughness of the target distribution density: $I = \mathbf{E}[(logf)'^2]$. And l measures mobility. So we have obtained that the optimal mobility l for convergence is *smaller* if the target density is rougher. Indeed, with roughness, one must proceed with small steps of caution to avoid brutal variation.

3 Practical consequences: Asymptotically Optimal Acceptance Rate

Now that we have a theoretical result, how do we implement it to ensure optimal convergence for the RWM algorithm? The answer lies with the Asymptotically Optimal Accceptance Rate, or AOAR.

By definition, the average acceptance rate is:

$$a_d(l) = \mathbf{E}[\alpha(X_t^n, Y_t^n)] = \mathbf{E}[\alpha(X_t^n, X_t^n + Z_{t+1}^n)] \text{ where } Z \sim \mathcal{N}(0, I_n)$$

$$= \iint \alpha(x^n, x^n + z^n)\, \Pi(x^n) dx^n \, pdf_{N(0,\sigma^2)} dz^n$$

A specific case of the Generator-based proof gives the *asymptotic* average acceptance rate: a(l) $= 2\phi(\frac{-l\sqrt{I}}{2})$.
Therefore, the *optimal asymptotic* average acceptance rate, or AOAR, is $a(\hat{l}) \approx 2\phi(-1.19) \approx 0.234$

Practically, this means tuning the proposal variance so that the algorithm accepts about 23% of proposed moves.

4 Further generalizations to date

For more detail, we refer the reader to *Optimal scaling for various Metropolis-Hastings algorithms* by Roberts, Rosenthal.

4.1 The independent target distribution model

In this case, independent components assure the product form of the target distribution, which was the structural backbone of the iid-case proof. However, their non-identical distribution explains the individual scaling terms in the target density here shown:

$\Pi(d, x^d) = \prod_{i=1}^{d} \frac{1}{\sqrt{\theta_i^2(d)}} f(\frac{x_i}{\sqrt{\theta_i^2(d)}})$

We shall not here expose the detailed results, for which we refer the reader to *Optimal scaling of Metropolis algorithms: Heading toward general target distributions*, by Bédard and Rosenthal. However, a word. The proceeding, sufficient regularity conditions, and main result of an AOAR=0.234, are similar to the iid case. The proposal variance scaling, instead of being in $O(\frac{1}{\sqrt{(d)}})$, requires the dimension power to be more tailored to the individual target components, as their scaling terms are not the same!

4.2 The infinite dimension Markov chain model

In this case, instead of considering a finite-dimension Markov Chain which asymptotically goes to the infinite dimension, we suppose the state space of the process to be of infinite dimension from the very beginning. The previous proofs relied on component-wise study, how do we make use of that? By giving the infinite-dimensional space a Hilbert structure, so that any vector can be projected via components unto a Hilbert base (for example, the eigenvector-base of a self-adjoint operator). For results and proof, we refer the reader to *Diffusion Limits Of The Random Walk Metropolis Algorithm In High Dimensions* by Mattingly, Pillai, Stuart.

5 Present Leads on generalization

There are many ways to go about optimizing the RWM algorithm and attempt to extract necessary hypotheses for our proof.

First, let us try to emulate RGG's proof with no particular condition on the target distribution and see what happens.

We start with the same generator: $< G_n, V > (x^n) = n\mathbf{E}[(V(Y^n) - V(x^n))\alpha(x^n, Y^n)]$

At this point, there are many options at hand.
The RGG-like option would be to look at the first component of the accelerated process: $U_{t,1}$. Because of possible correlation, we wouldn't be to 'insert' an expectancy like so:
$< G_n, V > (x^n) = n\mathbf{E}_{Y_1}[(V(Y^n) - V(x^n))\mathbf{E}_{Y_{2,\dots,n}}\alpha(x^n, Y^n)]$

but we can emulate this formula, making use of V's dependence on x_1 only, by using conditional expectancy:
$$< G_n, V > (x^n) = n\mathbf{E}\Big[\mathbb{E}\big[(V(Y^n)-V(x^n))\alpha(x^n, Y^n)|y_1\big]\Big] = n\mathbf{E}\Big[(V(Y^n)-V(x^n))\mathbb{E}\big[\alpha(x^n, Y^n)|y_1\big]\Big]$$

The Taylor development of the core function V is classic: there is not much to say on its behalf.

