SAMPLING FROM MULTIMODAL DISTRIBUTIONS WITH WARM STARTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling from multimodal distributions is a central challenge in Bayesian inference and machine learning. In light of hardness results for sampling—classical MCMC methods, even with tempering, can suffer from exponential mixing times—a natural question is how to leverage additional information, such as a warm start point for each mode, to enable faster mixing across modes. For this problem, we prove the first polynomial-time bound that works in a general setting, under a natural assumption that each component contains significant mass relative to the others when tilted towards the corresponding warm start point. For this, we introduce a modified version of the Annealed Leap-Point Sampler (ALPS) (Tawn et al., 2021; Roberts et al., 2022). Similarly to ALPS, we define distributions tilted towards a mixture centered at the warm start points, and at the coldest level, use teleportation between warm start points to enable efficient mixing across modes. In contrast to ALPS, our method does not require Hessian information at the modes, but instead estimates component partition functions via Monte Carlo. This additional estimation step is critical in allowing the algorithm to handle target distributions with more complex geometries besides approximate Gaussian. For the proof, we show convergence results for Markov processes when only part of the stationary distribution is well-mixing and estimation for partition functions for individual components of a mixture. We numerically evaluate our algorithm's mixing performance on a mixture of heavy-tailed distributions, comparing it against the ALPS algorithm on the same distribution.

1 INTRODUCTION

A key task in statistics and machine learning is sampling from a probability distribution known up to normalization, $\pi(x) \propto e^{-V(x)}$. The standard approach of Markov Chain Monte Carlo (MCMC) is to define a Markov chain with stationary distribution $\pi(x)$. The time it takes for MCMC methods to produce an approximate sample from $\pi(x)$ depends on the mixing time of the underlying Markov chain. Unfortunately, in many applications, the target distribution $\pi(x)$ is multimodal, which causes Markov chains with local moves to mix slowly, as transitions between different modes rarely occur; this is the general phenomenon of metastability Bovier et al. (2002).

Modern MCMC methods such as simulated tempering Marinari & Parisi (1992), parallel tempering (also known as replica exchange) Swendsen & Wang (1986), Sequential Monte Carlo (also known as particle filtering) Del Moral et al. (2006), and annealed importance sampling Neal (2001) attempt to speed up sampling by running a Markov chain with a sequence of interpolating distributions $p_\beta(x) \propto e^{-\beta V(x)}$ or $p_\beta(x) \propto \pi(x)^\beta p_0(x)^{1-\beta}$ at varying inverse temperatures β. The idea is at high temperatures the Markov chain can more easily mix between modes of the target distribution.

Recent analysis of these methods has shown that it is possible to obtain non-asymptotic mixing time bounds for multimodal stationary distributions with polynomial dependence on parameters, but only under restrictive assumptions. Indeed, there are simple families of multimodal distributions with bottlenecks arising from low-weight components, which require exponentially many queries to sample Ge et al. (2018a) (see also Example B.1). This motivates the search for algorithms that leverage more information, such as approximate location of modes, which we term *warm starts*. We formalize the problem of sampling with warm starts, and prove that our algorithm has polynomial running times under general assumptions.

To motivate the algorithm, we note another approach to multimodal sampling is using mode jump samplers such as Tjelmeland & Hegstad (2001); Ibrahim (2009); Lindsey et al. (2022). These algorithms address poor mixing due to multimodality by allowing samples to jump (teleport) between modes of the target distribution. However, in high dimensions the Markov process can have very low acceptance rates when jumping between modes, because arbitrary distributions will in general have exponentially small overlap even when superimposed. Tawn et al. (2021) cleverly combine tempering with teleportation in the Annealed Leap-Point Sampler (ALPS) which, given a warm starts $x_1, \dots, x_M$ to the modes, anneals the target distribution to colder temperature. At the coldest temperature, the distribution is peaked around the warm start points, and samples can leap from mode to mode of the peaked distributions with high acceptance probability. Note that annealing to cold temperatures is exactly the opposite of how tempering methods typically function.

In this paper, we prove non-asymptotic bounds in total variation (TV) for a modified version of the ALPS algorithm (Roberts et al., 2022). We make some modifications for technical convenience of the analysis, and one critical modification that prevents bottlenecks in modes at different levels from arising under general assumptions. Instead of using weight-preserving power tempering Tawn et al. (2020b), we let the intermediate distributions be $\pi_\beta(x) \propto \pi(x) \cdot \sum_{k=1}^{M} w_{\beta,k} q_\beta(x - x_k)$ with $w_{\beta,k}$ dynamically chosen by the algorithm. This allows eliminating bottlenecks without requiring the components are approximately Gaussian, and does not require Hessian information.

Importantly, our results are free of functional inequalities that depend on the global geometry of the target density. Instead, we prove upper bounds on mixing time for the underlying Markov process in the algorithm in terms of local Poincaré constants (capturing local mixing) alone. Our analysis proceeds through a Markov chain decomposition theorem Madras & Randall (2002); Ge et al. (2018a), which requires us to bound the Poincaré constant of a certain projected chain (capturing mixing between components). This Poincaré constant is bounded by appropriate algorithmic choice of level and component weights $r_i, w_{i,k}$ and temperature ladder β_i.

We overcome two new technical challenges in the analysis. First, the tempering scheme can create bad components, so we develop new theoretical analyses for Markov chains that show mixing in the "good" part of the stationary distribution. Second, in addition to estimating the partition function of the tempered distributions π_{β_i} for each level i to balance the levels (via r_i), we also need to estimate the partition functions for the components $\pi_{\beta_i,k}$ of the mixture, in order to balance the modes (via $w_{i,k}$) and avoid a bottleneck in the projected chain. We show that the partition functions can be approximated using Monte Carlo; the proof requires a technically involved analysis due to possible interference between different components.

