Proceedings of Machine Learning Research vol 291:1–63, 2025 38th Annual Conference on Learning Theory

Efficiently learning and sampling multimodal distributions with data-based initialization

Frederic Koehler FKOEHLER@UCHICAGO.EDU
University of Chicago

Holden Lee HLEE283@JHU.EDU
John Hopkins University

Thuy-Duong Vuong TDVUONG@BERKELEY.EDU
UC Berkeley

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider the problem of sampling a multimodal distribution with a Markov chain given a small number of samples from the stationary measure. Although mixing can be arbitrarily slow, we show that if the Markov chain has a kth order spectral gap, initialization from a set of $\tilde{O}(k/\varepsilon^2)$ samples from the stationary distribution will, with high probability over the samples, efficiently generate a sample whose conditional law is ε-close in TV distance to the stationary measure. In particular, this applies to mixtures of k distributions satisfying a Poincaré inequality, with faster convergence when they satisfy a log-Sobolev inequality. Our bounds are stable to perturbations to the Markov chain, and in particular work for Langevin diffusion over $\mathbb{R}^d$ with score estimation error, as well as Glauber dynamics combined with approximation error from pseudolikelihood estimation. This justifies the success of data-based initialization for score matching methods despite slow mixing for the data distribution, and improves and generalizes the results of Koehler and Vuong (2023) to have linear, rather than exponential, dependence on k and apply to arbitrary semigroups. As a consequence of our results, we show for the first time that a natural class of low-complexity Ising measures can be efficiently learned from samples.

Keywords: Learning, sampling, Markov chain, multimodal distribution, Ising model

1. Introduction

Since its introduction in 1953 by Metropolis et al. (1953), Markov-Chain Monte Carlo (MCMC) has become the one of the dominant approaches to sampling and integration of high-dimensional distributions in Bayesian statistics, computational physics, biostatistics, astronomy, machine learning, and many other areas. Typically, in MCMC we sample from a distribution of interest by simulating a Markov chain which converges to the correct stationary measure.

One of the key mathematical questions concerning a Markov chain is its *mixing time*—how quickly does the process forget its initialization and converge to stationarity? While some Markov chains rapidly mix to their stationary distributions, it is well-known that in many other cases, the existence of "bottlenecks" (i.e., sparse cuts) between modes leads to slow or "torpid" mixing, oftentimes exponential in the dimension of the problem. In general, whenever the distribution of interest is supported on multiple well-separated clusters or modes, standard MCMC methods like Metropolis-Hastings, Langevin dyamics, Glauber dynamics, and so on which make "local" moves will get stuck in the first cluster they reach. This phenomena is often referred to as metastability in the literature (Gayrard et al., 2004, 2005).

A large body of research in MCMC, both in theory and practice, is on developing ways to overcome this difficulty. For example, popular methods such as simulated tempering (Marinari and Parisi, 1992) and parallel tempering (Swendsen and Wang, 1986) attempt to improve connectivity by varying the temperature of the system. In other situations, alternative Markov chains can be constructed which are able to cross between the modes — for example, the celebrated Swendsen-Wang dynamics (Swendsen and Wang, 1987) which are provably able to sample from the ferromagnetic Ising model at all temperatures in polynomial time (Jerrum and Sinclair, 1993; Guo and Jerrum, 2017). In other cases, the sampling problem is provably hard (see e.g. Sly and Sun (2012); Galanis et al. (2016)) so no computationally efficient Markov chain could possibly mix rapidly to the stationary measure unless $\mathsf{P} = \mathsf{NP}$.

Hence, sampling can be computationally intractable in general. However, many cases where standard Markov chains are known to fail actually correspond to very simple distributions. For example, even a simple mixture of two well-separated Gaussians leads to exponentially large mixing time for the Langevin dynamics (in the separation distance). Previous work in MCMC theory has studied in depth some of these failure cases (e.g. in the Curie-Weiss model, see related work below) and developed specialized solutions to resolve the mixing time issue in a particular setting. We might hope that a more general approach can resolve the difficulty with multimodality for a large class of models.

Our contribution. In this work, we develop general tools to analyze MCMC chains in multimodal situations. Analogous to the role of the spectral gap in the unimodal setting, the key mathematical object in our theory is the *higher-order spectral gap* of the transition matrix or generator of a Markov process. Looking at such a notion of gap is very natural from the perspective of higher-order analogues of Cheeger's inequality (Lee et al., 2014; Louis et al., 2012; Gharan and Trevisan, 2014; Miclo, 2014). These results roughly tell us that if the vertices of a graph can be separated into a small number of well-connected parts whose boundaries are sparse cuts, then there has to be a corresponding gap in the spectrum of the Laplacian after a small number of eigenvalues, and vice versa.

When we have only a higher-order, rather than standard, spectral gap, we cannot hope for rapid mixing of the dynamics from an arbitrary initialization. However, in some applications there is a natural candidate for a warm start for the dynamics. In particular, in the application of density estimation or *generative modeling*, a distribution is learned from access to samples from the ground truth distribution. In these settings, samples from the ground truth are available, which naturally suggests the idea of *data-based initialization* — starting the dynamics from the empirical measure. The idea of data-based initialization has appeared in the empirical machine learning literature in many different forms, for example as a part of the mechanics of "contrastive divergence" training for energy-based methods and other approximations to Maximum Likelihood Estimation. For a few related references in the empirical literature, see Hinton (2002); Xie et al. (2016); Gao et al. (2018); Nijkamp et al. (2019, 2020); Wenliang et al. (2019), and see also Koehler and Vuong (2023) for more discussion. In particular, the terminology of "data-based initialization" is as used in Nijkamp et al. (2020).

Intuitively, if the underlying distribution is a mixture distribution, then each mixture component will have a roughly proportional representation in the samples from the empirical measure, so we might hope that the dynamics run for a polynomial amount of time can actually recover the ground truth distribution. A direct analysis along these lines was done in Koehler and Vuong (2023) in the case of a mixture of strongly log-concave distributions; however, handling the behavior of overlapping clusters complicated the analysis and ultimately led to a poor (exponential) dependence on the number of clusters or mixture components in the distribution.

From here on, we revisit the analysis of data-based initialization from the spectral perspective. This yields a much more elegant proof with dramatically improved quantitative dependencies. Our approach applies to general Markov semigroups, and in particular lets us prove new results for the Glauber dynamics as well as the Langevin dynamics. We also can easily obtain natural extensions of our results, such as more rapid mixing under a component-wise log-Sobolev inequality. Crucially for density estimation applications, we also show that our results are robust to perturbations in the Markov chain, which is very important when the chain transitions are themselves estimated from data. The quantitative improvements in this theory are key to an illustrative new application—a new result for learning a class of Ising models well beyond the regime where previous approaches were known to succeed.

1.1. Main results

The heart of this work in the following theorem which holds in a very general setting—it applies to all Markov semigroups, and shows that a higher-order spectral gap implies rapid mixing from data-based initialization.

Theorem 1 (Theorem 9, simplified) *Let $P_t = e^{t\mathscr{L}}$ be a reversible Markov semigroup with self-adjoint generator $\mathscr{L}$ and stationary measure π defined over $\mathcal{D}$. Suppose that the generator satisfies $\lambda_{k+1}(-\mathscr{L}) \geq \alpha$ and there are constants t_0, R such that*

$$(\textit{warm start after time } t_0) \quad \forall y \in \mathcal{D}, \quad \chi^2(\delta_y P_{t_0} \| \pi) \leq R.$$

Let $\mu_0 = \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$ where $Y_1, \dots, Y_n \sim \pi$ are independent samples, and define $\mu_t = \mu_0 P_t$. Then with probability $\geq 1-\delta$, for $n = \Omega\left(\frac{k}{\varepsilon_{\mathrm{TV}}^2}\ln\left(\frac{k}{\delta}\right)\right)$ and $t \geq t_0 + \frac{1}{\alpha}\ln\left(\frac{4R}{\varepsilon_{\mathrm{TV}}^2}\right)$, we have

$$\mathrm{TV}(\mu_t, \pi) \leq \varepsilon_{\mathrm{TV}}.$$

In fact, we only need a weaker version of the warm-start condition, which is important in some applications—see Theorem 9 and Theorem 16 for the more general result. A higher-order spectral gap is easy to show for Glauber or Langevin dynamics if our distribution is a mixture of well-connected components—see Lemma 10.

Langevin dynamics on mixtures. Our first application is to Langevin dynamics on mixture distributions. In this application, we include a perturbation analysis which shows that an L_2 approximate score function suffices for sampling—this is important if the score function is learned from data via score matching (see e.g. Hyvärinen (2005)).

Theorem 2 (Langevin with score matching, Theorem 26, Theorem 30, simplified) *Assume $\pi = \sum_{i=1}^k p_i \pi_i$, where π_i are $O(1)$-smooth and the means of the π_i are at distance $\lesssim \sqrt{d}$. Suppose that the approximate score function s satisfies*

$$\mathbb{E}_\pi \left\| s - \nabla \ln \pi \right\|^2 \leq \varepsilon_{\mathrm{score}}^2.$$

Let $\mu_0 = \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$, *where* $Y_1, \dots, Y_n \sim \pi$ *are independent samples, and suppose* $n = \Omega\left(\frac{k\ln(k/\delta)}{\varepsilon_{\mathrm{TV}}^2}\right)$ *for a parameter* $\delta > 0$. *Let* $(\bar{X}_t)_t$ *be the continuous Langevin diffusion wrt* π *initialized at* μ_0 *and driven by* s*, so it satisfies the SDE*

$$d\bar{X}_t = s(\bar{X}_t)dt + \sqrt{2}\,dW_t$$

for an independent Brownian motion W_t*, and* $\bar{X}_0 \sim \mu_0$. *Let* μ_t *be the law of* $\bar{X}_t$ *conditioned on the empirical samples* $Y_1, \dots, Y_n$*, i.e.,* $\mu_t = \mathcal{L}(\bar{X}_t | Y_1, \dots, Y_n)$.

Suppose that either:

1. *Each* π_i *satisfies a Poincaré inequality with constant* $\le \frac{1}{\alpha}$*, and* $T = \Omega\left(\frac{1}{\alpha}\left(d + \ln\left(\frac{k}{\varepsilon_{\mathrm{TV}}}\right)^2\right)\right)$.
2. *Each* π_i *satisfies a log-Sobolev inequality with constant* $\le \frac{1}{\alpha}$*, and* $T = \Omega\left(\frac{1}{\alpha}\ln\left(\frac{dk}{\varepsilon_{\mathrm{TV}}}\right)\right)$.

Then with probability at least $1 - \delta$ *over the randomness of* $Y_1, \dots, Y_n$*,*

$$\mathrm{TV}(\mu_T, \pi) \le \sqrt{T}\varepsilon_{\text{score}} + \varepsilon_{\mathrm{TV}}$$

The above result is stated for the continuous-time Langevin diffusion. We also prove a version of this result for its discrete-time analog, Langevin Monte Carlo in Theorem 31. We note that there are previous results that quantify mixing up to multimodality in the distribution (e.g., Balasubramanian et al. (2022) for convergence in Fisher information for averaged LMC, or Tzen et al. (2018); Zhang et al. (2017)); our innovation is to show that data-based initialization can lead to the much stronger condition of mixing. This is analogous to the difference between finding a stationary point vs. a global optimum in nonconvex optimization.

Remark 3 (Matching sample complexity lower bound) *It is a classical fact that* $\Theta(k/\varepsilon_{\mathrm{TV}}^2)$ *samples are needed to learn a distribution on the alphabet* $\{1, \dots, k\}$ *within total variation distance* $\varepsilon_{\mathrm{TV}}$ *(see e.g. Han et al. (2015) for references). This is a special case of our problem: it corresponds to the case where* π *is a mixture of* k *known components with disjoint support where only the mixing weights are unknown. (Note that the score function in this case does not depend on the mixing weights.) Thus, the dependence on* n *in both of the previous theorems is optimal up to the* $\log$ *factor in* k.

Remark 4 (Using score matching in Gaussian mixture models) *Recent works (Chen et al., 2024; Gatmiry et al., 2024) show how to learn a mixture of well-conditioned Gaussians from samples using a computationally efficient score matching approach. The key step is to show that the score function of such a distribution can be well-approximated using a piecewise-polynomial function, which can be efficiently estimated from data. Because mixtures of Gaussians are closed under convolution with noise, this implies that they can use the learned score functions at different noise levels to approximately sample via a denoising diffusion process (see e.g. Benton et al. (2023); Chen et al. (2023a) and references therein). Since they show that the score function can be accurately estimated from samples (at least for a slightly noised version of the distribution, which is close in TV—see Proposition 2.1 and the proof of Theorem 4.1 in Chen et al. (2023a)), their score function estimate could be combined with our method (data-based initialization of the vanilla Langevin dynamics) to give an alternative and arguably simpler algorithm.*

Glauber dynamics and an application to learning. A similar result to Theorem 30 holds for Glauber dynamics (see Lemma 20)—in this context, we also show that the dynamics are robust to a small error in the KL divergence, which would occur if we estimate the dynamics via pseudolikelihood estimation (Besag, 1975). Pseudolikelihood is a very classical and popular method, and the exact analogue of score matching for Glauber—see Hyvärinen (2007); Koehler et al. (2022a) for more discussion.

As a concrete end-to-end application of the result for Glauber dynamics, we prove a new theorem about learning a large class of Ising models: those which are in some sense low complexity or approximately low rank. This class of models has been extensively studied in probability theory due to its connection to mean-field approximation—see the discussion of related work below. As we discuss therein, this class of models is well outside of the realm where previous learning results (e.g. Wu et al. (2019); Gaitonde and Mossel (2024)) can be applied.

Theorem 5 (Learning approximate low-rank Ising models, Theorem 40, simplified) *Suppose π is an Ising model, i.e. a probability measure on $\{\pm 1\}^n$ satisfying*

$$\pi(x) \propto \exp\left(\frac{1}{2}\langle x, Jx\rangle + \langle h, x\rangle\right)$$

for some symmetric interaction matrix $J \in \mathbb{R}^{n\times n}$ and external field vector $h \in \mathbb{R}^n$.

Suppose J has eigenvalues (ordered from the largest)

$$\lambda_1 \geq \cdots \lambda_r > 1 - \frac{1}{c} \geq \lambda_{r+1} \geq \cdots \geq \lambda_n,$$

for a constant $c > 1$ and $-\sum_{j:\lambda_j<0} \lambda_j = O(1)$. Given $(nr\lambda_1)^{O(r)}/\varepsilon_{\mathrm{TV}}^4$ samples from $\pi_{J,h}$, with high probability over the samples, the distribution produced by pseudolikelihood estimation and Glauber dynamics from data-based initialization is within TV error $\varepsilon_{\mathrm{TV}}$ of $\pi_{J,h}$.

See the full statement of the theorem for more details and the precise definition of pseudolikelihood estimation.

1.2. Other related work

Learning Markov random fields from samples. There have been too many works on learning Markov random fields from samples to give an exhaustive list, so we instead summarize some of the most recent and directly relevant works. Information-theoretically, it is known that density estimation of Ising models in TV distance can be done with polynomial dependence on the dimension n and target accuracy ϵ (Devroye et al., 2020). However, all existing results for learning Ising models with computationally efficient algorithms, which typically use some variant of pseudolikelihood estimation and include works such as Ravikumar et al. (2010); Lokhov et al. (2018); Bresler et al. (2017); Klivans and Meka (2017); Wu et al. (2019); Gaitonde and Mossel (2024), either restrict the model to be in a high-temperature regime, or have an exponential dependence on some parameter in the sample complexity—typically the maximum ℓ_1-norm of any row of the interaction matrix. In many examples this ℓ_1-norm is polynomial in the dimension (see e.g. Anari et al. (2024a); Gaitonde and Mossel (2024) for more discussion), and it is always linear in the "inverse temperature" β of the system. So in particular, all of these results have an exponential dependence on β. In contrast, our result has polynomial dependence on these parameters when the interaction matrix is (approximately) low rank. See Remark 41 for a much more detailed discussion of the limitations of previous techniques.

Tempering and annealing on multimodal distributions. The empirical community has developed many algorithms for sampling from multimodal distributions. This includes *tempering* methods such as simulated tempering (Marinari and Parisi, 1992) and parallel tempering (Swendsen and Wang, 1986), which involve constructing a Markov chain which varies the temperature of the system, as well as *annealing* or *sequential* methods such as sequential Monte Carlo (Liu and Chen, 1998) and annealed importance sampling (Neal, 2001), which vary the temperature unidirectionally over time.

Efficient theoretical guarantees are known only in special cases; a necessary condition for all known results is that the distribution is decomposable into parts which are not too imbalanced (do not have a bottleneck) between different temperatures. For simulated tempering, these results can be used to show that simulated tempering with Langevin dynamics can sample from an isotropic mixture of Gaussians (or more generally, translates of a fixed log-concave distribution) (Ge et al., 2018a,b). A version of simulated tempering was also used in Koehler et al. (2022b) to sample from multimodal Ising models. There are analogous results for sampling from multimodal distributions under stronger assumptions for parallel tempering (Woodard et al., 2009; Lee and Shen, 2023) and sequential Monte Carlo (Schweizer, 2012; Paulin et al., 2018; Mathews and Schmidler, 2022; Lee and Santana-Gijzen, 2024).

However, even simple multimodal distributions can violate the condition of balance between temperatures and cause sampling to be provably hard. In particular, (Ge et al., 2018b, Appendix F) show an exponential query complexity lower bound for sampling from a L^∞-perturbed mixture of two Gaussians with different covariances. Here, the key difficulty in their setting is finding the components, whereas our results apply without issue. Simulated tempering also fails for the mean-field Potts models (Bhatnagar and Randall, 2004) (whereas our method will succeed, see Theorem 39).

The problem of sampling from multimodal distributions given *warm starts* to the different modes has also been considered, for which the Annealed Leap Point Sampler (Tawn et al., 2021; Roberts et al., 2022) has been proposed. We note this is a weaker notion of "advice" than ours, and guarantees have only been given in a limit where the components are approximately Gaussian.

Glauber dynamics in multimodal/metastable settings. There has been a lot of work on understanding the behavior of Glauber dynamics with respect to metastability in spin systems, especially for complete graph models, random graphs, and the square lattice. See e.g. Gheissari and Sinclair (2022); Blanca et al. (2024); Bovier et al. (2021); Cuff et al. (2012); Levin et al. (2010); Ding et al. (2009a,b); Galanis et al. (2024) for rigorous results. For the most part, this literature focuses on settings where there are a small number of metastable states which can be explicitly characterized, often taking advantage of symmetry considerations. For example, when the Ising model has no external field, it is symmetric under interchange of $+$ and $-$, so initializations like $(1/2)(\delta_{\mathbf{1}} + \delta_{-\mathbf{1}})$ are natural and have been studied in some of the aforementioned works. Generally speaking, our results hold in a general setting and do not rely on any structure of the underlying distribution besides the higher-order eigenvalue gap, but unlike those works we do not obtain explicit characterizations of the metastable states. In Section F, we show how to combine our general technique with results from this literature in the case of the Curie-Weiss model, which lets us make more precise statements about the spectrum of the generator and mixing from non-data-based initialization.

Learning from dynamics. There have also been recent works on learning a distribution *from* the Glauber dynamics, rather than from i.i.d. samples. For example, see Bresler et al. (2017); Gaitonde et al. (2024). The work (Jayakumar et al., 2024) is in spirit closely related—they study learning the Ising model when given i.i.d. samples from a metastable state (i.e. a region where Glauber dynamics becomes trapped for a long time). These results do not have new implications for the i.i.d. setting we study, but combining the ideas from our work with this setting may be an interesting direction for future work.

Low-complexity Ising models. The class of Ising models which are close to being low rank is significant, because informally these are the models for which the "naive mean-field approximation" from statistical physics is appropriate. (Naive mean field is, generally speaking, the most common type of variational approximation when doing variational inference in practice. See e.g. Wainwright et al. (2008) for some background.) One significance of this class of models is that it includes models in all of the high temperature, critical temperature, and low temperature regimes, and in particular many settings where the mixing time of natural Markov chains are exponential in the dimension due to multimodality or metastability. There have been many works in probability theory studying the structural properties of this class of models and generalizations—see e.g. Basak and Mukherjee (2017); Eldan (2018); Eldan and Gross (2018); Jain et al. (2019); Austin (2019); Augeri (2021).

As in Koehler et al. (2022b), the class of models we consider for learning are somewhat broader, in that we allow the bulk of the spectrum to have diameter at most 1 instead of requiring the bulk to be asymptotically negligible; this means the class also includes some models where naive mean field approximation is highly inaccurate (e.g. sparse models and spin glasses at high temperature).

Theory for score matching. There has been a lot of recent work on score matching, diffusion models, and related topics which we cannot exhaustively survey; instead, we mention a few relevant works. Denoising diffusion models are a popular approach to generative modeling which use approximations to the score function of the distribution convolved with different levels of Gaussian noise; recent works showed that these methods are robust to L_2-approximation of the score function (see e.g. Chen et al. (2023b,a); Benton et al. (2023) and references within). The method we study is different in that only the original ("vanilla") score function is needed; see Koehler and Vuong (2023) for more discussion. One of the motivations for score matching is that it can be easier to compute than the maximum likelihood estimator; for example, in some cases computing the MLE is NP-hard but vanilla score matching is statistically effective and computationally efficient (Pabbaraju et al., 2024; Hyvärinen, 2005, 2007). When the vanilla score function is estimated from data, it turns out that data-based initialization is not only needed for computational reasons—if Langevin dynamics on the estimated score function is run to stationarity, then in many multimodal settings the resulting distribution will be a poor estimate of the ground truth (Lee et al., 2022; Koehler et al., 2022a; Balasubramanian et al., 2022). If the distribution is unimodal, more specifically satisfies the log-Sobolev inequality, then the same work shows that vanilla score matching is statistically efficient even when Langevin is run to stationarity.

Concurrent work. During the preparation of this manuscript, we were made aware of independent and concurrent work (Huang et al., 2024) which also gives guarantees for sampling mixture distributions using Langevin with data-based initialization, with a different proof technique based on weak Poincaré inequalities. We note that their number of samples in the data-based initialization (see (Huang et al., 2024, Theorem 5.1)) has $\frac{1}{p_*}$ dependence on the minimum mixture weight p_*, while ours depends only on the number of components k.

1.3. Technical overview

The proof of our main result, Theorem 9, is conceptually simple. We aim to prove that the process initialized at the empirical samples contracts in χ^2-divergence, which corresponds to the contraction in $L^2(\pi)$-norm where π is the stationary distribution. When the Markov process has a higher order spectral gap, i.e., $\lambda_{k+1}(-\mathscr{L}) \geq \alpha$ where $\mathscr{L}$ is the generator of the Markov process and $\lambda_{k+1}(-\mathscr{L})$ is the $(k+1)$-th smallest eigenvalue, this contraction holds for any function ϕ which is orthogonal to the subspace V spanned by the eigenfunctions corresponding to the k smallest eigenvalues of $-\mathscr{L}$. The conclusion still holds if the projection of ϕ to V has a sufficiently small norm, a condition we name *eigenfunction balanced* (see Definition 12).

Hence, to prove rapid mixing from data-based initialization, we only need to show that the empirical distribution satisfies this condition. This is also the key technical challenge of our main result. For ease of presentation, we summarize our argument in the simpler finite-dimensional setting. We observe that when ϕ corresponds to the empirical distribution, the *expected* projection[1] is precisely zero, due to the orthogonality of the eigenfunctions/eigenvectors. Our task is thus to establish a strong (Chernoff-like) concentration bound for this projection norm. A key technical difficulty is that we do not have high moment bounds; we only have a second moment bound due to orthonormality. A second moment bound typically leads to a weaker concentration bound, where the number of samples n has linear dependency on the failure probability δ, instead of the expected $\log(1/\delta)$ dependency.

For concreteness, let π be the stationary distribution of the Markov process and $(f_i)_{i=1}^{\infty}$ be an orthonormal basis of eigenvectors of $-\mathscr{L}$ with eigenvalues $\lambda_1 = 0 \leq \lambda_2 \leq \lambda_3 \leq \cdots$. Then for $y \sim \pi$ and $i > 1$

$$\mathbb{E}_{y\sim\pi}[\langle \frac{\mathrm{d}\delta_y}{\mathrm{d}\pi}, f_i\rangle_{L^2(\pi)}] = \mathbb{E}_{y\sim\pi}[f_i(y)] = \langle f_i, f_1\rangle_{L^2(\pi)} = 0$$

since $f_1 \equiv 1$. We can also bound the second moment of the projection by

$$\mathbb{E}_{y\sim\pi}[\langle \frac{\mathrm{d}\delta_y}{\mathrm{d}\pi}, f_i\rangle^2_{L^2(\pi)}] = \mathbb{E}_{y\sim\pi}[f_i^2(y)] = \langle f_i, f_i\rangle_{L^2(\pi)} = 1.$$

More generally, for $\phi = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ where $y_1, \cdots, y_n \sim \pi$ are i.i.d.,

$$\mathbb{E}_{y_1,\ldots,y_n\sim\pi \text{ i.i.d}}[\langle \frac{\mathrm{d}\phi}{\mathrm{d}\pi}, f_i\rangle_{L^2(\pi)}] = 0 \quad \text{and} \quad \mathbb{E}_{y_1,\ldots,y_n\sim\pi \text{ i.i.d}}[\langle \frac{\mathrm{d}\phi}{\mathrm{d}\pi}, f_i\rangle^2_{L^2(\pi)}] = 1/n.$$

A naive application of standard concentration inequality such as Chebyshev's inequality gives us the following

$$\mathbb{P}_{y_1,\cdots,y_n\sim\pi \text{ i.i.d.}}[|\langle \frac{\mathrm{d}\phi}{\mathrm{d}\pi}, f_i\rangle_{L^2(\pi)}| \geq \varepsilon_{\mathrm{TV}}] \leq \delta$$

when $n \geq \frac{1}{\varepsilon_{\mathrm{TV}}^2 \delta}$, which has sub-optimal dependency on δ.

We obtain an exponential improvement on the δ-dependency by restricting the domain to those with bounded projection norm. We then show a strong concentration bound on this restricted domain using standard concentration inequalities for bounded random variables (Bernstein's inequality). This implies that the process initialized at samples from the restricted domain is rapidly mixing.

1. taken over the randomness of the empirical dataset

By an appropriate choice of parameters, we can ensure that most of the samples will be from the restricted domain, and thus the result followed from a standard comparison argument.

