J. Appl. Prob. **58,** 83–105 (2021)
*doi:*10.1017/jpr.2020.83

SIMPLE CONDITIONS FOR METASTABILITY OF CONTINUOUS MARKOV CHAINS

OREN MANGOUBI,* *Worcester Polytechnic Institute*
NATESH PILLAI,** *Harvard University*
AARON SMITH,*** *University of Ottawa*

Abstract

A family $\{Q_\beta\}_{\beta\geq 0}$ of Markov chains is said to exhibit *metastable mixing* with *modes* $S_\beta^{(1)}, \ldots, S_\beta^{(k)}$ if its spectral gap (or some other mixing property) is very close to the worst conductance $\min\left(\Phi_\beta\left(S_\beta^{(1)}\right), \ldots, \Phi_\beta\left(S_\beta^{(k)}\right)\right)$ of its modes for all large values of β. We give simple sufficient conditions for a family of Markov chains to exhibit metastability in this sense, and verify that these conditions hold for a prototypical Metropolis–Hastings chain targeting a mixture distribution. The existing metastability literature is large, and our present work is aimed at filling the following small gap: finding sufficient conditions for metastability that are easy to verify for typical examples from statistics using well-studied methods, while at the same time giving an asymptotically exact formula for the spectral gap (rather than a bound that can be very far from sharp). Our bounds from this paper are used in a companion paper (O. Mangoubi, N. S. Pillai, and A. Smith, arXiv:1808.03230) to compare the mixing times of the Hamiltonian Monte Carlo algorithm and a random walk algorithm for multimodal target distributions.

Keywords: Metastability; Markov chain Monte Carlo (MCMC); spectral gap; multimodal distribution.

2010 Mathematics Subject Classification: Primary 60J05
Secondary 65C40

1. Introduction

It is well known that Markov chains targeting multimodal distributions, such as those that appear in mixture models, will often mix very slowly. Of course, some algorithms are still faster than others, and the present paper is motivated by the problem of comparing different MCMC (Markov chain Monte Carlo) algorithms in this 'highly multimodal' regime. We provide a step in this direction by finding some simple sufficient conditions under which we can find an explicit formula for the spectral gap for MCMC algorithms on multimodal target distributions. To be slightly more precise, we consider a sequence of Markov transition kernels $\{Q_\beta\}_{\beta\geq 0}$ with state space Ω partitioned into pieces $\Omega = \sqcup_{i=1}^k S_\beta^{(i)}$. One of our main results, Lemma 2,

Received 7 May 2019; revision received 15 June 2020.
* Postal address: Worcester Polytechnic Institute, 100 Institute Road, Worcester, Massachusetts, USA. Email address: omangoubi@gmail.com
** Postal address: Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA.
*** Postal address: Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa ON K1N 7N5, Canada.

gives sufficient conditions under which the spectral gap λ_β of Q_β is asymptotically given by the worst-case conductance $\Phi_{\min}(\beta) = \min\left(\Phi_\beta\left(S_\beta^{(1)}\right), \ldots, \Phi_\beta\left(S_\beta^{(k)}\right)\right)$, in the sense

$$\lim_{\beta\to\infty} \frac{\log(\lambda_\beta)}{\log(\Phi_{\min}(\beta))} = 1. \tag{1.1}$$

The main heuristic behind our calculations is that, in the highly multimodal regime, a Markov chain with strongly multimodal stationary distribution will mix *within* its starting mode before travelling between modes. When this occurs, we say that the Markov chain exhibits *metastable behaviour*, and the mixing properties of the Markov chain are often determined by the rate of transition between modes at stationarity (see, e.g., [2] and the references therein for an introduction to metastability). As a prototypical example, we consider the simple mixture of two Gaussians,

$$\pi_\sigma = \frac{1}{2}\mathcal{N}\left(-1, \sigma^2\right) + \frac{1}{2}\mathcal{N}\left(1, \sigma^2\right), \tag{1.2}$$

for $\sigma > 0$. When σ is close to 0, the usual tuning heuristic for the random walk Metropolis–Hastings (MH) algorithms (see, e.g., [19]) suggests using a proposal distribution with standard deviation on the order of σ, such as

$$K_\sigma(x, \cdot) = \mathcal{N}(x, \sigma^2).$$

Informally, an MH chain $\{X_t\}_{t\geq 0}$ with proposal distribution K_σ, target distribution π_σ, and initial point $X_0 \in [-2, -0.5]$ in one of the modes will evolve according to the following three stages:

1. For t very small, the law of the chain X_t, $\mathcal{L}(X_t)$, will depend quite strongly on the starting point X_0.
2. For $\sigma^{-1} \ll t \ll \mathrm{e}^{c_1\sigma^{-2}}$ and $c_1 > 0$ small, the chain will have mixed very well on its first mode and is very unlikely to have ever left the interval $(-\infty, -0.1)$, so that
$$\|\mathcal{L}(X_t) - \mathcal{N}\left(-1, \sigma^2\right)\|_{\mathrm{TV}} \ll \mathrm{e}^{-c_2\sigma^{-1}}$$
for some $c_2 > 0$. Note that $\mathcal{L}(X_t)$ is close to $\mathcal{N}\left(-1, \sigma^2\right)$, which is *not* its stationary measure π_σ.
3. For $t \gg \mathrm{e}^{c_3\sigma^{-2}}$, the chain will have mixed well on the entire state space in the sense that
$$\|\mathcal{L}(X_t) - \pi_\sigma\|_{\mathrm{TV}} \ll \mathrm{e}^{-c_3\sigma^{-1}}$$
for some $c_3 > 0$.

In the context of this example, the main result of our work is a straightforward way to verify that there is a sharp transition around $t \approx \mathrm{e}^{\frac{1}{2}\sigma^{-2}}$, so that we may take $c_1 = c_3 = \frac{1}{2}$ in this heuristic description (see Theorems 1 and 2 for a precise statement). In the notation of (1.1), we can take the parameter β that indexes our chains to be equal to σ^{-1}. We view β as indexing 'how multimodal' a chain is, while in this particular example σ^{-1} measures both the width of each mode *and* how well separated they are.

We believe that these scaling exponents c_1, c_3 are natural ways to measure performance in the highly multimodal regime, as they seem to capture the most important differences in algorithm performance; see our companion paper [11] for further discussion of this point and relationships to the literature on optimal scaling and lifted Markov chains.

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

1.1. Related work

Our work is closely related to two large pieces of the Markov chain literature: *decomposition bounds* (see, e.g., [5, 10, 17, 23, 24]) and *metastability bounds* (see, e.g., the popular books [2, 16] and the references within the recent articles [1, 6]). It is far beyond the scope of the present article to give a comprehensive survey. Instead, we describe the small gap in the literature that the present article fills.

Our main goal was to find sufficient conditions for metastability that are both easy to verify for typical statistical examples, and still give an asymptotically exact formula for the spectral gap. Such bounds would allow a (relatively) simple way to compare the asymptotic performance of different algorithms. Despite the size of the literature on metastability, we were not able to find results that met both of these criteria.

Much of the existing work on decomposition bounds applies to examples from statistics in a straightforward way ([23, 24] are both applied in this context, but the other references above can be extended to this setting as well). However, because the resulting bound involves a *product* of two terms that can be small, it is not straightforward to use these bounds to actually compute the asymptotics of the spectral gap—we merely obtain bounds.

The existing work from the metastability community tends to give much sharper bounds, and many results on metastability can be used to compute the asymptotics of the spectral gap. However, these results are typically not aimed at a statistical audience, and they can be difficult to apply to the Markov chains that appear in computational statistics. Most Markov chains used in this setting are discrete-time chains on unbounded, continuous state spaces; they are also typically geometrically ergodic but not uniformly ergodic. To our knowledge, the most recent large-scale text on metastability [2] does not directly give bounds on *any* Markov chains in this setting. Other recent surveys also seem to omit this setting.

Of course, this does not mean that the ideas in the book (and the larger literature on metastability) cannot be applied at all. Some of the classical approaches to metastability can be extended to this setting without much difficulty, but the methods that seem simplest to extend are based on assumptions that seem difficult to verify for statistical examples. For example, approaches based on the analysis of 'typical' trajectories between modes may require a detailed understanding of these trajectories, and it is usually quite difficult to compute these trajectories.

Our results allow the computation of the asymptotics of the spectral gap, and the assumptions are stated in terms that should be familiar to the statistics community. In particular, our assumption about control of the tails is stated in terms of control of hitting times and a Lyapunov condition, which are quite similar to conditions in the dominant drift-and-minorization approach to convergence analysis of Markov chains in statistics [15, 21]. While the basic idea of metastability is not new, we believe that such 'in-between' results are a useful compromise that focuses on the most relevant properties for comparison of Markov chains targeting multimodal distributions that arise in a statistical context. Although our introduction focuses on a simple Metropolis–Hastings algorithm with a one-dimensional stationary measure based on Gaussians, the main conditions we study here are also fairly easy to verify for a wide variety of other Markov chains. Towards the end of this paper we illustrate metastability with a simple example related to a high-dimensional Gibbs-like sampler for the problem of estimating the volume of a set, and our companion paper [11] uses our main result to prove metastability for the Hamiltonian Monte Carlo algorithm in a mixture setting.

1.2. Guide to the paper

In Section 2 we review the basic notation and definitions, and also provide some simple bounds. Our main results on metastability are presented in Section 3. Finally, we

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

give an illustrative one-dimensional Gaussian application in Section 4 and an illustrative high-dimensional Gibbs-like sampler in Section 5.

2. Preliminaries

2.1. Basic notation

Throughout the remainder of the paper we denote by π the smooth density function of a probability distribution on a convex subset of $\mathbb{R}^d$. We denote by $\mathcal{L}(X)$ the distribution of a random variable X. Similarly, if μ is a probability measure, we write '$X \sim \mu$' for 'X has distribution μ'. We use $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ to denote probabilities and expectations. We will often consider a Markov chain $\{X_t\}_{t\geq 0}$ sampled from a transition kernel L and with initial distribution $X_0 \sim \mu$; when we wish to emphasize the role of the starting distribution, we will write $\mathbb{P}_\mu[\cdot]$ and $\mathbb{E}_\mu[\cdot]$. In a slight abuse of notation, we write $\mathbb{P}_x[\cdot]$ and $\mathbb{E}_x[\cdot]$ in the special case that μ is a point mass concentrated at $x \in \Omega$.

For two nonnegative functions or sequences f,g, we write $f = O(g)$ as shorthand for the statement: there exist constants $0 < C_1, C_2 < \infty$ such that, for all $x > C_1$, $f(x) \leq C_2\, g(x)$. We write $f = \Omega(g)$ for $g = O(f)$, and we write $f = \Theta(g)$ if both $f = O(g)$ and $g = O(f)$. Relatedly, we write $f = o(g)$ as shorthand for the statement: $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$. Finally, we write $f = \tilde{O}(g)$ if there exist constants $0 < C_1, C_2, C_3 < \infty$ such that, for all $x > C_1$, $f(x) \leq C_2\, g(x) \log(x)^{C_3}$, and write $f = \tilde{\Omega}(g)$ for $g = \tilde{O}(f)$. As shorthand, we say that a function f is 'bounded by a polynomial' if there exists a polynomial g such that $f = O(g)$.

