Fit Like You Sample: Sample-Efficient Score Matching From Fast Mixing Diffusions

Anonymous Author(s)
Affiliation
Address
email

Abstract

We show a close connection between the mixing time of a *broad class of* Markov processes with generator $\mathcal{L}$ and stationary distribution p, and an appropriately chosen *generalized score matching loss* that tries to fit $\frac{\mathcal{O}p}{p}$. In the special case of $\mathcal{O} = \nabla_x$, and $\mathcal{L}$ being the generator of Langevin diffusion, this generalizes and recovers the results from Koehler et al. [2022]. This allows us to adapt techniques to speed up Markov chains to construct better score-matching losses. In particular, *"preconditioning"* the diffusion can be translated to an appropriate *"preconditioning"* of the score loss. Lifting the chain by adding a temperature like in simulated tempering can be shown to result in a Gaussian-convolution annealed score matching loss, similar to Song and Ermon [2019]. Moreover, we show that if the distribution being learned is a finite mixture of Gaussians in d dimensions with a shared covariance, the sample complexity of annealed score matching is polynomial in the ambient dimension, the diameter of the means, and the smallest and largest eigenvalues of the covariance—obviating the Poincaré constant-based lower bounds of the basic score matching loss shown in Koehler et al. [2022].

1 Introduction

Energy-based models (EBMs) are parametric families of probability distributions parametrized up to a constant of proportionality, namely $p_\theta(x) \propto \exp(E_\theta(x))$ for some energy function $E_\theta(x)$. Fitting θ from data by using the standard approach of maximizing the likelihood of the training data with a gradient-based method requires evaluating $\nabla_\theta \log Z_\theta = \mathbb{E}_{p_\theta}[\nabla_\theta E_\theta(x)]$ — which cannot be done in closed form, and instead Markov Chain Monte Carlo methods are used. Score matching [Hyvärinen, 2005] obviates the need to estimate a partition function, by instead fitting the score of the distribution $\nabla_x \log p(x)$. While there is algorithmic gain, the statistical cost can be substantial. In recent work, Koehler et al. [2022] show that score matching is statistically much less efficient (i.e. the estimation error, given the same number of samples is much bigger) than maximum likelihood when the distribution being estimated has poor isoperimetric properties (i.e. a large Poincaré constant). However, even very simple multimodal distributions like a mixture of two Gaussians with far away means—have a very large Poincaré constant. As many distributions of interest (e.g. images) are multimodal in nature, the score matching estimator is likely to be statistically untenable.

The seminal paper by Song and Ermon [2019] proposes a way to deal with multimodality and manifold structure in the data by annealing: namely, estimating the scores of convolutions of the data distribution with different levels of Gaussian noise. The intuitive explanation they propose is that the distribution smoothed with more Gaussian noise is easier to estimate (as there are no parts of the distribution that have low coverage by the training data), which should help estimate the score at lower levels of Gaussian noise. However, making this either quantitative or formal seems very challenging.

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2023. Do not distribute.

In this paper, we show that there is a deep connection between the *mixing time* of a broad class of continuous, time-homogeneous Markov processes with stationary distribution p and generator $\mathcal{L}$, and the *statistical efficiency* of an appropriately chosen generalized score matching loss [Lyu, 2012] that tries to match $\frac{\mathcal{O}p}{p}$. This "dictionary" allows us to design score losses with better statistical behavior, by adapting techniques for speeding up Markov chain convergence — e.g. preconditioning a diffusion and lifting the chain by introducing additional variables. To summarize our contributions:

1. **A general framework** for designing generalized score matching losses with good sample complexity from fast-mixing diffusions. Precisely, for a broad class of diffusions with generator $\mathcal{L}$ and Poincaré constant C_P, we can choose a linear operator $\mathcal{O}$, such that the generalized score matching loss $\frac{1}{2}\mathbb{E}_p \left\| \frac{\mathcal{O}p}{p} - \frac{\mathcal{O}p_\theta}{p_\theta} \right\|_2^2$ has statistical complexity that is a factor C_P^2 worse than that of maximum likelihood. (Recall, C_P characterizes the mixing time of the Markov process with generator $\mathcal{L}$ in chi-squared distance.) In particular, for diffusions that look like "preconditioned" Langevin, this results in "appropriately preconditioned" score loss.
2. We analyze a **lifted** diffusion, which introduces a new variable for temperature and provably show **statistical benefits of annealing for score matching**. Precisely, we exhibit continuously-tempered Langevin, a Markov process which mixes in time $\text{poly}(D, d, 1/\lambda_{\min}, \lambda_{\max})$ for finite mixtures of Gaussians in ambient dimension d with identical covariances whose smallest and largest eigenvalues are lower and upper bounded by $\lambda_{\min}$ and $\lambda_{\max}$ respectively, and means lying in a ball of radius D. (Note, the bound has no dependence on the number of components.) Moreover, the corresponding generalized score matching loss is a form of annealed score matching loss [Song and Ermon, 2019, Song et al., 2020], with a particular choice of weighing for the different amounts of Gaussian convolution. This is the first result formally showing the statistical benefits of annealing for score matching.

Our work draws on and brings together, theoretical developments in understanding score matching, as well as designing and analyzing faster-mixing Markov chains based on strategies in annealing. We discuss these related lines of work here in Appendix J.

2 Preliminaries

The conventional score-matching objective [Hyvärinen, 2005] is defined as

$$D_{SM}(p, q) = \frac{1}{2}\mathbb{E}_p \|\nabla_x \log p - \nabla_x \log q\|_2^2 = \frac{1}{2}\mathbb{E}_p \left\| \frac{\nabla_x p}{p} - \frac{\nabla_x q}{q} \right\|_2^2 \tag{1}$$

Note, in this notation, the expression is asymmetric: p is the data distribution, q is the distribution that is being fit. Written like this, it is not clear how to minimize this loss, when we only have access to data samples from p. The main observation of Hyvärinen [2005] is that the objective can be rewritten (using integration by parts) in a form that is easy to fit given samples:

$$D_{SM}(p, q) = \mathbb{E}_{X\sim p}\left[\text{Tr}\, \nabla_x^2 \log q + \frac{1}{2}\|\nabla_x \log q\|^2\right] + K_p \tag{2}$$

where K_p is some constant independent of q. Generalized Score Matching, first introduced in Lyu [2012], generalizes ∇_x to an arbitrary linear operator $\mathcal{O}$:

