Restricted Spectral Gap Decomposition for Simulated Tempering Targeting Mixture Distributions

Jhanvi Garg
Department of Statistics
Texas A&M University
College Station, TX 77843
gargjhanvi@tamu.edu

Krishnakumar Balasubramanian
Department of Statistics
University of California
Davis, CA 95616
kbala@ucdavis.edu

Quan Zhou
Department of Statistics
Texas A&M University
College Station, TX 77843
quan@stat.tamu.edu

Abstract

Simulated tempering is a widely used strategy for sampling from multimodal distributions. In this paper, we consider simulated tempering combined with an arbitrary local Markov chain Monte Carlo sampler and present a new decomposition theorem that provides a lower bound on the restricted spectral gap of the algorithm for sampling from mixture distributions. By working with the restricted spectral gap, the applicability of our results is extended to broader settings such as when the usual spectral gap is difficult to bound or becomes degenerate. We demonstrate the application of our theoretical results by analyzing simulated tempering combined with random walk Metropolis–Hastings for sampling from mixtures of Gaussian distributions. Our complexity bound scales polynomially with the separation between modes, logarithmically with $1/\varepsilon$, where ε denotes the target accuracy in total variation distance, and exponentially with the dimension d.

1 Introduction

Efficient sampling from complex distributions is a foundational problem with numerous applications across various fields, including computational statistics [Robert et al., 1999, Liu and Liu, 2001, Brooks et al., 2011, Owen, 2013], Bayesian inference [Gelman et al., 2013], statistical physics [Newman and Barkema, 1999, Landau and Binder, 2021], and finance [Dagpunar, 2007]. These distributions are often multimodal, reflecting underlying heterogeneity in the data. Sampling from such distributions presents challenges closely related to those encountered in non-convex optimization, where objective functions with multiple local minima require methods capable of effectively exploring the solution space. While discretizations of Langevin dynamics (for a comprehensive overview, see Chewi [2024]) excel in log-concave settings where gradients reliably guide the sampler toward a single mode, they tend to be less effective in multimodal landscapes, which require strategies capable of navigating between separated modes. The assumption of dissipativity, which limits the growth rate of the potential, has been widely used in previous works to establish better convergence rates for Langevin Monte Carlo (LMC) in such settings [Raginsky et al., 2017, Durmus and Moulines, 2017, Erdogdu et al., 2018, Erdogdu and Hosseinzadeh, 2021, Mou et al., 2022, Mousavi-Hosseini et al., 2023]. More recently, Balasubramanian et al. [2022] characterized the performance of averaged LMC for target densities that are only Hölder continuous, without relying on functional inequalities or curvature-based assumptions; they measured the convergence rate using the weaker Fisher information metric.

For distributions that deviate significantly from log-concavity and exhibit numerous deep modes, additional techniques are often required to ensure efficient sampling. A comprehensive discussion of the fundamental challenges in sampling from multimodal distributions, as well as an overview of major types of Markov chain Monte Carlo (MCMC) algorithms designed for this purpose—including

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

parallel tempering, mode jumping, and Wang–Landau methods—can be found in Łatuszyński et al. [2025]. The recent work of Koehler et al. [2024] addresses the problem of sampling from a multimodal distribution by initializing the sampler using a small number of stationary samples. They show that, under the assumption of a k-th order spectral gap, the chain efficiently yields an ε-accurate sample in total variation (TV) distance with $\widetilde{O}(k/\varepsilon^2)$ samples, and they further extend the result to scenarios where the score (i.e., the drift term of LMC) is estimated. Lee and Santana-Gijzen [2024] studied the sequential Monte Carlo (SMC) method [Liu and Chen, 1998, Del Moral et al., 2006, Chopin et al., 2020, Syed et al., 2024] and derived the complexity result for mixture target distributions. Additionally, denoising-diffusion-based samplers have been proposed for sampling from non-log-concave targets without relying on isoperimetric inequalities [Huang et al., 2023, 2024, He et al., 2024b]. These methods reverse the Ornstein–Uhlenbeck process and require estimating score functions (i.e., gradients of the log-density) via importance sampling, which becomes particularly challenging in high-dimensional settings.

A widely used strategy for tackling multimodality is annealing or tempering, which leverages a sequence of distributions to gradually explore complex landscapes. Guo et al. [2024] provided a non-asymptotic analysis of annealed Langevin Monte Carlo [Neal, 2001], highlighting its provable benefits for non-log-concave sampling. They demonstrated that a simple annealed Langevin Monte Carlo algorithm achieves ε^2 accuracy in Kullback-Leibler divergence with an oracle complexity of $\widetilde{O}\left(d/\varepsilon^6\right)$. Chehab et al. [2024] develops a comprehensive theoretical framework for tempered Langevin dynamics, providing convergence guarantees in Kullback-Leibler divergence across a variety of tempering schedules. Simulated tempering [Marinari and Parisi, 1992] has also been introduced as a method to promote transitions between modes. By dynamically adjusting the "temperature" of the distribution, simulated tempering effectively smooths the energy landscape, allowing the sampler to escape local modes and traverse between high-probability regions. The sampler gradually returns to the original distribution as the temperature decreases. A version of simulated tempering was also used in Koehler et al. [2022] to sample from multimodal Ising models. Parallel tempering [Swendsen and Wang, 1986] shares a similar mechanism with simulated tempering, but runs multiple chains in parallel at different temperatures, allowing exchanges between them to improve mixing. Zheng [2003] established that the spectral gap of simulated tempering chain is bounded below by a multiple of the spectral gap of parallel tempering chain and a bound depending on the overlap between distributions at adjacent temperatures. Woodard et al. [2009] provided a lower bound on the spectral gap for a simulated tempering chain using the state space decomposition technique [Madras and Randall, 2002]. This method analyzes the probability flow by dividing the state space into subsets and studying transitions within and between them. However, determining an optimal partition into subsets often requires complex spectral partitioning arguments, and estimating the spectral gap through conductance methods can introduce a squared-factor loss due to Cheeger's inequality.

An alternative approach is to decompose the Markov chain directly rather than the state space, particularly when the target distribution is a mixture distribution or closely resembles one. This method can yield potentially tighter spectral gap bounds by separately analyzing two types of chains: local chains that efficiently explore each mixture component at every temperature level and a projected chain that governs transitions between different components and temperature levels. This decomposition technique was introduced in Ge et al. [2018] as a framework for bounding the spectral gap of the simulated tempering chain combined with LMC for mixtures of strongly log-concave distributions that are translates of each other. They established that the runtime required to reach an ε TV distance from the target distribution depends polynomially on the dimension d, the mode separation D, and the inverse accuracy $1/\varepsilon$.

We introduce a new Markov chain decomposition theorem for discrete-time chains, in contrast to the continuous-time framework adopted by Ge et al. [2018]. Since spectral gap bounds of discrete-Markov chains (e.g. Metropolis–Hastings algorithms) over unbounded Euclidean spaces is often difficult to obtain, we propose to directly decompose the restricted spectral gap [Atchadé, 2021], which, roughly speaking, can be thought of as the spectral gap of the Markov chain restricted to a subset of the state space. Geometric tools such as the path methods developed by Yuen [2000] can be used to lower bound the restricted spectral gap. Intuitively, if each local chain mixes fast in a large subset of the space, where the target distribution concentrates, and the projected chain also mixes well, then the overall simulated tempering chain should converge fast. The projected chain can be

constructed in many ways. We present a construction which differs substantially from that of Ge et al. [2018]; it arises naturally from the structure of the algorithm and significantly simplifies the proof.

The remainder of the paper is organized as follows. In Section 2, we present our decomposition theorem. In Section 3, we apply this theorem to analyze the simulated tempering combined with the random walk Metropolis–Hastings (STMH) algorithm for sampling from mixtures of Gaussian distributions. In Section 4, we empirically validate our convergence guarantee for the STMH algorithm by sampling from a two-dimensional Gaussian mixture. Finally, Section 5 summarizes our key contributions and highlights promising directions for future research. Detailed proofs are provided in the appendix.

2 A New Decomposition Theorem

2.1 Notation and Definitions

We begin by introducing the necessary notation that will be used throughout the paper. We adopt the convention that uppercase letters denote probability distributions (or transition kernels) and the corresponding lowercase letters their densities (or transition densities). For example, P will be used to denote the probability distribution with density p. We use $\mathcal{L}^2(\Omega, \Pi)$ to denote the space of all real-valued functions defined on Ω that are square-integrable with respect to a measure Π.

Definition 1 (Restricted Spectral Gap)**.** *Let K be a Markov transition kernel with state space Ω and stationary probability measure Π. Let $\Omega^0 \subseteq \Omega$ be measurable such that $\Pi(\Omega^0) > 0$. The Ω^0-restricted spectral gap of K, denoted by $\mathrm{SpecGap}_{\Omega^0}(K)$, is defined as*

$$\mathrm{SpecGap}_{\Omega^0}(K) = \inf_{g \in \mathcal{L}^2(\Omega,\Pi)} \frac{\mathcal{E}_{\Omega^0}(g, g; \Pi, K)}{\mathrm{Var}_{\Pi,\Omega^0}(g)},$$

where

$$\mathcal{E}_{\Omega^0}(g, g; \Pi, K) = \frac{1}{2} \int_{\Omega^0} \int_{\Omega^0} \big[g(\omega') - g(\omega)\big]^2 \, \Pi(\mathrm{d}\omega)\, K(\omega, \mathrm{d}\omega'),$$
$$\mathrm{Var}_{\Pi,\Omega^0}(g) = \frac{1}{2} \int_{\Omega^0} \int_{\Omega^0} \big[g(\omega') - g(\omega)\big]^2 \, \Pi(\mathrm{d}\omega)\, \Pi(\mathrm{d}\omega').$$

We will refer to $\mathcal{E}_{\Omega^0}(g, g; \Pi, K)$ as the Ω^0-restricted Dirichlet form and omit Π, K when they are clear from the context. When $\Omega^0 = \Omega$, $\mathrm{SpecGap}_{\Omega^0}(K)$ is known as the spectral gap of K, and we simply write $\mathrm{SpecGap}(K) = \mathrm{SpecGap}_{\Omega}(K), \mathcal{E}(g, g) = \mathcal{E}_{\Omega}(g, g)$ and $\mathrm{Var}_{\Pi}(g) = \mathrm{Var}_{\Pi,\Omega_0}(g)$. Note that $\mathrm{Var}_{\Pi}(g)$ equals the variance of $g(\omega)$ with $\omega \sim \Pi$. Intuitively, the spectral gap quantifies how rapidly a Markov chain mixes: a larger gap corresponds to faster convergence to stationarity distribution. The restricted spectral gap generalizes this idea by measuring the rate of mixing in a subset Ω^0 of the state space.

Simulated tempering can be viewed as a bivariate Markov chain. The key idea is to augment the state space with a temperature index that allows the chain to move between a family of distributions, ranging from the target distribution (low temperature) to flattened versions (high temperature). At higher temperatures the landscape is smoother, enabling easier transitions between modes, while samples from the target are obtained by collecting states only at the lowest temperature. Formally, it is defined as follows.

Definition 2 (Simulated Tempering Markov Chain)**.** *Let $L \geq 2$ be an integer and $[L] = \{1, 2, \ldots, L\}$. Let $\lambda \in (0, 1)$ and $(r_i)_{i\in[L]}$ be constants such that $r_i > 0$ for each i, and $\sum_{i=1}^{L} r_i = 1$. Let M_i, for each $i \in [L]$, be the transition kernel of a Markov chain with state space $\mathcal{X}$ and stationary density p_i. The simulated tempering Markov chain has state space $[L] \times \mathcal{X}$. Denote its transition kernel by*

$$M = M\left((p_i)_{i=1}^{L}, (r_i)_{i=1}^{L}, (M_i)_{i=1}^{L}, \lambda\right),$$

which has density

$$m((i, x), (i', x')) = \begin{cases} (1-\lambda) m_i(x, x'), & \text{if } i = i', x \neq x', \\ \frac{\lambda}{2} a((i, x), (i', x)), & \text{if } i' = i \pm 1, x = x', \end{cases}$$

where

$$a((i, x), (i', x)) = \min\left\{\frac{r_{i'} p_{i'}(x)}{r_i p_i(x)}, 1\right\}. \tag{1}$$

In words, the simulated tempering Markov chain evolves as follows. Given current state (i, x), we sample $u \sim \text{Bernoulli}(\lambda)$. If $u = 0$, draw x' from $M_i(x, \cdot)$ and move to (i, x'). If $u = 1$, propose $i' = i \pm 1$, each with equal probability $1/2$, and accept the proposal with probability $a((i, x), (i', x))$ (if $i' \notin [L]$, the proposal is always rejected). It is easy to check that the stationary distribution P of the simulated tempering Markov chain M has density given by

$$p(i, x) = r_i p_i(x).$$

2.2 Decomposition of the Simulated Tempering Markov chain

Consider the simulated tempering Markov chain given in Definition 2. Let $\mathcal{X}^0 \subset \mathcal{X}$ be a measurable subset of $\mathcal{X}$, and our goal is to derive a decomposition theorem for the spectral gap of M restricted to $[L] \times \mathcal{X}^0$. We assume that the distributions $(P_i)_{i\in[L]}$ satisfies the following assumption.

Assumption 1. *For each $i \in [L]$, the stationary density p_i can be expressed as a mixture of n component densities:*

$$p_i(x) = \sum_{j=1}^{n} w_{(i,j)} p_{(i,j)}(x),$$

where $w_{(i,j)} \geq 0$ for each $j \in [n]$ and $\sum_{j=1}^{n} w_{(i,j)} = 1$.

For simplicity, in Assumption 1 we assume that each stationary density p_i can be expressed as a mixture of simpler component densities, an assumption also made in Ge et al. [2018]. Under Assumption 1, we can introduce a latent variable J so that given $I = i, J = j$, X is drawn from the component distribution $P_{(i,j)}$. The density of the target distribution is augmented to $p(i, j, x) = r_i w_{(i,j)} p_{(i,j)}(x)$. The overall strategy of our decomposition is similar to Theorem 6.3 of Ge et al. [2018] in that we directly decompose the Dirichlet form of M into the Dirichlet forms of Markov chains $(M_{(i,j)})_{i\in[L],j\in[n]}$, where $M_{(i,j)}$ has stationary distribution $P_{(i,j)}$. We assume that these chains are chosen such that the following assumption is satisfied.

Assumption 2. *Let $\mathcal{E}_{i,\mathcal{X}^0}$ denote the $\mathcal{X}^0$-restricted Dirichlet form of M_i. For each $i \in [L], j \in [n]$, let $\mathcal{E}_{(i,j),\mathcal{X}^0}$ be the $\mathcal{X}^0$-restricted Dirichlet form of a Markov chain $M_{(i,j)}$ on $\mathcal{X}$ with stationary density $p_{(i,j)}$. For each i and any $g_i \in \mathcal{L}^2(\mathcal{X}, P_i)$,*

$$\sum_{j=1}^{n} w_{(i,j)} \mathcal{E}_{(i,j),\mathcal{X}^0}(g_i, g_i) \leq C_1 \mathcal{E}_{i,\mathcal{X}^0}(g_i, g_i), \tag{2}$$

where $C_1 > 0$ is some constant. Further, each $M_{(i,j)}$ satisfies the following inequality with some constant $C_2 > 0$:

$$\text{Var}_{P_{(i,j)},\mathcal{X}^0}(g_i) \leq C_2 \mathcal{E}_{(i,j),\mathcal{X}^0}(g_i, g_i). \tag{3}$$

Assumption 2 assumes a lower bound on the mixing rate of the local chains $M_{(i,j)}$, each targeting a single component $p_{(i,j)}$. Efficient mixing within each component is essential, as poor local mixing can limit global exploration and slow the overall convergence of the simulated tempering chain.

Remark 1. In Ge et al. [2018], each M_i is a continuous-time Langevin diffusion with stationary distribution P_i. In this case, condition (2) can be easily satisfied by letting $M_{(i,j)}$ be the Langevin diffusion having stationary distribution $P_{(i,j)}$, since

$$\mathcal{E}_i(g_i, g_i) = \int \|\nabla g_i\|^2 p_i \, \mathrm{d}x = \int \|\nabla g_i\|^2 \sum_{j=1}^{n} w_{(i,j)} p_{(i,j)} \, \mathrm{d}x = \sum_{j=1}^{n} w_{(i,j)} \mathcal{E}_{(i,j)}(g_i, g_i),$$

which yields Equation (2) with $C_1 = 1$. However, in our setting each M_i is a discrete-time Markov chain (e.g. a Metropolis–Hastings algorithm), and finding Markov chains $M_{(i,j)}$ satisfying (2) may not be trivial.

Remark 2. Condition (3) implies that the $\mathcal{X}^0$-restricted spectral gap of $M_{(i,j)}$ is at least C_2^{-1}. In contrast, Ge et al. [2018] requires each "local chain" to have a positive spectral gap. By weakening this condition, we can develop a decomposition theorem that is particularly useful in settings where the usual spectral gap is difficult to bound over the entire state space.

Loosely speaking, condition (2) allows us to view the dynamics of M_i as governed by the hidden variable J. Conditional on $J = j$, the behavior of M_i can be approximated by $M_{(i,j)}$. The constant C_2 in (3) measures the convergence rate of each $M_{(i,j)}$ on $\mathcal{X}^0$. To bound the convergence rate of the simulated tempering Markov chain M, we need one more assumption characterizing the transitions between any (i, j) and (i', j'). To this end, we construct a projected chain as follows.

Definition 3. *Let M be the simulated tempering Markov chain given in Definition 2 and Assumption 1 hold. Define*

$$a((i,j,x),(i',j',x)) = \min\left\{\frac{r_{i'}w_{(i',j')}p_{(i',j')}(x)}{r_i w_{(i,j)} p_{(i,j)}(x)},\ 1\right\}, \quad p_i(j|x) = \frac{w_{(i,j)}p_{(i,j)}(x)}{p_i(x)}.$$

Define the projected chain with transition matrix $\overline{M}$ on $[L] \times [n]$ by

$$\overline{M}((i,j),(i',j')) = \begin{cases} (1-\lambda)\displaystyle\int_{\mathcal{X}^0} \frac{p_{(i,j)}(x)}{P_{(i,j)}(\mathcal{X}^0)}\, p_i(j'|x)\mathrm{d}x, & \text{if } i = i',\ j \neq j', \\ \frac{\lambda}{2}\displaystyle\int_{\mathcal{X}^0} \frac{p_{(i,j)}(x)}{P_{(i,j)}(\mathcal{X}^0)} a((i,j,x),(i',j,x))\mathrm{d}x, & \text{if } i' = i \pm 1,\ j = j', \\ 1 - \sum_{(k,l)\neq(i,j)} \overline{M}((i,j),(k,l)), & \text{if } (i',j') = (i,j). \end{cases}$$

It is easy to prove that $\overline{M}$ is indeed a transition rate matrix (i.e., all entries are non-negative and each row sums to one.) By checking the detailed balance condition, one obtains the following result.