1.1 Sampling with different kinds of advice

We are interested in the problem of approximately sampling from $\pi(x) \propto e^{-V(x)}$ which is multimodal. A common way of formalizing the multimodality is to assume that $\pi = \sum_{i=1}^{m} w_i \pi_i$, where each component π_i satisfies a functional inequality; that is, the natural Markov chain on the space mixes rapidly. We classify approaches to this problem depending on the strength of extra information, or *advice* that we are given. Here, we focus on approaches with theoretical guarantees.

No advice. Without extra information, guarantees are available only under strong conditions. Early work gives guarantees for simulated and parallel tempering assuming suitable decompositions (Madras & Randall, 2002; Woodard et al., 2009a). Ge et al. (2018a) show that simulated tempering combined with Langevin dynamics works for a mixture of translates of distributions satisfying a log-Sobolev inequality (e.g. a mixture of Gaussians with equal covariance); this is generalized to other Markov processes by Garg et al. (2025). For sequential Monte Carlo, Paulin et al. (2018); Mathews & Schmidler (2024) show guarantees for multimodal distributions but require separation between modes. Lee & Santana-Gijzen (2024) allow a general mixture but assume component weights do not change between temperatures, which is relaxed by Han et al. (2025).

An inherent challenge that leads to restrictive assumptions in the above results is the following: in general, a component can have smaller weights at higher temperatures, creating a "bottleneck" that prevents samples from moving into that mode. In simple terms, it is generally difficult to find a mode. A simple example is that of two Gaussians with different covariances. Woodard et al. (2009b) observe exponential lower bounds for simulated and parallel tempering in this setting. More

generally, considering a family of perturbations of such distributions, no algorithm can generate a sample within constant TV distance with sub-exponentially many queries to π or $\nabla \ln \pi$ (Ge et al., 2018a). Reweighting is a possible solution (Tawn et al., 2020b) but relies on components being located and approximable by nice distributions such as gaussians.

Strong advice. Given strong advice in the form of a few samples from the target distribution, Koehler & Vuong (2023); Koehler et al. (2025); Gay et al. (2025) show that the problem is generically solvable: for a mixture with m components, given $O(m/\epsilon^2)$ samples, a fresh sample within distance ϵ in TV can be generated by simply running the Markov chain starting from a random sample; this is termed data-based initialization. This framework works for both continuous and discrete settings. Although the assumption is strong, it is reasonable in the setting of generative modeling, when a dataset of samples is given and the task is to learn to generate new samples.

Weak advice. Given the impossibility results in the setting of no advice and the lack of strong advice in many problems, it is natural to try for general results given weaker information. As mode location is an inherent challenge, a natural assumption to isolate the search problem from the sampling problem is to assume we already have *warm starts* to the modes, e.g. obtained by multiple runs of optimization. Tawn et al. (2021) introduce the annealed-leap point sampler, which combines tempering towards a mixture of peaked distributions, with teleportation, and gives asymptotic analysis in the limit as the modes become gaussian (Roberts et al., 2022). Another kind of information which can be considered as weak advice is that of a few *reaction coordinates* that are assumed to be the main obstacle to fast mixing; algorithms can take advantage of this by stratifying the landscape and forcing exploration in those directions. Examples include umbrella sampling Torrie & Valleau (1977); Thiede et al. (2016) (with analysis in Dinner et al. (2020)), the Wang–Landau algorithm Wang & Landau (2001), and adaptive biasing force Darve & Pohorille (2001).

Theoretical tools. We highlight some theoretical tools that are useful for analyzing sampling for multimodal distributions. Firstly, Markov chain decomposition theorems Madras & Randall (2002); Woodard et al. (2009a); Ge et al. (2018a) or two-scale functional inequalities Otto & Reznikoff (2007); Grunewald et al. (2009); Lelièvre (2009); Chen et al. (2021) show that functional inequalities for a Markov chain or process hold given that they hold locally (restricted to some component or coordinate) and that they hold for a projected process that tracks flow or closeness between the components. A number of works quantify and apply local mixing: Although Langevin diffusion does not generally converge quickly, Balasubramanian et al. (2022) show it is efficient to sample with small relative Fisher information, which for a mixture, corresponds to local mixing within modes but not necessarily global mixing between modes. Huang et al. (2025) show that sampling is possible under weak Poincaré inequalities Andrieu et al. (2023) when a warm start can be maintained. Finally, partition function estimation using simulated annealing and Monte Carlo Dyer et al. (1991); Štefankovič et al. (2009) is well-studied theoretically.

1.2 Problem Statement & Assumptions

We address the problem of sampling from a target distribution $\pi(x) \propto e^{-V(x)}$ with oracle access to the target distribution $\pi(x)$ by utilizing a set of *warm start points* $\{x_1, \ldots, x_M\}$. (We will formally define this below.) These can be obtained as prior information or from multiple runs of optimization algorithms. close to local maxima of the target distribution $\pi(x)$.