The term that 'changes' from RGG's proof is the one we will have to concentrate on: $\mathbb{E}[\alpha(x^n, Y^n)|y_1]$. It is a *conditional* inner expectancy, therefore a random variable, not a scalar. However, to get to that random variable, we *start* with:
$w(a) = \mathbb{E}[\alpha(x^n, Y^n)|Y_1 = a]$ which is a scalar, knowing that $w(Y_1) = \mathbb{E}[\alpha(x^n, Y^n)|Y_1]$

$w(a) = \mathbb{E}[1 \wedge exp(g(Y_n) - g(X_n))|Y_1 = a] = \mathbb{E}[1 \wedge exp(\nabla g_{x_n}.(Y_n - x_n) + (Y_n - x_n)^T \nabla^2 g_{x_n}(Y_n - x_n) + O(||Y_n - x_n||^3)|Y_1 = a]$ where $||\ .\ ||$ is a 'norme d'algèbre'.

In order to follow RGG's steps, we would need something computable like: $\mathbb{E}[1 \wedge exp(G)]$ where $G \sim \mathcal{N}(0, \sigma^2)$. Let us see if asymptotically, we can replace the second order term with a constant, as we did in RGG's case.

$\mathbb{E}[1 \wedge exp(\frac{\partial g}{\partial x_1}(x^n)(a - x_1) + \sum_{i=2}^n \frac{\partial g}{\partial x_i}(x^n)(Y_i - x_i) + \frac{1}{2}\frac{\partial^2 g}{\partial x_1^2}(x^n)(a - x_1)^2 + \sum_{i=2}^n \frac{\partial g}{\partial x_1 x_i}(x^n)(Y_i - x_i)(a - x_1) + \sum_{2 \le i \le j \le n} \frac{\partial g}{\partial x_i x_j}(x^n)(Y_i - x_i)(Y_j - x_j) + O(||Y_n - x_n||^3)]$

What do we know? The first-order terms $(Y_i - x_i)$ are Gaussian, the squared second-order terms $(Y_i - x_i)^2$ follow a χ^2, yet regarding the cross second-order terms $(Y_i - x_i)(Y_j - x_j), i \neq j$, we have a product of two iid Gaussians - and that is complicated. In this first draft, let us rid ourselves of the cross-terms by annulling the coefficients before them. We hope to find a more sophisticated hypothesis than $(\frac{\partial g}{\partial x_i x_j}(x^n) = 0, i \neq j)$, perhaps a sort of 'tail condition' where i and j have to be 'far away' and we take the sup with respect to x, but setting them to 0 will do for the moment. This should emulate a sort of 'independence' between the variables of the target distribution, mirroring the iid case.

We end up with:

$\mathbb{E}[1 \wedge exp(\frac{\partial g}{\partial x_1}(x^n)(a - x_1) + \sum_{i=2}^n \frac{\partial g}{\partial x_i}(x^n)(Y_i - x_i) + \frac{1}{2}\frac{\partial^2 g}{\partial x_1^2}(x^n)(a - x_1)^2 + \sum_{i=2}^n \frac{\partial g}{\partial x_1 x_i}(x^n)(Y_i - x_i)(a - x_1) + \sum_{2 \le i \le n} \frac{\partial^2 g}{\partial x_i^2}(x^n)(Y_i - x_i)^2 + O(||Y_n - x_n||^3)]$

Once again, why bother with the χ^2 when the sum of second-order terms may simplify asymptotically, by some law-of-large-numbers-like property, from $\sum_{2 \le i \le n} \frac{\partial^2 g}{\partial x_i^2}(x^n)(Y_i - x_i)^2$ to $\sum_{2 \le i \le n} \frac{\partial^2 g}{\partial x_i^2}(x^n)\frac{\sigma^2}{2}$. This requires a solid proof, but seems very feasible. So for the purposes of moving forward, let us say that asymptotically we get:

$\mathbb{E}[1 \wedge exp(\frac{\partial g}{\partial x_1}(x^n)(a-x_1) + \sum_{i=2}^n \frac{\partial g}{\partial x_i}(x^n)(Y_i - x_i) + \frac{1}{2}\frac{\partial^2 g}{\partial x_1^2}(x^n)(a-x_1)^2 + \sum_{i=2}^n \frac{\partial g}{\partial x_1 x_i}(x^n)(Y_i - x_i)(a-x_1) + \sum_{2 \leq i \leq n} \frac{\partial^2 g}{\partial x_i^2}(x^n)\frac{\sigma^2}{2} + O(||Y_n - x_n||^3)]$

