Dynamic sampling

1 Introduction

Sampling is the central technical step in many Monte Carlo computations. A simple sampler, as we have seen, is a procedure that produces independent samples in time on the order of the complexity of the system. There are many probability distributions for which no simple sampler is known. For example, a sampler based on rejection for a system with n components easily could have an acceptance probability on the order of e^{-Cn}. While such a sampler in principle is "correct" (producing independent samples from the correct distribution), it is impractical for large n.

For example, suppose $X = (X_1 < X_2 < \cdots < X_n)$ is uniformly distributed except for the constraints that $X_{k+1} \geq X_k + r$, $X_0 \geq 0$, and $X_n \leq Rn$ (of course $R > r$). The X_k could represent the left endpoints of n rods of length r that are not allowed to overlap but otherwise are completely random. One sampling idea might be to choose n independent uniformly distributed points in $[0, nR]$, sort them, then reject if there are any overlaps. It is not hard to see that if we fix r and R while $n \to \infty$, it is exponentially unlikely to have a success.

Dynamic sampling is based on a different idea. Suppose X is an n component random variable with probability distribution $f(x)$. There may be a way to perturb X to get a different sample, X', that also is from the distribution f. For example, suppose X is an allowed rod configuration as above. We can choose a k, fix the X_j for $j \neq k$ (i.e. set $X'_j = X_j$ for $j \neq k$), and let X'_k be a uniform random variable in the interval $I_k = (X_{k-1} + r, X_{k+1} - r)$. We know that I_k is not empty because $X_k \in I_k$ since X was an allowed configuration. The reader should try to check that if $X \sim f$, then $X' \sim f$. We will verify this later through the principle of detailed balance.

Starting now and for the rest of this section, we use superscripts to distinguish samples and subscripts to label components of a sample. Thus, X_k is component k of configuration X, and X^m is the m^{th} sample configuration, $X^m = (X_1^m, \ldots, X_n^m)$. In the rod system, we could start with an initial configuration, X^0, resample the first component to get X^1 with $X_j^1 = X_j^0$ if $j > 1$ but $X_1^1 \neq X_0^0$. After resampling X_2, we would get X^2 with $X_j^2 = X_j^0$ for $j > 2$ but $X_1^2 \neq X_1^0$ and $X_2^2 \neq X_2^0$, etc. Eventually, we would get X^n with $X_k^n \neq X_k^0$ for all k. Since the distribution f is preserved at each stage, X^n has the correct distribution if X^0 does.

It is important to recognize that even though every component of X^n is different from the corresponding component of X^0, X^n is not independent of X^0. For example, suppose $X_{n-1}^n \leq X_n^0$. If $X_n^0 \ll nR$ (unlikely but possible), then $X_{n-1}^n \ll nR$, and $X_{n-2}^{2n} \ll nR$ (assuming we start over after n), etc. The memory of a dynamic Monte Carlo sampler can extend over not just one, but many sweeps. One of the main issues in dynamic Monte Carlo is the long persistence of correlations between samples, measuring it, understanding it, and reducing it.

Suppose we are trying to calculate $A = E_f[V(X)]$. The dynamic Monte Carlo procedure is to generate a sequence of L samples then apply the estimator

$$A \approx \widehat{A}_L = \frac{1}{L} \sum_{m=1}^{L} V(X^m) \,. \tag{1}$$

There are many issues to think about. We need conditions under which $\widehat{A}_L \to A$ as $L \to \infty$ even though the X^m are not independent. We must be able to do this even if X^0 does not have the f distribution, since the point of dynamic Monte Carlo is that sampling f is hard. Finally, we need error bars for (1) that take into account correlation between the samples.

A dynamic sampler is an iteration $X^{m+1} = \Phi(X^m, \xi^m)$, where $\Phi(x, \xi)$ is some deterministic function (or procedure) and the ξ_m are independent random variables that may be sampled through a simple sampler. The X^m form a Markov chain and dynamic sampling often is called *Markov chain Monte Carlo*, or *MCMC*. For example, in the rod example above, (with a change of notation), we get X^{m+1} from X^m be resampling each of the X_k once. Then ξ_k^m is the standard uniform random variable used to resample We say that Φ preserves f if $X' = \Phi(X, \xi)$ has $X' \sim f$ if $X \sim f$. The basic theory of Markov chains implies that if Φ preserves f and satisfies certain non-degeneracy conditions, then the distribution of X^m converges to f as $m \to \infty$ and $\widehat{A}_L \to A$ as $L \to \infty$. We outline this theory below.

Next, we give some strategies for designing such functions Φ given f. The most famous are the *Metropolis algorithm* and *heat bath* (also called the *Gibbs sampler*). The rod example above is an instance of heat bath. The main ideas are *partial resampling* and detailed balance. Sometimes we can find more sophisticated dynamic samplers that have less correlation among the samples. Two ways to achieve this are expanding the state space (the system with more components may have fewer constraints and be easier to sample), and *hybridization* (alternating different functions Φ that reduce correlations in different ways.

Dynamic Monte Carlo codes need a way to estimate error bars. These are determined by the *static variance* $\sigma^2 = \text{var}_f(V(X))$ and the *autocorrelation time*, τ. Simple samplers have $\tau = 1$ so only σ^2 matters. The autocorrelation time is a measure of how may applications of Φ (iterations) it takes to produce an effectively independent sample of f. More precisely, the error bar for simple sampling,

$$A = \widehat{A}_L \pm \frac{\widehat{\sigma}}{\sqrt{L}} \,,$$

is replaced by

$$A = \widehat{A}_L \pm \frac{\widehat{\sigma}}{\sqrt{L/\tau}} \,. \tag{2}$$

This says that the *effective sample size*, the effective number of independent samples, is $L_{\text{eff}} = L/\tau$. The penalty for dynamic Monte Carlo is that you have to increase the run length by a factor of τ longer to get the same statistical error as you would have from a simple sampler. On the other hand, if the best simple

sampler is exponentially inefficient, a dynamic sampler, even with a large τ, is preferable. We will discuss ways to estimate τ for the error bar (2).

There is much work by people in fields ranging from theoretical computer science to applied mathematics to theoretical physics and statistics estimating τ and related quantities. Much beautiful mathematics has been developed for this purpose. We discuss some of that at the end.

2 Theory of Markov chain Monte Carlo

The theory of dynamic samplers is part of the theory of Markov chains. The general theory of Markov chains is a little technical so we start with the simplest case, the Markov chain with a finite state space.

The state space will be called $\mathcal{S}$, and states will be denoted by Greek letters, $\alpha \in \mathcal{S}$, etc. If there are s states in all ($|\mathcal{S}| = s$), then the transition matrix is an $s \times s$ matrix, P, with entries $p_{\alpha\beta} = P(\alpha \to \beta)$. Consistent with notation in the introduction (but inconsistent with that of the rest of the world), write $X^m \in \mathcal{S}$ for the state after m steps. The probability distribution of X^m will be denoted $P(X^m = \alpha) = f(\alpha, m)$. Writing f^m for the n component row vector with components $f(\alpha, m)$, we have the forward equation $f^{m+1} = f^m P$.

