The Annals of Applied Probability
2014, Vol. 24, No. 1, 131–149
DOI: 10.1214/12-AAP918

MINIMISING MCMC VARIANCE VIA DIFFUSION LIMITS, WITH AN APPLICATION TO SIMULATED TEMPERING

BY GARETH O. ROBERTS AND JEFFREY S. ROSENTHAL[1]

University of Warwick and University of Toronto

We derive new results comparing the asymptotic variance of diffusions by writing them as appropriate limits of discrete-time birth–death chains which themselves satisfy 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, for example, [7, 24]. 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 $\operatorname{Var}_\pi(h,P):=\lim_{n\to\infty}\frac{1}{n}\operatorname{Var}_\pi(\sum_{i=1}^{n}h(X_i))$ (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, that is, how to choose the Markov chain optimally; see, for example, [10, 15]. 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 $\operatorname{Var}_\pi(h,P_1)\le\operatorname{Var}_\pi(h,P_2)$ for all $L^2(\pi)$ functionals $h:\mathcal{X}\to\mathbf{R}$, that is, 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 [18] for finite state spaces, and by Tierney [25] for general state spaces (see also [15, 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 $P_1(x,A)\ge P_2(x,A)$ for all $x\in\mathcal{X}$ and $A\in\mathcal{F}$ with $x\notin A$, that is, that P_1 *dominates* P_2 *off the diagonal*.

Received March 2012; revised December 2012.

[1]Supported in part by NSERC of Canada.

MSC2010 subject classifications. Primary 60J22; secondary 62M05, 62F10.

Key words and phrases. Markov chain Monte Carlo, simulated tempering, optimal scaling, diffusion limits.

Meanwhile, diffusion limits have become a common way to establish asymptotic comparisons of MCMC algorithms [2–5, 20–22]. 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 [20], and Langevin algorithms [21]), this is easy since one diffusion is simply a time-change of the other. But more general diffusion comparisons are less clear; for example, the processes' spectral gaps

$$1 - \sup\left\{\int h(y)P(x, dy) : \int h(y)\pi(dy) = 0, \int h^2(y)\pi(dy) = 1\right\}$$

can be ordered directly by using Dirichlet forms, but this does not lead to bounds on 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 $P_1 \succeq P_2$, that is, P_1 is more efficient than P_2. (We note that Mira and Leisen [12, 17] 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 [10, 14], 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, 19], that a particular choice of temperatures (which leads to an asymptotic temperature-swap acceptance rate of 0.234) maximises the asymptotic L^2 jumping distance, that is, $\lim_{n\to\infty} \mathbf{E}(|X_n - X_{n-1}|^2)$. (Indeed, this result has already influenced adaptive MCMC algorithms for simulated tempering; see, for example, [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 nonexplosive Langevin diffusions X^σ on $\mathcal{X}$ with stationary density π, satisfying

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

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 (similarly 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 $f : \mathcal{X} \to \mathbf{R}$,

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

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" than 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 [20–22]), since then $X_t^{\sigma_1}$ has the same distribution as $X_{ct}^{\sigma_2}$ where $c = \sigma_1/\sigma_2 > 1$; that is, 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 nonconstant functions, then the 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 $\mathcal{X}$; that is, there is $L < \infty$ with

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

(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; that is, 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 $P(i/m, (i+1)/m) + P(i/m, (i-1)/m) \le 1$].

In terms of $Z^{m,\sigma}$, we then let $\{Y^\sigma_{m,t}\}_{t\ge 0}$ be the continuous-time version of $Z^{m,\sigma}$, speeded up by a factor of $m^2S/2$, that is, defined by $Y^\sigma_{m,t} = Z^{m,\sigma}_{\lfloor m^2St/2\rfloor}$ for $t \ge 0$. (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^σ_m 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 $\sigma_1(x) \ge \sigma_2(x)$ for all $x \in \mathbf{R}$. Then $Z^{m,\sigma_1} \succeq Z^{m,\sigma_2}$.*

PROOF. By inspection, the fact that $\sigma_1(x) \ge \sigma_2(x)$ implies that

$$\mathbf{P}\big(Z^{m,\sigma_1}_{(i+1)/m} = j+1 \mid Z^{m,\sigma_1}_{i/m} = j\big) \ge \mathbf{P}\big(Z^{m,\sigma_2}_{(i+1)/m} = j+1 \mid Z^{m,\sigma_2}_{i/m} = j\big)$$

and

$$\mathbf{P}\big(Z^{m,\sigma_1}_{(i+1)/m} = j-1 \mid Z^{m,\sigma_1}_{i/m} = j\big) \ge \mathbf{P}\big(Z^{m,\sigma_2}_{(i+1)/m} = j-1 \mid Z^{m,\sigma_2}_{i/m} = j\big).$$