An important case satisfying the higher-order spectral gap is when the stationary distribution π admits a decomposition into a *mixture* of distributions each satisfying a Poincaré inequality (see Lemma 10). For this case, our mixing time bound matches the state-of-the-art worst-case-start mixing for a single distribution satisfying the Poincaré inequality (see Theorem 26(1)). If we further assume that each component of the mixture satisfies the stronger log-Sobolev inequality, then using the hyper-contractivity argument from Lee and Santana-Gijzen (2024), we can obtain a tighter bound on the mixing time. For the continuous Langevin diffusion on mixtures of distributions satisfying log-Sobolev inequality, our mixing time bound has an optimal dependency on all parameters (see Theorem 30).

Markov chain perturbation. In many applications, we suffer errors when implementing the Markov process. For example, we can only implement a discretization of the continuous Langevin diffusion, i.e., the Langevin Monte-Carlo. Furthermore, in learning problems, we might not have access to the chain transition probabilities but need to estimate them from data, e.g., using score matching or pseudolikelihood estimation. We show that our analysis is robust to such perturbations.

In particular, we establish that data-based initialization is a natural and elegant way to exploit the guarantees on transition probabilities estimators learned from data, which is typically of the form $\mathbb{E}_{x\sim\pi}[\mathrm{dist}(\hat{u}(x), u(x))] \le \varepsilon_{\text{score}}$ where $\hat{u}, u$ stand for the estimated and the true transition probabilities, respectively. A priori, it is unclear how to use such a guarantee in the initial stage of the Markov process when the distribution is very far from the stationary distribution π. Even when π has rapid convergence to stationarity from *worst-case* start, most previous works (Lee et al., 2022) can only handle such perturbation by assuming that the process is initialized at a a distribution ν s.t. $\chi^2(\nu||\mu)$ is small. However, the empirical distribution *does not* satisfy this condition. For example, if π a continuous distribution and ν is the empirical distribution then $\chi^2(\nu||\pi) = +\infty$ since ν is discrete. If π is the uniform distribution over the hypercube $\{\pm 1\}^d$ and ν is the empirical distribution formed by $n = o(2^d)$ samples, then $\chi^2(\nu||\pi) = \Omega(2^d)$. Nevertheless, the empirical distribution is closely related to π in an average sense, and Koehler and Vuong (2023) exploited this fact in their perturbation analysis; however, their analysis is lossy, incurring an extra $\mathrm{poly}(1/\varepsilon_{\mathrm{TV}})$ factor in the number of samples.

In Theorem 21, we directly reduce the perturbation analysis from processes initialized at data samples to the perturbation analysis when initialized at *the stationary distribution*. Unlike Koehler and Vuong (2023), we do not incur any loss in the number of samples (see Theorem 31). The key idea is to bound the expected TV distance between two processes X_t and $\tilde{X}_t$ when initialized at a data-sample $Y \sim \pi$ by the KL-divergence when initialized at the stationary distribution π by applying Jensen's inequality, Pinsker's inequality and the chain rule for KL-divergence.

$$\mathbb{E}_{Y\sim\pi}[\mathrm{TV}(\mathcal{L}((X_t^Y)_{0\le t\le T}), \mathcal{L}((\tilde{X}_t^Y)_{0\le t\le T}))] \le \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathcal{L}((X_t^\pi)_{0\le t\le T})\|\mathcal{L}((\tilde{X}_t^\pi)_{0\le t\le T}))}$$

We note that we can bound the second moment using a similar argument. We can then obtain a strong concentration bound using Bernstein's inequality and the fact that TV-distance is bounded by 1.

Finally, we perform perturbation analysis for processes initialized at the *stationary distribution*. For continuous processes such as the Langevin diffusion with smooth stationary distribution, such a bound follows from the Girsanov's theorem (see Lemma 18). In our application, the stationary distribution is not necessarily smooth, but is a mixture of smooth distributions satisfying a Poincaré inequality. In that case, we can establish quantitatively similar results using higher moment bounds implied by the Poincaré inequality on each component (see Lemma 19). For the Glauber dynamics, we derive a qualitatively similar result in Lemma 20 when the transition probabilities, i.e., the conditional marginals, are estimated using pseudo-likelihood.

Application to low-complexity Ising models. The learning result follows from our general theory provided we can: (1) prove such a distribution has a higher-order spectral gap, and (2) estimate the needed Glauber transitions from data. For (1), we do this using a two-step argument; first, we use techniques from Koehler et al. (2022b) and a result from Anari et al. (2024b) to prove that such an Ising model can be approximately decomposed into a small mixture of rapidly mixing Ising models. Because our approximate decomposition has a density which is within a constant factor of the true model, we can boost this to an exact mixture decomposition into a mixture of Poincaré distributions by using the robustness of spectral gap to small changes of measure. For problem (2) of learning the transitions from data, this is exactly the problem solved by pseudolikelihood estimation, which we can analyze using standard symmetrization techniques from statistical learning theory.

1.4. Organization

We cover the mathematical preliminaries in Section A. Section B is devoted to the proof of our main result, Theorem 9. In Section C, we show the robustness of Langevin and Glauber dynamics in our setting to perturbations in the dynamics. Section D covers the application of our theory to Langevin dynamics on mixtures, and Section E illustrates the application to Glauber dynamics for sampling and learning Ising models. Finally, in Section F we give some examples of non-sample initializations which also satisfy eigenfunction balance.

2. Overview of the Appendix

In this section, we highlight some key points in the mathematical analysis from the appendix. All of the content here is repeated in more detail in the appendix (so that the appendix is self-contained).

The following general theorem analyzes the performance of a generic Markov chain with a higher-order spectral gap, showing it will mix well when started from a small number of samples from the true distribution. A key assumption for the theorem is the guarantee that dynamics run for some initial time period t_0, started from a typical sample, reaches a distribution with bounded χ^2-divergence to the stationary measure π. Proving this assumption requires a different analysis for different Markov chains. We will later show how to apply this result in natural settings like mixtures of log-concave measures and in statistical physics models.

Theorem *[Theorem 9 of Appendix] Let $\mathscr{L}$ be the self-adjoint generator of a reversible Markov semigroup $P_t = e^{t\mathscr{L}}$ with stationary distribution π defined over $\mathcal{D}$. Suppose that $-\mathscr{L}$ satisfies the kth order spectral gap condition from Definition 8, $\lambda_{k+1}(-\mathscr{L}) \geq \alpha > 0$, where $k \geq 1$.*

For $y \in \mathcal{D}$ and $t \geq 0$, let $\rho_t^y = \delta_y P_t$ be the marginal law of the Markov chain generated by $\mathscr{L}$ initialized at δ_y. Let $y_1, \ldots, y_n$ be i.i.d. samples from π and let U_{sample} be the multiset of $y_1, \ldots, y_n$. Consider the Markov process generated by $\mathscr{L}$ initialized at $\mu_0 = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, and let $\mu_t = \mu_0 P_t = \frac{1}{n}\sum_{j=1}^n \rho_t^{y_j}$ be the marginal law of the process at time t conditional on U_{sample}.

Suppose there exists time $t_0 \ge 0$, parameter $R \ge 0$ and a set Ω_{bd} such that

$$\forall y \in \Omega_{\text{bd}} : \chi^2(\rho_{t_0}^y || \pi) \le R,$$

and $\pi(\Omega_{\text{bd}}^c) \le \frac{\varepsilon_{\text{TV}}^2}{16k}$.

Then for $t \ge t_0$, with probability $\ge 1 - k\exp\left(-\Omega\left(\frac{n\varepsilon_{\text{TV}}^2}{k}\right)\right)$,

$$\text{TV}(\mu_t, \pi) \le \sqrt{\frac{\varepsilon_{\text{TV}}^2}{4} + e^{-\alpha(t-t_0)}R} + \frac{\varepsilon_{\text{TV}}}{16k}.$$

In particular, for $n = \Omega\left(\frac{k}{\varepsilon_{\text{TV}}^2}\ln\left(\frac{k}{\delta}\right)\right)$ and

$$t \ge T := t_0 + \frac{1}{\alpha}\ln\left(\frac{2R}{\varepsilon_{\text{TV}}^2}\right),$$

with probability at least $1-\delta$ over the randomness of U_{sample},

$$\text{TV}(\mu_t, \pi) \le \varepsilon_{\text{TV}}.$$

2.1. Eigenvalue gap for mixtures

We now show that the assumption in Theorem 9 is satisfied in the case of mixtures of distributions satisfying a Poincaré inequality. More precisely, this holds for arbitrary semigroups as long as the Dirichlet form of the mixture dominates the average Dirichlet form of the components (Equation (2) below); this holds automatically for common semigroups like the Langevin and Glauber dynamics.

Lemma *[Eigenvalue gap for Markov chain on mixtures, Lemma 10 of Appendix] Let $\mathscr{L}$, and $\mathscr{L}_i, 1 \le i \le k$ be generators of Markov processes on a measurable space $(\mathcal{D}, \mathscr{F})$ with stationary distributions π and $\pi_i, 1 \le i \le k$, respectively. Suppose that there are weights $w_i > 0$ such that*

$$\pi = \sum_{i=1}^k w_i \pi_i \tag{1}$$

$$\forall f \in \text{Dom}(\mathscr{L}) \subseteq L_2(\pi), \quad \langle f, -\mathscr{L}f\rangle_\pi \ge \sum_{i=1}^k w_i \langle f, -\mathscr{L}_i f\rangle_{\pi_i}. \tag{2}$$

Suppose that $\lambda_2(-\mathscr{L}_i) \ge \alpha$ for each $1 \le i \le k$ (i.e., each π_i satisfies a Poincaré inequality with constant $\frac{1}{\alpha}$). Then $\lambda_{k+1}(-\mathscr{L}) \ge \alpha$.

In particular, (2) holds in the following cases.

1. *$\mathcal{D} = \mathbb{R}^k$ and $\mathscr{L}$, $\mathscr{L}_i, 1 \le i \le k$ are the generators of Langevin diffusion with stationary distributions π, $\pi_i, 1 \le i \le k$.*
2. *$\mathcal{D} = \bigotimes_{i=1}^n \Omega_i$ and $\mathscr{L}$, $\mathscr{L}_i, 1 \le i \le k$ are the generators of Glauber dynamics with stationary distributions π, $\pi_i, 1 \le i \le k$.*

If the generator does not have a discrete spectrum, the assumption and conclusion should be interpreted in the sense of Definition 8; see the discussion after the definition about our notational convention.

2.2. Example application

Pseudolikelihood estimator. It is helpful to first recall the definition of the pseudolikelihood estimator Besag (1975) in the case of the Ising model. Recall that in the Ising model $\pi(x) \propto \exp(\langle x, Jx\rangle/2 + \langle h, x\rangle)$, $\pi(X_i = x_i \mid X_{\sim i}) \propto \exp(J_{i,\sim i} \cdot X_{\sim i} x_i + h_i x_i)$ and taking into account the normalizing constant, we have $\pi(X_i = x_i \mid X_{\sim i}) = \frac{1}{1+\exp(-2J_{i,\sim i}\cdot X_{\sim i} x_i - 2h_i x_i)}$. Therefore

$$\ln \pi(X_i = x_i \mid X_{\sim i}) = -\ln\left(1 + \exp(-2J_{i,\sim i} \cdot X_{\sim i} x_i - 2h_i x_i)\right),$$

and this is what is known as a logistic regression model. (See e.g. McCullagh and Nelder (2019).) The pseudolikelihood of an outcome $x \in \{\pm 1\}^n$ under the model π is the sum of these conditional likelihooods over the choice of index i, i.e.

$$\sum_{i=1}^{n} \ln \pi(X_i = x_i \mid X_{\sim i} = x_{\sim i})$$

and the pseudolikelihood estimator is given by optimizing this objective averaged over a dataset.

Theorem *[Theorem 40 of Appendix] Fix $c \in [2, \infty)$. Suppose that π is a probability measure on $\{\pm 1\}^n$ satisfying*

$$\pi(x) \propto \exp\left((1/2)\langle x, Jx\rangle + \langle h, x\rangle\right)$$

for some symmetric interaction matrix J and external field vector h. Suppose that J has r eigenvalues greater than $1 - 1/c$, including its top eigenvalue λ_1, and the rest are at most $1 - 1/c$.

Suppose $X^{(0)}, \dots, X^{(m)} \sim \pi$ are i.i.d. samples; let $\hat{\mathbb{E}}$ denote the corresponding empirical expectation over the empirical measure $\hat{\pi}$, i.e. $\hat{\mathbb{E}}$ is the average over the sample set. Let $Y^{(0)}, \dots, Y^{(m)} \sim \pi$ be another set of iid samples at denote their empirical measure by $\hat{\pi}_2$. Furthermore, suppose that $R \geq 0$ is such that

$$\max_i \sum_j |J_{ij}| + |h_i| \leq R. \tag{3}$$

Define the constrained pseudolikelihood estimator by

$$\hat{\rho} = \arg\max_{\rho \in \mathcal{P}_R} \sum_{i=1}^{n} \hat{\mathbb{E}}_X \ln \rho(X \mid X_{\sim i})$$

where $\mathcal{P}_R$ is a set consisting of Ising models (i.e. measures on $\{\pm 1\}^n$ with quadratic log-likelihood) with parameters satisfying the convex constraint (24) (so in particular, it can be optimized in polynomial time).

With probability at least $1 - \delta$ over the randomness of the sample set, for any

$$t \geq T = \exp\left[O\left(\int_0^{T_0} c e^{cs \operatorname{Tr}(J_-)} ds\right)\right]\left(2Rn + \ln\left(\frac{4k\ln(k/\delta)}{m}\right)\right)$$

where $T_0 = 1 - 1/c - \lambda_{\min}(J)$ we have

$$\mathrm{TV}(\mathcal{L}(X_t^{\hat{\pi}_2} \mid \hat{\pi}, \hat{\pi}_2), \pi) \leq \sqrt{tRn\sqrt{\ln(2n/\delta)/m}} + 4\sqrt{k\ln(k/\delta)/m}.$$

where

$$k = O\left(\lambda_1 n + \sqrt{\lambda_1 n r} + \sqrt{\lambda_1 n r \ln(\sqrt{n})}\right)^{O(r)} = \tilde{O}(\lambda_1 n)^{O(r)}$$

and $X_t^{\hat{\pi}_2}$ is the output of the $\hat{\rho}$-Glauber diffusion initialized at $\hat{\pi}_2$ and run for time t.

References

Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Universality of spectral independence with applications to fast mixing in spin glasses. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5029–5056. SIAM, 2024a.

Nima Anari, Frederic Koehler, and Thuy-Duong Vuong. Trickle-down in localization schemes and applications. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1094–1105, 2024b.

Fanny Augeri. A transportation approach to the mean-field approximation. *Probability Theory and Related Fields*, 180(1):1–32, 2021.

Tim Austin. The structure of low-complexity gibbs measures on product spaces. 2019.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 2006.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 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*, pages 2896–2923. PMLR, 2022.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Anirban Basak and Sumit Mukherjee. Universality of the mean-field for the potts model. *Probability Theory and Related Fields*, 168:557–600, 2017.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

Nayantara Bhatnagar and Dana Randall. Torpid mixing of simulated tempering on the potts model. In *SODA*, volume 4, pages 478–487. Citeseer, 2004.

Antonio Blanca, Reza Gheissari, and Xusheng Zhang. Mean-field potts and random-cluster dynamics from high-entropy initializations. *arXiv preprint arXiv:2404.13014*, 2024.

Anton Bovier, Saeda Marello, and Elena Pulvirenti. Metastability for the dilute curie–weiss model with glauber dynamics. 2021.

Guy Bresler, David Gamarnik, and Devavrat Shah. Learning graphical models from the glauber dynamics. *IEEE Transactions on Information Theory*, 64(6):4072–4080, 2017.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023b.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Sinho Chewi, Murat A. Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of langevin monte carlo from poincaré to log-sobolev, 2021.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Paul Cuff, Jian Ding, Oren Louidor, Eyal Lubetzky, Yuval Peres, and Allan Sly. Glauber dynamics for the mean-field potts model. *Journal of Statistical Physics*, 149:432–477, 2012.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rates of normal and ising undirected graphical models. 2020.

Jian Ding, Eyal Lubetzky, and Yuval Peres. Censored glauber dynamics for the mean field ising model. *Journal of Statistical Physics*, 137:407–458, 2009a.

Jian Ding, Eyal Lubetzky, and Yuval Peres. The mixing time evolution of glauber dynamics for the mean-field ising model. *Communications in Mathematical Physics*, 289(2):725–764, 2009b.

Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. *Geometric and Functional Analysis*, 28(6):1548–1596, 2018.

Ronen Eldan and Renan Gross. Decomposition of mean-field gibbs distributions into product measures. 2018.

Richard S Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, 2007.

Jason Gaitonde and Elchanan Mossel. A unified approach to learning ising models: Beyond independence and bounded width. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 503–514, 2024.

Jason Gaitonde, Ankur Moitra, and Elchanan Mossel. Efficiently learning markov random fields from dynamics. *arXiv preprint arXiv:2409.05284*, 2024.

Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Inapproximability of the partition function for the antiferromagnetic ising and hard-core models. *Combinatorics, Probability and Computing*, 25(4):500–559, 2016.

Andreas Galanis, Leslie Ann Goldberg, and Paulina Smolarova. Planting and mcmc sampling from the potts model. *arXiv preprint arXiv:2410.14409*, 2024.

Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. Learning generative convnets via multi-grid modeling and sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9155–9164, 2018.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Véronique Gayrard, Anton Bovier, Michael Eckhoff, and Markus Klein. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Véronique Gayrard, Anton Bovier, and Markus Klein. Metastability in reversible diffusion processes ii: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

Rong Ge, Holden Lee, and Andrej Risteski. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 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.

Shayan Oveis Gharan and Luca Trevisan. Partitioning into expanders. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1256–1266. SIAM, 2014.

Reza Gheissari and Alistair Sinclair. Low-temperature ising dynamics with random initializations. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1445–1458, 2022.

Heng Guo and Mark Jerrum. Random cluster dynamics for the ising model is rapidly mixing. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1818–1827. SIAM, 2017.

Brian C Hall. *Quantum theory for mathematicians*, volume 267. Springer Science & Business Media, 2013.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2291–2295. IEEE, 2015.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X. Wu. Weak poincaré inequalities, simulated annealing, and sampling from spherical spin glasses. In *submission*, 2024.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Aapo Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on neural networks*, 18(5):1529–1531, 2007.

Vishesh Jain, Frederic Koehler, and Andrej Risteski. Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1226–1236, 2019.

Abhijith Jayakumar, Andrey Y Lokhov, Sidhant Misra, and Marc Vuffray. Discrete distributions are learnable from metastable samples. *arXiv preprint arXiv:2410.13800*, 2024.

Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

Frederic Koehler and Thuy-Duong Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2023.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*, 2022a.

Frederic Koehler, Holden Lee, and Andrej Risteski. Sampling approximately low-rank ising models: Mcmc meets variational methods, 2022b.

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

Holden Lee and Zeyu Shen. Improved bound for mixing time of parallel tempering. *arXiv preprint arXiv:2304.01303*, 2023.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

James R Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

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

David A Levin, Malwina J Luczak, and Yuval Peres. Glauber dynamics for the mean-field ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146: 223–265, 2010.

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

Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791, 2018.

Anand Louis, Prasad Raghavendra, Prasad Tetali, and Santosh Vempala. Many sparse cuts via higher eigenvalues. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 1131–1140, 2012.

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

Joseph Mathews and Scott C. Schmidler. Finite sample complexity of sequential monte carlo estimators on multimodal target distributions, 2022.

Peter McCullagh and John Nelder. *Generalized linear models*. Routledge, 2019.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Laurent Miclo. On hyperboundedness and spectrum of markov operators. *Inventiones mathematicae*, 200:311 – 343, 2014. URL https://api.semanticscholar.org/CorpusID:122909896.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

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

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 32, 2019.

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5272–5280, 2020.

Chirag Pabbaraju, Dhruv Rohatgi, Anish Prasad Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Daniel Paulin, Ajay Jasra, and Alexandre Thiery. Error bounds for sequential monte carlo samplers for multimodal distributions, 2018.

Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. 2010.

Michael Reed and Barry Simon. *Methods of modern mathematical physics: Functional analysis*, volume 1. Gulf Professional Publishing, 1980.

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.

Narayana P Santhanam and Martin J Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

Nikolaus Schweizer. Non-asymptotic error bounds for sequential mcmc methods in multimodal settings, 2012.

Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 361–369. IEEE, 2012.

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

Robert H Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in monte carlo simulations. *Physical review letters*, 58(2):86, 1987.

Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

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

Gerald Teschl. *Mathematical methods in quantum mechanics*, volume 157. American Mathematical Soc., 2014.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Belinda Tzen, Tengyuan Liang, and Maxim Raginsky. Local optimality and generalization guarantees for the langevin algorithm via empirical metastability. In *Conference On Learning Theory*, pages 857–875. PMLR, 2018.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Li Wenliang, Danica J Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746. PMLR, 2019.

Dawn B. Woodard, Scott C. Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Probab.*, 19(2):617–640, 2009. ISSN 1050-5164,2168-8737. doi: 10.1214/08-AAP555. URL https://doi.org/10.1214/08-AAP555.

Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. *Advances in Neural Information Processing Systems*, 32, 2019.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644. PMLR, 2016.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022. PMLR, 2017.

Appendix A. Preliminaries

For a vector-valued function f, let $f_{r+1:s}(y)$ define the vector $(f_{r+1}(y), \ldots, f_s(y)) \in \mathbb{R}^{s-r}$.

A.1. Functional inequalities

For nonnegative $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, define the entropy of f with respect to probability distribution π to be

$$\mathrm{Ent}_\pi[f] = \mathbb{E}_\pi[f \ln(f/\mathbb{E}_\pi[f])].$$

Functional inequalities for the Langevin diffusion. We say π satisfies a log-Sobolev inequality (LSI) with constant C_{LS} with respect to the Langevin semigroup if for all smooth functions f,

$$\mathrm{Ent}_\pi[f^2] \leq 2C_{\mathrm{LS}}\mathbb{E}_\pi[\|\nabla f\|^2]$$

and π satisfies a Poincaré inequality (PI) with constant C_{P} if

$$\mathrm{Var}_\pi[f] \leq C_{\mathrm{P}}\mathbb{E}_\pi[\|\nabla f\|^2].$$

The log-Sobolev inequality implies the Poincaré inequality: $C_{\mathrm{P}} \leq C_{\mathrm{LS}}$. Due to the Bakry-Émery criterion Bakry and Émery (2006), if π is α-strongly log-concave then π satisfies LSI with constant $C_{\mathrm{LS}} = 1/\alpha$.

LSI and PI are equivalent to statements about exponential ergodicity of the continuous-time Langevin diffusion, which is defined by the Stochastic Differential Equation (SDE)

$$d\bar{X}_t^\pi = \nabla \log \pi(\bar{X}_t^\pi)\, dt + \sqrt{2}\, dB_t. \tag{4}$$

Specifically, let π_t denote the law of the diffusion at time t initialized from π_0. Then a LSI is equivalent to the inequality

$$\mathrm{KL}(\pi_t \| \pi) \leq \exp(-2t/C_{\mathrm{LS}})\, \mathrm{KL}(\pi_0 \| \pi)$$

holding for an arbitrary initial distribution π_0. Similarly, a PI is equivalent to

$$\chi^2(\pi_t \| \pi) \leq \exp(-2t/C_{\mathrm{P}})\chi^2(\pi_0 \| \pi).$$

Here $\mathrm{KL}(P\|Q) = \mathbb{E}_P[\ln \frac{dP}{dQ}]$ is the Kullback-Leibler divergence and $\chi^2(P\|Q) = \mathbb{E}_Q[(\frac{dP}{dQ} - 1)^2]$ is the χ^2-divergence.

Markov semigroups and generators. Our results apply to general diffusions besides Langevin. In particular, they also have interesting and useful applications for the Glauber dynamics or Gibbs sampler, which resamples the coordinates of the random vector one at a time. In order to state results in the appropriate level of generality, we recall the definition of a Markov semigroup. See Bakry et al. (2014); Van Handel (2014) for more background.

A stochastic process $(X_t)_{t\geq 0}$ is a *time-homogenous Markov process* if for every $t \geq 0$, there exists a linear operator P_t such that for all $s \geq 0$ and bounded measurable functions f,

$$\mathbb{E}[f(X_{s+t}) \mid (X_r)_{r\leq s}] = (P_t f)(X_s).$$

We define the corresponding operator on measures μP_t or $P_t^* \mu$ as the distribution of X_{s+t} when $X_s \sim \mu$; this leads to the identity $\int \mu(P_t f) = \int (\mu P_t) f$ where f is bounded measurable and μ is a measure.

Suppose that the stochastic process has stationary measure π. Given such a process, we can define its infinitesimal generator as $\mathscr{L}$ such that $P_t = e^{t\mathscr{L}}$. We say P_t is *reversible* if $\mathscr{L}$ is a self-adjoint operator on $L_2(\pi)$ (this is discussed more later). The *Dirichlet form* is then defined as

$$\mathcal{E}(f, g) = -\langle f, \mathscr{L} g\rangle_{L_2(\pi)}.$$

The *spectral gap* of the generator is the maximal $\gamma \geq 0$ such that for all test functions f,

$$\gamma \operatorname{Var}(f) \leq \mathcal{E}(f, f).$$

To be consistent with our notation above, we call $1/\gamma$ the Poincaré constant; the above inequality is the general form of the Poincaré inequality for Markov semigroups. As before, the Poincaré constant characterizes ergodicity in χ^2-divergence.

For example, for the Langevin diffusion described above, the generator is

$$\mathscr{L} f = \langle \nabla \log \pi, \nabla f\rangle_\pi + \Delta f \tag{5}$$

where $\Delta f = \sum_i \partial_i^2 f$ is the usual Laplacian, and we can compute that $\mathcal{E}(f, f) = \mathbb{E}\|\nabla f\|^2$. Similarly, the semigroup for the Glauber dynamics with stationary measure π over a measure space $\bigotimes_{i=1}^n \Omega_i$ is generated by

$$\mathscr{L} = \sum_{i=1}^n (E_i - I)$$

where $E_i f = \mathbb{E}_{X\sim\pi}[f(X) \mid X_{\sim i}]$ for all $1 \leq i \leq n$, and $X_{\sim i}$ denotes all the coordinates of X except i. Note that one unit of continuous time for the Glauber dynamics corresponds to Poisson(n) discrete updates. The generator of the Glauber dynamics is a self-adjoint operator on $L_2(\pi)$ and the generator of the Langevin dynamics is essentially self-adjoint; we define these terms more precisely below.