We are looking for criteria under which a Markov chain forgets its initial state. One necessary condition is that it should be possible, eventually, to get from any state to any other. A transition $\alpha \to \beta$ is *possible* if $p_{\alpha\beta} \neq 0$. A possible path of length $k+1$ from α to β is a sequence of states $\gamma_j \in \mathcal{S}$, for $j = 0, \ldots, k$, with $\gamma_0 = \alpha$ and $\gamma_k = \beta$ so that each of the transitions $\gamma_j \to \gamma_{j+1}$ is possible. If $X^0 = \gamma_0$, then the probability that $X^k = \gamma_k$ for $k = 1, 2, \ldots, k$ is non zero if and only if γ is possible. A Markov chain is called *irreducible* or *indecomposable* if for every pair $\alpha \in \mathcal{S}$ and $\beta \in \mathcal{S}$, there is a possible path of some length (which may depend on α and β) from α to β.

Periodicity is another form of long term memory. State α has *period* r if every possible path from α to α has a length that is a multiple of r. If state $\alpha \in \mathcal{S}$ has a period greater than one, then some memory of the initial state persists for ever. If $X^0 = \alpha$ and m is not a multiple of r, then it is impossible to have $X^m = \alpha$. For example, suppose the state $X = (X_1, \cdots, X_n)$ is a sequence of *spins* that take possible values $X_k = +1$ or $X_k = -1$ (called *spin up* and *spin down* respectively) only. Then the number of states is $s = |\mathcal{S}| = 2^n$. Suppose one step in the Markov chain chooses a *site*, k at random (each site equally likely, all choices independent), and *flips* the spin ($X'_k = -X_k$). For this chain, every state has period 2 because the parity (even or odd) of the number of up spins changes each step. The Markov chain is *aperiodic* if every state has period $r = 1$.

A row vector is a probability vector if none of its components is negative and the components sum to one. The fundamental theorem of Markov chains states that if the chain is aperiodic and irreducible, then there is a unique probability row vector, f, so that

$$f \;=\; fP\,. \tag{3}$$

We say that f is the *invariant* or *steady state*[1] probability distribution for P. Moreover, if f^0 is any probability vector and $f^{m+1} = f^m P$ for all $m > 0$, then $f^m \to f$ as $m \to \infty$. For Monte Carlo applications we say it like this. We want to sample a distribution, f. We construct a Markov chain that has f as an invariant distribution. If that Markov chain is irreducible and aperiodic, then no matter how we choose X^0, the distribution of X^m will converge to f as $m \to \infty$.

This theorem sometimes is called the *Perron Frobenius* theorem because of a generalization due to Perron and Frobenius. They studied the eigenvalue problem for matrices, A, (like P) with all nonnegative entries. They proved that if A is aperiodic and irreducible and λ is the eigenvalue of A with maximum modulus ($|\lambda| = \max$), then λ is a positive real number and a simple eigenvalue (one dimensional eigenspace, no Jordan blocks). We give a proof here that works for Markov chain transition matrices only. The more general theorem (see, e.g. **Matrix Theory** by Gantmacher) is not more difficult, but is not stated in probabilistic terms.

We give a proof that relies on duality theory of the matrix eigenvalue problem. If λ_1 is an eigenvalue of P, then there is a right eigenvector, v, with $Pv = \lambda_1 v$. If $\lambda_2 \neq \lambda_1$ is another eigenvalue ant g is a left eigenvector ($gP = \lambda_2 g$), then $gv = 0$. In the present case, the column vector $v = \mathbf{1}$, whose entries are $v_\alpha = 1$ for all α, satisfies $Pv = v$ because $\sum_\beta p_{\alpha\beta} = 1$ for all α. Therefore, there is a left eigenvector, f, with $fP = f$. If the entries of f all have the same sign, we may assume (flipping the sign if necessary) that they all are non-negative and that $\sum_\alpha f_\alpha = 1$. This shows that any Markov chain with a finite state space has at least one invariant probability distribution.

To show that the invariant distribution is unique and stable, we need to use the hypotheses that P is aperiodic and irreducible. One of the several ways to do this is to examine a left eigenvector $gP = \lambda g$ with $\lambda \neq 1$ and show that $gP^m \to 0$ as $m \to \infty$. This would imply that there are no eigenvalues $\lambda \neq 1$ with $|\lambda| \geq 1$. For a finite state space, it implies further that there is a positive *spectral gap*[2], $\rho > 0$, with

$$\max_{\lambda \neq 1} |\lambda| = 1 - \rho \; .$$

The main idea is to find cancellation in the sum that computes[3] $g^m = gP^m$, which cancellation implies that

$$\sum_\alpha |g_\alpha^m| \; < \; \sum_\alpha |g_\alpha| \; .$$

If $gP = \lambda g$, this forces $|\lambda| < 1$, which is the conclusion we are looking for. The first step is to show that entries of g have different signs. That is simply because

[1] Physical scientists reserve the term *equilibrium* for steady states that satisfy detailed balance, see below.

[2] The set of eigenvalues of a matrix or operator are called the *spectrum* of that matrix or operator. There are eigenvalue problems in quantum mechanics that predict the colors materials will glow when heated (iron red, copper green, etc).

[3] Note the conflict of notation. On the left g^m is iterate m of the recurrence relation $g^{k+1} = g^k P$. On the right P^m is the matrix P to the power m.

$g\mathbf{1} = \sum_\alpha g_\alpha = 0$. Next, note that

$$|g^m_\alpha| \;=\; \left|\sum_\beta g_\beta p^m_{\beta\alpha}\right| \;\le\; \sum_\beta |g_\beta|\;\left|p^m_{\beta\alpha}\right| \quad . \tag{4}$$

Now P^m is a Markov chain transition matrix, so $\left|p^m_{\beta\alpha}\right| = p^m_{\beta\alpha}$ and $\sum_\alpha p^m_{\beta\alpha} = 1$ (for all β). Therefore

$$\|g^m\|_{L^1} \;=\; \sum_\alpha |g^m_\alpha| \;\le\; \sum_\alpha \sum_\beta |g_\beta|\; p^m_{\beta\alpha} \;=\; \sum_\beta |g_\beta| \;=\; \|g\|_{L^1} \quad . \tag{5}$$

The inequality (5) will be strict if any of the inequalities (4) is strict. If there is equality in (4) for each α, then for each α, all the terms in the sum $\sum_\beta g_\beta p^m_{\beta\alpha}$ must have the same sign, either all positive or all negative. By hypothesis, we can find β with $g_\beta > 0$ and γ with $g_\gamma < 0$. So, the conclusion follows from the

Lemma. If P is aperiodic and irreducible then for any α, β, and γ there is an $m > 0$ so that $p^m_{\beta\alpha} > 0$ and $p^m_{\gamma\alpha} > 0$.