Problem 1.1. *Suppose we are given a set of "warm starts" $\{x_1, \ldots, x_M\}$ to the modes of a target distribution $\pi(x) = \sum_{k=1}^{M} \alpha_k \pi_k(x)$. Assume query access to $\pi(x)$ up to a normalization constant, and possibly $\nabla \ln \pi$ (in the case of $\mathbb{R}^d$). Produce a sample that is ϵ-close in total variation distance to $\pi(x)$.*

Note that we only assume the existence of a decomposition of π, not that the π_k are known. To introduce our algorithm, we fix a family of density functions q_β on X, which have the property $q_\beta \to \delta_0$ weakly as $\beta \to \infty$ and $q_\beta = 1$ if $\beta = 0$. For example, the q_β could be Gaussians in $\mathbb{R}^d$ or product distributions on the hypercube. We will apply simulated tempering to the sequence of

distributions (for β ranging from 0 to very large)

$$\tilde{p}_\beta(x) \propto \pi(x) \sum_{k=1}^{M} w_{\beta,k} q_\beta(x - x_k),$$

for some weights $w_{\beta,k}$ estimated by the algorithm. (On the hypercube, addition is understood in $\mathbb{Z}/2$.) Essentially, we tilt the target distribution towards the set of warm start points. At the coldest level, the distribution becomes approximately a mixture of Dirac deltas, and because of their similar shape, the teleportation step of our algorithm allows samples to move between the warm start points.

We will make the following assumptions on $\pi(x)$ and its components $\pi_k(x)$. The general idea of the warm start assumption (part 2 below) is that a significant portion of the mass should be located in the component that corresponds to the product between the component $\pi_k(x)$ and the q_β centered at the corresponding warm start point x_k. We call this the *tilt* towards x_k of the distribution π_k. In addition, we assume that each of these tilts satisfy a Poincaré inequality.

Assumption 1.1. *Suppose that $\pi(x)$ is a distribution on Ω, $q_\beta : \Omega \to \mathbb{R}, \beta \geq 0$ are functions with $q_0 \equiv 1$. Fix a way of associating a distribution $p(x)$ on Ω with a Markov process with generator $\mathscr{L}_p$ that has p as stationary distribution.*

1. *(Mixture distribution) The target distribution $\pi(x)$ is a mixture distribution $\pi(x) = \sum_{k=1}^{M} \alpha_k \pi_k(x)$, where $\alpha_k \geq 0$ and π_k is a probability distribution.*
2. *(x_i are warm starts) For each $i \in [M]$ and for every $\beta \geq 0$,*
$$\int_X \alpha_i \pi_i(x) q_\beta(x - x_i) dx \geq c_{tilt} \int_X \pi(x) q_\beta(x - x_i) dx.$$
3. *(Local mixing) For all $i \in [M]$, $p_{\beta,i}(x) \propto \pi_i(x) q_\beta(x - x_i)$ satisfies a Poincaré inequality of the form*
$$\mathrm{Var}_{p_{\beta,i}}(f) \leq C_{\mathrm{P}} \mathscr{E}_{p_{\beta,i}}(f, f),$$
where $\mathscr{E}_\pi(f, f) = -\langle f, \mathscr{L}_\pi f \rangle_\pi$ is the Dirichlet form and $\mathscr{L}_\pi$ is the generator of the Markov process with stationary distribution π.
4. *(Markov chain decomposes) Whenever $p = \sum_{i=1}^{m} a_i p_i$, $a_i \geq 0$ is a mixture distribution on Ω, the generators decompose: $\langle f, \mathscr{L}_p f \rangle_p \leq \sum_{i=1}^{m} a_i \langle f, \mathscr{L}_{p_i} f \rangle_{p_i}$.*

Many common Markov chains have generators which satisfy the last assumption, for example, Langevin diffusion or the Metropolis random walk on $\mathbb{R}^d$ and Glauber dynamics on product spaces. See Lee & Santana-Gijzen (2024) for a complete discussion with proofs. We work with continuous-time Markov processes for technical convenience; any discrete time Markov chain can be converted to a continuous-time by letting the waiting time between jumps be exponential random variables. For a discussion of the limitations of the warm start assumption, see Section 6. For the main theorem, we will make some additional assumptions on the tempering scheme in Assumptions 3.1.

2 Algorithms

2.1 Ingredients: Simulated tempering and teleportation

To introduce our main algorithm, we first define simulated tempering and the leap point process. These two algorithms will be the primary components of our main algorithm.

Simulated tempering is a classical approach to sampling from multimodal target distributions (Marinari & Parisi, 1992), where the target distribution is (typically) tempered to smoother (high-temperature) distributions that allow mixing between modes. Particles are allowed to transition between temperatures (with appropriate Metropolis-Hastings acceptance ratio) in addition to moving within their current temperature, and we take the samples that are at the desired temperature.

In our setting, the target measure is instead tempered to more peaked (colder temperature) distributions, and then at the coldest temperature the leap point process is applied to transition particles to

different components of the mixture measure. This works particularly well given a set of warm starts $\{x_k\}_{k=1}^M$ since the mixture $\pi_0(x)$ can be tempered to peak around each x_i and then a teleportation map $g_{jj'}$ can be defined to move points around x_j to around $x_{j'}$. We define simulated tempering generally to apply for any specified sequence of distributions.

Definition 2.1. *Given a sequence of Markov processes M_i with stationary distributions p_i, $1 \leq i \leq L$ on state space Ω and level weights r_i, we define* **simulated tempering** *to be the process on $\Omega \times [L]$ as follows. At each level $i \in [L]$ we are given a Markov process M_i with stationary $p_i(x)$. Then the simulated tempering Markov process is defined as follows:*

1. *Evolve $(x, i) \in \Omega$ according to M_i.*
2. *Propose jumps with rate λ. When a jump is proposed, leap to i' with probability*

$$\frac{1}{2} \min\left\{\frac{r_{i'} p_{i'}(x)}{r_i p_i(x)}, 1\right\}, \quad i' = i \pm 1; \tag{2.1}$$

otherwise stay at $i \in [L]$.