2.2. Cheeger's inequality and the spectral gap

We recall the basic definitions used to measure the efficiency of MCMC algorithms. Let L be a reversible transition kernel with unique stationary distribution μ on $\mathbb{R}^d$. It is common to view L as an operator from $L_2(\pi)$ to itself via the formula

$$(Lf)(x) = \int_{y\in\mathbb{R}^d} L(x, \mathrm{d}y) f(y).$$

The constant function is always an eigenfunction of this operator, with eigenvalue 1. We define the space $W^\perp = \left\{f \in L_2(\mu)\colon \int_x f(x)\mu(\mathrm{d}x) = 0\right\}$ of functions that are orthogonal to the constant function, and denote by $L^\perp$ the restriction of the operator L to the space $W^\perp$. We then define the *spectral gap* ρ of L by the formula

$$\rho = 1 - \sup\left\{|\lambda| : \lambda \in \mathrm{Spectrum}(L^\perp)\right\},$$

where Spectrum refers to the usual spectrum of an operator. If $L^\perp$ has a largest eigenvalue $|\lambda|$ (for example, if L is a matrix of a finite-state-space Markov chain), then $\rho = 1 - |\lambda|$.

Cheeger's inequality [3, 7] provides bounds for the spectral gap in terms of the ability of L to move from any set to its complement in a single step. This ability is measured by the conductance Φ, which is defined by the pair of equations

$$\Phi = \inf_{S\in\mathcal{A}\,:\, 0<\mu(S)<\frac{1}{2}} \Phi(S),$$

$$\Phi(S) = \frac{\int_x \mathbb{1}\{x \in S\} L(x, S^{\mathrm{c}})\mu(\mathrm{d}x)}{\mu(S)},$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

where $\mathcal{A} = \mathcal{A}(\mathbb{R}^d)$ denotes the usual collection of Lebesgue-measurable subsets of $\mathbb{R}^d$. Cheeger's inequality for reversible Markov chains, first proved in [7], gives

$$\frac{\Phi^2}{2} \leq \rho \leq 2\Phi. \tag{2.1}$$

2.3. Traces and hitting times

We recall some standard definitions related to Markov processes.

Definition 1. (*Trace and restriction chains.*) Let L be the transition kernel of a reversible ergodic Markov chain on state space Ω with stationary measure μ, and let $S \subset \Omega$ be a measurable subset with $\mu(S) > 0$. Let $\{X_t\}_{t\geq 0}$ be a Markov chain evolving according to L, and iteratively define

$$c_0 = \inf\{t \geq 0 : X_t \in S\},$$
$$c_{i+1} = \inf\{t > c_i : X_t \in S\}.$$

Then $\hat{X}_t = X_{c_t}$, $t \geq 0$, is the *trace of* $\{X_t\}_{t\geq 0}$ *on* S. Note that $\{\hat{X}_t\}_{t\geq 0}$ is a Markov chain with state space S, and so this procedure also defines a transition kernel with state space S. We call this kernel the *trace of the kernel* L *on* S.

Similarly, define the restriction $\mu|_S$ of μ to S by

$$\mu|_S(A) = \frac{\mu(S \cap A)}{\mu(S)}$$

for measurable $A \subset \Omega$. Define the restriction $L|_S$ of L to S to be the Metropolis–Hastings kernel with proposal distribution L and target distribution $\mu|_S$, that is, the transition kernel given by $L|_S(x, A) = L(x, A)$ for all $x \in S$ and measurable $A \subset S\backslash\{x\}$, and by $L|_S(x, \{x\}) = L(x, \{x\}) + L(x, S^c)$. We call $L|_S$ the *restriction of the kernel* L *to* S.

We note that both the *trace* and *restriction* kernels are defined here purely to help us write conditions that are easier to verify in practice. In particular, although $L|_S$ is a Metropolis–Hastings kernel, we do not assume in any sense that the kernel L itself is a practical Metropolis–Hastings algorithm in statistics.

Definition 2. (*Hitting time.*) Let $\{X_t\}_{t\geq 0}$ be a Markov chain with initial point $X_0 = x$ and let S be a measurable set. Then $\tau_{x,S} = \inf\{t \geq 0 : X_t \in S\}$ is called the *hitting time* of S from x. When the starting point $X_0 = x$ is already fixed, we sometimes use the shorthand τ_S.

3. Generic metastability bounds

Denote by $\{Q_\beta\}_{\beta\geq 0}$ the transition kernels of a collection of reversible ergodic Markov chains with stationary measures $\{\pi_\beta\}_{\beta\geq 0}$ on common state space Ω, which we take to be a convex subset of $\mathbb{R}^d$. Throughout the remainder of the paper we will always use the subscript β to indicate which chain is being used; for example, $\Phi_\beta(S)$ is the conductance of the set S with respect to the chain Q_β, ρ_β is the spectral gap of Q_β, and so on.

Our two main results are:

1. In Lemma 1, we fix a set $S \subset \Omega$ and give sufficient conditions for the *worst-case* hitting time of S^c from S to be bounded by the *average-case* hitting time $\Phi_\beta(S)$.

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

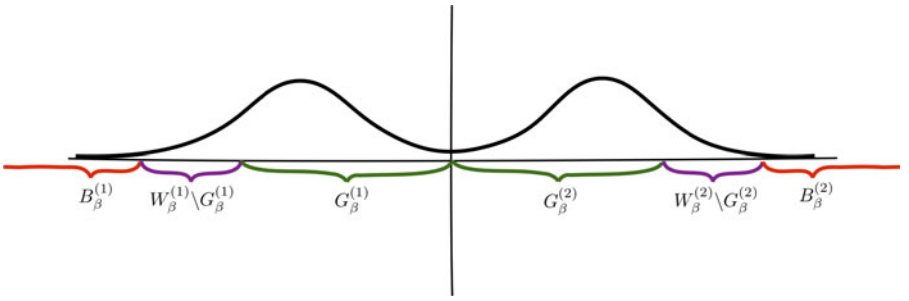

FIGURE 1. A cartoon plot of the target density f_σ with the regions illustrated. Note that we have substantially compressed the regions so that they are all visible; in a scale drawing, B_β would not be visible.

2. In Lemma 2, we consider sufficient conditions on the entire partition $S^{(1)}, \ldots, S^{(k)}$ to ensure that the *spectral gap* of Q_β is approximately equal to the *worst-case conductance* $\min_{1\le i\le k} \Phi_\beta(S^{(i)})$.

3.1. Metastability and hitting times

The main point of our first set of assumptions is to guarantee that the Markov chain cannot get 'stuck' for a long time before mixing within a mode S. Fix $S \subset \Omega$ with $\inf_{\beta\ge 0} \pi_\beta(S) \equiv c_1 > 0$. We will also define a sequence of sets $G_\beta, B_\beta, W_\beta \subset S$ indexed by $\beta \ge 0$ that satisfy $G_\beta \subset W_\beta$ and $B_\beta \subset W_\beta^{\mathrm{c}} \cap S$.

In the following assumption we think of the set G_β as the points that are 'deep within' the mode S, the points B_β as the points that are 'far in the tails' of the target distribution, and the 'covering set' W_β as a way of separating these two regions. To help with intuition, a reasonable choice of these sets in the special case of the mixture of Gaussians in (1.2) is illustrated in Figure 1. These sets are $S = (-\infty, 0)$, $G_\beta = (-\sigma^{-9}, 0)$, $W_\beta = (-\sigma^{-10}, 0)$, and $B_\beta = (-\infty, -\sigma^{-10}]$. Note that Figure 1 shows *two* collections of these sets. The first covers S; the second is associated with S^{c}, which is needed by Lemma 2 below. Also note that, as this example suggests, there is a great deal of flexibility in designing these sets; changing, e.g., 9 to 11.3 and 10 to 11.7 in these definitions would not substantially change our analysis.

Our assumptions are as follows.

Assumptions 1. *We assume the following all hold for $\beta > \beta_0$ sufficiently large:*

1. *Small conductance: There exists some $c > 0$ such that $\Phi_\beta(S) \le \mathrm{e}^{-c\beta}$.*

2. *Rapid mixing within G_β: Let $\hat{Q}_\beta$ be the restriction of Q_β to S. There exists some function r_1 bounded by a polynomial such that*

$$\sup_{x\in G_\beta} \left\| \hat{Q}_\beta^{r_1(\beta)}(x, \cdot) - \pi_\beta|_S(\cdot) \right\|_{\mathrm{TV}} \le \beta^{-2}\Phi_\beta(S). \tag{3.1}$$

3. *Never stuck in $W_\beta \backslash G_\beta$: There exists some function r_2 bounded by a polynomial such that*

$$\sup_{x\in W_\beta\backslash G_\beta} \mathbb{P}_x[\tau_{G_\beta\cup S^{\mathrm{c}}} > r_2(\beta)] \le \beta^{-2}\Phi_\beta(S).$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

4. *Never hitting W_β^c: We have*

$$\sup_{x\in G_\beta} \mathbb{P}_x\big[\tau_{W_\beta^c} < \min(r_1(\beta)+r_2(\beta)+1,\, \tau_{S^c})\big] \le \Phi_\beta(S)^4. \tag{3.2}$$

It is important to note that the first inequality in (3.2) is strict; one use of this assumption is to check that $\tau_{W_\beta^c} = \tau_{S^c}$ occurs with high probability for 'typical' starting points in G_β.

Remark 1. As is common in the metastable literature, there are a wide variety of tweaks to these assumptions that would give similar results. For example, it is possible to get a very similar result if we replace the chain $\hat{Q}_\beta$ by another chain that closely approximates the dynamics of Q_β on S (e.g. the trace of Q_β on S rather than the restriction), at some small cost in checking that periodicity does not pose a problem. Similarly, the power 4 appearing in (3.2) could be decreased at some small cost to the other constants appearing in this assumption.

Under these assumptions, we have the following conclusion.