Definition 1. *Let $\mathcal{F}^1$ and $\mathcal{F}^m$ be the space of all scalar-valued and m-variate functions of $x \in \mathbb{R}^d$, respectively. The Generalized Score Matching (GSM) loss with a general linear operator $\mathcal{O} : \mathcal{F}^1 \to \mathcal{F}^m$ is defined as $D_{GSM}(p, q) = \frac{1}{2}\mathbb{E}_p \left\| \frac{\mathcal{O}p}{p} - \frac{\mathcal{O}q}{q} \right\|_2^2$.*

In this paper, we will be considering operators $\mathcal{O}$, such that $(\mathcal{O}g)(x) = B(x)\nabla g(x)$. In other words, the generalized score matching loss will have the form:

$$D_{GSM}(p, q) = \frac{1}{2}\mathbb{E}_p \|B(x)(\nabla_x \log p - \nabla_x \log q)\|_2^2 \tag{3}$$

This can intuitively be thought of as a "preconditioned" version of the score matching loss, notably with a preconditioner function $B(x)$ that is allowed to change at every point x. The generalized score matching loss can also be turned into an expression that doesn't require evaluating the pdf of the data distribution (or gradients thereof), using a similar "integration-by-parts" identity: For the special case of the family of operators $\mathcal{O}$ in (3), the objective has the form (the proof is provided in Appendix B):

Lemma 1 (Integration by parts for the GSM in (3))**.** *The generalized score matching objective in* (3) *satisfies* $D_{GSM}(p,q) = \frac{1}{2}\left[\mathbb{E}_p\|B(x)\nabla_x \log q(x)\|^2 + 2\mathbb{E}_p \mathrm{div}\left(B(x)^2\nabla_x \log q(x)\right)\right] + K_p$.

We also introduce some key definitions related to diffusion processes. More detailed preliminaries are in Section A.

Definition 2 (Markov semigroup)**.** *We say that a family of functions* $\{P_t(x,y)\}_{t\geq 0}$ *on a state space* Ω *is a Markov semigroup if* $P_t(x,\cdot)$ *is a distribution on* Ω *and* $P_{t+s}(x,dy) = \int_\Omega P_t(x,dz)P_s(z,dy)$. *for all* $x, y \in \Omega$ *and* $s, t \geq 0$. *Finally, we say that* $p(x)$ *is a stationary distribution if* $X_0 \sim p$ *implies that* $X_t \sim p$ *for all* t.

A particularly important class of time-homogeneous Markov processes is given by Itô diffusions, namely stochastic differential equations of the form $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ for a *drift* function b, and a *diffusion coefficient* function. In fact, a classical result due to Dynkin (Rogers and Williams [2000], Theorem 13.3) states that **any** "sufficiently regular" time-homogeneous Markov process (specifically, a process whose semigroup is Feller-Dynkin) can be written in the above form. We will be interested in Itô diffusions, whose stationary distribution is a given distribution $p(x) \propto \exp(-f(x))$. Perhaps the most well-known example of such a diffusion is Langevin diffusion, namely $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$. In fact, a completeness result due to Ma et al. [2015] states that we can characterize **all** Itô diffusions whose stationary distribution is $p(x) \propto \exp(-f(x))$:

Theorem 1 (Itô diffusions with a given stationary distribution, Ma et al. [2015])**.** *Any Itô diffusion with stationary distribution* $p(x) \propto \exp(-f(x))$ *can be written in the form:*

$$dX_t = \left(-(D(X_t) + Q(X_t))\nabla f(X_t) + \Gamma(X_t)\right) dt + \sqrt{2D(X_t)}dB_t \tag{4}$$

where $\forall x \in \mathbb{R}^d, D(x) \in \mathbb{R}^{d\times d}$ *is a positive-definite matrix,* $\forall x \in \mathbb{R}^d, Q(x)$ *is a skew-symmetric matrix,* D, Q *are differentiable, and* $\Gamma_i(x) := \sum_j \partial_j(D_{ij}(x) + Q_{ij}(x))$.

Intuitively, $D(x)$ can be viewed as "reshaping" the diffusion, whereas Q and Γ are "correction terms" to the drift so that the stationary distribution is preserved. Versions of the SDEs we consider have appeared in the literature under various names, e.g., Riemannian Langevin [Girolami and Calderhead, 2011] and preconditioned Langevin [Hairer et al., 2007, Beskos et al., 2008], Fisher-adaptive Langevin [Titsias, 2023]. We finally recall a few objects related to the mixing time of Markov processes:

Definition 3. *The generator* $\mathcal{L}$ *corresponding to Markov semigroup is* $\mathcal{L}g = \lim_{t\to 0}\frac{P_t g - g}{t}$. *Moreover, if* p *is the unique stationary distribution, the Dirichlet form and the variance are respectively*

$$\mathcal{E}(g,h) = -\mathbb{E}_p\langle g, \mathcal{L}h\rangle \text{ and } \mathrm{Var}_p(g) = \mathbb{E}_p(g - \mathbb{E}_p g)^2 \text{ and denote } \mathcal{E}(g) := \mathcal{E}(g,g)$$

Definition 4 (Poincaré inequality)**.** *A continuous-time Markov process satisfies a Poincaré inequality with constant* C *if for all functions* g *such that* $\mathcal{E}(g)$ *is defined (finite), we have* $\mathcal{E}(g) \geq \frac{1}{C}\mathrm{Var}_p(g)$. *We will abuse notation, and for a Markov process with stationary distribution* p*, denote by* C_P *the* Poincaré constant of p*, the smallest* C *such that above Poincaré inequality is satisfied.*

The Poincaré inequality implies exponential ergodicity for the χ^2-divergence, namely: $\chi^2(p_t, p) \leq e^{-2t/C_P}\chi^2(p_0, p)$ where p is the stationary distribution of the chain and p_t is the distribution after running the Markov process for time t, starting at p_0.