It is simple to check that the stationary distribution is $p(x, i) = \sum_{j=1}^L r_j p_j(x) I\{i = j\}$ Marinari & Parisi (1992); Neal (1996). In our case, the p_i will be chosen as $\tilde{p}_{\beta_i}$. For ease of notation, we will overload notation by replacing β_i by i in subscripts, e.g., $\tilde{p}_i := \tilde{p}_{\beta_i}$.

For the definition of the leap point process, we assume we are given a set of teleportation functions $g_{jj'}$; for $\mathbb{R}^d$, a simple choice is translation between modes, $g_{jj'}(x) = x + x_j - x_{j'}$.

Definition 2.2. *Given a set of teleportation functions g_{ij}, $i, j \in [M]$ satisfying Definition D.1 and a Markov process P with stationary distribution q, define the* **leap point process** *on the state space Ω as follows.*

1. *Evolve $x \in \Omega$ according to P.*
2. *Propose leaps with rate γ. When a leap is proposed, choose j and j' uniformly and leap to $g_{jj'}(x)$ with probability*

$$\frac{1}{M} \min\left\{\frac{g_{\#}^{jj'} q(x)}{q(x)}, 1\right\}, \quad \forall j' \neq j; \tag{2.2}$$

otherwise stay at x.

Note that we define j to be randomly sampled, which differs slightly from from Roberts et al. (2022); Tawn et al. (2020a; 2021), where the current position x of the Markov chain is assigned to a mode $j = \arg\min_k d(x, x_k)$; this is only for ease of analysis.

Our Markov chain uses simulated tempering to perform temperature swaps and employs the leap point process at the coldest temperature to mix between modes. Formally, we define our process on the level of the generators by adding together the original generator, the simulated tempering jumps, and the leaps at the coldest level ($i = 1$). We defer a formal treatment to Section D.1. See Algorithm 2 for pseudocode of simulation.

2.2 Weight Estimation & Level Balance

A challenge for tempering algorithms is the potential for bottlenecks to prevent the chain from exploring the entire state space. These bottlenecks form when the probability mass of specific modes becomes vanishingly small at certain levels. Our algorithm explicitly addresses this by iteratively estimating the weights—modal and level—to maintain the following balance condition.

Definition 2.3. *We define* **component balance** *with constant C_1 for partition functions $Z_{i,k} = \int_\Omega \alpha_k \pi_k(x) q_i(x - x_k) dx$ and weights $\{w_{i,k}\}_{i \in [L], k \in [M]}$ to be the condition*

$$\frac{w_{i,k} Z_{i,k}}{w_{i,k'} Z_{i,k'}} \in \left[\frac{1}{C_1}, C_1\right] \text{ for all } i \in [L], k, k' \in [M]. \tag{2.3}$$

We define **level balance** *with constant C_2 for partition functions $Z_i = \int_\Omega \pi(x) \cdot \sum_k w_{ik} q_i(x - x_k) dx$ and weights $\{r_i\}_{i=1}^L$ as*

$$\frac{r_i Z_i}{r_{i'} Z_{i'}} \in \left[\frac{1}{C_2}, C_2\right] \text{ for all } i, i' \in [L]. \tag{2.4}$$

Since level balance is enforced between each level, no exponentially bad bottlenecks can form between the coldest and warmest levels. Example B.2 illustrates a simple setting (a mixture of two Gaussians with different covariances) where exponential bottlenecks form between cold and warm levels.

2.3 Main Algorithm

Our algorithm is an inductive process which uses an auxiliary variable $\beta_1 > \cdots > \beta_L = 0$ to define a sequence of distributions $\{\sum_{k=1}^M w_{i,k} \pi_{i,k}(x)\}_{i=1}^L$ which temper peaked multimodal distributions to the target distribution. The main Algorithm 1 will inductively run Algorithm 2 (vanilla ALPS) to level l and then approximate the weights of the component functions at $l+1$ via Algorithm 3 (reweighting via partition function estimation). To start off, Algorithm 1 requires an estimation of the partition functions of $\tilde{\pi}_{1,k}$. In Section I, we show that under appropriate conditions, for large enough β_1, these estimates can be obtained.

Algorithm 2 (with all levels) is run once all the weights are learned; this is akin to a vanilla version of ALPS (Roberts et al., 2022). As described in Section 2.1, it incorporates simulated tempering to transition between adjacent temperature levels and the leap point process at the coldest level to transition between modes.

Algorithm 3 runs Algorithm 2 to level l to acquire N samples at the l-th level. Then the weights at level $l+1$ are approximated as $w_{l+1,k} = \left(\frac{1}{N}\sum_{j=1}^N \frac{\pi(x_j) q_{l+1}(x_j - x_k)}{\tilde{p}(x_j, i_j)} I\{i_j = l\}\right)^{-1}$ and $r_{l+1}^{(l)} = \left(\frac{1}{N}\sum_{i=1}^N \frac{\pi(x_j) \cdot \left(\sum_k w_{l+1,k} q_{l+1}(x_j - x_k)\right)}{\tilde{p}(x_j, i_j)} I\{i_j = l\}\right)^{-1}$. After these weights have been estimated, Algorithm 3 re-runs the chain, this time to level $l+1$, acquiring samples at levels. Then the level weights are adjusted via empirical occupancy, i.e., $r_i^{(l+1)} = r_i^{(l)} \Big/ \frac{1}{N}\sum_{j=1}^N I\{i_j = i\}$.

For clarity, we note our algorithm and analysis is akin to the vanilla version of ALPS (Roberts et al., 2022). This has the core algorithmic ideas of, but is different from the full version Tawn et al. (2021), which is equipped with online mode location and parallel tempering.