A.2. Spectral theory

Because operators like the generator of the Langevin diffusion are (essentially) self-adjoint operators on infinite-dimensional spaces, we will need to use some terminology and tools from functional analysis to analyze them in full generality. Readers who are only interested in operators with discrete spectrum (e.g., those on finite-dimensional spaces, or for Langevin dynamics on distributions with sufficiently rapid tail decay—see Corollary 4.10.9 of Bakry et al. (2014)) can largely ignore this section.

In the general case the spectrum of any self-adjoint operator can be decomposed into two disjoint parts: a discrete spectrum (the only part present in the finite-dimensional case) and an essential spectrum, which are both defined precisely below. For example, the spectrum of the negative Laplacian in Euclidean space is $[0, \infty)$ so it is all essential spectrum. For a simple example in our context, the spectrum of the generator of the Langevin diffusion on the symmetric exponential density $\mu(x) = \frac{1}{2}e^{-|x|}$ is not discrete. (This is also true if the density is modified in a neighborhood of zero to be smooth. See Section 4.4.1 of Bakry et al. (2014) for discussion.) Nevertheless, this density admits a spectral gap in the appropriate sense, which we also define below.

First, we briefly review the most relevant facts from functional analysis, see e.g. Reed and Simon (1980); Bakry et al. (2014); Hall (2013); Teschl (2014) for details. For an operator A defined on $\mathrm{Dom}(A)$, we let A^* denote its adjoint. We say that operator A is *self-adjoint* if $A = A^*$ and $\mathrm{Dom}(A) = \mathrm{Dom}(A^*)$. An operator is said to be *essentially self-adjoint* if its closure is self-adjoint.

Theorem 6 (Special case of Corollary 3.2.2 of Bakry et al. (2014)) *Suppose that μ is a probability measure on $\mathbb{R}^n$ with a smooth density. Then the generator $\mathscr{L}$ of the corresponding Langevin diffusion, defined on the set of smooth compactly supported functions, is essentially self-adjoint.*

Theorem 7 (Spectral theorem, Theorem VIII.6 of Reed and Simon (1980)) *There is a one-to-one correspondence between self-adjoint operators A and projection-valued measures π on a Hilbert space H, given by*

$$A = \int \lambda \, d\pi_\lambda.$$

See also Chapter A.4 of Bakry et al. (2014). With the notation of the above theorem, we have the functional calculus $g(A) = \int g(\lambda)\, d\pi_\lambda$ and in particular

$$P_t f = e^{t\mathscr{L}} f = \int e^{t\lambda} f \, d\pi_\lambda;$$

see Chapter VIII.3 of Reed and Simon (1980) or Chapter A.4 of Bakry et al. (2014). This enables us to generalize results from the case of a completely discrete spectrum in a natural way.

Discrete and essential spectrum. The support of the projection-valued measure π is called the *spectrum* of the operator A and denoted $\sigma(A)$. Furthermore, the spectrum can be decomposed as a disjoint union

$$\sigma(A) = \sigma_d(A) \cup \sigma_{\text{ess}}(A)$$

where $\sigma_d(A)$ is called the *discrete spectrum*, and consists of isolated points in the spectrum which are also required to be eigenvalues of A with finite multiplicity, and $\sigma_{\text{ess}}(A) = \sigma(A) \setminus \sigma_d(A)$ is called the *essential spectrum*. Note that the discrete spectrum may be a proper subset of the point spectrum of A, i.e., of the full set of eigenvalues of A.

Example: finite-dimensional case. For example, in the finite-dimensional case where A is an operator on $\mathbb{R}^n$ the above formulation of the spectral theorem reduces to the fact that

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T$$

where λ_i is the ith eigenvalue, v_i is the ith eigenvector, and $v_i v_i^T$ is the (rank one) projection operator onto the span of v_i. The spectrum will just be the set of eigenvalues of A, which is necessarily discrete.

Higher-order spectral gap. For finite dimensional operators, the natural higher-order spectral gap assumption is that $\lambda_{k+1}(A) \geq \alpha$ for some $k \geq 1$ and $\alpha > 0$. We can now write down the general analogue for infinite-dimensional operators.

Definition 8 *We say a positive-semidefinite self-adjoint operator A has a spectral gap after eigenvalue k (or a kth order spectral gap) of size at least α for $k \geq 1$ and $\alpha > 0$ if*

$$\sigma_{\text{ess}}(A) \subset [\alpha, \infty)$$

and the projection $\pi([0,\alpha))$ has rank at most k, where π is the spectral measure. (In other words, the latter condition says there are at most k eigenvalues, counted with multiplicity, in $[0,\alpha)$.) We will also write this condition as $\lambda_{k+1}(A) \geq \alpha$.

The higher-order spectral gap defined in the above sense has the same min-max variational characterization as in the finite dimensional case: see Theorem 4.10 of Teschl (2014). For example, this means that the characterization of the spectral gap in terms of the Poincaré functional inequality holds in general.

Similar to Teschl (2014), we will generally use the following notational convention for infinite-dimensional operators where the spectrum is bounded below: we write that the spectrum of a self-adjoint operator A below the essential spectrum is $\lambda_1 \leq \lambda_2 \leq \cdots$, where this is a list first consisting of the ordered elements of the discrete spectrum below $\inf \sigma_{\text{ess}}(A)$, and then if the first list is finite, this is followed by $\inf \sigma_{\text{ess}}(A)$ repeated infinitely many times. This convention generally enables us to write the same statement for finite and infinite-dimensional settings.

Appendix B. Markov chains with data-based initialization: Mixing given warm start

The following general theorem analyzes the performance of a generic Markov chain with a higher-order spectral gap, showing it will mix well when started from a small number of samples from the true distribution. A key assumption for the theorem is the guarantee that dynamics run for some initial time period t_0, started from a typical sample, reaches a distribution with bounded χ^2-divergence to the stationary measure π. Proving this assumption requires a different analysis for different Markov chains. We will later show how to apply this result in natural settings like mixtures of log-concave measures and in statistical physics models.

Theorem 9 *Let $\mathscr{L}$ be the self-adjoint generator of a reversible Markov semigroup $P_t = e^{t\mathscr{L}}$ with stationary distribution π defined over $\mathcal{D}$. Suppose that $-\mathscr{L}$ satisfies the kth order spectral gap condition from Definition 8, $\lambda_{k+1}(-\mathscr{L}) \geq \alpha > 0$, where $k \geq 1$.*

For $y \in \mathcal{D}$ and $t \geq 0$, let $\rho_t^y = \delta_y P_t$ be the marginal law of the Markov chain generated by $\mathscr{L}$ initialized at δ_y. Let $y_1, \ldots, y_n$ be i.i.d. samples from π and let U_{sample} be the multiset of $y_1, \ldots, y_n$. Consider the Markov process generated by $\mathscr{L}$ initialized at $\mu_0 = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, and let $\mu_t = \mu_0 P_t = \frac{1}{n}\sum_{j=1}^n \rho_t^{y_j}$ be the marginal law of the process at time t conditional on U_{sample}.

Suppose there exists time $t_0 \geq 0$, parameter $R \geq 0$ and a set Ω_{bd} such that

$$\forall y \in \Omega_{\text{bd}} : \chi^2(\rho_{t_0}^y || \pi) \leq R,$$

and $\pi(\Omega_{\text{bd}}^c) \leq \frac{\varepsilon_{\text{TV}}^2}{16k}$.

Then for $t \geq t_0$, with probability $\geq 1 - k\exp\left(-\Omega\left(\frac{n\varepsilon_{\text{TV}}^2}{k}\right)\right)$,

$$\text{TV}(\mu_t, \pi) \leq \sqrt{\frac{\varepsilon_{\text{TV}}^2}{4} + e^{-\alpha(t-t_0)}R} + \frac{\varepsilon_{\text{TV}}}{16k}.$$

In particular, for $n = \Omega\left(\frac{k}{\varepsilon_{\mathrm{TV}}^2}\ln\left(\frac{k}{\delta}\right)\right)$ *and*

$$t \ge T := t_0 + \frac{1}{\alpha}\ln\left(\frac{2R}{\varepsilon_{\mathrm{TV}}^2}\right),$$

with probability at least $1-\delta$ *over the randomness of* U_{sample},

$$\mathrm{TV}(\mu_t, \pi) \le \varepsilon_{\mathrm{TV}}.$$

B.1. Eigenvalue gap for mixtures

We now show that the assumption in Theorem 9 is satisfied in the case of mixtures of distributions satisfying a Poincaré inequality. More precisely, this holds for arbitrary semigroups as long as the Dirichlet form of the mixture dominates the average Dirichlet form of the components (Equation (7) below); this holds automatically for common semigroups like the Langevin and Glauber dynamics.

Lemma 10 (Eigenvalue gap for Markov chain on mixtures) *Let* $\mathscr{L}$*, and* $\mathscr{L}_i, 1 \le i \le k$ *be generators of Markov processes on a measurable space* $(\mathcal{D}, \mathscr{F})$ *with stationary distributions* π *and* $\pi_i, 1 \le i \le k$*, respectively. Suppose that there are weights* $w_i > 0$ *such that*

$$\pi = \sum_{i=1}^{k} w_i \pi_i \tag{6}$$

$$\forall f \in \mathrm{Dom}(\mathscr{L}) \subseteq L_2(\pi), \quad \langle f, -\mathscr{L} f\rangle_\pi \ge \sum_{i=1}^{k} w_i \langle f, -\mathscr{L}_i f\rangle_{\pi_i}. \tag{7}$$

Suppose that $\lambda_2(-\mathscr{L}_i) \ge \alpha$ *for each* $1 \le i \le k$ *(i.e., each* π_i *satisfies a Poincaré inequality with constant* $\frac{1}{\alpha}$*). Then* $\lambda_{k+1}(-\mathscr{L}) \ge \alpha$*.*

In particular, (2) holds in the following cases.

1. $\mathcal{D} = \mathbb{R}^k$ *and* $\mathscr{L}$*,* $\mathscr{L}_i, 1 \le i \le k$ *are the generators of Langevin diffusion with stationary distributions* π*,* $\pi_i, 1 \le i \le k$*.*
2. $\mathcal{D} = \bigotimes_{i=1}^n \Omega_i$ *and* $\mathscr{L}$*,* $\mathscr{L}_i, 1 \le i \le k$ *are the generators of Glauber dynamics with stationary distributions* π*,* $\pi_i, 1 \le i \le k$*.*

If the generator does not have a discrete spectrum, the assumption and conclusion should be interpreted in the sense of Definition 8; see the discussion after the definition about our notational convention.

Proof This is a generalization of the proof for Langevin dynamics in (Ge et al., 2018a, Lemma 6.1)[2]. Let $V = \mathrm{span}\left\{\frac{d\pi_i}{d\pi} : 1 \le i \le k\right\}$, and take $f \in \mathrm{Dom}(\mathscr{L}) \subseteq L_2(\pi)$ such that $f \in V^\perp$ (with respect to $\langle \cdot, \cdot \rangle_\pi$). This means that

$$\mathbb{E}_{\pi_i} f = \mathbb{E}_\pi \frac{d\pi_i}{d\pi} f = 0 \text{ and } \mathbb{E}_\pi f = 0.$$

2. See the arxiv version http://www.arxiv.org/abs/1710.02736

Then by the law of total variance and the spectral gap assumption for each π_i,

$$\begin{aligned}\|f\|_\pi^2 = \mathrm{Var}_\pi(f) &= \sum_{i=1}^k w_i(\mathbb{E}_{\pi_i} f - \mathbb{E}_\pi f)^2 + \sum_{i=1}^k w_i \,\mathrm{Var}_{\pi_i}(f) \\ &\le 0 + \sum_{i=1}^k w_i \frac{1}{\alpha} \langle f, -\mathscr{L}_i f\rangle_{\pi_i} \\ &\le \frac{1}{\alpha} \langle f, -\mathscr{L} f\rangle_\pi .\end{aligned}$$

Since we proved this for all f orthogonal to a subspace of dimension k, the desired conclusion follows from the variational characterization of eigenvalues,

$$\lambda_{k+1}(-\mathscr{L}) = \max_{\substack{\text{subspace } S \subseteq L^2(\pi) \\ \dim S = k}} \min_{f \perp_\pi S} \frac{\langle f, -\mathscr{L} f\rangle_\pi}{\|f\|_\pi^2} \ge \alpha.$$

Finally, we note that for Langevin diffusion, (7) holds with equality:

$$\int_{\mathbb{R}^n} \|\nabla f\|^2 \, d\pi = \sum_{i=1}^k \int_{\mathbb{R}^n} w_i \|\nabla f\|^2 \, d\pi_i.$$

For Glauber dynamics, (7) holds as well—see Lee and Santana-Gijzen (2024) or Lemma 29 of Anari et al. (2024b). In short, it follows from the equation $\mathcal{E}(f, f) = \sum_{i=1}^n \mathbb{E}\,\mathrm{Var}(f(X) \mid X_{\sim i})$ for the Dirichlet form of the Glauber dynamics, combined with the law of total variance over the choice of mixture component. ■

Remark 11 *The comparison of Dirichlet forms, Equation (7), also holds for a Metropolis-Hastings chain with fixed proposal (Lee and Santana-Gijzen, 2024).*

B.2. Mixing from a balanced initialization

To prove Theorem 9, we use the following Lemma 13 which says that as long as a set of points $y_1, \dots, y_n$ satisfy a natural balance condition (9) defined in terms of the eigenfunctions of the generator, a higher-order spectral gap leads to rapid contraction of the χ^2-divergence along the diffusion. Then in Lemma 15, we show the natural balance condition is satisfied with high probability using matrix Bernstein.

Definition 12 *Let $\mathscr{L}$ be the self-adjoint generator of a reversible Markov semigroup on $\mathcal{D}$, and let $f_1 \equiv 1, f_2, \dots, f_k$ be the eigenfunctions corresponding to (discrete) eigenvalues $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_k$ of $-\mathscr{L}$. A distribution μ_0 on $\mathcal{D}$ is (k, ε)-**eigenfunction balanced** if*

$$\|\mathbb{E}_{Y \sim \mu_0}[f_{2:k}(Y)]\| \le \varepsilon.$$

Lemma 13 (Mixing under eigenfunction balance) *Let $\mathscr{L}$ denote the generator of a Markov semigroup $P_t = e^{t\mathscr{L}}$ with stationary distribution π defined over $\mathcal{D}$, and suppose that $-\mathscr{L}$ satisfies the kth order spectral gap condition from Definition 8, $\lambda_{k+1}(-\mathscr{L}) \ge \alpha > 0$.*

Consider an initial distribution μ_0 and define $\mu_t = \mu_0 P_t$. Let $\varepsilon > 0$ and $t_0 \geq 0$ be arbitrary such that $\chi^2(\mu_{t_0}||\pi) < \infty$ and

$$\|\mathbb{E}_{Y\sim\mu_0}\left[f_{2:k}(Y)\right]\| \leq \varepsilon. \tag{8}$$

Then for $t \geq t_0$,

$$\chi^2(\mu_t\|\pi) \leq \varepsilon^2 + e^{-\alpha(t-t_0)}\chi^2(\mu_{t_0}\|\pi).$$

In particular, for $t \geq t_0 + \frac{1}{\alpha}\ln\left(\frac{\chi^2(\mu_{t_0}||\pi)}{\varepsilon^2}\right)$, we have that $\chi^2(\mu_t\|\pi) \leq 2\varepsilon^2$. In particular, if $\mu_0 = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ for some $y_1, \ldots, y_n \in \mathcal{D}$ then Equation (8) becomes

$$\left\|\frac{1}{n}\sum_{j=1}^n f_{2:k}(y_j)\right\| \leq \varepsilon. \tag{9}$$

The mixing time depends on $\log\chi^2(\mu_{t_0}||\pi)$, which can be bounded by standard techniques if π is a discrete distribution (see Section E for an example application when π is supported on the discrete hypercube $\{\pm 1\}^d$). When π is continuous with unbounded support, $\log\chi^2(\mu_{t_0}||\pi)$ can be unbounded at $t_0 = 0$. Fortunately, when π is a mixture distribution where each component satisfying functional inequality, we can bound $\log\chi^2(\mu_{t_0}||\pi)$ for relatively small t_0 using concentration properties of the components (see Section D.1).

Proof First we prove the result in the special case of a discrete spectrum. Let $(f_i)_{i=1}^\infty$ be the eigenfunctions of $\mathscr{L}$ with eigenvalues $-\lambda_1 = 0 > -\lambda_2 \geq -\lambda_3 \geq \cdots$. Then for $t \geq t_0$, we have

$$\frac{d\delta_y P_t}{d\pi}(x) = \sum_{i=1}^\infty e^{-\lambda_i t} f_i(y) f_i(x).$$

Informally, this is because we would have $\frac{d\delta_y}{d\pi}(x) = \sum_{i=1}^\infty f_i(y)f_i(x)$, if we use the "calculation" $\left\langle \frac{d\delta_y}{d\pi}, f\right\rangle_\pi = f(y)$. More formally, this is because by functional calculus, $P_t = e^{-t\mathscr{L}} = \sum_{i=1}^\infty e^{-\lambda_i t} f_i f_i^*$ where $f_i^* g = \langle f, g\rangle_{L_2(\pi)}$, i.e. f_i^* is the $L_2(\pi)$-adjoint of f_i. Hence $\langle \frac{d\delta_y P_t}{d\pi}, g\rangle_{L_2(\pi)} = \delta_y P_t g = \sum_{i=1}^\infty e^{-\lambda_i t} f_i(y)\langle f_i, g\rangle_{L_2(\pi)}$ for all $g \in L_2(\pi)$, so as claimed $\frac{d\delta_y P_t}{d\pi}(x) = \sum_{i=1}^\infty e^{-\lambda_i t} f_i(y) f_i(x)$.

By linearity, we have

$$\frac{d\mu_0 P_t}{d\pi}(x) = \sum_{i=1}^\infty e^{-\lambda_i t}\mathbb{E}_{y\sim\mu_0}[f_i(y)]f_i(x) = \sum_{i=1}^\infty e^{-\lambda_i t} C_i f_i(x) = 1 + \sum_{i=2}^\infty e^{-\lambda_i t} C_i f_i(x)$$

where $C_i = \mathbb{E}_{y\sim\mu_0}[f_i(y)]$ and the final equality follows from $\lambda_1 = 0$ and $f_1 \equiv 1$.

We know that $\langle f_i, f_j\rangle_{L^2(\pi)} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$ thus

$$\chi^2(\mu_t\|\pi) = \left\|\frac{d\mu_t}{d\pi} - 1\right\|_{L^2(\pi)}^2 = \|g_t\|_{L^2(\pi)}^2 + \|h_t\|_{L^2(\pi)}^2.$$

where $g_t = \sum_{i=2}^{k} e^{-\lambda_i t} C_i f_i(x)$ and $h_t = \sum_{i=k+1}^{\infty} e^{-\lambda_i t} C_i f_i(x)$. Next, we bound

$$\begin{aligned} \|g_t\|_{L^2(\pi)}^2 &= \left\| \sum_{i=2}^{k} e^{-\lambda_i t} C_i f_i \right\|_{L^2(\pi)}^2 = \sum_{i=2}^{k} e^{-2\lambda_i t} \cdot \|C_i f_i\|_{L^2(\pi)} = \sum_{i=2}^{k} e^{-2\lambda_i t} \cdot |C_i|^2 \\ &\le \sum_{i=2}^{k} |C_i|^2 = \|\mathbb{E}_{y\sim\mu_0} f_{2:k}(y)\|^2 \le \varepsilon^2, \end{aligned} \tag{10}$$

where the first inequality follows from $\lambda_i \ge 0$, the second equality from orthogonality of the eigenvectors f_i, the third equality from $\|f_i\|_{L^2(\pi)} = 1$, and the fourth equality from $\mathbb{E}_{y\sim\mu_0} f_{2:k}(y) = [C_2 \dots C_k]^\intercal$. Further,

$$\begin{aligned} \|h_t\|_{L^2(\pi)}^2 = \sum_{i=k+1}^{\infty} e^{-2\lambda_i t} \cdot |C_i|^2 &\le e^{-\alpha(t-t_0)} \cdot \sum_{i=k+1}^{\infty} e^{-2\lambda_i t_0} \cdot |C_i|^2 \\ &= e^{-\alpha(t-t_0)} \|h_{t_0}\|_{L^2(\pi)}^2 \le e^{-\alpha(t-t_0)} \chi^2(\mu_{t_0} || \pi) \end{aligned}$$

where the inequality follows from $\lambda_i \ge \alpha$ for $i \ge k+1$. Thus

$$\chi^2(\mu_t \| \pi) \le \varepsilon^2 + e^{-\alpha(t-t_0)} \chi^2(\mu_{t_0} || \pi) \le 2\varepsilon^2$$

for $t \ge t_0 + \frac{1}{\alpha} \ln \left(\frac{\chi^2(\mu_{t_0} || \pi)}{\varepsilon^2} \right)$.

In the case of a general operator satisfying Definition 8, the argument works the same way except that we replace the tail sum $\sum_{i=k+1}^{\infty} e^{-\lambda_i t} f_i f_i^*$ in the representation of $e^{t\mathcal{L}}$ by the integral $\int e^{-t\lambda} d\pi_\lambda$ over the projection-valued measure π corresponding to $\mathcal{L}$. By using the functional calculus, we can still prove in the same way that $\|h_t\|_{L^2}^2 \le e^{-(t-t_0)/C} \|h_{t_0}\|_{L^2}^2$, which controls the contribution to the χ^2-divergence from the non-leading eigenvalues. ■

B.3. Ensuring balance

We will ensure the balance condition (9) when initialized at the empirical distribution using the following.

Theorem 14 (Matrix Bernstein, (Tropp et al., 2015, Theorem 6.1.1)) *Let S_k, $1 \le k \le n$ be independent random matrices with dimension $d_1 \times d_2$, with $\mathbb{E} S_k = 0$ and $\|S_k\| \le L$ for each k. Let $Z = \sum_{k=1}^{n} S_k$, and define*

$$v(Z) = \max \left\{ \left\| \sum_{k=1}^{n} \mathbb{E} S_k S_k^* \right\|, \left\| \sum_{k=1}^{n} \mathbb{E} S_k^* S_k \right\| \right\}.$$

Then

$$\Pr(\|Z\| \ge t) \le (d_1 + d_2) \exp\left(-\frac{t^2/2}{v(Z) + Lt/3} \right).$$

Lemma 15 (Eigenfunction balance with high probability) *Let $\mathscr{L}$ denote a Markov chain generator with stationary distribution $\pi : \mathcal{D} \to \mathbb{R}_{\geq 0}$. Let $(f_i)_{i=1}^{\infty}$ be the eigenfunctions of $\mathscr{L}$ with eigenvalues $-\lambda_1 = 0 > -\lambda_2 \geq -\lambda_3 \geq \cdots$. Fix $k \geq 2$ and recall that $f_{2:k}(y) = (f_2(y), \ldots, f_k(y)) \in \mathbb{R}^k$. Let $\varepsilon \in (0, \frac{1}{2}]$. Let*

$$\Omega = \{y \in \mathcal{D} : \|f_{2:k}(y)\| \leq \frac{\sqrt{k-1}}{\varepsilon}\}.$$

Then $\pi(\Omega^c) \leq \varepsilon^2$.

