Minimising MCMC Variance via Diffusion Limits, with an Application to Simulated Tempering

by

Gareth O. Roberts[1] and Jeffrey S. Rosenthal[2]

(Version of March 28, 2012.)

Abstract. We derive new results comparing the asymptotic variance of diffusions, by writing them as appropriate limits of discrete-time birth-death chains and taking limits of Peskun orderings. We then apply our results to simulated tempering algorithms, to establish which choice of inverse temperatures minimises the asymptotic variance of all functionals and thus leads to the most efficient MCMC algorithm.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms are very widely used to approximately compute expectations with respect to complicated high-dimensional distributions (see e.g. [25, 7]). Specifically, if a Markov chain $\{X_n\}$ has stationary distribution π on state space $\mathcal{X}$, and $h : \mathcal{X} \to \mathbf{R}$ with $\pi|h| < \infty$, then $\pi(h) := \int h(x)\,\pi(dx)$ can be estimated by $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$ for suitably large n. This estimator is unbiased if the chain is started in stationarity (i.e. if $X_0 \sim \pi$), and in any case has bias only of order $1/n$. Furthermore, it is consistent provided the Markov chain is ϕ-irreducible. Thus, the efficiency of the estimator is often measured in terms of the asymptotic variance $\mathbf{Var}_\pi(h, P) := \lim_{n\to\infty} \frac{1}{n}\mathbf{Var}_\pi\Big(\sum_{i=1}^{n} h(X_i)\Big)$ (where the subscript π indicates that $\{X_n\}$ is in stationarity): the smaller the variance, the better the estimator.

An important question in MCMC research is how to *optimise* it, i.e. how to choose the Markov chain optimally (see e.g. [10, 16]). This leads to the question of how to *compare* different Markov chains. Indeed, for two different ϕ-irreducible Markov chain kernels P_1 and P_2 on $\mathcal{X}$, both having the same invariant probability measure π, we say that P_1 *dominates* P_2 *in the efficiency ordering*, written $P_1 \succeq P_2$, if $\mathbf{Var}_\pi(h, P_1) \le \mathbf{Var}_\pi(h, P_2)$ for all $L^2(\pi)$

[1]Department of Statistics, University of Warwick, CV4 7AL, Coventry, U.K. Email: g.o.roberts@lancaster.ac.uk.

[2]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Email: jeff@math.toronto.edu. Web: http://probability.ca/jeff/ Supported in part by NSERC of Canada.

functionals $h : \mathcal{X} \to \mathbf{R}$, i.e. if P_1 is "better" than P_2 in the sense of being uniformly more efficient for estimating expectations of functionals.

It was proved by Peskun [19] for finite state spaces, and by Tierney [26] for general state spaces (see also [17, 16]), that if P_1 and P_2 are discrete-time Markov chains which are both reversible with respect to the same stationary distribution π, then a sufficient condition for $P_1 \succeq P_2$ is that $P_1(x, A) \geq P_2(x, A)$ for all $x \in \mathcal{X}$ and $A \in \mathcal{F}$ with $x \notin A$, i.e. that P_1 *dominates* P_2 *off the diagonal*.

Meanwhile, diffusion limits have become a common way to establish asymptotic comparisons of MCMC algorithms [21, 22, 23, 2, 3, 4, 5]. Specifically, if $P_{1,d}$ and $P_{2,d}$ are two different Markov kernels in dimension d (for $d = 1, 2, 3, \ldots$), with diffusion limits $P_{1,*}$ and $P_{2,*}$ respectively as $d \to \infty$, then one way to show that $P_{1,d}$ is more efficient than $P_{2,d}$ for large d is to prove that $P_{1,*}$ is more efficient that $P_{2,*}$. This leads to the question of how to establish that one diffusion is more efficient than another. In some cases (e.g. random-walk Metropolis [21], and Langevin algorithms [22]), this is easy since one diffusion is simply a time-change of the other. But more general diffusion comparisons are less clear (the spectral gaps can be ordered directly using Dirichlet forms, but this does not bound the asymptotic variances).

In this paper, we develop (Section 2) a new comparison of asymptotic variance of diffusions. Specifically, we prove (Theorem 1) that if P_i are Langevin diffusions with respect to the same stationary distribution π, with variance functions σ_i^2 (for $i = 1, 2$), then if $\sigma_1^2(x) \geq \sigma_2^2(x)$ for all x, then $P_1 \succeq P_2$, i.e. P_1 is more efficient than P_2. (We note that that Mira and Leisen [12, 18] extended the Peskun ordering in an interesting way to continuous-time Markov processes on finite state spaces, and on general state spaces when the processes have generators which can be represented as $G_i f(x) = \int f(y)\, Q_i(x, dy)$ and which satisfy the condition that $Q_1(x,\ A \setminus \{x\}) \geq Q_2(x,\ A \setminus \{x\})$ for all x and A. However, their results do not appear to apply in our context, since generators of diffusions involve differentiation and thus do not admit such representation.)

We then consider (Section 3) simulated tempering algorithms [14, 10], and in particular the question of how best to choose the intermediate temperatures. It was previously shown in [1], generalising some results in the physics literature [11, 20], that a particular choice of temperatures (which leads to an asymptotic temperature-swap acceptance rate of 0.234) maximises the L^2 jumping distance. (Indeed, this result has already influenced adaptive MCMC algorithms for simulated tempering; see e.g. [9].) However, the previous papers did *not* prove a diffusion limit, nor did they provide any comparisons of Markov chain variances. In this paper, we establish (Theorem 6) diffusion limits for certain simulated tempering algorithms. We then apply our diffusion comparison results to prove (Theorem 7) that the given choice of temperatures does indeed minimise the asymptotic variance of all functionals.

2 Comparison of Diffusions

Let $\pi : \mathcal{X} \to (0, \infty)$ be a C^1 target density function, where $\mathcal{X}$ is either $\mathbf{R}$ or some finite interval $[a, b]$. We shall consider non-explosive Langevin diffusions X^σ on $\mathcal{X}$ with stationary density π, satisfying

$$dX_t^\sigma = \sigma(X^\sigma)\, dB_t + \left(\frac{1}{2}\, \sigma^2(X_t^\sigma) \log \pi'(X_t^\sigma) + \sigma(X_t^\sigma)\sigma'(X_t^\sigma)\right) dt\,, \tag{1}$$

for some C^1 function $\sigma : \mathcal{X} \to [\underline{k}, \overline{k}]$ for some fixed $0 < \underline{k} < \overline{k} < \infty$, and with reflecting boundaries at a and b in the case $\mathcal{X} = [a, b]$.

