Skew Brownian Motion and Complexity of the ALPS Algorithm

Gareth O. Roberts[1], Jeffrey S. Rosenthal[2], and Nicholas G. Tawn[3]

[1]*Department of Statistics, University of Warwick, United Kingdom, CV4 7AL,* Gareth.O.Roberts@warwick.ac.uk

[2]*Department of Statistical Sciences, University of Toronto, 100 St. George Street, Room 6018, Toronto, Ontario, Canada M5S 3G3,* jeff@math.toronto.edu

[3]*Department of Statistics, University of Warwick, United Kingdom, CV4 7AL,* n.tawn.1@warwick.ac.uk

(September, 2020)

Abstract

Simulated tempering is a popular method of allowing MCMC algorithms to move between modes of a multimodal target density π. The paper [24] introduced the Annealed Leap-Point Sampler (ALPS) to allow for rapid movement between modes. In this paper, we prove that, under appropriate assumptions, a suitably scaled version of the ALPS algorithm converges weakly to skew Brownian motion. Our results show that under appropriate assumptions, the ALPS algorithm mixes in time $O(d[\log d]^2)$ or $O(d)$, depending on which version is used.

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms [5], such as the Metropolis-Hastings algorithm [14, 8], are very widely used to explore and sample complicated high-dimensional target probability distributions, but they have a tendency to get stuck in local modes which limits their effectiveness. Annealing and tempering methods [15, 9, 1, 7, 13] attempt to overcome this problem by raising the target probability's density to an inverse-temperature power $\beta > 0$. Small values $\beta \ll 1$, corresponding to hot temperatures, lead to flatter target densities which can be explored more easily. Then, returns to $\beta = 1$ correspond to the original target density and can thus be "counted" as correct sampling.

Despite the tremendous success of tempering, these methods suffer from deficiencies, especially in high dimensions. In particular, tempering of distributions does not usually preserve the relative mass contained in each of the modes. This was addressed in [26], which provides a methodology which overcomes the weight instability problem as long as all modes

look reasonably Gaussian. Unfortunately, in applications this is often not the case, since modes often exhibit significant skewness.

An alternative approach, the Annealed Leap-Point Sampler (ALPS), was introduced in [24]. This algorithm instead considers very *large* values $\beta \gg 1$, corresponding to very peaked target densities at very cold temperatures. (Large β are often used for optimisation purposes, but are not normally used by sampling algorithms.) The resulting sharp peaks then become approximately Gaussian, thus facilitating simpler ways of moving between them. Furthermore, a weight-preserving transformation is performed to approximately preserve the probabilistic weight of each peak upon tempering. However, an important theoretical question concerns the extra computational overhead associated with this scheme, which is what we address here.

In this paper, we study the ALPS algorithm in terms of diffusion limits as the dimension d tends to infinity, an established technique for establishing complexity order of MCMC algorithms [18, 19, 20, 4, 21]. We prove that, under appropriate assumptions, a suitably scaled version of the ALPS algorithm converges to *skew Brownian motion* (e.g. [11]), as explained in Theorem 5 below. This limit will allow us to draw conclusions about the computational complexity of our algorithm, and show that under appropriate assumptions, as the dimension $d \to \infty$ the ALPS algorithm mixes in time $O(d[\log d]^2)$ or $O(d)$, depending on which version is used.

2 Background

In this section, we briefly present some background and context for the stochastic processes we shall study herein. Readers already familiar with these topics can skip this section.

Markov chain Monte Carlo (MCMC) algorithms run a Markov chain which converges to a stationary probability density π, thus facilitating sampling from that distribution (and, ultimately, estimating its probabilities and expected values). They are extremely popular in a wide variety of domains (see e.g. [5] and the many references therein).

The most basic version of MCMC is the *Metropolis algorithm* [14, 8]. From a given state x, it proceeds by first *proposing* to move to a new state y, and then either *accepting* that proposal (i.e., moving to y), or *rejecting* that proposal (i.e., staying at x). The *acceptance probability* is given by the minimum of 1 and the ratio $\pi(y) \,/\, \pi(x)$. Provided that the proposal densities are symmetric (i.e., have the same probability of proposing y from x, as of proposing x from y), this procedure ensures that the resulting Markov chain will be reversible with respect to π, and thus have π as its stationary density.

MCMC algorithms can have problems moving between different modes of π. To deal with this, various forms of *tempering* [7, 13] consider different powers π^β of the target density, where $\beta \leq 1$ is an *inverse-temperature*. Then $\beta = 1$ corresponds to the desired distribution so those are the only samples which are "counted", but small positive values $\beta \ll 1$ make the density flatter and thus much easier to traverse. (The related method of *simulated annealing* [15, 9, 1] instead lets β to grow to large values for optimisation purposes, i.e. to find maximum values of the target density, as opposed to sampling from its distribution.)

One problem with tempering is that small values of β change the relative weights of

the modes of π, so that previously important modes could get ignored when exploring at small inverse-temperatures. To correct for this, the paper [26] introduced *weight-preserving* tempering transitions, which adjust the tempering transformations to avoid this problem and preserve the same relative weights of all of the modes.

For any MCMC algorithm, an important question is how quickly it converges to its stationary distribution π. While there have been many attempts to bound such convergence times directly (see e.g. [22] and the references therein), much of the effort has been focused on questions of *computational complexity*, i.e. how the algorithm's running time grows as a function of other parameters (dimension, size of data, etc.).

One particularly promising, though technically challenging, approach to determining the computational complexity of Metropolis algorithms is through the use of *diffusion limits*. Just as simple symmetric random walk converges to Brownian motion under appropriate rescaling, so certain transforms of some Metropolis algorithms will converge to Langevin diffusions. This was exploited originally in the paper [18], to derive complexity and optimality results for ordinary random-walk-based Metropolis algorithms, and was later generalised to many other contexts (see e.g. [20] and the references therein).

3 The ALPS Algorithm

As discussed above, tempering methods for MCMC usually use *small* positive values of $\beta \ll 1$, to make the target distribution flatter, and thus allow for easier mixing between modes. By contrast, the paper [24] introduced the *Annealed Leap-Point Sampler* (ALPS) algorithm, which instead uses *large* values $\beta \gg 1$ (while still using weight-preserving tempering transitions as in [26] so the modes retain their same relative masses). Such transformations make the modes of π even more separated. However, under certain smoothness and integrability assumptions, they also make each mode approximately Gaussian. This allows for auxiliary Markov chain steps which move between the different modes (which are now similarly-shaped). Then, as usual, only samples in the original temperature $\beta = 1$ are "counted" as actual samples from π.

To illustrate the idea of this algorithm, consider the following illustrative example in dimension $d = 5$. Suppose the target density π on $\mathbf{R}^5$ is a mixture of two *skew-normal* modes centered at $(-20, -20, -20, -20, -20)$ and $(20, 20, 20, 20, 20)$ respectively, with scalings 1 and 2 respectively, and with skew parameter $\alpha = 10$, so for all $\theta \in \mathbf{R}^5$,

$$\pi(\theta) \;=\; (0.7) \prod_{i=1}^{5} 2\,\phi\big(\theta_i + 20\big)\,\Phi\big(10(\theta_i + 20)\big) \;+\; (0.3) \prod_{i=1}^{5} \phi\big(\frac{1}{2}(\theta_i - 20)\big)\,\Phi\big(5(\theta_i - 20)\big)\,,$$

where as usual $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^{x} \phi(u)\,du$; see Figure 1.

In such an example, it is very easy for a Markov chain to mix separately *within* either of the two modes. The challenge is to move between the modes (which is virtually impossible for a typical fixed-temperature Metropolis algorithm even in this simple 5-dimensional example). The ALPS algorithm introduces a special move so that at very large inverse-temperature values $\beta \gg 1$, the chain can exploit the near-Gaussianity of each of the modes to easily jump

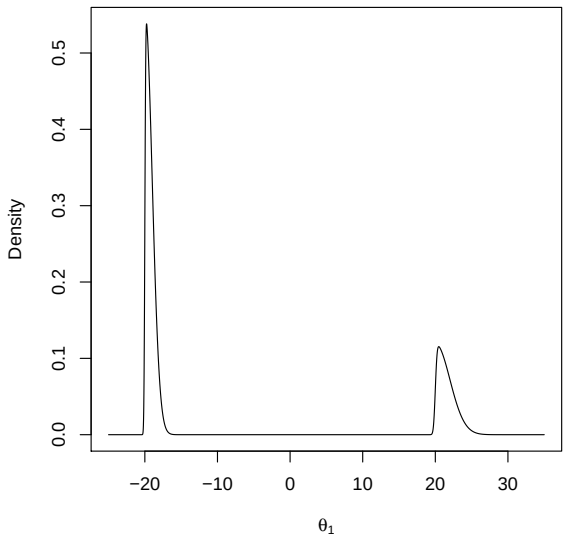

Figure 1: The θ_1 marginal of the target density in the illustrative example.

between them. To illustrate this, Figure 2 shows a trace plot of the inverse-temperature values β while the algorithm proceeds, and also indicates by colour which of the two modes the chain is in (i.e., closest to). As can be seen from the plot, the chain stays in the same mode for long periods of time, and only switches modes when the values of β are very large at which point it jumps to either mode with its correct probability. (Note that this description is for the "vanilla" version of ALPS; see Remark 1 below.)