Proof. The statement $p^m_{\beta\alpha} > 0$ is equivalent to the statement that there is a possible path of length m from β to α. We must show that there is an m so that there are possible paths both of length m from β and γ to α. Since P is irreducible, there is a possible path of some length from β to α and one from γ to α. Therefore, the lemma follows from the statement that for any α there is a k so that if $l \ge k$ there is a path of length l from α to α. This is equivalent to the statement that $p^l_{\alpha\alpha} > 0$ for all $l \ge k$.

This is an elementary fact about numbers. If there is a path of length l_1 and one of length l_2 (both from α to α), then there is a path of length $l_1 + l_2$ (do one path then the other). This completes the proof of the lemma.

This discussion is closed by noting that it shows that f with $fP = f$ (the dual eigenvector to $v = \mathbf{1}$ with $Pv = v$) has all entries of the same sign. Indeed, if there are entries of f with both signs, then there must be cancellation in (4).

Now, let $f^0(\alpha) = \Pr(X(0) = \alpha)$ be arbitrary and run the chain. If $f^m(\alpha) = \Pr(X(m) = \alpha)$, then $f^m = f^0 P^m$, as before. The large time behavior of f^m may be studied using the eigenvector decomposition of f^0, which takes the form

$$f^0 \;=\; a_1 f + \sum_{j=2}^{s} a_j g_j \;,$$

where the g_j are eigenvectors ($g_j P = \lambda_j g_j$) or generalized Jordan block vectors corresponding to eigenvalues with $|\lambda_j| \le 1 - \rho$. The coefficients a_j are found using the normalized dual eigenvectors, v_j with $Pv_j = \lambda_j v_j$ (or generalized eigenvectors), $a_j = f^0 v_j$. This implies that $a_1 = f^0\mathbf{1} = \sum_\alpha f^0_\alpha = 1$. Setting $a_1 = 1$ above gives

$$f^m \;=\; f + \sum_{j=2}^{s} a_j \lambda_j^m g_j \;=\; f \;+\; O\left((1-\rho)^m\right) \;.$$

This implies that the distribution of $X(m)$ converges exponentially to the invariant steady state distribution, f. This is the basis of dynamic, or Markov chain Monte Carlo.

The theory of Markov chains applies also to the case when the state space is infinite but discrete (countable). The definitions of acyclic and irreducible are the same. The basic theorem that is important for Monte Carlo still is true as stated: if the Markov chain preserves the probability distribution f, then the probability distribution of X^m converges to f as $t \to \infty$. There are two differences, however. There is no guarantee of a spectral gap and there are examples where the convergence $f^m \to f$ is not exponentially fast. Slow convergence for infinite or very large finite state space systems is a major problem for some dynamic Monte Carlo algorithms.

The other difference is less important for Monte Carlo: even if the chain is acyclic and irreducible, there may be no steady state probability distribution. A Markov chain that has a steady state probability distribution is called *positive recurrent*. Suppose we pick $\alpha_0 \in \mathcal{S}$ and let τ_k be the successive times, m, for which $X^m = \alpha_0$. These are called *recurrence times* or *renewal times*. It is a theorem that a Markov chain with a countable state space has a steady state probability distribution if and only if the expected return time is finite: $E\left[\tau_{k+1} \mid \mathcal{F}_{\tau_k}\right] < \infty$. This is the origin of the term positive recurrence. The other possibilities are that the expected value of, for example, τ_1 is infinite but τ_1 itself is finite. This is called *null recurrence* and ordinary symmetric random walk on the integers is an example. The other possibility is that there is a positive probability never to return to α_0. This is called *transience*, and biased random walk on the integers (say left with probability 2/3 and right with probability 1/3) is an example. The possibilities are very important in the theory of Markov chains, but less important for dynamic samplers, since we usually design a Markov chain to preserve a given distribution.

The theory of Markov chains on continuous state spaces is more subtle. The simplest case to describe is when the distribution of $y = X^{m+1}$ given $x = X^m$ is given by a conditional probability density, $p(x, y$. If $f^m(x)$ is the probability density for X^m,then $f^{m+1}(x) = \int f^m(y)p(y, x)dy$. In keeping with our notation for matrices, we write this as $f^{m+1} = f^m P$. Putting the vector f^m on the left of the integral operator P means that we integrate with respect to the first argument of P. The balance condition for a steady state probability density is $f = fP$ as before.

More generally, it may be that the conditional distribution of X^{m+1} given X^m is singular and given by a probability measure rather than a density. That is, for each x there is a probability measure, written $p(x, dy)$ (or something like that) so that if $f(dx)$ is the probability measure describing the distribution of X^m, then $f^{m+1}(dy) = \int f^m(dx)p(x, dy)$. There is some mathematical subtlety in this formula, as the integration is with respect to the probability measure $f^m(dx)$, and the integrand, $p(x, dy)$, is measure valued. For the integral to make sense, this measure valued function of x must be measurable with respect to the $\sigma-$algebra used for f^m.

As an example, consider the rod example with two rods of which only the

first will move. Call the old positions (x_1, x_2) and the new positions (y_1, y_2). The distribution of y_1 is uniform in the interval $[1, x_2 - r]$, so its density is $\frac{1}{x_2 - r}\mathbf{1}_{[0,x_2-r]}(y_1)$. Since $y_2 = x_2$, its "density" is $\delta(y_2 - x_2)$ (actually, the "$\delta-$measure" or "$\delta-$mass" at x_2). The overall transition measure is

$$p(x_1, x_2, y_1, y_2) \;=\; \frac{1}{x_2 - r}\mathbf{1}_{[0,x_2-r]}(y_1)\delta(y_2 - x_2)\;. \tag{6}$$

Integrating with respect to the second variable in the transition density or transition measure defines a different operator $(Pu)(x) = \int p(x, y)u(y)dy$ or $(Pu)(x) = \int p(x, dy)u(y)$. This is the operator that arises in the backward equation. If we define

$$u^m(x) \;=\; E\left[\, V\left(X^T\right) \mid X^m = x \,\right]\;,$$

then the functions u^m satisfy the recurrence relation $u^m = Pu^{m+1}$. This is the same relation that holds for matrices and vectors in the case of a discrete or finite state space.

3 Detailed balance, Metropolis

Start in the setting of finite state space, $\mathcal{S}$, and a given probability distribution f_α. We want a Markov chain matrix (*stochastic matrix*), P, that satisfies the balance condition (3). Of course, this P should be easy to implement. One way to satisfy the balance conditions is to require them in the stronger form of *detailed balance*: for each pair α and β,

$$f_\alpha p_{\alpha\beta} \;=\; f_\beta p_{\beta\alpha} \tag{7}$$

The left side of (7) is the probability, in the steady state, of observing an $\alpha \to \beta$ transition: first you must choose α (probability $= f_\alpha$), then you must choose to make a transition to β (probability $= p_{\alpha\beta}$). Detailed balance is the statement that in f, the probability of observing a given transition (e.g. $\alpha \to \beta$) is equal to the probability of observing the reverse (e.g. $\beta \to \alpha$).