For two such diffusions X^{σ_1} and X^{σ_2}, we write (similar to the above) that $X^{\sigma_1} \succeq X^{\sigma_2}$, and say that X^{σ_1} *dominates* X^{σ_2} *in the efficiency ordering*, if for all $L^2(\pi)$ functionals $f : \mathcal{X} \to \mathbf{R}$,

$$\lim_{T\to\infty} T^{-1/2}\, \mathrm{Var}\left(\int_0^T f(X_s^{\sigma_1})\, ds\right) \;\leq\; \lim_{T\to\infty} T^{-1/2}\, \mathrm{Var}\left(\int_0^T f(X_s^{\sigma_2})\, ds\right)\ .$$

We wish to argue that if $\sigma_1(x) \geq \sigma_2(x)$ for all x, then $X^{\sigma_1} \succeq X^{\sigma_2}$. Intuitively, this is because X^{σ_1} "moves faster" that X^{σ_2}, while maintaining the same stationary distribution. Indeed, if σ_1 and σ_2 are constants, then this result is trivial (and implicit in earlier works [21, 22, 23]), since then $X_t^{\sigma_1}$ has the same distribution as $X_{ct}^{\sigma_2}$ where $c = \sigma_1/\sigma_2 > 1$, i.e. X^{σ_1} accomplishes the same sampling as X^{σ_2} in a shorter time so it must be more efficient. However, if σ_1 and σ_2 are non-constant functions, then comparison of X^{σ_1} and X^{σ_2} is less clear.

To make theoretical progress, we assume:

(A1) π is log-Lipschitz function on $\mathcal{X}$, i.e. there is $L < \infty$ with

$$|\log \pi(y) - \log \pi(x)| \;\leq\; L\,|y - x|\,, \qquad x, y \in \mathcal{X}\,. \tag{2}$$

(A2) Either **(a)** $\mathcal{X}$ is a bounded interval $[a, b]$, and the diffusions X^σ have reflecting boundaries at a and b; or **(b)** $\mathcal{X}$ is all of $\mathbf{R}$, and π has exponentially-bounded tails, i.e. there is $0 < K < \infty$ and $r > 0$ such that

$$\pi(x + y) \;\leq\; \pi(x)e^{-ry}\,, \qquad x > K,\; y > 0$$

and

$$\pi(x - y) \;\leq\; \pi(x)e^{-ry}\,, \qquad x < -K,\; y > 0\,.$$

In case A2(b), we can then find sufficiently large $q \geq K$ such that

$$\sum_{\substack{i \\ |i/m| \geq q}} \pi(i/m) \ \leq \ (1/4) \sum_i \pi(i/m) \,, \qquad \text{for all } m \in \mathbf{N} \tag{3}$$

(where the sums in (3) must be finite due to (2)), and then set

$$Q \ = \ \inf\{\pi(x) : |x| \leq q+1\} \,, \tag{4}$$

which must be positive by continuity of π and compactness of the interval $[-q-1, \ q+1]$.

Our main result is then the following.

Theorem 1. *If X^{σ_1} and X^{σ_2} are two Langevin diffusions of the form (1) with respect to the same density π, with variance functions σ_1 and σ_2 respectively, and if $\sigma_1(x) \geq \sigma_2(x)$ for all $x \in \mathcal{X}$, then assuming (A1) and (A2), we have $X^{\sigma_1} \succeq X^{\sigma_2}$.*

2.1 Proof of Theorem 1

To prove Theorem 1, we introduce auxiliary processes for each $m \in \mathbf{N}$. Given $\sigma : \mathcal{X} \to \mathbf{R}$, let $S = 2\overline{k}e^L$, and let $Z^{m,\sigma}$ be a discrete-time birth and death process on the discrete state space $\mathcal{X}_m := \{i/m; \ i \in \mathbf{Z}\}$ in case A2(b), or $\mathcal{X}_m := \{i/m; i \in \mathbf{Z}\} \cap [a,b]$ in case A2(a), with transition probabilities given by

$$P(i/m, \ (i+1)/m) \ = \ \frac{1}{2S}\left(\sigma^2(i/m) + \frac{\sigma^2((i+1)/m) \ \pi((i+1)/m)}{\pi(i/m)}\right) \,,$$
$$P(i/m, \ (i-1)/m) \ = \ \frac{1}{2S}\left(\sigma^2(i/m) + \frac{\sigma^2((i-1)/m) \ \pi((i-1)/m)}{\pi(i/m)}\right) \,,$$

and

$$P(i/m, \ i/m) \ = \ 1 - P(i/m, \ (i+1)/m) - P(i/m, \ (i-1)/m) \,.$$

(In case A2(a), any transitions which would cause the process to move out of the interval $[a,b]$ are instead given probability 0.) These transition rates are chosen to satisfy detailed balance with respect to the stationary distribution π_m on $\mathcal{X}_m$ given by $\pi_m(i/m) = \pi(i/m)/\sum_{x\in\mathcal{X}_m}\pi(x)$ (and S is chosen to be large enough to ensure that $P(i/m, \ (i+1)/m) + P(i/m, \ (i-1)/m) \leq 1$).

In terms of $Z^{m,\sigma}$, we then let $\{Y^{\sigma}_{m,t}\}_{t\geq 0}$ be the continuous-time version of $Z^{m,\sigma}$, speeded up by a factor of $m^2S/2$, i.e. defined by $Y^{\sigma}_{m,t} = Z^{m,\sigma}_{\lfloor m^2St/2\rfloor}$ for $t \geq 0$. (Here and throughout, $\lfloor r\rfloor$ is the floor function which rounds r down to the next integer, e.g. $\lfloor 6.8\rfloor = 6$ and $\lfloor -2.1\rfloor = -3$). It then follows that $Y_{m,t}$ converges to $X^{m,\sigma}$, as stated in the following lemma (whose proof is deferred until the end of the paper, since it uses similar ideas to those of the following section).

Lemma 2. *Assuming (A1) and (A2), as $m \to \infty$, the processes Y_m^σ converge weakly (in the Skorokhod topology) to X^σ.*

We then apply the usual discrete-time Peskun ordering to the $Z^{m,\sigma}$ processes, as follows.