Lemma 1. *The stationary distribution of $\overline{M}$ is given by*

$$\overline{P}(i,j) = \frac{r_i w_{(i,j)} P_{(i,j)}(\mathcal{X}^0)}{P([L] \times \mathcal{X}^0)}.$$

Proof. See Appendix B. □

Assumption 3. *Let $\overline{\mathcal{E}}$ be the Dirichlet form of $\overline{M}$. Then, for any $\overline{g}\colon [L] \times [n] \to \mathbb{R}$, $\overline{M}$ satisfies the following inequality for some constant $C_3 > 0$:*

$$\mathrm{Var}_{\overline{P}}(\overline{g}) \leq C_3 \overline{\mathcal{E}}(\overline{g}, \overline{g}). \tag{4}$$

Assumption 3 assumes a lower bound on the mixing rate of the projected chain. Intuitively, if the components are well-separated, transitions between them becomes rare, resulting in slow mixing of the projected chain and, consequently, slow overall convergence of the simulated tempering chain. Conversely, when components are closer and transitions are more likely, the chain mixes more rapidly.

Remark 3. Our construction of the projected chain $\overline{M}$ is significantly different from that in Ge et al. [2018]. In particular, Ge et al. [2018] defined the transition between j, j' by

$$\overline{M}((i,j),(i,j')) \propto \frac{w_{(i,j')}}{\chi^2_{\max}(p_{(i,j)} \| p_{(i,j')})}$$

where $\chi^2_{\max}(P\|Q) := \max\{\chi^2(P\|Q), \chi^2(Q\|P)\}$ for two distributions P and Q. In Lemmas 4 and 5 in Appendix B, we show how to bound the Dirichlet form of our projected chain $\overline{M}$, denoted by $\overline{\mathcal{E}}$. Compared to the approach of Ge et al. [2018], our argument for bounding $\overline{\mathcal{E}}$ is more straightforward and yields simpler and equally tight bound on the Poincaré constant of the simulated tempering chain in the decomposition theorem.

We are now ready to state our main decomposition theorem, which relates the restricted spectral gap of M to $\mathrm{SpecGap}_{\mathcal{X}^0}(M_{(i,j)})$ for $i \in [L], j \in [n]$ and $\mathrm{SpecGap}(\overline{M})$. Roughly speaking, this result guarantees that the restricted spectral gap of M is $\Omega((C_1C_2C_3)^{-1})$.

Theorem 1. *Consider the simulated tempering Markov chain M given in Definition 2. Suppose Assumptions 1, 2 and 3 hold, and define*

$$\theta = P([L] \times \mathcal{X}^0), \quad \phi = \min_{i,j} P_{(i,j)}(\mathcal{X}^0).$$

Then, we have $\mathrm{SpecGap}_{[L]\times\mathcal{X}^0}(M) \geq C_M^{-1}$ where

$$C_M = \max\left\{3\theta C_3,\ \frac{\theta C_1 C_2}{\phi(1-\lambda)}\left((2+\lambda)C_3 + 1\right)\right\} \mathcal{E}_{[L]\times\mathcal{X}^0}(g,g). \tag{5}$$

Proof. See Appendix B. □

When $\mathcal{X}^0 = \mathcal{X}$, we obtain a bound on the spectral gap of M, which is similar in spirit to the continuous-time decomposition theorem of Ge et al. [2018], though the two results are not directly comparable due the discrete-time setting and our construction of $\overline{M}$.

Corollary 1. *Consider the simulated tempering Markov chain M given in Definition 2. Suppose Assumptions 1, 2 and 3 hold with $\mathcal{X}^0 = \mathcal{X}$. Then,* $\mathrm{SpecGap}(M) \geq C_M^{-1}$ *where C_M is given by* (5)*.*

Proof. This immediately follows from Theorem 1. □

2.3 Mixing Time Bounds

Given two probability distributions Π_1 and Π_2 with densities π_1 and π_2, we define their total variation distance by $\|\Pi_1 - \Pi_2\|_{\mathrm{tv}} = \int |\pi_1(x) - \pi_2(x)| \mathrm{d}x$, which takes values in $[0, 2]$. In addition to Assumptions 1 to 3, we need to further require that M is both reversible and lazy, where "lazy" means that $M((i, x), \{(i, x)\}) \geq 1/2$ for any $(i, x) \in [L] \times \mathcal{X}$. Both conditions are very mild and commonly used in the sampling literature. Then, if M has a positive spectral gap, M converges exponentially fast to its stationary distribution P in TV distance [Montenegro et al., 2006]. It was shown in Atchadé [2021] that a positive restricted spectral gap can also be used to obtain an upper bound on the mixing time. The following lemma follows from the result of Atchadé [2021], where we also characterize the exponential convergence rate at each temperature level.

Lemma 2. *Assume that the simulated tempering Markov chain M, defined in Definition 2, is reversible and lazy, with stationary distribution P. Let P^0 denote the initial distribution, and suppose that P^0 is absolutely continuous with respect to P. Define f_0 by*

$$P^0(i, \mathrm{d}x) = f_0(i, x)\, P(i, \mathrm{d}x).$$

Let P^N denote the distribution of the chain after N steps, and for each temperature level $i \in [L]$, define the conditional distribution $P_i^N(\mathrm{d}x) \propto P^N(i, \mathrm{d}x)$. Fix $\varepsilon \in (0, 1)$. Suppose that there exist constants $B > 1$, $\infty \geq q > 2$ and a measurable set $\mathcal{X}^0 \subset \mathcal{X}$ such that

1. $\|f_0\|_{\mathcal{L}^q(P)} \leq B$*, where* $\|\cdot\|_{\mathcal{L}^q(P)}$ *denotes the $\mathcal{L}^q$-norm w.r.t. P,*
2. $P([L] \times \mathcal{X}^0) \geq 1 - (\frac{\varepsilon^2}{20B^2})^{q/(q-2)}$,
3. $\mathrm{SpecGap}_{[L]\times\mathcal{X}^0}(M) \geq C_M^{-1}$.

Then, for $N \geq C_M \log(2B^2/\varepsilon^2)$, we have

$$\|P^N - P\|_{\mathrm{tv}} \leq \varepsilon, \text{ and } \|P_i^N - P_i\|_{\mathrm{tv}} \leq \frac{3\varepsilon}{2 \min_{k\in[L]} r_k} \text{ for all } i \in [L]. \tag{6}$$

Proof. See Appendix B. □

The constant C_M controls the rate of convergence of the simulated tempering chain to its stationary distribution in total variation distance; larger values of C_M correspond to slower convergence.

3 Analysis of the Simulated Tempering Metropolis—Hastings Algorithm for Multivariate Gaussian Mixtures

Let function $f\colon \mathbb{R}^d \to \mathbb{R}$ be defined by

$$f(x) = -\log\left\{\sum_{i=1}^{n} w_i e^{-\frac{1}{2}(x-\mu_i)^\top \Sigma^{-1}(x-\mu_i)}\right\}, \tag{7}$$

where $\mu_i \in \mathbb{R}^d$ for each i, $\Sigma \in \mathbb{R}^{d\times d}$ is a positive definite matrix, and $w_i > 0$ for each i such that $\sum_{i=1}^n w_i = 1$. We want to sample from a probability distribution P^* with probability density function $p^*(x) \propto e^{-f(x)}$. The target density p^* corresponds to a mixture of n Gaussian components,

where each component has mean μ_i, weight w_i, and shared covariance matrix Σ. We assume that we do not have access to the individual means $(\mu_i)_{i=1}^n$ and weights $(w_i)_{i=1}^n$, but we can evaluate $f(x)$ at any point $x \in \mathbb{R}^d$; the problem would become trivial if $(\mu_i, w_i)_{i=1}^n$ are known.

We illustrate the use of our decomposition theorem by analyzing the sampling complexity of the simulated tempering Metropolis–Hastings (STMH) algorithm for target distribution P^*.

Definition 4 (STMH)**.** *Given a target probability density $p^*(x) \propto \exp(-f(x))$, let*

$$M^* = M\left((p_i^*)_{i=1}^L, (r_i)_{i=1}^L, (M_i^*)_{i=1}^L, \lambda\right),$$

denote the simulated tempering Markov chain defined in Definition 2. Here,

$$p_i^*(x) \propto \exp(-\beta_i f(x)), \tag{8}$$

where $0 < \beta_1 < \cdots < \beta_L = 1$ is a sequence of inverse temperatures. The transition kernel M_i^* is that of the Metropolis–Hastings algorithm targeting p_i^* with a symmetric Gaussian proposal density $q(x, y) = \mathcal{N}(y; x, \eta I)$, where $\eta > 0$ is the step size. The constants r_i are set proportional to $Z_i/\widehat{Z}_i$, where Z_i is the (unknown) normalizing constant of p_i^*, and $\widehat{Z}_i$ is its estimate. Specifically, we set

$$r_i = \frac{Z_i/\widehat{Z}_i}{\sum_{k=1}^L Z_k/\widehat{Z}_k}$$

so that $\sum_{i=1}^L r_i = 1$.

We define $r_i \propto Z_i/\widehat{Z}_i$ because the true normalizing constants Z_i (also known as partition functions, where Z is viewed as a function of the temperature index i) are typically unknown in practice. When implementing the STMH algorithm, the acceptance probability in Equation (1) is given by

$$a\big((i,x),(i',x)\big) = \min\left\{\frac{\widehat{Z}_i \exp(-\beta_{i'} f(x))}{\widehat{Z}_{i'} \exp(-\beta_i f(x))}, 1\right\}.$$

This choice ensures that the acceptance probability depends only on the estimated normalizing constants $\widehat{Z}_i$ and not the true values Z_i, thereby making the algorithm implementable even when Z_i are unknown. Since acceptance probability depends only on the ratio $\widehat{Z}_i/\widehat{Z}_{i'}$, it is sufficient to estimate the normalizing constants up to a common multiplicative factor. We set $\widehat{Z}_1 = 1$ and estimate the other normalizing constants using the inductive strategy considered by Ge et al. [2018]: for $\ell = 1, \ldots, L-1$, we run STMH using ℓ inverse temperatures $\beta_1 < \cdots < \beta_\ell$ and use the samples at temperature level ℓ to compute the estimate $\widehat{Z}_{\ell+1}$. This estimation procedure is summarized in Algorithm 2 in Appendix A. After obtaining all estimates $(\widehat{Z}_i)_{i\in[L]}$, we run STMH using all L temperature levels, and the samples collected at the L-th temperature level can be used to approximate the original target density $p^*(x) \propto e^{-f(x)}$. Algorithm 1 in Appendix A summarizes the STMH algorithm given in Definition 4, assuming all partition functions are known.

The following theorem provides a mixing time bound for the STMH algorithm targeting the distribution P^*. Directly applying the spectral gap decomposition in Corollary 1 is very challenging, since it remains unclear how to effectively bound the spectral gap of the Metropolis–Hastings chain over an unbounded domain. In contrast, such bounds are more tractable within bounded regions [Yuen, 2000]. This motivates the use of the decomposition of restricted spectral gap given in Theorem 1.

Theorem 2. *Let $f(x)$ be defined as in Equation* (7)*, and define*

$$D := \max\left\{\max_{k\in[n]} \|\mu_k\|, \sqrt{\gamma_{\min}}\right\}, \quad w_{\min} := \min_{1\le j\le n} w_j, \qquad \kappa := \frac{\gamma_{max}}{\gamma_{min}},$$

where γ_{max} and γ_{min} denote the largest and smallest eigenvalues of the covariance matrix Σ, respectively. Then, assuming d is fixed, STMH algorithm can be used to generate a sample from a distribution that is within ε total variation distance of the target distribution P^*, in time

$$\mathrm{poly}\left(\frac{1}{w_{\min}}, D, \kappa\right) \cdot \log^3\left(\frac{1}{\varepsilon}\right).$$

Proof. See Appendix C, where a precise version of the time complexity bound is also provided. □

Remark 4. The parameter D captures the spread and separation of the Gaussian mixture components since $\max_{i \neq j} \|\mu_i - \mu_j\| \leq 2D$ by the triangle inequality.

Remark 5. The time complexity in Theorem 2 is polynomial in the separation parameter D, logarithmic in the inverse accuracy $1/\varepsilon$, and exponential in the dimension d. In fixed-dimensional settings—where d is constant—this result provides high-accuracy guarantees, showing that STMH is particularly effective for sampling from such mixture distributions. For comparison, the proximal sampler incurs a time complexity with exponential dependence on D, since the log-Sobolev constant and Poincaré constant for a Gaussian mixture distribution decays exponentially in D; see Schlichting [2019]. To further illustrate the scaling behavior of the proximal sampler, consider a mixture of two Gaussians with identical covariance matrices and means located at $(-m/2, \ldots, -m/2)$ and $(m/2, \ldots, m/2)$, respectively. In this case, the separation becomes $D = m\sqrt{d}$, which implies that the complexity of the proximal sampler scales exponentially in both m and d. The simulated tempering Langevin Monte Carlo (STLMC) algorithm of Ge et al. [2018] admits an upper bound that scales polynomially with D, d, and $1/\varepsilon$. In contrast, the upper bound we obtain for STMH exhibits only a logarithmic dependence on $1/\varepsilon$, representing a state-of-the-art theoretical guarantee among existing known upper bounds. Table 1 summarizes the theoretical complexity of STMH in comparison with several other sampling algorithms. It can be seen that no algorithm dominates STMH in terms of the complexity dependence on D and $1/\varepsilon$. Our upper bound on the time complexity of STMH has exponential dependence on d, which is likely due to the use of the random walk Metropolis–Hastings (RWMH) sampler for each local chain M_i. We conjecture that by replacing it with proximal sampler [Chen et al., 2022, He et al., 2024a] or Metropolis-adjusted Langevin algorithm [Wu et al., 2022], the resulting simulated tempering algorithm may achieve a better complexity dependence on d.

Remark 6. Theorem 2 establishes convergence guarantees for STMH when sampling from a mixture of Gaussians with a shared covariance matrix. This result can be naturally extended to target distributions that are sufficiently close to such mixtures, following an approach similar to that of Ge et al. [2018]. In such cases, the time complexity would additionally depend on closeness between the actual distribution and the Gaussian mixture approximation. We also anticipate that similar techniques can be adapted to handle mixtures of log-concave distributions that are translates to each other, or more broadly, distributions that are well-approximated by such mixtures. However, as demonstrated by Ge et al. [2018], some seemingly mild violations of the assumptions, such as component covariance matrices differing by a constant factor, can lead to exponential time complexity for any reasonable algorithm with similar guarantees.

Algorithm	d	D	$1/\varepsilon$
STLMC [Ge et al., 2018]	poly	poly	poly
LMC [Vempala and Wibisono, 2019]	d	$\exp(D^2)^2$	$1/\varepsilon^2$
Annealed LMC [Guo et al., 2024]	poly	poly	poly
Proximal Sampler [Fan et al., 2023]	$d^{1/2}$	$\exp(D^2)$	$\log(1/\varepsilon)$
LMC (data-initialized) [Koehler et al., 2024]	poly	$\exp(D^2)^2$	$1/\varepsilon^2$
STMH	exp	poly	$\log^2(1/\varepsilon)$

Table 1: Dependence of time complexity on d, D and ε for sampling from densities of the form in (7).

The proof of Theorem 2 is divided into several steps. In order to apply the decomposition arguments developed in Section 2, we first define an approximate STMH chain that satisfies Assumption 1.

Definition 5 (Approximate STMH)**.** *Let*

$$\widetilde{M} = M\left((\widetilde{p}_i)_{i=1}^L, (r_i)_{i=1}^L, (\widetilde{M}_i)_{i=1}^L, \lambda\right),$$

denote the simulated tempering Markov chain defined in Definition 2, where

$$\widetilde{p}_i(x) \propto \sum_{j=1}^n w_j \exp\left\{-\frac{\beta_i}{2}(x-\mu_j)^\top \Sigma^{-1}(x-\mu_j)\right\}, \tag{9}$$

and transition kernel $\widetilde{M}_i$ is that of the Metropolis–Hastings algorithm targeting $\widetilde{p}_i$, with a symmetric Gaussian proposal density $q(x,y) = \mathcal{N}(y; x, \eta I)$ where step size $\eta > 0$. The weights $(r_i)_{i=1}^L$ and λ are the same as in Definition 4.

The stationary densities $\widetilde{p}_i$ in Definition 5 are mixtures of Gaussian densities; denote the component distributions by $\widetilde{P}_{(i,j)}$. This enables us to apply Theorem 1. It is important to note that this approximate STMH sampler cannot be implemented in practice, as it requires knowledge of the component distributions, which is typically unavailable. We show that, for some $\mathcal{X}^0 \subset \mathbb{R}^d$, the $\mathcal{X}^0$-restricted spectral gap of the STMH chain constructed in Definition 4 is comparable to that of the approximate STMH chain given in Definition 5 (see Lemma 9 in Appendix C.1). This comparison result is a discrete-time extension of the argument used in Ge et al. [2018], which suggests that adjusting the temperature is roughly equivalent to modifying the variance of a Gaussian distribution. As a result, it suffices to obtain a lower bound on the $\mathcal{X}^0$-restricted spectral gap for the approximate STMH chain. Note that since $\beta_L = 1$, $p_L = \widetilde{p}_L = p^*$.