3 Main results

The first main result is a general framework that provides a bound on the sample complexity of a generalized score matching objective under the assumption that an appropriate Markov process mixes fast. We let n denote the number of samples, and $\hat{\mathbb{E}}$ will denote an empirical average, that is the expectation over the n training samples. We show:

Theorem 2 (Main, sample complexity bound)**.** *Consider an Itô diffusion of the form* (4) *with stationary distribution* $p(x) \propto \exp(-f(x))$ *and Poincaré constant* C_P *with respect to the generator of the Itô diffusion. Consider the generalized score matching loss with operator* $(\mathcal{O}g)(x) := \sqrt{D(x)}\nabla g(x)$*, namely* $D_{GSM}(p,q) = \frac{1}{2}\mathbb{E}_p\left\|\sqrt{D(x)}\left(\nabla_x \log p - \nabla_x \log q\right)\right\|_2^2$. *Suppose we are optimizing this loss over a parametric family* $\{p_\theta : \theta \in \Theta\}$ *satisfying:*

1. *(Asymptotic normality) Let Θ^* be the set of global minima of the generalized score matching loss D_{GSM}, that is $\Theta^* = \{\theta^* : D_{GSM}(p, p_{\theta^*}) = \min_{\theta \in \Theta} D_{GSM}(p, p_\theta)\}$. Suppose the generalized score matching loss is asymptotically normal: namely, for every $\theta^* \in \Theta^*$, and every sufficiently small neighborhood S of θ^*, there exists a sufficiently large n, such that there is a unique minimizer $\hat{\theta}_n$ of $\hat{\mathbb{E}} l_\theta(x)$ in S, where*[1] $l_\theta(x) := \frac{1}{2}\left[\|\sqrt{D(x)}\nabla_x \log p_\theta(x)\|^2 + 2div\left(D(x)\nabla_x \log p_\theta(x)\right)\right]$. *Furthermore, assume $\hat{\theta}_n$ satisfies $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Gamma_{SM})$.*
2. *(Realizibility) At any $\theta^* \in \Theta^*$, we have $p_{\theta^*} = p$.*

Then, we have:
$$\|\Gamma_{SM}\|_{OP} \leq 2C_P^2 \|\Gamma_{\boldsymbol{MLE}}\|_{OP}^2 [\|cov(\nabla_\theta \nabla_x \log p_\theta(x) D(x) \nabla_x \log p_\theta(x))\|_{OP} + \|cov(\nabla_\theta \nabla_x \log p_\theta(x)^\top div(D(x)))\|_{OP} + \|cov(\nabla_\theta \operatorname{Tr}[D(x)\nabla_x^2 \log p_\theta(x))\|_{OP}]$$

The two terms on the right hand sides qualitatively capture two intuitive properties needed for good sample complexity: the factor involving covariances can be thought of as a smoothness term capturing the regularity of the score; the C_P term captures how the error compounds as we "extrapolate" the score into a probability density. Note that if we know $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Gamma_{SM})$, we can extract bounds on the expected ℓ_2^2 distance between $\hat{\theta}_n$ and θ^*. Namely, from Markov's inequality (see e.g., Remark 4 in Koehler et al. [2022]), we have for sufficiently large n, with probability at least 0.99 it holds that $\|\hat{\theta}_n - \theta^*\|_2^2 \leq \frac{\operatorname{Tr}(\Gamma_{SM})}{n}$.

The second result is translating another technique used to speed up Markov chains to statistical benefits of score matching: "lifting" the Markov chain by introducing additional variables (e.g., momentum in underdamped Langevin, temperature in tempering techniques) can be used to design better score losses to deal with multimodality in the data distribution. Precisely, we introduce a diffusion we call *Continuously Tempered Langevin Dynamics*, and we show it mixes in time $\text{poly}(D, d)$ for a mixture of K Gaussians (with identical covariance) in d dimensions, and means in a ball of radius D, with *no dependence* on the number of components K. Precisely:

Assumption 1. *Let $p_0 := \mathcal{N}(0, \Sigma)$. We will assume the data distribution p is a K-Gaussian mixture $p = \sum_{i=1}^K w_i p_i$, where $p_i(x) = p_0(x - \mu_i)$, i.e. a shift of the distribution p_0 so its mean μ_i. We assume $\max_i \|\mu_i\|_2 \leq D$ and denote $\lambda_{\min} := \lambda_{\min}(\Sigma), \lambda_{\max} := \lambda_{\max}(\Sigma), w_{\min} := \min_i w_i, w_{\max} := \max_i w_i$. Let $\Sigma_\beta := \Sigma + \beta\lambda_{\min} I_d$.*

Note, mixtures of Gaussians are universal approximators, if we consider a mixture with sufficiently many components [Alspach and Sorenson, 1972]. Note also we are just saying that the data distribution p can be described as a mixture of Gaussians, we are not saying anything about the parametric family we are fitting when optimizing the score matching loss. We will consider the following SDE:

Definition 5 (Continuously Tempered Langevin Dynamics (CTLD)). *We will consider an SDE over a temperature-augmented state space, that is a random variable $(X_t, \beta_t), X_t \in \mathbb{R}^d, \beta_t \in \mathbb{R}^+$, defined as*
$$\begin{cases} dX_t = \nabla_x \log p^\beta(X_t)dt + \sqrt{2}dB_t \\ d\beta_t = \nabla_\beta \log r(\beta_t)dt + \nabla_\beta \log p^\beta(X_t)dt + \nu_t L(dt) + \sqrt{2}dB_t \end{cases}$$
where $r : [0, \beta_{\max}] \to \mathbb{R}$ is defined as $r(\beta) \propto \exp\left(-\frac{7D^2}{\lambda_{\min}(1+\beta)}\right), \beta_{\max} := \frac{14D^2}{\lambda_{\min}} - 1$, and $p^\beta := p * \mathcal{N}(0, \beta\lambda_{\min} I_d)$. $L(dt)$ is a measure supported on the boundary of the interval $[0, \beta_{\max}]$ and ν_t is the unit normal at the endpoints of the interval, s.t. the stationary distribution is $p(x, \beta) = r(\beta)p^\beta(x)$ [Saisho, 1987].

The main result on the mixing time of CTLD is the following:

Theorem 3 (Poincaré constant of CTLD). *Under Assumption 1, the Poincaré constant of CTLD C_P enjoys the upper bound $C_P \lesssim D^{22} d^2 \lambda_{\max}^9 \lambda_{\min}^{-2}$.*

Leveraging our "dictionary" between mixing time and sample complexity of generalized score matching losses, we can show an asymptotic sample complexity bound for the corresponding score matching loss, which scales polynomially in $D, d, \lambda_{\max}, 1/\lambda_{\min}$, circumventing the Poincaré constant-based lower bounds of basic score matching in Koehler et al. [2022]. Due to space constraints, the details and formal statement are relegated to Appendix D.

[1]The notation $\text{div}D(x)$ denotes the divergence of the vector field $\mathbb{R}^d \to \mathbb{R}^d$, s.t. $\text{div}D(x)_i = \sum_j \partial_j D_{ji}(x)$

References

Daniel Alspach and Harold Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, 17(4):439–448, 1972.