Lemma 3. *Suppose that $\sigma_1(x) \geq \sigma_2(x)$ for all $x \in \mathbf{R}$. Then $Z^{m,\sigma_1} \succeq Z^{m,\sigma_2}$.*

Proof. By inspection, the fact that $\sigma_1(x) \geq \sigma_2(x)$ implies that

$$\mathbf{P}(Z^{m,\sigma_1}_{(i+1)/m} = j+1 \,|\, Z^{m,\sigma_1}_{i/m} = j) \;\geq\; \mathbf{P}(Z^{m,\sigma_2}_{(i+1)/m} = j+1 \,|\, Z^{m,\sigma_2}_{i/m} = j)$$

and

$$\mathbf{P}(Z^{m,\sigma_1}_{(i+1)/m} = j-1 \,|\, Z^{m,\sigma_1}_{i/m} = j) \;\geq\; \mathbf{P}(Z^{m,\sigma_2}_{(i+1)/m} = j-1 \,|\, Z^{m,\sigma_2}_{i/m} = j)\,.$$

It follows that Z^{m,σ_1} dominates Z^{m,σ_2} off the diagonal. The usual discrete-time Peskun ordering [19, 26] thus implies that $Z^{m,\sigma_1} \succeq Z^{m,\sigma_2}$. ∎

To continue, let

$$V_*(f,\sigma) \;:=\; \lim_{T\to\infty} T^{-1}\mathrm{Var}_\pi\left(\int_0^T f(X_s^\sigma)ds\right),$$

which we assume satisfies the usual relation

$$V_*(f,\sigma) \;=\; \int_{-\infty}^{\infty} \mathrm{Cov}_\pi(f(X_0^\sigma), f(X_s^\sigma))\,ds\,.$$

Also, let

$$V_m(f,\sigma) \;:=\; \lim_{n\to\infty} n^{-1}\mathrm{Var}_\pi\left(\sum_{i=1}^{mn} f(Z_i^{m,\sigma})\right),$$

which we assume satisfies the usual relation

$$V_m(f,\sigma) \;=\; \sum_{i=-\infty}^{\infty} \mathrm{Cov}_\pi(f(Z_0^{m,\sigma}), f(Z_i^{m,\sigma}))\,.$$

(In both cases, the subscript π indicates that the process is assumed to be in stationarity, all the way from time $-\infty$ to ∞.) We then have the following.

Lemma 4. *Let G_m be the spectral gap of the process $Z^{m,\sigma}$. Assume there is some constant $g > 0$ such that $G_m \geq g/m^2$ for all m. Then for all bounded functions $f : \mathbf{R} \to \mathbf{R}$, $\lim_{m\to\infty}(m^2 S/2)V_m(f,\sigma) = V_*(f,\sigma)$.*

Proof. Let

$$A_{m,t} \;=\; \mathbf{Cov}_\pi[f(Z_0^{m,\sigma}), f(Z_{\lfloor m^2 St/2\rfloor}^{m,\sigma})]\,.$$

And, let

$$A_{*,t} \;=\; \mathbf{Cov}_\pi[f(X_0^\sigma), f(X_t^\sigma)]\,.$$

Then

$$V_*(f,\sigma) \;=\; \int_\infty^\infty A_{*,t}\,dt\,,$$

and (since $\lfloor m^2 St/2\rfloor$ is a step-function of t, with steps of size $m^2 S/2$)

$$V_m(f,\sigma) \;=\; \frac{\int_{-\infty}^\infty A_{m,t}\,dt}{m^2 S/2}\,.$$

Now, by Lemma 2, since f is bounded,

$$\lim_{m\to\infty} A_{m,t} \;=\; A_{*,t}\,.$$

Now, let $v = \mathbf{Var}_\pi[f(X)]$, and write F for the forward operator corresponding to the chain $Z^{m,\sigma}$. Then using Lemma 2.3 of [13], since F is reversible, we have for all $m \in \mathbf{N}$ and $t \geq 0$ that

$$\begin{aligned}
A_{m,t} \;&=\; \mathbf{Cov}_\pi[f(Z_0^{m,\sigma}), f(Z_{\lfloor m^2 St/2\rfloor}^{m,\sigma})] \\
\leq\; &\sup\left\{\mathbf{Cov}_\pi[h(Z_0^{m,\sigma}), h(Z_{\lfloor m^2 St/2\rfloor}^{m,\sigma})] \;:\; h \in L^2(\pi),\; \mathbf{Var}_\pi[h(X)] = v\right\} \\
&=\; v\,\|F^{\lfloor m^2 St/2\rfloor}\| \;=\; v\,\|F\|^{\lfloor m^2 St/2\rfloor} \;=\; v\,(1-G_m)^{\lfloor m^2 St/2\rfloor} \\
&\leq\; v\,(1-g/m^2)^{\lfloor m^2 St/2\rfloor} \;\leq\; v\,(e^{-g/m^2})^{m^2 St/2} \;=\; v\,e^{-gSt/2}\,.
\end{aligned}$$

Hence,

$$V_m(f,\sigma) \;=\; \int_{-\infty}^\infty A_{m,t}\,dt \;\leq\; 2\int_0^\infty A_{m,t}\,dt \;\leq\; 4v/gS \;<\; \infty\,.$$

Hence, by the Dominated Convergence Theorem,

$$\lim_{m\to\infty}\int_{-\infty}^\infty A_{m,t}\,dt \;=\; \lim_{m\to\infty}\int_{-\infty}^\infty A_{*,t}\,dt\,,$$

i.e.

$$\lim_{m\to\infty}(m^2 S/2)\;V_m(f,\sigma) \;=\; V_*(f,\sigma)\,,$$

as claimed. ∎

To make use of Lemma 4, we need to bound the spectral gaps of the $Z^{m,\sigma}$ processes. We do this using a capacitance argument (see e.g. [24]). Let

$$\kappa_m = \inf_{\substack{A\subseteq\mathcal{X}_m \\ 0<\pi(A)\leq 1/2}} \frac{1}{\pi_m(A)} \sum_{x\in A} P_m(x, A^C)\pi_m(x)$$

be the capacitance of $Z^{m,\sigma}$. We prove:

Lemma 5. *The capacitance κ_m satisfies that*

$$\kappa_m \;\geq\; \min\Big(\frac{ke^{-L}r}{2m},\; \frac{Qke^{-2L/m}}{2m}\Big)\,,$$

where the quantities L and Q are defined in (2) and (4), respectively, and where the bound reduces to simply $\kappa_m \geq \frac{ke^{-L}r}{2m}$ in case A2(a).

Proof. We consider two different cases (only the second of which can occur in case A2(a)):

(i) $\exists a \in A$ with $|a| \leq q$. Then, since $\pi_m(A) \leq 1/2$, there is $j \in \mathbf{Z}$ with $|j/m| \leq q$ and $j/m \in A$ and either $(j+1)/m \in A^C$ or $(j-1)/m \in A^C$. Assume WOLOG that $(j+1)/m \in A^C$. We will need the following estimate on $\sum_{j\in\mathbf{Z}} \pi(j/m)$. For $x \in [i/m, (i+1)/m)$,

$$\pi(x) \geq (\pi(i/m)e^{-L(x-i/m)}$$

so that

$$\int_{i/m}^{(i+1)/m} \pi(x) \;\geq\; \pi(i/m) \int_0^{1/m} e^{-Lu}du \;=\; \pi(i/m)\left(\frac{1-e^{-L/m}}{L}\right)$$

$$=\; \pi(i/m)\, e^{-L/m}\left(\frac{e^{L/m}-1}{L}\right) \;\geq\; \pi(i/m)\, e^{-L/m}\left(\frac{L/m}{L}\right) \;=\; \frac{\pi(i/m)\, e^{-L/m}}{m}\,.$$