Figure 2 illustrates that the key to the ALPS algorithm's success is moving rapidly between the large $\beta = \beta_{max} = 256$ values (which allow for mixing between the modes) and the small $\beta = 1$ value (which can be "counted" as a sample from π). However, it is not clear how quickly such mixing takes place, and in particular how it changes depending on the target π and dimension d. To study this, we would like to prove a diffusion limit of a suitably scaled version of the β process, but it is not clear from Figure 2 what sort of limiting diffusive behaviour is available.

To make further progress, we consider a suitable transformation of β. Namely, we replace β by $s \log(\beta_{max}/\beta)$, where $s = 1$ if the chain is in mode 1 or $s = -1$ if the chain is in mode 2. The resulting process is shown in Figure 3, which suggests that this modified functional does indeed start to resemble a diffusive process. Indeed, away from the special mode-hopping value 0, the process looks something like ordinary Brownian motion. In fact, we shall prove below (Theorem 5) that under appropriate assumptions and scalings, this modified process converges to a skew Brownian motion.

More precisely, we shall prove diffusion limits for suitably rescaled versions of the ALPS algorithm, as the dimension $d \to \infty$. We shall assume that the ALPS algorithm can easily jump between modes when it reaches the sufficiently large inverse-temperature $\beta_{max}^{(d)}$, but that it is stuck within one mode whenever $\beta < \beta_{max}^{(d)}$. We therefore focus on how the

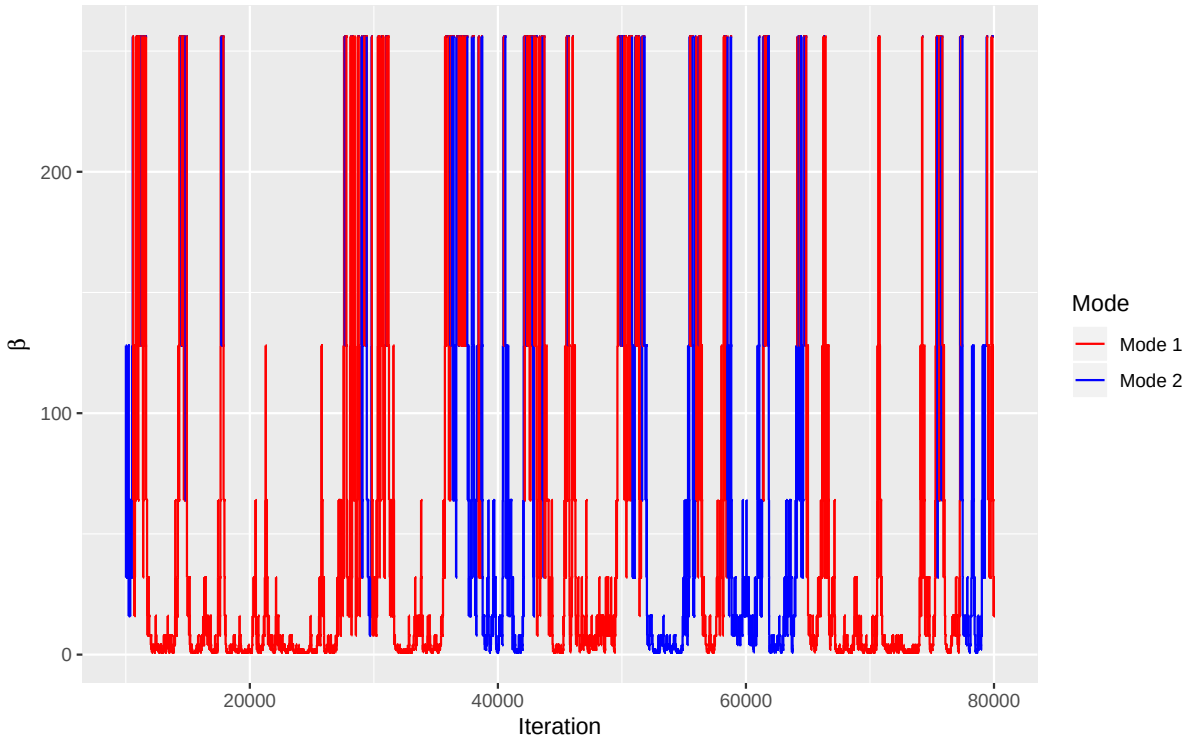

Figure 2: Trace plot of the β values in the illustrative example, coloured to indicate whether the chain is in the first mode (red) or second mode (blue).

inverse-temperatures β themselves are updated by the algorithm. In particular, we will prove (Theorem 5) that a particular rescaling of the β process converges to *skew Brownian motion* [11]. This will in turn allow us to derive computational complexity results (Section 6).

Remark 1 The ALPS algorithm studied herein differs in certain ways from the full algorithm run in actual applications as in [24]. For example, we assume the process mixes perfectly between modes when $\beta = \beta_{max}^{(d)}$ and not at all when $\beta < \beta_{max}^{(d)}$, while in practice it would mix better and better at higher β values but never perfectly. Also, the ALPS algorithm actually uses *parallel tempering*, in which a separate chain is run at each temperature and their values are swapped; the single β process studied herein can then be thought of as following which of the chains is currently carrying state information between larger and smaller inverse-temperatures and thus facilitating mixing (cf. Section 4 of [2]). Finally, the full ALPS algorithm in [24] also makes use of the QuanTA transformation [25], an additional affine transformation to increase the efficiency of the temperature-swap moves, which we omit in the "vanilla" version of ALPS described here; we discuss the effect of this extra QuanTA transformation in Corollary 7 below.

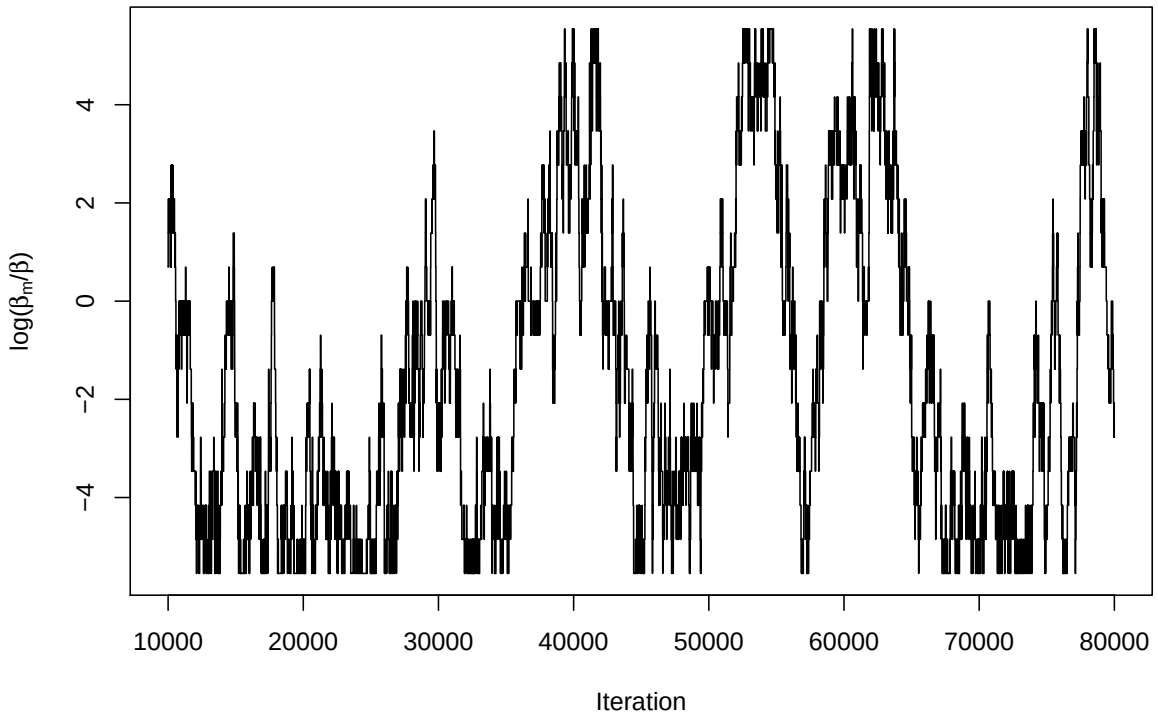

Figure 3: A trace plot of the transformed values $\log(\beta_{max}/\beta)$ in the illustrative example, multiplied by -1 when the chain is in mode 2.