Next, we apply Theorem 1 to derive a lower bound on the $\mathcal{X}^0$-restricted spectral gap of the approximate STMH chain. This involves three steps. First, we verify inequality (2) in Appendix C.2.1 by comparing the transition density of the chain $\widetilde{M}_{(i,j)}$ with that of $\widetilde{M}_i$. Second, we need to compute the $\mathcal{X}^0$-restricted spectral gap of the "local chain" $\widetilde{M}_{(i,j)}$, which we define as the random walk Metropolis–Hastings algorithm targeting $\widetilde{P}_{(i,j)}$, the Gaussian distribution with mean μ_j and covariance matrix $\beta_i^{-1}\Sigma$. It is well known that for strongly log-concave target distributions, the Metropolis–Hastings algorithm should mix fast [Johndrow and Smith, 2018]. To explicitly compute the bound, we apply the path method on continuous spaces proposed by [Yuen, 2000], which is detailed in Appendix C.2.2. Finally, the projected chain captures transitions between mixture components and temperature levels. Intuitively, it should mix fast because (i) if β_1 is sufficiently small, the component distributions $\widetilde{P}_{(1,1)}, \ldots, \widetilde{P}_{(1,n)}$ overlap significantly, and (ii) if β_i/β_{i-1} is not too big, then $\widetilde{P}_{(i,j)}$ and $\widetilde{P}_{(i-1,j)}$ also share substantial overlap. To compute a lower bound on the spectral gap for the projected chain, we apply the well-known canonical path method [Levin and Peres, 2017]; see Appendix C.2.4.

To conclude the proof, we derive error bounds on the estimated partition functions in Appendix C.4.3.

4 Simulation Study

To numerically investigate the complexity of the STMH algorithm, we perform a simulation study with target distribution being a symmetric two-dimensional Gaussian mixture distribution, whose density is given by

$$p^*(x) = \frac{1}{2}\mathcal{N}\left(x; -\frac{D}{2\sqrt{2}} \cdot 1_2, I_2\right) + \frac{1}{2}\mathcal{N}\left(x; \frac{D}{2\sqrt{2}} \cdot 1_2, I_2\right),$$

where I_2 is the 2×2 identity matrix and $1_2 = (1,1)^\top$. The parameter D controls the separation between the two components. We vary the parameter D to explore how increasing the separation between the modes affects the convergence behavior of the algorithm. For each value of D, we run the STMH algorithm with the parameters specified in Appendix C.4.1 and initialized at $(10, 10)$. To assess convergence, we monitor how quickly the empirical mean of the samples, denoted by $\hat{\mu}$, approaches the true mean of the target distribution, $(0, 0)$. If the chain has not yet mixed, it tends to spend more time near one mode, resulting in an empirical mean that deviates from the target mean. Moreover, according to Nishiyama [2022], when the empirical and target distribution have the same covariance matrix, we can lower bound their TV distance by $\|\hat{\mu}\|^2/(C + \|\hat{\mu}\|^2)$ for some constant $C > 0$, which is a monotone increasing function of $\|\hat{\mu}\|^2$. This justifies the use of $\|\hat{\mu}\|$ as a measure of convergence. To reduce variability, we repeat the simulation $20,000$ times for each fixed D and average the empirical means over all runs. To provide a benchmark, we also run the baseline Metropolis–Hastings (MH) algorithm under the same setup and compare its convergence behavior with that of STMH. All simulations were performed on a standard consumer-grade CPU with parallelization and completed within approximately six hours.

In the left panel of Figure 1, we plot the number of steps required for the empirical mean to fall below 0.1 for both STMH and the baseline MH algorithm. The error bars are obtained by computing the 95% coverage interval of the empirical mean at each step, and then determining the corresponding number of steps needed for the lower and upper bounds of the interval to cross the 0.1 threshold. The results show that this number grows approximately linearly with D^2 for STMH, consistent with the theoretical bound in Theorem 2. In contrast, for the baseline MH algorithm the number of

steps increases exponentially with D^2, reflecting its slower convergence, which is consistent with the well-known behavior of MH. To analyze how the mixing time depends on the threshold ε, we fix $D = 30$ and plot how $\|\hat{\mu}\|$ varies with the number of steps N in the right panel of Figure 1. For STMH, the results show an approximately linear relationship between the number of steps and $\log^2\big(1/\|\hat{\mu}\|\big)$, consistent with the theoretical bound in Theorem 2. In contrast, for the baseline MH algorithm the relationship is closer to logarithmic, implying a polynomial dependence of the mixing time on $1/\varepsilon$, which aligns with the well-known limitations of MH in achieving high accuracy. The error bars represent 95% coverage intervals of the empirical mean, with the delta method applied to obtain the corresponding intervals for $\log^2\big(1/\|\hat{\mu}\|\big)$. Overall, these simulations demonstrate that STMH provides more efficient sampling from this Gaussian mixture distribution compared to MH.

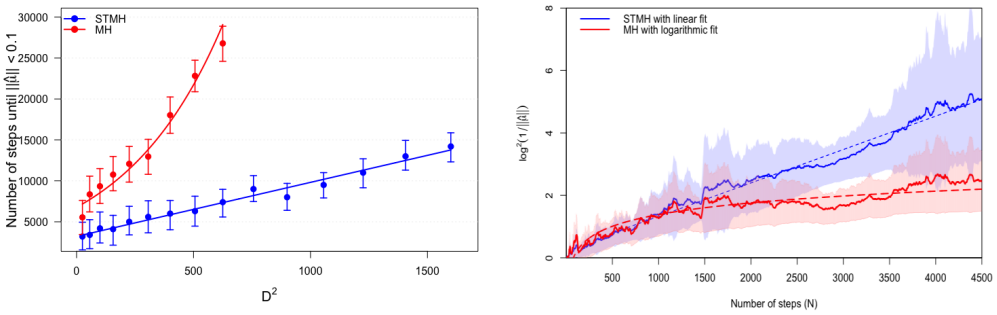

Figure 1: Left: number of steps until $\|\widehat{\mu}\| < 0.1$ versus D^2. Right: $\log^2\left(1/\|\widehat{\mu}\|\right)$ versus the number of steps N for $D = 30$.

5 Concluding Remarks

Simulated tempering addresses the challenge of sampling from multimodal distributions. In this work, we develop a general theoretical framework for analyzing simulated tempering and demonstrate its effectiveness through a detailed analysis of simulated tempering combined with the Metropolis–Hastings algorithm for sampling from Gaussian mixtures. Our framework can be used to analyze simulated tempering combined with other local MCMC samplers, such as the Metropolis-adjusted Langevin algorithm (MALA) and proximal algorithms, but verifying Assumptions 2 and 3 and computing the constants C_1, C_2, C_3 in these settings may be more involved. In particular, since it has been shown in the literature that the dimensional dependence of the complexity of MALA for log-concave target distributions is $\tilde{\Theta}(\sqrt{d})$ [Chewi et al., 2021, Wu et al., 2022], it would be interesting to investigate if the complexity of simulated tempering combined with MALA has a similar polynomial dependence on d, improving on the exponential dependence in our result. Another promising direction for future work is to generalize our techniques to a broader class of target distributions beyond Gaussian mixtures. The argument of Ge et al. [2018] can be used to extend our result to target distributions that are "sufficiently close" to Gaussian mixtures. Finally, while this work primarily focuses on establishing lower bounds on the spectral gap, an interesting direction for future work is to investigate the tightness of these bounds by also deriving upper bounds on the spectral gap for simulated tempering Markov chains.

6 Acknowledgments

Jhanvi Garg and Quan Zhou were supported in part by NSF grants DMS-2245591 and DMS-2311307. Krishnakumar Balasubramanian was supported in part by NSF grant DMS-2413426.

References

Yves F Atchadé. Approximate spectral gaps for Markov chain mixing times in high dimensions. *SIAM Journal on Mathematics of Data Science*, 3(3):854–872, 2021.

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

Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

Hyunwoong Chang and Quan Zhou. Dimension-free relaxation times of informed MCMC samplers on discrete spaces. *arXiv preprint arXiv:2404.03867*, 2024.

Omar Chehab, Anna Korba, Austin Stromme, and Adrien Vacher. Provable convergence and limitations of geometric tempering for Langevin dynamics. *arXiv preprint arXiv:2410.09697*, 2024.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

Sinho Chewi. Log-concave sampling, 2023. *Book draft available at https://chewisinho. github. io*, 2024.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021.

Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

John S Dagpunar. *Simulation and Monte Carlo: With applications in finance and MCMC*. John Wiley & Sons, 2007.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, pages 1551–1587, 2017.

Murat A Erdogdu and Rasa Hosseinzadeh. On the convergence of Langevin Monte Carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR, 2021.

Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1473–1521. PMLR, 2023.

Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering Langevin Monte Carlo ii: An improved proof using soft Markov chain decomposition. *arXiv preprint arXiv:1812.00793*, 2018.

A Gelman et al. Bayesian data analysis, 3rd: Boca Raton. *Texts in Statistical Science*, 2013.

Wei Guo, Molei Tao, and Yongxin Chen. Provable benefit of annealed Langevin Monte Carlo for non-log-concave sampling. *arXiv preprint arXiv:2407.16936*, 2024.

Ye He, Alireza Mousavi-Hosseini, Krishna Balasubramanian, and Murat A Erdogdu. A separation in heavy-tailed sampling: Gaussian vs. stable oracles for proximal samplers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL `https://openreview.net/forum?id=zuwLGhgxtQ`.

Ye He, Kevin Rojas, and Molei Tao. Zeroth-order sampling methods for non-log-concave distributions: Alleviating metastability by denoising diffusion. *arXiv preprint arXiv:2402.17886*, 2024b.

Xunpeng Huang, Hanze Dong, HAO Yifan, Yian Ma, and Tong Zhang. Reverse diffusion Monte Carlo. In *The Twelfth International Conference on Learning Representations*, 2023.

Xunpeng Huang, Difan Zou, Hanze Dong, Yi-An Ma, and Tong Zhang. Faster sampling without isoperimetry via diffusion-based Monte Carlo. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2438–2493. PMLR, 2024.

James E Johndrow and Aaron Smith. Fast mixing of Metropolis-Hastings with unimodal targets. *arXiv preprint arXiv:1806.07047*, 2018.

Frederic Koehler, Holden Lee, and Andrej Risteski. Sampling approximately low-rank Ising models: MCMC meets variational methods. In *Conference on Learning Theory*, pages 4945–4988. PMLR, 2022.

Frederic Koehler, Holden Lee, and Thuy-Duong Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. *arXiv preprint arXiv:2411.09117*, 2024.

David Landau and Kurt Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2021.

Krzysztof Łatuszyński, Matthew T Moores, and Timothée Stumpf-Fétizon. MCMC for multi-modal distributions. *arXiv preprint arXiv:2501.05908*, 2025.

Holden Lee and Matheau Santana-Gijzen. Convergence bounds for sequential Monte Carlo on multimodal distributions using soft decomposition. *arXiv preprint arXiv:2405.19553*, 2024.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Jun S Liu and Rong Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044, 1998.

Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Neal Madras and Dana Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, pages 581–606, 2002.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics letters*, 19(6):451, 1992.

Ravi Montenegro, Prasad Tetali, et al. Mathematical aspects of mixing times in Markov chains. *Foundations and Trends® in Theoretical Computer Science*, 1(3):237–354, 2006.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *Bernoulli*, 28(3): 1577–1601, 2022.

Alireza Mousavi-Hosseini, Tyler K Farghly, Ye He, Krishna Balasubramanian, and Murat A Erdogdu. Towards a complete analysis of Langevin Monte Carlo: Beyond Poincaré inequality. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1–35. PMLR, 2023.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

Mark EJ Newman and Gerard T Barkema. *Monte Carlo methods in statistical physics*. Clarendon Press, 1999.

Tomohiro Nishiyama. Lower bounds for the total variation distance given means and variances of distributions. *arXiv preprint arXiv:2212.05820*, 2022.

Art B Owen. Monte Carlo theory, methods and examples, 2013.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

André Schlichting. Poincaré and log–Sobolev inequalities for mixtures. *Entropy*, 21(1):89, 2019.

Robert H Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

Saifuddin Syed, Alexandre Bouchard-Côté, Kevin Chern, and Arnaud Doucet. Optimised annealed sequential Monte Carlo samplers. *arXiv preprint arXiv:2408.12057*, 2024.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Dawn B Woodard, Scott C Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2): 617–640, 2009.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270): 1–63, 2022.

Wai Kong Yuen. Applications of geometric bounds to the convergence rate of Markov chains on R^n. *Stochastic processes and their applications*, 87(1):1–23, 2000.

Zhongrong Zheng. On swapping and simulated tempering algorithms. *Stochastic Processes and their Applications*, 104(1):131–154, 2003.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction clearly state the main contributions and are consistent with the theoretical and empirical results presented in the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work in the conclusion section, highlighting potential extensions and directions for future research.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are accompanied by clearly stated assumptions and complete, rigorous proofs. Proofs provided in the appendix are referenced appropriately in the main text.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides all necessary details to reproduce the experiments, including clear algorithm descriptions and parameter settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is included as a ZIP file (code.zip) to reproduce the experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper does not involve standard training/test splits. However, all algorithmic parameters are clearly specified.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars that are appropriately defined and clearly presented.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiment was run on a standard consumer-grade CPU with parallelization and completed within approximately six hours.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research fully adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper presents foundational theoretical results on sampling from complex, multimodal distributions. While the work is not tied to a specific application or deployment, it has implications for domains such as Bayesian inference, generative modeling, and scientific computing. Given its theoretical nature, we do not foresee any direct societal impacts at this stage.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release any models or datasets that pose a risk of misuse or require safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use any external assets such as third-party code, datasets, or pretrained models.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not introduce or release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The research does not use LLMs as part of its core methods.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendices

A Simulated Tempering Metropolis–Hastings algorithm

Algorithm 1 Simulated Tempering Metropolis–Hastings Algorithm

Input: function f, inverse temperatures $\beta_1, \ldots, \beta_\ell$, partition function estimates $\widehat{Z}_1, \ldots, \widehat{Z}_\ell$, number of steps N, step size η, rate λ, initial covariance matrix Σ_0.
Sample $x_0 \sim \mathcal{N}(0, \Sigma_0)$
$i \leftarrow 1, x \leftarrow x_0, n \leftarrow 0$
while $n < N$ **do**
    Sample $u \sim \text{Bernoulli}(\lambda)$.
    if $u = 0$ **then**
        Propose $x' \sim \mathcal{N}(x, \eta I)$
        Sample $v \sim \text{Uniform}(0, 1)$
        if $v < \min\left\{1, \frac{e^{-\beta_i f(x')}}{e^{-\beta_i f(x)}}\right\}$ **then**
            $x \leftarrow x'$
        end if
    else
        Propose $i' = i \pm 1$, each with probability $1/2$
        if $1 \le i' \le \ell$ **then**
            Sample $v \sim \text{Uniform}(0, 1)$
            if $v < \min\left\{1, \frac{e^{-\beta_{i'} f(x)}/\widehat{Z}_{i'}}{e^{-\beta_i f(x)}/\widehat{Z}_i}\right\}$ **then**
                $i \leftarrow i'$
            end if
        end if
    end if
    $n \leftarrow n + 1$
end while
Output: Sample (x, i) collected at the N^{th} step.

Algorithm 2 Partition Function Estimation

Input: function f, inverse temperature sequence $\beta_1 < \cdots < \beta_L$ and number of samples s.
$\widehat{Z}_1 \leftarrow 1$
for $\ell = 1$ **to** L **do**
    Repeat Algorithm 1 until s samples $(x_j)_{j=1}^s$ are obtained at temperature level ℓ.
    $\widehat{Z_{\ell+1}} \leftarrow (\widehat{Z_\ell}/s) \sum_{j=1}^s e^{(-\beta_{\ell+1}+\beta_\ell) f(x_j)}$
end for

Remark 7. Algorithm 1 is always run for a fixed number of steps N and returns the sample obtained at the final step. In Algorithm 2, if this sample is not from the desired temperature level, Algorithm 1 is simply re-run for another N steps.

B Proofs for Section 2

B.1 Proof of Lemma 1

Proof. Clearly, $\sum_{i,j} \overline{P}(i,j) = 1$. Hence, it only remains to check the detailed balance condition

$$\overline{P}((i,j))\,\overline{M}((i,j),(i',j')) = \overline{P}((i',j'))\,\overline{M}((i',j'),(i,j))$$

for two types of moves. First, let i be fixed and consider $j \neq j'$. Then,

$$r_i w_{(i,j)} P_{(i,j)}(\mathcal{X}^0)\overline{M}((i,j),(i,j')) = r_i(1-\lambda)\int_{\mathcal{X}^0} w_{(i,j)} p_{(i,j)}(x)\frac{w_{(i,j')}p_{(i,j')}(x)}{p_i(x)}\mathrm{d}x,$$

which is clearly symmetric with respect to j and j'. Similarly, if j is fixed and $i' = i \pm 1$ (assuming both $i, i' \in [L]$), we have

$$\begin{aligned} r_i w_{(i,j)} P_{(i,j)}(\mathcal{X}^0)\overline{M}((i,j),(i',j)) &= \frac{\lambda}{2}\int_{\mathcal{X}^0} r_i w_{(i,j)} p_{(i,j)}(x) a((i,j,x),(i',j,x))\mathrm{d}x \\ &= \frac{\lambda}{2}\int_{\mathcal{X}^0} \min\left\{r_i w_{(i,j)} p_{(i,j)}(x), r_{i'} w_{(i',j)} p_{(i',j)}(x)\right\}\mathrm{d}x, \end{aligned}$$

which is symmetric with respect to i and i'. □

B.2 Proof of Theorem 1

We first prove an auxiliary lemma on the Dirichlet form of the simulated tempering Markov chain.