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

Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. 2017.

Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32, 2019.

Mario Bebendorf. A note on the poincaré inequality for convex domains. *Zeitschrift für Analysis und ihre Anwendungen*, 22(4):751–756, 2003.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE, 2010.

Alexandros Beskos, Gareth Roberts, Andrew Stuart, and Jochen Voss. Mcmc methods for diffusion bridges. *Stochastics and Dynamics*, 8(03):319–350, 2008.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yu Cao, Jianfeng Lu, and Lihan Wang. On explicit l 2-convergence rate estimate for underdamped langevin dynamics. *Archive for Rational Mechanics and Analysis*, 247(5):90, 2023.

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

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022.

Zongchen Chen and Santosh S Vempala. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

G Constantine and T Savits. A multivariate faa di bruno formula with applications. *Transactions of the American Mathematical Society*, 348(2):503–520, 1996.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710. PMLR, 2017.

Persi Diaconis and Daniel Stroock. Geometric bounds for eigenvalues of markov chains. *The annals of applied probability*, pages 36–61, 1991.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering langevin monte carlo ii: An improved proof using soft markov chain decomposition. *arXiv preprint arXiv:1812.00793*, 2018.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011.

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

M Hairer, AM Stuart, and J Voss. Analysis of spdes arising in path sampling part ii: The nonlinear case. *The Annals of Applied Probability*, 17(5/6):1657–1706, 2007.

Björn Holmquist. The d-variate vector hermite polynomial of order k. *Linear algebra and its applications*, 237:155–190, 1996.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

Koji Hukushima and Koji Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

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

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

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

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

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

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

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Siwei Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

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

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

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34: 25359–25369, 2021.

Ankur Moitra and Andrej Risteski. Fast convergence for langevin diffusion with manifold structure. *arXiv preprint arXiv:2002.05576*, 2020.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

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

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

Chirag Pabbaraju, Dhruv Rohatgi, Anish Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *arXiv preprint arXiv:2306.01993*, 2023.

L Chris G Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge university press, 2000.

Yasumasa Saisho. Stochastic differential equations for multi-dimensional domain with reflecting boundary. *Probability Theory and Related Fields*, 74(3):455–477, 1987.

Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257, 2001.

Umut Simsekli, Roland Badeau, Taylan Cemgil, and Gaël Richard. Stochastic quasi-newton langevin monte carlo. In *International Conference on Machine Learning*, pages 642–651. PMLR, 2016.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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

Henry Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.

Michalis K Titsias. Optimal preconditioning and fisher adaptive langevin sampling. *arXiv preprint arXiv:2305.14442*, 2023.

Alexis Akira Toda. Operator reverse monotonicity of the inverse. *The American Mathematical Monthly*, 118(1):82–83, 2011.

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

Dawn Woodard, Scott Schmidler, and Mark Huber. Sufficient conditions for torpid mixing of parallel and simulated tempering. 2009a.

Dawn B Woodard, Scott C Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. 2009b.

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Yun Yang and David B Dunson. Sequential markov chain monte carlo. *arXiv preprint arXiv:1308.3861*, 2013.

Part I

Appendix

Table of Contents

A Preliminaries

A.1 Markov Chain Decomposition Theorems

Our mixing time bounds for the Continuously Tempered Langevin Dynamics will heavily use decomposition theorems to bound the Poincaré constant. These results "decompose" the Markov chain by partitioning the state space into sets, such that: (1) the mixing time of the Markov chain inside the sets is good; (2) the "projected" chain, which transitions between sets with probability equal to the probability flow between sets, also mixes fast.

In particular, we recall the following two results:

Theorem 4 (Decomposition of Markov Chains, Theorem 6.1 in Ge et al. [2018])**.** *Let $M = (\Omega, \mathcal{L})$ be a continuous-time Markov chain with stationary distribution p and Dirichlet form $\mathcal{E}(g, g) = -\langle g, \mathcal{L}g\rangle_p$. Suppose the following hold.*

1. *The Dirichlet form for $\mathcal{L}$ decomposes as $\langle f, \mathcal{L}g\rangle_p = \sum_{j=1}^m w_j \langle f, \mathcal{L}_j g\rangle_{p_j}$, where*

$$p = \sum_{j=1}^{m} w_j p_j$$

   and $\mathcal{L}_j$ is the generator for some Markov chain M_j on Ω with stationary distribution p_j.

2. *(Mixing for each M_j) The Dirichlet form $\mathcal{E}_j(f, g) = -\langle f, \mathcal{L}g\rangle_{p_j}$ satisfies the Poincaré inequality*

$$\mathrm{Var}_{p_j}(g) \le C\mathcal{E}_j(g, g).$$

3. *(Mixing for projected chain) Define the χ^2-projected chain $\bar{M}$ as the Markov chain on $[m]$ generated by $\bar{\mathcal{L}}$, where $\bar{\mathcal{L}}$ acts on $g \in L^2([m])$ by*

$$\bar{\mathcal{L}}\bar{g}(j) = \sum_{1 \le k \le m, k \ne j} [\bar{g}(k) - \bar{g}(j)]\bar{P}(j, k), \text{ where } \bar{P}(j, k) = \frac{w_k}{\max\{\chi^2(p_j, p_k), \chi^2(p_k, p_j), 1\}}.$$

   Let $\bar{p}$ be the stationary distribution of $\bar{M}$. Suppose $\bar{M}$ satisfies the Poincaré inequality $\mathrm{Var}_{\bar{p}}(\bar{g}) \le \bar{C}\bar{\mathcal{E}}(g, g)$.

Then M satisfies the Poincaré inequality

$$\mathrm{Var}_p(g) \le C\left(1 + \frac{\bar{C}}{2}\right)\mathcal{E}(g, g).$$

The Poincaré constant bounds we will prove will also use a "continuous" version of the decomposition Theorem 4, which also appeared in Ge et al. [2018]:

Theorem 5 (Continuous decomposition theorem, Theorem D.3 in Ge et al. [2018])**.** *Consider a probability measure π with C^1 density on $\Omega = \Omega^{(1)} \times \Omega^{(2)}$, where $\Omega^{(1)} \subseteq \mathbb{R}^{d_1}$ and $\Omega^{(2)} \subseteq \mathbb{R}^{d_2}$ are closed sets. For $X = (X_1, X_2) \sim P$ with probability density function p (i.e., $P(dx) = p(x)\,dx$ and $P(dx_2|x_1) = p(x_2|x_1)\,dx_2$), suppose that*