4 Assumptions

We consider a version of the Annealed Leap-Point Sampler (ALPS) algorithm of [24]. We assume the chain always mixes immediately within each mode, but the chain can only jump between modes when at the sufficiently cold inverse-temperature $\beta = \beta_{max}^{(d)}$, at which point it immediately jumps to any of its modes with the correct probability weight.

We assume as in [24] that our collection of inverse-temperatures is given by $1 = \beta_0^{(d)} < \beta_1^{(d)} < \ldots < \beta_{k(d)}^{(d)} \approx \beta_{max}^{(d)}$, Similar to [2] and [21], following [16] and [10], we assume that the inverse temperatures are related by

$$\beta_i = \beta_{i-1} + \ell(\beta_{i-1})/d^{1/2} \tag{1}$$

for some fixed C^1 function ℓ. Indeed, it is shown in [2, 21] that in the single-mode case, under a strong assumption about iid targets (see below), it is optimal to choose

$$\ell(\beta) \;=\; I^{-1/2}(\beta)\,\ell_0 \tag{2}$$

for some fixed constant $\ell_0 \doteq 2.38$, where $I(\beta) = \mathrm{Var}_{x\sim f^\beta}(\log f(x))$.

We note that [24] also makes use of the *QuanTA Algorithm* of [25], which modifies the temperature-swap moves for greater efficiency. In our analysis below, we do *not* make use of QuanTA, though we do comment on its effect in Corollary 7 below.

To facilitate theoretical analysis, we assume that the target density π is a mixture distribution, given by

$$\pi(x) \;\propto\; \sum_{j=1}^{J} h_j(x) \;=\; \sum_{j=1}^{J} w_j g_j(x) \tag{3}$$

where $g_j(x)$ is a normalised target density and w_j are weights. Then for each inverse-temperature $\beta \geq 1$, we set

$$\pi_\beta(x) \;\propto\; \sum_{j=1}^{J} f_j(x,\beta) \;=\; \sum_{j=1}^{J} W_{(j,\beta)} \frac{[g_j(x)]^\beta}{\int [g_j(x)]^\beta dx} \tag{4}$$

for appropriate weights $W_{(j,\beta)}$. We assume for simplicity (though see Remark 8 below) that we have just $J = 2$ modes, of weights w_1 and $w_2 = 1 - w_1$ respectively. We assume we have some way of allocating each state x to one of the two modes, e.g. to whichever mode it is closer to (if the modes are well-separated then the precise mechanism for this does not matter). We further assume that

$$\pi_\beta(x) \;\propto\; w_1 \, \frac{[g_1(x)]^\beta}{\int [g_1(x)]^\beta dx} \;+\; w_2 \, \frac{[g_2(x)]^\beta}{\int [g_2(x)]^\beta dx} \;\equiv\; w_1 \, g_1^\beta(x) + w_2 \, g_2^\beta(x) \tag{5}$$

for each β, where we use the same values of w_1 and w_2 for each inverse-temperature due to the weight preserving properties of the ALPS algorithm. Furthermore, we assume as in the original MCMC diffusion limit results [18] that each of the individual components g_i consists of iid components in d-dimensions, i.e. that each $g_i(x) = \prod_{j=1}^d \overline{g}_i(x_j)$ for some fixed one-dimensional density function $\overline{g}_i$, thus allowing us to apply the diffusion-limit results of [21] within each individual target mode.

A useful situation to consider is the *Exponential Power Family* special case in which each of the two mixture component factors g_j is of the form $g_j(x) \propto e^{-\lambda_j |x|^{r_j}}$ for some $\lambda_j, r_j > 0$. If so, then for each individual mode we have $I(\beta) = 1/r_j\beta^2$. The corresponding choice of ℓ from (2) would then be $\ell(\beta) = \beta/\sqrt{r_i}$ in mode i. This includes the Gaussian case, for which $r_1 = r_2 = 2$ and $\lambda_j = 1/\sigma_j^2$.

5 Main Results

We now state various weak convergence results for various transformations of our process. Let $\beta_t^{(d)}$ be the inverse temperature at time t for the d-dimensional process. Let $\beta_{N(dt)}^{(d)}$ be a continuous-time version of the $\beta_t^{(d)}$ process, sped up by a factor of d, where $\{N(t)\}$ is an independent standard rate-1 Poisson process. To combine the two modes into one single process, we further augment this process by multiplying it by -1 when the algorithm's state

is allocated to the second mode, while leaving it positive (unchanged) when state is allocated to the first mode. Thus define

$$X_t^{(d)} \;=\; \begin{cases} \beta^{(d)}_{N(dt)}\,, & \text{in mode 1} \\ -\beta^{(d)}_{N(dt)}\,, & \text{in mode 2} \end{cases} \tag{6}$$

Our first diffusion limit result (proved in Section 7), following [21], states that within each mode, the inverse temperature process behaves identically to the case where there is only one mode (i.e. $J = 1$). To state it, we extend the definition of I to $I(\beta) = \mathrm{Var}_{x\sim f_1^{\beta}}(\log f_1(x))$ for $\beta > 0$, and $I(\beta) = \mathrm{Var}_{x\sim f_2^{|\beta|}}(\log f_2(x))$ for $\beta < 0$, so that positive values correspond to the first mode while negative values correspond to the second mode.

Theorem 2 *Assume the target π is of the form (3), with $J = 2$ modes of weights w_1 and $w_2 = 1 - w_1$, with inverse weights chosen as in (1). Then up until the first time it reaches 1 or $\beta^{(d)}_{max}$, the process $\{X_t^{(d)}\}$ defined by (6) converges weakly as $d \to \infty$ to a fixed diffusion process X, which for $X^{(d)} > 0$ satisfies*

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

The same equation holds for $X_t < 0$, except with the sign of the drift reversed.

As a check, (7) satisfies the general relation $\mu(x) = \frac{1}{2}\sigma^2(x)\frac{d}{dx}\log\pi(x) + \sigma(x)\sigma'(x)$. This implies that π is *locally invariant* for $X^{(d)}$, meaning formally that its generator G satisfies that $\pi\,(Gf)(x) = 0$ for appropriate smooth f and for x in the interior of the domain, or informally that π is stationary for $X^{(d)}$ locally within each mode, as we would expect.

On the other hand, Theorem 2 describes only what happens on each mode separately; it says nothing about the mode-switching process itself. Moreover, its state space $(-\infty, -1] \cup [1, \infty)$ is not connected. In fact, we will see below that as $d \to \infty$, the value $\beta^{(d)}_{max}$ will go to infinity and hence never be reached in finite time. To resolve these issues, we make several transformations on the $X_t^{(d)}$ process. First, for $|x| \geq 1$, we define

$$h(x) \;=\; \int_1^{|x|} \frac{1}{\ell(u)}du\,.$$

We then set

$$H_t^{(d)} \;=\; \mathrm{sign}(X_t^{(d)})\;\left[1 + \frac{h\left(X^{(d)}_{t\,(h(\beta^{(d)}_{max}))^2}\right)}{h(\beta^{(d)}_{max})}\right]\,. \tag{8}$$

Hence, $1 \le H_t^{(d)} \le 2$ in the first mode, and $-1 \ge H_t^{(d)} \ge -2$ in the second mode. Also, $H_t^{(d)}$ is sped up by a factor of $h(\beta_{max}^{(d)})^2$ from $X_t^{(d)}$, and hence moves at Poisson rate $d\,h(\beta_{max}^{(d)})^2$. These new processes $H_t^{(d)}$ satisfy the following.

Theorem 3 *Under the set-up and assumptions of Theorem 2, on $(-2,-1)\cup(1,2)$ (i.e., away from its boundary points), as $d \to \infty$ the process $\{H_t^{(d)}\}$ converges weakly in the Skorokhod topology to a limiting diffusion H which satisfies*

$$dH_t \;=\; \left[2\,\Phi\left(\frac{-\ell(X_t)I^{1/2}(X_t)}{2}\right)\right]^{1/2} dB_t + \;\ell(X_t)\left[\Phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)\right]' dt\,. \qquad (9)$$

Furthermore, H leaves constant (uniform) densities locally invariant.