Lemma 3. *The $[L]\times\mathcal{X}^0$-restricted Dirichlet form $\mathcal{E}_{[L]\times\mathcal{X}^0}$ of the simulated tempering Markov chain M, defined in Definition 2, can be expressed by*

$$\mathcal{E}_{[L]\times\mathcal{X}^0}(g,g) = (1-\lambda)\sum_{i=1}^{L} r_i\,\mathcal{E}_{i,\mathcal{X}^0}(g_i,g_i) + \lambda\,\mathcal{E}^I_{\mathcal{X}^0}(g,g),$$

where $\mathcal{E}_{i,\mathcal{X}^0}$ is the $\mathcal{X}^0$-restricted Dirichlet form of the Markov chain M_i, $g \in \mathcal{L}^2([L]\times\mathcal{X}, P)$ with $g_i(x) = g(i,x)$ for each $i \in [L]$ and

$$\mathcal{E}^I_{\mathcal{X}^0}(g,g) = \frac{1}{4}\sum_{i,i'\in[L]:\ i'=i\pm1}\int_{\mathcal{X}^0}\big(g(i,x)-g(i',x)\big)^2 r_i\,p_i(x)\,a\big((i,x),(i',x)\big)\,\mathrm{d}x.$$

Proof. Since the stationary density of M is $p(i,x) = r_i p_i(x)$ and either x or i is fixed in each simulated tempering iteration, the restricted Dirichlet form $\mathcal{E}_{[L]\times\mathcal{X}^0}(g,g)$ can be expressed by

$$\begin{aligned} \mathcal{E}_{[L]\times\mathcal{X}^0}(g,g) = &\frac{1}{2}\sum_{i=1}^{L}\int_{\mathcal{X}^0}\int_{\mathcal{X}^0}\big(g(i,x)-g(i,y)\big)^2 r_i\,p_i(x)\,M\big((i,x),(i,\mathrm{d}y)\big)\,\mathrm{d}x \\ &+\frac{1}{2}\sum_{i,i'\in[L]:\ i'=i\pm1}\int_{\mathcal{X}^0}\big(g(i,x)-g(i',x)\big)^2 r_i\,p_i(x)\,M\big((i,x),(i',x)\big)\,\mathrm{d}x. \end{aligned} \tag{10}$$

Since $M\big((i,x),(i,\mathrm{d}y)\big) = (1-\lambda)M_i(x,\mathrm{d}y)$ and M_i has stationary density p_i,

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{L}\int_{\mathcal{X}^0}\int_{\mathcal{X}^0}\big(g(i,x)-g(i,y)\big)^2 r_i\,p_i(x)\,M\big((i,x),(i,\mathrm{d}y)\big)\,\mathrm{d}x \\ &= \frac{1-\lambda}{2}\sum_{i=1}^{L} r_i\int_{\mathcal{X}^0}\int_{\mathcal{X}^0}\big(g_i(x)-g_i(y)\big)^2 p_i(x)\,M_i\big(x,\mathrm{d}y\big)\mathrm{d}x = (1-\lambda)\sum_{i=1}^{L} r_i\,\mathcal{E}_{i,\mathcal{X}^0}\big(g_i,g_i\big). \end{aligned}$$

For the second term on the right-hand side of (10), note that

$$M\big((i,x),(i',x)\big) = \frac{\lambda}{2}a((i,x),(i',x)),$$

where the acceptance probability function a is given by (1). A straightforward calculation then concludes the proof of the lemma. □

Next, we prove two key lemmas about the Dirichlet form $\overline{\mathcal{E}}$ of the Markov chain $\overline{M}$ constructed in Definition 3. Let $\theta = P([L] \times \mathcal{X}^0)$. Recall that under Assumption 1, we can augment the stationary density to

$$p(i, j, x) = r_i w_{(i,j)} p_{(i,j)}(x).$$

We still denote the corresponding probability measure by P. Let P_0 denote the conditional probability measure given $X \in \mathcal{X}^0$, whose density is given by

$$p_0(i, j, x) = \frac{r_i w_{(i,j)} p_{(i,j)}(x)}{\theta}.$$

The Dirichlet form $\overline{\mathcal{E}}$ can be expressed by

$$\overline{\mathcal{E}}(\overline{g}, \overline{g}) = \overline{\mathcal{E}}^J(\overline{g}, \overline{g}) + \overline{\mathcal{E}}^I(\overline{g}, \overline{g}),$$

where

$$\overline{\mathcal{E}}^J(\overline{g}, \overline{g}) = \frac{1}{2\theta} \sum_{i=1}^{L} \sum_{j,j'=1}^{n} (\overline{g}(i, j) - \overline{g}(i, j'))^2 \, r_i w_{(i,j)} P_{(i,j)}(\mathcal{X}^0) \, \overline{M}((i, j), (i, j')),$$

$$\overline{\mathcal{E}}^I(\overline{g}, \overline{g}) = \frac{1}{2\theta} \sum_{j=1}^{n} \sum_{i,i' \in [L] \colon i' = i \pm 1} (\overline{g}(i, j) - \overline{g}(i', j))^2 \, r_i w_{(i,j)} P_{(i,j)}(\mathcal{X}^0) \, \overline{M}((i, j), (i', j)).$$

Lemma 4. *Suppose Assumption 1 holds. For any $g \in \mathcal{L}^2([L] \times \mathcal{X}, P)$, define $g_i \colon \mathcal{X} \to \mathbb{R}$ by $g_i(x) = g(i, x)$, and define $\overline{g} \colon [L] \times [n] \to \mathbb{R}$ by*

$$\overline{g}(i, j) = \int_{\mathcal{X}^0} g(i, x) \frac{p_{(i,j)}(x)}{P_{(i,j)}(\mathcal{X}^0)} \mathrm{d}x.$$

Then,

$$\overline{\mathcal{E}}^J(\overline{g}, \overline{g}) \leq \frac{2(1 - \lambda)}{\theta} \sum_{i=1}^{L} \sum_{j=1}^{n} \frac{r_i w_{(i,j)}}{P_{(i,j)}(\mathcal{X}^0)} \operatorname{Var}_{P_{(i,j)}, \mathcal{X}^0}(g_i).$$

Proof. For every $x \in \mathcal{X}$, $i \in [L]$ and every pair $j, j' \in [n]$, the following inequality holds

$$\left(\overline{g}(i, j) - \overline{g}(i, j')\right)^2 \leq 2 \left[\left(\overline{g}(i, j) - g(i, x)\right)^2 + \left(\overline{g}(i, j') - g(i, x)\right)^2 \right].$$

Hence, using the expression for $\overline{M}((i, j), (i, j'))$, we get

$$\begin{aligned}
\overline{\mathcal{E}}^J(\overline{g}, \overline{g}) &= \frac{1 - \lambda}{2\theta} \sum_{i=1}^{L} \sum_{j,j'=1}^{n} (\overline{g}(i, j) - \overline{g}(i, j'))^2 r_i w_{(i,j)} \int_{\mathcal{X}^0} p_{(i,j)}(x) p_i(j' | x) \, \mathrm{d}x \\
&\leq (1 - \lambda) \sum_{i=1}^{L} \sum_{j,j'=1}^{n} r_i w_{(i,j)} \int_{\mathcal{X}^0} \left[\left(\overline{g}(i, j) - g(i, x)\right)^2 + \left(\overline{g}(i, j') - g(i, x)\right)^2 \right] \frac{p_{(i,j)}(x)}{\theta} p_i(j' | x) \, \mathrm{d}x \\
&= (1 - \lambda) \mathsf{E}_{\tilde{P}} \left[\left(\overline{g}(I, J) - g(I, X)\right)^2 + \left(\overline{g}(I, J') - g(I, X)\right)^2 \right], \qquad (11)
\end{aligned}$$

where $\tilde{P}$ denotes the joint probability measure of (I, J, J', X) with density

$$\tilde{p}(i, j, j', x) = \frac{r_i w_{(i,j)} p_{(i,j)}(x)}{\theta} p_i(j' | x) = \frac{r_i w_{(i,j)} w_{(i,j')} p_{(i,j)}(x) p_{(i,j')}(x)}{\theta \, p_i(x)},$$

for $i \in [L], j, j' \in [n], x \in \mathcal{X}^0$. Hence, under $\tilde{P}$, the joint distribution of (I, J, X) and that of (I, J', X) are both given by P_0, and thus

$$\mathsf{E}_{\tilde{P}} \left[\left(\overline{g}(I, J) - g(I, X)\right)^2 + \left(\overline{g}(I, J') - g(I, X)\right)^2 \right] = 2 \mathsf{E}_{P_0} \left[\left(\overline{g}(I, J) - g(I, X)\right)^2 \right]. \qquad (12)$$

Since $\overline{g}(i,j) = \mathsf{E}_{P_0}[g(I,X) \,|\, I=i, J=j]$, we find that

$$\begin{aligned}
&\mathsf{E}_{P_0}\left[\big(\overline{g}(I,J) - g(I,X)\big)^2\right] = \mathsf{E}_{P_0}\left[\mathrm{Var}_{P_0}(g(I,X) \,|\, I,J)\right] \\
&= \sum_{i=1}^{L}\sum_{j=1}^{n} \frac{r_i w_{(i,j)} P_{(i,j)}(\mathcal{X}^0)}{\theta}\, \mathrm{Var}_{P_0}(g(I,X) \,|\, I=i, J=j) \\
&= \sum_{i=1}^{L}\sum_{j=1}^{n} \frac{r_i w_{(i,j)}}{\theta\, P_{(i,j)}(\mathcal{X}^0)}\, \mathrm{Var}_{P_{(i,j)},\mathcal{X}^0}(g_i), \qquad (13)
\end{aligned}$$

where in the last step we have used

$$\mathrm{Var}_{P_0}(g(I,X) \,|\, I=i, J=j) = \frac{1}{2}\int_{\mathcal{X}^0\times\mathcal{X}^0} [g(i,x)-g(i,y)]^2 \frac{p_{(i,j)}(x)}{P_{(i,j)}(\mathcal{X}^0)} \frac{p_{(i,j)}(y)}{P_{(i,j)}(\mathcal{X}^0)} \mathrm{d}x\mathrm{d}y.$$

The claim then follows from (11), (12) and (13). □

Lemma 5. *Consider the setting of Lemma 4. We also have*

$$\overline{\mathcal{E}}^I(\overline{g},\overline{g}) \le \frac{3\lambda}{\theta}\sum_{i=1}^{L}\sum_{j=1}^{n} \frac{r_i w_{(i,j)}}{P_{(i,j)}(\mathcal{X}^0)}\, \mathrm{Var}_{P_{(i,j)},\mathcal{X}^0}(g_i) + \frac{3\lambda}{\theta}\mathcal{E}^I_{\mathcal{X}^0}(g,g).$$

Proof. For every $x \in \mathcal{X}$, every pair $i, i' \in [L]$, and each $j \in [n]$,

$$\big(\overline{g}(i,j) - \overline{g}(i',j)\big)^2 \le 3\Big[\big(\overline{g}(i,j) - g(i,x)\big)^2 + \big(g(i,x) - g(i',x)\big)^2 + \big(\overline{g}(i',j) - g(i',x)\big)^2\Big].$$

Then, using the definition of $\overline{M}((i,j),(i',j))$ for $i' = i \pm 1$, we get

$$\begin{aligned}
\overline{\mathcal{E}}^I(\overline{g},\overline{g}) &= \frac{\lambda}{4\theta}\sum_{i,i'\in[L]:\, i'=i\pm1}\sum_{j=1}^{n} (\overline{g}(i,j)-\overline{g}(i',j))^2 r_i w_{(i,j)} \int_{\mathcal{X}^0} p_{(i,j)}(x)\, a((i,j,x),(i',j,x))\mathrm{d}x \\
&\le \frac{3\lambda}{2}\mathsf{E}_{\tilde{P}}\Big[\big(\overline{g}(I,J) - g(I,X)\big)^2 + \big(g(I,X) - g(I',X)\big)^2 + \big(\overline{g}(I',J) - g(I',X)\big)^2\Big],
\end{aligned}$$

where $\tilde{P}$ is the probability measure of (I, I', J, X) with density

$$\tilde{p}(i,i',j,x) = \begin{cases} \dfrac{r_i w_{(i,j)} p_{(i,j)}(x)\, a\big((i,j,x),(i',j,x)\big)}{2\theta}, & \text{if } i' = i \pm 1, \\ 1 - \tilde{p}(i,i+1,j,x) - \tilde{p}(i,i-1,j,x), & \text{if } i' = i, \\ 0, & \text{otherwise.} \end{cases}$$

That is, we first draw $I, J, X \sim P_0$ and then update I' by proposing $I' = I \pm 1$ with equal probability and accept it with probability $a((i,j,x),(i',j,x))$. Since the update for I' given I, J, X is reversible with respect to P_0, we also have $(I', J, X) \sim P_0$. Hence,

$$\mathsf{E}_{\tilde{P}}\Big[\big(\overline{g}(I,J) - g(I,X)\big)^2\Big] = \mathsf{E}_{\tilde{P}}\Big[\big(\overline{g}(I',J) - g(I',X)\big)^2\Big] = \mathsf{E}_{P_0}\Big[\big(\overline{g}(I,J) - g(I,X)\big)^2\Big]$$

which has been characterized in (13). Finally,

$$\begin{aligned}
&\mathsf{E}_{\tilde{P}}\Big[\big(g(I,X) - g(I',X)\big)^2\Big] \\
&= \frac{1}{2\theta}\sum_{i,i'\in[L]:\, i'=i\pm1}\sum_{j=1}^{n}\int_{\mathcal{X}^0} (g(i,x)-g(i',x))^2 r_i w_{(i,j)} p_{(i,j)}(x)\, a((i,j,x),(i',j,x))\mathrm{d}x \\
&= \frac{1}{2\theta}\sum_{i,i'\in[L]:\, i'=i\pm1}\sum_{j=1}^{n}\int_{\mathcal{X}^0} (g(i,x)-g(i',x))^2 \min\left\{r_i w_{(i,j)} p_{(i,j)}(x),\, r_{i'} w_{(i',j)} p_{(i',j)}(x)\right\}\mathrm{d}x \\
&\le \frac{1}{2\theta}\sum_{i,i'\in[L]:\, i'=i\pm1}\int_{\mathcal{X}^0} (g(i,x)-g(i',x))^2 \min\left\{r_i p_i(x),\, r_{i'} p_{i'}(x)\right\}\mathrm{d}x \\
&= \frac{2}{\theta}\mathcal{E}^I_{\mathcal{X}^0}(g,g),
\end{aligned}$$

where $\mathcal{E}^I_{\mathcal{X}^0}(g,g)$ is defined in Lemma 3. Note that in the inequality above, we have used that $\sum_j \min\{a_j, b_j\} \le \min\{\sum_j a_j, \sum_j b_j\}$ for two non-negative sequences a_j, b_j. □

Proof of Theorem 1. Fix an arbitrary $g \in \mathcal{L}^2([L]\times\mathcal{X}, P)$. Define, for each i, $g_i\colon \mathcal{X} \to \mathbb{R}$ by $g_i(x) = g(i,x)$, and $\overline{g}\colon [L]\times[n] \to \mathbb{R}$ by

$$\overline{g}(i,j) = \int_{\mathcal{X}^0} g(i,x)\frac{p_{(i,j)}(x)}{P_{(i,j)(\mathcal{X}^0)}}\mathrm{d}x.$$

Note that $\overline{g}(i,j)$ is the conditional expectation of $g(I,X)$ given $I=i$ and $J=j$ under the joint probability measure P_0, and $\overline{P}(i,j)$ is the marginal probability of $I=i, J=j$ under P_0. Hence, by the law of total variance, Assumption 3 and Equation 13, we find that

$$\begin{aligned}\mathrm{Var}_{P_0}(g) &= \mathrm{Var}_{\overline{P}}(\overline{g}) + \sum_{i=1}^{L}\sum_{j=1}^{n}\overline{P}(i,j)\,\mathrm{Var}_{P_0}(g(I,X)\,|\,I=i,J=j)\\ &\le C_3\overline{\mathcal{E}}(\overline{g},\overline{g}) + \sum_{i=1}^{L}\sum_{j=1}^{n}\frac{r_i w_{(i,j)}}{\theta P_{(i,j)}(\mathcal{X}^0)}\,\mathrm{Var}_{P_{(i,j)},\mathcal{X}^0}(g_i).\end{aligned}$$

Using $P_{(i,j)}(\mathcal{X}^0) \ge \phi$ and Assumption 2,

$$\sum_{i=1}^{L}\sum_{j=1}^{n}\frac{r_i w_{(i,j)}}{\theta P_{(i,j)}(\mathcal{X}^0)}\,\mathrm{Var}_{P_{(i,j)},\mathcal{X}^0}(g_i) \le \frac{C_2}{\theta\phi}\sum_{i=1}^{L} r_i \sum_{j=1}^{n} w_{(i,j)}\mathcal{E}_{(i,j),\mathcal{X}^0}(g_i,g_i) \le \frac{C_1C_2}{\theta\phi}\sum_{i=1}^{L} r_i\mathcal{E}_{i,\mathcal{X}^0}(g_i,g_i).$$

Recall that $\overline{\mathcal{E}}(\overline{g},\overline{g}) = \overline{\mathcal{E}}^J(\overline{g},\overline{g}) + \overline{\mathcal{E}}^I(\overline{g},\overline{g})$. By Lemma 4,

$$\overline{\mathcal{E}}^J(\overline{g},\overline{g}) \le \frac{2(1-\lambda)}{\theta}\sum_{i=1}^{L}\sum_{j=1}^{n}\frac{r_i w_{(i,j)}}{P_{(i,j)}(\mathcal{X}^0)}\,\mathrm{Var}_{P_{(i,j)},\mathcal{X}^0}(g_i) \le \frac{2(1-\lambda)C_1C_2}{\theta\phi}\sum_{i=1}^{L} r_i\mathcal{E}_{i,\mathcal{X}^0}(g_i,g_i).$$

By Lemma 5,

$$\begin{aligned}\overline{\mathcal{E}}^I(\overline{g},\overline{g}) &\le \frac{3\lambda}{\theta}\sum_{i=1}^{L}\sum_{j=1}^{n}\frac{r_i w_{(i,j)}}{P_{(i,j)}(\mathcal{X}^0)}\,\mathrm{Var}_{P_{(i,j)},\mathcal{X}^0}(g_i) + \frac{3\lambda}{\theta}\mathcal{E}^I_{\mathcal{X}^0}(g,g)\\ &\le \frac{3\lambda C_1C_2}{\theta\phi}\sum_{i=1}^{L} r_i\mathcal{E}_{i,\mathcal{X}^0}(g_i,g_i) + \frac{3\lambda}{\theta}\mathcal{E}^I_{\mathcal{X}^0}(g,g)\end{aligned}$$

Hence,

$$\frac{1}{\theta^2}\,\mathrm{Var}_{P,[L]\times\mathcal{X}^0}(g) = \mathrm{Var}_{P_0}(g) \le \frac{3\lambda C_3}{\theta}\mathcal{E}^I_{\mathcal{X}^0}(g,g) + \frac{C_1C_2\left[(2+\lambda)C_3+1\right]}{\theta\phi}\sum_{i=1}^{L} r_i\mathcal{E}_{i,\mathcal{X}^0}(g_i,g_i).$$

Comparing with Lemma 3, we obtain the Poincaré inequality for M

$$\mathrm{Var}_{P,[L]\times\mathcal{X}^0}(g) \le \max\left\{3\theta C_3,\ \frac{\theta C_1C_2}{\phi(1-\lambda)}\left((2+\lambda)C_3+1\right)\right\}\mathcal{E}_{[L]\times\mathcal{X}^0}(g,g)$$

which concludes the proof of the theorem. □

B.3 Proof for the Mixing Times

We first recall the mixing time bound given in Atchadé et al. [2011] using restricted spectral gaps.