Therefore summing both sides over all $i \in \mathbf{Z}$,

$$1 \;=\; \int_{-\infty}^{\infty} \pi(x)\, dx \;\geq\; \frac{e^{-L/m}}{m} \sum_{i\in\mathbf{Z}} \pi(i/m)\,,$$

whence

$$\sum_{i\in\mathbf{Z}} \pi(i/m) \;\leq\; m\, e^{L/m}\,.$$

Then

$$\sum_{x\in A} P_m(x, A^C)\, \pi_m(x) \;\geq\; \pi_m(j/m)\, P_m(j/m, (j+1)/m)$$

$$=\; \pi_m(j/m)(1/2)\sigma^2(j/m)e^{-L/m} \;\geq\; (\pi(j/m)/m)(k/2)e^{-2L/m}$$

$$\geq Qke^{-2L/m}/2m\,.$$

(ii) $A \subseteq (-\infty, q) \cup (q, \infty)$. Let $a \in A$ with $\pi(a) = \max\{\pi(x) : x \in A\}$. Assume WOLOG that $a > 0$. Then

$$\sum_{x\in A} P_m(x, A^C)\pi_m(x) \geq \pi_m(a)P_m(a, a-(1/m))$$

$$\geq ke^{-L/m}\pi(a)/\sum_{\substack{i\\ |i/m|\geq a}} \pi(i/m) \geq ke^{-L/m}\pi(a)/[2\sum_{j=0}^{\infty} \pi(a)e^{-rj/m}]$$

$$= \frac{1}{2} k e^{-L/m} [1 - e^{-r/m}] \leq \frac{1}{2} k e^{-L} (r/m).$$

Thus, in either case, the conclusion of the lemma is satisfied. ∎

Now, it is known (e.g. [24]) that the spectral gap can be bounded in terms of the capacitance, specifically that $G_m \geq \kappa_m^2/2$. Thus, for $m \geq 1$,

$$G_m \ \geq \ [\min(\frac{1}{2} k e^{-L}(r/m), \ Q k e^{-2L/m}/2m)]^2/2$$

$$\geq \ [\min(\frac{1}{2} k e^{-L}(r/m), \ Q k e^{-2L}/2m)]^2/2 \ = \ g/m^2 \,,$$

where $g = [\min(\frac{1}{2} k e^{-L} r, \ Q k e^{-2L}/2)]^2/2 > 0$. This together with Lemma 2 shows that the conditions of Lemma 4 are satisfied. Hence, by Lemma 4, $\lim_{m\to\infty}(m^2 S/2) V_m(f,\sigma) = V_*(f,\sigma)$ for all bounded functions f.

On the other hand, by Lemma 3, $Z^{m,\sigma_1} \succeq Z^{m,\sigma_2}$, i.e. $V_m(f,\sigma_1) \leq V_m(f,\sigma_2)$. Hence, for all bounded functions f,

$$V_*(f,\sigma_1) \ = \ \lim_{m\to\infty} (m^2 S/2) V_m(f,\sigma_1) \ \leq \ \lim_{m\to\infty} (m^2 S/2) V_m(f,\sigma_2) \ = \ V_*(f,\sigma_2)\,. \tag{5}$$

Finally, if f is in L^2 but not bounded, then letting

$$f_m(x) \ = \ \begin{cases} m, & f(x) > m \\ f(x), & -m \leq f(x) \leq m \\ -m, & f(x) < -m \end{cases}$$

we have by the Monotone (or Dominated) Convergence Theorem that $V_*(f,\sigma_1) = \lim_{m\to\infty} V_*(f_m,\sigma_1)$ and $V_*(f,\sigma_2) = \lim_{m\to\infty} V_*(f_m,\sigma_2)$. Hence, it follows from (5) that $V_*(f,\sigma_1) \leq V_*(f,\sigma_2)$ for all $L^2(\pi)$ functions f. That is, $X^{\sigma_1} \succeq X^{\sigma_2}$, thus proving Theorem 1.

3 Simulated Tempering Diffusion Limit

We now apply our results to a version of the Simulated Tempering algorithm. Specifically, following [1], we consider a d-dimensional target density

$$f_d(x) \ = \ e^{dK} \prod_{i=1}^{d} f(x_i)\,, \tag{6}$$

for some unnormalised one-dimensional density function $f : \mathbf{R} \to [0,\infty)$, where $K = -\log(\int f(x)dx)$ is the corresponding normalising constant. (Although (6) is a very restrictive assumption, it is known [21, 23] that conclusions drawn from this special case are often approximately applicable in much broader contexts.) We consider simulated tempering in d dimensions, with inverse-temperatures chosen as follows: $\beta_0^{(d)} = 1$, and $\beta_{i+1}^{(d)} = \beta_i^{(d)} - \frac{\ell(\beta_i^{(d)})}{d^{1/2}}$ for some fixed C^1 function $\ell : [0,1] \to \mathbf{R}$. (The question then becomes, what is the optimal choice of ℓ.) As for when to stop adding new temperature values, we fix some $\chi \in (0,1)$ and keep going until the temperatures drop below χ, i.e. we stop at temperature $\beta_{k(d)}^{(d)}$ where $k(d) = \sup\{i : \beta_i^{(d)} \geq \chi\}$.

We shall consider a joint process $(y_n^{(d)}, X_n)$, with $X_n \in \mathbf{R}^d$, and with $y_n^{(d)} \in E_d := \{\beta_i^{(d)};\ 0 \leq i \leq k(d)\}$ defined as follows. If $y_{n-1} = \beta_i^{(d)}$ (where $0 < i < k(d)$), then chain proceeds by choosing $X_{n-1} \sim f^{\beta}$, then proposing Z_n to be β_{i+1} or β_{i-1} with probability $1/2$ each, and then accepting Z_n with the usual Metropolis acceptance probability. (If we propose to move to $\beta_{-1}^{(d)}$ or $\beta_{k(d)+1}^{(d)}$, then we automatically reject.) We assume [1] that the chain then immediately jumps to stationary at the new temperature, i.e. that mixing within a temperature is infinitely more efficient than mixing between temperatures.

The process $(y_n^{(d)}, X_n)$ is thus a Markov chain on the state space $E_d \times \mathbf{R}^d$, with joint stationary density given by

$$f_d(\beta, x) \ = \ e^{dK(\beta)} \prod_{i=1}^{d} f^{\beta}(x_i)\,,$$

where $K(\beta) = -\log \int f^{\beta}(x)dx$ is the normalising constant.