It is easy to see that detailed balance implies balance, just sum over β in (7) and use the fact that, for each α, $\sum_\beta p_{\alpha\beta} = 1$. You get

$$f_\alpha \;=\; \sum_\beta f_\beta p_{\beta\alpha}\;,$$

which is the balance condition (3). The relation between balance and detailed balance might be clearer if we leave out the term $\beta = \alpha$ on both sides. On the left we have

$$f_\alpha \;=\; f_\alpha p_{\alpha\alpha} \;+\; \sum_{\beta\neq\alpha} f_\alpha p_{\alpha\beta}\;.$$

On the right we have

$$f_\alpha p_{\alpha\alpha} \;+\; \sum_{\beta\neq\alpha} f_\beta p_{\beta\alpha}\;.$$

The balance condition then becomes

$$\sum_{\beta \neq \alpha} f_\alpha p_{\alpha\beta} \;=\; \sum_{\beta \neq \alpha} f_\beta p_{\beta\alpha} \; .$$

The left side is the probability of observing an $\alpha \to$ (not α) transition, while the right is the probability of a (not α) $\to \alpha$ transition. Ordinary balance is the fact that for each α, the probability of an inbound transition is the same as that of an outbound transition. Detailed balance is the more restrictive statement that for every β, the probability of an inbound transition from β is the same as the probability of an outbound transition to β.

A simple system that has balance but not detailed balance has three states $1, 2, 3$ with transition probabilities $P(1 \to 2) = P(2 \to 3) = P(3 \to 1) = 3/4$ and $P(2 \to 1) = P(3 \to 2) = P(1 \to 3) = 1/4$, with no other transitions allowed. The invariant probability distribution is $f_1 = f_2 = f_3 = 1/3$. This satisfies $f = fP$, but the probability of observing $1 \to 2$ is $(1/3) \cdot (3/4) = 1/4$ while the probability of the reverse is $(1/3) \cdot (1/4) = 1/12$.

Physical scientists use the term *equilibrium* only for system that satisfy detailed balance. Other systems are merely in *steady state*, or perhaps *statistical steady state* to emphasize that the state changes from time to time but the statistical distribution of states in not changing. Detailed balance is a fundamental principle of equilibrium statistical physics. An equilibrium is a system that has nothing (energy, particles, ...) flowing through it. The distribution of air molecules in a room would be in equilibrium if the walls of the room would not conduct heat. The earth is not in equilibrium: It receives energy from the sun and radiates it into space – a non-equilibrium steady state. Non-equilibrium systems can be complex in ways that are not allowed for equilibrium systems. The energy flowing through the earth, being absorbed in one way and radiated in another, is necessary for the complexity of life.

The *Metropolis* algorithm uses rejection to enforce detailed balance. A trial move is proposed. If the move is rejected, then $X^{m+1} = X^m$. We start with *proposal* probabilities

$$T_{\alpha\beta} \;=\; \Pr\big(\text{ propose } X^{m+1} = \beta \;\mid X^m = \alpha\big)$$

In principle, the $T_{\alpha\beta}$ only need to satisfy $T_{\alpha\beta} \geq 0$ (for all α and β) and $\sum_\beta T_{\alpha\beta} = 1$ (for all α). In practice, it must be possible to program an efficient sampler of $T_{\alpha\beta}$ – pick a random β with probability $T_{\alpha\beta}$. The *acceptance probabilities*, $R_{\alpha\beta}$ can be any probabilities $0 \leq R_{\alpha\beta} \leq 1$. The algorithm is: if $X^m = \alpha$, first choose $\beta \in \mathcal{S}$ by sampling the probability distribution $T_{\alpha\beta}$. The accept β with probability $R_{\alpha\beta}$. If β is accepted, then $X^{m+1} = \beta$. Otherwise the proposed move is rejected and $X^{m+1} = \alpha$. The probability of $\alpha \to \beta$ is, for $\alpha \neq \beta$,

$$p_{\alpha\beta} \;=\; T_{\alpha\beta} R_{\alpha\beta} \; . \tag{8}$$

If the desired probability distribution, f_α, and proposal probabilities $T_{\alpha\beta}$ are given, it is possible to choose the acceptance probabilities $R_{\alpha\beta}$ to satisfy

detailed balance. For a given pair α and β, the detailed balance condition (7) becomes

$$f_\alpha T_{\alpha\beta} R_{\alpha\beta} \;=\; f_\beta T_{\beta\alpha} R_{\beta\alpha} \;,$$

which implies that

$$\frac{R_{\alpha\beta}}{R_{\beta\alpha}} \;=\; \frac{f_\beta T_{\beta\alpha}}{f_\alpha T_{\alpha\beta}} \;. \tag{9}$$

This single condition does not determine the two numbers $R_{\alpha\beta}$ and $R_{\beta\alpha}$ completely. We can multiply both by the same factor and preserve (9). The Metropolis idea is to choose that factor as large as possible so that rejection is as unlikely as possible. That says that $\max\bigl(R_{\alpha\beta}, R_{\beta\alpha}\bigr) = 1$. This is achieved by

$$R_{\alpha\beta} \;=\; \min\left(\frac{f_\beta T_{\beta\alpha}}{f_\alpha T_{\alpha\beta}}, 1\right) \;. \tag{10}$$

For example, if

$$\frac{f_\beta T_{\beta\alpha}}{f_\alpha T_{\alpha\beta}} \;\leq 1 \;,$$

then

$$R_{\alpha\beta} \;=\; \frac{f_\beta T_{\beta\alpha}}{f_\alpha T_{\alpha\beta}} \quad \text{and} \quad R_{\beta\alpha} = 1 \;,$$

so (10) is satisfied.

The version of Metropolis just described is due to Hastings and often is called Metropolis–Hastings. The original Metropolis (which more properly is called $\mathrm{MR}^2\mathrm{T}^2$ for Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller) was a special case and was described differently. Suppose $H(\alpha)$ represents the energy of state α (any function can be an energy function). The *Gibbs– Boltzmann* probability distribution for temperature T is

$$f_\alpha \;=\; \frac{1}{Z} e^{-H(\alpha)/kT} \;. \tag{11}$$

Here k is *Boltzmann's constant*, which is a conversion factor from temperature (in degrees) to some units of energy.

The probability distribution (11) says that the probability of configuration α is depends on the energy of that configuration. Configurations with less energy are more likely. How much more likely depends on the temperature. For large T there is less penalty for large $H(\alpha)$, but for low temperature (small T) low energy states are more strongly preferred. For example, the minimum energy configuration of a collection of atoms may be a regular crystal. For low temperature, only approximately crystalline states have significant probability. At higher temperature, a greater variety of configurations are likely, which correspond to disordered liquid or gas states.