Lemma 1. (Hitting times and conductance.) *Let Assumptions 1 hold, and fix a point x that is in G_β for all $\beta > \beta_0(x)$ sufficiently large. Then, for all $\epsilon > 0$,*

$$\mathbb{P}_x\left[\frac{\log(\tau_{S^c})}{-\log(\Phi_\beta(S))} > 1+\epsilon\right] = o(1).$$

Proof. Fix $p \in W_\beta$, and define $T = T(\beta) = r_1(\beta)+r_2(\beta)+1$. Throughout the following argument we denote by $\{X_t\}_{t\ge 0}$ a Markov chain sampled from Q_β (with starting point indicated by subscripts). We then calculate:

$$\begin{aligned}
\mathbb{P}_p[\tau_{S^c} \le T] &\ge \inf_{q\in G_\beta} \mathbb{P}_q[\tau_{S^c} \le r_1(\beta)+1] - \mathbb{P}_p[\tau_{G_\beta\cup S^c} > r_2(\beta)] \\
&\ge \mathbb{P}_{\pi_\beta|_S}[X_1 \in S^c] - \sup_{q\in G_\beta} \big\|\hat{Q}_\beta^{r_1(\beta)}(q,\cdot) - \pi_\beta|_S(\cdot)\big\|_{\mathrm{TV}} - \mathbb{P}_p[\tau_{G_\beta\cup S^c} > r_2(\beta)] \\
&= \Phi_\beta(S) - \sup_{q\in G_\beta} \big\|\hat{Q}_\beta^{r_1(\beta)}(q,\cdot) - \pi_\beta|_S(\cdot)\big\|_{\mathrm{TV}} - \mathbb{P}_p[\tau_{G_\beta\cup S^c} > r_2(\beta)],
\end{aligned}$$

where the first inequality follows from using the Markov property at the stopping time $\min(\tau_{G_\beta\cup S^c}, r_2(\beta))$, and the second is the triangle inequality (for the total variation metric on distributions). Applying Assumptions 1.2 and 1.3, the two negative terms in this expression are both bounded by $\beta^{-2}\Phi_\beta(S)$. This implies that

$$\mathbb{P}_p[\tau_{S^c} \le T] \ge \Phi_\beta(S)\big(1-2\beta^{-2}\big). \tag{3.3}$$

If we also have $p \in G_\beta \subset W_\beta$, then applying Assumption 1.4 gives

$$\mathbb{P}_p[\tau_{W_\beta^c} < \min(T,\, \tau_{p,S^c})] \le \Phi_\beta(S)^4. \tag{3.4}$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

We now iteratively apply (3.3) and (3.4) to control the behaviour of $\{X_t\}_{t\geq 0}$ over longer time intervals. More precisely, for all $k \in \mathbb{N}$ and starting points $p \in G_\beta$,

$$\begin{aligned}
\mathbb{P}_p[\tau_{S^c} > kT] = & \\
& \mathbb{P}_p[\tau_{S^c} > kT | \tau_{S^c} > (k-1)T,\ X_{(k-1)T} \in W_\beta]\mathbb{P}_p[\tau_{S^c} > (k-1)T,\ X_{(k-1)T} \in W_\beta] \\
& + \mathbb{P}_p[\tau_{S^c} > kT | \tau_{S^c} > (k-1)T,\ X_{(k-1)T} \in W_\beta^c]\mathbb{P}_p[\tau_{S^c} > (k-1)T,\ X_{(k-1)T} \in W_\beta^c] \\
& \leq \mathbb{P}_p[\tau_{S^c} > kT | \tau_{S^c} > (k-1)T,\ X_{(k-1)T} \in W_\beta]\mathbb{P}_p[\tau_{S^c} > (k-1)T] \\
& + \mathbb{P}_p[\tau_{S^c} > (k-1)T,\ X_{(k-1)T} \in W_\beta^c] \\
& \leq (1 - \Phi_\beta(S)(1-2\beta^{-2}))\mathbb{P}_p[\tau_{S^c} > (k-1)T] + \mathbb{P}_p[\tau_{S^c} > (k-1)T,\ X_{(k-1)T} \in W_\beta^c] \\
& \leq (1 - \Phi_\beta(S)(1-2\beta^{-2}))\mathbb{P}_p[\tau_{S^c} > (k-1)T] + k\Phi_\beta(S)^4,
\end{aligned}$$

where (3.3) is used in the second-last line and (3.4) is used in the last line. Iterating and collecting terms, this gives

$$\mathbb{P}_p[\tau_{S^c} > kT] \leq \left(1 - \Phi_\beta(S)\left(1 - 2\beta^{-2}\right)\right)^k + k^2\Phi_\beta(S)^4. \tag{3.5}$$

Fix any $\epsilon > 0$ and take $k = \lceil \beta\Phi_\beta(S)^{-1} \rceil$. By Assumption 1.1 and the fact that r_1, r_2 are bounded by polynomials,

$$kT = \lceil \beta\Phi_\beta(S)^{-1} \rceil (r_1(\beta) + r_2(\beta) + 1) \leq \Phi_\beta(S)^{-1}(5\beta(r_1(\beta) + r_2(\beta))) \leq \Phi_\beta(S)^{-1-\epsilon},$$

for all $\beta > \beta_0(\epsilon)$ sufficiently large. Fix x as in the statement of the lemma; we can use this bound on kT along with (3.5) to conclude that

$$\begin{aligned}
\mathbb{P}_x\left[\frac{\log(\tau_{S^c})}{-\log(\Phi_\beta(S))} > 1 + \epsilon\right] &= \mathbb{P}_x[\tau_{S^c} > \Phi_\beta(S)^{-1-\epsilon}] \\
&\leq \mathbb{P}_x[\tau_{S^c} > kT] \\
&\leq \left(1 - \Phi_\beta(S)\left(1 - 2\beta^{-2}\right)\right)^k + k^2\Phi_\beta(S)^4 = o(1).
\end{aligned}$$

This completes the proof of the lemma. □

3.2. Metastability and spectral gaps

If we can partition the state space of a Markov chain into a collection of sets $S^{(1)}, \ldots, S^{(k)}$ satisfying Assumptions 1, we typically expect the spectral gap of the Markov chain to be entirely determined by the typical transition rates between these sets. However, we must rule out a few possible sources of bad behaviour:

1. Very slow mixing in the 'tails' of the distribution could have an impact on the spectral gap.
2. A typical transition from one mode could land far out in the tails of the mode being entered, causing the walk to get 'stuck'.
3. The transitions between modes might exhibit near-periodic behaviour, even if the Markov chain is not exactly periodic.

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

4. There might be metastability among *collections* of modes. For example, there might be some $I \subset \{1, 2, \ldots, k\}$ for which $\Phi_\beta(\cup_{i\in I} S^{(i)})$ is much smaller than $\min_{1\le i\le k} \Phi_\beta(S^{(i)})$.

Although detailed discussion of metastability is beyond the scope of the present paper, the first three types of behaviour can all cause the spectral gap to be very different from the prediction given by our metastability heuristic. The fourth behaviour simply says that you have chosen the 'wrong' partition of the state space, and that you should check the conditions again after joining several pieces of the partition together.

The following assumptions rule out these new complications.

Assumptions 2. *Let $\Omega = \sqcup_{i=1}^k S^{(i)}$ be a partition of Ω into k pieces. Set $\Phi_{\min} = \min(\Phi_\beta(S^{(1)}), \ldots, \Phi_\beta(S^{(k)}))$ and $\Phi_{\max} = \max(\Phi_\beta(S^{(1)}), \ldots, \Phi_\beta(S^{(k)}))$. We assume the following.*

1. *Metastability of sets: Each set $S^{(i)}$ satisfies Assumptions 1 (with $\Phi_\beta(S)$ replaced by $\Phi_{\max}$ in Assumption 1.1 and replaced by $\Phi_{\min}$ in Assumptions 1.2–4). We use the superscript (i) to extend the notation of that assumption in the obvious way.*

2. *Lyapunov control of tails: Denote by $B_r(x)$ the ball of radius $r > 0$ around a point $x \in \Omega$. Assume there exist $0 < m, M < \infty$ satisfying*

$$\cup_{i=1}^k W_\beta^{(i)} \subset B_M(0), \qquad B_m(0) \subset \cup_{i=1}^k G_\beta^{(i)}.$$

Assume there exist a collection of privileged points $s_i \in G_\beta^{(i)}$ such that the function $V_\beta(x) = \exp\{\beta \min_{1\le i\le k} \|x - s_i\|\}$ satisfies

$$(Q_\beta V_\beta)(x) \le \left(1 - \frac{1}{r_3(\beta)}\right) V_\beta(x) + r_4 e^{\ell\beta}$$

for all $x \in \Omega$, where r_3, r_4 are bounded by polynomials and $0 \le \ell < m$.

3. *Never hitting W_β^c: We have the following variant of (3.2):*

$$\sup_{x\in\cup_{i=1}^k G_\beta^{(i)}} \mathbb{P}\Big[\tau_{x,\left(\cup_{i=1}^k W_\beta^{(i)}\right)^c} < \Phi_{\min}^{-2}\Big] \le \Phi_{\min}^4. \tag{3.6}$$

4. *Non-periodicity: For all $1 \le i \ne j \le k$,*

$$\inf_\beta \inf_{x\in S^{(i)}} Q_\beta(x, S^{(i)}) \equiv c_2^{(i)} > 0, \tag{3.7}$$

$$\sup_{x\in S^{(i)}} Q_\beta\big(x, S^{(j)} \setminus G_\beta^{(j)}\big) < \Phi_{\min}^4. \tag{3.8}$$

5. *Connectedness: There exists some r_5 bounded by a polynomial such that the graph with vertex set $\{1, 2, \ldots, k\}$ and edge set*

$$\left\{(i,j) : \min\left(\inf_{x\in S^{(i)}} \mathbb{P}\big[X_{\tau_{x,(S^{(i)})^c}} \in S^{(j)}\big], \inf_{x\in S^{(j)}} \mathbb{P}\big[X_{\tau_{x,(S^{(j)})^c}} \in S^{(i)}\big]\right) \ge r_5(\beta)\right\} \tag{3.9}$$

is connected.

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

Lemma 2. (Spectral gap and conductance.) *Let Assumptions 2 hold. Denote by* λ_β *and* Φ_β *the spectral gap and conductance of* Q_β*. Then*

$$\lim_{\beta\to\infty} \frac{\log(\lambda_\beta)}{\log(\Phi_{\min})} = \lim_{\beta\to\infty} \frac{\log(\Phi_\beta)}{\log(\Phi_{\min})} = 1. \tag{3.10}$$

Proof. Define the candidate 'small set' $R = \cup_{i=1}^k G_\beta^{(i)}$.

For convenience, we define $T_{\max} = T_{\max}(\beta) \equiv \Phi_{\min}^{-1.5}$ to be the longest timescale of interest in this problem; we note that any mixing behaviour should occur on this timescale, while on the other hand there should be no entrances to the 'bad' set $W \equiv \left(\cup_{i=1}^k W_\beta^{(i)}\right)^{\mathrm{c}}$. In order to reduce notational clutter, we will frequently use q with a subscript to refer to a function that is bounded by a polynomial and whose specific values are not of interest.

We will begin by estimating the mixing rate of Q_β for Markov chains started at points $x, y \in R$. We do this by coupling Markov chains $\{X_t\}_{t=0}^{T_{\max}}$, $\{Y_t\}_{t=0}^{T_{\max}}$ started at $X_0 = x$, $Y_0 = y$ for some $x, y \in R$ and trying to force them to collide. (Note that the Markov chains are only defined up until this 'maximal time' $T_{\max}$.) This saves us from having to either explicitly write $\min(\cdot, T_{\max})$, or add extremely small terms that correspond to the probability that various times exceed $T_{\max}$, in essentially all of the following calculations. This choice has virtually no other impact. Roughly speaking, we will make the following two calculations:

1. If we run the two chains independently, the time it takes for them to both be in $G_\beta^{(i)}$, for the *same i simultaneously*, is not too much larger than the conjectured relaxation time $\Phi_{\min}^{-1}$.
2. If we run two chains started on the same good set $G_\beta^{(i)}$, the two chains will couple long before either one transitions from $G_\beta^{(i)}$ to another mode.

We now give some further details, following this sketch. Let $x, y \in R$.