We now prove that the $\{y_n^{(d)}\}$ process has a diffusion limit (similar to random-walk Metropolis and Langevin algorithms, see [21, 22, 23]), and furthermore the asymptotic variance of the algorithm is minimised by choosing the function ℓ to leads to an asymptotic temperature acceptance rate $\doteq 0.234$. Specifically, we prove the following:

Theorem 6. *Under the above assumptions, the $\{y_n^{(d)}\}$ inverse-temperature process, when speeded up by a factor of d, converges in the Skorokhod topology as $d \to \infty$ to a diffusion limit $\{X_t\}_{t \geq 0}$ satisfying*

$$dX_t = \left[2\ell^2 \Phi\left(\frac{-\ell I^{1/2}}{2}\right)\right]^{1/2} dB_t + \left[\ell(X)\ell'(X)\Phi\left(\frac{-I^{1/2}\ell}{2}\right) - \ell^2\left(\frac{\ell I^{1/2}}{2}\right)'\phi\left(\frac{-I^{1/2}\ell}{2}\right)\right] dt\,, \tag{7}$$

for X_t in $(\chi, 1)$ with reflecting boundaries at both χ and 1. Furthermore, the speed of this diffusion is maximised, and the asymptotic variance of all L^2 functionals is minimised, when the function ℓ is chosen so that the asymptotic temperature acceptance rate is equal to 0.234 (to three decimal places).

Then, combining Theorem 6 and Theorem 1, we immediately obtain:

Theorem 7. *For the above simulated tempering algorithm, for any L^2 functional f, the choice of ℓ which minimises the limiting asymptotic variance $V_*(f) = \lim_{m\to\infty} V_m(f)$, is the same as the choice which maximises $\sigma(x)$, i.e. is the choice which leads to an asymptotic temperature acceptance probability of 0.234 (to three decimal places).*

Remark. In this context, it was proved in [1] that as $d \to \infty$, the choice of ℓ leading to an asymptotic temperature acceptance rate $\doteq 0.234$ maximises the expected squared jumping distance of the $\{y_n^{(d)}\}$ process. However, that left open the question of whether that choice would also minimise the asymptotic variance for any L^2 function. That question is resolved by Theorem 7.

3.1 Proof of Theorem 6

The key computation for proving Theorem 6 will be given next, but first we require some additional notation. We let $\mathrm{int}(E_d)$ denote $E_d \setminus \{1, \beta_{k(d)}^{(d)}\}$. We also denote by $G^{(d)}$ the generator of the inverse-temperature process $\{y_n^{(d)}\}$, and set H to be the set of all functions $h \in C^2[\chi, 1]$ with $h'(\chi) = h'(1) = 0$. We also let G^* be the generator of the diffusion given in (7), defined for all functions $h \in H$:

$$G^* h \;=\; \frac{\sigma^2(x) h''(x)}{2} + \mu(x) h'(x)\,, \quad h \in H\,, \tag{8}$$

where

$$\mu(x) \;=\; \ell(x)\ell'(x)\Phi\left(\frac{-I^{1/2}\ell}{2}\right) - \ell^2 \left(\frac{\ell I^{1/2}}{2}\right)' \phi\left(\frac{-I^{1/2}\ell}{2}\right)$$

and

$$\sigma^2(x) \;=\; 2\ell^2 \Phi\left(\frac{-\ell I^{1/2}}{2}\right)\,. \tag{9}$$

Note that $\{(h, G^* h);\ h \in H\}$ forms a core for the generator of the diffusion process described above in (7), see for example Chapter 8 of [8].

To proceed, we apply the powerful weak convergence theory of [8]. We do this using a technique for limiting reflecting processes similar to the arguments in Ward and Glynn [27]. Specifically, from Theorems 1.6.1 and 4.2.11 of [8], we must show that for any pair (h, G^*h) with $h \in H$, there exists a sequence $(h_d, dG^{(d)}h_d)_{d\in\mathbf{N}}$ such that

$$\lim_{d\to\infty} \sup_{x\in E_d} |h(x) - h_d(x)| \;=\; 0 \tag{10}$$

and

$$\lim_{d\to\infty} \sup_{x\in E_d} |G^* h(x) - dG^{(d)} h_d(x)| \;=\; 0\,. \tag{11}$$

To establish this convergence on $int(E_d)$, we can simply let $h_d = h$ (see Lemma 8 below). However, to establish the convergence on the boundary of E_d (Lemma 9), we need to modify h slightly (without destroying the convergence on $int(E_d)$). We do this as follows. First, given any $h \in H$, we let

$$\overline{h}_d(x) \;=\; h\left(\gamma_d(x)\right) \,,$$

where

$$\gamma_d(x) \;=\; \frac{(1-\chi)x + \chi - \chi_d}{1-\chi_d} \,,$$

so that $\overline{h}_d$ is just like h except "stretched" to be defined on $[\chi_d, 1]$ instead of just on $[\chi, 1]$. Here we set $\chi_d = \beta^{(d)}_{k(d)}$, and $\chi_d^+ = \beta^{(d)}_{k(d)-1}$; thus $\chi_d \le \chi \le \chi_d^+$. Notice that since $\chi_d \to \chi$ as $d \to \infty$, $\overline{h}_d$ and its first and second derivatives converge to h and its corresponding derivatives uniformly for $x \in [\chi_d, 1]$ as $d \to \infty$.

Finally, given the function h, we let $\eta(x)$ to be any smooth function: $[\chi, 1] \to \mathbf{R}$ satisfying

$$\eta'(\chi) = h''(\chi)\ell(\chi)/2 \text{ and } \eta'(1) = h''(1)\ell(1)/2 \,,$$

and then set

$$h_d(x) \;=\; \overline{h}_d(x) + d^{-1/2}\eta(\gamma_d(x)) \;=\; h(\gamma_d(x)) + d^{-1/2}\eta(\gamma_d(x)) \,,$$

so that $h_d(x)$ is similar to $\overline{h}_d(x)$ except with the addition of a separate $O(d^{-1/2})$ term (which will only be relevant at the boundary points, i.e. in Lemma 9 below). In particular, (10) certainly holds.

In light of the above discussion, Theorem 6 will follow by establishing (11), which is done in Lemmas 8 and 9 below.