The normalization constant, $Z(T)$, is the *partition function*. In typical physical applications, the energy function and temperature are known, but the partition function is not. One of the challenges of sampling (11) is that Z is not

known. The Metropolis version had (unnecessarily) $T_{\alpha\beta} = T_{\beta\alpha}$ In this case, the ratio in (10) is

$$\frac{f_\beta}{f_\alpha} \;=\; e^{-(H(\beta)-H(\alpha))/kT} \;=\; e^{-\Delta H/kT} \;.$$

The Metropolis version of the rejection algorithm, then, is: first choose β using probabilities $T_{\alpha\beta}$, then compute the energy difference $\Delta H = H(\beta) - H(\alpha)$. If $\Delta H < 0$, the new state (β) is more likely than the old state – accept the move and take $X^{m+1} = \beta$. If $\Delta H > 0$, the new state is less likely. Accept it with probability $e^{-\Delta H/kT}$. None of these computations require us to know the partition function.

As an example, suppose $\mathcal{S} = Z$ (the integers) and $H(\alpha) = \alpha^2$, so that[4] $f_\alpha = \frac{1}{Z} e^{-\alpha^2/kT}$. Suppose the trial move is to move left or right with probability half: $T_{\alpha\beta} = \frac{1}{2}$ if $|\alpha - \beta| = 1$ and $T_{\alpha\beta} = 0$ otherwise. One step of the Metropolis algorithm would be to propose $X^{m+1} = X^m \pm 1$ (equal probabilities), compute ΔH (which is negative if the proposed move brings X closer to the origin), accept if ΔH is negative, accept with probability $e^{-\Delta H/kT}$ if ΔH is negative, and otherwise take $X^{m_1} = X^m$. Although the trial move Markov chain is cyclic with period 2, the full Markov chain with rejection is acyclic because of the possibility of rejection.

The efficiency of this method or sampling f is temperature dependent. If T is small, f is concentrated on small integers and not very many steps of the algorithm can take us from any likely state to any other one. For high temperature, the distribution of f is broader and more steps may be needed to get from one somewhat likely state to another. We will give a more quantitative treatment of this later. While the effectiveness of Metropolis or other samplers usually is temperature dependent, most samplers have more trouble at low temperature than at high temperature.

The principle of detailed balance also applied to continuous probability distributions and densities. Suppose $f(x)$ is a probability density we wish to sample. The detailed balance condition relative to probability density f is that the probability density to go from x to y is the same as the density to go from y to x:

$$f(x)p(x,y) = f(y)p(y,x) \;. \tag{12}$$

Integrating the detailed balance condition with respect to x and using the fact that $\int p(y,x)dx = 1$ for all y (p is a probability density with respect to its second argument) gives the balance condition from the detailed balance condition. We saw that the transition probability kernel for the Ornstein Uhlenbeck process satisfies detailed balance (12).

It may be unclear how to apply the symmetry condition (12) when the transition density is replaced by a more general transition probability measure. For this is may be helpful to use the reformulation of detailed balance as a symmetry condition of operators, as we did for Ornstein Uhlenbeck. One way to

[4]Note the conflict of notation. The letter Z represents both the partition function normalization constant and the integers.

do this is just to multiply the supposed pointwise identity (12) by *test functions* $u(x)$ and $v(y)$ and integrate

$$\int \int u(x) f(x) p(x,y) v(y) \, dx dy \;=\; \int \int u(x) f(y) p(y,x) v(y) \, dx dy \tag{13}$$

In case p is a measure as a function of y, the dy parts of these integrals are with respect to that measure.

As for the Ornstein Uhlenbeck process, detailed balance in general can be interpreted as the operator determined by p being symmetric (or self-adjoint) in the weighted L^2 inner product, weighted by f. In the discrete case, the weighted inner product is

$$\langle u, v \rangle_f \;=\; \sum_\alpha u_\alpha v_\alpha f_\alpha \; .$$

Let A be an $s \times s$ matrix. The adjoint of A in the f weighted inner product, A_f^*, is determined by the requirement that

$$\langle u, Av \rangle_f \;=\; \langle A_f^* u, v \rangle_f \; ,$$

for all u and v. In particular, P is self-adjoint in the f weighted inner product if

$$\langle u, Pv \rangle_f \;=\; \langle Pu, v \rangle_f \; ,$$

for all u and v. If this is expressed in components using the summation convention (e.g. $(Pv)_\alpha = p_{\alpha\beta} v_\beta$), the result is

$$u_\alpha p_{\alpha\beta} v_\beta f_\alpha \;=\; p_{\alpha\beta} u_\beta v_\alpha f_\alpha \; ,$$

for any sets of numbers u_α and v_α. Interchanging the labels α and β on the right gives

$$u_\alpha v_\beta \left(p_{\alpha\beta} f_\alpha \right) \;=\; u_\alpha v_\beta \left(p_{\beta\alpha} f_\beta \right) \; .$$

Since the u_α and v_β are arbitrary, this implies that $p_{\alpha\beta} f_\alpha = p_{\beta\alpha} f_\beta$ for all α and β. This is the detailed balance condition (7).

The story is the same for continuous probability. The weighted L^2 inner product is

$$\langle u, v \rangle_f \;=\; \int u(x) v(x) f(x) \, dx \; .$$

The transition density or transition measure defines an operator through

$$(Pv)(x) \;=\; \int p(x,y) v(y) \, dy \; ,$$

in the case of a bona fide probability density and a related expression in case p is a measure as a function of y. The detailed balance condition is equivalent to the requirement that

$$\langle u, Pv \rangle_f \;=\; \langle Pu, v \rangle_f \; . \tag{14}$$

In fact, this is exactly (13).

4 Partial resampling

Looking for a Markov chain that preserves f is the same as looking for random *moves* that leave f invariant. Often the moves change X in a small way, such as changing only one component of X as in the rod example above. Suppose P_k changes X_k in some way without changing the other components, such as uniform resampling of X_k in the rod example. The moves P_k, for different k, must be combined in some way to create a Markov chain that is irreducible (and aperiodic). We use the term *partial resampling* for the strategy of combining a collection of simple moves to make an effective dynamic sampler.

Suppose we have a collection of moves P_k each of which preserves f ($f = fP_k$). There are at least two ways to combine the P_k to make an irreducible Markov chain. One is to do them in some order, say first P_1, then P_2, and so on. Since each of these moves preserves f, doing them in order also will preserve f. This gives a Markov chain with transition matrix $P_1 P_2 \cdots P_n$. In our notation, this corresponds to first doing P_1 then P_2, and so on.

Even if the individual moves satisfy detailed balance, the composite resampler P may not. Mathematically, we can see this by noting that the product of symmetric matrices need not be symmetric. A simple example is sampling the set of permutations of three letters, a, b, c, with all permutations being equally likely. Let P_1 interchange the first two. This satisfies detailed balance because

$$\Pr(\,(X_1, X_2, X_3) \to X_2, X_1, X_3)\,) \;=\; \Pr(\,(X_2, X_1, X_3) \to (X_1, X_2, X_3)\,)\;.$$

It also works to have P_1 interchange X_1 and X_2 with a given probability. Let P_2 interchange X_2 and X_3, again possibly with a probability possibly less than one. Both P_1 and P_2 have detailed balance, but the composite $P_1 P_2$ does not. One could see this by writing the transitions matrices (6×6 for the 6 dimensional state space of all permutations of 3 items). Another way to see is is to note that under $P_1 P_2$ the transformation

$$(\underline{a}, \underline{b}, c) \;\stackrel{P_1}{\to}\; (b, \underline{a}, \underline{c}) \;\stackrel{P_2}{\to}\; (b, c, a)$$

is possible, but the reverse, $(b, c, a) \to (a, b, c)$, cannot be achieved first by interchanging the first pair ((b, a) in this case) then the second pair.