Lemma 6 (Atchadé et al. [2011])**.** *Let K be a lazy, reversible Markov transition kernel on a state space Ω, with stationary distribution Π. Suppose the initial distribution Π_0 is absolutely continuous with respect to Π, and define*

$$f_0(\omega)\Pi(\mathrm{d}\omega) = \Pi_0(\mathrm{d}\omega).$$

Assume there exist constants $B > 1$ and $q > 2$ such that $\|f_0\|_{\mathcal{L}^q(\Pi)} \leq B$, where $\|\cdot\|_{\mathcal{L}^q(\Pi)}$ denotes the $\mathcal{L}^q$-norm with respect to Π. Let $\varepsilon \in (0,1)$. Further, suppose there exists a measurable subset $\Omega^0 \subseteq \Omega$ such that

$$\Pi(\Omega^0) \geq 1 - \Big(\frac{\varepsilon}{20B^2}\Big)^{q/(q-2)}.$$

Then, for

$$N \geq \frac{1}{\mathrm{SpecGap}_{\Omega^0}(K)} \log\left(\frac{2B^2}{\varepsilon^2}\right),$$

the total variation distance between the distribution of the Markov chain K after N steps and its stationary distribution Π is at most ε.

Proof of Lemma 2. The first part of Equation (6) follows directly from Lemma 6. To prove the second part of (6), we first note that the TV distance between P^N and P admits the following lower bound

$$\|P^N - P\|_{\text{tv}} = \sum_{i=1}^{L} \int \big|P^N(i, \mathrm{d}x) - P(i, \mathrm{d}x)\big| \geq \int \big|p^N(i,x) - r_i p_i(x)\big| \mathrm{d}x, \quad \text{for all } i \in [L].$$

For each $i \in [L]$, let $r_{i,N} = P^N(i, \mathcal{X})$. The TV distance between P_i^N and P_i is bounded by

$$\begin{aligned}
\|P_i^N - P_i\|_{\text{tv}} &= \Big\| r_{i,N}^{-1} P^N(i,\cdot) \;-\; P_i \Big\|_{\text{tv}} = \int \Big| r_{i,N}^{-1} p^N(i,x) \;-\; p_i(x) \Big| \,\mathrm{d}x \\
&\leq \int \Big| r_{i,N}^{-1} p^N(i,x) \;-\; r_i^{-1} p^N(i,x) \Big| \,\mathrm{d}x + \int \big| r_i^{-1} p^N(i,x) \;-\; p_i(x) \big| \,\mathrm{d}x \\
&\leq \int \Big| r_{i,N}^{-1} p^N(i,x) \;-\; r_i^{-1} p^N(i,x) \Big| \,\mathrm{d}x + r_i^{-1} \|P^N - P\|_{\text{tv}},
\end{aligned}$$

where the first inequality follows from the triangle inequality. For the first term in the last expression, using $r_{i,N} = P^N(i, \mathcal{X})$ we get

$$\int \Big| r_{i,N}^{-1} p^N(i,x) \;-\; r_i^{-1} p^N(i,x) \Big| \,\mathrm{d}x = r_i^{-1} \left| r_i - r_{i,N} \right| \leq \frac{r_i^{-1}}{2} \|P^N - P\|_{\text{tv}},$$

where in the last step we use $r_i = P(A)$ and $r_{i,N} = P^N(A)$ with $A = \{i\} \times \mathcal{X}$.

Combining the above two displayed inequalities and using the first part of Equation (6), we get

$$\|P_i^N - P_i\|_{\text{tv}} \;\leq\; \frac{3}{2r_i}\varepsilon \leq \frac{3}{2 \min_{k \in [L]} r_k}\varepsilon.$$

This completes the proof. □

C Appendix for Section 3

C.1 Comparison of STMH chain with approximate STMH chain

To compare the STMH chain defined in Definition 4 with the approximate STMH chain in Definition 5, it suffices to compare the stationary density p_i^* with $\widetilde{p}_i$ and transition kernel M_i^* with $\widetilde{M}_i$. For the former, we use Lemma 7.3 of Ge et al. [2018], which shows that varying the temperature is roughly the same as changing the variance of a Gaussian distribution. For the latter, we derive a bound in Lemma 8.

Lemma 7 (Lemma 7.3 of Ge et al. [2018])**.** *Let $0 < \beta \leq 1$, and suppose $w_1, \ldots, w_n > 0$ are weights such that $\sum_{i=1}^n w_i = 1$. Define the density functions*

$$\pi(x) \propto \left(\sum_{i=1}^n w_i \pi_i(x) \right)^\beta \quad \textit{and} \quad \widetilde{\pi}(x) \propto \sum_{i=1}^n w_i \pi_i^\beta(x),$$

where $\pi_1, \ldots, \pi_n$ are component densities. Then,

$$w_{\min} \cdot \widetilde{\pi}(x) \leq \pi(x) \leq \frac{1}{w_{\min}} \cdot \widetilde{\pi}(x),$$

where $w_{\min} := \min_{1 \leq i \leq n} w_i$.

Lemma 8. *For each $i \in [L]$, let M_i^* be the transition kernel defined in Definition 4, with transition density m_i^*. Also, let $\widetilde{M}_i$ be the transition kernel defined in Definition 5, with transition density $\widetilde{m}_i$. Assume that $w_{\min} := \min_{1 \leq j \leq m} w_j > 0$. Then, for all $x \neq y \in \mathbb{R}^d$, the following inequality holds*

$$\widetilde{m}_i(x, y) \;\leq\; \frac{1}{w_{\min}^2} \, m_i^*(x, y).$$

Proof. Let q denote the symmetric Gaussian proposal density used in the Metropolis–Hastings algorithms M_i^* and $\widetilde{M}_i$. Then, for $x \neq y$, the transition densities are given by

$$m_i^*(x, y) = q(x, y) \, \alpha_i^*(x, y), \quad \widetilde{m}_i(x, y) = q(x, y) \, \widetilde{\alpha}_i(x, y),$$

where

$$\alpha_i^*(x, y) = \min\left\{1, \; \frac{p_i^*(y)}{p_i^*(x)}\right\}, \quad \widetilde{\alpha}_i(x, y) = \min\left\{1, \; \frac{\widetilde{p}_i(y)}{\widetilde{p}_i(x)}\right\}.$$

Using Lemma 7, we have

$$\widetilde{p}_i(y) \leq \frac{1}{w_{\min}} p_i^*(y) \quad \text{and} \quad \widetilde{p}_i(x) \geq w_{\min} \, p_i^*(x),$$

which gives

$$\frac{\widetilde{p}_i(y)}{\widetilde{p}_i(x)} \leq \frac{1}{w_{\min}^2} \cdot \frac{p_i^*(y)}{p_i^*(x)}.$$

Hence,

$$\widetilde{m}_i(x, y) \leq \frac{1}{w_{\min}^2} q(x, y) \, \alpha_i^*(x, y) = \frac{1}{w_{\min}^2} m_i^*(x, y),$$

which completes the proof. □

From now on, we assume that $\mathcal{X}^0 \subseteq \mathbb{R}^d$ is a measurable subset. Our next result, Lemma 9, shows that it suffices to obtain a lower bound on the $[L] \times \mathcal{X}^0$-restricted spectral gap of the approximate STMH chain in order to derive a corresponding bound for the STMH chain.

Lemma 9. *Let M^* be the STMH chain defined in Definition 4, and let $\widetilde{M}$ be the approximate STMH chain defined in Definition 5. Assume that the mixture weights satisfy $w_{\min} := \min_{1 \leq j \leq n} w_j > 0$. Then, the $[L] \times \mathcal{X}^0$-restricted spectral gaps of M^* and $\widetilde{M}$ satisfy the inequality*

$$\mathrm{SpecGap}_{[L] \times \mathcal{X}^0}(\widetilde{M}) \;\leq\; \frac{1}{w_{\min}^5} \, \mathrm{SpecGap}_{[L] \times \mathcal{X}^0}(M^*).$$

Proof. Let p^* and $\widetilde{p}$ denote the stationary densities of the Markov chains M^* and $\widetilde{M}$, respectively. Then, for all $i \in [L]$ and $x \in \mathbb{R}^d$, we have

$$p^*(i,x) = r_i p_i^*(x) \quad \text{and} \quad \widetilde{p}(i,x) = r_i \widetilde{p}_i(x).$$

By Lemma 7, we have

$$w_{\min}\, \widetilde{p}_i(x) \le p_i^*(x) \le w_{\min}^{-1}\, \widetilde{p}_i(x) \tag{14}$$

for every (i,x), and the same inequality holds for $p^*(i,x)$ and $\widetilde{p}(i,x)$ since the weights $(r_i)_{i=1}^L$ are the same for P^* and $\widetilde{P}$. Let $\mathcal{E}^*_{[L]\times\mathcal{X}^0}, \widetilde{\mathcal{E}}_{[L]\times\mathcal{X}^0}$ denote the $[L]\times\mathcal{X}^0$-restricted Dirichlet forms associated with M^* and $\widetilde{M}$ respectively. Fix a function $g \in \mathcal{L}^2([L]\times\mathcal{X}, \widetilde{p})$ and define $g_i(x) := g(i,x)$ for each (i,x). By Definition 1, it suffices to show that

$$\mathrm{Var}_{P^*,[L]\times\mathcal{X}^0}(g) \le \frac{1}{w_{\min}^2}\mathrm{Var}_{\widetilde{P},[L]\times\mathcal{X}^0}(g), \text{ and } \mathcal{E}^*_{[L]\times\mathcal{X}^0}(g,g) \le \frac{1}{w_{\min}^3}\widetilde{\mathcal{E}}_{[L]\times\mathcal{X}^0}(g,g).$$

For the first inequality, it follows from (14) that

$$\begin{aligned}
\mathrm{Var}_{P^*,[L]\times\mathcal{X}^0}(g) &\le \frac{1}{2w_{\min}^2}\sum_{i=1}^L\sum_{j=1}^L \int_{\mathcal{X}^0}\int_{\mathcal{X}^0} \big(g(i,x)-g(j,y)\big)^2\, \widetilde{p}(i,x)\, \widetilde{p}(j,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{w_{\min}^2}\mathrm{Var}_{\widetilde{P},[L]\times\mathcal{X}^0}(g).
\end{aligned}$$

By Lemma 3, we have

$$\widetilde{\mathcal{E}}_{[L]\times\mathcal{X}^0}(g,g) = (1-\lambda)\sum_{i=1}^L r_i\, \widetilde{\mathcal{E}}_{i,\mathcal{X}^0}(g_i,g_i) + \lambda\, \widetilde{\mathcal{E}}^I_{\mathcal{X}^0}(g,g), \tag{15}$$

and $\mathcal{E}^*_{[L]\times\mathcal{X}^0}$ can be decomposed analogously. We will bound the two terms on the right-hand side of Equation (15) separately. For the first term, we apply Lemmas 7 and 8 to get

$$\begin{aligned}
\widetilde{\mathcal{E}}_{i,\mathcal{X}^0}(g_i,g_i) &= \frac{1}{2}\int_{\mathcal{X}^0}\int_{\mathcal{X}^0} \big(g_i(x)-g_i(y)\big)^2\, \widetilde{p}_i(x)\, \widetilde{M}_i(x,\mathrm{d}y)\,\mathrm{d}x \\
&\le \frac{1}{2w_{\min}^3}\int_{\mathcal{X}^0}\int_{\mathcal{X}^0} \big(g_i(x)-g_i(y)\big)^2\, p_i(x)\, M_i(x,\mathrm{d}y)\,\mathrm{d}x \\
&= \frac{1}{w_{\min}^3}\mathcal{E}^*_{i,\mathcal{X}^0}(g_i,g_i).
\end{aligned}$$

For the second term, we apply Lemma 7 to get

$$\begin{aligned}
\widetilde{\mathcal{E}}^I_{\mathcal{X}^0}(g,g) &= \frac{1}{4}\sum_{i,i'\in[L]\,:\,i'=i\pm1}\int_{\mathcal{X}^0} \big(g_i(x)-g_{i'}(x)\big)^2 \min\left\{r_i\widetilde{p}_i(x),\, r_{i'}\widetilde{p}_{i'}(x)\right\}\,\mathrm{d}x \\
&\le \frac{1}{4\,w_{\min}}\sum_{i,i'\in[L]\,:\,i'=i\pm1}\int_{\mathcal{X}^0} \big(g_i(x)-g_{i'}(x)\big)^2 \min\left\{r_i p_i(x),\, r_{i'}p_{i'}(x)\right\}\,\mathrm{d}x \\
&= \frac{1}{w_{\min}}\,\mathcal{E}^{*,I}_{\mathcal{X}^0}(g,g).
\end{aligned}$$

Combining both bounds, we obtain that $\mathcal{E}^*_{[L]\times\mathcal{X}^0}(g,g) \le w_{\min}^{-3}\widetilde{\mathcal{E}}_{[L]\times\mathcal{X}^0}(g,g)$, which concludes the proof. □

C.2 Restricted Spectral Gap of the Approximate STMH Chain

We begin by introducing some notation. For each $i \in [L]$ and $j \in [n]$, define the j-th component of the density $\widetilde{p}_i$ as

$$\widetilde{p}_{(i,j)}(x) \propto \exp\left\{-\frac{\beta_i}{2}(x-\mu_j)^\top\Sigma^{-1}(x-\mu_j)\right\}, \tag{16}$$

so that $\widetilde{p}_i$ is a weighted mixture of the $\widetilde{p}_{(i,j)}$'s:

$$\widetilde{p}_i(x) \propto \sum_{j=1}^{n} w_j \widetilde{p}_{(i,j)}(x).$$

Let $\widetilde{M}_{(i,j)}$ denote the Metropolis–Hastings transition kernel targeting $\widetilde{p}_{(i,j)}$, with a symmetric Gaussian proposal density $q(x, y) = \mathcal{N}(y; x, \eta I)$, where $\eta > 0$ is the step size. We set

$$\mathcal{X}^0 := \left\{ x \in \mathbb{R}^d \ : \ \|x\| \leq R \right\},$$

where $R > 0$ is a fixed radius. To obtain a lower bound on the $[L] \times \mathcal{X}^0$-restricted spectral gap of the approximate STMH chain, we invoke Theorem 1. Assumption 1 holds by our construction of $\widetilde{P}$. The following lemmas verify the other two assumptions required for this theorem.

C.2.1 Validation of Condition (2) in Assumption 2

Lemma 10. *For each $i \in [L]$, let $\widetilde{p}_i$ be the density defined in Equation* (9)*, and let $g_i \in \mathcal{L}^2(\mathcal{X}, \widetilde{p}_i)$. Then the following inequality holds*

$$\sum_{j=1}^{n} w_j \, \widetilde{\mathcal{E}}_{(i,j),\mathcal{X}^0}(g_i, g_i), \ \leq \ \widetilde{\mathcal{E}}_{i,\mathcal{X}^0}(g_i, g_i) \qquad \forall i \in [L],$$

where $\widetilde{\mathcal{E}}_{(i,j),\mathcal{X}^0}$ denotes the $\mathcal{X}^0$-restricted Dirichlet form of the kernel $\widetilde{M}_{(i,j)}$, and $\widetilde{\mathcal{E}}_{i,\mathcal{X}^0}$ denotes the $\mathcal{X}^0$-restricted Dirichlet form of the kernel $\widetilde{M}_i$, as defined in Definition 5.

In particular, for the approximate STMH chain $\widetilde{M}$ defined in Definition 5, condition (2) *holds with constant $C_1 = 1$.*

Proof. For any nonnegative real sequences $\{a_j\}$ and $\{b_j\}$, we have the inequality $\min\left\{\sum_j a_j, \ \sum_j b_j\right\} \ \geq \ \sum_j \min\{a_j, \ b_j\}$. Applying this to $\widetilde{p}_i = \sum_{j=1}^{n} w_j \widetilde{p}_{(i,j)}$, we obtain

$$\min\left\{\widetilde{p}_i(x), \widetilde{p}_i(z)\right\} \ \geq \ \sum_{j=1}^{n} w_j \min\left\{\widetilde{p}_{(i,j)}(x), \widetilde{p}_{(i,j)}(z)\right\}, \tag{17}$$

for all $x, z \in \mathbb{R}^d$ and $i \in [L]$. Let q denote the symmetric Gaussian proposal density associated with the Metropolis–Hastings algorithms $\widetilde{M}_i$ and $\widetilde{M}_{(i,j)}$. Then, applying Equation (17), we obtain

$$\begin{aligned}
\widetilde{\mathcal{E}}_{i,\mathcal{X}^0}(g_i, g_i) &= \frac{1}{2} \int_{\mathcal{X}^0} \int_{\mathcal{X}^0} \left(g_i(x) - g_i(z)\right)^2 q(x, z) \min\left\{\widetilde{p}_i(x), \widetilde{p}_i(z)\right\} \, \mathrm{d}x \, \mathrm{d}z \\
&\geq \frac{1}{2} \int_{\mathcal{X}^0} \int_{\mathcal{X}^0} \left(g_i(x) - g_i(z)\right)^2 q(x, z) \sum_{j=1}^{n} w_j \min\left\{\widetilde{p}_{(i,j)}(x), \widetilde{p}_{(i,j)}(z)\right\} \, \mathrm{d}x \, \mathrm{d}z \\
&= \sum_{j=1}^{n} w_j \widetilde{\mathcal{E}}_{(i,j),\mathcal{X}^0}(g_i, g_i).
\end{aligned}$$

This completes the proof of the Lemma. □

C.2.2 Validation of Condition (3) in Assumption 2

We lower bound the $\mathcal{X}^0$-restricted spectral gap of each Metropolis–Hastings chain $\widetilde{M}_{(i,j)}$ using the path method of Yuen [2000] in the following lemma.