We have finally obtained:

$$\begin{cases} \mathbb{E}[1 \wedge exp(X + O(||Y_n - x_n||^3)] \\ X \sim \mathcal{N}(\mu_n, \Sigma_n^2) \\ \mu_n(a) = \frac{\partial g}{\partial x_1}(x^n)(a-x_1) + \frac{1}{2}\frac{\partial^2 g}{\partial x_1^2}(x^n)(a-x_1)^2 + \sum_{2 \leq i \leq n} \frac{\partial^2 g}{\partial x_i^2}(x^n)\frac{\sigma^2}{2} \\ \Sigma_n^2 = \sum_{i=2}^n \frac{\partial g}{\partial x_i}(x^n)^2 \end{cases}$$

Supposing that the 3rd-order remainder phases out asymptotically, we recall the useful lemma allowing use to compute the expression, and obtain:

$\Phi(\frac{\mu_n(a)}{\Sigma_n}) + exp(\mu_n(a) + \frac{\Sigma_n^2}{2})\Phi(-\Sigma_n - \frac{\mu_n(a)}{\Sigma_n})$

To reach our *conditional expectancy*, we replace $Y_1 = a$ with Y_1: the parameter $\mu_n(a)$, a scalar, thus becomes the random variable $\mu_n(Y_1)$ which schematically $\sim \mathcal{N}(\sum_{2 \leq i \leq n} \frac{\partial^2 g}{\partial x_i^2}(x^n)\frac{\sigma^2}{2}, \frac{\partial g}{\partial x_1}(x^n)^2\sigma^2) + \frac{\sigma^2}{2}\frac{\partial^2 g}{\partial x_1^2}(x^n)\chi^2(1)$

It is possible to follow the RGG scheme of proof adapted to this situation further, but computation becomes more difficult.

6 Concepts to get acquainted with

Definition: Tail-Field
$\tau = \cap_{n\in\mathbb{N}}\sigma((X_k)_{k\in[n,\infty[}) = \cap_{n=0}^{\infty}\sigma(X_n, X_{n+1}, ...) = \cap_{n=0}^{\infty}\sigma\big(\cup_{k=n}^{\infty}\sigma(X_k)\big)$

Interpretation:

- sort of a limsup version for fields
- events of Ω for which the realization *depends* on the values of the X_i but is *independent* of any finite subset of these X_i

---

Definition: Trivial Tail-Field
Tail field that is P-trivial, that contains events that are almost sure or negligible

Interpretation: generalizes traditional notion of independent random variables
Understanding it through the 0-1 theorems:

- Kolmogorov's 0-1 Law
  The tail sigma-field of a sequence of independent random variables is trivial.
- Hewitt-Savage 0-1 Law
  The sigma-field of exchangeable events, a generalization of the tail-field defined by events invariant under permutation, of a sequence of iid random variables is trivial.

Links with other notions of independence:

- Strong-mixing - Trivial Tail-Field equivalence

  [Lindvall(2002)]

  definition of strong mixing: $\sup\big|P(A\cap B) - P(A)P(B)\big| \to 0$ as $s\to\infty$
  where $A\in\sigma(X_{-\infty}, ..., X_t)$ and $B\in\sigma(X_{t+s}, ..., X_\infty)$

  definition of mixing: $\lim_{n\to\infty}\sup_{A\in F_n}\big|P(A\cap B) - P(A)P(B)\big| = 0$

  interpretation of strong mixing: "for any two states of the system = realizations of the random variables, when given a sufficient amount of time between the two states, the occurrence of the states is independent"

  [Samorodnitsky(2016)], p44 and whereabouts

  "a stationary Gaussian process is mixing iff its correlation function asymptotically vanishes"

  of a stationary stochastic process, if the tail sigma-field is trivial, then the process is ergodic and mixing.

  "This statement makes it possible to view the triviality of the tail sigma-field as a kind of uniform mixing"

7 Acknowledgements

To Prof. Jeff Rosenthal, my earnest gratitude for your patient listening, your generosity with your time, and for taking me in.

To Prof. Mylène Bédard, an extended thanks for the clarity of your written explanations that helped my understanding.