One view of partial resampling is as follows. Suppose $X = (X_1, \ldots, X_n)$ and we want to sample the distribution $f(x_1, \ldots, x_n)$. Let

$$f_k(x_k \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n) \;=\; \frac{f(x_1, \ldots, x_n)}{Z(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)}$$

be the conditional distribution of X_k with all other components fixed. If X is a sample of f and X'_k is a sample of

$$f_k(x_k \mid X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)\;,$$

then $X' = (X_1, \ldots, X_{k-1}, X'_k, X_{k+1}, \ldots, X_n)$ also is a sample of f.

It is not necessary that the new X'_k is independent (or conditionally independent given the other components) of X_k. For example, we could resample X_k using a Metropolis proposal and rejection strategy. This would give a positive probability that $X' = X_k$.

The terms *heat bath* or *Gibbs sampler* are used for partial resampling strategies that give an X'_k that is independent of X_k, of course conditionally given the values of the X_j for $j \neq k$. This may seem to be the optimal resampling strategy, since it introduces the most new randomness. For example, if we would resample X_k repeatedly using Metropolis, then after many resamplings, we would get an effectively independent resample, which is what heat bath does in one step. However, there is a counter-example below to the statement that heat bath is always better than strategies that make X'_k dependent on X_k. Moreover, the extra work to go from dependent Metropolis resampling to fully independent heat bath may not produce gains that justify the cost. If we have just resampled X_k using Metropolis, should we resample X_k again to get a more independent resample, or would it do more good to resample X_{k+1} instead?

If we have partial resamplers P_k, we must assemble them to create an overall resampler that is irreducible. One could do the partial resamplings in a specified order, say, $P = P_1 \cdots P_n$, a strategy called *sweeping*, or *systematic scan*. One also could choose k at random (all choices independent) and perform P_k. This is called *random site updating*, or *random scan*. It has the feature that (if each k is equally likely)

$$P_{\text{rs}} \;=\; \frac{1}{n}\sum_{k=1}^{n} P_k$$

is self adjoint if each of the individual P_k satisfies detailed balance (This can be argued directly.).

4.1 Heat bath resampling of Gaussians

It is possible to analyze the heat bath algorithm applied to a multivariate Normal. It is unlikely that one would actually sample a multivariate normal in this way, given the alternatives (Choleski for moderate n, multi-grid for very large n). Still, the analysis gives much insight into the strengths and weaknesses of partial resampling strategies for other distributions that lack very effective samplers. In this sense, the Gaussian sampling problem is a *model problem*. We study it because we want some insight into how it might work in other situations.

Suppose $X \in R^n$ is Gaussian with probability density $f(x) = \frac{1}{Z} e^{-x^* H x/2}$ (H symmetric and positive definite). The conditional density of X_k is a one dimensional Gaussian whose mean and variance are easy to calculate. We will see that the conditional mean, $\mu_k(X_1, \ldots, X_{k-1}, X_{k+1}, X_n)$, is a linear function of the other component, while the conditional variance, σ_k^2, is independent of X. This means that we can resample X_k using

$$X'_k \;=\; \mu_k(X_1, \ldots, X_{k-1}, X_{k+1}, X_n) \;+\; \sigma_k Z_k \;, \tag{15}$$

where the Z_k are independent standard normals.

The formulas for μ_k and σ_k may be derived as follows. The Gaussian probability density may be written out as (using the summation convention, and h_{jk} for the (j,k) entry of H)

$$f(x) \;=\; \frac{1}{Z} \exp\left(-x_i x_j h_{ij}/2\right) \;.$$

Within the exponential, the terms that depend on x_k are $x_k^2 h_{kk}/2$, and

$$\sum_{j \neq k} x_j x_k h_{jk} \;=\; \left(\sum_{j \neq k} x_j h_{jk} \right) x_k \;=\; l x_k$$

(the factor of 2 disappears since the terms $x_j x_k h_{jk}$ and $x_k x_j h_{kj}$ are equal). Therefore (completing the square)

$$\begin{aligned} f(x) &= e^{-l x_k - h_{kk} x_k^2/2} \;\times\; (\text{indep of } x_k) \\ &= e^{-h_{kk}(x_k + l/h_{kk})^2/2} \;\times\; (\text{also indep of } x_k) \;. \end{aligned}$$

Collecting these calculations gives, $\sigma_k^2 = 1/h_{kk}$ and

$$\mu_k \;=\; \frac{-1}{h_{kk}} \left(\sum_{j \neq k} h_{kj} x_j \right) \;. \tag{16}$$

Suppose we sweep through the components doing a heat bath resampling of each one, starting with $X^m = (X_1^m, \ldots, X_n^m)$. Resampling X_1 gives, using (15) and (16),

$$X_1^{m+1} \;=\; \frac{-1}{h_{11}} \left(\sum_{j=2}^{n} h_{1j} X_j^m \right) \;+\; \frac{1}{\sqrt{h_{11}}} Z_1^m \;.$$

When we then resample X_2, we use the current configuration, which uses the new X_1 and the old X_3, etc. The result is

$$X_2^{m+1} \;=\; \frac{-1}{h_{22}} \left(h_{21} X_1^{m+1} + \sum_{j=3}^{n} h_{2j} X_j^m \right) \;+\; \frac{1}{\sqrt{h_{22}}} Z_2^m \;.$$

In general, we have

$$X_k^{m+1} \;=\; \frac{-1}{h_{kk}} \left(\sum_{j=1}^{k-1} h_{kj} X_j^{m+1} + \sum_{j=k+1}^{n} h_{kj} X_j^m \right) \;+\; \frac{1}{\sqrt{h_{kk}}} Z_k^m \;. \tag{17}$$

One complete sweep of the *single site* (single component) heat bath algorithm is to apply (17) for $k = 1$, $k = 2$, …, $k = n$.