Lemma 11. *Let $0 < \eta \leq R^2$. For each $i \in [L]$ and $j \in [n]$, the Markov chain $\widetilde{M}_{(i,j)}$ admits the following lower bound on its $\mathcal{X}^0$-restricted spectral gap*

$$\mathrm{SpecGap}_{\mathcal{X}^0}\left(\widetilde{M}_{(i,j)}\right) \ \geq \ \frac{\gamma_{\min}^{d/2} \eta^{3/2}}{13 R^{d+3}}.$$

In particular, for the approximate STMH chain $\widetilde{M}$ defined in Definition 5, condition (3) *holds with constant*

$$C_2 = \frac{13R^{d+3}}{\gamma_{\min}^{d/2}}.$$

Proof. We use the linear path method described in Section 2 of Yuen [2000]. This approach also extends to the restricted spectral gap setting; see, for example, Atchadé [2021] and Chang and Zhou [2024] where the canonical path method has been adapted to the restricted spectral gap in discrete spaces. For each pair $(x, y) \in \mathcal{X}^0 \times \mathcal{X}^0$, we construct a linear path connecting x to y, with all intermediate points lying in $\mathcal{X}^0$. Fix a step size $\delta > 0$, and define the number of steps along the path by

$$b_{xy} \coloneqq \left\lceil \frac{\|x - y\|}{\delta} \right\rceil.$$

The path is then given by

$$\gamma_{xy} = \left(\gamma_{xy}^{(0)}, \ldots, \gamma_{xy}^{(b_{xy})}\right),$$

where

$$\gamma_{xy}^{(i)} \coloneqq \frac{(b_{xy} - i)x + iy}{b_{xy}}, \quad \text{for } 0 \le i \le b_{xy}.$$

Let $\Gamma \coloneqq \{\gamma_{xy} : (x, y) \in \mathcal{X}^0 \times \mathcal{X}^0\}$ denote the collection of all such paths, and let E denote the set of all edges that appear in at least one path $\gamma_{xy} \in \Gamma$. The capacity of an edge $(u, v) \in E$ is given by

$$T_{(i,j)}(u, v) \coloneqq \widetilde{p}_{(i,j)}(u)\, \widetilde{m}_{(i,j)}(u, v) = \widetilde{p}_{(i,j)}(u)\, q(u, v) \min\left\{1,\ \frac{\widetilde{p}_{(i,j)}(v)}{\widetilde{p}_{(i,j)}(u)}\right\},$$

where $\widetilde{m}_{(i,j)}$ denotes the transition density corresponding to the kernel $\widetilde{M}_{(i,j)}$, and $q(u, v)$ is the Gaussian proposal density associated with kernel $\widetilde{M}_{(i,j)}$. As shown in Section 2 of Yuen [2000], the set of paths Γ satisfies the regularity conditions and, for any $(u, v) \in \gamma_{xy}$, the associated Jacobian satisfies $J_{x,y}(u, v) = b_{xy}^d$ (see [Yuen, 2000, page 5] for details). Then, by Theorem 2.1 and Corollary 2.4 in Yuen [2000], we have

$$\mathrm{SpecGap}_{\mathcal{X}^0}\left(\widetilde{M}_{(i,j)}\right) \ge \frac{1}{A} \tag{18}$$

where

$$A = \operatorname*{ess\,sup}_{(u,v)\in E} \left\{ \frac{1}{T_{(i,j)}(u, v)} \sum_{\gamma_{xy} \ni (u,v)} |\gamma_{xy}|\, \widetilde{p}_{(i,j)}(x)\, \widetilde{p}_{(i,j)}(y)\, b_{xy}^d \right\},$$

and $|\gamma_{xy}|$ denotes the length of the path γ_{xy}. Since $\tilde{p}_{(i,j)}$ is log-concave, for any $(u, v) \in \gamma_{xy}$, we have

$$\min\left\{\widetilde{p}_{(i,j)}(x),\, \widetilde{p}_{(i,j)}(y)\right\} \le \min\left\{\widetilde{p}_{(i,j)}(u),\, \widetilde{p}_{(i,j)}(v)\right\}.$$

Hence, $T_{(i,j)}(u, v) \ge q(u, v) \min\{\widetilde{p}_{(i,j)}(x),\, \widetilde{p}_{(i,j)}(y)\}$, and we can upper bound A as

$$A \le b^{d+1} \cdot \operatorname*{ess\,sup}_{(u,v)\in E} \left\{ q(u, v)^{-1} \sum_{\gamma_{x,y} \ni (u,v)} \widetilde{p}_{(i,j)}(z_{x,y}) \right\}, \tag{19}$$

where $b \coloneqq \max_{(x,y)\in\mathcal{X}^0\times\mathcal{X}^0} b_{x,y}$ and $z_{x,y}$ is defined as

$$z_{x,y} \coloneqq \begin{cases} x, & \text{if } \max\{\widetilde{p}_{(i,j)}(x), \widetilde{p}_{(i,j)}(y)\} = \widetilde{p}_{(i,j)}(x), \\ y, & \text{otherwise.} \end{cases}$$

Note that

$$\widetilde{p}_{(i,j)}(z_{x,y}) \ \le\ \frac{\beta_i^{d/2}}{(2\pi\gamma_{\min})^{d/2}},$$

and the proposal density $q(u, v)$ is given by

$$q(u, v) \ =\ \frac{1}{(2\pi\eta)^{d/2}} \exp\left(-\frac{\|v - u\|^2}{2\eta}\right).$$

Substituting this into Equation (19), we obtain

$$A \;\leq\; \frac{\beta_i^{d/2}\eta^{d/2}b^{d+3}}{\gamma_{\min}^{d/2}} \cdot \operatorname*{ess\,sup}_{(u,v)\in E}\left\{\exp\left(\frac{\|v-u\|^2}{2\eta}\right)\right\},$$

where we have also used that an edge (u, v) belongs to at most b^2 paths in Γ. Choose step size $\delta = \sqrt{5\eta}$, which yields

$$b \leq \left\lceil \frac{2R}{\sqrt{5\eta}} \right\rceil \leq \frac{R}{\sqrt{\eta}}.$$

Since $\beta_i \leq 1$, we obtain that

$$A \;\leq\; \frac{\beta_i^{d/2}R^{d+3}}{\gamma_{\min}^{d/2}\eta^{3/2}}e^{2.5} \leq \frac{13R^{d+3}}{\gamma_{\min}^{d/2}\eta^{3/2}}.$$

From Equation (18), we get

$$\operatorname{SpecGap}_{\mathcal{X}^0}\big(\widetilde{M}_{(i,j)}\big) \;\geq\; \frac{1}{A} \;\geq\; \frac{\gamma_{\min}^{d/2}\eta^{3/2}}{13R^{d+3}},$$

which concludes the proof. □

C.2.3 Auxiliary Lemmas

To verify Assumption 3 and compute the constant C_3 in condition (4), we will need several lemmas. The proof of Lemma 12 is omitted.

Lemma 12 (Canonical Paths Bound)**.** *Let $\mathcal{S}$ be a finite state space, and let K be the transition kernel of a reversible Markov chain on $\mathcal{S}$ with stationary distribution Π and Dirichlet form $\mathcal{E}$. For each pair of distinct states $x, y \in \mathcal{S}$, let γ_{xy} denote a path from x to y consisting of valid transitions under K, i.e.,*

$$x = x_0 \to x_1 \to x_2 \to \cdots \to x_{n-1} \to x_n = y.$$

Let $\Gamma = \{\gamma_{xy} : x, y \in \mathcal{S},\ x \neq y\}$ be the collection of such paths for all distinct pairs (x, y). The edge congestion associated with Γ is defined as

$$\rho_e(\Gamma) = \max_{\substack{u,v\in\mathcal{S}\\ K(u,v)>0}} \frac{1}{\Pi(u)K(u,v)} \sum_{\substack{(u,v)\in\gamma_{xy}\\ \gamma_{xy}\in\Gamma}} \Pi(x)\Pi(y)|\gamma_{xy}|,$$

where $|\gamma_{xy}|$ denotes the length of the path γ_{xy}. Then, for any function $g : \mathcal{S} \to \mathbb{R}$, the following Poincaré inequality holds

$$\operatorname{Var}_\Pi(g) \leq \rho_e(\Gamma)\,\mathcal{E}(g, g).$$

Lemma 13. *Let $\Pi, \tilde{\Pi}$ be two probability distributions (absolutely continuous with respect to each other) with density function $\pi, \tilde{\pi}$ respectively. Then,*

$$\int \min\{\pi(x), \tilde{\pi}(x)\}\,\mathrm{d}x = 1 - \frac{1}{2}\|\Pi - \tilde{\Pi}\|_{\mathrm{tv}} \geq 1 - \sqrt{\frac{1}{2}\mathrm{KL}(\Pi \,|\, \tilde{\Pi})}.$$

Lemma 14. *Let $|\Sigma|$ denote the determinant of a matrix Σ. The Kullback-Leiber divergence between two d-dimensional Gaussian distributions with equal means is given by*

$$\mathrm{KL}\left(N(\mu, \Sigma_1) \,|\, N(\mu, \Sigma_2)\right) = \frac{1}{2}\left\{\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \operatorname{tr}(\Sigma_2^{-1}\Sigma_1)\right\}.$$

Lemma 15. *Let $D \coloneqq \max\left\{\max_{k\in[n]}\|\mu_k\|,\ \sqrt{\gamma_{\min}}\right\}$. For each $i \in [L]$, $j \in [n]$, define*

$$\widetilde{p}_i(j \,|\, x) := \frac{w_j\,\widetilde{p}_{(i,j)}(x)}{\sum_{k=1}^{n} w_k\,\widetilde{p}_{(i,k)}(x)} \qquad \text{for all } x \in \mathbb{R}^d,$$

where $\widetilde{p}_{(i,j)}(x)$ denotes the density defined in Equation (16)*. Then, for all $i \in [L]$, $j \in [n]$, and $x \in \mathbb{R}^d$, the following inequalities hold*

$$\widetilde{p}_{(i,j)}(x) \geq \left(\frac{\beta_i}{2\pi\gamma_{\max}}\right)^{d/2} \exp\left(-\frac{\beta_i}{2\gamma_{\min}}(\|x\| + D)^2\right), \tag{20}$$

$$\widetilde{p}_i(j \,|\, x) \geq w_j \exp\left(-\frac{\beta_i}{\gamma_{\min}}(\|x\| + D)^2\right). \tag{21}$$

Proof. To prove Equation (20), we write

$$\widetilde{p}_{(i,j)}(x) \geq \left(\frac{\beta_i}{2\pi\gamma_{\max}}\right)^{d/2} \exp\left(-\frac{\beta_i}{2}(x-\mu_j)^\top \Sigma^{-1}(x-\mu_j)\right),$$

where the inequality follows from the bound $|\Sigma| \leq \gamma_{\max}^d$. Next, we use the inequality

$$\|(x-\mu_j)^\top \Sigma^{-1}(x-\mu_j)\| \leq \frac{1}{\gamma_{\min}}\|x-\mu_j\|^2 \leq \frac{1}{\gamma_{\min}}(\|x\|+\|\mu_j\|)^2 \leq \frac{1}{\gamma_{\min}}(\|x\|+D)^2$$

to obtain

$$\widetilde{p}_{(i,j)}(x) \geq \left(\frac{\beta_i}{2\pi\gamma_{\max}}\right)^{d/2} \exp\left(-\frac{\beta_i}{2\gamma_{\min}}(\|x\|+D)^2\right).$$

This establishes Equation (20). To prove Equation (21), we define the function $\widetilde{J} : \mathbb{R}^d \to [n]$ by

$$\widetilde{J}(x) := \arg\max_{k\in[n]} \widetilde{p}_{(i,k)}(x).$$

It follows that for every $k \in [n]$,

$$\widetilde{p}_{(i,k)}(x) \leq \widetilde{p}_{(i,\widetilde{J}(x))}(x),$$

and therefore,

$$\sum_{k=1}^n w_k\, \widetilde{p}_{(i,k)}(x) \leq \sum_{k=1}^n w_k\, \widetilde{p}_{(i,\widetilde{J}(x))}(x) = \widetilde{p}_{(i,\widetilde{J}(x))}(x).$$

Substituting this upper bound into the definition of $\widetilde{p}_i(j \mid x)$, we get

$$\widetilde{p}_i(j \mid x) \geq \frac{w_j \widetilde{p}_{(i,j)}(x)}{\widetilde{p}_{(i,\widetilde{J}(x))}(x)}. \tag{22}$$

To simplify the ratio of Gaussian densities on the right-hand side, we expand the expression explicitly as

$$\frac{\widetilde{p}_{(i,j)}(x)}{\widetilde{p}_{(i,\widetilde{J}(x))}(x)} = \exp\left\{-\beta_i(\mu_{\widetilde{J}(x)}-\mu_j)^\top\Sigma^{-1}x - \frac{\beta_i}{2}\left(\mu_j^\top\Sigma^{-1}\mu_j - \mu_{\widetilde{J}(x)}^\top\Sigma^{-1}\mu_{\widetilde{J}(x)}\right)\right\}.$$

By definition of D, we have $\|\mu_{\widetilde{J}(x)} - \mu_j\| \leq 2D$. Using the Cauchy–Schwarz inequality, we obtain

$$\|(\mu_{\widetilde{J}(x)}-\mu_j)^\top\Sigma^{-1}x\| \leq \|\mu_{\widetilde{J}(x)}-\mu_j\| \cdot \|\Sigma^{-1}x\| \leq \frac{2D}{\gamma_{\min}}\|x\|,$$

and similarly,

$$\left\|\mu_j^\top\Sigma^{-1}\mu_j - \mu_{\widetilde{J}(x)}^\top\Sigma^{-1}\mu_{\widetilde{J}(x)}\right\| \leq \frac{1}{\gamma_{\min}}\left(\|\mu_j\|^2 + \|\mu_{\widetilde{J}(x)}\|^2\right) \leq \frac{2D^2}{\gamma_{\min}}.$$

Putting these together, we get

$$\frac{\widetilde{p}_{(i,j)}(x)}{\widetilde{p}_{(i,\widetilde{J}(x))}(x)} \geq \exp\left(-\frac{\beta_i}{\gamma_{\min}}(2D\|x\|+D^2)\right) \geq \exp\left(-\frac{\beta_i}{\gamma_{\min}}(\|x\|+D)^2\right). \tag{23}$$

Equations (22) and (23) together prove Equation (21). This completes the proof. □

C.2.4 Validation of Condition (4) in Assumption 3

Let $\widehat{M}$ denote the projected chain associated with the approximate STMH chain $\widetilde{M}$. We next establish a lower bound on the spectral gap of $\widehat{M}$ using the canonical paths method. Recall that we define $\mathcal{X}^0 = \{x \in \mathbb{R}^d : \|x\| \leq R\}$.

Lemma 16. *Let $\widetilde{M}$ denote the approximate STMH chain defined in Definition 5. Define the following parameters*

$$D \coloneqq \max\left\{\max_{k\in[n]}\|\mu_k\|,\, \sqrt{\gamma_{\min}}\right\}, \quad r \coloneqq \frac{\min_{i\in[L]} r_i}{\max_{i\in[L]} r_i}.$$

Suppose the following conditions hold

(i) Let $R \geq \sqrt{d}D$ be such that $P_{(i,j)}(\mathcal{X}^0) \geq 3/4$ for all $i \in [L]$ and $j \in [n]$,

(ii) $\beta_1 = \Theta\left(\gamma_{\min}/D^2\right)$ and $\beta_1 \leq 1$,

(iii) $\beta_{i+1}/\beta_i \leq 1 + 1/\sqrt{d}$ for all $i \in [L-1]$.

Let $\widehat{M}$ be the projected chain defined in Definition 3 associated with $\widetilde{M}$. Under these conditions, $\widehat{M}$ satisfies the spectral gap bound

$$\mathrm{SpecGap}(\widehat{M}) \;\geq\; \frac{3\min\{(1-\lambda),\lambda\}\, r^2}{64L^2\kappa^{d/2}\exp(cd)},$$

where $c > 0$ is a fixed constant. In particular, for the approximate STMH chain $\widetilde{M}$, condition (4) *holds with constant*

$$C_3 = \frac{64L^2\kappa^{d/2}\exp(cd)}{3\min\{(1-\lambda),\lambda\}\, r^2}.$$

Proof. We construct the canonical paths as follows. Fix two arbitrary states $x = (i,j)$, $y = (i',j') \in [L] \times [n]$ with $i \leq i'$.

(a) If $j = j'$, let γ_{xy} be $(i,j) \to (i+1,j) \to \ldots \to (i',j)$.

(b) If $j \neq j'$, let γ_{xy} be $(i,j) \to (i-1,j) \to \ldots \to (1,j) \to (1,j') \to (2,j') \to \ldots \to (i',j')$.