- *The marginal distribution of X_1 satisfies a Poincaré inequality with constant C_1.*
- *For any $x_1 \in \Omega^{(1)}$, the conditional distribution $X_2|X_1 = x_1$ satisfies a Poincaré inequality with constant C_2.*

Then π satisfies a Poincaré inequality with constant

$$\tilde{C} = \max\left\{C_2\left(1 + 2C_1\left\|\int_{\Omega^{(2)}} \frac{\|\nabla_{x_1} p(x_2|x_1)\|^2}{p(x_2|x_1)} dx_2\right\|_{L^\infty(\Omega^{(1)})}\right), 2C_1\right\}$$

A.2 Asymptotic efficiency

We will need a classical result about asymptotic convergence of M-estimators, under some mild identifiability and differentiability conditions. For this section, n will denote the number of samples, and $\hat{\mathbb{E}}$ will denote an empirical average, that is the expectation over the n training samples. The following result holds:

Lemma 2 (Van der Vaart [2000], Theorem 5.23)**.** *Consider a loss $L : \Theta \mapsto \mathbb{R}$, such that $L(\theta) = \mathbb{E}_p[\ell_\theta(x)]$ for $l_\theta : \mathcal{X} \mapsto \mathbb{R}$. Let Θ^* be the set of global minima of L, that is*

$$\Theta^* = \{\theta^* : L(\theta^*) = \min_{\theta \in \Theta} L(\theta)\}$$

Suppose the following conditions are met:

- *(Gradient bounds on l_θ) The map $\theta \mapsto l_\theta(x)$ is measurable and differentiable at every $\theta^* \in \Theta^*$ for p-almost every x. Furthermore, there exists a function $B(x)$, s.t. $\mathbb{E}B(x)^2 < \infty$ and for every θ_1, θ_2 near θ^*, we have:*
$$|l_{\theta_1}(x) - l_{\theta_2}(x)| < B(x)\|\theta_1 - \theta_2\|_2$$
- *(Twice-differentiability of L) $L(\theta)$ is twice-differentiable at every $\theta^* \in \Theta^*$ with Hessian $\nabla^2_\theta L(\theta^*)$, and furthermore $\nabla^2_\theta L(\theta^*) \succ 0$.*
- *(Uniform law of large numbers) The loss L satisfies a uniform law of large numbers, that is*
$$\sup_{\theta \in \Theta} \left| \hat{\mathbb{E}} l_\theta(x) - L(\theta) \right| \xrightarrow{p} 0$$

Then, for every $\theta^* \in \Theta^*$, and every sufficiently small neighborhood S of θ^*, there exists a sufficiently large n, such that there is a unique minimizer $\hat{\theta}_n$ of $\hat{\mathbb{E}} l_\theta(x)$ in S. Furthermore, $\hat{\theta}_n$ satisfies:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla^2_\theta L(\theta^*))^{-1} Cov(\nabla_\theta \ell(x;\theta^*))(\nabla^2_\theta L(\theta^*))^{-1}\right)$$

A.3 Hermite Polynomials

To obtain polynomial bounds on the moments of derivatives of Gaussians, we will use the known results on multivariate Hermite polynomials.

Definition 6 (Hermite polynomial, [Holmquist, 1996])**.** *The multivariate Hermite polynomial of order k corresponding to a Gaussian with mean 0 and covariance Σ is given by the Rodrigues formula:*

$$H_k(x;\Sigma) = (-1)^k \frac{(\Sigma \nabla_x)^{\otimes k} \phi(x;\Sigma)}{\phi(x;\Sigma)}$$

where $\phi(x;\Sigma)$ is the pdf of a d-variate Gaussian with mean 0 and covariance Σ, and $\otimes$ denotes the Kronecker product.

Note that $\nabla_x^{\otimes k}$ can be viewed as a formal Kronecker product, so that $\nabla_x^{\otimes k} f(x)$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a C^k-smooth function gives a d^k-dimensional vector consisting of all partial derivatives of f of order up to k.

Proposition 1 (Integral representation of Hermite polynomial, [Holmquist, 1996])**.** *The Hermite polynomial H_k defined in Definition 6 satisfies the integral formula:*

$$H_k(x;\Sigma) = \int (x + iu)^{\otimes k} \phi(u;\Sigma) du$$

where $\phi(x;\Sigma)$ is the pdf of a d-variate Gaussian with mean 0 and covariance Σ.

Note, the Hermite polynomials are either even functions or odd functions, depending on whether k is even or odd:

$$H_k(-x;\Sigma) = (-1)^k H_k(x;\Sigma) \tag{5}$$

This property can be observed from the Rodrigues formula, the fact that $\phi(\cdot;\Sigma)$ is symmetric around 0, and the fact that $\nabla_{-x} = -\nabla_x$.

We establish the following relationship between Hermite polynomial and (potentially mixed) derivatives in x and μ, which we will use to bound several smoothness terms appearing in Section H.

Lemma 3. *If $\phi(x;\Sigma)$ is the pdf of a d-variate Gaussian with mean 0 and covariance Σ, we have:*

$$\frac{\nabla_\mu^{k_1}\nabla_x^{k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (-1)^{k_2}\mathbb{E}_{u\sim\mathcal{N}(0,\Sigma)}[\Sigma^{-1}(x-\mu+iu)]^{\otimes(k_1+k_2)}$$

where the left-hand-side is understood to be shaped as a vector of dimension $\mathbb{R}^{d^{k_1+k_2}}$.

Proof. Using the fact that $\nabla_{x-\mu} = \nabla_x$ in Definition 6, we get:

$$H_k(x-\mu;\Sigma) = (-1)^k\frac{(\Sigma\nabla_x)^{\otimes k}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)}$$

Since the Kronecker product satisfies the property $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$, we have $(\Sigma\nabla_x)^{\otimes k} = \Sigma^{\otimes k}\nabla_x^{\otimes k}$. Thus, we have:

$$\frac{\nabla_x^k\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (-1)^k(\Sigma^{-1})^{\otimes k}H_k(x-\mu;\Sigma) \tag{6}$$

Since $\phi(\mu-x;\Sigma)$ is symmetric in μ and x, taking derivatives with respect to μ we get:

$$H_k(\mu-x;\Sigma) = (-1)^k\frac{(\Sigma\nabla_\mu)^k\phi(\mu-x;\Sigma)}{\phi(\mu-x;\Sigma)}$$

Rearranging again and using (5), we get:

$$\frac{\nabla_\mu^k\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (\Sigma^{-1})^{\otimes k}H_k(x-\mu;\Sigma) \tag{7}$$