Part 1: Time to be in same good set simultaneously. We will run the chains independently until the first time $\psi_1 = \inf\left\{t \geq 0 : \text{there exists } 1 \leq i \leq k \text{ such that } X_t, Y_t \in G_\beta^{(i)}\right\}$ that they are both in the same 'good' part of the partition. Define $\psi_2 = \inf\left\{t \geq 0 : \text{there exists } 1 \leq i \leq k \text{ such that } X_t, Y_t \in S^{(i)}\right\}$, the first time that $\{X_t\}$, $\{Y_t\}$ are both in the same part of our partition. For convenience, set $c_2 = \min(c_2^{(1)}, \ldots, c_2^{(k)}, 0.5)$.

Let $1 \leq a, b \leq k$ satisfy $x \in G_\beta^{(a)}$, $y \in G_\beta^{(b)}$ and let G be the connected graph whose existence is guaranteed in (3.9). Since G is a connected graph on k vertices, there is a path γ in G from a to b of length $|\gamma| \leq k-1$. We next consider the event $\mathcal{E}$ that, the first $|\gamma|$ times that $\{X_t\}_{t\geq 0}$ or $\{Y_t\}_{t\geq 0}$ changes parts of the partition, they go towards each other along the path γ. By (3.9), $\mathbb{P}[\mathcal{E}] \geq r_5(\beta)^{k-1}$.

Bounding the amount of time spent in each part of the partition along this path by (3.5) and (3.7), this implies that

$$\mathbb{P}\left[\psi_2 \leq q_1(\beta)\Phi_{\min}^{-1}\right] \geq \frac{1}{2} c_2 r_5(\beta)^{k-1} \tag{3.11}$$

for some function q_1 that is bounded by a polynomial. By (3.8),

$$\mathbb{P}[\psi_1 = \psi_2] \geq 1 - T_{\max}\Phi_{\min}^4 \geq 1 - \Phi_{\min}^2.$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

Combining this with (3.11),

$$\begin{aligned}\mathbb{P}\big[\psi_1 \le q_1(\beta)\Phi_{\min}^{-1}\big] &\ge \mathbb{P}\big[\psi_2 \le q_1(\beta)\Phi_{\min}^{-1}\big] - \mathbb{P}[\psi_1 \ne \psi_2] \\ &\ge \frac{1}{2}c_2 r_5(\beta)^{k-1} - \Phi_{\min}^2 \equiv \frac{1}{q_2(\beta)},\end{aligned} \tag{3.12}$$

where we note that q_2 is bounded by a polynomial.

Part 2: Mixing from same good set. If $\psi_1 \ge T_{\max} - r_1(\beta)$, continue to evolve $\{X_t\}_{t\ge 0}$, $\{Y_t\}_{t\ge 0}$ independently. Otherwise, let $1 \le i \le k$ satisfy $X_{\psi_1},\ Y_{\psi_1} \in S^{(i)}$. We then let $\{\hat{X}_t\}_{t\ge 0}$, $\{\hat{Y}_t\}_{t\ge 0}$ be Markov chains evolving according to the Metropolis–Hastings kernel with proposal distribution Q_β and target distribution $\pi_\beta|_{S^{(i)}}$. We give these chains initial points $\hat{X}_0 = X_{\psi_1}$, $\hat{Y}_0 = Y_{\psi_1}$ and couple them according to a maximal $r_1(\beta)$-step coupling (that is, a coupling that maximizes $\mathbb{P}[\hat{X}_{r_1(\beta)} = \hat{Y}_{r_1(\beta)}]$; such a coupling is known to exist [4]).

We next observe that the following informal algorithm gives a valid coupling of the Markov chains $\{X_t\}_{t=\psi_1}^{\psi_1+r_1(\beta)}$, $\{\hat{X}_t\}_{t=0}^{r_1(\beta)}$:

1. Run the full Markov chain $\{X_t\}_{t=\psi_1}^{\psi_1+r_1(\beta)}$ according to Q_β.
2. For all $t < \tau_{\text{bad}} \equiv \inf\{s : X_{\psi_1+s} \notin S^{(i)}\}$, set $\hat{X}_t = X_{\psi_1+t}$.
3. If $\tau_{\text{bad}} < r_1(\beta)$, continue to evolve $\{\hat{X}_s\}_{s=\tau_{\text{bad}}}^{r_1(\beta)}$ independently of $\{X_t\}_{t=0}^{r_1(\beta)}$. (Note that the particular choice made in this third step will not influence the analysis—we could make any measurable choice here.)

We couple the pair of chains $\{X_t\}_{t=\psi_1}^{\psi_1+r_1(\beta)}$, $\{\hat{X}_t\}_{t=0}^{r_1(\beta)}$ this way, and we couple $\{Y_t\}_{t=\psi_1}^{\psi_1+r_1(\beta)}$, $\{\hat{Y}_t\}_{t=0}^{r_1(\beta)}$ analogously. Under these couplings, we have

$$\begin{aligned}\mathbb{P}[X_{\psi_1+r_1(\beta)} \ne Y_{\psi_1+r_1(\beta)}] &\le \mathbb{P}[\hat{X}_{r_1(\beta)} \ne \hat{Y}_{r_1(\beta)}] + \mathbb{P}[X_{\psi_1+r_1(\beta)} \ne \hat{X}_{r_1(\beta)}] \\ &\quad + \mathbb{P}[Y_{\psi_1+r_1(\beta)} \ne \hat{Y}_{r_1(\beta)}] \\ &\le \beta^{-2}\Phi_{\max} + 2\Phi_{\min}^4,\end{aligned}$$

where the first term is bounded by (3.1) and the last two terms are bounded by (3.6).

Combining this with (3.12), we conclude that

$$\mathbb{P}[X_{T_1} = Y_{T_1}] = (1 - o(1)), \tag{3.13}$$

where $T_1 = \lceil q_1(\beta)\Phi_{\min}^{-1} + r_1(\beta)\rceil$.

This completes the proof of our two-stage analysis, as (3.13) gives a useful minorization bound for $x, y \in R$. Note that (3.13) is very close to a minorization condition in the sense of [21] for the small set R. Applying the closely related Lemma A.11 of [12] (see also [13]), the minorization bound (3.13), and the drift bound in Assumptions 2.2, we find

$$\|Q_\beta^t(x,\cdot) - \pi_\beta(\cdot)\|_{\text{TV}} \le M(\beta, x)\exp\left\{-\frac{t\Phi_{\min}}{q_4(\beta)}\right\},$$

where $q_4(\cdot) \ge 1$ and, for each x, $M(\cdot, x)$ is bounded by a polynomial.

By Theorem 2.1 of [20], this implies that

$$\lambda_\beta \ge \frac{\Phi_{\min}}{q_4(\beta)}. \tag{3.14}$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

By (2.1), the conductance Φ_β of Q_β satisfies

$$\lambda_\beta \leq 2\Phi_\beta \leq 2\Phi_{\min}. \tag{3.15}$$

Combining (3.14) and (3.15), we conclude that

$$\frac{\Phi_{\min}}{q_4(\beta)} \leq \frac{\Phi_\beta}{q_4(\beta)}, \qquad \lambda_\beta \leq 2\Phi_{\min} \leq 2\Phi_\beta.$$

This immediately implies the limit in (3.10), completing the proof of the lemma. □

4 Application to mixtures of Gaussians

We define the usual random walk Metropolis algorithm.

Definition 3. *(Random walk Metropolis algorithm.)* The transition kernel K of the *random walk Metropolis algorithm* with step size $\sigma > 0$ and target distribution π on $\mathbb{R}^d$ with density ρ is given by the following algorithm for sampling $X \sim K(x, \cdot)$:

1. Sample $\epsilon_1 \sim \mathcal{N}(0, \sigma^2)$ and $U_1 \sim \text{Unif}[0, 1]$.
2. If
$$U < \frac{\rho(x+\epsilon_1)}{\rho(x)},$$
set $X = x + \epsilon_1$, otherwise set $X = x$.

For $\sigma > 0$, define the mixture distribution

$$\pi_\sigma = \frac{1}{2}\mathcal{N}(-1, \sigma^2) + \frac{1}{2}\mathcal{N}(1, \sigma^2)$$

and denote its density by f_σ. Let K_σ be the kernel from Definition 3 with step size σ and target distribution π_σ.

Denote by λ_σ the relaxation time of K_σ (the reciprocal of the spectral gap of K_σ), and denote by $\Phi_\sigma = \Phi(K_\sigma, (-\infty, 0))$ the Cheeger constant associated with kernel K_σ and set $(-\infty, 0)$.

We will state our two main results about this walk; the proofs are deferred until both results have been stated. First, we have an asymptotic formula for the Cheeger constant.

Theorem 1. *The Cheeger constant Φ_σ satisfies*

$$\lim_{\sigma \to 0} (-2\sigma^2) \log(\Phi_\sigma) = 1. \tag{4.1}$$

For fixed $x \in (-\infty, 0)$, let $\{X_t^{(\sigma)}\}_{t \in \mathbb{N}}$ be a Markov chain with transition kernel K_σ and initial point $X_1^{(\sigma)} = x$. Define the hitting time $\tau_x^{(\sigma)} = \inf\{t > 0 : X_t^{(\sigma)} \notin (-\infty, 0)\}$.

We also have the following estimate of the spectral gap and the hitting time.

Theorem 2. *For all $\epsilon > 0$ and fixed $x \in (-\infty, 0)$, the hitting time $\tau_x^{(\sigma)}$ satisfies*

$$\lim_{\sigma \to 0} \mathbb{P}\left[\frac{\log(\tau_x^{(\sigma)})}{\log(\Phi_\sigma)} < 1 + \epsilon\right] = 1$$

and the relaxation time satisfies

$$\lim_{\sigma \to 0} \frac{\log(\lambda_\sigma)}{\log(\Phi_\sigma)} = \lim_{\sigma \to 0} \frac{\log(\lambda_\sigma)}{\log(\Phi(K_\sigma))} = 1.$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

Remark 2. This result implies that the Cheeger constant $\Phi(K_\sigma)$ of K_σ is close to the bottleneck ratio $\Phi_\sigma = \Phi(K_\sigma, (-\infty, 0))$ associated with the set $(-\infty, 0)$, at least for σ very small. The set $(-\infty, 0)$ is of course a natural guess for the set with the 'worst' conductance, though we do not know of any simple argument that would actually prove this. In some sense this is the motivation for the approach taken in this paper: it can be very hard to guess a good partition, even in a very simple example!

We begin by proving Theorem 1.