To make further progress, we now assume a *Proportionality Condition*, that the quantities corresponding to $I(\beta)$ are proportional to each other in the two modes. That is, we assume there is a fixed C^1 function $I_0 : \mathbb{R}_+ \to \mathbb{R}_+$, and positive constants r_1 and r_2, such that we have $I(\beta) = I_0(\beta)/r_1$ in the first mode, and $I(\beta) = I_0(|\beta|)/r_2$ in the second mode. (It follows from Section 2.4 of [2] that in the Exponential Power Family case, $I(\beta) = 1/r_1\beta^2$ for $\beta > 0$ and $I(\beta) = 1/r_2\beta^2$ for $\beta < 0$, so this Proportionality Condition holds in that case.) Also, inspired by the optimal choice (2) for the inverse-temperature spacing function ℓ in the single-mode case, we assume that

$$\ell(\beta) \;=\; I_0^{-1/2}(\beta)\,\ell_0 \qquad (10)$$

for some fixed constant $\ell_0 > 0$. In this case, $\ell(X_t)\,I^{1/2}(X_t) = \ell_0 r_1^{1/2}$ for $X_t < 0$, and $\ell(X_t)\,I^{1/2}(X_t) = \ell_0 r_2^{1/2}$ for $X_t > 0$, with $\Big[\ell(X_t)\,I^{1/2}(X_t)\Big]' = 0$ for all $X_t \neq 0$. Hence, Theorem 3 immediately gives:

Corollary 4 *Under the set-up and assumptions of Theorem 2, assuming the above Proportionality Condition and the choice (10), then as $d \to \infty$, the process $\{H_t^{(d)}\}$ converges weakly in the Skorokhod topology to a limit process H on $(-2,-1)$ and on $(1,2)$, i.e. away from its boundary points. Furthermore, H is a diffusion, with drift 0, and with diffusion coefficient which is constant on each of the two intervals $(-2,-1)$ and $(1,2)$, i.e.*

$$dH_t \;=\; s(H_t)\;dB_t\,.$$

Here $s(H_t) = s_1$ for $H_t \in (1,2)$, and $s(H_t) = s_2$ for $H_t \in (-2,-1)$, where

$$s_i \;:=\; \left[2\,\Phi\left(-\frac{1}{2}\ell_0 r_i^{1/2}\right)\right]^{1/2}.$$

Next, we need to join up the two parts of the domain $[-2,-1]\cup[1,2]$ of the process $H_t^{(d)}$. Now, the original process can jump between modes when at the coldest temperture $\beta_{max}^{(d)}$, corresponding to the values ± 2 for the transformed process $H_t^{(d)}$. Hence, we let

$$Z_t^{(d)} \;=\; 2\;\mathrm{sign}(H_t^{(d)}) - H_t^{(d)} \;=\; \begin{cases} 2 - H_t^{(d)}\,, & H_t^{(d)} \ge 1,\ \text{i.e. in mode 1} \\ -2 - H_t^{(d)}\,, & H_t^{(d)} \le -1,\ \text{i.e. in mode 2} \end{cases}$$

so that $Z_t^{(d)}$ has domain $[-1, 1]$ with mode-switching at 0.

However, by Corollary 4, the limit of the process $Z_t^{(d)}$ will still have diffusion coefficient s_1 or s_2 on its positive and negative parts. We thus rescale the process by setting

$$W_t^{(d)} \;=\; s(Z_t^{(d)})^{-1}\, Z_t^{(d)}\,,$$

so that $W_t^{(d)}$ has domain $[-\frac{1}{s_2}, \frac{1}{s_1}]$, and limit which is actual Brownian motion on each of $(-\frac{1}{s_2}, 0)$ and $(0, \frac{1}{s_1})$. The precise limit of this process requires the notion of *skew Brownian motion*, a generalisation of usual Brownian motion that, intuitively, behaves just like a Brownian motion except that the sign of each excursion from 0 is chosen using an independent Bernoulli random variable; for further details and constructions and discussion see e.g. [11]. In terms of skew Brownian motion, we have:

Theorem 5 *Under the set-up and assumptions of Theorem 2, assuming the above Proportionality Condition and the choice (10), as $d \to \infty$ the process $\{W_t^{(d)}\}$ converges weakly in the Skorokhod topology to a limit process W which is skew Brownian motion on $[-\frac{1}{s_2}, \frac{1}{s_1}]$, with reflecting boundaries at $-\frac{1}{s_2}$ and at $\frac{1}{s_1}$, and with excursion probabilities at 0 proportional to $w_1 s_1$ (to go positive) and $w_2 s_2$ (to go negative), respectively.*

6 Computational Complexity

Theorem 5 has implications for the computational complexity of the ALPS algorithm. Indeed, it shows that the limiting process W does not depend at all on the dimension d, and hence has convergence time $O(1)$ as $d \to \infty$. However, W was derived from the processes $H_t^{(d)}$ and $Z_t^{(d)}$, which sped up time by a factor of $(h(\beta_{max}^{(d)}))^2$ from the process $X_t^{(d)}$, which itself sped up time by a factor d. That is, W was sped up by a total factor of $d[h(\beta_{max}^{(d)})]^2$. So, in the original scaling, the convergence time is $O(d[h(\beta_{max}^{(d)})]^2)$. This raises the question of how $h(\beta_{max}^{(d)})$ grows as a function of d.

Now, it is proven in [24] that for the ALPS process to mix modes efficiently, we need the maximum inverse-temperature value $\beta_{max}^{(d)}$ to grow linearly with dimension, i.e. we need to choose $\beta_{max}^{(d)} \propto d$. Furthermore, in the Exponential Power Family case, as mentioned, $I(\beta) \propto 1/\beta^2$, so $\ell(\beta) = I_0^{-1/2}(\beta)\,\ell_0 \propto \beta$. It then follows that

$$h(x) \;=\; \int_1^{|x|} \frac{1}{\ell(u)}\, du \;\propto\; \int_1^{|x|} \frac{1}{u}\, du \;=\; \log|x|\,.$$

Hence, $h(\beta_{max}^{(d)}) \propto \log(d)$, so the complexity order is $O\big(d\,[\log d]^2\big)$. That is, for the inverse temperature process to hit $\beta_{max}^{(d)}$ and hence mix modes, takes $O\big(d\,[\log d]^2\big)$ iterations.

If we are not in the Exponential Power Family case, then it may no longer be true that $I(\beta) \propto 1/\beta^2$. However, as $d, \beta \to \infty$, under appropriate smoothness assumptions the densities in the different modes will become approximately Gaussian, which corresponds to the Exponential Power Family case with $r = 2$. And, it is proven in equation (66) of [25] that if the first four moments converge to those of a Gaussian, then $2\beta^2 I(\beta) \to 1$, i.e.

approximately $I(\beta) \propto 1/\beta^2$. Hence, from (10), approximately $\ell(\beta) \propto \beta$, so again $h(\beta_{max}^{(d)}) \propto \log(d)$, and the complexity order is still $O\big(d\,[\log d]^2\big)$ as before. We summarise this conclusion as follows.

Corollary 6 *Under the set-up and assumptions of Theorem 2, assuming the above Proportionality Condition and the choice (10), if either* **(a)** *the densities of the two modes of π are in the Exponential Power Family, or* **(b)** *the two modes' first four moments each converge to those of a Gaussian as $d, \beta \to \infty$, then the number of iterations required for the algorithm to converge to stationarity is $O\big(d\,[\log d]^2\big)$.*

In a different direction, the paper [25] introduces a *QuanTA Algorithm*, which modifies parallel tempering's usual temperature-swap moves by adjusting the x space in order to to permit larger moves in the inverse temperature space. As a result of this, the resulting $\ell(\beta)$ function is proportional to $\beta^{k/2}$ for some $k > 2$ (instead of just to β). In this case,

$$h(\beta_{max}^{(d)}) \;=\; \int_1^{\beta_{max}^{(d)}} \frac{1}{\ell(u)}\,du \;\le\; \int_1^{\infty} \frac{1}{\ell(u)}\,du \;\propto\; \int_1^{\infty} u^{-k/2}\,du \;=\; (k/2) - 1 \;<\; \infty\,,$$

so that $h(\beta_{max}^{(d)})$ is $O(1)$ rather than $O(\log d)$. This means that the convergence complexity $O(d[h(\beta_{max}^{(d)}]^2)$ becomes simply $O(d)$, i.e. the $[\log d]^2$ factor vanishes. We summarise this observation as follows.

Corollary 7 *Under the set-up and assumptions of Corollary 6, if we instead run the version of the ALPS algorithm which uses the QuanTA modification of [25], then the number of iterations required for the algorithm to converge to stationarity is simply $O\big(d\big)$.*

Corollaries 6 and 7 are notable since if the modes are well-separated, then ordinary random-walk Metropolis algorithms might converge exponentially slowly. On the other hand, *within* each mode, Metropolis algorithms typically mix in time $O(d)$ [18], so our results say that the ALPS between-mode mixing times are comparable to typical within-mode mixing times.