Combining (6) and (7), we get:

$$\begin{aligned}
\frac{\nabla_\mu^{k_1}\nabla_x^{k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} &= (-1)^{k_2}\frac{\nabla_\mu^{k_1}[(\Sigma^{-1})^{\otimes k_2}H_{k_2}(x-\mu;\Sigma)\phi(x-\mu;\Sigma)]}{\phi(x-\mu;\Sigma)}\\
&= (-1)^{k_2}\frac{\nabla_\mu^{k_1}[\nabla_\mu^{k_2}\phi(x-\mu;\Sigma)]}{\phi(x-\mu;\Sigma)}\\
&= (-1)^{k_2}\frac{\nabla_\mu^{k_1+k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)}\\
&= (-1)^{k_2}(\Sigma^{-1})^{\otimes(k_1+k_2)}H_{k_1+k_2}(x-\mu;\Sigma)
\end{aligned}$$

Applying the integral formula from Proposition 1, we have:

$$\frac{\nabla_\mu^{k_1}\nabla_x^{k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (-1)^{k_2}\int[\Sigma^{-1}(x-\mu+iu)]^{\otimes(k_1+k_2)}\phi(u;\Sigma)\,du$$

as we needed. ☐

Now we are ready to obtain an explicit polynomial bound for the mixed derivatives for a multivariate Gaussian with mean μ and covariance Σ. We have the following bounds:

Lemma 4. *If $\phi(x;\Sigma)$ is the pdf of a d-variate Gaussian with mean 0 and covariance Σ, we have:*

$$\left\|\frac{\nabla_\mu^{k_1}\nabla_x^{k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)}\right\|_2 \lesssim \|\Sigma^{-1}(x-\mu)\|_2^{k_1+k_2} + d^{(k_1+k_2)/2}\lambda_{\min}^{-(k_1+k_2)/2}$$

where the left-hand-side is understood to be shaped as a vector of dimension $\mathbb{R}^{d^{k_1+k_2}}$.

Proof. We start with Lemma 3 and use the convexity of the norm

$$\left\|\frac{\nabla_\mu^{k_1}\nabla_x^{k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)}\right\|_2 \leq \mathbb{E}_{u\sim\mathcal{N}(0,\Sigma)}\|[\Sigma^{-1}(x-\mu+iu)]^{\otimes(k_1+k_2)}\|_2$$

Bounding the right-hand side, we have:

$$\begin{aligned}\mathbb{E}_{u\sim\mathcal{N}(0,\Sigma)}\|[\Sigma^{-1}(x-\mu+iu)]^{\otimes(k_1+k_2)}\|_2 &\lesssim \|\Sigma^{-1}(x-\mu)\|_2^{k_1+k_2}+\mathbb{E}_{u\sim\mathcal{N}(0,\Sigma)}\|\Sigma^{-1}u\|_2^{k_1+k_2}\\ &= \|\Sigma^{-1}(x-\mu)\|_2^{k_1+k_2}+\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}\|\Sigma^{-\frac{1}{2}}z\|_2^{k_1+k_2}\\ &\leq \|\Sigma^{-1}(x-\mu)\|_2^{k_1+k_2}+\|\Sigma^{-\frac{1}{2}}\|_{OP}^{k_1+k_2}\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}\|z\|_2^{k_1+k_2}\end{aligned}$$

Applying Lemma 28 yields the desired result. ☐

Similarly, we can bound mixed derivatives involving a Laplacian in x:

Lemma 5. *If $\phi(x;\Sigma)$ is the pdf of a d-variate Gaussian with mean 0 and covariance Σ, we have:*

$$\left\|\frac{\nabla_\mu^{k_1}\Delta_x^{k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)}\right\| \lesssim \sqrt{d^{k_2}}\|\Sigma^{-1}(x-\mu)\|_2^{k_1+2k_2}+d^{(k_1+3k_2)/2}\lambda_{\min}^{-(k_1+2k_2)/2}$$

Proof. By the definition of a Laplacian, and the AM-GM inequality, we have, for any function $f:\mathbb{R}^d\to\mathbb{R}$

$$\begin{aligned}(\Delta^k f(x))^2 &= \left(\sum_{i_1,i_2,\ldots,i_k=1}^{d}\partial_{i_1}^2\partial_{i_2}^2\cdots\partial_{i_k}^2 f(x)\right)^2\\ &\leq d^k\sum_{i_1,i_2,\ldots,i_k=1}^{d}\left(\partial_{i_1}^2\partial_{i_2}^2\cdots\partial_{i_k}^2 f(x)\right)^2\\ &\leq d^k\|\nabla_x^{2k}f(x)\|_2^2\end{aligned}$$

Thus, we have

$$\left\|\frac{\nabla_\mu^{k_1}\Delta_x^{k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)}\right\|_2 \leq \sqrt{d^{k_2}}\left\|\frac{\nabla_\mu^{k_1}\nabla_x^{2k_2}\phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)}\right\|_2$$

Applying Lemma 4, the result follows.

☐

A.4 Logarithmic derivatives

Finally, we will need similar bounds for logarithmic derivatives—that is, derivatives of $\log p(x)$, where p is a multivariate Gaussian.

We recall the following result, which is a consequence of the multivariate extension of the Faá di Bruno formula:

Proposition 2 (Constantine and Savits [1996], Corollary 2.10)**.** *Consider a function $f:\mathbb{R}^d\to\mathbb{R}$, s.t. f is N times differentiable in an open neighborhood of x and $f(x)\neq 0$. Then, for any multi-index $I\in\mathbb{N}^d$, s.t. $|I|\leq N$, we have:*

$$\partial_{x_I}\log f(x)=\sum_{k,s=1}^{|I|}\sum_{p_s(I,k)}(-1)^{k-1}(k-1)!\prod_{j=1}^{s}\frac{\partial_{l_j}f(x)^{m_j}}{f(x)^{m_j}}\frac{\prod_{i=1}^d(I_i)!}{m_j!l_j!^{m_j}}$$

where $p_s(I,k)=\{\{l_i\}_{i=1}^s\in(\mathbb{N}^d)^s,\{m_i\}_{i=1}^s\in\mathbb{N}^s: l_1\prec l_2\prec\cdots\prec l_s,\sum_{i=1}^s m_i=k,\sum_{i=1}^s m_i l_i=I\}$.