Define γ_{yx} as the reverse of γ_{xy}. Let Γ denote the collection of such paths over all distinct pairs (x,y). Let $i \in [L]$, $j \in [n]$, and $i' = i \pm 1 \in [L]$. From the definition of $\widehat{M}$, we have

$$\widehat{M}((i,j),(i',j)) = \frac{\lambda}{2}\int_{\mathcal{X}^0} \frac{\widetilde{p}_{(i,j)}(x)}{\widetilde{P}_{(i,j)}(\mathcal{X}^0)} \cdot \widetilde{a}\big((i,j,x),(i',j,x)\big)\,\mathrm{d}x,$$

where the acceptance probability is given by

$$\widetilde{a}\big((i,j,x),(i',j,x)\big) = \min\left\{\frac{r_{i'}\,\widetilde{p}_{(i',j)}(x)}{r_i\,\widetilde{p}_{(i,j)}(x)},\ 1\right\}.$$

Hence, the probability of transitioning from state (i,j) to $(i-1,j)$ under the projected chain $\widehat{M}$ is given by

$$\widehat{M}((i,j),(i-1,j)) = \frac{\lambda}{2\widetilde{P}_{(i,j)}(\mathcal{X}^0)}\int_{\mathcal{X}^0} \min\left\{\frac{r_{i-1}}{r_i}\,\widetilde{p}_{(i-1,j)}(x),\, \widetilde{p}_{(i,j)}(x)\right\}\mathrm{d}x,$$

for all $i \in \{2,\ldots,L\}$ and $j \in [n]$. Since $r_{i-1}/r_i \geq r$ by definition and $\widetilde{P}_{(i,j)}(\mathcal{X}^0) \leq 1$, we have

$$\widehat{M}((i,j),(i-1,j)) \geq \frac{\lambda r}{2}\int_{\mathcal{X}^0} \min\left\{\widetilde{p}_{(i-1,j)}(x),\, \widetilde{p}_{(i,j)}(x)\right\}\mathrm{d}x. \tag{24}$$

By Lemma 13 and Lemma 14,

$$\begin{aligned}\int_{\mathbb{R}^d} \min\left\{\widetilde{p}_{(i-1,j)}(x),\, \widetilde{p}_{(i,j)}(x)\right\}\mathrm{d}x &\geq 1 - \sqrt{\frac{1}{2}\mathrm{KL}(\widetilde{P}_{(i,j)} \,|\, \widetilde{P}_{(i-1,j)})} \\ &\geq 1 - \frac{\sqrt{d}}{2}\sqrt{f\left(\frac{\beta_i}{\beta_{i-1}}\right)},\ \text{where } f(x) = x - 1 - \log x.\end{aligned}$$

For $x \geq 1$, we have $f(x) \leq (x-1)^2/2$. Hence, if $\beta_i/\beta_{i-1} - 1 = 1/\sqrt{d}$, then

$$\int_{\mathbb{R}^d} \min\left\{\widetilde{p}_{(i-1,j)}(x),\, \widetilde{p}_{(i,j)}(x)\right\}\mathrm{d}x \geq \frac{1}{2}.$$

Condition (i) implies

$$\int_{\mathcal{X}^0} \min\left\{\widetilde{p}_{(i-1,j)}(x),\, \widetilde{p}_{(i,j)}(x)\right\}\mathrm{d}x \geq \frac{1}{4}.$$

Substituting the above bound in Equation (24), we obtain

$$\widehat{M}((i,j),(i-1,j)) \geq \frac{\lambda r}{8}.$$

Similarly, we can derive the bound

$$\widehat{M}((i,j),(i+1,j)) \geq \frac{\lambda r}{8}, \qquad \text{for all } i \in [L-1],\ j \in [n].$$

Next, we derive a lower bound on the transition probability from $(1,j)$ to $(1,j')$ in the projected chain $\widehat{M}$, which is given by

$$\widehat{M}((1,j),(1,j')) = (1-\lambda)\int_{\mathcal{X}^0} \frac{\widetilde{p}_{(1,j)}(x)}{\widetilde{P}_{(1,j)}(\mathcal{X}^0)} \cdot \widetilde{p}_1(j' \,|\, x)\,\mathrm{d}x, \qquad j,j' \in [n].$$

By applying Lemma 15 and noting that $\widetilde{P}_{(1,j)}(\mathcal{X}^0) \leq 1$, we obtain

$$\widehat{M}((1,j),(1,j')) \geq (1-\lambda)w_{j'}\left(\frac{\beta_1}{2\pi\gamma_{\max}}\right)^{d/2}\int_{\mathcal{X}^0}\exp\left(-\frac{2\beta_1}{\gamma_{\min}}(\|x\|+D)^2\right)\mathrm{d}x.$$

By condition (ii), there exist fixed constants $c_1, c_2 > 0$ such that

$$c_1\,\frac{\gamma_{\min}}{D^2} < \beta_1 < c_2\,\frac{\gamma_{\min}}{D^2}.$$

Let $\mathcal{X}^D \coloneqq \{x = (x_1,\ldots,x_d) \in \mathbb{R}^d : |x_i| \leq D \text{ for all } i \in [d]\} \subseteq \mathcal{X}^0$. Then for any $x \in \mathcal{X}^D$, we have $\|x\| + D \leq 2\sqrt{d}D$, which implies

$$\exp\left(-\frac{2\beta_1}{\gamma_{\min}}(\|x\|+D)^2\right) \geq \exp\left(-\frac{8\beta_1 dD^2}{\gamma_{\min}}\right) \geq \exp(-8c_2 d).$$

Therefore, we obtain the following lower bound

$$\widehat{M}((1,j),(1,j')) \geq (1-\lambda)w_{j'}(c_1)^{d/2}\left(\frac{\gamma_{\min}}{2\pi\gamma_{\max}D^2}\right)^{d/2}\exp(-8c_2d)\cdot \mathrm{Vol}(\mathcal{X}^D),$$

where $\mathrm{Vol}(\mathcal{X}^D)$ denotes the volume of the cube $\mathcal{X}^D$. Substituting $\mathrm{Vol}(\mathcal{X}^D) = (2D)^d$, we get

$$\begin{aligned}\widehat{M}((1,j),(1,j')) &\geq (1-\lambda)w_{j'}(c_1)^{d/2}\left(\frac{\gamma_{\min}}{2\pi\gamma_{\max}D^2}\right)^{d/2}\exp(-8c_2d)\cdot(2D)^d \\ &\geq (1-\lambda)w_{j'}\cdot\kappa^{-d/2}\cdot\exp(-cd), \end{aligned} \tag{25}$$

where $c > 0$ is a fixed constant.

Let γ_{xy} be a path between any two vertices $x, y \in [L]\times[n]$. Then, $|\gamma_{xy}| \leq 2L$. We now derive an upper bound on the edge congestion $\rho_e(\Gamma)$ defined in Lemma 12. Let $z, w \in [L]\times[n]$.

(a) Let $z = (i,j)$ and $w = (i-1,j)$. Then the edge (z,w) is used only by paths between vertices x and y such that one lies in the set $S \coloneqq \{i,\ldots,L\}\times\{j\}$, and the other in S^c. Therefore, its contribution to the edge congestion is bounded by

$$\frac{\sum_{(x,y)\in\Gamma:((i,j),(i-1,j))\in\gamma_{xy}}|\gamma_{xy}|\,\widehat{P}(x)\,\widehat{P}(y)}{\widehat{P}((i,j))\,\widehat{M}((i,j),(i-1,j))} \leq \frac{(2L)\,\widehat{P}(S)\,\widehat{P}(S^c)}{\widehat{P}((i,j))\,\widehat{M}((i,j),(i-1,j))},$$

where

$$\widehat{P}((\ell,k)) = r_\ell w_k \frac{\widetilde{P}_{(\ell,k)}(\mathcal{X}^0)}{\widetilde{P}([L]\times\mathcal{X}^0)}, \qquad \ell \in [L],\ k \in [n]$$

denotes the stationary distribution of the projected chain $\widehat{M}$. We have the following bounds

$$\frac{\widehat{P}(S)}{\widehat{P}((i,j))} = \frac{\widehat{P}(\{i,\ldots,L\}\times\{j\})}{\widehat{P}((i,j))} \leq \frac{4L}{3r}, \qquad \widehat{P}(S^c) \leq 1, \qquad \widehat{M}((i,j),(i-1,j)) \geq \frac{\lambda r}{8},$$

where the first bound follows from condition (i). Combining these, we conclude

$$\frac{\sum_{(x,y)\in\Gamma:((i,j),(i-1,j))\in\gamma_{xy}}|\gamma_{xy}|\,\widehat{P}(x)\,\widehat{P}(y)}{\widehat{P}((i,j))\,\widehat{M}((i,j),(i-1,j))} \leq \frac{64L^2}{3\lambda r^2}.$$

Similarly, we obtain the same bound for $z = (i,j)$ and $w = (i+1,j)$.

(b) Let $z = (1, j)$ and $w = (1, j')$. Then the edge (z, w) is used only by paths between vertices x and y such that one of them lies in the set $[L] \times \{j\}$ and the other in $[L] \times \{j'\}$. Therefore, its contribution to edge congestion is bounded by

$$\frac{\sum_{(x,y)\in\Gamma:((1,j),(1,j'))\in\gamma_{xy}} |\gamma_{x,y}|\, \widehat{P}(x)\, \widehat{P}(y)}{\widehat{P}((1,j))\, \widehat{M}((1,j),(1,j'))} \leq \frac{(2L)\, \widehat{P}([L]\times\{j\})\, \widehat{P}([L]\times\{j'\})}{\widehat{P}((1,j))\, \widehat{M}((1,j),(1,j'))}.$$

We now bound each term on the right-hand side. By condition (i), we have

$$\widehat{P}([L]\times\{j\}) \leq \frac{4L}{3r}\, \widehat{P}((1,j)), \qquad \widehat{P}([L]\times\{j'\}) \leq \frac{4w_{j'}}{3},$$

and from Equation (25), we have

$$\widehat{M}((1,j),(1,j')) \geq (1-\lambda) w_{j'} \cdot \kappa^{-d/2} \cdot \exp(-cd).$$

Combining these, we obtain

$$\frac{\sum_{(x,y):((1,j),(1,j'))\in\gamma_{xy}} |\gamma_{xy}|\, \widehat{P}(x)\, \widehat{P}(y)}{\widehat{P}((1,j))\, \widehat{M}((1,j),(1,j'))} \leq \frac{32L^2\kappa^{d/2}\exp(cd)}{9(1-\lambda)r}.$$

Let $\lambda \in (0,1)$ be a fixed constant. Thus, the edge congestion associated with Γ is bounded by

$$\rho_e(\Gamma) \leq \frac{64L^2\kappa^{d/2}\exp(cd)}{3\min\{(1-\lambda),\lambda\}\, r^2}.$$

From Lemma 12, projected chain $\widehat{M}$ satisfies the Poincaré inequality

$$\mathrm{Var}_{\widehat{P}}(\widehat{g}) \leq \frac{64L^2\kappa^{d/2}\exp(cd)}{3\min\{(1-\lambda),\lambda\}\, r^2}\, \widehat{\mathcal{E}}(\widehat{g},\widehat{g}), \quad \forall\, \widehat{g}\colon [L]\times[n] \to \mathbb{R},$$

where $\widehat{\mathcal{E}}$ denotes the Dirichlet form associated with the projected chain $\widehat{M}$. This completes the proof of the lemma. □

C.2.5 Restricted Spectral Gap Bound

We now invoke Theorem 1 to bound the $[L] \times \mathcal{X}^0$-restricted spectral gap of the approximate STMH chain $\widetilde{M}$, as formalized in the next lemma.

Lemma 17. *Let $\widetilde{M}$ denote the approximate STMH chain, as defined in Definition 5, with λ being a fixed constant. Under the same conditions as in Lemma 16, the $[L] \times \mathcal{X}^0$-restricted spectral gap of $\widetilde{M}$ admits the following lower bound*

$$\mathrm{SpecGap}_{[L]\times\mathcal{X}^0}(\widetilde{M}) \;\geq\; \Omega\left(\frac{\gamma_{\min}^{d/2} r^2 \eta^{3/2}}{R^{d+3} L^2 \kappa^{d/2}\exp(cd)}\right),$$

where $c > 0$ is a fixed constant.

Proof. For the approximate STMH chain $\widetilde{M}$, Assumption 2 is satisfied with constants $C_1 = 1$ and

$$C_2 = \frac{13R^{d+3}}{\gamma_{\min}^{d/2}\eta^{3/2}}.$$

Additionally, Assumption 3 holds with

$$C_3 = \frac{64L^2\kappa^{d/2}\exp(cd)}{3\min\{(1-\lambda),\lambda\}\, r^2},$$

where $c > 0$ is a fixed constant. Since λ is treated as fixed, combining these with Theorem 1 completes the proof of the lemma. □

C.3 Restricted Spectral Gap of the STMH chain

We now establish a lower bound on the $[L] \times \mathcal{X}^0$-restricted spectral gap of the STMH chain M^*, as formalized in next lemma.

Lemma 18. *Let M^* be the STMH chain, as defined in Definition 4, with λ being a fixed constant. Under the same conditions as in Lemma 16, the $[L] \times \mathcal{X}^0$-restricted spectral gap of M^* admits the lower bound*

$$\mathrm{SpecGap}_{[L]\times\mathcal{X}^0}(M^*) \;\geq\; \Omega\left(\frac{w_{\min}^5 \gamma_{\min}^{d/2} r^2 \eta^{3/2}}{R^{d+3} L^2 \kappa^{d/2} \exp(cd)}\right),$$

where $c > 0$ is a fixed constant.

Proof. This follows directly from Lemma 9 and Lemma 17. □

C.4 Estimation of Partition Functions

C.4.1 Assumptions on the Parameters

We now describe how to choose the algorithm parameters—number of temperature levels L, inverse temperature sequence $(\beta_i)_{i=1}^L$, temperature-swap rate λ, proposal step size η, initial covariance matrix Σ_0, number of iterations N—so that the STMH algorithm achieves the asserted time complexity. Recall that $\kappa = \gamma_{\max}/\gamma_{\min}$.

$$L = \Theta\left[\kappa\left\{D^2 + \log w_{\min}^{-1} + d\,(1 + \log\kappa)\right\}\log\left(\frac{D^2}{\gamma_{\min}}\right) + 1\right], \tag{26}$$

$$\beta_1 = \Theta\left(\frac{\gamma_{\min}}{D^2}\right),\ \frac{\beta_{i+1}}{\beta_i} \leq \min\left\{1 + \frac{1}{\sqrt{d}},\ \frac{\gamma_{\min}}{D^2 + 2\gamma_{\max} d\nu}\right\} \text{ for } i \in [L-1], \tag{27}$$

$$\text{where } \nu = 1 + \log\kappa + \frac{2}{d}\log\left(\frac{2}{w_{\min}}\right),$$

$$R = D + \sqrt{d\kappa D^2} + \sqrt{2\kappa D^2 \log\left(\frac{20\, e^6 L^2 \kappa^d}{w_{\min}^2 \varepsilon}\right)}, \tag{28}$$

$$N \geq \frac{C' L^4 R^d \kappa^{d/2} \exp(c'd)}{\gamma_{\min}^{d/2} w_{\min}^5} \log\left(\frac{L^2\kappa^d}{\varepsilon^2 w_{\min}^2}\right),\ \text{for some fixed constants } c', C' > 0, \tag{29}$$

$$\Sigma_0 = \sigma_0^2 I,\ \text{where } \sigma_0^2 = \Theta\left(\frac{\gamma_{\min}}{\beta_1}\right), \tag{30}$$

$$\lambda \text{ is any fixed constant,} \tag{31}$$

$$\eta \geq R^2. \tag{32}$$

C.4.2 Auxiliary Lemmas

Lemma 19. *Let $L > 0$ be an integer. Assume the partition–function estimates $\widehat{Z}_1, \ldots, \widehat{Z}_L$ satisfy*

$$\frac{\widehat{Z}_i/Z_i}{\widehat{Z}_1/Z_1} \in \left[(1 - \tfrac{1}{L})^{i-1},\, (1 + \tfrac{1}{L})^{i-1}\right], \qquad \textit{for all } i \in [L]. \tag{33}$$

Define

$$r_i \;:=\; \frac{Z_i/\widehat{Z}_i}{\sum_{k=1}^{L} Z_k/\widehat{Z}_k}, \qquad \textit{for all } i \in [L].$$

Then,

$$\frac{e^{-2}}{L} \;\leq\; r_i \;\leq\; \frac{e^2}{L}, \qquad \textit{for all } i \in [L]. \tag{34}$$

Moreover,

$$r := \frac{\min_{i\in[L]} r_i}{\max_{i\in[L]} r_i} \geq e^{-4}.$$

Proof. For each $i \in [L]$, define $b_i := Z_i/\widehat{Z}_i$, and denote their sum by S,

$$S := \sum_{k=1}^{L} b_k.$$

Then $r_i = b_i/S$. From Equation (33) we have, for every $i \in [L]$,

$$(1+\tfrac{1}{L})^{-(i-1)} \;\leq\; \frac{b_i}{b_1} \;\leq\; (1-\tfrac{1}{L})^{-(i-1)}, \tag{35}$$

which gives

$$L\, b_1\, (1+\tfrac{1}{L})^{-(L-1)} \;\leq\; S \;\leq\; L\, b_1\, (1-\tfrac{1}{L})^{-(L-1)}. \tag{36}$$

Combining Equations (35) with (36), we get

$$r_i = \frac{b_i}{S} \;\geq\; \frac{b_1(L+1)^{-(L-1)}}{L\, b_1\, (L-1)^{-(L-1)}} = \frac{1}{L}\Big(\frac{L-1}{L+1}\Big)^{L-1},$$

and

$$r_i \;\leq\; \frac{b_1(L-1)^{-(L-1)}}{L\, b_1\, (L+1)^{-(L-1)}} = \frac{1}{L}\Big(\frac{L+1}{L-1}\Big)^{L-1}.$$

Define

$$C_L \;:=\; \Big(\frac{L+1}{L-1}\Big)^{L-1}.$$

Taking logarithms and using $\log(1+x) \leq x$ for all $x > -1$, we obtain

$$\log C_L = (L-1)\, \log\Big(1+\tfrac{2}{L-1}\Big) \;\leq\; (L-1)\Big(\tfrac{2}{L-1}\Big) \;\leq\; 2,$$

so $C_L \leq e^2$. Likewise $C_L^{-1} \geq e^{-2}$. Therefore

$$\frac{e^{-2}}{L} \;\leq\; r_i \;\leq\; \frac{e^2}{L}, \qquad \text{for all } i \in [L],$$

establishing (34). This further implies that $r \geq e^{-4}$, completing the proof of the lemma. □

Lemma 20. *Let $L > 0$ be an integer, and suppose β_1 and σ_0 satisfy Equation* (27) *and Equation* (30)*, respectively. Then the initial density is given by*

$$p^0(1,\cdot) = \mathcal{N}\left(0, \sigma_0^2 I\right), \qquad p^0(i,\cdot) = 0 \quad \textit{for all } i \in [L]\setminus\{1\},$$

with the corresponding distribution denoted by P^0. The stationary density of the STMH chain M^*, as defined in Definition 4, is

$$p(i,x) = r_i\, p_i^*(x), \quad i \in [L],\, x \in \mathbb{R}^d,$$

where $(r_i)_{i=1}^L$ are defined in Lemma 19, and the component densities $(p_i^*)_{i\in[L]}$ are given in Equation (8)*. Let P denote the corresponding stationary distribution. Define $f_0 := dP^0/dP$. Then,*

$$\|f_0\|^2_{\mathcal{L}^2(P)} \leq \frac{c_1\, \exp(c_2 d)\, L\, \kappa^{d/2}}{w_{\min}},$$

for some fixed constants $c_1, c_2 > 0$.