It follows that Z^{m,σ_1} dominates Z^{m,σ_2} off the diagonal. The usual discrete-time Peskun ordering [18, 25] thus implies that $Z^{m,\sigma_1} \succeq Z^{m,\sigma_2}$. $\square$

To continue, let

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

which we assume satisfies the usual relation

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

Also, let

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

which we assume satisfies the usual relation

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

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

PROOF. Let

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

and let

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

Then

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

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

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

Now, by Lemma 2, since f is bounded,

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

To continue, let F be the forward operator corresponding to the chain $Z^{m,\sigma}$, that is, $Fh(x) = \mathbf{E}[h(Z_1^{m,\sigma}) \mid Z_0^{m,\sigma} = x]$. Then since F is reversible, it follows from Lemma 2.3 of [13] that

$$\|F^t\| = \|F\|^t = \sup\{\operatorname{Cov}_\pi\big[h_1(Z_0^{m,\sigma}), h_2(Z_t^{m,\sigma})\big] : \operatorname{Var}_\pi(h_1) = \operatorname{Var}_\pi(h_2) = 1\}.$$

Letting $v = \mathrm{Var}_\pi[f(X)]$, we then have, for all $m \in \mathbf{N}$ and $t \ge 0$, that

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

Hence,

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

Hence, by the dominated convergence theorem,

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

that is,

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

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, for example, [23]. Let

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

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

LEMMA 5. *The capacitance κ_m satisfies that*

$$\kappa_m \ge \min\biggl(\frac{ke^{-L}r}{2m}, \frac{Qke^{-2L/m}}{2m}\biggr),$$

where the quantities L and Q are defined in (2) *and* (4), *respectively, and where the bound reduces to simply $\kappa_m \ge \frac{ke^{-L}r}{2m}$ 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 $|a| \le q$. Then, since $\pi_m(A) \le 1/2$, there is $j \in \mathbf{Z}$ with $|j/m| \le q$ 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

$$\begin{aligned}
\int_{i/m}^{(i+1)/m} \pi(x) &\geq \pi(i/m)\int_0^{1/m} e^{-Lu}\,du = \pi(i/m)\left(\frac{1-e^{-L/m}}{L}\right) \\
&= \pi(i/m)e^{-L/m}\left(\frac{e^{L/m}-1}{L}\right) \geq \pi(i/m)e^{-L/m}\left(\frac{L/m}{L}\right) \\
&= \frac{\pi(i/m)e^{-L/m}}{m}.
\end{aligned}$$

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

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

Then

$$\begin{aligned}
\sum_{x\in A} P_m(x, A^C)\pi_m(x) &\geq \pi_m(j/m)P_m(j/m, (j+1)/m) \\
&= \pi_m(j/m)(1/2)\sigma^2(j/m)e^{-L/m} \\
&\geq (\pi(j/m)/m)(k/2)e^{-2L/m} \\
&\geq Qke^{-2L/m}/2m.
\end{aligned}$$

(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

$$\begin{aligned}
\sum_{x\in A} P_m(x, A^C)\pi_m(x) &\geq \pi_m(a)P_m(a, a-(1/m)) \\
&\geq ke^{-L/m}\pi(a)\Big/ \sum_{\substack{i \\ |i/m|\geq a}} \pi(i/m) \\
&\geq ke^{-L/m}\pi(a)\Big/\left[2\sum_{j=0}^{\infty}\pi(a)e^{-rj/m}\right] \\
&= \frac{1}{2}ke^{-L/m}[1-e^{-r/m}] \leq \frac{1}{2}ke^{-L}(r/m).
\end{aligned}$$

Thus, in either case, the conclusion of the lemma is satisfied. $\square$

Now, it is known (e.g., [23]) that the spectral gap can be bounded in terms of the capacitance, specifically that $G_m \geq \kappa_m^2/2$. Thus, for $m \geq 1$,

$$\begin{aligned} G_m &\geq \bigl[\min\bigl(\tfrac{1}{2}ke^{-L}(r/m), Qke^{-2L/m}/2m\bigr)\bigr]^2/2 \\ &\geq \bigl[\min\bigl(\tfrac{1}{2}ke^{-L}(r/m), Qke^{-2L}/2m\bigr)\bigr]^2/2 \\ &= g/m^2, \end{aligned}$$