Proof of Theorem 1. Let $\{X_t\}_{t\geq 0}$ be a Markov chain with transition kernel K_σ, started at $X_0 \sim \pi_\sigma$, and drawn according to the stationary distribution. Denote by ϕ_σ the density of the Gaussian with variance σ^2. Defining the set $\mathcal{E} = \{X_0 < -\sigma^{-1}\} \cup \{|X_1 - X_0| > \sigma^{-1}\}$, we have

$$\begin{aligned}
\mathbb{P}[\{X_0 < 0\} \cap \{X_1 > 0\} \cap \mathcal{E}^c] &\leq \int_{-\sigma^{-1}}^{0} \int_{0}^{\sigma^{-1}} f_\sigma(x)\phi_\sigma(y-x)\,\mathrm{d}x\,\mathrm{d}y \\
&\leq 2\int_{-\sigma^{-1}}^{0} \int_{0}^{\sigma^{-1}} \phi_\sigma(1+x)\phi_\sigma(y-x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{2}{\pi\sigma^2}\int_{-\sigma^{-1}}^{0} \int_{0}^{\sigma^{-1}} \exp\left\{-\frac{1}{2\sigma^2}((1+x)^2 + (y-x)^2)\right\}\mathrm{d}x\,\mathrm{d}y \\
&\leq \frac{2}{\pi\sigma^2}\int_{-\sigma^{-1}}^{0} \int_{0}^{\sigma^{-1}} \mathrm{e}^{-\frac{1}{2\sigma^2}} = \frac{2}{\pi\sigma^4}\mathrm{e}^{-\frac{1}{2\sigma^2}}.
\end{aligned}$$

We also have the simple bound

$$\begin{aligned}
\mathbb{P}[\mathcal{E}] &\leq \mathbb{P}[X_0 < -\sigma^{-1}] + \mathbb{P}[|X_1 - X_0| > \sigma^{-1}] \\
&\leq 2\int_{-\infty}^{-\sigma^{-1}} \phi_\sigma(x)\,\mathrm{d}x + 2\int_{\sigma^{-1}}^{\infty} \phi_\sigma(x)\,\mathrm{d}x \\
&\leq \frac{4}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\sigma^{-1}-1)^2}{2\sigma^2}\right\} \leq \frac{4}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{3}\sigma^{-3}},
\end{aligned}$$

where the last inequality holds for all σ sufficiently small. Putting these two bounds together, we have, for all $\sigma > 0$ sufficiently small,

$$\mathbb{P}[X_0 < 0,\ X_1 > 0] \leq \frac{1}{\pi\sigma^4}\mathrm{e}^{-\frac{1}{2\sigma^2}} + \frac{4}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{3}\sigma^{-3}}.$$

Taking logs, this immediately proves that $\lim_{\sigma\to 0}(-2\sigma^2)\log(\Phi_\sigma) \leq 1$, the desired upper bound on the left-hand side of (4.1). To prove the lower bound on this quantity, begin by defining the intervals $I_\sigma = (-2\sigma^{20}, -\sigma^{20})$ and $J_\sigma = (\sigma^{10}, 2\sigma^{10})$. Since $\sigma^{20} \ll \sigma^{10}$ for σ small, we have, for sufficiently small $\sigma > 0$,

$$\inf_{x\in I_\sigma, y\in J_\sigma} \frac{f_\sigma(y)}{f_\sigma(x)} \geq 1.$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

Informally, this means that any proposed step from I_σ to J_σ will be accepted. Thus, letting $Y \sim \mathcal{N}(0, \sigma^2)$ be independent of X_1, we have

$$\begin{aligned}\mathbb{P}[X_0 < 0, X_1 > 0] &\geq \mathbb{P}[X_0 \in I_\sigma, X_1 \in J_\sigma] \\ &\geq \mathbb{P}[X_0 \in I_\sigma] \inf_{x \in I_\sigma} \mathbb{P}[X_0 + Y \in J_\sigma \mid X_0 = x] \\ &\geq \left(\frac{\sigma^{20}}{4\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2\sigma^2}}\right) \times \left(\frac{\sigma^{10}}{4\sqrt{2\pi}}\right),\end{aligned}$$

where the last inequality holds for all $\sigma > 0$ sufficiently small. Taking logs, this proves that $\lim_{\sigma \to 0} (-2\sigma^2) \log(\Phi_\sigma) \geq 1$, completing the proof of (4.1). □

Proof of Theorem 2. We defer some of the longer exact calculations in this proof to Appendix A, retaining here the key steps that might be used to prove similar metastability results for other Markov chains. To prove Theorem 2, it is enough to verify Assumptions 1 and 2 for the sets $S^{(1)} = (-\infty, 0)$ and $S^{(2)} = [0, \infty)$, with the decomposition of $S^{(1)}$ into $G_\beta^{(1)} = (-\sigma^{-9}, 0)$, $W_\beta^{(1)} = (-\sigma^{-10}, 0)$, $B_\beta^{(1)} = (-\infty, -\sigma^{-10})$, and $G_\beta^{(2)}$, $W_\beta^{(2)}$, and $B_\beta^{(2)}$ defined analogously (see Figure 1). Note that Assumption 1.1 follows immediately from (4.1), which we have already proved.

Denote by $\hat{K}_\sigma$ the Metropolis–Hastings transition kernel on $(-\infty, 0)$ that has as its proposal kernel K_σ and as its target distribution the density $\hat{\rho}_\sigma(x) = 2f_\sigma(x)$, $x \in (-\infty, 0)$.

We begin by proving some stronger Lyapunov-like bounds for K_σ and $\hat{K}_\sigma$.

Lemma 3. *Let $V_\sigma(x) = \mathrm{e}^{\sigma^{-1} \min(\|x-1\|, \|x+1\|)}$. Then there exist $0 < \alpha \leq 1$, $0 \leq M, C < \infty$ such that, for all $K \in \{K_\sigma, \hat{K}_\sigma\}$ and $x \in (-\infty, -M\sigma)$,*

$$(KV_\sigma)(x) \leq (1 - \alpha)V_\sigma(x) + C. \tag{4.2}$$

Furthermore, Assumption 2.2 holds.

Proof. The proof is deferred to Appendix A. □

We next check the main condition.

Theorem 3. *With notation as above, Assumption 1.2 is satisfied.*

Proof. We begin with a weak estimate of mixing from within a good set.

Lemma 4. *Fix $0 < \delta < \frac{1}{20}$. With notation as above, there exist some constants $0 < a_1, A_1 < \infty$ such that*

$$\sup_{-\sigma^{-11} < x, y < -\delta} \left\| \hat{K}_\sigma^T(x, \cdot) - \hat{K}_\sigma^T(y, \cdot) \right\|_{\mathrm{TV}} \leq A_1 \mathrm{e}^{-a_1 \sigma^{-1}} \tag{4.3}$$

for $T > A_1 \sigma^{-a_1}$.

Proof. The proof is deferred to Appendix A. □

Note that this bound is not good enough for our conclusions, since our upper bound $\mathrm{e}^{-a_1\sigma^{-1}}$ is still very large compared to the conductance of interest. We improve the bound by iterating it several times.

Lemma 5. *Fix $0 < \delta < \frac{1}{20}$. There exist constants $0 < a_2, A_2 < \infty$ depending only on δ such that*

$$\sup_{-\sigma^{-9} < x, y < -\delta} \left\| \hat{K}_\sigma^S(x, \cdot) - \hat{K}_\sigma^S(y, \cdot) \right\|_{\mathrm{TV}} \leq A_2 \mathrm{e}^{-a_2 \sigma^{-5}} \tag{4.4}$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

for $S = \lceil A_2 \sigma^{-a_2} \rceil$. *Furthermore, there exist constants* $0 < a_3, A_3 < \infty$ *such that*

$$\sup_{x \in (-\delta, 0)} \mathbb{P}[\tau_{x,(-\delta,0)^c} < A_3 \sigma^{-a_3}] \geq 1 - \mathrm{e}^{-\sigma^{-10}}. \tag{4.5}$$

Proof. The proof is deferred to Appendix A. □

Fix $\delta = 0.01$. Combining (4.4) with the bound in (4.5) on the length of excursions above $-\delta$ completes the proof of the Theorem 3.

Lemma 6. *With notation as above, Assumptions 1.3, 1.4, and 2.3 hold.*

Proof. These all follow immediately from Lemma 3 and the definition of our partition.

Next, note that Assumption 2.1 holds by the symmetry of $S^{(1)}$, $S^{(2)}$ and the fact that we have already checked Assumptions 1.

Thus, it remains only to check Assumption 2.4.

Lemma 7. *With notation as above, Assumption 2.4 holds.*

Proof. It is immediately clear that $K_\sigma(x, (-\infty, x]) \geq \frac{1}{2}$ for all $x \in \mathbb{R}$, which implies (3.7). Standard Gaussian inequalities imply that $\sup_{x<0} K_\sigma(x, (\sigma^{10}, \infty)) \leq \mathrm{e}^{-\sigma^9}$ for $\sigma < \sigma_0$ sufficiently small. Combining this with (4.1) completes the proof of (3.8). □

Since we have verified all the assumptions of Lemmas 1 and 2, applying them completes the proof of Theorem 2.

5. Application to sampling from high-dimensional sets

We give an example with a very different flavour from that in Section 4: an analysis of multimodality in a simple Gibbs-like walk for sampling from the uniform distribution on a high-dimensional set. To keep the analysis as simple and easy-to-read as possible, we consider a very simple situation; we briefly discuss some generalizations in Remark 4.

We begin by defining a simple variant of the hit-and-run sampler introduced in [22]. Roughly speaking, this algorithm proceeds by choosing a random direction at every step and sampling uniformly along this line. More precisely, denote by λ the Lebesgue measure on $\mathbb{R}^d$ and fix an open set $\mathcal{C} \subset \mathbb{R}^d$ with $0 < \lambda(\mathcal{C}) < \infty$. We define the associated sampler L with target distribution equal to the uniform distribution on $\mathcal{C}$ by the following algorithm for sampling $X \sim L(x, \cdot)$:

1. Sample v uniformly on the unit sphere $\{y \in \mathbb{R}^n : \|y\| = 1\}$.
2. Define $\ell' = \{x + sv : s \in \mathbb{R}^n\} \cap \mathcal{C}$. Note that ℓ' is a union of line segments and includes x; let ℓ be the connected component of ℓ' that contains x.
3. Return $X \sim \mathrm{Unif}(\ell)$.

We now set up the main problem. Roughly speaking, we will check that this algorithm exhibits metastability when $\mathcal{C}$ is the union of two moderately rounded sets that have small intersection. A concrete family of such sets is given in Example 1.

Proceeding more formally, fix constants $r, R \in \mathbb{R}$ such that $0 < r < R$, and $\Delta > 0$. Let $\hat{\mathsf{A}}_1, \hat{\mathsf{A}}_2, \ldots \subseteq \mathbb{R}^d$ and $\hat{\mathsf{K}}_1, \hat{\mathsf{K}}_2, \ldots \subseteq \mathbb{R}^d$ be two sequences of convex bodies. Suppose that, for each $\beta \in \mathbb{N}$, the convex bodies $\hat{\mathsf{A}}_\beta$ and $\hat{\mathsf{K}}_\beta$ each contain a (possibly different) ball of radius r, and are each contained in a (possibly different) ball of radius R. For each $\beta \in \mathbb{N}$, we define

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

$\mathsf{A}_\beta := \hat{\mathsf{A}}_\beta + B(0, \Delta)$ and $\mathsf{K}_\beta := \hat{\mathsf{K}}_\beta + B(0, \Delta)$, where '+' denotes the Minkowski sum. Finally, let $\mathcal{C}_\beta = \mathsf{A}_\beta \cup \mathsf{K}_\beta$, and let Q_β be the hit-and-run sampler associated with the set $\mathcal{C}_\beta$.

For $c > 0$ and $S \subset \mathbb{R}^d$, we define the c-interior of S by $S^{(-c)} = \{x \in \mathbb{R}^d : \inf_{y \notin S} \|x - y\| > c\}$.