We close with a remark about generalisations to more than two modes:

Remark 8 *(More than Two Modes.)* For simplicity, all of the above analysis was done assuming a mixture of just $J = 2$ modes. However, a similar analysis works more generally. Indeed, suppose π is a mixture of $J > 2$ modes, of weights $w_1, w_2, \ldots, w_J \ge 0$ where $\sum_{i=1}^{J} w_i = 1$. Then when β_t reaches $\beta_{max}^{(d)}$, the process chooses one of the J modes with probability w_i. This implies that $\{W_t\}$ will converge to a Brownian motion not on $[-\frac{1}{s_2}, \frac{1}{s_1}]$, but rather on a "star" shape with J different line segments all meeting at the origin (corresponding, in the original scaling, to $\beta_{max}^{(d)}$). And, each time W_t reaches the origin, it chooses one of the J line segments with probabilities w_i. This process is called *Walsh's Brownian motion*, see e.g. [3]. (The case $J = 2$ but $w_1 \neq 1/2$ corresponds to skew Brownian motion as above.) For this generalised process, a theorem similar to Theorem 5 can then be proven by similar methods, leading to the same complexity bound of $O\big(d\,[\log d]^2\big)$ iterations (or $O(d)$ iterations if using QuanTA) when $J > 2$ as well.

7 Theorem Proofs

In this section, we prove the theorems stated above.

7.1 Proof of Theorem 2

Since mixing between modes is only possible at $\beta_{max}^{(d)}$, the dynamics for other β will be identical to the single mode case ($J = 1$) as covered in [21]. It therefore follows directly from Theorem 6 of [21] that as $d \to \infty$, the process $\{X_t\}$ converges weakly, at least on $X_t > 0$, to a diffusion limit $\{X_t\}_{t\geq 0}$ satisfying (7). The result for $X_t < 0$ follows similarly.

7.2 Proof of Theorem 3

We assume $x \in (1, 2)$; the proof for $x \in (-2, -1)$ is virtually identical. Here $H_t = h(X_t)$, where $h'(x) = \ell(x)^{-1}$, and $h''(x) = -\ell'(x)\ell(x)^{-2}$. Hence, by Ito's Formula,

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

In this last equation, the second and fourth terms cancel. Also, since $\Phi' = \phi$, it follows from the chain rule that the third term can be written as

$$-\ell^2(X_t)\left[\Phi\left(\frac{-I^{1/2}(X_t)\ell(X_t)}{2}\right)\right]' dt\,.$$

This gives (9). Then, writing everything in terms of $H_t = h(X_t)$, this becomes

$$\begin{aligned}
dH_t &= \left[2\,\Phi\left(\frac{-\ell(h^{-1}(H_t))I^{1/2}(h^{-1}(H_t))}{2}\right)\right]^{1/2} dB_t \\
&\quad +\ell(h^{-1}(H_t))\left[\Phi\left(\frac{-I^{1/2}(h^{-1}(H_t))\ell(h^{-1}(H_t))}{2}\right)\right]' dt\,.
\end{aligned}$$

Now, a diffusion of the form $dH_t = \sigma(H_t)dB_t + \mu(H_t)dt$ has locally invariant distribution π provided that $\frac{1}{2}(\log\pi)'\sigma^2 + \sigma\sigma' = \mu$. That holds for constant π if $\sigma\sigma' = \mu$. In this case, we compute that

$$\begin{aligned}
\sigma\sigma' \;&=\; \frac{1}{2}(\sigma^2)' \;=\; \frac{1}{2}\frac{d}{dH}\left[2\,\Phi\left(\frac{-\ell(h^{-1}(H))I^{1/2}(h^{-1}(H))}{2}\right)\right] \\
&=\; \frac{1}{2}\left(\frac{dH}{dX}\right)^{-1}\frac{d}{dX}\left[2\,\Phi\left(\frac{-\ell(X)I^{1/2}(X)}{2}\right)\right] \\
&=\; \frac{1}{2}\left(\ell(X)^{-1}\right)^{-1}\left[2\,\Phi\left(\frac{-\ell(X)I^{1/2}(X)}{2}\right)\right]' \\
&=\; \ell(X)\left[\Phi\left(\frac{-\ell(X)I^{1/2}(X)}{2}\right)\right]' \;=\; \mu\,,
\end{aligned}$$

thus showing that H leaves constant densities locally invariant.

7.3 Proof of Theorem 5

Let $w_{min}^{(d)} = -\frac{1}{s_2}$ and $w_{max}^{(d)} = \frac{1}{s_1}$ be the endpoints of the domain of W. By Corollary 4, $dH_t = s(H_t)\,dB_t$ in the interior of its domain. Since $W_t = s(H_t)^{-1}\,H_t$, it follows that W_t behaves like Brownian motion on $(-w_{min}^{(d)}, 0)$ and on $(0, w_{max}^{(d)})$. It remains to show that the process converges weakly to skew Brownian motion, including at the boundary points $W_t = 0, w_{min}^{(d)}, w_{max}^{(d)}$. We prove this result using infinitesimal generators, as we now explain.

7.3.1 Method of Proof: Generators

To prove the weak convergence, it suffices by Corollary 8.7 of Chapter 4 of [6] to show (similar to previous proofs of diffusion limits of MCMC algorithms in [18, 19, 4]) that the *infinitesimal generator* $G^{(d)}$ of the process $W^{(d)}$ converges uniformly in x as $d \to \infty$ to the generator G^* of skew Brownian motion, when applied to a *core* $\mathcal{D}$ of functionals, i.e. that

$$\lim_{d\to\infty}\;\sup_{x\in[w_{min}^{(d)},w_{max}^{(d)}]}\left|G^{(d)}f(x) - G^*f(x)\right| \;=\; 0\,,\qquad f\in\mathcal{D}\,,$$

where

$$G^{(d)}f(x) \;:=\; \lim_{\delta\searrow 0}\frac{\mathbf{E}[f(W_\delta^{(d)})\,|\,W_0^{(d)} = x] - f(x)}{\delta}\,.$$

To this end, let $\mathcal{D}$ be the set of all functions $f : [-w_{min}^{(d)}, w_{max}^{(d)}] \to \mathbb{R}$ which are continuous and twice-continuously-differentiable on $[w_{min}^{(d)}, 0]$ and also on $[0, w_{max}^{(d)}]$, with matching one-sided second derivatives $f''^{+}(0) = f''^{-}(0)$, and skewed one-sided first derivatives satisfying $w_1 s_1 f'^{+}(0) = w_2 s_2 f'^{-}(0)$, and $f'(w_{max}^{(d)}) = f'(w_{min}^{(d)}) = 0$. Then it follows from e.g. [12] and Exercise 1.23 of Chapter VII of [17]) that the generator of skew Brownian motion (with excursion weights proportional to $w_1 s_1$ and $w_2 s_2$ respectively, and with reflections at $w_{min}^{(d)}$

and $w_{max}^{(d)}$) satisfies that $G^* f(x) = \frac{1}{2} f''(x)$ for all $f \in \mathcal{D}$, where $f''(0)$ represents the common value $f''^+(0) = f''^-(0)$. Furthermore, $\mathcal{D}$ is clearly dense (in the sup norm) in the set of all $C^2[w_{min}^{(d)}, w_{max}^{(d)}]$ functions, so in the language of [6], $\mathcal{D}$ serves as a core of functions for which it suffices to prove that the generators converge.

It follows from Corollary 4, as discussed above, that for any fixed $f \in \mathcal{D}$,

$$\lim_{d\to\infty} \sup_{w \in \left(w_{min}^{(d)}, w_{max}^{(d)}\right) \setminus \{0\}} |G^{(d)} f(w) - G^* f(w)| = 0\,. \tag{11}$$

That is, the generators do converge uniformly to G^*, as required, at least for $w \neq 0, w_{min}^{(d)}, w_{max}^{(d)}$, i.e. avoiding the mode-hopping value 0 and the reflecting boundaries $w_{min}^{(d)}$ and $w_{max}^{(d)}$. To complete the proof, it suffices to prove that (11) also holds at $w = 0, w_{min}^{(d)}, w_{max}^{(d)}$, i.e. to prove

$$\lim_{d\to\infty} G^{(d)} f(0) \;\equiv\; G^* f(0) \;=\; \frac{1}{2} f''(0)\,, \tag{12}$$

$$\lim_{d\to\infty} G^{(d)} f(w_{min}^{(d)}) \;\equiv\; G^* f(w_{min}^{(d)}) \;=\; \frac{1}{2} f''(w_{min}^{(d)})\,, \tag{13}$$

and

$$\lim_{d\to\infty} G^{(d)} f(w_{max}^{(d)}) \;\equiv\; G^* f(w_{max}^{(d)}) \;=\; \frac{1}{2} f''(w_{max}^{(d)})\,. \tag{14}$$

7.3.2 Verification of (13) and (14)

The proofs of (13) and (14) are virtually identical, so here we prove (14).