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

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

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

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

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

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 [2–5, 20, 22] 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; that is, 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 the 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 finally accepting Z_n with the usual Metropolis acceptance probability. (A proposed move to $\beta_{-1}^{(d)}$ or $\beta_{k(d)+1}^{(d)}$ is automatically rejected.) We assume, as in [1], that the chain then immediately jumps to stationary at the new temperature, that is, 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 [20–22]), and furthermore the asymptotic variance of the algorithm is minimised by choosing the function ℓ that 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*

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

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 Theorems 1 and 6, 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 $\sigma(x)$, that is, 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, the question of whether that choice would also minimise the asymptotic variance for any L^2 function was left open. 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 $h \in H$, by

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

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

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

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 [26]. We first note that by page 17 and Chapter 8 of [8], the set $\{(h, G^*h); h \in H\}$ forms a core for the generator of the diffusion process described above in (7) (i.e., the closure of the restriction of the generator to that set is again equal to the generator itself). Hence, by Theorems 1.6.1 and 4.2.11 of [8], we need to 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

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

and

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

To establish this convergence on $\operatorname{int}(E_d)$, 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 $\operatorname{int}(E_d)$]. We do this as follows. First, given any $h \in H$, we let

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

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

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

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 $h \in H$,*

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

and

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

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 $\stackrel{r(d)}{\approx}$ to mean that the expansion holds up to terms of order $1/r(d)$, uniformly for $x\in E_d$, as $d\to\infty$ [e.g., $LHS\stackrel{d}{\approx}RHS$ 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 $\beta_i^{(d)}\in\mathrm{int}(E_d)$:

$$\begin{aligned}
G^{(d)}h(\beta_i^{(d)}) &\stackrel{d}{\approx} \frac{h'(\beta_i^{(d)})}{2}\left[\alpha^+(\beta_{i+1}^{(d)}-\beta_i^{(d)})+\alpha^-(\beta_{i-1}^{(d)}-\beta_i^{(d)})\right]\\
&\quad+\frac{h''(\beta_i^{(d)})}{2}\left[(\beta_{i+1}^{(d)}-\beta_i^{(d)})^2\alpha^+\right]\\
&\stackrel{d}{\approx} \frac{h'(\beta_i^{(d)})}{2}\left[\alpha^+(\beta_{i+1}^{(d)}-\beta_i^{(d)})+\alpha^-(\beta_{i-1}^{(d)}-\beta_i^{(d)})\right]\\
&\quad+\frac{h''(\beta_i^{(d)})}{2}\left[(\beta_{i+1}^{(d)}-\beta_i^{(d)})^2\alpha^+\right]\\
&= \frac{h'(\beta_i^{(d)})}{2}\,\frac{\alpha^-\ell(\beta_{i-1}^{(d)})-\alpha^+\ell(\beta_i^{(d)})}{d^{1/2}}\\
&\quad+\frac{h''(\beta_i^{(d)})}{2}\left[\frac{\ell(\beta_i^{(d)})^2\alpha^+}{d}\right],
\end{aligned}$$

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

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

It follows, as in [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 $\underline{\varepsilon}=\beta_{i-1}^{(d)}-\beta_i^{(d)}=\underline{\ell}/d^{1/2}$, and $\varepsilon=\beta_i^{(d)}-\beta_{i+1}^{(d)}=\ell/d^{1/2}$, and $I=I(\beta)$ and $K''=K''(\beta)$ and $K'''=K'''(\beta)$.