Lemma 8. *For all $h \in H$,*

$$\lim_{d\to\infty} \sup_{x \in int(E_d)} |dG^{(d)}h(x) - G^* h(x)| = 0 \,, \tag{12}$$

and

$$\lim_{d\to\infty} \sup_{x \in int(E_d)} |dG^{(d)}h_d(x) - G^* h(x)| = 0 \,. \tag{13}$$

Proof. We begin with a Taylor series expansion for $G^{(d)}$. Since the computations shall get somewhat messy, we wish to keep only higher-order terms, so for simplicity we shall use the notation $\overset{r(d)}{\approx}$ to mean that the expansion holds up to terms of order $1/r(d)$, uniformly for $x \in E_d$, as $d \to \infty$. (For example, $LHS \overset{d}{\approx} RHS$ means that $\lim_{d\to\infty} \sup_{x\in E_d} d(LHS -$

$RHS) = 0$.) Then for bounded C^2 functionals h, we have (combining the two h'' terms together) that for $\beta_i^{(d)} \in int(E_d)$:

$$G^{(d)}h(\beta_i^{(d)}) \overset{d}{\approx} \frac{h'(\beta_i^{(d)})}{2}[\alpha^+(\beta_{i+1}^{(d)} - \beta_i^{(d)}) + \alpha^-(\beta_{i-1}^{(d)} - \beta_i^{(d)})] + \frac{h''(\beta_i^{(d)})}{2}\Big[(\beta_{i+1}^{(d))} - \beta_i^{(d)})^2\alpha^+\Big]$$

$$\overset{d}{\approx} \frac{h'(\beta_i^{(d)})}{2}[\alpha^+(\beta_{i+1}^{(d)} - \beta_i^{(d)}) + \alpha^-(\beta_{i-1}^{(d)} - \beta_i^{(d)})] + \frac{h''(\beta_i^{(d)})}{2}\Big[(\beta_{i+1}^{(d))} - \beta_i^{(d)})^2\alpha^+\Big]$$

$$= \frac{h'(\beta_i^{(d)})}{2}\frac{\alpha^-\ell(\beta_{i-1}^{(d)}) - \alpha^+\ell(\beta_i^{(d)})}{d^{1/2}} + \frac{h''(\beta_i^{(d)})}{2}\left[\frac{\ell(\beta_i^{(d)})^2\alpha^+}{d}\right],$$

where α^+ is the probability of accepting an upwards move, and α^- is the probability of accepting a downwards move.

To continue, we let $g = \log f$, and

$$M(\beta) \;=\; \mathbf{E}^\beta(g) \;=\; \frac{\int \log f(x) f^\beta(x)dx}{\int f^\beta(x)dx},$$

and

$$I(\beta) \;=\; \mathbf{Var}^\beta(g) \;=\; \frac{\int (\log f(x))^2 f^\beta(x)dx}{\int f^\beta(x)dx} - M(\beta)^2 .$$

It follows [1] that $M'(\beta) = I(\beta)$ and $K'(\beta) = -M(\beta)$, so $K''(\beta) = -M'(\beta) = -I(\beta)$. We also define $\overline{g} = g - M(\beta)$.

For shorthand, we write $\beta = \beta_i^{(d)}$, and $\ell = \ell(\beta_i^{(d)})$, and $\underline{\ell} = \ell(\beta_{i-1}^{(d)})$, and $\underline{\epsilon} = \beta_{i-1}^{(d)} - \beta_i^{(d)} = \underline{\ell}/d^{1/2}$, and $\epsilon = \beta_i^{(d)} - \beta_{i+1}^{(d)} = \ell/d^{1/2}$, and $I = I(\beta)$ and $K'' = K''(\beta)$ and $K''' = K'''(\beta)$.

Then, with $X \sim f^\beta$,

$$\alpha^- \;=\; \mathbf{E}\Big[1 \wedge \frac{f_d^{\beta+\underline{\epsilon}}(X)e^{dK(\beta+\underline{\epsilon})}}{f_d^\beta(X)e^{dK(\beta)}}\Big]$$

$$= \mathbf{E}\Big[1 \wedge \exp\Big((K(\beta+\underline{\epsilon}) - K(\beta))d + \underline{\epsilon} d M(\beta) + \underline{\epsilon}\sum_{i=1}^d \overline{g}(X_i)\Big)\Big]$$

$$\overset{d^{1/2}}{\approx} \mathbf{E}\Big[1 \wedge \exp\Big(\frac{d\underline{\epsilon}^2}{2}K'' + \frac{d\underline{\epsilon}^3}{6}K''' + N(0,\; I\underline{\epsilon}^2 d)\Big)\Big]$$

$$= \mathbf{E}\Big[1 \wedge \exp\Big(\frac{\underline{\ell}^2}{2}K'' + \frac{\underline{\epsilon}\underline{\ell}^2}{6}K''' + N(0,\; I\underline{\ell}^2)\Big)\Big]$$

$$= \Phi\Big(-\frac{I^{1/2}\underline{\ell}}{2} + \frac{\underline{\epsilon}\underline{\ell}K'''}{6I^{1/2}}\Big)$$

$$+\exp(\underline{\epsilon}\underline{\ell}^2K'''/6)\,\Phi\Big(-\frac{I^{1/2}\underline{\ell}}{2} - \frac{\underline{\epsilon}\underline{\ell}K'''}{6I^{1/2}}\Big). \tag{14}$$

Similarly,

$$\alpha^+ \;=\; \mathbf{E}\Big[1 \wedge \frac{f_d^{\beta-\epsilon}(X)e^{dK(\beta-\epsilon)}}{f_d^\beta(X)e^{dK(\beta)}}\Big]$$

$$
\begin{aligned}
&= \mathbf{E}\Big[1 \wedge \exp\Big((K(\beta-\epsilon) - K(\beta))d - \epsilon d M(\beta) - \epsilon \sum_{i=1}^{d} \overline{g}(X_i)\Big)\Big] \\
&\overset{1}{\approx} \mathbf{E}\Big[1 \wedge \exp\Big(\frac{d\epsilon^2}{2} K'' - N(0,\ I\epsilon^2 d)\Big)\Big] \\
&= \mathbf{E}\Big[1 \wedge \exp\Big(\frac{\ell^2}{2} I - \frac{\epsilon \ell^2}{6} K''' - N(0,\ I\ell^2)\Big)\Big] \\
&= \Phi\Big(-\frac{I^{1/2}\ell}{2} - \frac{\epsilon \ell K'''}{6 I^{1/2}}\Big) \\
&\quad + \exp(-\epsilon \ell^2 K'''/6)\, \Phi\Big(-\frac{I^{1/2}\ell}{2} - \frac{\epsilon \ell K'''}{6 I^{1/2}}\Big)\,.
\end{aligned}
$$

So

$$
\begin{aligned}
\alpha^+(\beta_i^{(d)}) &\overset{d^{1/2}}{\approx} \Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2} - \frac{\epsilon \ell K'''(\beta_i^{(d)})}{6 I^{1/2}(\beta_i^{(d)})}\Big) \\
&+ \exp(-\epsilon \ell^2(\beta_i^{(d)}) K'''(\beta_i)/6) \Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2} + \frac{\epsilon \ell K'''(\beta_i^{(d)})}{6 I^{1/2}(\beta_i^{(d)})}\Big)\,.
\end{aligned}
$$