If the original inverse-temperature process $\beta_t^{(d)}$ proposes to move in time 1 from inverse-temperature $1 + 0 = 1$ to $1 + \ell(1)d^{-1/2}$, then by (8), the $H_t^{(d)}$ process proposes to move at Poisson rate $[d\, h(\beta_{max}^{(d)})^2]$ from $1 + \frac{0}{h(\beta_{max}^{(d)})} = 1$ to

$$1 + \frac{h\big(1 + \ell(1)d^{-1/2}\big)}{h(\beta_{max}^{(d)})} \;=\; 1 + \frac{1}{h(\beta_{max}^{(d)})} \int_1^{1+\ell(1)d^{-1/2}} \frac{1}{\ell(u)}\, du$$

which to first order as $d \to \infty$ is equal to

$$1 + \frac{1}{h(\beta_{max}^{(d)})} (\ell(1)d^{-1/2}) \frac{1}{\ell(1)} \;=\; 1 + \frac{d^{-1/2}}{h(\beta_{max}^{(d)})}\,.$$

Simultaneously, the $Z_t^{(d)}$ process proposes to move from $2-1 = 1$ to $2-[1+d^{-1/2}/h(\beta_{max}^{(d)})] = 1 - d^{-1/2}/h(\beta_{max}^{(d)})$, and the $W_t^{(d)}$ process proposes to move from $w_{max}^{(d)}$ to

$$(w_{max}^{(d)}) - d^{-1/2}/s_1 h(\beta_{max}^{(d)})\,.$$

Let A be the probability that the original $\beta_t^{(d)}$ process accepts a move from 1 to $1+\ell(1)d^{-1/2}$. Then since $\beta_t^{(d)}$ proposes to move from 1 to $1 + \ell(1)d^{-1/2}$ with probability 1/2, it actually

moves from 1 to $1+\ell(1)d^{-1/2}$ with probability $A/2$, otherwise it stays at 1. So, correspondingly, $W_t^{(d)}$ moves from $w_{max}^{(d)}$ to $(w_{max}^{(d)}) - d^{-1/2}/s_1 h(\beta_{max}^{(d)})$. Furthermore, recall that $W_t^{(d)}$ moves at Poisson rate $[d\,h(\beta_{max}^{(d)})^2]$, so it moves from $w_{max}^{(d)}$ to $(w_{max}^{(d)}) - d^{-1/2}/s_1 h(\beta_{max}^{(d)})$ at rate $[d\,h(\beta_{max}^{(d)})^2](A/2)$. However, we instead consider a minor modification of the process $W_t^{(d)}$ which speeds up time by a factor of 2 whenever it is at $w_{max}^{(d)}$, i.e. it moves from there at Poisson rate $[d\,h(\beta_{max}^{(d)})^2](A)$. This is equivalent to the original $\beta_t^{(d)}$ process always proposing a positive move from 1, instead of proposing either a positive or a negative (always-rejected) move with probability 1/2 each. We show in Section 8 below that this minor modification will not change the limiting distribution of the $W_t^{(d)}$, and thus does not affect the proof.

Thus, to first order as $\delta \searrow 0$ [i.e., up to $o(1)$ errors], our modified process $W_t^{(d)}$ will move from $w_{max}^{(d)}$ to $(w_{max}^{(d)}) - d^{-1/2}/s_1 h(\beta_{max}^{(d)})$ at Poisson rate $[d\,h(\beta_{max}^{(d)})^2](A)$. Hence, setting $x = w_{max}^{(d)} = 1/s_1$, we have that

$$\frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta} \;=\; [d\,h(\beta_{max}^{(d)})^2](A)\Big[f\Big((w_{max}^{(d)}) - d^{-1/2}/s_1 h(\beta_{max}^{(d)})\Big) - f(x)\Big] + o(1)\,.$$

Then, taking a Taylor series expansion around $x = w_{max}^{(d)} = 1/s_1$,

$$\begin{aligned}\frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta} \;&=\; -[d\,h(\beta_{max}^{(d)})^2](A)[d^{-1/2}/s_1 h(\beta_{max}^{(d)})] f'(w_{max}^{(d)}) \\ &\quad +\; \frac{1}{2}[d\,h(\beta_{max}^{(d)})^2](A)[d^{-1/2}/s_1 h(\beta_{max}^{(d)})]^2 f''(w_{max}^{(d)}) \;+\; O(d^{-1/2}) + o(1) \\ &=\; -[A d^{1/2} h(\beta_{max}^{(d)})/s_1]\, f'(w_{max}^{(d)}) \;+\; \frac{1}{2}[A/s_1^2]\, f''(w_{max}^{(d)}) \;+\; O(d^{-1/2}) + o(1)\,,\end{aligned}$$

Since $f \in \mathcal{D}$, we have $f'(w_{max}^{(d)}) = 0$, so the first term vanishes. Furthermore, it is shown in [26] that as $d \to \infty$,

$$A \;\to\; 2\,\Phi\left(\frac{-\ell_0}{2\sqrt{r_1}}\right) \;=\; s_1^2\,.$$

Hence,

$$\frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta} \;=\; 0 \;+\; \frac{1}{2}\,[1]\, f''(w_{max}^{(d)}) \;+\; O(d^{-1/2}) + o(1)\,,$$

so that

$$\lim_{d\to\infty} G^{(d)} f(w_{max}^{(d)}) \;=\; \lim_{d\to\infty} \lim_{\delta\searrow 0} \frac{\mathbf{E}[f(W_\delta^{(d)}) \,|\, W_0^{(d)} = x] - f(x)}{\delta} \;=\; \frac{1}{2} f''(w_{max}^{(d)}) \;=\; G^*(w_{max}^{(d)})\,,$$

as required.

7.3.3 Verification of (12)

To prove (12), note that if the original inverse-temperature process $\beta_t^{(d)}$ proposes to move in time 1 from $\beta_{max}^{(d)}$ to $\beta_{max}^{(d)} - \ell(\beta_{max}^{(d)})d^{-1/2}$ in one of the two modes (with probabilities w_1

and w_2 respectively), then by (8) the $H_t^{(d)}$ process proposes to move at rate $[d\,h(\beta_{max}^{(d)})^2]$ from $1+\frac{h(\beta_{max}^{(d)})}{h(\beta_{max}^{(d)})}=2$ to

$$\pm\left[1+\frac{h\big(\beta_{max}^{(d)}-\ell(\beta_{max}^{(d)})d^{-1/2}\big)}{h(\beta_{max}^{(d)})}\right] \;=\; \pm\left[2-\frac{\int_{\beta_{max}^{(d)}-\ell(\beta_{max}^{(d)})d^{-1/2}}^{\beta_{max}^{(d)}}\frac{1}{\ell(u)}\,du}{h(\beta_{max}^{(d)})}\right]$$

$$\approx\; \pm\left[2-\big(\ell(\beta_{max}^{(d)})d^{-1/2}\big)\,\frac{1}{\ell(\beta_{max}^{(d)})}\right] \;=\; \pm(2-d^{-1/2})\,.$$

Simultaneously, the $Z_t^{(d)}$ process proposes to move from $2-2=0$ to $\pm2-[\pm(2-d^{-1/2})]=\pm d^{-1/2}$, and the $W_t^{(d)}$ process proposes to move from 0 to either $d^{-1/2}/s_1$ or $-d^{-1/2}/s_2$. Hence, similar to the above (but without the minor modification), with $x=0$ we have to first order as $\delta\searrow 0$ that

$$\begin{aligned}&\frac{\mathbf{E}[f(W_\delta)\,|\,W_0=x]-f(x)}{\delta}\\ &\quad=\;[d\,h(\beta_{max}^{(d)})^2]\left(w_1\alpha_1\left[f\left(d^{-1/2}/s_1\right)-f(0)\right]\;+\;w_2\alpha_2\left[f\left(-d^{-1/2}/s_2\right)-f(0)\right]\right)+o(1)\,,\end{aligned}\tag{15}$$

where α_i is the acceptance probability for the original process to accept a proposal to increase the inverse-temperature from $\beta_{max}^{(d)}$ to $\beta_{max}^{(d)}-\ell(\beta_{max}^{(d)})d^{-1/2}$ in mode i. Now, the argument in [26] shows that as $d\to\infty$ we have

$$\alpha_i\;\to\;2\,\Phi\left(\frac{-\ell_0}{2\sqrt{r_i}}\right)\;=\;s_i^2\,,\qquad i=1,2\,.$$