Let $y_1, \ldots, y_n \sim \pi$ be iid. Consider $\tilde{\Omega} \subseteq \Omega$ and $U = \{y_j : y_j \in \tilde{\Omega}\}$. If $\pi(\tilde{\Omega}^c) \leq 2\varepsilon^2$ then

$$\Pr\left(\left\|\frac{1}{|U|}\sum_{y_j \in U} f_{2:k}(y_j)\right\| \geq 4\varepsilon\sqrt{k}\right) \leq k\exp\left[-\Omega(n\varepsilon^2)\right].$$

Proof Since the eigenfunctions have norm 1, $\mathbb{E}_\pi \|f_{2:k}(y)\|^2 = \sum_{i=2}^k \mathbb{E}_{y\sim\pi}[f_i(y)^2] = \sum_{i=2}^k ||f_i||_\pi^2 = k-1$. By Markov's inequality,

$$\pi(\Omega^c) < \frac{\mathbb{E}_{y\sim\pi}\|f_{2:k}(y)\|^2}{(k-1)/\varepsilon^2} = \frac{k-1}{(k-1)/\varepsilon^2} = \varepsilon^2.$$

We now proceed to prove the main claim. Let $\pi' = \pi|_{\tilde{\Omega}} = \frac{\pi(\cdot\cap\tilde{\Omega})}{\pi(\tilde{\Omega})}$. We claim that $\|\mathbb{E}_{\pi'}[f_{2:k}]\| \leq \varepsilon\sqrt{8(k-1)}$. Indeed,

$$\begin{aligned}
\|\mathbb{E}_{\pi'}[f_{2:k}(y)]\| &= \frac{1}{\pi(\tilde{\Omega})}\left\|\int_{\tilde{\Omega}} f_{2:k}(y)\pi(y)dy\right\| = \frac{1}{\pi(\tilde{\Omega})}\left\|\int_{\tilde{\Omega}^c} f_{2:k}(y)\pi(y)dy\right\| \\
&\leq \frac{1}{\pi(\tilde{\Omega})}\sqrt{\int_{\tilde{\Omega}^c}\pi(y)dy}\cdot\sqrt{\int_{\tilde{\Omega}^c}\pi(y)\|f_{2:k}(y)\|^2 dy} \\
&\leq 2\sqrt{\pi(\tilde{\Omega}^c)}\cdot\sqrt{\mathbb{E}_\pi[\|f_{2:k}\|^2]} \leq \varepsilon\sqrt{8(k-1)} \leq \sqrt{2(k-1)} \qquad (11)
\end{aligned}$$

where the second equality is due to $\mathbb{E}_\pi[f_i] = 0$ (by the orthogonality of eigenfunctions), the first inequality follows from Cauchy-Schwarz, and the second from $\pi(\tilde{\Omega}) \geq 1/2$. To use Matrix Bernstein, we calculate

$$\begin{aligned}
\left\|\mathbb{E}_{\pi'}(f_{2:k} - \mathbb{E}_{\pi'}f_{2:k})(f_{2:k} - \mathbb{E}_{\pi'}f_{2:k})^\top\right\| &\leq \mathbb{E}_{\pi'}(f_{2:k} - \mathbb{E}_{\pi'}f_{2:k})^\top(f_{2:k} - \mathbb{E}_{\pi'}f_{2:k}) \\
&\leq \mathbb{E}_{\pi'}\|f_{2:k}\|^2 \leq \frac{1}{\pi(\tilde{\Omega})}\mathbb{E}_\pi\|f_{2:k}\|^2 \\
&\leq 2\mathbb{E}_\pi\|f_{2:k}\|^2 = 2(k-1) \leq 2k. \qquad (12)
\end{aligned}$$

Conditioned on $|U| = m$, $y_j \in U$ are independent draws from π'. Applying Theorem 14 to $f_{2:k}(y_j) - \mathbb{E}_{\pi'}f_{2:k}$ and using (12) gives us

$$\begin{aligned}
\Pr\left(\left\|\frac{1}{|U|}\sum_{y_j\in U}[f_{2:k}(y_j) - \mathbb{E}_{\pi'}f_{2:k}]\right\| \geq \varepsilon\sqrt{k}\,\middle|\,|U| = m\right) &\leq k\exp\left[-\Omega\left(\frac{m\varepsilon^2 k}{k + \frac{\sqrt{k}}{\varepsilon}\cdot\varepsilon\sqrt{k}}\right)\right] \\
&\leq k\exp\left[-\Omega(m\varepsilon^2)\right]
\end{aligned}$$

Using (11) ($\|\mathbb{E}_{\pi'}[f_{2:k}(y)]\| < \varepsilon\sqrt{8(k-1)}$), we obtain

$$\Pr\left(\left\|\frac{1}{|U|}\sum_{y_j\in|U|} f_{2:k}(y_j)\right\| \geq 4\varepsilon\sqrt{k}\,\middle|\,|U|=m\right) \leq k\exp\left[-\Omega(m\varepsilon^2)\right].$$

Finally, we note that by Hoeffding's inequality

$$\Pr\left(|U| \leq \frac{1}{4}n\right) \leq \exp\left(-\Omega(n)\right).$$

Taking a union bound and adjusting constants appropriately then gives the result. ■

B.4. Proof of main theorem

Proof [Proof of Theorem 9]

We prove the first statement. Let $\varepsilon = \frac{\varepsilon_{\text{TV}}}{8\sqrt{k}} < 1/2$. As in Lemma 15, define

$$\Omega = \{y \in \mathcal{D} : \|f_{2:k}(y)\| \leq \frac{\sqrt{2(k-1)}}{\varepsilon}\},$$

let $\tilde{\Omega} = \Omega \cap \Omega_{\text{bd}}$, and let $U = U_{\text{sample}} \cap \tilde{\Omega}$. By Lemma 15, $\pi(\Omega^c) \leq \varepsilon^2$, thus $\pi(\tilde{\Omega}^c) \leq 2\varepsilon^2 = \frac{\varepsilon_{\text{TV}}^2}{32k}$ by a union bound and the definition of Ω_{bd}.

Let $\mathcal{E}_1$ be the event that

$$\left\|\frac{1}{|U|}\sum_{y_j\in U} f_{2:k}(y_j)\right\| \leq 4\varepsilon\sqrt{k},$$

which happens with probability at least $1 - k\exp(-\Omega(n\varepsilon^2))$ by Lemma 15.

Let μ_t' be the distribution at time t of the process associated with the Markov semigroup initialized at $\mu_0' = \frac{1}{|U|}\sum_{y_j\in U}\delta_{y_j}$. By Lemma 52 and the definition of U, for

$$\chi^2(\mu_{t_0}'||\pi) \leq \max_{y_j\in U}\chi^2(\rho_{t_0}^{y_j}||\pi) \leq R.$$

By Lemma 13, if $\mathcal{E}_1$ holds, then for $t \geq t_0$,

$$\chi^2(\mu_t'\|\pi) \leq 16\varepsilon^2 k + e^{-\alpha(t-t_0)}R = \frac{\varepsilon_{\text{TV}}^2}{4} + e^{-\alpha(t-t_0)}R.$$

We can rewrite $\mu_0 = (1-p)\mu_0' + p\mu_0''$ where $\mu_0' = \sum_{y_j\in U}\delta_{y_j}$, $\mu_0'' = \sum_{y_j\notin U}\delta_{y_j}$ and $p = \frac{|U_{\text{sample}}\setminus U|}{n}$. Let $\mathcal{E}_2$ be the event $p \leq 4\varepsilon^2$. Since $\pi(\tilde{\Omega}^c) \leq 2\varepsilon^2$, by Chernoff's bound,

$$\Pr\left(\mathcal{E}_2\right) \geq 1 - \exp(-\Omega(n\varepsilon^2)).$$

Let μ_t, μ_t', μ_t'' be the distribution at time t of the process driven by P_t initialized at μ_0, μ_0', μ_0'', respectively; then $\mu_t = (1-p)\mu_t' + p\mu_t''$. With probability $1 - k\exp(-\Omega(n\varepsilon^2))$ over the random choice of the samples y_j, both $\mathcal{E}_1$ and $\mathcal{E}_2$ hold, and thus by triangle inequality (for details, see (Koehler and Vuong, 2023, Proposition 9))

$$\begin{aligned}\mathrm{TV}(\mu_t,\pi) &\le (1-p)\,\mathrm{TV}(\mu_t',\pi) + p\,\mathrm{TV}(\mu_t'',\pi)\\ &\le \sqrt{\frac{\varepsilon_{\mathrm{TV}}^2}{4} + e^{-\alpha(t-t_0)}R} + p \le \sqrt{\frac{\varepsilon_{\mathrm{TV}}^2}{4} + e^{-\alpha(t-t_0)}R} + \frac{\varepsilon_{\mathrm{TV}}}{16k}.\end{aligned}$$

When $t \ge t_0 + \frac{1}{\alpha}\ln\left(\frac{2R}{\varepsilon_{\mathrm{TV}}^2}\right)$, this is $\le \varepsilon_{\mathrm{TV}}$. When $n = \Omega\left(\frac{k}{\varepsilon_{\mathrm{TV}}^2}\ln\left(\frac{k}{\delta}\right)\right)$, the probability of $\mathcal{E}_1 \cup \mathcal{E}_2$ is $\ge 1-\delta$. ■

B.5. Error in sampled distribution

We will also need the following generalization of Theorem 9, which shows it is robust to TV error in the sampled distribution.

Theorem 16 *Keep the setting of Theorem 9, but assume that $y_1,\dots,y_n$ are drawn iid from ν. Then for $n = \Omega\left(\frac{k}{\varepsilon_{\mathrm{TV}}^2}\ln\left(\frac{k}{\delta}\right)\right)$, $t \ge T := t_0 + \frac{1}{\alpha}\ln\left(\frac{4R}{\varepsilon_{\mathrm{TV}}^2}\right)$ and $\mathrm{TV}(\pi,\nu) \le \frac{\varepsilon_{\mathrm{TV}}}{16}$, with probability $\ge 1-\delta$ over the randomness of the sample,*

$$\mathrm{TV}(\mu_t,\pi) \le \varepsilon_{\mathrm{TV}}.$$

Proof We use a coupling argument. Let $y_1^\pi,\dots,y_n^\pi \sim \pi$ be iid, and couple these variables with $y_1,\dots,y_n$ such that $\mathbb{P}(y_i^\pi \ne y_i) = \mathrm{TV}(\mu,\nu)$. By a Chernoff bound,

$$\mathbb{P}\left(|\{i : y_i^\pi \ne y_i\}| \le \frac{\varepsilon_{\mathrm{TV}}}{8}n\right) \ge 1 - e^{-\Omega(\varepsilon_{\mathrm{TV}}n)}.$$

Thus for $n = \Omega\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\ln\left(\frac{1}{\delta}\right)\right)$, this is $\ge 1-\frac{\delta}{2}$. Under this event, by triangle inequality

$$\mathrm{TV}\left(\mu_T, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}^\pi P_T\right) = \mathrm{TV}\left(\frac{1}{n}\sum_{i=1}^n \delta_{y_i}P_T, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}^\pi P_T\right) \le \frac{\varepsilon_{\mathrm{TV}}}{8}.$$

Also with probability $\ge 1-\frac{\delta}{2}$, by Theorem 9,

$$\mathrm{TV}\left(\frac{1}{n}\sum_{i=1}^n \delta_{y_i}^\pi P_T, \pi\right) \le \sqrt{\frac{\varepsilon_{\mathrm{TV}}^2}{4} + e^{-\alpha(t-t_0)}R} + \frac{\varepsilon_{\mathrm{TV}}}{16k} \le \frac{7\varepsilon_{\mathrm{TV}}}{8}.$$

The theorem follows from a union bound and triangle inequality. ■

Appendix C. Markov chain perturbation

We show that for both Langevin and Glauber dynamics, if the dynamics are perturbed within some error, and the chain is started at the original stationary distribution π, then the KL divergence between π and the resulting distribution grows at most linearly in this perturbation. By an averaging and concentration argument, this then implies that for most sets of samples drawn from π, the distribution starting from those samples also stays close.

C.1. Error growth when starting from stationary distribution

For Langevin dynamics, the appropriate notion of error is the L^2 error in the score estimate, and we bound the growth in error using Girsanov's Theorem, similar to the analysis for reverse diffusion in Chen et al. (2023b).

Remark 17 (Connection to score matching) *A L^2-accurate score function can naturally be obtained from score matching; in particular, Hyvärinen (2005) showed via integration by parts that*

$$\frac{1}{2}\mathbb{E}_{X\sim p}[\|\nabla \log p(X) - \nabla \log q(X)\|^2] + K_p = \mathbb{E}_{X\sim p}\left[\operatorname{Tr}\nabla^2 \log q + \frac{1}{2}\|\nabla \log q\|^2\right]$$

where K_p is a constant independent of p. If we consider the empirical analogue of the right hand side, then this gives Hyvärinen's score matching objective which can be optimized over a finite dataset, and standard tools from statistical learning theory can be used to recover bounds on the population score matching loss, i.e. the L^2-error of the score estimate (see e.g. proof of Theorem 1 of Koehler et al. (2022a)).

We let $\bar{X}_{s,t}$ denote Langevin diffusion with score estimate s, satisfying the SDE

$$d\bar{X}_{s,t} = s(\bar{X}_{s,t})\,dt + \sqrt{2}\,dB_t.$$

This reduces to the usual Langevin diffusion with stationary distribution $\pi \propto e^{-V}$ when $s = -\nabla V$. We denote the discretized Langevin dynamics with score estimate s and given step size h by $X_{s,t}$, and extend it to continuous time by interpolation:

$$dX_{s,t} = s(X_{s,\lfloor t/h\rfloor h})\,dt + \sqrt{2}\,dB_t.$$

In both cases, we denote the initial distribution (the distribution at $t = 0$) as a superscript.

Lemma 18 *Let $\pi \propto e^{-V}$ where V is β-smooth. Let s be such that $\mathbb{E}_\pi \|s - (-\nabla V)\|^2 \le \varepsilon_{\text{score}}^2$. Then we have*

$$\begin{aligned}
\mathrm{KL}(\mathcal{L}((\bar{X}^\pi_{-\nabla V,t})_{0\le t\le T})\|\mathcal{L}((\bar{X}^\pi_{s,t})_{0\le t\le T})) &\le T\varepsilon_{\text{score}}^2 \\
\mathrm{KL}(\mathcal{L}((\bar{X}^\pi_{-\nabla V,t})_{0\le t\le T})\|\mathcal{L}((X^\pi_{s,t})_{0\le t\le T})) &\le 8T(h^2 d\beta + hd)\beta^2 + 2T\varepsilon_{\text{score}}^2 \\
&\quad \text{when } T = Nh,\ N \in \mathbb{N}.
\end{aligned}$$

Proof By Girsanov's Theorem (we do not need to check Novikov's condition since we can use (Chen et al., 2023b, Eq (5.5) and Theorem 9)), we have that

$$\mathrm{KL}(\mathcal{L}((\bar{X}^\pi_{-\nabla V,t})_{0\le t\le T})\|\mathcal{L}((\bar{X}^\pi_{s,t})_{0\le t\le T})) \le \int_0^T \mathbb{E}\left\|\nabla \ln \pi(\bar{X}^\pi_{-\nabla V,t}) - s(\bar{X}^\pi_{-\nabla V,t})\right\|^2 dt \le T\varepsilon_{\text{score}}^2$$

since the distribution of $\bar{X}^{\pi}_{-\nabla V,t}$ is π. Let $t_k = kh$. Write $\bar{X}^{\pi}_{-\nabla V,t}$ as $\bar{X}_t$ for short. Also by Girsanov's Theorem,

$$\begin{aligned}
\mathrm{KL}(\mathcal{L}((\bar{X}^{\pi}_{-\nabla V,t})_{0\le t\le T})\|\mathcal{L}((X^{\pi}_{s,t})_{0\le t\le T})) &\le \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left\|\nabla \ln \pi(\bar{X}_t) - s(\bar{X}_{t_k})\right\|^2\,dt \\
&\le 2\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}\left\|\nabla \ln \pi(\bar{X}_t) - \nabla \ln \pi(\bar{X}_{t_k})\right\|^2 + \left\|\nabla \ln \pi(\bar{X}_{t_k}) - s(\bar{X}_{t_k})\right\|^2\,dt \qquad (13)\\
&\le 2\left(\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \beta^2\mathbb{E}\left\|\bar{X}_t - \bar{X}_{t_k}\right\|^2\,dt\right) + 2T\varepsilon^2_{\text{score}}. \qquad (14)
\end{aligned}$$

Now

$$\begin{aligned}
\mathbb{E}\left\|\bar{X}_t - \bar{X}_{t_k}\right\|^2 &= \left\|\int_{t_k}^{t} -\nabla V(\bar{X}_s)\,ds + \int_{t_k}^{t}\sqrt{2}\,dB_t\right\|^2 \\
&\le 2\mathbb{E}\left\|\int_{t_k}^{t}\nabla V(\bar{X}_s)\,ds\right\|^2 + 4(t-t_k)d \\
&\le 2(t-t_k)^2\mathbb{E}\left\|\nabla V(\bar{X}_s)\right\|^2 + 4(t-t_k)d \\
&\le 2(t-t_k)^2\beta d + 4(t-t_k)d
\end{aligned}$$

using the bound $\mathbb{E}_\pi\left\|-\nabla V\right\|^2 \le \beta d$ from Lemma 61. Hence

$$\int_{t_k}^{t_{k+1}}\mathbb{E}\left\|\bar{X}_t - \bar{X}_{t_k}\right\|^2\,dt \le 2h^3\beta d + 4h^2 d. \qquad (15)$$

Combining (14) and (15) gives

$$\mathrm{KL}(\mathcal{L}((\bar{X}^{\pi}_{-\nabla V,t})_{0\le t\le T})\|\mathcal{L}((\bar{X}^{\pi}_{s,t})_{0\le t\le T})) \le 2\cdot\frac{T}{h}\cdot h\cdot(2h^2 d\beta + 4hd)\beta^2 + 2T\varepsilon^2_{\text{score}}.$$

■

When π is not smooth but is a mixture of smooth distributions π_i with good tail bounds, we have a similar result.

Lemma 19 *Let $\pi \propto e^{-V} = \sum_i p_i\pi_i$ where each $\pi_i = \exp(-V_i)$ is β-smooth, $p_i > 0$, and $\sum_i p_i = 1$. Let $G(x) = \max_i\left\|\nabla V_i(x)\right\|$. Suppose $\mathbb{E}_\pi[G(x)^6] \le \tilde{G}^6$.*

Let s be such that $\mathbb{E}_\pi\left\|s - (-\nabla V)\right\|^2 \le \varepsilon^2_{\text{score}}$. Then we have

$$\mathrm{KL}(\mathcal{L}((\bar{X}^{\pi}_{-\nabla V,t})_{0\le t\le T})\|\mathcal{L}((\bar{X}^{\pi}_{s,t})_{0\le t\le T})) \le T\varepsilon^2_{\text{score}}$$

$$\mathrm{KL}(\mathcal{L}((\bar{X}^{\pi}_{-\nabla V,t})_{0\le t\le T})\|\mathcal{L}((X^{\pi}_{s,t})_{0\le t\le T})) \lesssim T\cdot\left(\beta^4(hd + h^6\tilde{G}^6 + h^3d^3) + (h^2\tilde{G}^2 + hd)\tilde{G}^4 + \varepsilon^2_{\text{score}}\right)$$

when $T = Nh$, $N \in \mathbb{N}$.

Proof Note that since $\|\nabla V(x)\| \le G(x)$, for $p \le 6$, $\mathbb{E}_\pi[\|\nabla V(x)\|^p] \le \mathbb{E}_\pi[G(x)^p] \le \tilde{G}^p$ by Hölder's inequality. Let $\bar{X}_t := \bar{X}^\pi_{-\nabla V,t}$. It suffices to bound (13). By the mean value inequality,

$$\left\|\nabla V(\bar{X}_t) - \nabla V(\bar{X}_{t_k})\right\| \le \left\|\bar{X}_t - \bar{X}_{t_k}\right\| \max_{y=\eta\bar{X}_t+(1-\eta)\bar{X}_{t_k},\eta\in[0,1]} \left\|\nabla^2 V(y)\right\|_{\text{op}}.$$

By Lemma 60 and Hölder's inequality,

$$\left\|\nabla^2 V(y)\right\|_{\text{op}} \le G(\bar{X}_t)^2 + G(\bar{X}_{t_k})^2 + \beta.$$

For $y = \eta\bar{X}_t + (1-\eta)X_{t_k}$, $\eta \in [0,1]$,

$$\begin{aligned}
\left\|\nabla^2 V(y)\right\|_{\text{op}} &\le \beta + G(y)^2 \\
&\le \beta + (G(\bar{X}_t) + \beta\left\|y - \bar{X}_t\right\|)^2 \\
&\le \beta + 2G(\bar{X}_t)^2 + 2\beta^2\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^2
\end{aligned}$$

Then,

$$\begin{aligned}
\left\|\nabla V(\bar{X}_t) - \nabla V(\bar{X}_{t_k})\right\|^2 &\le \left\|\bar{X}_{t_k} - \bar{X}_t\right\|^2 \left(\beta^2 + 2G(\bar{X}_t)^2 + 2\beta^2\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^2\right)^2 \\
&\le 3\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^2 \left(\beta^4 + 4G(\bar{X}_t)^4 + 4\beta^4\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^4\right)
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{E}\left\|\nabla V(\bar{X}_t) - \nabla V(\bar{X}_{t_k})\right\|^2 \\
&\le 3\beta^4\mathbb{E}[\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^2] + 12\beta^4\mathbb{E}[\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^6] + 12\mathbb{E}[\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^2 G(\bar{X}_t)^4] \\
&\le 3\beta^4\mathbb{E}[\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^2] + 12\beta^4\mathbb{E}[\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^6] + 12\mathbb{E}[\left\|\bar{X}_{t_k} - \bar{X}_t\right\|^6]^{1/3}\tilde{G}^4
\end{aligned}$$

where the last step uses Hölder's inequality.

We bound for $p \ge 1$

$$\begin{aligned}
\mathbb{E}\left\|\bar{X}_t - \bar{X}_{t_k}\right\|^{2p} &= \mathbb{E}\left[\left\|\int_{t_k}^t -\nabla V(\bar{X}_s)\,ds + \int_{t_k}^t \sqrt{2}\,dB_t\right\|^{2p}\right] \\
&\le 2^{2p-1}\mathbb{E}\left[\left\|\int_{t_k}^t \nabla V(\bar{X}_s)\,ds\right\|^{2p}\right] + 2^{3p-1}(t-t_k)^p d^p \\
&\le 2^{2p-1}(t-t_k)^{2p}\mathbb{E}\left\|\nabla V(\bar{X}_s)\right\|^{2p} + 2^{3p-1}(t-t_k)^p d^p \\
&\le 2^{2p-1}(t-t_k)^{2p}\tilde{G}^{2p} + 2^{3p-1}(t-t_k)^p d^p
\end{aligned}$$

For $p = 1$, using the bound $\mathbb{E}\|\nabla V\|^2 \le \beta d$ from Lemma 61 gives

$$\mathbb{E}\left\|\bar{X}_t - \bar{X}_{t_k}\right\|^2 \le 2(t-t_k)^2\beta d + 4(t-t_k)d$$

Hence,

$$\mathbb{E}\left\|\nabla V(\bar{X}_t) - \nabla V(\bar{X}_{t_k})\right\|^2 \lesssim \beta^4(hd + h^6\tilde{G}^6 + h^3d^3) + (h^2\tilde{G}^2 + hd)\tilde{G}^4.$$

Finally, substituting into (13) gives the desired bound. ■

Next we show a similar error bound for the Gibbs sampler (Glauber dynamics) that if it is executed with a distribution μ which approximately matches the conditional law of ν. This is useful because pseudolikelihood estimation can naturally produce such a μ from samples.

Lemma 20 *Suppose $t \geq 0$, $\varepsilon > 0$ and that ν and μ are distributions on a product space $\Sigma_1 \otimes \cdots \otimes \Sigma_n$ such that*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X\sim\nu} \operatorname{KL}(\nu(X_i = \cdot \mid X_{\sim i}), \mu(X_i = \cdot \mid X_{\sim i})) \leq \varepsilon. \tag{16}$$

Let $X^{(0)}, \dots, X^{(t)}$ is the trajectory of the ν-Gibbs sampler initialized from $X^{(0)} \sim \nu$, and let $\tilde{X}^{(0)}, \dots, \tilde{X}^{(t)}$ be the trajectory of the μ-Gibbs sampler with the same initialization $\tilde{X}^{(0)} = X^{(0)}$. Then the KL divergence between the law of the trajectories can be bounded as

$$\operatorname{KL}(\mathcal{L}(X^{(0)}, \dots, X^{(t)}), \mathcal{L}(\tilde{X}^{(0)}, \dots, \tilde{X}^{(t)})) \leq t\varepsilon. \tag{17}$$

Furthermore, for any $\delta > 0$, for the trajectories conditioned on the initial point $X^{(0)} \sim \nu$, we have

$$\nu\left(\operatorname{KL}(\mathcal{L}(X^{(0)}, \dots, X^{(t)} \mid X^{(0)}), \mathcal{L}(\tilde{X}^{(0)}, \dots, \tilde{X}^{(t)} \mid \tilde{X}^{(0)} = X^{(0)}) > t\varepsilon/\delta\right) \leq \delta. \tag{18}$$

Proof By the chain rule for the KL divergence (see e.g. Cover (1999)) and the fact that the dynamics are Markovian, we have

$$\begin{aligned}
&\operatorname{KL}(\mathcal{L}(X^{(0)}, \dots, X^{(t)}), \mathcal{L}(\tilde{X}^{(0)}, \dots, \tilde{X}^{(t)})) \\
&= \sum_{s=1}^{t} \mathbb{E}_{X^{(s-1)}\sim\nu} \operatorname{KL}(\nu(X^{(s)} = \cdot \mid X^{(s-1)}), \mu(\tilde{X}^{(s)} = \cdot \mid \tilde{X}^{(s-1)} = X^{(s-1)}) \\
&\leq t\varepsilon
\end{aligned}$$

where we used the assumption and the fact that a step of the Gibbs sampler picks a coordinate $i \sim Uni[n]$ and then samples from the corresponding conditional law.

The second claim in the theorem follows by applying Markov's inequality, using by the chain rule that

$$\begin{aligned}
&\operatorname{KL}(\mathcal{L}(X^{(0)}, \dots, X^{(t)}), \mathcal{L}(\tilde{X}^{(0)}, \dots, \tilde{X}^{(t)})) \\
&= \mathbb{E}_{X^{(0)}\sim\nu} \operatorname{KL}(\mathcal{L}(X^{(0)}, \dots, X^{(t)} \mid X^{(0)}), \mathcal{L}(\tilde{X}^{(0)}, \dots, \tilde{X}^{(t)} \mid \tilde{X}^{(0)} = X^{(0)})).
\end{aligned}$$

■

C.2. Error growth when starting from samples

Theorem 21 *Suppose that we have a bound*

$$\operatorname{KL}(\mathcal{L}((X_t^\pi)_{0\leq t\leq T}) \| \mathcal{L}((\tilde{X}_t^\pi)_{0\leq t\leq T})) \leq T\varepsilon \leq 1,$$

where X_t^π and $\tilde{X}_t^\pi$ denote Markov chains or processes X_t and $\tilde{X}_t$ initialized at π (the index set can be $\mathbb{N}_0$ or $\mathbb{R}_{\geq 0}$).