To first order this expression can be approximated as

$$
\alpha^+(\beta_i^{(d)}) \overset{1}{\approx} 2\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\,.
$$

Next, we note that in the current setting, β is itself marginally a Markov chain with uniform stationary distribution among all temperatures. In fact it is a birth and death process, and hence reversible. So, by detailed balance,

$$
\alpha^- = \alpha^+(\beta_i^{(d)} - \ell/\sqrt{d})\,.
$$

Therefore

$$
\begin{aligned}
\alpha^-(\beta_i^{(d)}) &= \alpha^+(\beta_i^{(d)} - \ell/\sqrt{d}) \\
&\overset{d^{1/2}}{\approx} \alpha^+(\beta_i^{(d)}) - \frac{(\ell(\beta_i^{(d)}) I^{1/2}(\beta_i^{(d)}))'}{2} \left(\frac{-\ell}{\sqrt{d}}\right) \phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2} - \frac{\epsilon \ell K'''(\beta_i^{(d)})}{6 I^{1/2}(\beta_i^{(d)})}\Big) \\
&- \exp(-\epsilon \ell^2(\beta_i^{(d)}) K'''(\beta_i)/6) \frac{(\ell(\beta_i^{(d)}) I^{1/2}(\beta_i^{(d)}))'}{2} \left(\frac{-\ell}{\sqrt{d}}\right) \phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2} + \frac{\epsilon \ell K'''(\beta_i^{(d)})}{6 I^{1/2}(\beta_i^{(d)})}\Big)
\end{aligned}
$$

Then, since $\underline{\ell} \overset{d^{1/2}}{\approx} \ell + \underline{\epsilon}\ell' \overset{d^{1/2}}{\approx} \ell + \epsilon \ell' = \ell + \frac{\ell\ell'}{d^{1/2}}$, we compute that

$$
\begin{aligned}
\mu(\beta_i^{(d)}) &\overset{d^{1/2}}{\approx} \frac{1}{2d^{1/2}}\Big[-\alpha^+\ell + \left(\ell + \frac{\ell\ell'}{d^{1/2}}\right) \times \\
&\Big(\alpha^+(\beta_i^{(d)}) - \frac{(\ell(\beta_i^{(d)}) I^{1/2}(\beta_i^{(d)}))'}{2} \left(\frac{-\ell}{\sqrt{d}}\right) \phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2} - \frac{\epsilon \ell K'''(\beta_i^{(d)})}{6 I^{1/2}(\beta_i^{(d)})}\Big) -
\end{aligned}
$$

$$\exp(-\epsilon\ell^2(\beta_i^{(d)})K'''(\beta_i)/6)\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2}\left(\frac{-\ell}{\sqrt{d}}\right)\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}+\frac{\epsilon\ell K'''(\beta_i^{(d)})}{6I^{1/2}(\beta_i^{(d)})}\Big)\Big)\Big].$$

Hence, ignoring all lower order terms,

$$\mu(\beta_i^{(d)}) \overset{d^{1/2}}{\approx} \frac{1}{2d^{1/2}}\Big[-\ell\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2}\left(\frac{-\ell}{\sqrt{d}}\right)\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}-\frac{\epsilon\ell K'''(\beta_i^{(d)})}{6I^{1/2}(\beta_i^{(d)})}\Big)$$

$$-\ell\exp\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2}\left(\frac{-\ell}{\sqrt{d}}\right)\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}+\frac{\epsilon\ell K'''(\beta_i^{(d)})}{6I^{1/2}(\beta_i^{(d)})}\Big)+\frac{2\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\ell\ell'}{d^{1/2}}\Big]$$

$$\overset{d^{1/2}}{\approx} \frac{1}{d}\Big[-\ell^2\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2}\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)+\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\ell\ell'\Big].$$

Similarly $\sigma^2(\beta_i^{(d)})$ is to first order

$$\frac{2\ell^2}{d}\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)$$

so that we can write (for $0 < \beta < 1$)

$$G^d h \overset{d}{\approx} \frac{1}{d}\Big(\ell^2\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)h''(\beta)$$

$$+\Big[\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\ell\ell'-\ell^2\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2}\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\Big]h'(\beta)\Big).$$

However this expression is just $d^{-1}G^*h$, thus establishing (12).

Finally, to establish (13), we note that in this case the terms $d^{-1/2}\eta(\gamma_d(x))$ and $\overline{h}_d(x) - h(x)$ are both lower-order and do not affect the limit. Hence, (13) follows directly from (12). ∎

The uniformity over $int(E_d)$ for h (as opposed to h_d) in the proof of Lemma 8 does not extend to the boundary of E_d. (If it did, then the proof of Theorem 6 would be complete simply by setting $h_d = h$ and applying Lemma 8.) However the following lemma shows that with the definition of h_d that we have used, the extension to the boundary does hold.

Lemma 9. *For all $h \in H$, for $x = 1$ and for $x = \chi_d$,*

$$\lim_{d\to\infty} |dG^{(d)}h_d(x) - G^*h(x)| = 0.$$

Proof. We prove the case when $x = \chi_d$; the case $x = 1$ is similar but somewhat easier (since then x does not depend on d).

Mimicking the Taylor expansion of Lemma 8,

$$G^{(d)}h_d(\chi_d) \overset{d}{\approx} \frac{h'_d(\chi_d)\alpha^-(\chi_d^+ - \chi_d)]}{2} + \frac{h''_d(\chi_d)}{4}\left[(\chi_d - \chi_d^+)^2\alpha^-\right]$$

$$= \frac{h_d'(\chi_d)}{2}\frac{\alpha^-\ell(\chi_d^+)}{d^{1/2}} + \frac{h_d''(\chi_d)}{4}\left[\frac{\ell(\chi_d)^2\alpha^-}{d}\right]$$

$$\overset{d}{\approx} \frac{\alpha^-\ell(\chi_d^+)}{2d^{1/2}}\left(h'(\chi) + \eta'(\chi)d^{-1/2}\right) + \frac{h_d''(\chi_d)}{4}\left[\frac{\ell(\chi_d)^2\alpha^-}{d}\right] .$$

Thus since $h'(\chi) = 0$, this expression equals

$$\frac{h_d''(\chi_d)}{2}\left[\frac{\ell(\chi_d)^2\alpha^-}{d}\right] .$$

Next we note from (14) that

$$\alpha^- \overset{1}{\approx} 2\,\Phi\left(-\frac{I^{1/2}\ell}{2}\right) .$$

Hence, the above results show that

$$\lim_{d\to\infty} dG_d h_d(\chi_d) = \ell^2(\chi)h''(\chi)\,\Phi\left(-\frac{I^{1/2}\ell}{2}\right) .$$

In light of the formulae (8) and (9), this completes the proof. ∎

Finally, we provide the missing proof from the previous section.