Algorithm 1 Main Algorithm: Simulated Tempering with Teleporting

INPUT: Temperature scale $\beta_1 > \beta_2 > \cdots > \beta_L = 0$ and weights $\{w_{1,k}\}_{k=1}^M$ satisfying level balance (2.3).
OUTPUT: A sample $x \in \mathbb{R}^d$

for $l = 1 \to L$ **do**
  Input weights $\{w_{i,k}\}$, $\{r_i^{(l)}\}_{i=1}^l$ for $i \in [1, l]$ and $k \in [1, M]$ with temperature scale $\beta_1 > \beta_2 > \cdots > \beta_l$, time T and rates λ, γ.
  if $l < L$ **then**
    Run Algorithm 3 (reweighting with partition function estimates) to obtain weights $\{w_{l+1,k}\}_{k=1}^M$ and $\{r_i^{(l+1)}\}_{i=1}^{l+1}$.
  else if $l = L$ **then**
    Run Algorithm 2 (vanilla ALPS) and return sample $x \in \mathbb{R}^d$.
  end if
end for

3 Main Result

We make some additional technical assumptions for the main theorem. We later show in Section I that these assumptions are satisfied in representative settings.

Assumption 3.1. *Defining* $\pi_{l,k}(x) \propto \alpha_k \pi_k(x) q_l(x - x_k)$, $\bar{\pi}_{l,k} = \pi(x) q_l(x - x_k)$, $Z_{l,k} = \int \bar{\pi}_{l,k}$ *for* $l \in [1, L]$, $k \in [1, M]$,

1. *(Closeness at adjacent temperatures)* $\chi^2(\pi_{l+1,k} || \pi_{l,k}) = O(1)$, $\chi^2(\frac{\bar{\pi}_{l+1,k}}{Z_{l+1,k}} \| \frac{\bar{\pi}_{l,k}}{Z_{l,k}}) = O(1)$.
2. *(Closeness for components at lowest temperature)* $\chi^2(\pi_{1,k} || \pi_{1,j}) = O(1)$.
3. *(Warmness of initial distribution) The initial distribution* $\nu_0(x, i)$ *satisfies* $\left\| \frac{\nu_0(x,i)}{p(x,i)} \right\|_\infty \leq U$.
4. *Component balance with constant* $O(1)$ *(Definition 2.3) is satisfied when* $L = 1$.

Given reasonable choices of tilting functions q_β, Assumption 1 requires the temperature ladder to be sufficiently closely spaced and Assumption 2 requires starting out at cold enough temperature so that teleportation is accepted with good probability. Assumption 3 requires the initialization of the samples to be close enough to the chain (this is possible by initializing at the lowest temperature, which is easily approximable). Assumption 4 is the base case of the inductive hypothesis and again depends on the lowest temperature distribution being approximable. As an example, we show these assumptions hold for Gaussian tilts on $\mathbb{R}^d$; we have not attempted to optimize the number of levels.

Proposition 3.2. *(Tempering by Gaussians) Let Assumptions 1.1 hold for* $\pi(x) = \sum_{k=1}^{M} \alpha_k \pi_k(x)$ *with* $\alpha_k \pi_k(x) = e^{-f_k(x)}$ *where* $f_k(x)$ *is L-smooth. In addition, assume that a log-Sobolev inequality holds with constant* C_{LS} *for* $\pi_{i,j,k} \propto \pi_j(x) \cdot q_i(x - x_k)$, *for all* $i \in [L], j, k \in [M]$. *Define* $q_i(x) = e^{-\beta_i \frac{\|x\|^2}{2}}$ *and the teleportation map* $g_{jj'}(x) = x - x_j + x_{j'}$. *Lastly, choose* $\Delta\beta = |\beta_i - \beta_{i+1}| = O(\frac{1}{C_{LS} d + r^2})$ *and* $\beta_1 = O(L^2 D^2 d)$, *with* $||x_j - x_j^*|| \leq D$, *where* x_j^* *is the true mode and* $||x_j - \mathbb{E}_{p_{i,j,k}} x|| \leq r$ *for all* $j, k \in [M]$. *Then Assumptions 3.1 hold with* $U = O\left(\frac{1}{c_{tilt}^2}\right)$ *and* $w_{1k} \propto \frac{1}{\pi(x_k)}$ *on a temperature schedule of* $\Omega(d^2)$ *levels.*

We now state our main theorem.

Theorem 3.3. *Suppose we are given a warm start of points* $\{x_1, \ldots, x_M\}$ *from a target distribution* $p(x)$. *Fix a family of density functions* $q_\beta, \beta > 0$ *on* X, *with* $q_\beta = 1$ *if* $\beta = 0$. *Suppose Assumptions 1.1 and 3.1 hold. Then Algorithm 1 with parameters*[1]

$$T = \Omega\left(\text{poly}\left(U, \tilde{C}, M, L, \frac{1}{c_{tilt}}, \frac{1}{\gamma}, \frac{1}{\lambda}, \frac{1}{\epsilon}\right)\right), \qquad N = \Omega\left(\text{poly}\left(L, M, U, \frac{1}{c_{tilt}}, \frac{1}{\delta}\right)\right)$$

produces samples from $\hat{p}(x)$ *such that with probability* $1 - \delta$, $TV(\hat{p}(x), \pi(x)) \leq \epsilon$.

4 Proof Overview

A standard approach to proving mixing time bounds for tempering Markov chains is to use a Markov decomposition theorem (Ge et al., 2018b). Decomposition theorems allow for mixing to be quantified in terms of mixing within the components and mixing within the projected chain defined through probability flow between components. However, in our setting, the tempered distributions $p_\beta(x) \propto \sum_j \alpha_j \pi_j(x) \cdot \sum_k w_{\beta,k} q_\beta(x - x_k)$ have cross terms of components tilted towards the wrong mode: $p_\beta(x) \propto \sum_k \alpha_k w_{\beta,k} \pi_k(x) q_\beta(x - x_k) + \sum_{j \neq k} \alpha_j w_{\beta,k} \pi_j(x) q_\beta(x - x_k) =: \tilde{p}_{\beta,0}(x) + \tilde{p}_{\beta,1}(x)$. The projected chain is no longer mixing on the entire distribution, only if we ignore the bad portion.