The $\prec$ ordering on multi-indices is defined as follows: $(a_1,a_2,\ldots,a_d):=a\prec b:=(b_1,b_2,\ldots,b_d)$ if:

1. *$|a|<|b|$*
2. *$|a|=|b|$ and $a_1<b_1$.*

3. $|a| = |b|$ *and* $\exists k >= 1$, *s.t.* $\forall j \leq k, a_j = b_j$ *and* $a_{k+1} < b_{k+1}$.

As a straightforward corollary, we have the following:

Corollary 1. *For any multi-index* $I \in \mathbb{N}^d$*, s.t.* $|I|$ *is a constant, we have*

$$|\partial_{x_I} \log f(x)| \lesssim \max\left(1, \max_{J \leq I} \left|\frac{\partial_J f(x)}{f(x)}\right|^{|I|}\right)$$

where $J \in \mathbb{N}^d$ *is a multi-index, and* $J \leq I$ *iff* $\forall i \in d, J_i \leq I_i$.

A.5 Moments of mixtures and the perspective map

The main strategy in bounding moments of quantities involving a mixture will be to leverage the relationship between the expectation of the score function and the so-called *perspective map*. In particular, this allows us to bound the moments of derivatives of the mixture score in terms of those of the individual component scores, which are easier to bound using the machinery of Hermite polynomials in the prior section.

Note in this section all derivatives are calculated at $\theta = \theta^*$ and therefore $p(x, \beta) = p_\theta(x, \beta)$.

Lemma 6. *(Convexity of perspective, Boyd and Vandenberghe [2004]) Let* f *be a convex function. Then, its corresponding* perspective map $g(u, v) := v f\left(\frac{u}{v}\right)$ *with domain* $\{(u, v) : \frac{u}{v} \in Dom(f), v > 0\}$ *is convex.*

We will apply the following lemma many times, with appropriate choice of differentiation operator D and power k.

Lemma 7. *Let* $D : \mathcal{F}^1 \to \mathcal{F}^m$ *be a linear operator that maps from the space of all scalar-valued functions to the space of m-variate functions of* $x \in \mathbb{R}^d$ *and let* θ *be such that* $p = p_\theta$*. For* $k \in \mathbb{N}$*, and any norm* $\|\cdot\|$ *of interest*

$$\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left\|\frac{(Dp_\theta)(x|\beta)}{p_\theta(x|\beta)}\right\|^k \leq \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\|\frac{(Dp_\theta)(x|\beta,i)}{p_\theta(x|\beta,i)}\right\|^k$$

Proof. Let us denote $g(u, v) := v\|\frac{u}{v}\|^k$. Note that since any norm is convex by definition, so is g, by Lemma 6. Then, we proceed as follows:

$$\begin{aligned}
\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left\|\frac{(Dp_\theta)(x|\beta)}{p_\theta(x|\beta)}\right\|^k &= \mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p(x|\beta)} \left\|\frac{(Dp_\theta)(x|\beta)}{p_\theta(x|\beta)}\right\|^k \\
&= \mathbb{E}_{\beta\sim r(\beta)} \int g((Dp_\theta)(x|\beta), p_\theta(x|\beta))dx \\
&= \mathbb{E}_{\beta\sim r(\beta)} \int g\left(\sum_{i=1}^K w_i (Dp_\theta)(x|\beta,i), \sum_{i=1}^K w_i p_\theta(x|\beta,i)\right) dx \quad (8) \\
&\leq \mathbb{E}_{\beta\sim r(\beta)} \int \sum_{i=1}^K w_i g((Dp_\theta)(x|\beta,i), p_\theta(x|\beta,i))dx \quad (9) \\
&= \mathbb{E}_{\beta\sim r(\beta)} \sum_{i=1}^K w_i \mathbb{E}_{x\sim p(x|\beta,i)} \left\|\frac{(Dp_\theta)(x|\beta,i)}{p_\theta(x|\beta,i)}\right\|^k \\
&\leq \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\|\frac{(Dp_\theta)(x|\beta,i)}{p_\theta(x|\beta,i)}\right\|^k
\end{aligned}$$

where (8) follows by linearity of D, and (9) by convexity of the function g. □

B Generators and score losses for diffusions

In this section, we derive several expressions about generators, Dirichlet forms, and associated generalized score matching losses for diffusions of the kind (4).

First, we derive the Dirichlet form of Itô diffusions of the form (4). Namely, we show:

Lemma 8 (Dirichlet form of continuous Markov Process)**.** *Suppose p vanishes at infinity. For an Itô diffusion of the form* (4)*, its Dirichlet form is:*

$$\mathcal{E}(g) = \mathbb{E}_p \|\sqrt{D(x)}\nabla g(x)\|_2^2$$

Proof. By Itô's Lemma, the generator $\mathcal{L}$ of the Itô diffusion (4) is:

$$(\mathcal{L}g)(x) = \langle -[D(x)+Q(x)]\nabla f(x) + \Gamma(x), \nabla g(x)\rangle + \operatorname{Tr}(D(x)\nabla^2 g(x))$$

The Dirichlet form is given by

$$\mathcal{E}(g) = -\mathbb{E}_p\langle \mathcal{L}g, g\rangle \quad = -\int p(x)\left[\underbrace{\langle -[D(x)+Q(x)]\nabla f(x) + \Gamma(x), \nabla g(x)\rangle}_{\mathrm{I}} + \underbrace{\operatorname{Tr}(D(x)\nabla^2 g(x))}_{\mathrm{II}}\right] g(x)dx$$

Expanding and using the definition of Γ, term I can be written as:

$$\begin{aligned}
\mathrm{I} = &\int p(x)\langle D(x)\nabla f(x), \nabla g(x)\rangle g(x)dx && (10)\\
&+\int p(x)\langle Q(x)\nabla_x f(x), \nabla g(x)\rangle g(x)dx && (11)\\
&-\int p(x)\sum_{i,j}\partial_j D_{ij}(x)\partial_i g(x) g(x)dx && (12)\\
&-\int p(x)\sum_{i,j}\partial_j Q_{ij}(x)\partial_i g(x) g(x)dx && (13)
\end{aligned}$$