Proof of Lemma 2. We first compute that to first order as $h \searrow 0$ and $m \to \infty$, writing $x = i/m$ and $e = 1/m$, we have

$$\mathbf{E}\Big(Y^\sigma_{m,t+h} - Y^\sigma_{m,t} \,\Big|\, Y^\sigma_{m,t} = \frac{i}{m}\Big)$$

$$\approx (\frac{m^2Sh}{2})(\frac{1}{m})(\frac{1}{2S}) \times \left[\sigma^2(\frac{i}{m}) + \frac{\pi(\frac{i+1}{m})\,\sigma^2(\frac{i+1}{m})}{\pi(\frac{i}{m})} - \sigma^2(\frac{i}{m}) - \frac{\pi(\frac{i-1}{m})\,\sigma^2(\frac{i-1}{m})}{\pi(\frac{i}{m})}\right]$$

$$= \frac{hm}{4}\left[\frac{\pi(x+e)\,\sigma^2(x+e)}{\pi(x)} - \frac{\pi(x-e)\,\sigma^2(x-e)}{\pi(x)}\right]$$

$$\approx \frac{hm}{4}\left[\frac{(\pi(x)+e\pi'(x))(\sigma^2(x)+e(\sigma^2)'(x)) - (\pi(x)-e\pi'(x))(\sigma^2(x)-e(\sigma^2)'(x))}{\pi(x)}\right]$$

$$= \frac{hm}{4}\left[\frac{2e\pi'(x)\,\sigma^2(x) + 2e\pi(x)\,(\sigma^2)'(x)}{\pi(x)}\right]$$

$$= \frac{hm}{4}(2e)\left[(\log\pi)'(x)\,\sigma^2(x) + 2\sigma(x)\,\sigma'(x)\right]$$

$$= h\left[\frac{1}{2}\,(\log\pi)'(x)\,\sigma^2(x) + \sigma(x)\,\sigma'(x)\right] ,$$

and also

$$\mathbf{E}((Y^\sigma_{m,t+h} - Y^\sigma_{m,t})^2 \mid Y^\sigma_{m,t} = \frac{i}{m}) \;\approx\; (\frac{m^2 Sh}{2})(\frac{1}{2S})(\frac{1}{m^2})[2\sigma^2(x) + 2\sigma^2(x)] \;=\; h\,[\sigma^2(x)]\,.$$

A comparison with (1) then shows that Y^σ_m satisfies the same first and second moment characteristics as X^σ_t, so that X^σ_t is indeed the correct putative limit.

In light of these calculations, the formal proof of this lemma then proceeds along standard lines. Indeed, case (a) is just a simpler version of the proof of Theorem 6 above. And, case (b) follows from standard arguments about using the uniform convergence of generators (e.g. [8], Chapter 8) to establish the approximation of birth and death processes by diffusions; see for example Theorem 4.1 of Chapter 5 on page 387 of [6]. ∎

References

[1] Y. Atchadé, G.O. Roberts, and J.S. Rosenthal (2011), Towards Optimal Scaling of Metropolis-Coupled Markov Chain Monte Carlo. Stat. and Comput. **21(4)**, 555–568.

[2] M. Bédard (2007), Weak convergence of Metropolis algorithms for non-iid target distributions. Ann. Appl. Prob. **17**, 1222–1244.

[3] M. Bédard (2008). Optimal acceptance rates for Metropolis algorithms: moving beyond 0.234. Stoch. Proc. Appl. **118**, 2198–2222.

[4] M. Bédard and J.S. Rosenthal (2008). Optimal scaling of Metropolis algorithms: heading toward general target distributions. Can. J. Stat. **36**, 483–503.

[5] A. Beskos, G.O. Roberts, and A.M. Stuart (2009), Optimal scalings of Metropolis-Hastings algorithms for non-product targets in high dimensions. Ann. Appl. Prob. **19**, 863–898.

[6] R.N. Bhattacharya and E.C. Waymire (1990), Stochastic processes with applications. Wiley & Sons, New York.

[7] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds. (2011), Handbook of Markov chain Monte Carlo. Chapman & Hall, New York.

[8] S.N. Ethier and T. G. Kurtz (1986), Markov processes: characterization and convergence. Wiley, Chichester, New York, Toronto.

[9] G. Fort, E. Moulines, and P. Priouret (2011), Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. Ann. Stat. **39(6)**, 3262-3289.

[10] C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci., Vol. **7**, No. **4**, 473-483.

[11] D.A. Kofke (2002), On the acceptance probability of replica-exchange Monte Carlo trials. J. Chem. Phys. **117**, 6911. Erratum: J. Chem. Phys. **120**, 10852.

[12] F. Leisen and A. Mira (2008), An extension of Peskun and Tierney orderings to continuous time Markov chains. Statistica Sinica **18**, 1641–1651.

[13] J.S. Liu, W. Wong, and A. Kong (1994), Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika **81**, 27-40.

[14] E. Marinari and G. Parisi (1992), Simulated tempering: a new Monte Carlo scheme. Europhys. Lett. **19**, 451–458.

[15] S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London. Available at probability.ca/MT.

[16] A. Mira (2001), Ordering and improving the performance of Monte Carlo Markov chains. Stat. Sci. **16**, 340–350.

[17] A. Mira and C.J. Geyer (2000), On non-reversible Markov chains. In N. Madras (ed.), Fields Institute Communications, Volume **26**: Monte Carlo Methods, pp. 93–108. Providence, RI: American Mathematical Society.

[18] A. Mira and F. Leisen (2009), Covariance ordering for discrete and continuous time Markov chains. Statistica Sinica **19**, 651–666.

[19] P.H. Peskun (1973), Optimum Monte Carlo sampling using Markov chains. Biometrika **60**, 607–612.

[20] C. Predescu, M. Predescu, and C.V. Ciobanu (2004), The incomplete beta function law for parallel tempering sampling of classical canonical systems. J. Chem. Phys. **120**, 4119.

[21] G.O. Roberts, A. Gelman, and W.R. Gilks (1997), Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Prob. **7**, 110–120.

[22] G.O. Roberts and J.S. Rosenthal (1998), Optimal scaling of discrete approximations to Langevin diffusions. J. Roy. Stat. Soc. B **60**, 255–268.

[23] G.O. Roberts and J.S. Rosenthal (2001), Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. **16**, 351–367.

[24] A. Sinclair (1992), Improved bounds for mixing rates of Markov chains and multicommodity flow. Combinatorics, Prob., Comput. **1**, 351–370.

[25] L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.

[26] L. Tierney (1998), A note on Metropolis-Hastings kernels for general state spaces. Ann. Appl. Prob. **8**, 1–9.

[27] A. Ward and P.W. Glynn (2003), A diffusion approximation for a Markovian queue with reneging. Queueing Systems: Theory and Applications **43**, 103–128.