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

$$
\begin{aligned}
\alpha^- &= \mathbf{E}\left[1 \wedge \frac{f_d^{\beta+\underline{\varepsilon}}(X)e^{dK(\beta+\underline{\varepsilon})}}{f_d^{\beta}(X)e^{dK(\beta)}}\right] \\
&= \mathbf{E}\left[1 \wedge \exp\left(\big(K(\beta+\underline{\varepsilon}) - K(\beta)\big)d + \underline{\varepsilon} dM(\beta) + \underline{\varepsilon}\sum_{i=1}^{d} \overline{g}(X_i)\right)\right] \\
&\overset{d^{1/2}}{\approx} \mathbf{E}\left[1 \wedge \exp\left(\frac{d\underline{\varepsilon}^2}{2}K'' + \frac{d\underline{\varepsilon}^3}{6}K''' + N\big(0, I\underline{\varepsilon}^2 d\big)\right)\right] \\
&= \mathbf{E}\left[1 \wedge \exp\left(\frac{\underline{\ell}^2}{2}K'' + \frac{\underline{\varepsilon\ell}^2}{6}K''' + N\big(0, I\underline{\ell}^2\big)\right)\right] \\
&= \Phi\left(-\frac{I^{1/2}\underline{\ell}}{2} + \frac{\underline{\varepsilon\ell}K'''}{6I^{1/2}}\right) \\
&\quad + \exp(\underline{\varepsilon\ell}^2 K'''/6)\Phi\left(-\frac{I^{1/2}\underline{\ell}}{2} - \frac{\underline{\varepsilon\ell}K'''}{6I^{1/2}}\right).
\end{aligned} \tag{14}
$$

Similarly,

$$
\begin{aligned}
\alpha^+ &= \mathbf{E}\left[1 \wedge \frac{f_d^{\beta-\varepsilon}(X)e^{dK(\beta-\varepsilon)}}{f_d^{\beta}(X)e^{dK(\beta)}}\right] \\
&= \mathbf{E}\left[1 \wedge \exp\left(\big(K(\beta-\varepsilon) - K(\beta)\big)d - \varepsilon dM(\beta) - \varepsilon\sum_{i=1}^{d} \overline{g}(X_i)\right)\right] \\
&\overset{1}{\approx} \mathbf{E}\left[1 \wedge \exp\left(\frac{d\varepsilon^2}{2}K'' - N\big(0, I\varepsilon^2 d\big)\right)\right] \\
&= \mathbf{E}\left[1 \wedge \exp\left(\frac{\ell^2}{2}I - \frac{\varepsilon\ell^2}{6}K''' - N\big(0, I\ell^2\big)\right)\right] \\
&= \Phi\left(-\frac{I^{1/2}\ell}{2} - \frac{\varepsilon\ell K'''}{6I^{1/2}}\right) \\
&\quad + \exp(-\varepsilon\ell^2 K'''/6)\Phi\left(-\frac{I^{1/2}\ell}{2} - \frac{\varepsilon\ell K'''}{6I^{1/2}}\right).
\end{aligned}
$$

Hence

$$
\begin{aligned}
\alpha^+\big(\beta_i^{(d)}\big) &\overset{d^{1/2}}{\approx} \Phi\left(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2} - \frac{\varepsilon\ell K'''(\beta_i^{(d)})}{6I^{1/2}(\beta_i^{(d)})}\right) \\
&\qquad + \exp\big(-\varepsilon\ell^2(\beta_i^{(d)})K'''(\beta_i)/6\big)\Phi\left(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2} + \frac{\varepsilon\ell K'''(\beta_i^{(d)})}{6I^{1/2}(\beta_i^{(d)})}\right).
\end{aligned}
$$

A first order approximation of this expression is

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

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,

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

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

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

Hence, ignoring all lower order terms,

$$
\begin{aligned}
\mu(\beta_i^{(d)}) \overset{d^{1/2}}{\approx} \frac{1}{2d^{1/2}}\Big[&-\ell\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2} \\
&\qquad \times\Big(\frac{-\ell}{\sqrt{d}}\Big)\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}-\frac{\varepsilon\ell K'''(\beta_i^{(d)})}{6I^{1/2}(\beta_i^{(d)})}\Big) \\
&-\ell\exp\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2} \\
&\qquad \times\Big(\frac{-\ell}{\sqrt{d}}\Big)\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}+\frac{\varepsilon\ell K'''(\beta_i^{(d)})}{6I^{1/2}(\beta_i^{(d)})}\Big) \\
&\qquad\qquad +\frac{2\Phi(-I^{1/2}(\beta_i^{(d)})\ell/2)\ell\ell'}{d^{1/2}}\Big] \\
\overset{d^{1/2}}{\approx} \frac{1}{d}\Big[&-\ell^2\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2}\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big) \\
&\qquad +\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\ell\ell'\Big].
\end{aligned}
$$

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 $0<\beta<1$)

$$
\begin{aligned}
G^d h \overset{d}{\approx} \frac{1}{d}\Big(&\ell^2\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)h''(\beta) \\
&+\Big[\Phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\ell\ell' \\
&\quad -\ell^2\frac{(\ell(\beta_i^{(d)})I^{1/2}(\beta_i^{(d)}))'}{2}\phi\Big(-\frac{I^{1/2}(\beta_i^{(d)})\ell}{2}\Big)\Big]h'(\beta)\Big).
\end{aligned}
$$

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). $\square$

The uniformity over $\mathrm{int}(E_d)$ 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 used here, the extension to the boundary does indeed hold.