Let $\hat{\pi}$ be the empirical distribution of m i.i.d. samples from π. Then with probability at least $1-\gamma$ over the randomness of $\hat{\pi}$, the law of the trajectories initialized at $\hat{\pi}$ satisfy

$$\mathrm{TV}(\mathcal{L}((X_t^{\hat{\pi}})_{0\le t\le T} \mid \hat{\pi}), \mathcal{L}((\tilde{X}_t^{\hat{\pi}})_{0\le t\le T} \mid \hat{\pi})) \le \sqrt{T\varepsilon} + \frac{2\log(1/\gamma)}{m}.$$

Proof By convexity of TV distance,

$$\mathrm{TV}(\mathcal{L}((X_t^{\hat{\pi}})_{0\le t\le T} \mid \hat{\pi}), \mathcal{L}((\tilde{X}_t^{\hat{\pi}})_{0\le t\le T} \mid \hat{\pi})) \le \frac{1}{m}\sum_{i=1}^{m} \mathrm{TV}(\mathcal{L}((X_t^{Y_i})_{0\le t\le T} \mid Y_i), \mathcal{L}((\tilde{X}_t^{Y_i})_{0\le t\le T} \mid Y_i))$$

where $\hat{\pi} = \frac{1}{m}\sum_{i=1}^{m} \delta_{Y_i}$, Y_i being independent draws from π. Now for $Y \sim \pi$,

$$\begin{aligned}
&\mathrm{Var}_{Y\sim\pi}(\mathrm{TV}(\mathcal{L}((X_t^{Y})_{0\le t\le T}), \mathcal{L}((\tilde{X}_t^{Y})_{0\le t\le T})))\\
&\le \mathbb{E}_{Y\sim\pi}[\mathrm{TV}(\mathcal{L}((X_t^{Y})_{0\le t\le T}), \mathcal{L}((\tilde{X}_t^{Y})_{0\le t\le T}))^2]\\
&\le \frac{1}{2}\mathbb{E}_{Y\sim\pi}[\mathrm{KL}(\mathcal{L}((X_t^{Y})_{0\le t\le T}) \| \mathcal{L}((\tilde{X}_t^{Y})_{0\le t\le T}))] && \text{by Pinsker's inequality}\\
&= \frac{1}{2}\,\mathrm{KL}(\mathcal{L}((X_t^{\pi})_{0\le t\le T}) \| \mathcal{L}((\tilde{X}_t^{\pi})_{0\le t\le T})) \le \frac{1}{2}T\varepsilon && \text{by chain rule for KL.}
\end{aligned}$$

This calculation also shows, by Jensen's inequality

$$\begin{aligned}
&\mathbb{E}_{Y\sim\pi}[\mathrm{TV}(\mathcal{L}((X_t^{Y})_{0\le t\le T}), \mathcal{L}((\tilde{X}_t^{Y})_{0\le t\le T}))]\\
&\qquad\le \sqrt{\mathbb{E}_{Y\sim\pi}[\mathrm{TV}(\mathcal{L}((X_t^{Y})_{0\le t\le T}), \mathcal{L}((\tilde{X}_t^{Y})_{0\le t\le T}))^2]} \le \sqrt{\frac{1}{2}T\varepsilon}.
\end{aligned}$$

Since TV distance is bounded by 1, by Bernstein's inequality,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} \mathrm{TV}(\mathcal{L}((X_t^{Y_i})_{0\le t\le T} \mid Y_i), \mathcal{L}((\tilde{X}_t^{Y_i})_{0\le t\le T} \mid Y_i)) \ge \sqrt{\frac{1}{2}T\varepsilon} + u\right) \le \exp\left[-\frac{mu^2/2}{\frac{1}{2}T\varepsilon + u/3}\right].$$

Taking $u = \max\left\{\sqrt{\frac{2T\varepsilon\ln(1/\gamma)}{m}}, \frac{4\ln(1/\gamma)}{3m}\right\}$, we get that this is $\le \gamma$. Hence, with probability $\ge 1-\gamma$,

$$\mathrm{TV}(\mathcal{L}((X_t^{\hat{\pi}})_{0\le t\le T} \mid \hat{\pi}), \mathcal{L}((\tilde{X}_t^{\hat{\pi}})_{0\le t\le T} \mid \hat{\pi})) \le \sqrt{\frac{1}{2}T\varepsilon} + \max\left\{\sqrt{\frac{2T\varepsilon\ln(1/\gamma)}{m}}, \frac{4\ln(1/\gamma)}{3m}\right\}.$$

Finally, using the inequality $2ab \le a^2 + b^2$,

$$\sqrt{\frac{2T\varepsilon\ln(1/\gamma)}{m}} \le \frac{1}{4}T\varepsilon + \frac{2\ln(1/\gamma)}{m} \le \frac{1}{4}\sqrt{T\varepsilon} + \frac{2\ln(1/\gamma)}{m}$$

giving the desired conclusion. ■

Appendix D. Application: Langevin dynamics with estimated score

In this section, we illustrate how to apply our general results to sample a mixture of Poincaré distributions with an approximate score function. We also show how to obtain stronger results when the components satisfy the stronger log-Sobolev inequality.

Assumption 1 (Mixture assumption) *Let $\pi = \sum_{i=1}^k p_i \pi_i$ be a mixture of distributions $\pi_i \propto e^{-V_i(x)}$, where $p_i > 0$ for each i and $\sum_{i=1}^k p_i = 1$. Suppose each V_i is β-smooth and satisfies either:*

- $\mathsf{PI}\left(\frac{1}{\alpha}\right)$*: A Poincaré inequality with constant $C_{\mathrm{P}} \le \frac{1}{\alpha}$.*
- $\mathsf{LSI}\left(\frac{1}{\alpha}\right)$*: A log-Sobolev inequality with constant $C_{\mathrm{LS}} \le \frac{1}{\alpha}$.*

We let $\kappa := \frac{\beta}{\alpha}$. Let $\overline{x}_i := \mathbb{E}_{\pi_i} x$ and $x_i^* \in \operatorname{argmin} V_i$. Suppose L is such that the means satisfy

$$\max_{i,j} \|\overline{x}_i - \overline{x}_j\| \le L.$$

Note that we assume a bound on the distance between means rather than modes, because it is more natural to establish concentration around the mean.

Given Assumption 1, we observe that from Lemma 55 in Section H, we obtain the concentration bound

$$\mathbb{P}\left(\min_i \|x - \mathbb{E}_{\pi_i} x\| \ge R_\varepsilon\right) \le \varepsilon, \quad \text{where } R_\varepsilon := \sqrt{\frac{1}{\alpha}} \left(\sqrt{d} + \ln\left(\frac{3}{\varepsilon}\right)\right). \tag{19}$$

D.1. Warm start

To apply our result for the Langevin dynamics initialized from samples[3], we need to control the Rényi distance to stationarity of the distribution of the Langevin diffusion initialized at a typical sample x. In Lemma 23, we give a simple upper bound on q-Rényi that depends polynomially on the ambience dimension d. When the stationary distribution is a mixture where each component satisfies a log-Sobolev inequality, we can boost Lemma 23 into a better bound with no dependency on the dimension (see Lemma 25) using hypercontractivity of the mixture (see Lemma 59).

Lemma 22 *Let ν be the measure of $\mathcal{N}(y, \sigma^2 I)$. If $\pi(x) \propto \exp(-V(x))$ is β-smooth and satisfies a Poincaré inequality with constant $C_{\mathrm{P}} \le \frac{1}{\alpha}$ and $\sigma^2\beta \le \frac{1}{2}$, then*

$$\mathcal{R}_\infty(\nu\|\pi) \le \sigma^2 \|\nabla V(y)\|^2 + V(y) - V(x^*) + 1 - \frac{1}{2}\ln(\pi d) + \frac{d}{2}\ln\left(\frac{e}{\alpha\sigma^2}\right).$$

3. Note that the mixing time in Theorem 9 depends on $\mathcal{R}_2(\nu||\mu) = \log \chi^2(\nu||\mu)$.

Proof We use β-smoothness of V to upper bound $e^{V(x)}$ and Lemma 56 to upper bound $\int_{\mathbb{R}^d} e^{-V(x)}\,dx$.

$$\begin{aligned}\frac{\nu(x)}{\pi(x)} &= (2\pi\sigma^2)^{-d/2}e^{-\frac{\|x-y\|^2}{2\sigma^2}+V(x)}\cdot\int_{\mathbb{R}^d} e^{-V(x)}\,dx\\ &\le (2\pi\sigma^2)^{-d/2}e^{-\frac{\|x-y\|^2}{2\sigma^2}+V(y)+\langle\nabla V(y),x-y\rangle+\frac{\beta}{2}\|x-y\|^2}e^{-V(x^*)}\frac{e}{\sqrt{\pi d}}\left(\frac{2e\pi}{\alpha}\right)^{d/2}\\ &= \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}-\beta\right)\left\|x-y-\frac{1}{\frac{1}{\sigma^2}-\beta}\nabla V(y)\right\|^2+\frac{1}{2\left(\frac{1}{\sigma^2}-\beta\right)}\|\nabla V(y)\|^2+V(y)-V(x^*)\right]\\ &\quad\cdot\frac{e}{\sqrt{\pi d}}\left(\frac{e}{\alpha\sigma^2}\right)^{d/2}\\ &\le \exp\left[\frac{\sigma^2}{2(1-\sigma^2\beta)}\|\nabla V(y)\|^2+V(y)-V(x^*)\right]\frac{e}{\sqrt{\pi d}}\left(\frac{e}{\alpha\sigma^2}\right)^{d/2}\\ &\le \exp\left[\sigma^2\|\nabla V(y)\|^2+V(y)-V(x^*)\right]\frac{e}{\sqrt{\pi d}}\left(\frac{e}{\alpha\sigma^2}\right)^{d/2}\end{aligned}$$

Taking the log gives the result. ∎

Lemma 23 *Suppose π satisfies Assumption 1 with the Poincaré inequality $\mathsf{PI}\left(\frac{1}{\alpha}\right)$. Let $\bar{\nu}_t, \nu_t$ be respectively the distribution at time t of the continuous Langevin diffusion and the LMC with step size h initialized at δ_x. Let $G(x) := \max_i \|\nabla V_i(x)\|$. Suppose $h \le 1/(30\beta)$. Then*

$$\mathcal{R}_q(\bar{\nu}_h||\nu_h) \lesssim q^2 h(G^2(x)+\beta^2 dh) \qquad \textit{for any } q\in\left(1,\frac{1}{10\beta h}\right)$$

$$\mathcal{R}_q(\nu_h||\pi) \lesssim h\max_i\|\nabla V_i(x)\|^2+\max_i(V_i(x)-V_i(x_i^*))+d\left(1+\ln\left(\frac{1}{\alpha h}\right)\right) \qquad \textit{for any } q\in(1,\infty]$$

and

$$\begin{aligned}\mathcal{R}_q(\bar{\nu}_h||\pi) &\lesssim q^4 h(G^2(x)+\beta^2 dh)+h\max_i\|\nabla V_i(x)\|^2\\ &\quad+\max_i(V_i(x)-V(x_i^*))+d\left(1+\ln\left(\frac{1}{\alpha h}\right)\right) \qquad \textit{for any } q\in\left(1,\sqrt{\frac{1}{10\beta h}}\right).\end{aligned}$$

Proof The first inequality follows from (Koehler and Vuong, 2023, Lemma 10) (α-strong log-concavity is not used in the proof). For the second inequality, note by weak-convexity of Renyi divergence (Lemma 52) and Lemma 22,

$$\begin{aligned}\mathcal{R}_q(\nu_h||\mu) &\le \max_i \mathcal{R}_q(\nu_h||\mu_i) \le \max_i \mathcal{R}_\infty(\nu_h||\mu_i)\\ &\lesssim \max_i\left(h\|\nabla V_i(x-h\nabla V(x))\|^2+V_i(x-h\nabla V(x))-V_i(x^*)+d\left(1+\ln\left(\frac{1}{\alpha h}\right)\right)\right).\end{aligned}$$

We have by using $h \le \frac{1}{30\beta}$ and $\|\nabla V(x)\| \le \max_j \|\nabla V_j(x)\|$ that

$$\begin{aligned}
\|\nabla V_i(x - h\nabla V(x))\|^2 &\lesssim \|\nabla V_i(x)\|^2 + \beta^2 h^2 \|\nabla V(x)\|^2 \lesssim \max_j \|\nabla V_j(x)\|^2 \\
V_i(x - h\nabla V(x)) &\le V_i(x) - h\langle \nabla V_i(x), \nabla V(x)\rangle + \frac{\beta}{2} h^2 \|\nabla V(x)\|^2 \\
&\le V_i(x) + \left(h + \frac{\beta h^2}{2}\right) \max_j \|\nabla V_j\|^2 \le V_i(x) + O\left(h \max_j \|\nabla V_j(x)\|^2\right)
\end{aligned}$$

so

$$\mathcal{R}_q(\nu_h||\mu) \lesssim h \max_i \|\nabla V_i(x)\|^2 + \max_i (V_i(x) - V_i(x_i^*)) + d\left(1 + \ln\left(\frac{1}{\alpha h}\right)\right).$$

By the weak triangle inequality Lemma 51, choosing $a = q$, $b = \frac{q}{q-1}$,

$$\begin{aligned}
\mathcal{R}_q(\bar{\nu}_h||\mu) &\le \frac{aq-1}{a(q-1)} \mathcal{R}_{aq}(\bar{\nu}_h||\nu_h) + \mathcal{R}_{b(q-1)+1}(\nu_h||\mu) \\
&\lesssim \mathcal{R}_{q^2}(\bar{\nu}_h||\nu_h) + \mathcal{R}_{q+1}(\nu_h||\mu).
\end{aligned}$$

Substituting in the previous two inequalities then gives the last bound. ■

Corollary 24 *Suppose π satisfies Assumption 1 with the Poincaré inequality $\mathsf{PI}\left(\frac{1}{\alpha}\right)$. Let $\varepsilon_1 \le \frac{1}{2}$, and let $R_\varepsilon = \frac{1}{\sqrt{\alpha}}\left(\sqrt{d} + \ln\left(\frac{3}{\varepsilon}\right)\right)$. Then*

$$\mathbb{P}_{x\sim\pi}\left(\max_j \|x - \overline{x}_j\| < R_{\varepsilon_1} + L\right) \ge 1 - \varepsilon_1,$$

and under this event, for $h = \frac{1}{50\beta}$, when $\bar{\nu}_0 = \delta_x$, for any $q \in (1, 2]$,

$$\mathcal{R}_q(\bar{\nu}_h\|\pi) \lesssim \kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 + d\ln\kappa.$$

Proof By Lemma 55,

$$\mathbb{P}_{x\sim\pi_i}(\|x - \overline{x}_i\| \ge R_{\varepsilon_1}) \le \varepsilon_1.$$

Hence, considering a draw over the mixture distribution,

$$\mathbb{P}_{x\sim\pi}(\exists i,\ \|x - \overline{x}_i\| < R_{\varepsilon_1}) \ge 1 - \varepsilon_1.$$

Under this event, for all j,

$$\begin{aligned}
\|x - \overline{x}_j\| &< R_{\varepsilon_1} + L \\
V_j(x) - V_j(\overline{x}_j) &\le \langle \nabla V_j(\overline{x}_j), x - \overline{x}_j \rangle + \frac{\beta}{2} \|x - \overline{x}_j\|^2 \\
&\lesssim \beta\sqrt{d/\alpha}(R_{\varepsilon_1} + L) + \beta(R_{\varepsilon_1} + L)^2 && \text{by Lemma 57} \\
&\lesssim \frac{\beta d}{\alpha} + \beta(R_{\varepsilon_1} + L)^2 \\
&\lesssim \kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 \\
V_j(x) - \min V_j &\lesssim \kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 && \text{by Lemma 58} \\
\|\nabla V_j(x)\| &\le \|\nabla V_j(x) - \nabla V_j(\overline{x}_j)\| + \|\nabla V_j(\overline{x}_j)\| \\
&\lesssim \beta(R_{\varepsilon_1} + L) + \beta\sqrt{d/\alpha} && \text{by Lemma 57} \\
&\lesssim \sqrt{\beta\kappa d} + \sqrt{\beta\kappa} \ln\left(\frac{1}{\varepsilon_1}\right) + \beta L.
\end{aligned}$$

Then for $\bar{\nu}_0 = \delta_x$, by Lemma 23, for $h < \frac{1}{40\beta}$,

$$\begin{aligned}
\mathcal{R}_q(\bar{\nu}_h \| \pi) \lesssim h(G^2(x) + \beta^2 dh) &+ h \max_j \|\nabla V_j(x)\|^2 \\
&+ \max_j (V_j(x) - V_j(x_i^*)) + d\left(1 + \ln\left(\frac{1}{\alpha h}\right)\right)
\end{aligned}$$

Choose $h = \frac{1}{50\beta}$. By the above, with probability $\ge 1 - \varepsilon_1$, this is bounded by

$$\begin{aligned}
&\lesssim h\left(\beta\kappa d + \beta\kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta^2 L^2 + \beta^2 dh\right) + \kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 + d \ln\left(\frac{1}{\alpha h}\right) \\
&\lesssim \kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 + d \ln \kappa.
\end{aligned}$$

∎

Lemma 25 *Suppose π satisfies Assumption 1 with the Poincaré inequality $\mathsf{PI}\left(\frac{1}{\alpha}\right)$. Suppose furthermore that each π_i satisfies $\mathsf{LSI}(C_{\mathrm{LS}})$. Let $\bar{\nu}_t$ be the distribution of the continuous Langevin diffusion wrt μ initialized at δ_x, where $\max_j \|x - \overline{x}_j\| < R_{\varepsilon_1} + L$. Then there is $t = O\left(C_{\mathrm{LS}} \ln(\kappa d \ln(1/\varepsilon_1) + \beta L^2)\right)$ such that*

$$\left\|\frac{d\bar{\nu}_t}{d\mu}\right\|_{L^2(\mu)} \le \left(\frac{1}{\min p_i}\right)^{1/2} e.$$

Proof By choosing $h = \frac{1}{50\beta}$ and $\varepsilon = \frac{c}{\kappa d + \kappa \ln(1/\varepsilon_1)^2 + \beta L^2 + d \ln \kappa}$ for small enough constant c, using Corollary 24,

$$\left\|\frac{d\bar{\nu}_h}{d\mu}\right\|_{L^{1+\varepsilon}(\mu)} = \exp\left[\varepsilon \mathcal{R}_{1+\varepsilon}(\bar{\nu}_h || \mu)\right] \le e.$$

Now choose $t_1 = \frac{C_{\text{LS}}}{2} \ln\left(\frac{1}{\varepsilon}\right)$, so $q(t_1) = 1 + \varepsilon e^{2t/C_{\text{LS}}} = 2$. We get by hypercontractivity for mixtures (Lemma 59) that

$$\left\| \frac{d\bar{\nu}_{h+t_1}}{d\mu} \right\|_{L^2(\mu)} \leq \left(\frac{1}{\min p_i} \right)^{\frac{1}{1+\varepsilon} - \frac{1}{2}} \left\| \frac{d\bar{\nu}_h}{d\mu} \right\|_{L^{1+\varepsilon}(\mu)} \leq \left(\frac{1}{\min p_i} \right)^{1/2} e$$

Finally note

$$\ln\left(\frac{1}{\varepsilon}\right) = O\left(C_{\text{LS}} \ln(\kappa d + \kappa \ln(1/\varepsilon_1)^2 + \beta L^2 + d \ln \kappa)\right) = O\left(C_{\text{LS}} \ln(\kappa d \ln(1/\varepsilon_1) + \beta L^2)\right).$$

■

D.2. Convergence for Langevin diffusion

In this subsection, we upper bound the convergence time of the continuous Langevin diffusion when the stationary distribution π is a mixture of distributions with each component satisfying a Poincare or log-Sobolev inequality. In the latter case, our bound in Theorem 26 depends on the minimum weight of the mixture p^*, but this dependency will be removed in Section D.3.

Theorem 26 *Suppose that π satisfies Assumption 1 with either the Poincaré inequality $\mathsf{PI}\left(\frac{1}{\alpha}\right)$ or log-Sobolev inequality $\mathsf{LSI}\left(\frac{1}{\alpha}\right)$. Let $\hat{\pi}$ be the uniform distribution over n i.i.d. samples from[4] π and $\widetilde{\nu}_T$ be the distribution at time t of the continuous Langevin diffusions initialized at the empirical distribution $\hat{\pi}$ driven by the score function s, where s satisfies $\mathbb{E}_\pi \|s - \nabla \ln \pi\|^2 \leq \varepsilon_{\text{score}}^2$.*

Suppose $n = \Omega\left(\frac{k}{\varepsilon_{\text{TV}}^2} \ln\left(\frac{k}{\delta}\right)\right)$ with appropriate constants.

1. *Under $\mathsf{PI}\left(\frac{1}{\alpha}\right)$, for $T \geq t_0 = \frac{1}{50\beta}$, with probability $\geq 1 - \delta$,*

$$\mathrm{TV}(\widetilde{\nu}_T, \pi) \lesssim \varepsilon_{\text{TV}} + e^{O\left(\kappa d + \kappa \ln\left(\frac{k}{\varepsilon_{\text{TV}}}\right)^2 + \beta L^2 + d \ln \kappa\right)} e^{-\alpha(T - t_0)} + \sqrt{T} \varepsilon_{\text{score}} + \frac{\ln(1/\delta)}{n}.$$

   In particular, when $T = \Theta\left(\frac{\kappa}{\alpha}\left(d + \ln\left(\frac{k}{\varepsilon_{\text{TV}}}\right)^2\right) + \kappa L^2 + \frac{d}{\alpha} \ln \kappa + \frac{1}{\alpha} \ln\left(\frac{1}{\varepsilon_{\text{TV}}}\right)\right)$, and $\varepsilon_{\text{score}} = O\left(\frac{\varepsilon_{\text{TV}}}{\sqrt{T}}\right)$ with appropriate constants, then this is $\leq \varepsilon_{\text{TV}}$.

2. *Under $\mathsf{LSI}\left(\frac{1}{\alpha}\right)$, there is $t_1 = O\left(\frac{1}{\alpha} \ln(\kappa d \ln(k/\varepsilon_{\text{TV}}) + \beta L^2)\right)$, such that for $T \geq t_1$, with probability $\geq 1 - \delta$, letting[5] $p^* = \min_i p_i$*

$$\mathrm{TV}(\widetilde{\nu}_T, \pi) \lesssim \varepsilon_{\text{TV}} + \left(\frac{1}{p^*}\right)^{1/2} e^{-\alpha(T - t_1)} + \sqrt{T} \varepsilon_{\text{score}} + \frac{\ln(1/\delta)}{n}.$$

   In particular, when $T = t_1 + \Theta\left(\frac{1}{\alpha} \ln\left(\frac{1}{\varepsilon_{\text{TV}} p^*}\right)\right)$, and $\varepsilon_{\text{score}} = O\left(\frac{\varepsilon_{\text{TV}}}{\sqrt{T}}\right)$ with appropriate constants, then this is $\leq \varepsilon_{\text{TV}}$.

4. as in Theorem 9

5. Recall from Assumption 1 that $\pi = \sum p_i \pi_i$

Proof Let $\bar{\nu}_t^{\hat{\pi}}$ be the distribution at time t of the continuous Langevin diffusions initialized at the empirical distribution $\hat{\pi}$ driven by the score function $\nabla \ln \pi$.

1. Let $R_\varepsilon = \frac{1}{\sqrt{\alpha}}\left(\sqrt{d} + \ln\left(\frac{3}{\varepsilon}\right)\right)$. Let $\varepsilon_1 = \frac{\varepsilon_{\text{TV}}^2}{64k}$. By Corollary 24, when $\bar{\nu}_0 = \delta_x$, with probability $\geq 1 - \varepsilon_1$ over $x \sim \pi$,

$$\ln(\chi^2(\bar{\nu}_h \| \pi) + 1) = \mathcal{R}_2(\bar{\nu}_h \| \pi) \lesssim \kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 + d \ln \kappa.$$

   By Theorem 9, with probability $\geq 1 - \frac{\delta}{2}$,

$$\mathrm{TV}(\bar{\nu}_T^{\hat{\pi}}, \pi) \leq \frac{\varepsilon_{\text{TV}}}{2} + e^{O\left(\kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 + d \ln \kappa\right)} e^{-\alpha(T-h)}.$$

   Now we consider score error. By Lemma 19 and Theorem 21, with probability $\geq 1 - \frac{\delta}{2}$,

$$\mathrm{TV}(\bar{\nu}_T^{\hat{\pi}}, \widetilde{\nu}_T) \leq \sqrt{2T}\varepsilon_{\text{score}} + \frac{2\ln(2/\delta)}{n}.$$

   By the triangle inequality, with probability $\geq 1 - \delta$,

$$\mathrm{TV}(\widetilde{\nu}_T, \pi) \leq \frac{\varepsilon_{\text{TV}}}{2} + e^{O\left(\kappa d + \kappa \ln\left(\frac{1}{\varepsilon_1}\right)^2 + \beta L^2 + d \ln \kappa\right)} e^{-\alpha(T-h)} + \sqrt{2T}\varepsilon_{\text{score}} + \frac{2\ln(2/\delta)}{n}.$$

   When $T = \Theta\left(\frac{\kappa}{\alpha}\left(d + \ln\left(\frac{1}{\varepsilon_1}\right)^2\right) + \kappa L^2 + \frac{d}{\alpha}\ln\kappa + \frac{1}{\alpha}\ln\left(\frac{1}{\varepsilon_{\text{TV}}}\right)\right)$, $\varepsilon_{\text{score}} = O\left(\frac{\varepsilon_{\text{TV}}}{\sqrt{T}}\right)$, and $\delta = e^{-O(n\varepsilon_{\text{TV}})}$ with appropriate constants, then this is $\leq \varepsilon_{\text{TV}}$.

2. Let $\bar{\nu}_t^x$ be the distribution at time t of the continuous Langevin diffusion initialized at δ_x driven by the score function $\nabla \ln \pi$. By Lemma 25, there is $t_1 = O\left(\frac{1}{\alpha}\ln(\kappa d \ln(1/\varepsilon_1) + \beta L^2)\right)$ such that with probability $\geq 1 - \varepsilon_1$ over $x \sim \pi$,

$$\chi^2(\bar{\nu}_{t_1}^x \| \pi) \leq \left\| \frac{d\bar{\nu}_{t_1}}{d\pi} \right\|_{L^2(\pi)} \leq \left(\frac{1}{\min p_i}\right)^{1/2} e.$$

   Set $\varepsilon_1 = O(\frac{\varepsilon_{\text{TV}}^2}{16k})$ with appropriate constant. By Theorem 9, with probability $\geq 1 - \frac{\delta}{2}$, for $T \geq t_1$,

$$\mathrm{TV}(\bar{\nu}_T, \pi) \leq \frac{\varepsilon_{\text{TV}}}{2} + \left(\frac{1}{\min p_i}\right)^{1/2} e \cdot e^{-\alpha(T-t_1)}.$$

   By the triangle inequality, with probability $\geq 1 - \delta$,

$$\mathrm{TV}(\widetilde{\nu}_T, \pi) \leq \frac{\varepsilon_{\text{TV}}}{2} + \left(\frac{1}{\min p_i}\right)^{1/2} e \cdot e^{-\alpha(T-t_1)} + \sqrt{2T}\varepsilon_{\text{score}} + \frac{2\ln(2/\delta)}{n}.$$

   When $T = t_1 + \Theta\left(\frac{1}{\alpha}\ln\left(\frac{1}{\varepsilon_{\text{TV}}p^*}\right)\right)$, $\varepsilon_{\text{score}} = O\left(\frac{\varepsilon_{\text{TV}}}{\sqrt{T}}\right)$, and $\delta = e^{-O(n\varepsilon_{\text{TV}})}$ with appropriate constants, then this is $\leq \varepsilon_{\text{TV}}$.