Proof. The $\mathcal{L}^2$-norm of f_0 is given by

$$\|f_0\|^2_{\mathcal{L}^2(P)} = \sum_{i=1}^{L} \int p(i,x)\, |f_0(i,x)|^2\, dx = \frac{1}{r_1} \int \frac{(p^0(1,x))^2}{p_1^*(x)}\, dx, \tag{37}$$

where the second equality follows from the fact that P^0 is supported only on $i = 1$. By Lemma 7, for all $x \in \mathbb{R}^d$, we have

$$p_1^*(x) \;\geq\; w_{\min}\, \widetilde{p}_1(x) = w_{\min} \sum_{j=1}^{n} w_j \widetilde{p}_{(1,j)}(x). \tag{38}$$

where $\widetilde{p}_1$ is defined in Equation (9) and $\widetilde{p}_{(1,j)}$ is defined in Equation (16). Substituting (38) into (37), we obtain

$$\|f_0\|^2_{\mathcal{L}^2(P)} \leq \frac{1}{r_1 w_{\min}} \int \frac{(p^0(1,x))^2}{\sum_{j=1}^n w_j \widetilde{p}_{(1,j)}(x)}\, dx.$$

Using the convexity of the χ^2-divergence, we further bound this as

$$\|f_0\|^2_{\mathcal{L}^2(P)} \leq \frac{1}{r_1 w_{\min}} \sum_{j=1}^{n} w_j \int \frac{(p^0(1,x))^2}{\widetilde{p}_{(1,j)}(x)}\, dx. \tag{39}$$

For each $j \in [n]$, the density $\widetilde{p}_{(1,j)}(x)$ can be lower bounded as

$$\widetilde{p}_{(1,j)}(x) \geq \left(\frac{\beta_1}{2\pi\gamma_{\max}}\right)^{d/2} \exp\left(-\frac{\beta_1}{2\gamma_{\min}}\|x-\mu_j\|^2\right). \tag{40}$$

Since $p^0(1,\cdot) \sim \mathcal{N}(0, \sigma_0^2 I)$, we have

$$(p^0(1,x))^2 = \left(\frac{1}{2\pi\sigma_0^2}\right)^d \exp\left(-\frac{1}{\sigma_0^2}\|x\|^2\right).$$

From Equation (30), we have fixed constants $0 < s_1, s_2 < 2$ such that

$$s_1 \frac{\gamma_{\min}}{\beta_1} \leq \sigma_0^2 \leq s_2 \frac{\gamma_{\min}}{\beta_1}.$$

This gives

$$(p^0(1,x))^2 \leq \left(\frac{\beta_1}{2\pi s_1 \gamma_{\min}}\right)^d \exp\left(-\frac{\beta_1}{s_2\gamma_{\min}}\|x\|^2\right). \tag{41}$$

Substituting Equations (40) and (41) into Equation (39), we obtain

$$\begin{aligned}
&\|f_0\|^2_{\mathcal{L}^2(P)} \\
&\leq \frac{\kappa^{d/2}}{s_1^d\, r_1 w_{\min}} \sum_{j=1}^{n} w_j \left(\frac{\beta_1}{2\pi\gamma_{\min}}\right)^{d/2} \exp\left(\frac{\beta_1}{(2-s_2)\gamma_{\min}}\|\mu_j\|^2\right) \int \exp\left(-\frac{\beta_1(2-s_2)}{2\gamma_{\min} s_2}\left\|x + \frac{s_2\,\mu_j}{2-s_2}\right\|^2\right) \mathrm{d}x \\
&= \frac{\kappa^{d/2}\, s_2^{d/2}}{s_1^d\,(2-s_2)^{d/2}\, r_1 w_{\min}} \sum_{j=1}^{n} w_j \exp\left(\frac{\beta_1}{(2-s_2)\gamma_{\min}}\|\mu_j\|^2\right).
\end{aligned}$$

From Equation (27), we have a fixed constant $s_3 > 0$ such that $\beta_1 \leq s_3\gamma_{\min}/D^2$. Substituting this, we get

$$\|f_0\|^2_{\mathcal{L}^2(P)} \leq \frac{\kappa^{d/2}\, s_2^{d/2}}{s_1^d\,(2-s_2)^{d/2}\, r_1 w_{\min}} \exp\left(\frac{s_3}{2-s_2}\right).$$

By Lemma 19, we have $r_1 \geq 1/(e^2 L)$. Substituting this into the above bound on $\|f_0\|^2_{\mathcal{L}^2(P)}$ proves the lemma. □

Lemma 21. *Let X follow the d-dimensional Gaussian distribution with mean μ and covariance matrix Σ. Denote the largest eigenvalue of Σ by $\|\Sigma\|$. Then,*

$$\mathbb{P}\left(\|X\| \leq \|\mu\| + \sqrt{d\|\Sigma\|} + \sqrt{2\|\Sigma\|\log(1/\varepsilon)}\right) \geq \varepsilon.$$

Proof. Using the standard concentration inequality for Lipschitz functions of Gaussian random vectors, we get

$$\mathbb{P}\left(\|X-\mu\| \geq \mathbb{E}(\|X-\mu\|) + t\right) \leq e^{-t^2/2\|\Sigma\|}.$$

Since $\mathbb{E}(\|X-\mu\|) \leq \sqrt{d\|\Sigma\|}$, we get

$$\mathbb{P}\left(\|X-\mu\| \geq \sqrt{d\|\Sigma\|} + t\right) \leq e^{-t^2/2\|\Sigma\|}.$$

Letting $t = \sqrt{2\gamma_{\max}\log(1/\varepsilon)}$ and applying triangle inequality, we get the asserted bound. □

Lemma 22. *Suppose $1 \leq \ell \leq L$, and let Algorithm 1 be run with the potential function $f(x)$ defined in Equation* (7)*, inverse temperatures $\beta_1 < \cdots < \beta_\ell$, and using the parameters specified in Equations* (27)*,* (28)*,* (29)*,* (30)*,* (31) *and* (32)*. Assume that the partition function estimates $\widehat{Z}_1, \ldots, \widehat{Z}_\ell$ satisfy Equation* (33)*. Let P^N denote the distribution obtained after running Algorithm 1 for N steps, and let P denote its stationary distribution. Then the total variation distance between P and P^N satisfies*

$$\|P - P^N\|_{\mathrm{tv}} \;\leq\; \varepsilon.$$

Proof. Under the assumptions of the lemma, and by Lemmas 18 and 19, the $[\ell] \times \mathcal{X}^0$-restricted spectral gap of M^* satisfies

$$\mathrm{SpecGap}_{[\ell]\times\mathcal{X}^0}(M^*) \;\geq\; \Omega\left(\frac{w_{\min}^5\,\gamma_{\min}^{d/2}}{R^d\,\ell^4\,\kappa^{d/2}\,\exp(cd)}\right),$$

where $c > 0$ is a fixed constant. Moreover, from Lemma 20, $\|f_0\|^2_{\mathcal{L}^2(P)}$ is bounded above by B, where

$$B = \frac{c_1\,\exp(c_2 d)\ell\kappa^{d/2}}{w_{\min}},$$

and $c_1, c_2 > 0$ are fixed constants. Applying Lemma 6 with the above parameters yields the desired total variation bound. □

Lemma 23. *Assume the same conditions and notations as in Lemma 22. Let P_ℓ^* denote the marginal stationary distribution at temperature level ℓ, with density $p_\ell^*(x) \propto \exp(-\beta_\ell f(x))$, and let P_ℓ^N denote the marginal distribution at level ℓ after running Algorithm 1 for N steps, with density $p_\ell^N(x) \propto P^N(\ell, x)$. Then the total variation distance between P_ℓ^* and P_ℓ^N is bounded by*

$$\|P_\ell^* - P_\ell^N\|_{\mathrm{tv}} \;\leq\; \frac{3e^2\ell}{2}\,\varepsilon.$$

Proof. By Lemma 19, we have $\min_{i\in[\ell]} r_i \geq 1/(e^2\ell)$. The proof now follows directly from Lemmas 2 and 22. □

In the following lemma, we analyze how many times Algorithm 1 must be re-run, with a fixed number of steps N, in order to obtain a sample from the desired temperature level.

Lemma 24. *Suppose the partition function estimates $\widehat{Z}_1, \ldots, \widehat{Z}_\ell$ satisfy Equation* (33)*. Let $I_N \in [\ell]$ denote the temperature index of the state returned after running Algorithm 1 for N steps. Suppose the algorithm is run independently T times, each for N steps. Then, for any fixed temperature level $k \in [\ell]$, if*

$$T \;\geq\; e^2\ell\log\left(\tfrac{1}{\alpha}\right), \qquad \alpha \in (0,1),$$

the probability that at least one of the T runs returns a sample from level k satisfies

$$\mathsf{P}\left(\exists\, t \in [T] \text{ such that } I_N^{(t)} = k\right) \;\geq\; 1-\alpha,$$

where $I_N^{(t)}$ is the temperature level returned in the t-th run.

Proof. Let $k \in [\ell]$. From Lemma 19, we have

$$\mathsf{P}(I_N \neq k) \;=\; 1 - \mathsf{P}(I_N = k) \;\leq\; 1 - \frac{1}{e^2\,\ell}.$$

Hence,

$$\mathsf{P}\left(\nexists\, t \in [T] \text{ such that } I_N^{(t)} = k\right) \leq \left(1 - \tfrac{1}{e^2\,\ell}\right)^T \;\leq\; \exp\!\left(-\tfrac{T}{e^2\,\ell}\right).$$

Setting this upper bound no larger than δ, and solving for T, completes the proof of the lemma. □

Assuming the partition function estimates satisfy Equation (33), we have shown that the algorithm reaches total variation distance at most ε within the time complexity specified in Equation (29). We now show that partition function estimates satisfy Equation (33). By combining these two components, we establish the overall time complexity for the complete algorithm.

Lemma 25. *Let $\delta \in (0,1)$ and $1 \leq \ell \leq L$. Suppose the parameters satisfy Equations* (26)*,* (27)*,* (28)*,* (30)*,* (31)*, and* (32)*, and assume that the partition function estimates $\widehat{Z}_1, \ldots, \widehat{Z}_\ell$ satisfy Equation* (33)*. Let $s = L^2 \log(1/\delta)$. Collect s samples from Algorithm 1, denoted by $(x_j)_{j=1}^s$. Define the next partition function estimate $\widehat{Z}_{\ell+1}$ as*

$$\widehat{Z}_{\ell+1} := \overline{r}\,\widehat{Z}_\ell, \quad \textit{where} \quad \overline{r} := \frac{1}{s}\sum_{j=1}^{s} \exp\!\big(-(\beta_{\ell+1} - \beta_\ell) f(x_j)\big).$$

Then, with probability at least $1-\delta$, the estimate $\widehat{Z}_{\ell+1}$ also satisfies Equation (33)*. In particular,*

$$\left|\frac{\widehat{Z}_{\ell+1}/Z_{\ell+1}}{\widehat{Z}_1/Z_1}\right| \in \left[\left(1 - \frac{1}{L}\right)^\ell, \left(1 + \frac{1}{L}\right)^\ell\right].$$

The proof of Lemma 25 requires the following results.

Lemma 26 (Lemma 9.1 of Ge et al. [2018])**.** *Suppose that P_1 and P_2 are probability measures on Ω with density functions (with respect to a reference measure)*

$$p_1(x) = \frac{g_1(x)}{Z_1}, \quad \textit{and} \quad p_2(x) = \frac{g_2(x)}{Z_2}.$$

Suppose $\widetilde{P}_1$ is a measure such that $\|\widetilde{P}_1 - P_1\|_{tv} < c/2C^2$, and $g_2(x)/g_1(x) \in [0, C]$ for all $x \in \Omega$. Given n samples $x_1, \ldots, x_n$ from $\widetilde{P}_1$, define the random variable

$$\overline{r} = \frac{1}{n}\sum_{i=1}^{n} \frac{g_2(x_i)}{g_1(x_i)}.$$

Let

$$r = \mathsf{E}_{x\sim P_1} \frac{g_2(x)}{g_1(x)} = \frac{Z_2}{Z_1}.$$

and suppose $r \geq 1/C$. Then with probability at least $1 - e^{-nc^2/(2C^4)}$,

$$\left|\frac{\overline{r}}{r} - 1\right| \leq c.$$

Lemma 27 (Lemma G.16 of Ge et al. [2018])**.** *Suppose that $f(x) = -\log\left[\sum_{i=1}^n w_i\, e^{-f_i(x)}\right]$, where $f_i(x) = f_0(x - \mu_i)$, and $f_0 \colon \mathbb{R}^d \to \mathbb{R}$ is a κ-strongly convex and K-smooth function. For any $a > 0$, let P_a denote the probability measure with density $p_a(x) \propto e^{-af(x)}$. Let Z_a be the corresponding normalization constant, given by $Z_a = \int_{\mathbb{R}^d} e^{-af(x)}\,\mathrm{d}x$. Suppose that $\|\mu_i\| \leq D$ for all $i \in [n]$, and let $\alpha, \beta > 0$. Let*

$$A = D + \frac{1}{\sqrt{\alpha\kappa}}\left(\sqrt{d} + \sqrt{d\log\left(\frac{K}{\kappa}\right) + 2\log\left(\frac{2}{w_{\min}}\right)}\right).$$

If $\alpha < \beta$, then

$$\min_{x\in\mathbb{R}^d} \frac{p_\alpha(x)}{p_\beta(x)} \geq \frac{Z_\beta}{Z_\alpha} \qquad \textit{and} \qquad \frac{Z_\beta}{Z_\alpha} \in \left[\frac{1}{2}e^{-\frac{1}{2}(\beta-\alpha)KA^2}, 1\right].$$

Proof of Lemma 25. By Equation (27) and Lemma 27, we have

$$\frac{\exp(-\beta_{\ell+1} f(x))}{\exp(-\beta_\ell f(x))} = \exp\left(-(\beta_{\ell+1} - \beta_\ell) f(x)\right) \in [0, 1/(2\,e)]$$

for all $\ell \in [L-1]$. Moreover, by substituting $\varepsilon = 4/(3\ell L)$ into Lemma 23, we obtain

$$\|P_\ell^* - P_\ell^{\widetilde{N}}\|_{\mathrm{tv}} \;\le\; \frac{2e^2}{L},$$

when

$$\widetilde{N} \geq \frac{C' L^4 R^d \kappa^{d/2} \exp(c'd)}{\gamma_{\min}^{d/2} w_{\min}^5} \log\left(\frac{L^4 \kappa^d}{w_{\min}^2}\right),$$

where $C', c' > 0$ are fixed constants. Next, by applying Lemma 26 with constants $C = 1/2e$ and $c = 1/L$, we obtain the following bound

$$\left|\frac{\widehat{Z}_{\ell+1}/Z_{\ell+1}}{\widehat{Z}_\ell/Z_\ell}\right| \in \left[1 - \frac{1}{L},\; 1 + \frac{1}{L}\right].$$

The lemma then follows by induction on ℓ. □

C.4.3 Proof of Theorem 2

Proof of Theorem 2. Let L denote the number of temperature levels defined in Equation (26). By applying Lemma 25 inductively with $\delta = \varepsilon/(4L)$, we obtain that, with probability at least $1 - \varepsilon/4$, the following bound holds

$$\frac{\widehat{Z}_\ell}{Z_\ell} \in \left[\left(1 - \frac{1}{L}\right)^{\ell-1}, \left(1 + \frac{1}{L}\right)^{\ell-1}\right] \cdot \frac{\widehat{Z}_1}{Z_1} \qquad \text{for all } \ell \in [L].$$

To ensure this guarantee, it suffices to generate $s = L^2 \log(4L/\varepsilon)$ samples from each temperature level $i \in [L]$, resulting in a total of $sL = L^3 \log(4L/\varepsilon)$ samples from Algorithm 1. Applying Lemma 24 with $\alpha = \varepsilon/(4L^4 \log(4L/\varepsilon))$, we obtain that, with probability at least $1 - \varepsilon/4$, we obtain s samples from each temperature level $i \in [L]$ by running Algorithm 1 for N steps (as defined in Equation (29)) and repeating this process independently T times, where

$$T = sL \cdot e^2 L \log\left(\frac{1}{\alpha}\right) = e^2 L^4 \log\left(\frac{4L}{\varepsilon}\right) \log\left(\frac{4L^4}{\varepsilon} \log\left(\frac{4L}{\varepsilon}\right)\right).$$

Hence, the total time complexity for getting partition function estimates is

$$T_{\text{partition}} = T \cdot N = \frac{C'\, L^8 R^d \kappa^{d/2} \exp(c'd)}{\gamma_{\min}^{d/2} w_{\min}^5} \log^3\left(\frac{L\kappa}{\varepsilon w_{\min}}\right),$$

where $c', C' > 0$ are fixed constants. By applying Lemma 23 and Lemma 24, we conclude that, with probability at least $1 - \varepsilon/4$, Algorithm 1 produces a sample from a distribution that is within total variation distance $\varepsilon/4$ of the target distribution P^* in time T_{sampling}, where

$$\begin{aligned} T_{\text{sampling}} &= e^2\, L\, \log\left(\frac{4}{\varepsilon}\right) \frac{C'' L^4 R^d \kappa^{d/2} \exp(c''d)}{\gamma_{\min}^{d/2} w_{\min}^5} \log\left(\frac{L^2 \kappa^d}{\varepsilon^2 w_{\min}^2}\right) \\ &= \frac{C' L^5 R^d \kappa^{d/2} \exp(c'd)}{\gamma_{\min}^{d/2} w_{\min}^5} \log^2\left(\frac{L\kappa}{\varepsilon w_{\min}}\right), \end{aligned}$$

where $c', C', c'', C'' > 0$ are fixed constants. The overall time complexity T consists of two components: the time to get partition function estimates, and the time to generate sample from the target distribution

$$T = T_{\text{partition}} + T_{\text{sampling}}.$$

This completes the proof of the theorem.

□