We assume the following conditions on our sequences of convex bodies $\hat{\mathsf{A}}_1, \hat{\mathsf{A}}_2, \ldots \subseteq \mathbb{R}^d$ and $\hat{\mathsf{K}}_1, \hat{\mathsf{K}}_2, \ldots \subseteq \mathbb{R}^d$ in order to guarantee metastability.

Assumptions 3. *We assume:*

1. *There exists a measurable set $S \subset \mathbb{R}^d$ such that, for all $\beta \in \mathbb{N}$,*

$$\mathsf{K}_\beta \cap S \subseteq \mathsf{A}_\beta \cap S,$$
$$\mathsf{A}_\beta \cap S^{\mathrm{c}} \subseteq \mathsf{K}_\beta \cap S^{\mathrm{c}},$$
$$\mathsf{A}_\beta^{\left(-\frac{1}{80\sqrt{d}}\Delta\right)} \subset S,$$
$$\mathsf{K}_\beta^{\left(-\frac{1}{80\sqrt{d}}\Delta\right)} \subset S^{\mathrm{c}}.$$

2. *Define the function*

$$f(\beta) := \frac{1}{\lambda(S \cap \mathcal{C}_\beta)} \int_{S \cap \mathcal{C}_\beta} Q_\beta(x, S^{\mathrm{c}})\, \mathrm{d}x,$$

which is just the conductance of the set $S \cap \mathcal{C}_\beta$ for the kernel Q_β. There exist constants $c_1, c_2, c_3 > 0$ such that $\mathrm{e}^{-c_1 \beta^{c_2}} < f(\beta) \leq \mathrm{e}^{-c_3 \beta}$ for all $\beta \in \mathbb{N}$.

The main result of this section is the following theorem.

Theorem 4. *Fix sequences of convex sets satisfying Assumptions 3, and let Q_β be the associated hit-and-run transition kernel. The spectral gap ρ_β and conductance Φ_β of Q_β satisfy*

$$\lim_{\beta \to \infty} \frac{\log(\rho_\beta)}{\log(\Phi_\beta(S))} = 1.$$

Example 1. As a concrete example, for any $\Delta < \frac{1}{10}$, consider the Δ-rounded simplex $\mathcal{D} = \{x \in \mathbb{R}^d : x[i] \geq 0 \text{ for all } i \in [d] \text{ and } \|x\|_1 \leq 1\} + B(0, \Delta)$. For all $\alpha \geq 0$, let $\mathsf{A}^\alpha = \alpha \mathbf{1} + \mathcal{D}$, where $\mathbf{1} = (1, \ldots, 1)^\top$ is the all-ones vector. Let $\mathsf{K}^\alpha = -\mathsf{A}^\alpha$, and let $\mathcal{C}^\alpha = \mathsf{A}^\alpha \cup \mathsf{K}^\alpha$. Note that these sets are clearly contained, for example, in the ball of radius $R = 2$ around the origin, and contain, for example, the ball of radius $r = \frac{1}{2d^2}$ around $(\frac{1}{2d}, \ldots, \frac{1}{2d})$.

We next check that this sequence satisfies Assumptions 3 for some appropriate choice of $\alpha = \alpha(\beta)$. Let S be the half-plane $S = \{x \in \mathbb{R}^d : \mathbf{1}^\top x \geq 0\}$. Define the function

$$F(\alpha) := \frac{1}{\lambda(S \cap \mathsf{A}^\alpha)} \int_{S \cap \mathsf{A}^\alpha} L^\alpha(x, S^{\mathrm{c}})\, \mathrm{d}x,$$

where L^α is the hit-and-run sampler associated with the set $\mathcal{C}^\alpha$.

We note that F is clearly continuous, strictly monotone decreasing on $[0, \Delta]$, and satisfies $f(0) > 0$ and $F(\Delta) = 0$. Thus, there exists a monotone increasing function $g : [0, \infty) \mapsto [0, \Delta]$ that satisfies $\mathrm{e}^{-c_1 x^{c_2}} < F(g(\beta)) \leq \mathrm{e}^{-c_3 x}$ for some $c_1, c_2, c_3 > 0$ and all $\beta \in [0, \infty)$.

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

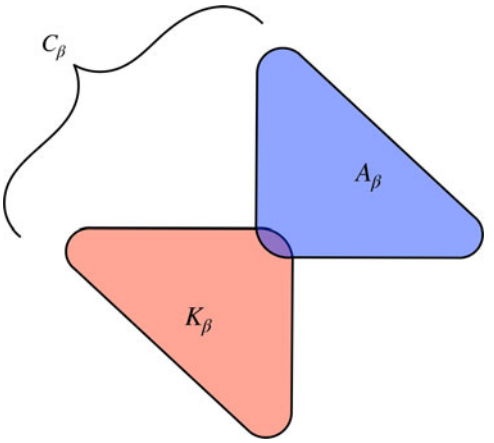

FIGURE 2. The convex sets A_β and K_β, and their union $\mathcal{C}_\beta = \mathsf{A}_\beta \cup \mathsf{K}_\beta$, considered in Example 1. Theorem 4 implies metastability of the hit-and-run Markov chain on the uniform distribution on the sets $\mathcal{C}_\beta$ as $\beta \to \infty$.

Define $\mathsf{A}_\beta := \mathsf{A}^{g(\beta)}$, $\mathsf{K}_\beta := \mathsf{K}^{g(\beta)}$, and $\mathcal{C}_\beta = \mathcal{C}^{g(\beta)} = \mathsf{A}^{g(\beta)} \cup \mathsf{K}^{g(\beta)}$ for all $\beta \in \mathbb{N}$ (see Figure 2). Then the sequence $\mathcal{C}_1, \mathcal{C}_2, \ldots$ satisfies the assumptions of Theorem 4.

Remark 3. Note that we did not need to explicitly compute either $\Phi_\beta(S)$ or ρ_β here! For this particular example, estimating $\Phi_\beta(S)$ is likely not too difficult, but in more complicated examples this can be useful.

Remark 4. In Theorem 4, we held the ambient dimension d, the radii of the contained and containing balls r, R, and the amount of 'rounding' Δ as constant for simplicity. We point out here that the proof of Theorem 4 more or less goes through for many sequences of convex bodies even when these constants are allowed to change (and indeed even when the amount of rounding Δ is 0). Essentially the only serious difficulty comes from the step at which we check that the mixing time of the walk restricted to each part of the partition is small. In the proof of Theorem 4, we checked this by invoking Corollary 1.2 of [9], the strongest result in this area that we are aware of. This theorem has basically two requirements for the sequence of convex bodies $\mathcal{K}_\beta$:

1. Denote by R_β the radius of the smallest ball containing $\mathcal{K}_\beta$, r_β the radius of the largest ball contained in $\mathcal{K}_\beta$, and d_β its dimension. Then $\frac{R_\beta d_\beta}{r_\beta}$ cannot grow more than polynomially quickly in β.
2. From *any* starting point, the hit-and-run kernel must jump at least some non-negligible distance with non-negligible probability, in a sense made formal by (B.4).

The first condition is often trivial to verify. The second condition seems trivial in moderate dimension, and is guaranteed by having a fixed dimension d and degree of rounding $\Delta > 0$. It can hold even as d goes to infinity and $\Delta = 0$, but it turns out to be more complicated in higher dimensions. We can see the basic difficulty quite clearly by considering the cube $[0, 1]^n$: if a hit-and-run walk is started at the origin, it will stay there until the proposed direction v has components that are either all positive or all negative. This has probability $2^{-(n-1)}$, so on

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

average the walk will not move *at all* for something on the order of 2^n steps. Thus, although the walk mixes in polynomial time *started from the point* $(0.5, 0.5, \ldots, 0.5)$, the *worst-case* mixing time is $\Omega(2^n)$. Thus, the worst-case mixing time can be quite bad for sequences of convex bodies with growing dimensions.

There is a large literature on dealing with 'corners' in high-dimensional bodies (with some discussion in [9] itself), but a detailed discussion is far beyond the scope of the present article. For this reason, we limit ourselves to convex bodies (such as Δ-rounded convex bodies) where it is easy to see that the walk gets 'far' from the boundary within a few steps with high probability.

Proof of Theorem 4. We verify the conditions in Assumptions 2 for the partition $S \sqcup S^{\mathrm{c}}$ of Ω, with trivial choice of sets $G_\beta = W_\beta \equiv S$, $B_\beta = \emptyset$.

Going down the list quickly, we look at the referenced parts of Assumptions 1:

1. This follows from the fact that $\mathrm{e}^{-c_1\beta^{c_2}} < f(\beta) \leq \mathrm{e}^{-c_3\beta}$ for all $\beta \in \mathbb{N}$.
2. Checking this is the main difficulty in proving the theorem. This follows from applying Lemma 8 in Appendix B with the choice $\epsilon = \beta^{-2}\Phi_\beta(S)$; note that the result is polynomial in β because of the assumption that $\Phi_\beta(S) \geq \mathrm{e}^{-c_1\beta^{c_2}}$ for some $c_1, c_2 > 0$.
3. $W_\beta \cup B_\beta \backslash G_\beta = \emptyset$, so this is immediate.
4. For the same reason as part 3, this is immediate.

We then verify the remaining parts of Assumptions 2:

1. This just refers to the parts of Assumptions 1 checked above.
2. Since Ω is compact, this is clear.
3. Again, since $W_\beta \cup B_\beta \backslash G_\beta = \emptyset$, this is immediate.
4. By the symmetry of the problem, (3.7) holds with constant at least $\frac{1}{2}$. The inequality (3.8) follows again from the fact that $W_\beta \cup B_\beta \backslash G_\beta = \emptyset$.
5. Since our partition has only two parts, this is immediate.

Having checked Assumptions 2, the result follows from an application of Lemma 2. □

Appendix A. Technical bounds from the proof of Theorem 2

We prove some technical lemmas that occur in the proof of Theorem 2.