The result of a complete sweep is that

$$X^{m+1} \;=\; AX^m \;+\; BZ^m \;, \tag{18}$$

for some matrices A and B, with $Z^m \in R^n$ being independent standard n component normals. We can see this by noting that all the operations in (17) are linear in the X^m, X^{m+1}, and Z^m. This linear iteration can be understood just as we understood the multi-dimensional Ornstein Uhlembeck process, and they depend entirely on the eigenvalues of A. Not only do they not depend on the matrix, B, they even do not depend on the distribution of the Z^m,

More precisely, let $f^m(x)$ be the probability density of X^m. Suppose the Z^m are independent samples of probability density $h(z)$. From (18) we have

$$f^{m+1}(x) = \int \delta(x - Ay - Bz) f^m(y) h(z)\, dy dz \;. \tag{19}$$

As for the general case, we can write this abstractly as $f^{m+1} = f^m P$. It is not necessary in this argument to write an explicit formula for transition probability density $p(x, y)$. Suppose A has n real eigenvalues and corresponding eigenvectors $Av_k = \mu_k v_k$. Let f be the invariant density for (19), so that if $f^m = f$ then $f^{m+1} = f$. It is not necessary to assume that f is Gaussian, though it will be if h is Gaussian. Let α_1, ..., α_n be non-negative integers, and $\lambda_\alpha = \alpha_1\mu_1 + \cdots + \alpha_n\mu_n$. Let

$$g_\alpha(x) \;=\; \bigl(v_1 \cdot \nabla\bigr)^{\alpha_1} \cdots \bigl(v_n \cdot \nabla\bigr)^{\alpha_n} f \;.$$

Precisely as for the Ornstein Uhlenbeck process, it is easy to see that these g_α are eigenfunctions (left eigenfunctions) of the operator P, $g_\alpha P = \lambda_\alpha g_\alpha$, in the sense that if $f^m = g_\alpha$ in (19), then

$$f^{m+1} \;=\; \lambda_\alpha g_\alpha \;.$$

5 Error bars and autocorrelation time

Estimating error bars for (1) is more complicated than in the case of uncorrelated samples. They have in common the central limit theorem, either the simple one for independent samples or the more subtle one for Markov chains. In both cases, the error bar, for large enough L, is determined by the variance of $\widehat{A}_L$, which can be estimated from Monte Carlo data. The difference is that the straightforward variance estimator for independent samples must be replaced by something more complicated.

The basic fact is that for large L, the variance of $\widehat{A}_L$ is given (approximately) by

$$\mathrm{var}\left(\widehat{A}_L\right) \;\approx\; \frac{D}{L} \;,$$

where D plays the role of $\sigma^2(V(X))$ for independent samples, and is given by the *Kubo* formula

$$D \;=\; \sum_{t=-\infty}^{\infty} \mathrm{cov}_f\left(V(X^0), V(X^t)\right) \;. \tag{20}$$

In this formula, we suppose that we start not at time $m = 0$, but some time in the distant past. In this way, X^t is defined for all t. An equivalent definition of the time t *covariance function*, which is more practical for computing, is

$$C(t) \;=\; \lim_{m\to\infty} \operatorname{cov}\left(V(X^m), V(X^{m+t}) \right) \;.$$

This limit should exist regardless of the starting conditions and is defined for t positive or negative. It is clear from this that

$$C(-t) \;=\; C(t) \;,$$

and that

$$C(0) \;=\; \operatorname{var}_f \left(V(X) \right) \;.$$

Therefore, we may write

$$D \;=\; \operatorname{var}_f \left(V(X) \right) \;+\; 2 \sum_{t=1}^{\infty} C(t) \;.$$

A final rewrite of (20) involves the *correlation function*

$$\begin{aligned} \rho(t) \;&=\; \operatorname{corr}_f \left(V(X^0), V(X^t) \right) \\ &=\; \frac{\operatorname{cov}_f \left(V(X^0), V(X^t) \right)}{\sqrt{\operatorname{var}_f \left(V(X^0) \right) \cdot \operatorname{var}_f \left(V(X^t) \right)}} \\ &=\; \frac{C(t)}{C(0)} \;. \end{aligned}$$

Dividing by $C(0)$ gives

$$D \;=\; \operatorname{var}_f \left(V(X) \right) \cdot \left(1 + 2 \sum_{t=1}^{\infty} \rho(t) \right) \;=\; \sigma^2 \tau \;, \tag{21}$$

where $\sigma^2 = \operatorname{var}_f \left(V(X) \right)$ is the *static* variance, and

$$\tau \;=\; 1 + 2 \sum_{t=1}^{\infty} \rho(t) \;, \tag{22}$$

is the *autocorrelation time* (more properly, em integrated autocorrelation time).

With all this notation, the dynamic Monte Carlo error bar may be stated

$$\operatorname{var}\left(\widehat{A}_L \right) \;\approx\; \frac{\sigma^2}{L_{\text{eff}}} \quad , \qquad L_{\text{eff}} \;=\; \frac{L}{\tau} \;. \tag{23}$$

This says that the error bar for dynamic Monte Carlo may be understood as the error bar for static Monte Carlo (independent samples) provided that we use the *effective sample size*, L_{eff}, which is the number of dynamic Monte Carlo steps divided by the integrated autocorrelation time. In other words, we can

think of τ as the number of dynamic Monte Carlo steps needed to produce an effectively independent sample.

One simple example is the one dimensional version of the linear Gaussian iteration (18), $X^{m+1} = aX^m + bZ^m$, with observable $V(x) = x$. If we start X^0 in the invariant distribution, then the formula

$$X^t = a^t X^0 + a^{t-1} bZ^0 + \cdots + bZ^{t-1}$$

makes it clear that

$$C(t) = E_f \left[X^t X^0 \right] = E_f \left[a^t X^0 X^0 \right] = a^t \sigma^2 \ .$$

Of course, it is necessary that $|a| < 1$ for the steady state to exist. In that case, we have

$$\tau = 1 + 2 \sum_{t=1}^{\infty} a^t = 1 + \frac{2a}{1-a} = \frac{1+a}{1-a} \ . \tag{24}$$

The bad case, from the point of view of Monte Carlo efficiency, is $a = 1 - \epsilon$, which has autocorrelation time $\tau \approx 2/\epsilon$. The correlation at that time is

$$\rho(\tau) \ \approx \ (1-\epsilon)^{2/\epsilon} \ \approx \ e^{-2} \ .$$

This shows that $X^{m+\tau}$ is not terribly close to being actually independent of X^m. The correlation coefficient is $1/e^2 \approx .15$. The term *effectively independent* does not mean *nearly independent*, but rather that the error bars have about the same size you would get from that many actually independent samples.

5.1 Estimating τ

From (23) we see that estimating error bars is equivalent to estimating σ^2, the static variance, and τ, the autocorrelation time. The static variance we can get as before:

$$\widehat{\sigma^2} \ = \ \frac{1}{L} \sum_{m=1}^{L} \left(V(X^m) - \widehat{A} \right)^2 \ . \tag{25}$$

There is no equally satisfactory way to estimate τ. We describe two somewhat unsatisfactory methods and wait for someone to produce something better (Hello, ... Godot?).