8 AppendixA

$W_n = \left|\sum_{i=2}^{n} \frac{1}{2} g''(x_i)(Y_i - x_i)^2 - \frac{l^2}{2(n-1)} g'(x_i)^2\right|$.

To simplify computation, let us write this generic coefficients that we can specify at the end: $W_n = \left|\sum_{i=2}^{n} A_i(Y_i - x_i)^2 - B_i\right|$

$W_n^2 = \sum_{i=2}^{n}\left[A_i(Y_i - x_i)^2 - B_i\right]^2 + 2\sum_{1\leq i<j\leq n}\left[A_i(Y_i - x_i)^2 - B_i\right]\left[A_i(Y_j - x_j)^2 - B_i\right] = squaredterms + crossterms$

Now we apply the expectancy.

- $\mathbf{E}[crossterms]$

  $\mathbf{E}[crossterms] = 2\sum_{2\leq i<j\leq n}\mathbf{E}\left[A_i(Y_i - x_i)^2 - B_i\right]\mathbf{E}\left[A_i(Y_j - x_j)^2 - B_i\right]$ by linearity of expectancy and independence of Gaussian jumps.

  We make appear the standard score form $(Y_i - x_i)^2 = \sigma^2(\frac{Y_i - x_i}{\sigma})^2$ to have a $\chi^2(1)$ and then, once again, apply the expectancy's linearity. Finally,

  $\mathbf{E}[crossterms] = 2\sum_{2\leq i<j\leq n}\left[A_i\sigma^2 - B_i\right]\left[A_j\sigma^2 - B_j\right]$

  We now replace the coefficients A_i and B_i, as well as σ, with their specific values and obtain:

  $\mathbf{E}[crossterms] = 2\frac{l^4}{4(n-1)^2}\sum_{2\leq i<j\leq n}\left[g''(x_i) - g'(x_i)^2\right]\left[g''(x_j) - g'(x_j)^2\right]$

- $\mathbf{E}[squaredterms]$

  It's overall the same idea as with the crossterms, except that we lose the term- independence.

  $\mathbf{E}[squareterms] = \sum_{i=2}^{n}\mathbf{E}\left[\left[A_i(Y_i - x_i)^2 - B_i\right]^2\right]$

  We develop the expression and apply the expectancy's linearity:

  $\mathbf{E}[squareterms] = \sum_{i=2}^{n} A_i^2\mathbf{E}[(Y_i - X_i)^4] - 2A_iB_i\mathbf{E}[(Y_i - X_i)^2] + B_i^2$

  We make once again appear the standard score form $(Y_i - x_i)^2 = \sigma^2(\frac{Y_i - x_i}{\sigma})^2$ to have a $\chi^2(1)$ and use $Var(X) = E(X^2) - E(X)^2$. So that:

  $\mathbf{E}[squareterms] = \sum_{i=2}^{n} A_i^2\sigma^4\left[Var((\frac{Y_i - x_i}{\sigma})^2) + \mathbf{E}[(\frac{Y_i - x_i}{\sigma})^2]^2\right] - 2A_iB_i\sigma^2\mathbf{E}[(\frac{Y_i - x_i}{\sigma})^2] + B_i^2 = 3A_i^2\sigma^4 - 2A_iB_i\sigma^2 + B_i^2$

  We now replace the coefficients A_i and B_i, as well as σ, with their specific values and obtain:

  $\mathbf{E}[squareterms] = \frac{l^4}{4(n-1)^2}\sum_{i=2}^{n} 2g''(x_i)^2 + \left[g''(x_i) - g'(x_i)^2\right]^2$

Now, we would have liked an expression with $g''(x_i) + g'(x_i)^2$ (the sum would asymptotically cancel out given that $g''(x_i) = (logf(x_i))'' = \frac{f''(x_i)}{f(x_i)} - g'(x_i)^2$ and $\mathbb{E}_f[\frac{f''}{f}(x_i)^4] < \infty$) rather than $g''(x_i) - g'(x_i)^2$. So we write: $a_i - b_i = a_i + b_i - 2b_i$, where $a_i = g''(x_i)$ and $b_i = g'(x_i)^2$.