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

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

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,

$$\begin{aligned} G^{(d)} h_d(\chi_d) &\overset{d}{\approx} \frac{h'_d(\chi_d)[\alpha^-(\chi_d^+ - \chi_d)]}{2} \\ &\quad + \frac{h''_d(\chi_d)}{4}\left[(\chi_d - \chi_d^+)^2 \alpha^-\right] \\ &= \frac{h'_d(\chi_d)}{2} \frac{\alpha^- \ell(\chi_d^+)}{d^{1/2}} + \frac{h''_d(\chi_d)}{4}\left[\frac{\ell(\chi_d)^2 \alpha^-}{d}\right] \\ &\overset{d}{\approx} \frac{\alpha^- \ell(\chi_d^+)}{2d^{1/2}} \left(h'(\chi) + \eta'(\chi) d^{-1/2}\right) \\ &\quad + \frac{h''_d(\chi_d)}{4}\left[\frac{\ell(\chi_d)^2 \alpha^-}{d}\right]. \end{aligned}$$

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 formulae (8) and (9), this completes the proof. $\square$

Finally, we provide the missing proof from Section 2.1.

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

$$\begin{aligned}
&\mathbf{E}\left(Y^\sigma_{m,t+h} - Y^\sigma_{m,t} \,\middle|\, Y^\sigma_{m,t} = \frac{i}{m}\right) \\
&\quad \approx \left(\frac{m^2 Sh}{2}\right)\left(\frac{1}{m}\right)\left(\frac{1}{2S}\right) \\
&\qquad \times \left[\sigma^2\left(\frac{i}{m}\right) + \frac{\pi((i+1)/m)\sigma^2((i+1)/m)}{\pi(i/m)}\right. \\
&\qquad\qquad \left. - \sigma^2\left(\frac{i}{m}\right) - \frac{\pi((i-1)/m)\sigma^2((i-1)/m)}{\pi(i/m)}\right] \\
&\quad = \frac{hm}{4}\left[\frac{\pi(x+e)\sigma^2(x+e)}{\pi(x)} - \frac{\pi(x-e)\sigma^2(x-e)}{\pi(x)}\right] \\
&\quad \approx \frac{hm}{4}\big[\big((\pi(x) + e\pi'(x))(\sigma^2(x) + e(\sigma^2)'(x)) \\
&\qquad\qquad - (\pi(x) - e\pi'(x))(\sigma^2(x) - e(\sigma^2)'(x))\big)/\pi(x)\big] \\
&\quad = \frac{hm}{4}\left[\frac{2e\pi'(x)\sigma^2(x) + 2e\pi(x)(\sigma^2)'(x)}{\pi(x)}\right] \\
&\quad = \frac{hm}{4}(2e)\big[(\log\pi)'(x)\sigma^2(x) + 2\sigma(x)\sigma'(x)\big] \\
&\quad = h\left[\frac{1}{2}(\log\pi)'(x)\sigma^2(x) + \sigma(x)\sigma'(x)\right]
\end{aligned}$$

and also

$$\begin{aligned}
&\mathbf{E}\left((Y^\sigma_{m,t+h} - Y^\sigma_{m,t})^2 \,\middle|\, Y^\sigma_{m,t} = \frac{i}{m}\right) \\
&\quad \approx \left(\frac{m^2 Sh}{2}\right)\left(\frac{1}{2S}\right)\left(\frac{1}{m^2}\right)\big[2\sigma^2(x) + 2\sigma^2(x)\big] = h\big[\sigma^2(x)\big].
\end{aligned}$$

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]. $\square$

4. Discussion. This paper has linked the usual Peskun ordering on asymptotic variance of discrete-time Markov chains, to asymptotic variance of diffusion processes. It has then applied these results to simulated tempering algorithms, by proving that the inverse-temperatures of such algorithms converge (in an appropriate limit) to a diffusion. By maximising the speed of the resulting diffusion, it has obtained results about the optimal choice of the temperature spacings.

We believe that Theorem 1 could be useful in other contexts as well, whenever we wish to compare two Langevin diffusion algorithms directly, or alternatively whenever we wish to compare two discrete-time processes which both have appropriate diffusion limits.