Proof of Lemma 3. Let $a = a_\sigma$ be the unique local minimum of f_σ in the interval $(-2, -0.5)$—it is clear one such exists for all $\sigma > \sigma_0$ sufficiently large, and that a_σ is within distance $O\big(\mathrm{e}^{-\frac{1}{3\sigma^2}}\big)$ of -1). Let Q_σ be the transition kernel given in Definition 3 with step size σ and target distribution $\mathcal{N}(a, 4\sigma^2)$. Let $L_\sigma(x) = \mathrm{e}^{-\sigma^{-1}\|x-1\|}$. By a standard computation (e.g. keeping track of the constants in the proof of Theorem 3.2 of [14]), there exist $0 < \alpha \leq 1$ and $0 \leq C < \infty$ such that $(Q_\sigma L_\sigma)(x) \leq (1-\alpha)L(x) + C$ for all $x \in \mathbb{R}$ and $Q \in \{Q_\sigma, \hat{Q}_\sigma\}$. Next, observe that

$$\inf_{x \in (-\infty, -10\sigma)} \frac{\mathrm{d}^2}{\mathrm{d}x^2} - \log(f_\sigma(x)) \geq \frac{1}{8\sigma^2}.$$

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

In particular, f_σ is strongly log-concave on the interval $(-\infty, -10\sigma)$ with the same parameter as the density of $\mathcal{N}(a, 4\sigma^2)$. Thus, if we fix $M > 0$ and $x \in (-\infty, -M\sigma)$ and let $X \sim K_\sigma(x, \cdot)$, we have, for $K \in \{K_\sigma, \hat{K}_\sigma\}$,

$$\begin{aligned}(KV_\sigma)(x) &\le (Q_\sigma L_\sigma)(x) + \mathbb{E}[V_\sigma(X)\mathbb{1}_{X>-10\sigma}] \\ &\le (1-\alpha)L_\sigma(x) + C + \mathbb{E}[V_\sigma(X)\mathbb{1}_{X>-10\sigma}] \\ &= (1-\alpha)V_\sigma(x) + C + \mathbb{E}[V_\sigma(X)\mathbb{1}_{X>-10\sigma}]. \end{aligned} \tag{A.1}$$

Let $Y \sim \mathcal{N}(0, \sigma^2)$. As $M \to \infty$, for $K = \hat{K}_\sigma$ we can then bound the last term by

$$\mathbb{E}[V_\sigma(X)\mathbb{1}_{X>-10\sigma}] \le V_\sigma(x)\mathbb{E}[\mathrm{e}^{\sigma^{-1}Y}\mathbb{1}_{x+Y\in[-10\sigma,0]}] = (1+o(1))V_\sigma(x). \tag{A.2}$$

Combining (A.1) and (A.2) completes the proof of (4.2) in the case $K = \hat{K}_\sigma$. In the case $K = K_\sigma$, we replace (A.4) by the similar bound

$$\begin{aligned}\mathbb{E}[V_\sigma(X)\mathbb{1}_{X>-10\sigma}] &\le V_\sigma(x)\mathbb{E}[\mathrm{e}^{\sigma^{-1}Y}\mathbb{1}_{x+Y\in[-10\sigma,10\sigma]}] + V_\sigma(x)\mathbb{P}[x+Y>10\sigma] \\ &= (1+o(1))V_\sigma(x)\end{aligned}$$

to obtain the same conclusion.

Finally, Assumption 2.2 immediately follows from (4.2) in the case $K = K_\sigma$ and the trivial inequality $\sup_{|x|\le M\sigma}(K_\sigma V_\sigma)(x) \le \mathrm{e}^{\sigma^{-3}}$ for any fixed M and all $\sigma < \sigma_0 = \sigma_0(A)$ sufficiently small.

Proof of Lemma 4. Fix α, C as in Lemma 3 and let μ be the uniform distribution on the interval $I = [-1 - \frac{10C}{\alpha}\sigma, -1 + \frac{10C}{\alpha}\sigma]$. We note that $\hat{K}_\sigma$ inherits the following *minorization condition* from the standard Gaussian:

$$\inf_{x\in I}\inf_{J\subset I}\hat{K}_\sigma(x, J) \ge \epsilon\mu(J) \tag{A.3}$$

for some $\epsilon > 0$ that does not depend on σ.

Fix $-\sigma^{-10} < x, y < -\delta$. Applying the popular 'drift-and-minorization' bound in Section 10 of [15], using the 'drift' bound in (4.2) and the 'minorization' bound in (A.3) gives a bound of the form

$$\sup_{-\sigma^{-10}<x,y<-\delta}\left\|\hat{K}_\sigma^T(x,\cdot) - \hat{K}_\sigma^T(y,\cdot)\right\|_{\mathrm{TV}} \le B_1\mathrm{e}^{-b_1\sigma^{-1}} + 2\sup_{-\sigma^{-10}<x<-\delta}\mathbb{P}[\tau_{x,(-M\sigma,\infty)} < t] \tag{A.4}$$

for all $T > B_1\sigma^{-b_1}$, where $0 < b_1, B_1$ are constants that do not depend on σ. Note that the second term on the right-hand side, which does not appear in [15], represents the possibility that a Markov chain ever escapes from the set $(-\infty, -M\sigma)$ on which the drift bound (4.2) holds.

Fix $-\sigma^{-10} < x < -\delta$ and let $\{X_t\}_{t\ge0}$ be a Markov chain with transition kernel $\hat{K}_\sigma$ and starting point $X_0 = x$. Let $\tau = \inf\{t \ge 0 : X_t > -M\sigma\}$. By (4.2), we have, for all $t \in \mathbb{N}$, $\mathbb{E}[V_\sigma(X_t)\mathbb{1}_{\tau\ge t-1}] \le (1-\alpha)^t V_\sigma(X_0) + \frac{C}{\alpha}$. Thus, by Markov's inequality,

$$\begin{aligned}\mathbb{P}[\tau \le t] &\le \mathrm{e}^{-\sigma^{-1}(-M\sigma+1)}\sum_{s=0}^{t}\left((1-\alpha)^t V_\sigma(X_0) + \frac{C}{\alpha}\right) \\ &\le t\,\mathrm{e}^{-\sigma^{-1}(-M\sigma+1)}\left(\mathrm{e}^{\sigma^{-1}(-\delta+1)} + \frac{C}{\alpha}\right).\end{aligned}$$

Combining this with (A.4) completes the proof of the lemma. □

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

Proof of Lemma 5. We denote by $\{X_t\}_{t\geq 0}$ a Markov chain with transition kernel $\hat{K}_\sigma$ and some starting point $X_0 = x$. To improve on the bound in Lemma 4, we must control what can occur when coupling does not happen quickly. There are two possibilities to control: the possibility that $\{X_t\}_{t\geq 0}$ goes above $-\delta$, and the possibility that it goes below $-\sigma^{-10}$. The latter is easier to control; by (4.2) and Markov's inequality, for all $\epsilon > 0$ there exist constants $c_1 = c_1(\epsilon)$, $C_1 = C_1(\epsilon) > 0$ such that

$$\begin{aligned}\sup_{|X_0|\leq \sigma^{-\beta}} \mathbb{P}\left[\min_{1\leq t\leq e^{\sigma^{-\beta}}} X_t < -\sigma^{-\beta-\epsilon}\right] &\leq e^{\sigma^{-\beta}} \sup_{|X_1|\leq \sigma^{-\beta}} \sup_{1\leq t\leq e^{\sigma^{-\beta}}} \frac{\mathbb{E}[e^{|X_t|}]}{e^{\sigma^{-\beta-\epsilon}}} \\ &\leq e^{\sigma^{-\beta}} \left(\frac{e^{\sigma^{-\beta}} + \alpha^{-1}C}{e^{\sigma^{-\beta-\epsilon}}}\right) \\ &\leq C_1 e^{-c_1\sigma^{-\beta-\epsilon}} \end{aligned} \tag{A.5}$$

uniformly in $\beta \geq 1$.

The possibility that $\{X_t\}_{t\geq 0}$ goes above $-\delta$ cannot be controlled in the same way, because it does not have negligible probability on the timescale of interest. Instead, we use the fact that X_t will generally exit the interval $(-\delta, 0)$ fairly quickly, often to the interval $(-\infty, -\delta)$.

To see this, fix $x \in (-\delta, 0)$ and let $\{X_t\}_{t\geq 0}$ have starting point $X_0 = x$. Next, let $\{\epsilon_t\}_{t\geq 1}$ be a sequence of independent and identically distributed $\mathcal{N}(0, \sigma^2)$ random variables and let $Y_t = X_0 + \sum_{s=1}^{t} \epsilon_t$. For $I \subset \mathbb{R}$, let $\psi_{x,I} = \inf\{t \geq 0 : Y_t \in I\}$ be the hitting time of I for the Markov chain $\{Y_t\}_{t\geq 0}$. Observing the forward mapping representation of K_σ in Definition 3, and that f_σ is monotone on $(-\delta, 0)$, it is clear that we can couple $\{X_t\}_{t\geq 0}$, $\{Y_t\}_{t\geq 0}$ so that

$$X_t \leq Y_t \text{ for all } 0 \leq t \leq \min(\tau_{x,(-\delta,0)^c}, \psi_{x,(-\delta,0)^c}). \tag{A.6}$$

But, by standard calculations for a simple random walk,

$$\sup_{x\in(-\delta,0)} \mathbb{P}[\psi_{x,(-\delta,0)^c} > C_2\sigma^{-c_2}] \leq \sigma^2, \qquad \inf_{x\in(-\delta,0)} \mathbb{P}[Y_{\psi_{x,(-\delta,0)^c}} < -\delta] > C_3\sigma \tag{A.7}$$

for some constants c_2, C_2, C_3 that do not depend on σ. (To see the first inequality in (A.7), note that a direct calculation for Gaussians gives $\sup_{x\in(-\delta,0)} \mathbb{P}[\psi_{x,(-\delta,0)^c} > C_2'\sigma^{-c_2'}] > C_2'' > 0$ for some C_2', $c_2'C_2'' > 0$; applying the strong Markov property to iterate this bound as in the proof of (3.5) gives the desired conclusion. The second inequality in (A.7) follows from the observation that $\mathbb{P}[Y_1 > C_3'\sigma] > C_3'' > 0$ for some constants C_3', $C_3'' > 0$ and then the well-known 'gambler's ruin' calculation (see, e.g., Section 10.14.4 of [18]).)

Combining (A.7) with (A.6) and noting that $\{X_t\}_{t\geq 0}$ never exits $(-\infty, 0)$ by construction, we find

$$\sup_{x\in(-\delta,0)} \mathbb{P}[\tau_{x,(-\delta,0)^c} < C_2\sigma^{-c_2}] \geq C_3\sigma - \sigma^2 = C_3(\sigma)(1 - o(1)).$$

Noting that these bounds are uniform over the starting point $X_0 \in (-\delta, 0)$, we find, for $k \in \mathbb{N}$, $\sup_{x\in(-\delta,0)} \mathbb{P}[\tau_{x,(-\delta,0)^c} < k\, C_2\sigma^{-c_2}] \geq 1 - (1 - C_3(\sigma) - \sigma^2)^k$. Taking k very large ($k > \sigma^{-12}$ suffices) gives $\sup_{x\in(-\delta,0)} \mathbb{P}[\tau_{x,(-\delta,0)^c} < C_4\sigma^{-c_4}] \geq 1 - e^{-\sigma^{-10}}$ for some constants $0 \leq c_4, C_4 < \infty$, which is exactly (4.5).

Combining the bound (4.3) on the mixing of $\hat{K}_\sigma$ on $(-\sigma^{-11}, -\delta)$ with the bound (A.5) on the possibility of excursions below $-\sigma^{-11}$ and the bound (4.5) on the length of excursions above $-\delta$ completes the proof of the lemma.

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

Appendix B. Technical bounds from the proof of Theorem 4

We obtain the required bound on the mixing time, which is essentially a corollary of [9, Theorem 1.1].

Lemma 8. *Fix sequences of convex sets satisfying Assumptions 3, and denote by L_β the hit-and-run kernel on $\mathcal{C}_\beta \cap S = \mathsf{A}_\beta \cap S$, with stationary measure π_β the uniform distribution on this set. Then there exists some constant $0 < C < \infty$ that does not depend on β (though depending on d) such that, for all $\epsilon > 0$ and all $T > C \log(\epsilon^{-1})$, $\sup_{x \in \mathcal{C}_\beta \cap S} \|L_\beta^T(x, \cdot) - \pi_\beta(\cdot)\|_{\mathrm{TV}} \leq \epsilon$.*

Proof. For $0 < \delta < 0.15$, define $\mathsf{A}_{\beta,\delta} = \{x \in \mathsf{A}_\beta \cap S : \inf_{y \notin \mathsf{A}_\beta \cap S} \|x - y\| < \delta\}$ to be the points at least distance δ from the boundary of $\mathsf{A}_\beta \cap S$. Recall from [9, Corollary 1.2] that $\sup_{x \in \mathsf{A}_{\beta,\delta}} \|L_\beta^t(x, \cdot) - \pi_\beta(\cdot)\|_{\mathrm{TV}} \leq 0.25$ for all $t > C_1(\delta) \equiv 4 \times 10^{11} n^3 \frac{\mathsf{R}^2}{\mathsf{r}^2} \log(4\delta^{-1})$.