We finally obtain for $\mathbf{E}[W_n^2]$:

- RGG's term (where in the original paper, l is taken equal to 1)
  $\frac{l^4}{4(n-1)^2} \sum_{i=2}^{n} \left[(a_i + b_i)^2 + 2a_i^2\right] + 2\frac{l^4}{4(n-1)^2} \sum_{2 \leq i < j \leq n} (a_i + b_i)(a_j + b_j)$
  $= \frac{l^4}{4(n-1)^2} \left[\sum_{i=2}^{n} a_i + b_i\right]^2 + \frac{2l^4}{4(n-1)^2} \sum_{i=2}^{n} a_i^2$
- Another term from the square part
  $\frac{l^4}{4(n-1)^2} \sum_{i=2}^{n} -4a_i b_i$
- Another term from the cross part
  $2\frac{l^4}{4(n-1)^2} \sum_{2 \leq i < j \leq n} 4b_i b_j - 2b_i(a_j + b_j) - 2b_j(a_i + b_i)$ the terms in $b_i b_j$ cancel out by symmetry
  $= \frac{l^4}{4(n-1)^2} \sum_{2 \leq i < j \leq n} -4[a_i b_j + a_j b_i]$

As $n \to \infty$, do the terms of correction to RGG's computation disappear?

For the additional term from the square part, we have a a sum of n bounded terms divided by n^2, so that goes to 0.

For the additional term from the cross part, we have a sum of $O(n^2)$ bounded terms divided by n^2. It is not obvious at first sight how such a term may be asymptotically flushed out.

9 AppendixB

By definition, the generator of the rescaled process Z^n is:
$< G_n, V > (Z^n) = n\mathbf{E}\Big[V(Z^n(s+\frac{1}{n})) - V(Z^n(s))\big)|Z^n(s) = constant\Big]$
The multiplying n, comes from the division by $\frac{1}{n}$ the time between two jumps, so between two consecutive states.

By writing the accelerated process in terms of the original one,
$< G_n, V > (Z^n) = n\mathbf{E}\Big[V(X^n(ns+1)) - V(X^n(ns))\big)|X^n(ns) = constant\Big]$

Since the generator of a time-homogeneous Markov process doesn't depend on s but on the last state, we can write
$< G_n, V > (U^n) = n\mathbf{E}\Big[V(X^n(s+1)) - V(X^n(s))\big)|X^n(s) = constant \;\; x^n(s)\Big]$ We write the constant with a lowercase.

By the law of total probabilities, we write the conditional event in the expectancy as a sum of the possibility that we move a step and of the possibility that we stay put:

$$V(X^n(s+1)) - V(X^n(s))\big|[X^n(s) = x^n(s)] = \begin{cases} V(Y^n(s+1)) - V(x^n(s)) & \text{with probability} \alpha(x^n(s), Y^n(s+1)) \\ V(X^n(s)) - V(x^n(s)) = 0 & \text{with probability} 1 - \alpha(x^n(s), Y^n(s+1)) \end{cases}$$

Hence $< G_n, V > (x^n) = n\mathbf{E}[(V(Y^n(s+1)) - V(x^n(s)))\alpha(x^n(s), Y^n(s+1))]$

10 AppendixC

We show that the generators of the linearly accelerated version and continuous-time Poisson version are the same.

To obtain the generator of the Poisson version, we start with that of the linearly accelerated version and add a condition: whereas in the deterministic case the process jumped for sure and arrived at Y with probability the α function, for the Poisson process there is the preliminary step of the probability that the process jumps at all, which is proportional to the length k of the time-segment we consider.

Thus, $< G_n, V > (x^n) = \lim_{k \to 0} k\mathbf{E}_Y[(V(Y^n(s+k)) - V(X^n(s)))\alpha(X^n(s), Y^n(s+k))(nk+o(k))] = n\mathbf{E}_Y[(V(Y^n) - V(X^n))\alpha(X^n, Y^n)]$

We recognize the same generator as for the linearly-accelerated version.

11 AppendixD

DEFINITION: time-homogeneous stochastic process

For any t>0, $P(X_{t_0+t}|X_{t_0})$ is independent of t_0
In other words, the transition probability between two states only depends on the time *difference* $(t - t_0)$ these two states, and not on the *beginning* time t_0.

Suppose for argument's sake that n=1. The evolution of states over time can be visualized with the first floor represents X_0, the second is X_1, etc.