Hence, taking a Taylor series expansion around $x=0$, we obtain from (15) that

$$\begin{aligned}&\frac{\mathbf{E}[f(W_\delta)\,|\,W_0=x]-f(x)}{\delta}\\ &\quad=\;d\,w_1s_1^2\left(d^{-1/2}/s_1\right)f'^{+}(0)+\;\frac{1}{2}d\,w_1s_1^2\left(d^{-1/2}/s_1\right)^2f''^{+}(0)\;+\;O(d\,d^{-3/2})+o(1)\\ &\quad-\;d\,w_2s_2^2\;\left(d^{-1/2}/s_2\right)f'^{-}(0)+\;\frac{1}{2}d\,w_2s_2^2\left(d^{-1/2}/s_2\right)^2f''^{-}(0)+\;O(d\,d^{-3/2})+o(1)\\ &\quad=\;d^{1/2}[w_1s_1f'^{+}(0)-w_2s_2f'^{-}(0)]\;+\;\frac{1}{2}[w_1f''^{+}(0)+w_2f''^{-}(0)]\;+\;O(d^{-1/2})+o(1)\,.\end{aligned}$$

Now, by the definition of $f\in\mathcal{D}$, $w_1s_1f'^{+}(0)-w_2s_2f'^{-}(0)=0$, and $w_1f''^{+}(0)+w_2f''^{-}(0)=(w_1+w_2)f''(0)=f''(0)$. Hence, we obtain finally that

$$\frac{\mathbf{E}[f(W_\delta)\,|\,W_0=x]-f(x)}{\delta}\;=\;\frac{1}{2}f''(0)+O(d^{-1/2})+o(1)\,,$$

so that

$$\lim_{d\to\infty}G^{(d)}f(0)\;=\;\lim_{d\to\infty}\lim_{\delta\searrow0}\frac{\mathbf{E}[f(W_\delta^{(d)})\,|\,W_0^{(d)}=x]-f(x)}{\delta}\;=\;\frac{1}{2}f''(0)\;=\;G^*(0)\,.$$

This establishes (12), and hence completes the proof of Theorem 5.

8 Modified Processes and Occupation Times

Recall that the proof of (14) in Section 7.3.2 above was actually for a minor modification of the process $W_t^{(d)}$, which speeds up time by a factor of 2 whenever it is in the state $w_{max}^{(d)}$. We now argue that this minor modification does not affect the limiting distribution. Indeed, since the modification corresponds to adjusting the rate of time, we can write the modified process as $\widehat{W}_t^{(d)} \equiv W_{\tau_d(t)}^{(d)}$, where $\tau_d(t)$ is the time scale including the occasional speedups. Clearly $\lim_{t \searrow 0} \tau_d(t) = 0$. Also, it follows from Proposition 12 below that the fraction of time that the original process spends at $w_{max}^{(d)}$ converges to 0 as $d \to \infty$. This implies that $\lim_{d\to\infty}(\tau_d(t)/t) = 1$. Since our process $W_t^{(d)}$ is continuous, this means that $\lim_{d\to\infty} |f(W_{\tau_d(t)}^{(d)}) - f(W_t^{(d)})| = 0$. That is, the two processes have the same limiting behaviour as $d \to \infty$. So, the diffusion limit is not affected by making our minor modification as above.

It remains to state and prove Proposition 12. We begin with a result about limiting probabilities for reflecting simple symmetric random walk.

Proposition 9 *Let $\{Y_n\}$ be reflecting simple symmetric random walk on the state space $\{0, 1, 2, \ldots, m\}$, i.e. a discrete-time birth-death Markov chain with transition probabilities $p_{i,i+1} = p_{i,i-1} = 1/2$ for $1 \leq i \leq m-1$, and $p_{0,1} = p_{m,m-1} = 1$. Then for all $m \in \mathbf{N}$ and all sufficiently large $n \in \mathbf{N}$, $\mathbf{P}(Y_n = 0) \leq (2/\sqrt{n}) + (1/m)$. Hence, $\lim_{n,m\to\infty} \mathbf{P}(Y_n = 0) = 0$.*

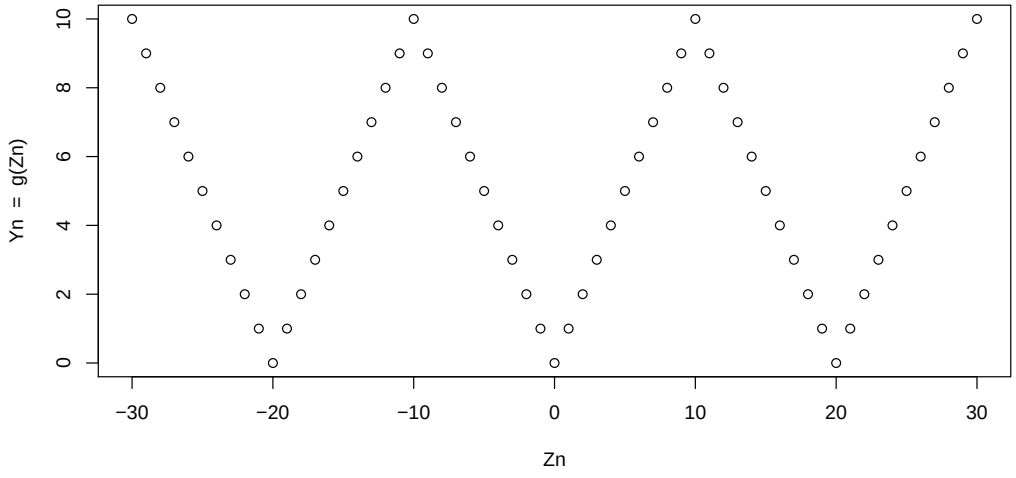

Figure 4: The lifting transformation function "g" (when $m = 10$).

Proof: We condition on $Y_0 = y$; the general case then follows by taking expectation with respect to Y_0. We "lift" $\{Y_n\}$ to $\mathbf{Z}$ by writing $Y_n = g(Z_n)$, where $\{Z_n\}$ is simple symmetric

random walk on *all* the integers $\mathbf{Z}$, and $g(z) = \min_j |z - 2jm|$ (see Figure 4). Then

$$\mathbf{P}_y[Y_n = 0] \;=\; \mathbf{P}_y[g(Z_n) = 0] \;=\; \sum_{j\in\mathbf{Z}} \mathbf{P}_y[Z_n = 2jm]$$

$$=\; \sum_{j\in\mathbf{Z}} \mathbf{P}_y\Big[\text{Binomial}(n,1/2) = \frac{n}{2} + \frac{y}{2} + jm\Big] \;=\; \sum_{j\in\mathbf{Z}} h\left(\frac{n}{2} + \frac{y}{2} + jm\right),$$

where $h(k) = \mathbf{P}[\text{Binomial}(n,1/2) = k]$. Now, h is maximised when $k = n/2$ (or $(n \pm 1)/2$ if n is odd), and decreases monotonically on either side of that. Hence, find $j_* \in \mathbf{N}$ with $\frac{y}{2} + (j_* - 1)m < 0 \le \frac{y}{2} + j_* m$. It follows from Stirling's Approximation (see e.g. [23]) that to first order as $n, k, n-k \to \infty$,

$$\mathbf{P}[\text{Binomial}(n,1/2) = k] \;\le\; e^{-2n[\frac{1}{2} - \frac{k}{n}]^2} \sqrt{1/2\pi k[1 - (k/n)]}\,,$$

so in particular

$$h\left(\frac{n}{2} + \frac{y}{2} + j_* m\right) \;\le\; \sqrt{2/\pi n} + o_n(1) \;\le\; 1/\sqrt{n}$$

for all sufficiently large n, and similarly for $h\left(\frac{n}{2} + \frac{y}{2} + (j_* - 1)m\right)$. Then, by monotonicity, we have for $j > j_*$ that

$$h\left(\frac{n}{2} + \frac{y}{2} + jm\right) \;\le\; \frac{1}{m}\Big[h\left(\frac{n}{2} + \frac{y}{2} + (j-1)m + 1\right) + h\left(\frac{n}{2} + \frac{y}{2} + (j-1)m + 2\right) + \ldots + h\left(\frac{n}{2} + \frac{y}{2} + jm\right)\Big].$$

Hence,

$$\sum_{j>j_*} h\left(\frac{n}{2} + \frac{y}{2} + jm\right) \;\le\; \frac{1}{m}\Big[h\left(\frac{n}{2} + \frac{y}{2} + 1\right) + h\left(\frac{n}{2} + \frac{y}{2} + 2\right) + h\left(\frac{n}{2} + \frac{y}{2} + 3\right) + \ldots\Big].$$