Applying the Markov property and then this bound (together with the well-known fact that TV distance to stationarity decays exponentially quickly, as shown, e.g., in [8, Lemmas 4.11, 4.12], we have, for all $t_1, t_2 \in \mathbb{N}$,

$$\begin{aligned}\sup_{x \in \mathsf{A}_\beta \cap S} \left\| L_\beta^{t_1+t_2}(x, \cdot) - \pi_\beta(\cdot) \right\|_{\mathrm{TV}} &\leq \sup_{x \in \mathsf{A}_\beta \cap S} \mathbb{P}_x[\tau_{\mathsf{A}_{\beta,\delta}} > t_1] + \sup_{x \in \mathsf{A}_{\beta,\delta}} \left\| L_\beta^{t_2}(x, \cdot) - \pi_\beta(\cdot) \right\|_{\mathrm{TV}} \\ &\leq \sup_{x \in \mathsf{A}_\beta \cap S} \mathbb{P}_x[\tau_{\mathsf{A}_{\beta,\delta}} > t_1] + 2^{-\left\lfloor \frac{t_2}{C_1} \right\rfloor}. \end{aligned} \tag{B.1}$$

Fix any point $x \in \mathsf{A}_\beta$. We now consider the hit-and-run step defined by sampling v uniformly on the $(d-1)$-dimensional unit sphere, setting $\ell = \{x + sv : s \in \mathbb{R}^d\} \cap \mathsf{A}_\beta$, and sampling $X \sim \mathrm{Unif}(\ell)$. Recall that $\mathsf{A}_\beta := \hat{\mathsf{A}}_\beta + B(0, \Delta)$ is the Minkowski sum of a convex body with a ball of radius Δ. This implies that x lies on the surface of some ball $\mathfrak{B} \subseteq \mathsf{A}_\beta$, where $\mathfrak{B}$ has radius $\frac{1}{2}\Delta$. Denote by n_x the normal vector to the tangent plane of $\mathfrak{B}$ at x, and let $v^\perp$ be the component of v in the direction of n_x. Since the vector v which determines the hit-and-run step is uniformly distributed on the $(d-1)$-dimensional unit sphere, standard concentration inequalities for the uniform distribution on the sphere imply that

$$\mathbb{P}\left[\|v^\perp\| \geq \frac{1}{3\sqrt{d}} \right] \geq \frac{1}{2}.$$

It is an exercise in two-dimensional Euclidean geometry to check that, when $\|v^\perp\| \geq \frac{1}{3\sqrt{d}}$, there exists some $z \in \ell$ such that $B_{\frac{1}{20\sqrt{d}}}(z) \subseteq \mathfrak{B} \subseteq \mathsf{A}_\beta$ (see Figure B1(a) for a picture proof of this fact). Denote by $\mathcal{E}$ the event that there exists $z \in \ell : B_{\frac{1}{20\sqrt{d}}}(z) \subseteq \mathfrak{B} \subseteq \mathsf{A}_\beta$; we have shown that

$$\mathbb{P}[\mathcal{E}] \geq \frac{1}{2}. \tag{B.2}$$

We now continue by analysing the distribution of X on the event $\mathcal{E}$, and fix the $z \in \ell$ whose existence is guaranteed as above. We note that $\mathcal{E}$ depends only on v. Let a,b be the two intersection points of ℓ with the convex body A_β, and consider the convex hull $\Gamma = \mathrm{Convex}\left(\{a\} \cup \{b\} \cup B_{\frac{1}{20\sqrt{d}}\Delta}(z)\right) \subset \mathsf{A}_\beta$ of these two points and the ball $B_{\frac{1}{20\sqrt{d}}\Delta}(z)$. Let m be the midpoint of the line segment $[z, a]$, and let M be the midpoint of the line $[z, b]$. Since the convex hull Γ consists of the union of two prisms which share a base consisting of a $(d-1)$-dimensional ball of radius $\frac{1}{20\sqrt{d}}\Delta$ centred at the point z, we can see that all the points on the

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press

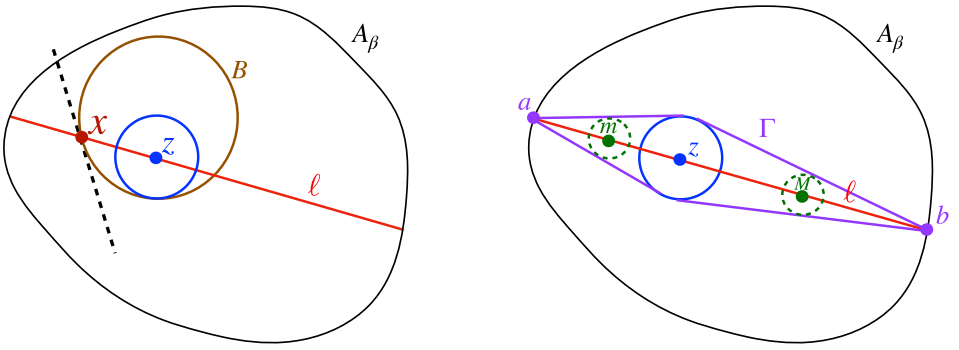

FIGURE B1. An illustration of the convex geometry in the proof of Lemma 8.

lines $[z, m]$ and $[z, M]$ are in the $\frac{1}{40\sqrt{d}}\Delta$-interior of $\Gamma \subseteq \mathsf{A}_\beta$; see Figure 3(b). On the event $\mathcal{E}$, then,

$$\mathbb{P}\left[X \in \mathsf{A}_\beta^{\left(-\frac{1}{40\sqrt{d}}\Delta\right)} \mid \mathcal{E}\right] \geq \mathbb{P}[X \in [z, m] \cup [z, M] \mid \mathcal{E}] \geq \frac{1}{2}.$$

Combining this with (B.2), we have shown that

$$\mathbb{P}\left[X \in \mathsf{A}_\beta^{\left(-\frac{1}{40\sqrt{d}}\Delta\right)}\right] \geq \frac{1}{4}. \tag{B.3}$$

Now, our assumption that $\mathsf{K}_\beta \cap S \subseteq \mathsf{A}_\beta \cap S$ implies that $\mathcal{C}_\beta \cap S = \mathsf{A}_\beta \cap S$. Moreover, by another one of our assumptions, S contains the $\frac{1}{80\sqrt{d}}\Delta$-interior of A_β. Therefore, applying (B.3) gives

$$\inf_x L_\alpha\left(x, \mathcal{C}_{\beta, \frac{1}{80\sqrt{d}}\Delta}\right) \geq \frac{1}{4}. \tag{B.4}$$

Applying this to (B.1) with the choice $\delta = \frac{1}{80\sqrt{d}}\Delta$, we find

$$\sup_{x \in \mathcal{C}_\beta \cap S} \left\| L_\beta^{t_1+t_2}(x, \cdot) - \pi_\beta(\cdot) \right\|_{\mathrm{TV}} \leq \left(1 - \frac{1}{4}\right)^{t_1} + 2^{-\left\lfloor \frac{t_2}{C_1} \right\rfloor}.$$

Choosing $t_1 = \left\lceil \frac{\log(\epsilon)}{\log(1-\frac{1}{4})} \right\rceil$ and $t_2 = \frac{\log(\epsilon)}{2C_1\left(\frac{1}{80\sqrt{d}}\Delta\right)}$ gives

$$\sup_{x \in \mathcal{C}_\beta \cap S} \left\| L_\beta^{t_1+t_2}(x, \cdot) - \pi_\beta(\cdot) \right\|_{\mathrm{TV}} \leq \epsilon.$$

□

Acknowledgements

We would like to thank Neil Shephard and Gareth Roberts for some questions about multimodal distributions and helpful comments on initial results.

NSP gratefully acknowledges the support of ONR. AS was supported by an NSERC Discovery grant. OM was supported by NSF-dms 1312831, by an NSERC Discovery grant, and by a Canadian Statistical Sciences Institute (CANSSI) Postdoctoral Fellowship.

References

[1] Beltrán, J. and Landim, C. (2015). A martingale approach to metastability. *Prob. Theory Relat. Fields* **161**, 267–307.

[2] Bovier, A. and Den Hollander, F. (2006). *Metastability: A Potential Theoretic Approach*. Springer, New York.

[3] Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the Laplacian. In *Problems in Analysis*, ed. R. C. Gunning, Princeton University Press, pp. 195–199.

[4] Griffeath, D. (1975). A maximal coupling for Markov chains. *Z. Wahrscheinlichkeitsth.* **31**, 95–106.

[5] Jerrum, M., Son, J.-B., Tetali, P. and Vigoda, E. (2004). Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains. *Ann. Appl. Prob.* **14**, 1741–1765.

[6] Landim, C. (2018). Metastable Markov chains. Preprint. arXiv:1807.04144

[7] Lawler, G. F. and Sokal, A. D. (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Am. Math. Soc.* **309**, 557–580.

[8] Levin, D. A., Peres, Y. and Wilmer, E. L. (2009). *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI.

[9] Lovász, J. and Vempala, S. (2006). Hit-and-run from a corner. *SIAM J. Comp.* **35**, 985–1005.

[10] Madras, N. and Randall, D. (2002). Markov chain decomposition for convergence rate analysis. *Ann. Appl. Prob.* **12**, 581–606.

[11] Mangoubi, O., Pillai, N. S. and Smith, A. (2018). Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? Preprint. arXiv:1808.03230.

[12] Mangoubi, O. and Smith, A. (2017). Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 1: Continuous dynamics. Preprint.

[13] Mangoubi, O. and Smith, A. (2017). Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. Preprint. arXiv:1708.07114.

[14] Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–121.

[15] Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Prob.* **4**, 981–1011.

[16] Olivieri, E. and Vares, M. E. (2005). *Large Deviations and Metastability*. Cambridge University Press.

[17] Pillai, N. S. and Smith, A. (2017). Elementary bounds on mixing times for decomposable Markov chains. *Stoch. Process. Appl.* **127**, 3068–3109.

[18] Resnick, S. I. (2013). *A Probability Path*. Springer, New York.

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

[20] Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Prob.* **2**, 13–25.

[21] Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Am. Statist. Assoc.* **90**, 558–566.

[22] Turchin, V. F. (1971). On the computation of multidimensional integrals by the Monte-Carlo method. *Theory Prob. Appl.* **16**, 720–724.

[23] Woodard, D., Schmidler, S. and Huber, M. (2009). Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electron. J. Prob.* **14**, 780–804.

[24] Woodard, D. B., Schmidler, S. C. and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Prob.* **19**, 617–640.

https://doi.org/10.1017/jpr.2020.83 Published online by Cambridge University Press