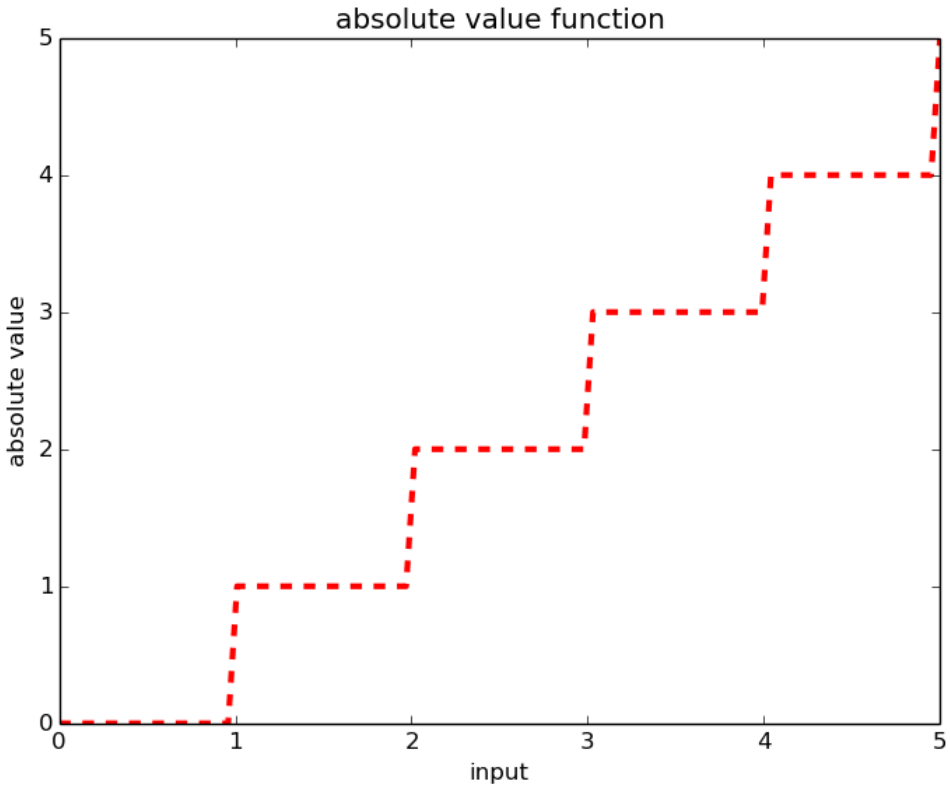

Figure 1: Floor function

The transition probability between times 0 and 0.8 is 1, because we're staying in the same state. However, between times 0.8 and 1.6, we're jumping to a new state Y_1 with a certain probability, given by the acceptance function $\alpha(X_0, Y_1)$ to be exact, which can very well be <1.

So even though the time difference between X_0 and $X_{0.8}$, and $X_{0.8}$ and $X_{1.6}$, is the same, the transition probability isn't. Therefore, *the linearly-accelerated process is not time-homogeneous.*

References

[Atchadé et al.(2011)Atchadé, Roberts, and Rosenthal] Yves F Atchadé, Gareth O Roberts, and Jeffrey S Rosenthal. Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Statistics and Computing*, 21(4):555–568, 2011.

[Bédard(2006)] Mylene Bédard. *On the robustness of optimal scaling for random walk Metropolis algorithms*, volume 68. 2006.

[Bédard and Rosenthal(2008)] Mylene Bédard and Jeffrey S Rosenthal. Optimal scaling of metropolis algorithms: Heading toward general target distributions. *Canadian Journal of Statistics*, 36(4):483–503, 2008.

[Lindvall(2002)] T. Lindvall. *Lectures on the Coupling Method*. Dover Books on Mathematics Series. Dover Publications, Incorporated, 2002. ISBN 9780486421452. URL https://books.google.ca/books?id=GUwyU1ypd1wC.

[Mengersen et al.(1996)Mengersen, Tweedie, et al.] Kerrie L Mengersen, Richard L Tweedie, et al. Rates of convergence of the hastings and metropolis algorithms. *The annals of Statistics*, 24 (1):101–121, 1996.

[Neal et al.(2011)] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

[Roberts and Rosenthal(1998)] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[Roberts and Smith(1994)] Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.

[Roberts et al.(1997)Roberts, Gelman, Gilks, et al.] Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

[Samorodnitsky(2016)] G. Samorodnitsky. *Stochastic Processes and Long Range Dependence*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 2016. ISBN 9783319455754. URL https://books.google.ca/books?id=39F5DQAAQBAJ.