To remedy this issue, we formulate χ^2 bounds that quantify the mixing of the whole chain on the good component, in terms of the mixing on the good component, Lemma F.1. We accompany this provide a lower bound on the portion of the mass from the Markov chain within the good component, Lemma F.2 and then use this to quantify the rate at which that portion converges to the good component itself. Since our Markov chain operates on the extended state space $\Omega \times [L]$, we generalize these results to the extended state space and quantify the amount of mass that mixes into the target level (which is entirely in the good part), Lemma F.3.

[1] L is the number of levels and M the number of warm start points. c_{tilt}, C_{ij} and $\tilde{C} = \max_{ij} C_{ij}$ are defined in Assumptions 1.1, and γ, λ are hyper-parameters defined in D.5.

This analysis quantifies the mixing of the whole chain by the Poincaré constant of the good component $C_{PI}(p_0(x,i))$. Now, working within the good component, we are able to bound this Poincaré constant using classical Markov decomposition theorems, Theorem E.1, which upper bound $C_{PI}(p_0(x,i))$ by the mixing within the modes and the mixing on the projected chain. The most difficult part is ensuring that there are no bottlenecks in the projected chain which would cause the Poincaré constant to explode.

The Poincaré constant of the projected chain is controlled by the probability flow between modes, which is ultimately determined by the modal and level balance (Definition 2.3). By defining the good component as $p_0(x,i) \propto \sum_i r_i \sum_k \alpha_k w_{\beta_i,k} \pi_k(x) q_{\beta_i}(x - x_k)$, we have control over both the level weights $\{r_i\}_{i\in[L]}$ and the modal weights $\{w_{i,k}\}_{i\in[L],k\in[M]}$—think of these as knobs that tune the probability flow between component measures. Good level balance is ensured by induction on the temperature levels, at each level showing that Definition 2.3 holds with $C_1 = \text{poly}(\frac{U}{c_{tilt}})$ and $C_2 = \text{poly}(\frac{U}{c_{tilt}})$ for weights $\{w_{i,k}\}_{i\in[L],k\in[M]}$ and $\{r_i\}_{i\in[L]}$ approximated by Monte Carlo averages is done in Theorem G.13, G.17.

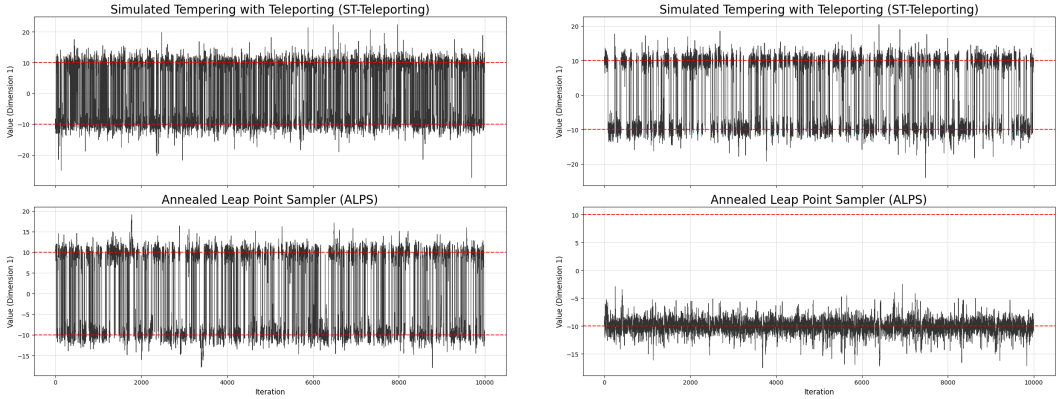
5 Experiments

We demonstrate that our ST Teleporting algorithm works well on a heavy-tailed mixture that the ALPS sampler Tawn et al. (2021) fails to sample from on as the dimension increases. First, we define the target distribution to be an unnormalized bimodal mixture of Student's t-distributions $\pi(x) \propto \sum_{k=1}^{2} (1 + \frac{\|x-\mu_k\|^2}{\nu})^{-\frac{\nu+d}{2}}$. We ran both our algorithm and the ALPS algorithm on this target distribution. We supplied the ALPS algorithm with the exact modes of the components μ_1, μ_2 and the Hessian information at those points, which is approximately $\frac{\nu+d}{\nu} I_d$.

We ran each algorithm for the same amount of time, with the same burn-in length and Metropolis random walk steps. The temperature swaps and mode jumps were simulated in discrete time using a Poisson process and we kept the hyperparameter for jump attempts and teleportation attempts the same across algorithms.

For both algorithms, we chose the coldest level to have $\beta_1 = 6$. For ALPS, we chose a geometric temperature ladder with 8 levels to the target level $\beta_8 = 1$ of $\beta_{\text{ALPS}} = (6, 4.65, 3.6, 2.78, 2.16, 1.67, 1.29, 1)$. We chose a similar ladder for ST teleporting; however the target level requires $\beta_8 = 0$ therefore we used the ladder $\beta_{\text{STTel}} = (6, 3.3, 1.75, .94, .51, .22, .1, 0)$.

ST Teleporting has one learning step where the modal weights and level weights are estimated. Running this with $N = 1500$ samples for the Monte Carlo estimates took 167 seconds when $d = 3$ and 178 seconds when $d = 10$.