∎

D.3. Removing dependence on the minimum weight

In this subsection, we improve the bound on the convergence time of the continuous Langevin diffusion in part 2 of Theorem 26 by removing the dependency on the minimum weight p^* of the mixture. This is achieved by comparing the Langevin diffusion wrt π and the Langevin diffusion wrt π' where π' is the submixture obtained by removing the low-weight components of π (see Theorem 30 for details).

We define some notation to refer to submixtures containing a subset of the components.

Definition 27 *Suppose that $\pi = \sum_{i=1}^k p_i\pi_i$, where $p_i > 0$ for each i and $\sum_{i=1}^k p_i = 1$. For set $S \subseteq [k]$, let $p_S = \sum_{i\in S} p_i$ and*

$$\pi_S = p_S^{-1} \sum_{i\in S} p_i\pi_i.$$

Note that we can write $\pi(x) = p_S\pi_S(x) + p_{S^c}\pi_{S^c}(x)$ where S^c is the complement of S. Let V_S be such that $\pi_S(x) = \exp(-V_S(x))$.

Lemma 28 *Suppose π satisfies Assumption 1 with the Poincaré inequality $\mathsf{PI}\left(\frac{1}{\alpha}\right)$. Let $G = \max_{1\le i\le k} \|\nabla V_i(x)\|$. Let $\tilde{G} = (\beta\kappa d)^{1/2} + \beta L$. Then for all $1 \le i \le k$ and $p = O(1)$,*

$$\mathbb{E}_{\pi_i} \|G(x)\|^{2p} \lesssim \tilde{G}^p \quad \text{and} \quad \mathbb{E}_{\pi} \|G(x)\|^{2p} \lesssim \tilde{G}^p.$$

Proof For all $1 \le j \le k$,

$$\begin{aligned}
\|\nabla V_j(x)\| &\le \|\nabla V_j(x) - \nabla V_j(\overline{x}_j)\| + \|\nabla V_j(\overline{x}_j)\| \\
&\le \beta \|x - \overline{x}_j\| + \|\nabla V_j(\overline{x}_j)\| \\
&\lesssim \beta \|x - \overline{x}_i\| + \beta \|\overline{x}_i - \overline{x}_j\| + \beta\sqrt{\frac{d}{\alpha}}
\end{aligned}$$

by Lemma 57. Then by (27),

$$\begin{aligned}
&\|G(x)\|^{2p} \lesssim \beta^{2p} \|x - \overline{x}_i\|^{2p} + \beta^{2p}L^{2p} + (\beta\kappa d)^p \\
\implies &\mathbb{E}_{\pi_i} \|G(x)\|^{2p} \lesssim \mathbb{E}_{\pi_i}[\|x - \bar{x}_i\|^{2p}] + (\beta\kappa d)^p + (\beta L)^{2p} \lesssim (\beta\kappa d)^p + (\beta L)^{2p}.
\end{aligned}$$

■

Lemma 29 (L^2 score error from removing small components) *Suppose π satisfies Assumption 1 with the Poincaré inequality $\mathsf{PI}\left(\frac{1}{\alpha}\right)$. Then*

$$\mathbb{E}_{\pi} \|\nabla V(x) - \nabla V_S(x)\|^2 \lesssim p_{S^c}(\beta\kappa d + \beta^2 L^2).$$

Proof First note that

$$\nabla V(x) - \nabla V_S(x) = \frac{\sum_{i\in S^c} p_i\pi_i(x)(\nabla V_i(x) - \nabla V_S(x))}{p_S\pi_{S^c}(x) + p_S\pi_S(x)}.$$

Letting $G(x) = \max_{1\le i\le k} \|\nabla V_i(x)\|$, note that $\|\nabla V(x)\|, \|\nabla V_S(x)\| \le G(x)$ by (Koehler and Vuong, 2023, Proposition 6). Thus, we have

$$\begin{aligned}\mathbb{E}_\pi \|\nabla V(x) - \nabla V_S(x)\|^2 &\le 4 \int \|G(x)\|^2 \frac{p_{S^c}\pi_{S^c}(x)}{p_{S^c}\pi_{S^c}(x) + p_S\pi_S(x)} \sum_{i\in S^c} p_i\pi_i(x)\, dx \\ &\le 4 \sum_{i\in S^c} p_i \mathbb{E}_{\pi_i} \|G(x)\|^2 . \end{aligned} \tag{20}$$

Then (20) combined with Lemma 28 give the desired bound. ■

Theorem 30 *Let $\hat\pi$ be the uniform distribution over n i.i.d. samples from*[6] *π and $\widetilde{\nu}_T$ be the distribution at time t of the continuous Langevin diffusion initialized at the empirical distribution $\hat\pi$ driven by s, where s satisfies $\mathbb{E}_\pi \|s - \nabla \ln \pi\|^2 \le \varepsilon_{\text{score}}^2$.*

Suppose that π satisfies Assumption 1 with the log-Sobolev inequality $\mathsf{LSI}\left(\frac{1}{\alpha}\right)$, and $n = \Omega\left(\frac{k}{\varepsilon_{\text{TV}}^2} \ln\left(\frac{k}{\delta}\right)\right)$ with appropriate constants. Let $\varepsilon_{\text{small}} \in (0, \frac{1}{2})$. There is

$$t_1 = O\left(\frac{1}{\alpha} \ln(\kappa d \ln(k/\varepsilon_{\text{TV}}) + \beta L^2)\right),$$

such that for $T \ge t_1$, with probability $\ge 1 - \delta$,

$$\mathrm{TV}(\widetilde{\nu}_T, \pi) \lesssim \varepsilon_{\text{TV}} + \left(\frac{k}{\varepsilon_{\text{small}}}\right)^{1/2} e^{-\alpha(T-t_1)} + \sqrt{T}\left(\varepsilon_{\text{score}} + \sqrt{\varepsilon_{\text{small}}(\beta\kappa d + \beta^2 L^2)}\right) + \frac{\ln(1/\delta)}{n} + \varepsilon_{\text{small}}.$$

In particular, when $T = t_1 + \Omega\left(\frac{1}{\alpha}\ln\left(\frac{k}{\varepsilon_{\text{TV}}\varepsilon_{\text{small}}}\right)\right)$, $\varepsilon_{\text{score}} = O\left(\frac{\varepsilon_{\text{TV}}}{\sqrt{T}}\right)$, and $\varepsilon_{\text{small}} = O\left(\frac{\varepsilon_{\text{TV}}^2}{T(\beta\kappa d + \beta^2 L^2)}\right)$ with appropriate constants, then this is $\le \varepsilon_{\text{TV}}$. The condition on $\varepsilon_{\text{small}}$ can be satisfied by taking $T = \Theta\left(\frac{1}{\alpha}\ln\left(\frac{k(\kappa^2 d + \beta\kappa L^2)}{\varepsilon_{\text{TV}}}\right)\right)$ and

$$\varepsilon_{\text{small}} = O\left(\frac{\varepsilon_{\text{TV}}^2 \alpha}{\ln\left(\frac{k\kappa d + \beta L^2}{\varepsilon_{\text{TV}}}\right)(\beta\kappa d + \beta^2 L^2)}\right) = O\left(\frac{\varepsilon_{\text{TV}}^2}{\ln\left(\frac{k\kappa^2 d + \beta\kappa L^2}{\varepsilon_{\text{TV}}}\right)(\kappa^2 d + \beta\kappa L^2)}\right)$$

for appropriate constants.

Proof Let

$$S = \left\{ i : p_i > \frac{\varepsilon_{\text{small}}}{k} \right\}.$$

Then $p_{S^c} \le \varepsilon_{\text{small}}$. We have samples from π, but treat them as samples from π_S with TV error $\le \varepsilon_{\text{small}}$. Comparing Langevin with ∇V and ∇V_S, by Lemma 29, the score error is

$$\begin{aligned}\mathbb{E}_{\pi_S} \|s(x) - (-\nabla V_S(x))\|^2 &\le 2\mathbb{E}_\pi \|s(x) - (-\nabla V_S(x))\|^2 \\ &\lesssim \mathbb{E}_\pi \|s(x) - (-\nabla V(x))\|^2 + \mathbb{E}_\pi \|\nabla V(x) - \nabla V_S(x)\|^2 \\ &\lesssim \varepsilon_{\text{score}}^2 + p_{S^c}(\beta\kappa d + \beta^2 L^2) \le \varepsilon_{\text{score}}^2 + \varepsilon_{\text{small}}(\beta\kappa d + \beta^2 L^2).\end{aligned}$$

6. as in Theorem 9

Let $\bar{\nu}_t$ be the distribution at time t of the continuous Langevin diffusion initialized at the empirical distribution $\hat{\pi}$ driven by the score functions $\nabla \ln \pi_S \equiv -\nabla V_S$.

Note that $\pi_S = \sum_{i\in S} p'_i \pi_i$ where $p'_i = p_i/p_S \ge p_i > \frac{\varepsilon_{\text{small}}}{k}$. As in the proof of Theorem 26(2), by Lemma 25 and Theorem 16, choosing t_1, T as in the theorem statement gives with probability $\ge 1 - \frac{\delta}{2}$ that

$$\mathrm{TV}(\bar{\nu}_T, \pi_S) \lesssim \varepsilon_{\text{TV}} + \left(\frac{k}{\varepsilon_{\text{small}}}\right)^{1/2} e^{-\alpha(T-t_1)}.$$

By Lemma 19 and Theorem 21, with probability $\ge 1 - \frac{\delta}{2}$,

$$\mathrm{TV}(\bar{\nu}_T, \widetilde{\nu}_T) \lesssim \sqrt{T}\left(\varepsilon_{\text{score}} + \sqrt{\varepsilon_{\text{small}}(\beta\kappa d + \beta^2 L^2)}\right) + \frac{\ln(1/\delta)}{n}.$$

Noting $\mathrm{TV}(\pi_S, \pi) \lesssim \varepsilon_{\text{small}}$, by the triangle inequality, with probability $\ge 1 - \delta$,

$$\mathrm{TV}(\widetilde{\nu}_T, \pi) \lesssim \varepsilon_{\text{TV}} + \left(\frac{k}{\varepsilon_{\text{small}}}\right)^{1/2} e^{-\alpha(T-t_1)} + \sqrt{T}\left(\varepsilon_{\text{score}} + \sqrt{\varepsilon_{\text{small}}(\beta\kappa d + \beta^2 L^2)}\right) + \frac{\ln(1/\delta)}{n} + \varepsilon_{\text{small}}.$$

Choosing parameters as in the theorem statement then makes this $\le \varepsilon_{\text{TV}}$. ■

D.4. Convergence for Langevin Monte Carlo

The Langevin Monte-Carlo (LMC) algorithm is obtained by discretizing the Langevin diffusion Equation (4). The LMC with score function s and step size $h > 0$ is defined by

$$dX_t = s(X_{\lfloor t/h \rfloor h})dt + \sqrt{2}dB_t \tag{21}$$

The score function s approximates the gradient of log-likehood of the target distribution π i.e. $\nabla \log \pi$.

In Theorem 31, we show that the LMC with appropriately chosen step size h becomes close to the target distribution π in total-variation distance after a polynomial number of steps.

Theorem 31 (Langevin Monte Carlo on mixture) *Let $\pi = \exp(-V(x)) = \sum_{i=1}^k p_i \pi_i$ be a mixture of distributions $\pi_i \propto \exp(-V_i(x))$ satisfying Assumption 1. Let $\tilde{G} = (\beta\kappa d)^{1/2} + \beta L$.*

Let $\hat{\pi}$ be the empirical distribution i.e. the uniform distribution over be the multiset U_{sample} of $y_1, \dots, y_n$ i.i.d. sampled from π.

Let s be such that $\mathbb{E}_\pi \|s - (-\nabla V)\|^2 \le \varepsilon_{\text{score}}^2$. Let $X_t^{\hat{\pi}}$ be the LMC driven by s with step size h initialized at $\hat{\pi}$. Fix $\delta > 0$. Suppose $n \ge \Omega\left(\frac{k}{\varepsilon_{\text{TV}}^2} \ln\left(\frac{k}{\delta}\right)\right)$. We state two results, corresponding to the two different functional inequalities for the mixture components.

1. *Under* PI $\left(\frac{1}{\alpha}\right)$, *suppose*

$$T = \Theta\left(\frac{\kappa}{\alpha}\left(d + \ln\left(\frac{k}{\varepsilon_{\text{TV}}}\right)^2\right) + \kappa L^2 + \frac{d}{\alpha}\ln\kappa + \frac{1}{\alpha}\ln\left(\frac{1}{\varepsilon_{\text{TV}}}\right)\right)$$

and

$$h = O\left(\frac{\varepsilon_{\text{TV}}^2}{d(\tilde{G}^4 + \beta^4)T}\right), \quad \varepsilon_{\text{score}} = O\left(\frac{\varepsilon_{\text{TV}}}{\sqrt{T}}\right)$$

with appropriate constants. Then with probability $\ge 1 - \delta$ over the randomness of U_{sample},

$$\mathrm{TV}(\mathcal{L}(X_T^{\hat{\pi}} \mid \hat{\pi}), \pi) \le \varepsilon_{\text{TV}}.$$

2. *Under* LSI $\left(\frac{1}{\alpha}\right)$, *suppose*

$$T = \Theta\left(\frac{1}{\alpha}\ln\left(\frac{k\kappa d + \beta L^2}{\varepsilon_{\text{TV}}}\right)\right), \; h = O\left(\frac{\varepsilon_{\text{TV}}^2}{d(\tilde{G}^4 + \beta^4)T}\right), \; \varepsilon_{\text{score}} = O\left(\frac{\varepsilon_{\text{TV}}}{\sqrt{T}}\right)$$

with appropriate constants. Then with probability $\geq 1 - \delta$ *over the randomness of* U_{sample}*,*

$$\text{TV}(\mathcal{L}(X_T^{\hat{\pi}} \mid \hat{\pi}), \pi) \leq \varepsilon_{\text{TV}}.$$

Remark 32 *When* $L \geq 1$, $\left\|\nabla^2 V\right\|_{OP}$ *can be as large as* $\beta_* = (\beta L)^2 = \Theta(\tilde{G}^2)$, *thus the step size* $h = \frac{1}{d(\tilde{G}^4+\beta^4)T} = \Theta(\frac{1}{d\beta_*^2 T})$ *is expected for Langevin Monte-Carlo.*

Proof The proof is similar to that of Theorem 26. We can compare between the Langevin Monte Carlo and the continuous Langevin diffusion by combining Lemmas 19, 28 and Theorem 21. Let $\bar{X}_t^{\hat{\pi}}, \bar{X}_t^{\pi}$ be the continuous Langevin diffusion initialized at $\hat{\pi}$ and π respectively. Let X_t^{π} be the LMC with step size h and score s initialized at π. Using Lemma 19 combined with the bound on the expected gradient from lemma 28, we have

$$\text{KL}(\mathcal{L}((X_t^{\pi})_{0\leq t\leq T})||\mathcal{L}((\bar{X}_T^{\pi})_{0\leq t\leq T})) \lesssim \sqrt{T\cdot\left(\beta^4(hd + h^6\tilde{G}^6 + h^3d^3) + (h^2\tilde{G}^2 + hd)\tilde{G}^4 + \varepsilon_{\text{score}}^2\right)}$$

Thus, by Theorem 21, with probability $\geq 1 - \frac{\delta}{2}$,

$$\begin{aligned}&\text{TV}(\mathcal{L}(\bar{X}_T^{\hat{\pi}} \mid \hat{\pi}), \mathcal{L}(X_T^{\hat{\pi}} \mid \hat{\pi}))\\ &\lesssim \sqrt{T\cdot\left(\beta^4(hd + h^6\tilde{G}^6 + h^3d^3) + (h^2\tilde{G}^2 + hd)\tilde{G}^4 + \varepsilon_{\text{score}}^2\right)} + \frac{\ln(1/\delta)}{n} \leq \frac{\varepsilon_{\text{TV}}}{2}\end{aligned}$$

by our choice of h, $\varepsilon_{\text{score}}$ and n.

Part 1 is by combining Theorem 26(1) and the triangle inequality. Indeed, the choice of T implies

$$\text{TV}(\mathcal{L}(X_T^{\hat{\pi}} \mid \hat{\pi}), \pi) \leq \varepsilon_{\text{TV}}/2 + \text{TV}(\mathcal{L}(\bar{X}_T^{\hat{\pi}} \mid \hat{\pi}), \pi) \leq \varepsilon_{\text{TV}}$$

Similarly, part 2 follows by combining Theorem 30 and the triangle inequality. ■

Appendix E. Application to low-complexity Gibbs measures

E.1. Multiplicative mixture approximation

The following result shows that for spin systems with pairwise/quadratic interactions, we can eliminate large positive eigenvalues in the interaction matrix at the cost of decomposing them as "small" mixture distributions. The main idea is to apply the Hubbard-Stratonovich transform to the large eigendirections of the interaction matrix, which naturally decomposes the distribution into an explicit infinite mixture, and then argue that this decomposition can be discretized to a finite mixture with small loss. This is inspired by Koehler et al. (2022b), where a related idea was used to build an approximate sampling algorithm in the case of the Ising model (i.e. μ_i supported on $\{\pm 1\}$). The argument here is done in a more general setting and crucially guarantees a multiplicative approximation to the true density.

Theorem 33 *Suppose that $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ is a product measure where each μ_i is supported within a Euclidean ball in $\mathbb{R}^d$ of radius D. Suppose that the measure π is absolutely continuous to μ with Radon-Nikodym derivative*

$$\frac{d\pi}{d\mu}(x) \propto \exp\left(\frac{1}{2}\langle x, Jx\rangle\right).$$

Suppose that $J = J_+ + \tilde{J}$ and that J_+ has rank r with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$. Then there exists a set S and a mixture distribution π_2 with mixing weights $p_h > 0$ for $h \in S$ such that:

$$\pi_2(x) = \sum_{h \in S} p_h \tilde{\pi}_h(x)$$

where $\tilde{\pi}_h$ is the probability measure with normalizing constant Z_h satisfying

$$\frac{d\tilde{\pi}_h}{d\mu}(x) = \frac{1}{Z_h} \exp\left(\langle x, \tilde{J}x\rangle + \langle h, x\rangle\right).$$

Furthermore,

$$\frac{d\pi_2}{d\pi}(x) \in [1/e^3, e^3]$$

for all x, and

$$|S| \leq \left(D^2 n\lambda_1 + D\sqrt{n}r\sqrt{\lambda_1} + \sqrt{\lambda_1 D^2 nr \ln(\lambda_r^{-1/2} + D\sqrt{n})}\right)^{O(r)}.$$

Remark 34 *In the greater than one-dimensional case, we view x as an $n \times d$ matrix and the inner product is the usual Frobenius inner product on matrices.*

Remark 35 *The proof can be straightforwardly modified to replace the interval $[1/e^3, e^3]$ by one of width $[1-\delta, 1+\delta]$ for any $\delta > 0$, with a polylogarithmic dependence on δ.*

Remark 36 (Exponential dependence on r is necessary) *Consider an Ising model which is a disjoint union of r low temperature Curie-Weiss models each on m spins. Since each Curie-Weiss model individually has a bimodal limiting distribution for the average magnetization $\frac{1}{m}\sum_i X_i$ (see e.g. Ellis (2007)) as $m \to \infty$, the joint distribution of the r magnetizations will have 2^r well-separated modes.*

Proof Using the moment generating function of a Gaussian distribution (Hubbard-Stratonovich identity) we have

$$e^{\|J_+^{1/2}x\|_2^2/2} = \mathbb{E}_{H \sim N(0,J_+)}\left[e^{\langle x, H\rangle}\right].$$

Therefore

$$\begin{aligned}
\frac{d\pi}{d\mu}(x) &\propto \mathbb{E}_{H \sim N(0,J_+)}\left[\exp\left(\frac{1}{2}\langle x, \tilde{J}x\rangle + \langle x, H\rangle\right)\right] \\
&\propto \mathbb{E}_{H \sim N(0,J_+)}\left[\frac{Z_H}{Z_H}\exp\left(\frac{1}{2}\langle x, \tilde{J}x\rangle + \langle x, H\rangle\right)\right] \\
&\propto \mathbb{E}_{H \sim N(0,J_+)}\left[Z_H \tilde{\pi}_H(x)\right] \\
&\propto \int_{\mathrm{im}(J_+)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+^{\dagger} H\rangle/2} dH
\end{aligned}$$

where the last integral is with respect to Lebesgue measure on the hyperplane $\mathrm{im}(J_+)$ ($\mathrm{im}(J_+)$ denotes the image or column space of J_+), and where $J_+^\dagger$ is the pseudoinverse of J_+. Recall here that

$$\frac{d\tilde{\pi}_H}{d\mu}(x) = \frac{1}{Z_H}\exp\left(\langle x, \tilde{J}x\rangle/2 + \langle H, x\rangle\right).$$

Let $R > 0$ to be fixed later and let $B_R(0)$ be the ball of radius R centered at the origin. We have that

$$\frac{d\pi}{d\mu}(x) \propto \int_{\mathrm{im}(J_+)\cap B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+^\dagger H\rangle/2} dH + \int_{\mathrm{im}(J_+)\setminus B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+^\dagger H\rangle/2} dH$$

and for any H, we know by the Cauchy-Schwarz inequality that

$$Z_H/Z_0 = \mathbb{E}_{X\sim\tilde{\pi}_0} e^{\langle H, X\rangle} \in [e^{-\|H\|_2 D\sqrt{n}}, e^{\|H\|_2 D\sqrt{n}}], \tag{22}$$

and similarly

$$\tilde{\pi}_H(x)/\tilde{\pi}_0(x) = \frac{e^{\langle H, X\rangle}}{Z_H} \in [e^{-2\|H\|_2 D\sqrt{n}}, e^{2\|H\|_2 D\sqrt{n}}]. \tag{23}$$

Let C_r be the surface area of the ($r-1$ dimensional) unit sphere in r dimensions. It follows that if J_+ has rank r with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ that

$$\begin{aligned}
\int_{\mathrm{im}(J_+)\setminus B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+^\dagger H\rangle/2} dH &\leq \tilde{\pi}_0(x) Z_0 \int_{\mathrm{im}(J_+)\setminus B_R(0)} e^{3\|H\|_2 D\sqrt{n} - \lambda_1^{-1}\|H\|_2^2/2} dH \\
&\leq C_r \tilde{\pi}_0(x) Z_0 \int_R^\infty s^{r-1} e^{3sD\sqrt{n} - \lambda_1^{-1}s^2/2} ds \\
&\leq C_r \tilde{\pi}_0(x) Z_0 e^{-\lambda_1^{-1}R^2/4} \int_R^\infty s^{r-1} e^{3sD\sqrt{n} - \lambda_1^{-1}s^2/4} ds \\
&\leq 2C_r \tilde{\pi}_0(x) Z_0 e^{-\lambda_1^{-1}R^2/4}
\end{aligned}$$

where in the first inequality we applied Equations (22) and (23), in the second inequality we rewrote the integral in terms of spherical coordinates, and the last step holds under the requirement $R = \Omega(D\sqrt{n}\lambda_1 + r\lambda_1)$.

By similar arguments, we have that

$$\begin{aligned}
\int_{\mathrm{im}(J_+)\cap B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+^\dagger H\rangle/2} dH &\geq \tilde{\pi}_0(x) Z_0 \int_{\mathrm{im}(J_+)\cap B_R(0)} e^{-3\|H\|_2 D\sqrt{n} - \lambda_r^{-1}\|H\|_2^2/2} dH \\
&\geq C_r \tilde{\pi}_0(x) Z_0 \int_0^R s^{r-1} e^{-3sD\sqrt{n} - \lambda_r^{-1}s^2/2} ds \\
&\geq C_r \tilde{\pi}_0(x) Z_0 \int_0^{1/(\lambda_r^{-1/2} + D\sqrt{n})} s^{r-1} e^{-3sD\sqrt{n} - \lambda_r^{-1}s^2/2} ds \\
&\geq e^{-4}(C_r/r)\tilde{\pi}_0(x) Z_0 [\lambda_r^{-1/2} + D\sqrt{n}]^{-r}.
\end{aligned}$$

Hence

$$\frac{\int_{\mathrm{im}(J_+)\setminus B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+ H\rangle/2} dH}{\int_{\mathrm{im}(J_+)\cap B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+ H\rangle/2} dH} \leq 2e^4 r[\lambda_r^{-1/2} + D\sqrt{n}]^r e^{-\lambda_1^{-1}R^2/4} < 1/4$$

provided that we additionally require $R = \Omega\left(\sqrt{r\lambda_1 \ln(\lambda_r^{-1/2} + D\sqrt{n})}\right)$.

Therefore if we define π_1 to be the probability measure with density

$$\frac{d\pi_1}{d\mu}(x) \propto \int_{\mathrm{im}(J_+)\cap B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+ H\rangle/2} dH$$

we have shown that $\frac{d\pi_1}{d\pi}(x) \in [3/4, 5/4]$ for all x. Let $\delta > 0$ and $N_\delta(R)$ to be an δ-net of $\mathrm{im}(J_+) \cap B_R(0)$. For any point $h \in N_\delta(R)$, define the set S_h to be the subset of points in $\mathrm{im}(J_+) \cap B_R(0)$ such that h is the closest point in the net, i.e. S_h is the Voronoi region of h. Define probability measure

$$\frac{d\pi_2}{d\mu}(x) \propto \sum_{h \in N_\delta(R)} \tilde{\pi}_h(x) \int_{S_h} Z_H e^{-\langle H, J_+ H\rangle/2} dH.$$

Since $\tilde{\pi}_h(x)/\tilde{\pi}_H(x) \in [e^{-2\delta D\sqrt{n}}, e^{2\delta D\sqrt{n}}]$, it follows that if $\delta = 1/2D\sqrt{n}$ then for any $H \in N_\delta(R)$, $\tilde{\pi}_h(x)/\tilde{\pi}_H(x) \in [1/e, e]$. Using that

$$\frac{d\pi_2}{d\pi_1}(x) \propto \frac{\sum_{h\in N_\delta(R)} \tilde{\pi}_h(x) \int_{S_h} Z_H e^{-\langle H, J_+ H\rangle/2} dH}{\int_{\mathrm{im}(J_+)\cap B_R(0)} \tilde{\pi}_H(x) Z_H e^{-\langle H, J_+ H\rangle/2} dH}$$

we conclude that $\frac{d\pi_2}{d\pi_1}(x) \in [1/e^2, e^2]$ for all x, hence $\frac{d\pi_2}{d\pi}(x) \in [1/e^3, e^3]$ for all x.