Of course, Theorem 1 requires assumptions (A1) and (A2). These are primarily just regularity assumptions, which would likely be satisfied in most applications of interest. On the other hand, the "exponentially-bounded tails" aspect of assumption (A2) is more than technical; rather, it provides us with some control over the extreme tail excursions of the processes which we consider, and we suspect that our limiting results might fail if no such control is provided.

Finally, our simulated tempering diffusion limit is only proven under the rather strong and artificial assumption (6) involving a product form of the target density. Indeed, this assumption is central to our method of proof. However, as mentioned earlier, it is known [2–5, 20, 22] that the general conclusions in this special case often hold in greater generality, either approximately in numerical simulation studies, or theoretically through more general methods of proof. In a similar spirit, we believe that the simulated tempering diffusion limit proven herein would approximately hold numerically in greater generality. In addition, it might be possible to prove a stronger version of our diffusion limit, with weaker assumptions, though such proofs would get rather technical and we do not pursue them here.

REFERENCES

[1] ATCHADÉ, Y. F., ROBERTS, G. O. and ROSENTHAL, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Stat. Comput.* **21** 555–568. MR2826692

[2] BÉDARD, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.* **17** 1222–1244. MR2344305

[3] BÉDARD, M. (2008). Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *Stochastic Process. Appl.* **118** 2198–2222. MR2474348

[4] BÉDARD, M. and ROSENTHAL, J. S. (2008). Optimal scaling of Metropolis algorithms: Heading toward general target distributions. *Canad. J. Statist.* **36** 483–503. MR2532248

[5] BESKOS, A., ROBERTS, G. and STUART, A. (2009). Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.* **19** 863–898. MR2537193

[6] BHATTACHARYA, R. N. and WAYMIRE, E. C. (1990). *Stochastic Processes with Applications*. Wiley, New York. MR1054645

[7] BROOKS, S., GELMAN, A., JONES, G. L. and MENG, X.-L., eds. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, Boca Raton, FL. MR2742422

[8] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and convergence*. Wiley, New York. MR0838085

[9] FORT, G., MOULINES, E. and PRIOURET, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann*. *Statist*. **39** 3262–3289. MR3012408

[10] GEYER, C. (1992). Practical Markov chain Monte Carlo. *Statist*. *Sci*. **7** 473–483.

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

[12] LEISEN, F. and MIRA, A. (2008). An extension of Peskun and Tierney orderings to continuous time Markov chains. *Statist*. *Sinica* **18** 1641–1651. MR2469328

[13] LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653

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

[15] MIRA, A. (2001). Ordering and improving the performance of Monte Carlo Markov chains. *Statist*. *Sci*. **16** 340–350. MR1888449

[16] MIRA, A. and GEYER, C. J. (2000). On non-reversible Markov chains. In *Monte Carlo Methods* (*Toronto*, *ON*, 1998). *Fields Institute Communications* **26** 95–110. Amer. Math. Soc., Providence, RI. MR1772309

[17] MIRA, A. and LEISEN, F. (2009). Covariance ordering for discrete and continuous time Markov chains. *Statist*. *Sinica* **19** 651–666. MR2514180

[18] PESKUN, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60** 607–612. MR0362823

[19] PREDESCU, C., PREDESCU, M. and CIOBANU, C. V. (2004). The incomplete beta function law for parallel tempering sampling of classical canonical systems. *J*. *Chem*. *Phys*. **120** 4119–4128.

[20] ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann*. *Appl*. *Probab*. **7** 110–120. MR1428751

[21] ROBERTS, G. O. and ROSENTHAL, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **60** 255–268. MR1625691

[22] ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist*. *Sci*. **16** 351–367. MR1888450

[23] SINCLAIR, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combin*. *Probab*. *Comput*. **1** 351–370. MR1211324

[24] TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann*. *Statist*. **22** 1701–1762. MR1329166

[25] TIERNEY, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann*. *Appl*. *Probab*. **8** 1–9. MR1620401

[26] WARD, A. R. and GLYNN, P. W. (2003). A diffusion approximation for a Markovian queue with reneging. *Queueing Syst*. **43** 103–128. MR1957808

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY, CV4 7AL
UNITED KINGDOM
E-MAIL: g.o.roberts@lancaster.ac.uk

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO, M5S 3G3
CANADA
E-MAIL: jeff@math.toronto.edu
URL: http://probability.ca/jeff/