(a) Student's t-distribution with $d = 3$

(b) Student's t-distribution with $d = 10$

Figure 1: The first component of $N = 10,000$ samples at the target level $\pi(x)$ with $\nu = 3$ and $(\mu_1, \mu_2) = (-10 \cdot \vec{1}_d, 10 \cdot \vec{1}_d)$.

Figure (1a) shows both algorithms traversing the state space, effectively reaching each modes as intended when the dimension is $d = 3$. After running both algorithms for $25,000$ potential temperature swaps the ALPS algorithm had a mode leap acceptance rate of .655 (249/380 attempts) and a temperature swap acceptance rate of .562 (14056/25000 attempts). By comparison, the ST teleporting algorithm had a mode swap acceptance rate of .753 (1148/1525 attempts) and a temperature swap rate of .765 (17973/25000 attempts). Notably, already in $d = 3$ the ALPS algorithm is struggling to transition samples to the coldest temperature. This leads to the ALPS algorithm having an ESS of 474 (4.7%) which is less than half that of ST Teleporting's ESS of 1014 (10.1%)—ESS is computed from samples at the target level.

Figure (1b) shows that when the dimension increases the ALPS algorithm fails to properly explore both modes. After running both algorithms for $25,000$ potential temperature swaps the ALPS algorithm had a mode leap acceptance rate of .25 (1/4 attempts) and a temperature swap rate of .092 (2293/25000 attempts). By comparison, ST teleporting had a mode leap acceptance rate of .745 (1356/1819 attempts) and a temperature swap rate of .728 (18201/25000 attempts).

This is what we would expect to see for ALPS on a heavy-tailed distribution. The temperature ladder $\pi_\beta(x) \propto \pi(x)^\beta \pi(\mu_k)^{1-\beta}$ sharpens the heavy tails of the target distribution too quickly causing samples at the target level on the tails to never accept temperature swaps. In contrast, our algorithm tames the tails lightly by the multiplicative factor of a flat gaussian mixture. This is failure in level re-weighting, this also holds when modal weights are varied.

6 Conclusion and Further Work

We prove the first general polynomial-time bounds for sampling from multimodal distributions under the "weak advice" of warm start points to the different modes. Our algorithm is a modified version of the ALPS algorithm Tawn et al. (2021) that is designed to work well on multi-modal target distributions with difficult geometries. The core innovation is a modified tempering schedule, $\pi_\beta(x) \propto \pi(x) \cdot \sum_k w_{\beta,k} q_\beta(x - x_k)$, where we estimate the weights via Monte Carlo simulation to keep components balanced. As our focus is on an initial theoretical analysis, there are several avenues for future theoretical and computational work towards making the algorithm practical.

Computationally speaking, our base algorithm has several computational inefficiencies, such as estimating the next level weights at every iteration from a fresh set of samples; it also requires warm start points to already be located. Hence, a beneficial modification would be to update weights and find additional modes in an online manner. As our algorithm is tailored for ease of theoretical analysis, we do not recommend it in its current form as a replacement for ALPS, and believe that a hybrid algorithm incorporating our approach to weight rebalancing may be ultimately more practical. We leave to future work the design of a more versatile and efficient algorithm which works on practical problems in high dimensions and with complex geometries.

Warm start assumption. An important limitation is our definition of a warm start point, in terms of the tilt having significant mass. Applied to components of different shapes (e.g., Gaussians in Proposition I.8), this may require separation conditions between the components. In order to loosen the definition of warm start point, we may need adaptive tilting schemes, e.g. Gaussians with covariances chosen adaptively. As a concrete theoretical problem to guide algorithm design, we propose this open question: *Is there a polynomial-time sampler for $\pi = \sum_{i=1}^M w_i \pi_i$ where each π_i is a log-concave distribution, given one sample $x_i \sim \pi_i$ from each component?*

Some more technical limitations of our warm start assumption is that (1) we currently assume a 1-to-1 correspondence between warm starts and modes, and (2) taking $\beta = 0$, each component is required to have mass that is lower-bounded. We can hope that this can be relaxed to making sure that all modes are covered (and allowing spurious points), and that modes having small mass can be disregarded.

Finally, theoretical analysis is also highly desirable for other algorithms in the weak advice setting, such as those based on stratification. Another promising direction is to combine information on warm starts with neural network flow-based methods for sampling Albergo et al. (2023); Vargas et al. (2023); Albergo & Vanden-Eijnden (2024), as well as learning the interpolation (Máté & Fleuret, 2023).

REFERENCES

Michael S Albergo and Eric Vanden-Eijnden. Nets: A non-equilibrium transport sampler. *arXiv preprint arXiv:2410.02711*, 2024.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Weak poincar\'e inequalities for markov chains: theory and applications. *arXiv preprint arXiv:2312.11689*, 2023.

Dominque Bakry, Ivan Gentil, and Michael Ledoux. *Analysis and geometry of Markov diffusion operators*. Springer International Publishing, 2014.

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*, pp. 2896–2923. PMLR, 2022.

Sergey G. Bobkov and Prasad Tetali. Modified logarithmic sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006. doi: 10.1007/s10959-006-0016-3. URL https://link.springer.com/article/10.1007/s10959-006-0016-3.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability and low lying spectra in reversible markov chains. *Communications in mathematical physics*, 228(2):219–255, 2002.

Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

Eric Darve and Andrew Pohorille. Calculating free energies using average force. *The Journal of chemical physics*, 115(20):9169–9183, 2001.

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. doi: 10.1111/j.1467-9868.2006.00553.x.