But $\sum_k h(k) = 1$, so by symmetry $\sum_{k>n/2} h(k) \le 1/2$, and so

$$h\left(\frac{n}{2} + \frac{y}{2} + 1\right) + h\left(\frac{n}{2} + \frac{y}{2} + 2\right) + h\left(\frac{n}{2} + \frac{y}{2} + 3\right) + \ldots \le 1/2\,.$$

Thus,

$$\sum_{j>j_*} h\left(\frac{n}{2} + \frac{y}{2} + jm\right) \;\le\; \frac{1}{2m}\,.$$

Similarly,

$$\sum_{j<j_*-1} h\left(\frac{n}{2} + \frac{y}{2} + jm\right) \;\le\; \frac{1}{2m}\,.$$

Therefore, for all sufficiently large n,

$$\sum_{j\in\mathbf{Z}} h\left(\frac{n}{2} + \frac{y}{2} + jm\right) \;\le\; (1/\sqrt{n}) + (1/\sqrt{n}) + \frac{1}{2m} + \frac{1}{2m} \;=\; (2/\sqrt{n}) + (1/m)\,,$$

as claimed. □

Remark 10 Similar arguments show that $\lim_{n,m\to\infty} \mathbf{P}(Y_n = z) = 0$ for any fixed number $z \in \mathbf{N}$, by replacing "$Z_n = 2jm$" by "$Z_n = 2jm + z$", and "$\frac{n}{2} + \frac{y}{2}$" by "$\frac{n}{2} + \frac{y}{2} - \frac{z}{2}$", throughout the proof, though we do not use that fact here.

Corollary 11 *Let $\{Y_n\}$ be as in Proposition 9. Let $N_0 = \#\{i : 0 \le i \le n-1,\ Y_i = 0\}$ be the occupation time of the state 0 before time n. Then as $n, m \to \infty$, the average occupation time N_0/n converges to 0 in probability.*

Proof: Let $I_i = \mathbf{1}_{Y_i=0}$ be the indicator function of the event $Y_i = 0$. Then by Proposition 9, $\lim_{n,m\to\infty} \mathbf{E}[I_n] = \lim_{n,m\to\infty} \mathbf{P}[Y_n = 0] = 0$. Hence, using the theory of Cesàro sums,

$$\lim_{n,m\to\infty} \mathbf{E}[N_0/n] \;=\; \lim_{n,m\to\infty} \mathbf{E}\Big[\sum_{i=0}^{n-1} I_i\Big]/n \;=\; \lim_{n,m\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mathbf{E}[I_i] \;=\; \lim_{n,m\to\infty} \mathbf{E}[I_n] \;=\; 0\,.$$

Hence, by Markov's inequality, since $N_0/n \ge 0$, for any $\epsilon > 0$ we have

$$\lim_{n,m\to\infty} \mathbf{P}[(N_0/n) > \epsilon] \;\le\; \lim_{n,m\to\infty} \mathbf{E}[N_0/n]/\epsilon \;=\; 0\,,$$

so that $N_0/n \to 0$ in probability, as claimed. □

Proposition 12 *Let $\{X_n\}$ be a discrete-time birth-death Markov chain on the state space $\{0, 1, 2, \ldots, m\}$, with transition probabilities satisfying that $p_{i,j} = 0$ whenever $|j - i| \ge 2$, $p_{i,i+1} = p_{i,i-1}$ for all $1 \le i \le m-1$, and $p_{i,i} \le 1 - a$ for some fixed constant $a > 0$. Let $N_0 = \#\{i : 0 \le i \le n-1,\ X_i = 0\}$. Then as $n, m \to \infty$, N_0/n converges to 0 in probability.*

Proof: Let $\{J_k\}$ be the *jump chain* of $\{X_n\}$, i.e. the Markov chain which copies $\{X_n\}$ except omitting immediate repetitions of the same state, and let $\{M_k\}$ count the number of repetitions. [For example, if the original chain $\{X_n\}$ began $\{X_n\} = (a, b, b, b, a, a, c, c, c, c, d, d, a, \ldots)$, then the jump chain $\{J_k\}$ would begin $\{J_k\} = (a, b, a, c, d, a, \ldots)$, and the corresponding multiplicity list $\{M_k\}$ would begin $\{M_k\} = (1, 3, 2, 4, 2, \ldots)$.] Then the assumptions imply that $\{J_k\}$ has the transition probabilities of reflecting simple symmetric random walk, as in Proposition 9 and Corollary 11 above.

Now, let $K(n)$ be the smallest integer with $M_1 + \ldots + M_{K(n)} \ge n$. Given J_k, the random variable M_k has the Geometric$(1 - p_{J_k J_k})$ distribution, so it is stochastically bounded above by the Geometric(a) distribution, from which it follows that $\lim_{n\to\infty} K(n) = \infty$ w.p. 1. Let $C_s = \#\{i : 0 \le i \le K(n),\ J_i = s\}$. Then Corollary 11 implies that $\lim_{n,m\to\infty}(C_0/K(n)) = 0$. On the other hand, N_0 is $\le$ a sum of C_0 independent Geometric$(1 - p_{00})$ random variables, so $\mathbf{E}[N_0 \,|\, C_0] = C_0/(1 - p_{00}) \le C_0/a$, and $\mathbf{P}[N_0 > 2C_0/a \,|\, C_0] \to 0$ as $n \to \infty$. Also, $M_1 + \ldots + M_{K(n)-1} \le n$, and each $M_i \ge 1$, so $n \ge K(n) - 1$. We therefore conclude that

$$\lim_{n,m\to\infty} \frac{N_0}{n} \;\le\; \lim_{n,m\to\infty} \frac{2C_0/a}{K(n)-1} \;=\; (2/a) \lim_{n,m\to\infty} \frac{C_0}{K(n)} \;=\; 0\,,$$

as claimed. □

Acknowledgements. We thank Alex Mijatovic and Neal Madras for very helpful comments related to Section 8 herein.

References

[1] Emile Aarts and Jan Korst. *Simulated Annealing and Boltzmann Machines*. New York, NY; John Wiley and Sons Inc., 1988.

[2] Yves F Atchadé, Gareth O. Roberts, and Jeffrey S. Rosenthal. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statistics and Computing*, 21(4):555–568, 2011.

[3] Martin T. Barlow, Jim Pitman, and Marc Yor. On Walsh's Brownian motions. *Séminaire de probabilités (Strasbourg)*, 23:275–293, 1989.

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

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

[6] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986.

[7] Charles J Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics*, 23:156–163, 1991.

[8] W. Keith Hastings. Monte Carlo Sampling Methods Using Markov chains and their Applications. *Biometrika*, 57(1):97–109, 1970.

[9] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[10] Aminata Kone and David A Kofke. Selection of Temperature Intervals for Parallel-Tempering Simulations. *The Journal of Chemical Physics*, 122(20):206101, 2005.

[11] Antoine Lejay. On the constructions of the skew Brownian motion. *Probability Surveys*, 3:413–466, 2006.

[12] Thomas M. Liggett. *Continuous Time Markov Processes: An Introduction*. American Mathematical Society, 2010.

[13] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

[14] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[15] M. Pincus. A Monte-Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems. *Journal of the Operations Research Society of America*, 18(6):967–1235, 1970.

[16] Cristian Predescu, Mihaela Predescu, and Cristian V Ciobanu. The Incomplete Beta Function Law for Parallel Tempering Sampling of Classical Canonical Systems. *The Journal of Chemical Physics*, 120(9):4119–4128, 2004.

[17] Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Springer, 3rd edition, 2004.

[18] Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

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

[20] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[21] Gareth O. Roberts and Jeffrey S. Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *The Annals of Applied Probability*, 24(1):131–149, 2014.

[22] Jeffrey S. Rosenthal. Quantitative convergence rates of Markov chains: A simple account. *Electronic Communications in Probability*, 7(13):123–128, 2002.

[23] Jeffrey S. Rosenthal. Maximum Binomial Probabilities and Game Theory Voter Models. *Advances and Applications in Statistics, to appear*, 2020.

[24] Nicholas G. Tawn, Sigurd Assing, Matt Moores, and Gareth O. Roberts. The Annealed Leap-Point Sampler (ALPS) for Multimodal Target Distributions. *In preparation*, 2020.

[25] Nicholas G. Tawn and Gareth O. Roberts. Accelerating Parallel Tempering: Quantile Tempering Algorithm (QuanTA). *Applied Probability Trust, to appear*, 2018.

[26] Nicholas G. Tawn, Gareth O. Roberts, and Jeffrey S. Rosenthal. Weight-preserving simulated tempering. *Statistics and Computing*, 30:27–41, 2020.