Finally, by standard covering number bounds (see e.g. Vershynin (2018)) we observe that π_2 is a mixture of

$$|N_\delta(R)| \le (R/\delta)^{O(r)}$$

many components. Taking

$$R = \Theta\left(\lambda_1 D\sqrt{n} + r\sqrt{\lambda_1} + \sqrt{\lambda_1 r \ln(\lambda_r^{-1/2} + D\sqrt{n})}\right)$$

yields that

$$R/\delta = D\sqrt{n} \cdot \Theta\left(D\sqrt{n}\lambda_1 + r\sqrt{\lambda_1} + \sqrt{\lambda_1 r \ln(\lambda_r^{-1/2} + D\sqrt{n})}\right)$$

as claimed. ■

E.2. Exact mixture decomposition and higher-order spectral gap

The previous section gave an approximate mixture decomposition, but when the mixture components satisfy a functional inequality, we now show that it leads to an exact mixture decomposition of components satisfying the same functional inequality. This in particular yields a higher-order spectral gap estimate.

We now give an example to illustrate the idea. For concreteness, we state the following result for Ising models, but as discussed after the proof it can be adapted to other similar models like the Potts model or $O(N)$ model, since Theorem 33 is a general result. For example, we illustrate the application of the result to prove a higher-order spectral gap for Glauber dynamics in the mean-field Potts model.

Theorem 37 *Let $c \in [1, \infty)$. Suppose that π is a probability measure on $\{\pm 1\}^n$ satisfying*

$$\pi(x) \propto \exp\left((1/2)\langle x, Jx\rangle + \langle h, x\rangle\right)$$

for some symmetric interaction matrix J and external field vector h. Suppose that J has r eigenvalues greater than $1 - 1/c$ and the rest are at most $1 - 1/c$, i.e. the eigenvalues of J are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 1 - 1/c \geq \lambda_{r+1} \geq \cdots \geq \lambda_n$. Then π admits a mixture decomposition

$$\pi(x) = \sum_{h \in S} q_h \overline{\pi}_h(x)$$

where $q_h \geq 0, \sum_h q_h = 1$,

$$|S| \leq \left(n\lambda_1 + \sqrt{n}r\sqrt{\lambda_1} + \sqrt{\lambda_1 n r \ln(\sqrt{\lambda_r^{-1/2}} + \sqrt{n})}\right)^{O(r)},$$

and each component $\overline{\pi}_h$ satisfies the Poincaré inequality for the continuous-time Glauber dynamics with constant

$$\exp\left[O\left(\int_0^{T_0} c e^{cs \operatorname{Tr}(J_-)} ds\right)\right]$$

where $-J_-$ is the negative definite part of J and $T_0 = 1 - 1/c - \lambda_{\min}(J)$. Furthermore, π satisfies the higher-order spectral gap inequality

$$\lambda_{|S|+1}(-\mathscr{L}) \geq \exp\left[-O\left(\int_0^{T_0} c e^{cs \operatorname{Tr}(J_-)} ds\right)\right]$$

where $\mathscr{L}$ is the generator of the continuous-time Glauber dynamics for π.

Proof By Theorem 33, there exists a distribution π_2 such that

$$\pi_2(x) = \sum_{h \in S} p_h \tilde{\pi}_h(x)$$

where $\tilde{\pi}_H$ is the probability measure satisfying

$$\tilde{\pi}_H(x) \propto \exp\left(\langle x, \tilde{J}x\rangle + \langle H, x\rangle\right)$$

and $\tilde{J} = \sum_{i=r+1}^n \lambda_i \phi_i \phi_i^\intercal$ where $\lambda_1 \geq \cdots \geq \lambda_r > 1 - 1/c \geq \lambda_{r+1} \geq \cdots \geq \lambda_n$ are the eigenvalues of J and ϕ_i are the eigenvector corresponds to λ_i.

Furthermore,

$$d\pi_2/d\pi(x) \in [1/e^3, e^3]$$

for all x, and

$$|S| \leq \left(n\lambda_1 + \sqrt{n}r\sqrt{\lambda_r} + \sqrt{\lambda_1 n r \ln(\lambda_r^{-1/2} + \sqrt{n})}\right)^{O(r)}.$$

It follows that

$$\pi(x) = \frac{\pi(x)}{\pi_2(x)}\pi_2(x) = \sum_{h \in S} p_h \frac{\pi(x)}{\pi_2(x)}\tilde{\pi}_h(x) = \sum_{h \in S} p_h \left(\sum_x \frac{\pi(x)}{\pi_2(x)}\tilde{\pi}_h(x)\right) \frac{\frac{\pi(x)}{\pi_2(x)}\tilde{\pi}_h(x)}{\sum_x \frac{\pi(x)}{\pi_2(x)}\tilde{\pi}_h(x)}.$$

By Theorem 103 of (Anari et al., 2024b, Appendix B), the measure $\tilde{\pi}_h$ satisfies the Poincaré inequality for the continuous-time Glauber dynamics with constant

$$\exp\left(\int_0^{T_0} c e^{cs \operatorname{Tr}(J_-)} ds\right)$$

where $-J_-$ is the negative definite part of $\tilde{J}$ (so J_- is positive definite), which is the same as the negative definite part of J and $T_0 = \lambda_{\max}(\tilde{J}) - \lambda_{\min}(\tilde{J}) \le 1 - 1/c - \lambda_{\min}(J)$. The probability density

$$\overline{\pi}_h = \frac{\frac{\pi(x)}{\pi_2(x)} \tilde{\pi}_h(x)}{\sum_x \frac{\pi(x)}{\pi_2(x)} \tilde{\pi}_h(x)}$$

has a density with respect to $\tilde{\pi}$ which is upper and lower bounded by a constant (since this is true for $\frac{\pi(x)}{\pi_2(x)}$, and using this we know the denominator is also upper and lower bounded by a constant), so by Proposition 4.2.7 of Bakry et al. (2014) we see that $\overline{\pi}_h$ satisfies the Poincaré inequality as stated in the theorem.

Finally, the higher-order spectral gap inequality follows by applying Lemma 10. ■

Remark 38 (Generalizations such as Potts model) *It is straightforward to extend the above result to similar models such as the Potts model under the same conditions on J. In the following theorem, for illustration, we prove an analogous result for the well-studied mean-field Potts model on q coclors. When $q = 2$, this is the same as the Curie-Weiss model, which is discussed more in Section F.1. When $q > 2$, the model has a qualitatively different behavior from the Curie-Weiss model which is called a "first-order phase transition". The implication for the Glauber dynamics is that in certain regimes, the dynamics with worst-case initialization can become trapped in a metastable region which has negligible mass under the true Gibbs measure; as our techniques rigorously show, this is not a problem for data-based initialization. See e.g. Cuff et al. (2012) for extensive discussion of this model.*

Theorem 39 (Application to Mean-field Potts Model) *Suppose that $q \ge 2$ and π is the distribution on $[q]^n$ with probability mass function*

$$\pi(x) \propto \exp\left(\frac{\beta}{2n} \sum_{i,j} 1(x_i = x_j)\right).$$

Then π admits a mixture decomposition

$$\pi(x) = \sum_{h \in S} q_h \overline{\pi}_h(x)$$

where $q_h \ge 0, \sum_h q_h = 1$,

$$|S| \le \left(n\beta + \sqrt{n} r \sqrt{\beta} + \sqrt{\beta n r \ln(n)}\right)^{O(r)},$$

and each component $\overline{\pi}_h$ satisfies the Poincaré inequality for the continuous-time Glauber dynamics with constant $O(1)$. Furthermore, π satisfies the higher-order spectral gap inequality

$$\lambda_{|S|+1}(-\mathscr{L}) = \Omega(1)$$

where $\mathscr{L}$ is the generator of the continuous-time Glauber dynamics for π.

Proof If $\beta < 1/2$ then the distribution satisfies the classical Dobrushin's condition (see e.g. path coupling discussion in Levin and Peres (2017)) so the conclusion is trivial. Otherwise, we can assume $\beta \geq 1/2$.

Note that we can rewrite

$$1(x_i = x_j) = \langle e_i, e_j \rangle$$

so we can view π as a distribution with spins in q dimensions. Therefore, we can apply Theorem 33 to get a multiplicative approximation to the distribution as a mixture of product measures, and then by applying the change of measure argument from the proof of Theorem 37 proves the result. ■

E.3. Fast mixing from data-based initialization & learning

As an application of our techniques, we now show how our ideas lead to provably learn a natural class of Ising models which were not covered by previous results. The algorithm is the natural combination of pseudolikelihood estimation with a warm-started Gibbs sampler, and hence is not much harder to implement than pseudolikelihood itself.

Pseudolikelihood estimator. It is helpful to first recall the definition of the pseudolikelihood estimator Besag (1975) in the case of the Ising model. Recall that in the Ising model $\pi(x) \propto \exp(\langle x, Jx \rangle/2 + \langle h, x \rangle)$,

$$\pi(X_i = x_i \mid X_{\sim i}) \propto \exp(J_{i,\sim i} \cdot X_{\sim i} x_i + h_i x_i)$$

and taking into account the normalizing constant, we have

$$\pi(X_i = x_i \mid X_{\sim i}) = \frac{1}{1 + \exp(-2J_{i,\sim i} \cdot X_{\sim i} x_i - 2h_i x_i)}.$$

Therefore

$$\ln \pi(X_i = x_i \mid X_{\sim i}) = -\ln\left(1 + \exp(-2J_{i,\sim i} \cdot X_{\sim i} x_i - 2h_i x_i)\right)$$

and this is what is known as a logistic regression model. (See e.g. McCullagh and Nelder (2019).) The pseudolikelihood of an outcome $x \in \{\pm 1\}^n$ under the model π is the sum of these conditional likelihooods over the choice of index i, i.e.

$$\sum_{i=1}^{n} \ln \pi(X_i = x_i \mid X_{\sim i} = x_{\sim i})$$

and the pseudolikelihood estimator is given by optimizing this objective averaged over a dataset.

Theorem 40 *Fix $c \in [2, \infty)$. Suppose that π is a probability measure on $\{\pm 1\}^n$ satisfying*

$$\pi(x) \propto \exp\left((1/2)\langle x, Jx \rangle + \langle h, x \rangle\right)$$

for some symmetric interaction matrix J and external field vector h. Suppose that J has r eigenvalues greater than $1 - 1/c$, including its top eigenvalue λ_1, and the rest are at most $1 - 1/c$.

Suppose $X^{(0)}, \dots, X^{(m)} \sim \pi$ are i.i.d. samples; let $\hat{\mathbb{E}}$ denote the corresponding empirical expectation over the empirical measure $\hat{\pi}$, i.e. $\hat{\mathbb{E}}$ is the average over the sample set. Let

$Y^{(0)}, \ldots, Y^{(m)} \sim \pi$ be another set of iid samples at denote their empirical measure by $\hat{\pi}_2$. Furthermore, suppose that $R \geq 0$ is such that

$$\max_i \sum_j |J_{ij}| + |h_i| \leq R. \tag{24}$$

Define the constrained pseudolikelihood estimator by

$$\hat{\rho} = \arg\max_{\rho \in \mathcal{P}_R} \sum_{i=1}^n \hat{\mathbb{E}}_X \ln \rho(X \mid X_{\sim i})$$

where $\mathcal{P}_R$ is a set consisting of Ising models (i.e. measures on $\{\pm 1\}^n$ with quadratic log-likelihood) with parameters satisfying the convex constraint (24) (so in particular, it can be optimized in polynomial time).

With probability at least $1-\delta$ over the randomness of the sample set, for any

$$t \geq T = \exp\left[O\left(\int_0^{T_0} ce^{cs\,\mathrm{Tr}(J_-)} ds\right)\right]\left(2Rn + \ln\left(\frac{4k\ln(k/\delta)}{m}\right)\right)$$

where $T_0 = 1 - 1/c - \lambda_{\min}(J)$ we have

$$\mathrm{TV}(\mathcal{L}(X_t^{\hat{\pi}_2} \mid \hat{\pi}, \hat{\pi}_2), \pi) \leq \sqrt{tRn\sqrt{\ln(2n/\delta)/m}} + 4\sqrt{k\ln(k/\delta)/m}.$$

where

$$k = O\left(\lambda_1 n + \sqrt{\lambda_1 n r} + \sqrt{\lambda_1 n r \ln(\sqrt{n})}\right)^{O(r)} = \tilde{O}(\lambda_1 n)^{O(r)}$$

and $X_t^{\hat{\pi}_2}$ is the output of the $\hat{\rho}$-Glauber diffusion initialized at $\hat{\pi}_2$ and run for time t.

After the proof of the result, we give a number of remarks to elaborate on its meaning and significance.

Proof [Proof of Theorem 40] By the standard symmetrization and bounded differences concentration argument from statistical learning theory (see e.g. Bartlett and Mendelson (2002)), we know that with probability at least $1-\delta$

$$\sum_{i=1}^n \mathbb{E}_{X\sim\pi} \ln \hat{\rho}(X \mid X_{\sim i}) \geq \sum_{i=1}^n \mathbb{E}_{X\sim\pi} \ln \pi(X \mid X_{\sim i}) - O\left(\mathcal{R}_m + Rn\sqrt{\ln(2/\delta)/m}\right) \tag{25}$$

where

$$\mathcal{R}_m = \mathbb{E}_{X^{(0)},\ldots,X^{(m)},\epsilon} \sup_{J,h} \frac{1}{mn} \sum_{a=1}^m \epsilon_a \sum_{i=1}^n \ell\left(J_{i,\sim i} \cdot X_{\sim i}^{(a)} X_i^{(a)} + h_i X_i^{(a)}\right)$$

is the Rademacher complexity of the loss class and $\ell(z) = \ln(1 + e^{2z})$ is the logistic loss, and we used that the loss in any example is bounded by $O(R)$ because the logistic loss is lipschitz and because of Holder's inequality. Applying Talagrand's contraction principle and Holder's inequality yields that

$$\mathcal{R}_m = O(Rn\sqrt{\ln(n)/m}),$$

see e.g. the proof of Theorem 47 in Anari et al. (2024a) for the details.

Using the Rademacher complexity bound and rearranging (25) yields

$$\sum_{i=1}^{n} \mathbb{E}\,\mathrm{KL}(\pi(\cdot \mid X_{\sim i}), \hat{\rho}(\cdot \mid X_{\sim i})) = \sum_{i=1}^{n} \mathbb{E}_{X\sim\pi} \ln \frac{\pi(X \mid X_{\sim i})}{\hat{\rho}(X \mid X_{\sim i})} = O\left(Rn\sqrt{\ln(2n/\delta)/m}\right)$$

and a bound on the left hand side is exactly what we need to apply Lemma 20, which in turn lets us apply Theorem 21 with $\varepsilon = O\left(Rn\sqrt{\ln(2n/\delta)/m}\right)$. The result is that for the empirical distribution $\hat{\pi}_2$ formed from m samples, with probability at least $1-\delta$,

$$\mathrm{TV}(\mathcal{L}((X_t^{\hat{\pi}_2})_{0\le t\le T} \mid \hat{\pi}, \hat{\pi}_2), \mathcal{L}((\tilde{X}_t^{\hat{\pi}_2})_{0\le t\le T} \mid \hat{\pi}, \hat{\pi}_2)) \le \sqrt{T\varepsilon} + \frac{2\ln(1/\delta)}{m}$$

where $(X_t^{\hat{\pi}_2})_t$ is the law of the $\hat{\rho}$-Glauber dynamics initialized from $\hat{\pi}_2$, and $\tilde{X}_t^{\hat{\pi}_2}$ is the law of the π-Glauber dynamics with the same initialization. Note that by the union bound, the two events we are using hold together with probability at least $1-2\delta$.

Recall from Theorem 37 that there exists

$$k = O\left(n\lambda_1 + \sqrt{n}r\sqrt{\lambda_1} + \sqrt{\lambda_1 nr\ln(\lambda_r^{-1/2} + \sqrt{n})}\right)^{O(r)}$$

such that

$$\lambda_k(-\mathscr{L}) \ge \exp\left[-O\left(\int_0^{T_0} ce^{c\,\mathrm{Tr}(J_-)}ds\right)\right]$$

where $\mathscr{L}$ is the generator of the Glauber dynamics.

It then follows our main result, Theorem 9, applied with $t_0 = 0$ we have that with probability at least $1-\delta$ over the randomness of the sample,

$$\mathrm{TV}(\mathcal{L}(\tilde{X}_T^{\pi} \mid \hat{\pi}, \hat{\pi}_2), \pi) \le \varepsilon_2$$

provided $t \ge T = (1/\lambda_k(-\mathscr{L}))\ln\left(\frac{4e^{2Rn}}{\varepsilon_2^2}\right)$ and $m = \Omega(\varepsilon_2^{-2}k\ln(k/\delta))$ where e^{2Rn} is an upper bound on the initial χ^2-distance between $\hat{\pi}_2$ and π. We take $\varepsilon_2 = \sqrt{k\ln(k/\delta)/m}$.

Thus, by the union bound, the data processing inequality, and the triangle inequality for total variation distance, we have with probability at least $1-3\delta$ that

$$\begin{aligned}
&\mathrm{TV}(\mathcal{L}((X_t^{\hat{\pi}_2} \mid \hat{\pi}, \hat{\pi}_2), \pi) \\
&\le \sqrt{t\varepsilon} + \frac{2\ln(1/\delta)}{m} + \varepsilon_2 \\
&= \sqrt{tRn\sqrt{\ln(2n/\delta)/m}} + \frac{2\ln(1/\delta)}{m} + \sqrt{k\ln(k/\delta)/m}.
\end{aligned}$$

The bound is trivial unless $m = \Omega(k\ln(k/\delta))$, so the middle term can be dominated by the right one by adjusting the multiplicative factor in front. ■

Remark 41 (Separation from previous results) *For example, consider the case where J is positive semidefinite and rank $r = O(1)$. Then for*

$$t = O(Rn + \ln(4/\varepsilon_{\text{TV}}^2))$$

and

$$m = \Omega(\max\{\varepsilon_{\text{TV}}^{-2} k \ln(k/\delta), \varepsilon_{\text{TV}}^{-4} (tRn)^2 \ln(2n/\delta)\}),$$

we are guaranteed that

$$\text{TV}(\mathcal{L}((X_t^{\hat{\pi}} \mid \hat{\pi}), \pi) \leq \varepsilon_{\text{TV}}.$$

Note that both k and m, the number of samples, are polynomial in all parameters as long as $r = O(1)$. In contrast, previous bounds for learning Ising models with computationally efficient algorithms have exponential dependence on R (see e.g. Lokhov et al. (2018); Gaitonde and Mossel (2024); Klivans and Meka (2017); Wu et al. (2019)), and in some simple examples of such Ising models (e.g. in Hopfield networks with finitely many memories Talagrand (2010)) R is polynomial in n. R can also be large simply if the external field of the model is large. Information-theoretically, such models can be learned with polynomial sample complexity without any dependence on R Devroye et al. (2020).

A key reason why previous analyses fail is that these works all seek to first show that the parameters are identifiable, and then argue that recovering the ground truth parameters also implies recovery in TV distance. However, there are models in our class which have very tiny TV distance despite having well-separated parameters. For example, consider the class of $n+1$ ferromagnetic Ising models with uniform edge weight R/n on a graph G, where G is either a clique or a clique with an unknown edge removed. Following the proof of Theorem 1 in Santhanam and Wainwright (2012), we know that the KL diveregence between any two models in this class is $e^{-\Omega(R)}$ (Lemma 2 of Santhanam and Wainwright (2012)), so by Fano's lemma $e^{\Omega(R)} \ln(n)$ samples from the model are needed to identify what the underlying graph G is. So for example, if we take $R = n$, we will need $e^{\Omega(n)}$ many samples to determine the underlying graph, or to determine the parameters of the model within ℓ_1 error $o(1)$. Nevertheless, our theorem shows that it is possible to learn the model to $o(1)$ error in TV from only $poly(n)$ samples.

Remark 42 *Continuing the discussion from the previous remark, we do not currently know if pseudolikelihood estimation alone can possibly achieve a similar guarantee. In general, the behavior of pseudolikelihood estimation is closely related to restricted versions of the log-Sobolev and Poincaré constants Koehler et al. (2022a), where the restriction is to functions related to the relative density between models in the class of distributions being fit.*

Based on an analogy to the lower bound results of Koehler et al. (2022a) for score matching, we expect that in variants of this setting where we introduce a small additional term in the log-likelihood that this approach will still work (this can be proved by a perturbation argument) but pseudolikelihood will fail.

Remark 43 *Theorem 40 also implies a fast-amortized way to produce many samples from the low-rank Ising model. For example, we can use Koehler et al. (2022b)'s algorithm to produce $y_1, \ldots, y_m$ i.i.d. sampled from a distribution ν with $\text{TV}(\nu, \pi) \leq \varepsilon_{\text{TV}}/16$, then run the Glauber dynamics initialized at a randomly chosen sample y_i with $i \sim \text{Uniform}([m])$. The output of Glauber is a distribution $\tilde{\pi} \equiv \tilde{\pi}_{\{y_i\}_{i\in[m]},(J,h)}$ parameterized by the samples $y_1, \cdots, y_m$ and the parameters (J, h)*

of the input Ising model, with the randomness coming from the Glauber updates. By Theorem 16, for any $\delta > 0$, with appropriate choice of m, we can ensure that $\mathbb{P}_{\{y_i\}_{i\in[m]}}[d_{TV}(\tilde{\pi}, \pi) \le \epsilon_{TV}] \ge 1 - \delta$. The amortized runtime (say, in the setting of Remark 41) is $O(Rn + \ln(\frac{1}{\varepsilon_{\text{TV}}}))$ since we can ignore the cost of producing the initial m samples i.e. the runtime cost to produce M samples is

$$T_{\text{init}} + O\left(MRn + \ln\left(\frac{1}{\varepsilon_{\text{TV}}}\right)\right).$$

We note that conditioned on $\{y_i\}_{i\in[m]}$, the M samples are i.i.d. samples from $\tilde{\pi}$ and can be used in various downstream tasks.

Appendix F. Examples with non-sample initialization

We have shown that MCMC approximately samples the target distribution in a polynomial time, as long as it is started from an initialization which approximately satisfies the eigenfunction balance condition (see Definition 12). Data-based initialization is one way to find such an initialization, but it is not the only way. We discuss this more in this section.

F.1. Example: Curie-Weiss

In this section, we show that the spectral properties of the Glauber dynamics in the low-temperature Curie-Weiss model ensure that a wide range of weakly balanced initializations mix to stationarity in polynomial time. This model is a good pedagogical example to consider, in particular because it has received a lot of previous in the rigorous literature on metastability of the Glauber dynamics.

As a reminder, it is known that the spectral gap of the Glauber dynamics is exponentially small in n for any fixed inverse temperature $\beta > 1$, while the mixing is polynomial time for $\beta \le 1$. Thus, the "low temperature" regime $\beta > 1$ is the one of interest for metastability. Informally, the bottleneck in the dynamics after the phase transition occurs because typical samples from the measure have either significant positive or negative magnetization, whereas the measure has very little support in the middle near zero magnetization. Our general result Theorem 37 applies to this model, and guarantees that there are $O_\beta(1)$ many exponentially small eigenvalues — now, building on results from the literature, we will show there is in fact only one bad nontrivial eigenvalue.

Levin et al. (2010) (see also Ding et al. (2009a)) proved via a coupling argument that a "censored" version of the Glauber dynamics for the low-temperature Curie-Weiss model mixes in $O_\beta(n \log n)$ steps; this dynamics is restricted to spins with positive magnetization. As a warmup, we observe this implies a restricted version of the Poincaré inequality for even functions:

Lemma 44 *Suppose n is odd, $\beta > 1$, and consider the Curie-Weiss model on $\{\pm 1\}^n$*

$$p(x) \propto \exp\left(\frac{\beta}{2n}\left(\sum_i x_i\right)^2\right).$$

Let f be an even function in the sense that $f(x) = f(-x)$ for all $x \in \{\pm 1\}^n$. Then

$$\operatorname{Var}(f) \le C(\beta)\mathcal{E}(f, f)$$

where $C(\beta) > 0$ is a constant independent of n.

Proof We show that this follows from the result of Levin et al. (2010). First, we recall that they proved $O_\beta(n \log n)$ time mixing of the following "censored" dynamics on the restriction of p to $\{x : \sum_i x_i > 0\}$, which has the following behavior in one step:

1. Given $x \sim \{\pm 1\}^n$, generate a proposal y via the standard Glauber dynamics.
2. If $\sum_i y_i > 0$, transition to y. If $\sum_i y_i < 0$, transition to $-y$.

By the well-known connection between mixing time and spectral gap, the mixing time result of Levin et al. (2010) in particular implies that the spectral gap of the generator of the continuous-time censored dynamics is $\Omega_\beta(1)$. Next, observe that if f is even, then $\mathrm{Var}(f) = \mathrm{Var}(f \mid \sum_i X_i > 0)$ by symmetry, and similarly observe that the Dirichlet form of the Glauber dynamics on the whole measure is equal to the Dirichlet form of the censored dynamics when restricted to even f. Hence, the conclusion follows immediately. ■

This is not enough to bound the higher-order spectral gap since it says nothing about odd functions; we now give a more sophisticated argument which does successfully bound the higher-order spectral gap.

Proposition 45 *Suppose n is odd, $\beta > 1$, and consider the Curie-Weiss model on $\{\pm 1\}^n$,*

$$p(x) \propto \exp\left(\frac{\beta}{2n}\left(\sum_i x_i\right)^2\right).$$

Let $\mathcal{L}$ be the generator of the continuous-time Glauber dynamics. Then $\lambda_3(-\mathcal{L}) = \Omega_\beta(1/n^3)$.

Proof Let f be an arbitrary function and observe by the law of total variance that

$$\mathrm{Var}(f) = \mathbb{E}\,\mathrm{Var}(f \mid \mathrm{sign}(\sum_i X_i)) + \mathrm{Var}(\mathbb{E}[f \mid \mathrm{sign}(\sum_i X_i)]).$$

Therefore,

$$\mathrm{Var}(f) \le C(\beta)[\mathcal{E}^+(f,f)/2 + \mathcal{E}^-(f,f)/2] + \mathrm{Var}(\mathbb{E}[f \mid \mathrm{sign}(\sum_i X_i)])$$

where $\mathcal{E}^+$ is the Dirichlet form of the censored dynamics from Levin et al. (2010), and $\mathcal{E}^-(f,f) = \mathcal{E}^+(x \mapsto f(-x), x \mapsto f(-x))$ is the Dirichlet form of the analogous dynamics on the set of spins with negative magnetization.