Aaron R Dinner, Erik H Thiede, Brian Van Koten, and Jonathan Weare. Stratification as a general variance reduction method for markov chain monte carlo. *SIAM/ASA journal on uncertainty quantification*, 8(3):1139–1188, 2020.

Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

Jhanvi Garg, Krishna Balasubramanian, and Quan Zhou. Restricted spectral gap decomposition for simulated tempering targeting mixture distributions. *arXiv preprint arXiv:2505.15059*, 2025.

William Gay, William He, Nicholas Kocurek, and Ryan O'Donnell. Sampling and identity-testing without approximate tensorization of entropy. *arXiv preprint arXiv:2506.23456*, 2025.

Rong Ge, Holden Lee, and Andrej Risteski. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. *Advances in neural information processing systems*, 31, 2018a.

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*, 2018b.

Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering langevin monte carlo ii: An improved proof using soft markov chain decomposition. 2018c.

Natalie Grunewald, Felix Otto, Cédric Villani, and Maria G Westdickenberg. A two-scale approach to logarithmic sobolev inequalities and the hydrodynamic limit. In *Annales de l'IHP Probabilités et statistiques*, volume 45, pp. 302–351, 2009.

Ruiyu Han, Gautam Iyer, and Dejan Slepčev. Polynomial complexity sampling from multimodal distributions using sequential monte carlo. *arXiv preprint arXiv:2508.02763*, 2025.

Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X Wu. Weak poincaré inequalities, simulated annealing, and sampling from spherical spin glasses. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 915–923, 2025.

Adriana Ibrahim. *New methods for mode jumping in Markov chain Monte Carlo algorithms*. PhD thesis, 2009.

Frederic Koehler and Thuy-Duong Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. *arXiv preprint arXiv:2310.01762*, 2023.

Frederic Koehler, Holden Lee, and Thuy-Duong Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3264–3326. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/koehler25a.html.

Holden Lee and Matheau Santana-Gijzen. Convergence bounds for sequential monte carlo on multimodal distributions using soft decomposition, 2024.

E. L. Lehmann. *Theory of point estimation*. January 1983. doi: 10.1007/978-1-4757-2769-2. URL https://doi.org/10.1007/978-1-4757-2769-2.

Tony Lelièvre. A general two-scale criteria for logarithmic sobolev inequalities. *Journal of Functional Analysis*, 256(7):2211–2221, 2009.

David Asher Levin, Yuval Peres, Elizabeth L. Wilmer, James Propp, and David B. Wilson. *Markov chains and mixing times*. American Mathematical Society, 2017.

Michael Lindsey, Jonathan Weare, and Anna Zhang. Ensemble markov chain monte carlo with teleporting walkers. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3):860–885, 2022.

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

Enzo Marinari and Giorgio Parisi. Simulated tempering: A new monte carlo scheme. *Europhysics Letters*, 19(6):451–458, 1992. doi: 10.1209/0295-5075/19/6/002.

Bálint Máté and François Fleuret. Learning interpolations between boltzmann densities. *arXiv preprint arXiv:2301.07388*, 2023.

Joseph Mathews and Scott C Schmidler. Finite sample complexity of sequential monte carlo estimators on multimodal target distributions. *The Annals of Applied Probability*, 34(1B):1199–1223, 2024.

Radford M Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Felix Otto and Maria G Reznikoff. A new criterion for the logarithmic sobolev inequality and two applications. *Journal of Functional Analysis*, 243(1):121–157, 2007.

Daniel Paulin, Ajay Jasra, and Alexandre H. Thiery. Error bounds for sequential Monte Carlo samplers for multimodal distributions. *The Annals of Applied Probability*, 28(3):1495–1535, 2018. doi: 10.1214/17-AAP1323.

Gareth O. Roberts, Jeffrey S. Rosenthal, and Nicholas G. Tawn. Skew brownian motion and complexity of the alps algorithm. *Journal of Applied Probability*, 59(3):777–796, 2022. doi: 10.1017/jpr.2021.78.

Daniel Štefankovič, Santosh Vempala, and Eric Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. *Journal of the ACM (JACM)*, 56(3):1–36, 2009.

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

Nicholas Tawn, Gareth Roberts, and Jeffrey Rosenthal. Weight-preserving simulated tempering. *Statistics and Computing*, 30, 02 2020a. doi: 10.1007/s11222-019-09863-3.

Nicholas G Tawn, Gareth O Roberts, and Jeffrey S Rosenthal. Weight-preserving simulated tempering. *Statistics and Computing*, 30(1):27–41, 2020b.

Nicholas G Tawn, Matthew T Moores, and Gareth O Roberts. Annealed leap-point sampler for multimodal target distributions. *arXiv preprint arXiv:2112.12908*, 2021.

Erik H Thiede, Brian Van Koten, Jonathan Weare, and Aaron R Dinner. Eigenvector method for umbrella sampling enables error analysis. *The Journal of chemical physics*, 145(8), 2016.

Håkon Tjelmeland and Bjørn Kåre Hegstad. Mode jumping proposals in mcmc. *Scandinavian Journal of Statistics*, 28(1):205–223, 2001. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616652.

Glenn M Torrie and John P Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of computational physics*, 23(2):187–199, 1977.

Francisco Vargas, Shreyas Padhy, Denis Blessing, and Nikolas Nüsken. Transport meets variational inference: Controlled monte carlo diffusions. *arXiv preprint arXiv:2307.01050*, 2023.

Fugao Wang and David P Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050, 2001.

Dawn Woodard, Scott Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19, 06 2009a. doi: 10.1214/08-AAP555.

Dawn B Woodard, Scott C Schmidler, and Mark Huber. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability [electronic only]*, 14:780–804, 2009b.