The method of *batched means* is the simplest and least accurate. This method divides the time series $V(X^m)$ into r sub-series, or *batches*, of length L/r. Batch B_j is the sub-sequence $m = (jL/r) + 1, \cdots, (j+1)r/L$. The batch means are

$$\mu_j = \frac{1}{L/r} \sum_{m \in B_j} V(X^m) \ . \tag{26}$$

Of course, the overall mean is the mean of the batch means, (1) is equivalent to:

$$\widehat{A}_L \ = \ \frac{1}{r} \sum_{j=1}^{r} \mu_j \ . \tag{27}$$

The idea is that if $L/r \gg \tau$, then the μ_j are nearly independent and identically distributed, independent enough so that the variance in (27) is approximately (assuming each of the μ_k has the same variance)

$$\mathrm{var}\left(\widehat{A}_L\right) \;\approx\; \frac{1}{r}\mathrm{var}\left(\mu_k\right) \;, \tag{28}$$

with[5]

$$\mathrm{var}\left(\mu_k\right) \;\approx\; \frac{1}{r-1}\sum_{j=1}^{r}\left(\mu_j - \widehat{A}_L\right)^2 \;. \tag{29}$$

The procedure is: Compute the batch means (26) and the overall mean (27), then estimate the variance of the batch means using the variance estimator for independent samples (28) and use this to estimate the variance of the overall mean (29).

Inaccuracy is one problem with the method of batched means. Its accuracy depends on r, the number of batches, which can be small. In principle, we might expect the accuracy of the variance estimate to depend on $L_{\mathrm{eff}} = L/\tau$. But we have to take the batch size $L/r \gg \tau$ to insure the μ_j are (almost) independent, so $r \ll L_{\mathrm{eff}}$, the number of batches is much smaller than the effective sample size. Being sure of the batch size is another problem. It would be prudent to combine the estimate (29) with the estimate of the static variance (25) to get a rudimentary estimate of the autocorrelation time using (23):

$$\widehat{\tau} \;=\; \frac{L\,\mathrm{var}\left(\widehat{A}_L\right)}{\widehat{\sigma^2}} \;.$$

If the batch size is too small relative to the estimated autocorrelation time, say $L/r < 5\widehat{\tau}$, the batch size is too small.

Estimating the covariance function may lead to a more accurate estimate of τ. A natural estimator is

$$\widehat{C}(t) \;=\; \frac{1}{L-t}\sum_{m=1}^{T-t}\left(V(X^{m+t}) - \widehat{A}\right)\left(V(X^m) - \widehat{A}\right) \;. \tag{30}$$

We then could estimate D using

$$\text{wrong} \qquad\qquad \widehat{D} \;=\; \widehat{C}(0) \;+\; 2\sum_{t=2}^{L}\widehat{C}(t) \;. \qquad\qquad \text{wrong}$$

A straightforward but lengthly calculation shows that this is an *inconsistent* estimator in the sense that $\mathrm{var}\left(\widehat{D_L}\right) \not\to 0$ as $L \to \infty$. All the $\widehat{C}(t)$ terms on the right contribute more noise than signal. For $t \gg \tau$, $C(t)$ is essentially zero,

[5] We write $r-1$ instead of the customary r in the denominator to emphasize that r may be so small that the difference between r and $r-1$ is significant.

but the estimate still has statistical error. All these statistical errors add up to something bigger than the signal in $\widehat{D}$.

The cure, obviously, is to exclude the values of $\widehat{C}(t)$ for $t \gg \tau$. We replace the inconsistent estimator with

$$\widehat{D} \;=\; \widehat{C}(0) \;+\; 2\sum_{t=2}^{T} \widehat{C}(t) \;, \tag{31}$$

where T is chosen to balance the bias from taking T small against the statistical error from taking T large. An ideal value for applications might be $T \approx 3\tau$. Of course, τ itself must be estimated, probably by

$$\widehat{\tau} \;=\; \frac{\widehat{D}}{\widehat{\sigma^2}} \;.$$

Using $T < L$ is called *windowing*, or *smoothing*. Taking the window size, T, to depend on the estimate of τ makes the procedure *self consistent*. You want the cutoff, T, in (31) to be related to the time scale of the time series $V(X^m)$, but you do not know that time scale in advance. Therefore, you estimate τ and D at the same time and stop when the cutoff time used to estimate D is consistent (self consistent) with the time scale, τ, implied by $\widehat{D}$.

The term *smoothing* comes from a more general problem that D estimation is a special case of, *spectral density* estimation. A time series, Y^m (defined for all integers, m) is *stationary* if the statistics of Y are the same as the statistics of a *shifted* sequence $\widetilde{Y}^m = Y^{m+t}$. The time series $Y^m = V(X^m)$ is supposed to become stationary after several autocorrelation times. Assuming $E[Y^m] = 0$ (like $V(X^m) - A$, for large m), the auto-covariance function is $C_Y(t) = E[V^0 Y^t]$. The spectral density is the discrete Fourier transform[6] of the auto-covariance function:

$$F(k) \;=\; \sum_{t=-\infty}^{\infty} e^{-ikt} C(t) \;. \tag{32}$$

The Kubo constant is a particular value: $D = F(0)$. Spectral density estimation is the problem if estimating the spectral density from a length L piece of the Y sequence. As discussed for D, the naive estimate that uses every available value of $\widehat{C}(t)$

$$\widehat{F}_0(k) = \sum_{t=-L}^{L} \widehat{C}(t) \;, \tag{33}$$

in (32) is an inconsistent estimator.

Of one plots $\widehat{F}_0(k)$, the striking thing is how noisy it is. It changes completely from one k value to the next. This motivates the attempt to improve the estimate through smoothing as we did in kernel density estimation. If $\phi(k)$ is

[6] You can skip this paragraph if you are not familiar with the discrete Fourier transform or FFT.

a smooth localized *mollifier*, then the *approximate identity* is $\phi_\epsilon(k) = \frac{1}{\epsilon}\phi(k/\epsilon)$, and the smoothed estimator is

$$\widehat{F}_\epsilon \;=\; \int_{k=-\pi}^{\pi} \widehat{F}_0(k-l)\phi_\epsilon(l)\,dl \tag{34}$$

Increasing ϵ in(34) gives more averaging together with more bias coming from the fact that

$$F(k) \;\neq\; F_\epsilon(k) \;=\; \int_{k=-\pi}^{\pi} F(k-l)\phi_\epsilon(l)\,dl\ ,$$

although the error is small if ϵ is small and ϕ satisfies moment conditions.

The relation to windowing comes from considering the Fourier transform of ϕ_ϵ, which means the numbers

$$w_\epsilon(t) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ikt}\phi_\epsilon(k)\,dk\ .$$

The Plancharel formula for Fourier series implies that

$$F_\epsilon(k) \;=\; \sum_{-\infty}^{\infty} e^{-ikt} w_\epsilon(t) C(t)\ .$$

In particular,

$$D \;\approx\; F_\epsilon(0) \;=\; \sum_{-\infty}^{\infty} w_\epsilon(t) C(t)\ .$$

If $w_\epsilon(t)$ is real and symmetric (it should be the resulting estimator is

$$\widehat{D} \;=\; \widehat{\sigma^2} \;+\; 2\sum_{t=1}^{L} w_\epsilon(t)\widehat{C}(t)\ . \tag{35}$$

This is a generalization of the simple windowed estimator (31), as we see by taking $w_\epsilon(t) = 1$ for $t \leq T$ and $w_\epsilon(t) = 0$ for $t > T$. Other window functions, $w(t)$, besides the step function window (31) might be more accurate. This is a topic that could use further research.