Next, we perform a comparison of Dirichlet forms, i.e. a routing argument, between the censored dynamics and the original Glauber dynamics. See Chapter 13 of Levin and Peres (2017) for more background on this technique. The censored dynamics has additional transitions not present in the Glauber dynamics, which are of the form $x \mapsto -y$ for x, y neighbors such that $\sum_i x_i = 1$ and $\sum_i y_i = -1$. Note that such a pair $(x, -y)$ will differ at all but one coordinate. We route such a transition to a path in the hypercube by replacing the coordinates of x by those of y one-by-one, left-to-right, skipping the one coordinate which is the same. Next, we need to bound the congestion of the path, i.e. how many paths can be routed over the same edge —- this is at most $2n+2$, because for a routing path going from x to $-y$ to cross over an edge (x', x'') where k is such that $x'_k \neq x''_k$, it must be that x is at most hamming distance 1 from the bitstring given by flipping the sign of all the coordinates of x' to the left of k, and analogously $-y$ is at most distance 1 from the bitstring given by flipping the coordinates of x'' to the right of k. Furthermore, for such a tuple $x, -y, x', x''$ we will have the following inequality of conductances:

$$\frac{p(x)p(-y)}{p(x)+p(-y)} \le \exp(4\beta)\frac{p(x')p(x'')}{p(x')+p(x'')}$$

because $p(x) = \min_{z \in \{\pm 1\}^n} p(z) \le p(x')$ and because $\frac{p(z)}{p(z)+p(z'')} \in [e^{-2\beta}, e^{2\beta}]$ for any pair z, z' at hamming distance one, by expanding the definition of p.

From this routing, it follows that

$$\mathcal{E}^+(f,f) \le C'(\beta)n^3\mathcal{E}(f,f)$$

for some constant $C'(\beta) > 0$ independent of n, and likewise for $\mathcal{E}^-$. Thus,

$$\mathrm{Var}(f) \le C''(\beta)n^3\mathcal{E}(f,f) + \mathrm{Var}(\mathbb{E}[f \mid \mathrm{sign}(\sum_i X_i)])$$

which proves the claim by the variational characterization of eigenvalues. ■

As a consequence, we can prove rapid mixing of Glauber dynamics from any initial distribution such that $\mathbb{E}_\rho f_2 = 0$. Furthermore, we can show some basic structural properties of the eigenfunction.

Proposition 46 *Fix $\beta > 1$. Require n is odd and consider the Curie-Weiss model on $\{\pm 1\}^n$*

$$p(x) \propto \exp\left(\frac{\beta}{2n}\left(\sum_i x_i\right)^2\right).$$

Let $\mathcal{L}$ be the generator of the continuous-time Glauber dynamics, and let f_2 to be the second eigenvector from the bottom of the spectrum of $\mathcal{L}$. Then f_2 depends on x only through $\sum_i x_i$ and, at least for sufficiently large n with respect to β, it is odd, i.e. $f_2(-x) = -f_2(x)$. Furthermore, the Glauber dynamics initialized from ρ satisfying $\mathbb{E}_\rho f_2 = 0$ will achieve total variation distance $\delta > 0$ from stationarity in time $O_\beta(n^3(n + \log(1/\delta)))$.

Proof The mixing statement follows from the previous proposition and our general results. The properties of f_2 follow because the dynamics respect the symmetries of permuting coordinates and reversing the role of $+$ and $-$, so applying these symmetries must send the eigenvector f_2 to a multiple of itself. Note that for all n larger than a constant, f_2 cannot be an even function (i.e. a positive multiple of itself under interchange of $+$ and $-$) because of Lemma 44, so it must be odd. ■

Remark 47 *We compare this result with what can be achieved using a different argument closely related to the Poincaré inequality for even functions. Using the main result of Levin et al. (2010) and making a coupling argument using symmetry, one could show that the Glauber dynamics mix in $O_\beta(n \log n)$ time from any* even *initialization, i.e. one such that $\rho(x) = \rho(-x)$. However, this is stricter than the minimal requirement that $\mathbb{E}_\rho f_2 = 0$ — the former corresponds to 2^{n-1} linearly independent constraints, whereas the latter is only a single linear constraint.*

F.2. Existence of perfectly balanced initializations with small support

The following elementary argument tells us that for any finite Markov chain with a higher-order spectral gap, there is an initialization with small support such that the dynamics rapidly mix to stationarity.

Proposition 48 *Suppose that $\mathcal{L}$ is the generator of a Markov semigroup on a finite state space Ω with unique stationary measure π and let $k \geq 1$ be arbitrary such that $\lambda_k(-\mathcal{L}) > 0$. Let $f_1, \ldots, f_k$ be the first k eigenfunctions of $\mathcal{L}$, starting from the bottom of the spectrum. There exists a distribution $\tilde{\rho}$ supported on at most $k-1$ points of Ω, such that $\mathbb{E}_{\tilde{\rho}}[f_i] = 0$ for all $1 < i < k$, and as a consequence $d_{TV}(\tilde{\rho}e^{tL}, \pi) \leq \delta$ provided that $t \geq (1/\lambda_k(-\mathcal{L}))\log(|\Omega|/\delta)$.*

Proof Consider the polytope of distributions ρ satisfying the usual constraints that $\sum_x \rho(x) = 1$, $\rho(x) \geq 0$ for all x, along with the additional constraint that $\mathbb{E}_\rho[f_i] = 0$ for all i from 2 to $k-1$. This polytope is nonempty because it contains the stationary distribution of the Markov chain. The existence then follows from complementary slackness — the polytope contains an extreme point $\tilde{\rho}$, and because of dimension counting, for all but $k-1$ many points x we must have that $\tilde{\rho}(x) = 0$. ■

Remark 49 *The previous result can be generalized to the case where the stationary distribution is not unique—in this case, the expected values of the eigenfunctions with eigenvalue zero should be selected to match the desired stationary measure.*

Remark 50 *In some cases it may be possible to explicitly solve for such a distribution. For example, in the Curie-Weiss model, the symmetrized all-ones initialization $(1/2)\delta_{\mathbf{1}} + (1/2)\delta_{-\mathbf{1}}$ is an explicit example of such a distribution $\tilde{\rho}$.*

Appendix G. Rényi divergence

The Rényi divergence, which generalizes the more well-known KL divergence, is a useful technical tool in the analysis of the Langevin diffusion, see e.g., Vempala and Wibisono (2019). The Rényi divergence of order $q \in (1, \infty)$ of μ from π is defined to be

$$\mathcal{R}_q(\mu||\pi) = \frac{1}{q-1}\ln \mathbb{E}_\pi\left[\left(\frac{d\mu}{d\pi}\right)^q\right] = \frac{1}{q-1}\ln \mathbb{E}_\mu\left[\left(\frac{d\mu}{d\pi}\right)^{q-1}\right].$$

The limit $\mathcal{R}_q$ as $q \to 1^+$ is the Kullback-Leibler divergence $\mathcal{D}_{\mathrm{KL}}(\mu||\pi) = \int \log\left(\frac{d\mu}{d\pi}\right) d\mu$, thus we write $\mathcal{R}_1(\cdot) = \mathcal{D}_{\mathrm{KL}}(\cdot)$. Rényi divergence increases as q increases, i.e. $\mathcal{R}_q \leq \mathcal{R}_{q'}$ for $1 \leq q \leq q'$. We note some basic facts about Rényi divergence.

Lemma 51 (Weak triangle inequality, (Mironov, 2017, Proposition 11)) *For $q > 1$ and any measure ν absolutely continuous with respect to measure μ, for any $a, b > 1$ such that $\frac{1}{a} + \frac{1}{b} = 1$,*

$$\mathcal{R}_q(\nu||\mu) \leq \frac{aq-1}{a(q-1)}\mathcal{R}_{aq}(\nu||\nu') + \mathcal{R}_{b(q-1)+1}(\nu'||\mu).$$

Lemma 52 (Weak convexity of Rényi entropy) *For $q > 1$, if μ is a convex combination of μ_i, i.e. $\mu = \sum_{i=1}^k p_i \mu_i$ for $p_i > 0$ such that $\sum_{i=1}^k p_i = 1$, and $\nu \ll \mu$, then*

$$\mathbb{E}_\nu \left[\left(\frac{d\nu}{d\mu} \right)^{q-1} \right] \le \sum_{i=1}^k p_i \mathbb{E}_\nu \left[\left(\frac{d\nu}{d\mu_i} \right)^{q-1} \right].$$

Consequently, $\mathcal{R}_q(\nu||\mu) \le \max_i \mathcal{R}_q(\nu||\mu_i)$. If instead $\mu \ll \nu$, we have $\mathcal{R}_q(\mu||\nu) \le \max_i \mathcal{R}_q(\mu_i||\nu)$.

Proof By Hölder's inequality, we have ν-almost everywhere that

$$\left(\sum_{i=1}^k p_i \frac{d\mu_i}{d\nu} \right)^{q-1} \left(\sum_{i=1}^k p_i \left(\frac{d\nu}{d\mu_i} \right)^{q-1} \right) \ge \left(\sum_{i=1}^k p_i \right)^q = 1.$$

Thus, we have ν-almost everywhere that

$$\left(\frac{d\nu}{d\mu} \right)^{q-1} \le \sum_{i=1}^k p_i \left(\frac{d\nu}{d\mu_i} \right)^{q-1}.$$

Taking expectation in ν gives the first statement. For the second statement, by bounding the convex combinations above with the maximal term,

$$\mathcal{R}_q(\nu||\mu) = \frac{\ln \mathbb{E}_\nu[(\frac{d\nu}{d\mu})^{q-1}]}{q-1} \le \frac{\ln(\max_i \mathbb{E}_\nu[(\frac{d\nu}{d\mu_i})^{q-1}])}{q-1} = \max_i \mathcal{R}_q(\nu||\mu_i).$$

For the final statement, by Jensen's inequality and convexity of x^q,

$$\ln \mathbb{E}_\nu \left[\left(\frac{d\mu}{d\nu} \right)^q \right] \le \ln \sum_{i=1}^k p_i \mathbb{E}_\nu \left[\left(\frac{d\mu_i}{d\nu} \right)^q \right] \le \ln \max_i \mathbb{E}_\nu \left[\left(\frac{d\mu_i}{d\nu} \right)^q \right] = \max_i \mathcal{R}_q(\mu_i||\nu).$$

■

Appendix H. Consequences of functional inequalities

In this section, we record properties of distribution satisfying a Poincaré or log-Sobolev inequalities, which would be useful in our analysis of the Langevin diffusion (see section D.1). Poincaré and log-Sobolev inequalities imply sub-exponential and sub-gaussian concentration of Lipschitz functions, respectively.

Lemma 53 ((Bakry et al., 2014, Pr. 4.4.2)) *Suppose that μ satisfies a Poincaré inequality with constant C_{P}. Let f be a 1-Lipschitz function. Then for any $t \in \left[0, \frac{2}{\sqrt{C_{\mathrm{P}}}}\right)$,*

$$\mathbb{E}_\mu e^{tf} \le \frac{2 + t\sqrt{C_{\mathrm{P}}}}{2 - t\sqrt{C_{\mathrm{P}}}} e^{t\mathbb{E}_\mu f}.$$

Lemma 54 ((Bakry et al., 2014, Pr. 5.4.1)) *Suppose that μ satisfies a log-Sobolev inequality with constant C_{LS}. Let f be a 1-Lipschitz function. Then*

1. *(Sub-exponential concentration) For any $t \in \mathbb{R}$,*

$$\mathbb{E}_\mu e^{tf} \le e^{t\mathbb{E}_\mu f + \frac{C_{\mathrm{LS}} t^2}{2}}.$$

2. *(Sub-gaussian concentration) For any $t \in \left[0, \frac{1}{C_{\mathrm{LS}}}\right)$,*

$$\mathbb{E}_\mu e^{\frac{tf^2}{2}} \le \frac{1}{\sqrt{1 - C_{\mathrm{LS}} t}} \exp\left[\frac{t}{2(1 - C_{\mathrm{LS}} t)} (\mathbb{E}_\mu f)^2\right].$$

From this, we can get concentration around the mean. First, we note we can bound the variance by the Poincaré constant: if μ satisfies a Poincaré inequality with constant C_{P},

$$\mathbb{E}_\mu[\|x - \mathbb{E}_\mu x\|^2] \le \sum_{i=1}^d \mathrm{Var}_\mu(x_i) \le C_{\mathrm{P}} d. \tag{26}$$

Lemma 55 *If μ satisfies a Poincaré inequality with constant C_{P}, then*

$$\mathbb{P}\left(\|x - \mathbb{E}_\mu x\| \ge \sqrt{C_{\mathrm{P}}}(\sqrt{d} + u)\right) \le 3e^{-u}.$$

and

$$\forall p \ge 1: \quad \mathbb{E}_{x\sim\mu}[\|x - \mathbb{E}_\mu x\|^p]^{1/p} = O(p\sqrt{C_{\mathrm{P}} d}). \tag{27}$$

Proof By Markov's inequality and Lemma 53, for $t \in \left[0, \frac{2}{\sqrt{C_{\mathrm{P}}}}\right)$,

$$\begin{aligned}\mathbb{P}\left(\|x - \mathbb{E}x\| \ge \sqrt{C_{\mathrm{P}}}(\sqrt{d} + u)\right) &= \mathbb{P}\left(e^{t\|x - \mathbb{E}x\|} \ge e^{t\sqrt{C_{\mathrm{P}}}(\sqrt{d}+u)}\right) \\ &\le \frac{\mathbb{E}e^{t\|x-\mathbb{E}x\|}}{e^{t\sqrt{C_{\mathrm{P}}}(\sqrt{d}+u)}} \le e^{t\mathbb{E}\|x-\mathbb{E}x\|} \frac{2 + t\sqrt{C_{\mathrm{P}}}}{2 - t\sqrt{C_{\mathrm{P}}}} e^{-t\sqrt{C_{\mathrm{P}}}(\sqrt{d}+u)}.\end{aligned}$$

Eq. (26) gives $\mathbb{E}\|x - \mathbb{E}x\| \le (\mathbb{E}\|x - \mathbb{E}x\|^2)^{1/2} \le \sqrt{C_{\mathrm{P}} d}$. Substituting $t = \frac{1}{\sqrt{C_{\mathrm{P}}}}$ gives the first inequality. From the above, we also have for $t = \frac{1}{\sqrt{C_{\mathrm{P}} d}}$

$$\mathbb{E}e^{t\|x-\mathbb{E}x\|} \le e^{t\mathbb{E}\|x-\mathbb{E}x\|} \frac{2 + t\sqrt{C_{\mathrm{P}}}}{2 - t\sqrt{C_{\mathrm{P}}}} \le 3e$$

thus the moment bound follows from (Vershynin, 2018, Proposition 2.7). ∎

Lemma 56 (Upper bound on partition function) *Suppose that $\pi \propto e^{-f}$ satisfies a Poincaré inequality with constant C_{P}. Then*

$$\int_{\mathbb{R}^d} e^{-f(x)}\, dx \le \frac{e}{\sqrt{\pi d}} (2e\pi C_{\mathrm{P}})^{d/2}$$

Proof Without loss of generality, assume $\mathbb{E}_\pi x = 0$. Consider a density q supported on the ball $B_R(0)$ of radius R centered at 0. such that $\int_{\mathbb{R}^d} q(x)\,dx = \int_{\mathbb{R}^d} e^{-f(x)}\,dx$ and $q(x) \equiv 1$ on the ball and 0 everywhere else. Then

$$\begin{aligned}\int_{x\in B_R(0)} (q(x) - e^{-f(x)}) \|x\|^2\,dx &\le R^2 \int_{x\in B_R(0)} (q(x) - e^{-f(x)})\,dx \\ &= R^2 \int_{x\notin B_R(0)} e^{-f(x)}\,dx \\ &\le \int_{x\notin B_R(0)} e^{-f(x)} \|x\|^2\,dx\end{aligned}$$

where the first inequality is due to $q(x) - e^{-f(x)} \ge 0$ for all $x \in B_R(0)$ and equality is due to

$$\int_{x\in B_R(0)} q(x)\,dx = \int_{x\in\mathbb{R}^d} q(x)\,dx = \int_{x\in B_R(0)} e^{-f(x)}\,dx + \int_{x\notin B_R(0)} e^{-f(x)}\,dx$$

Hence

$$\frac{\int_{\mathbb{R}^d} e^{-f(x)} \|x\|^2\,dx}{\int_{\mathbb{R}^d} e^{-f(x)}\,dx} \ge \frac{\int_{\mathbb{R}^d} q(x) \|x\|^2\,dx}{\int_{\mathbb{R}^d} q(x)\,dx} = \frac{\int_0^R r^{d-1} r^2\,dr}{\int_0^R r^{d-1}\,dr} = \frac{d}{d+2} R^2.$$

This implies $R \le \sqrt{C_{\mathrm{P}}(d+2)}$. Then by the volume of a n-ball and Stirling's formula,

$$\begin{aligned}\int_{\mathbb{R}^d} e^{-f(x)}\,dx = \int_{\mathbb{R}^d} q(x)\,dx &= \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} \left(C_{\mathrm{P}}(d+2)\right)^{d/2} \\ &\le \frac{1}{\sqrt{\pi d}} \left(\frac{2e\pi C_{\mathrm{P}}(d+2)}{d}\right)^{d/2} \le \frac{e}{\sqrt{\pi d}} (2e\pi C_{\mathrm{P}})^{d/2}. \end{aligned} \tag{28}$$

■

Lemma 57 *Suppose that $\pi \propto e^{-V}$ satisfies a Poincaré inequality with constant C_{P} and V is β-smooth. Then*

$$\|\nabla V(\mathbb{E}_\pi x)\| \le 2\sqrt{\beta}(\sqrt{d} + \sqrt{\beta C_{\mathrm{P}}}(\sqrt{d} + \ln 6)) \lesssim \beta\sqrt{C_{\mathrm{P}} d}.$$

Proof By Lemma 55, $\|x - \mathbb{E}_\pi x\| \le \sqrt{C_{\mathrm{P}}}(\sqrt{d} + \ln 6)$ with probability $\ge \frac{1}{2}$. Under this event, $\|\nabla V(\mathbb{E}_\pi x)\| \le \|\nabla V(x)\| + \beta\sqrt{C_{\mathrm{P}}}(\sqrt{d} + \ln 6)$. Hence

$$\frac{1}{2} \|\nabla V(\mathbb{E}_\pi x)\| \le \mathbb{E}_\pi[\|\nabla V\|] + \beta\sqrt{C_{\mathrm{P}}}(\sqrt{d} + \ln 6) \le \sqrt{\beta d} + \beta\sqrt{C_{\mathrm{P}}}(\sqrt{d} + \ln 6)$$

where the last inequality follows from $\mathbb{E}_\pi[\|\nabla V\|] \le \mathbb{E}_\pi[\|\nabla V\|^2]^{1/2} \le \sqrt{\beta d}$ by lemma 61.

The final inequality holds after observing that $\beta C_P \gtrsim 1$, which can be proven (for example) by combining the Cramer-Rao bound (viewing X as an unbiased estimator for the mean of the family of translates of the distribution π, see e.g. Van der Vaart (2000)) and the Poincaré inequality to show that $1/\beta \lesssim \mathrm{Var}(x_1) \lesssim C_P$. ■

Lemma 58 *Suppose that $\pi \propto e^{-V}$ satisfies a Poincaré inequality with constant C_{P} and V is β-smooth. Let $R_\varepsilon = \sqrt{C_{\mathrm{P}}}\left(\sqrt{d} + \ln\left(\frac{3}{\varepsilon}\right)\right)$. Then*

$$V(\mathbb{E}_\pi x) - \min V \le R_{1/2}\left\|\nabla V(\overline{x})\right\| + \frac{\beta}{2}R_{1/2}^2 + \ln 2 - \frac{d}{2}\ln\left(\frac{\beta}{2\pi}\right) \lesssim \beta C_{\mathrm{P}} d.$$

Proof Let $x^* \in \operatorname{argmin} V$ and $\overline{x} = \mathbb{E}_\pi x$. We upper and lower bound $\int_{\mathbb{R}^d} e^{-V(x)}\,dx$. For the lower bound, since V is β-smooth, letting $x^* \in \operatorname{argmin} V$, we have

$$V(x) - V(x^*) \le \langle x - x^*, \nabla V(x^*)\rangle + \beta\left\|x - x^*\right\|/2 = \beta\left\|x - x^*\right\|/2$$

since $\nabla V(x^*) = 0$. Thus

$$\int_{\mathbb{R}^d} e^{-V(x)}\,dx \ge e^{-\min V}\int_{\mathbb{R}^d} \exp(-\beta\left\|x - x^*\right\|/2)dx = e^{-\min V}\left(\frac{\beta}{2\pi}\right)^{\frac{d}{2}}.$$

For the upper bound, since by Lemma 55, $\left\|x - \overline{x}\right\| \le R_{1/2}$ with probability $\ge \frac{1}{2}$, we have

$$\int_{\mathbb{R}^d} e^{-V(x)}\,dx \le 2\int_{B_{R_{1/2}}(\overline{x})} e^{-V(x)}\,dx \le 2e^{-(V(\overline{x}) - R_{1/2}\|\nabla V(\overline{x})\| - \frac{\beta}{2}R_{1/2}^2)},$$

where the last inequality follows from

$$\begin{aligned} V(x) - V(\bar{x}) &\ge \langle \nabla V(\overline{x}), x - \overline{x}\rangle - \frac{\beta}{2}\left\|x - \overline{x}\right\|^2 \ge \left\|\nabla V(\overline{x})\right\|\left\|x - \overline{x}\right\| - \frac{\beta}{2}\left\|x - \bar{x}\right\|^2 \\ &\ge \left\|\nabla V(\overline{x})\right\| R_{1/2} - \frac{\beta}{2}R_{1/2}^2. \end{aligned}$$

Putting these inequalities together and taking the logarithm gives the result. ■

Appendix I. Inequalities for mixture distributions

In this section, we record several properties of a mixture of distributions, which would be useful in our analysis of the Langevin diffusion. It is well-known that the log-Sobolev inequality and hypercontractivity inequalities are equivalent, see e.g. Van Handel (2014). The following lemma gives a weaker version of hypercontractivity which is valid for mixtures of distributions satisfying the log-Sobolev inequality, which depends on the minimum weight in the mixture.

Lemma 59 (Hypercontractivity for mixtures, (Lee and Santana-Gijzen, 2024, Lemma 26)) *Let P_t be a reversible Markov process with stationary distribution $\pi = \sum_k w_k \pi_k$. Let $q(t) = 1 + (p - 1)e^{2t/C^*}$ where $C^* = \max_k c_k$. Assume that the following hold.*

1. *There exists a decomposition of the form,*

$$\langle f, \mathscr{L} f\rangle_\pi \le \sum_{k=1}^m w_i \langle f, \mathscr{L}_k f\rangle_{\pi_k}.$$

2. *For each π_k there exists a log-Sobolev inequality of the form,*

$$\mathrm{Ent}_{\pi_k}[f^2] \le 2c_k \cdot \mathscr{E}_{\pi_k}(f,f).$$

Let $f > 0$ and $w^* = \min_k w_k$. Then $\frac{\|P_t f\|_{L^{q(t)}(\pi)}}{(w^*)^{\frac{1}{q(t)}}}$ is a non-increasing function in t:

$$\frac{\|P_t f\|_{L^{q(t)}(\pi)}}{(w^*)^{\frac{1}{q(t)}}} \le \frac{\|P_0 f\|_{L^{q(0)}(\pi)}}{(w^*)^{\frac{1}{q(0)}}} = \frac{\|f\|_{L^p(\pi)}}{(w^*)^{\frac{1}{p}}}$$

and

$$\|P_t f\|_{L^{q(t)}(\pi)} \le \theta(q(t),p)\|f\|_{L^p(\pi)}$$

where $\theta(q,p) = \left(\frac{1}{w^*}\right)^{\frac{1}{p}-\frac{1}{q}}$.

The following two lemmas concern smoothness properties of the mixture distribution given smoothness of the components.

Lemma 60 (Hessian bound for mixture) *Suppose $\pi = \exp(-V(x)) = \sum_{i=1}^k p_i\pi_i$ where each π_i is β-smooth. Then*

$$-(\beta + G(x)^2)I \preceq \nabla^2 V(x) \preceq \beta I$$

where $G(x) = \max_i \|\nabla V_i(x)\|$.

Proof We note that

$$\nabla V^2(x) = \frac{\sum_{i=1}^k p_i\pi_i(x)\nabla^2 V_i(x)}{\pi(x)} - \frac{\sum_{i=1}^k p_i\pi_i(x)(\nabla V_i(x) - \nabla V_j(x))(\nabla V_i(x) - \nabla V_j(x))^\top}{4\pi^2(x)}. \tag{29}$$

The claim follows from $-\beta I \preceq \nabla^2 V_i(x) \preceq \beta I$ and

$$0 \preceq (\nabla V_i(x) - \nabla V_j(x))(\nabla V_i(x) - \nabla V_j(x))^\top \preceq \|\nabla V_i(x) - \nabla V_j(x)\|^2 I \preceq 4G(x)^2 I.$$

∎

Lemma 61 *If $\pi = \exp(-V(x)) = \sum_{i=1}^k p_i\pi_i$ where each π_i is β-smooth, then $\mathbb{E}_\pi[\|\nabla V(x)\|^2] \le \beta d$.*

Proof This follows from (Chewi et al., 2021, Lemma 16) i.e.

$$\mathbb{E}_\pi[\|\nabla V(x)\|^2] = \mathbb{E}_\pi[\Delta V(x)]$$

and noting that

$$\Delta V(x) = \mathrm{Tr}(\nabla^2 V(x)) \le \mathrm{Tr}\left(\frac{\sum_i p_i\pi_i(x)\nabla V_i^2(x)}{\pi(x)}\right) = \frac{\sum_i p_i\pi_i(x)\,\mathrm{Tr}(\nabla^2 V_i(x))}{\pi(x)} \le \beta d$$

where the first inequality is due to

$$\nabla^2 V \preceq \frac{\sum_i p_i\pi_i(x)\nabla^2 V_i(x)}{\pi(x)},$$

which is implied by Equation (29) and the last inequality is due to smoothness of each π_i. ∎