We will simplify term II via a sequence of integration by parts:

$$\begin{aligned}
\mathrm{II} &= -\int p(x)\operatorname{Tr}(D(x)\nabla^2 g(x)) g(x)dx\\
&= -\int p(x)\left(\sum_{i,j} D_{ij}(x)\partial_{ij} g(x)\right) g(x)dx\\
&= -\sum_{i,j}\int p(x)D_{ij}(x)g(x)\partial_{ij}g(x)dx\\
&= -\sum_{i,j}\left(p(x)D_{ij}(x)g(x)\partial_i g(x)\Big|_{x=-\infty}^{\infty} - \int \partial_j[p(x)D_{ij}(x)g(x)]\partial_i g(x)dx\right)\\
&= \sum_{i,j}\int \partial_j[p(x)D_{ij}(x)g(x)]\partial_i g(x)dx\\
&= \sum_{i,j}\int \partial_j p(x)D_{ij}(x)g(x)\partial_i g(x)dx && (14)\\
&+ \sum_{i,j}\int p(x)\partial_j D_{ij}(x)g(x)\partial_i g(x)dx && (15)\\
&+ \sum_{i,j}\int p(x)D_{ij}(x)\partial_j g(x)\partial_i g(x)dx && (16)
\end{aligned}$$

463 The term (14) cancels out with term (10), so that we get:

$$\begin{aligned}
&\sum_{i,j}\int \partial_j p(x) D_{ij}(x) g(x) \partial_i g(x) dx \\
&= \sum_{i,j}\int p(x)\partial_j \log p(x) D_{ij}(x) g(x)\partial_i g(x) dx \\
&= -\int p(x)\langle D(x)\nabla_x f(x), \nabla_x g(x)\rangle g(x) dx
\end{aligned}$$

464 The term (15) cancels out with the term (12).

465 For term (11),

$$\begin{aligned}
&\int p(x)\langle Q(x)\nabla_x f(x), \nabla_x g(x)\rangle g(x) dx \\
&= -\int \langle Q(x)\nabla_x p(x), \nabla_x g(x)\rangle g(x) dx \\
&= \int \langle \nabla_x p(x), Q(x)\nabla_x g(x)\rangle g(x) dx \\
&= \int \sum_{i,j} \partial_j p(x) Q_{ji}(x) \partial_i g(x) g(x) dx \\
&= -\int \sum_{i,j} \partial_j p(x) Q_{ij}(x) \partial_i g(x) g(x) dx
\end{aligned}$$

466 Combining term (11) and term (13),

$$\begin{aligned}
&\int p(x)\langle Q(x)\nabla_x f(x), \nabla_x g(x)\rangle g(x) dx - \int p(x)\sum_{i,j}\partial_j Q_{ij}(x)\partial_i g(x) g(x) dx \\
&= -\int \sum_{i,j}[\partial_j p(x) Q_{ij}(x) + p(x)\partial_j Q_{ij}(x)]\partial_i g(x) g(x) dx \\
&= -\sum_{i,j}\int \partial_j [p(x) Q_{ij}(x)]\partial_i g(x) g(x) dx \\
&= -\sum_{i,j}\left(p(x) Q_{ij}(x)\partial_i g(x) g(x)\Big|_{x=-\infty}^{\infty} - \int p(x) Q_{ij}(x)\partial_j[\partial_i g(x) g(x)] dx\right) \\
&= \sum_{i,j}\int p(x) Q_{ij}(x)[\partial_{ij} g(x) g(x) + \partial_i g(x)\partial_j g(x)] dx \\
&= \frac{1}{2}\sum_{i,j}\int p(x)\{Q_{ij}(x)[\partial_{ij} g(x) g(x) + \partial_i g(x)\partial_j g(x)] + Q_{ji}(x)[\partial_{ji} g(x) g(x) + \partial_j g(x)\partial_i g(x)]\} dx \\
&= \frac{1}{2}\sum_{i,j}\int p(x)\{Q_{ij}(x)[\partial_{ij} g(x) g(x) + \partial_i g(x)\partial_j g(x)] - Q_{ij}(x)[\partial_{ji} g(x) g(x) + \partial_j g(x)\partial_i g(x)]\} dx \\
&= 0
\end{aligned}$$

467 In the end, we are only left with term (16):

$$\begin{aligned}
\mathcal{E}(g) &= \sum_{i,j}\int p(x) D_{ij}(x)\partial_j g(x)\partial_i g(x) dx \\
&= \int p(x)\langle \nabla_x g(x), D(x)\nabla_x g(x)\rangle dx \\
&= \mathbb{E}_p \|\sqrt{D(x)}\nabla_x g(x)\|_2^2
\end{aligned}$$

$\square$

We also calculate the integration by parts version of the generalized score matching loss for (3).

Lemma 9 (Integration by parts for the GSM in (3))**.** *Suppose p vanishes at infinity. The generalized score matching objective in* (3) *satisfies the equality*

$$D_{GSM}(p,q) = \frac{1}{2}\left[\mathbb{E}_p\|B(x)\nabla \log q\|^2 + 2\mathbb{E}_p div\left(B(x)^2\nabla \log q\right)\right] + K_p$$

Proof. Expanding the squares in (3), we have:

$$D_{GSM}(p,q) = \frac{1}{2}\left[\mathbb{E}_p\|B(x)\nabla \log p\|^2 + \mathbb{E}_p\|B(x)\nabla \log q\|^2 - 2\mathbb{E}_p\langle B(x)\nabla \log p, B(x)\nabla \log q\rangle\right]$$

The cross-term can be rewritten using integration by parts as:

$$\begin{aligned}
\mathbb{E}_p\langle B(x)\nabla \log p, B(x)\nabla \log q\rangle &= \int_x \langle \nabla p, B(x)^2 \nabla \log q\rangle \\
&= -\int_x p(x)\text{div}\left(B(x)^2\nabla \log q\right) \\
&= -\mathbb{E}_p\text{div}\left(B(x)^2\nabla \log q\right)
\end{aligned}$$

$\square$

C A Framework for Analyzing Generalized Score Matching

First, by way of remarks, some conditions for asymptotic normality can be readily obtained by applying standard results from asymptotic statistics (e.g. Van der Vaart [2000], Theorem 5.23, reiterated as Lemma 2 for completeness). From that lemma, when an estimator $\hat{\theta} = \arg\min \hat{\mathbb{E}} l_\theta(x)$ is asymptotically normal, we have $\sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla_\theta^2 L(\theta^*))^{-1}\text{Cov}(\nabla_\theta \ell(x;\theta^*))(\nabla_\theta^2 L(\theta^*))^{-1}\right)$, where $L(\theta) = \mathbb{E}_\theta l(x)$. Therefore, to bound the spectral norm of Γ_{SM}, we need to bound the Hessian and covariance terms in the expression above. The latter turns out to be a fairly straightforward calculation (Lemma 12). The bound on the Hessian is where the connection to the Poincaré constant manifests:

Lemma 10 (Bounding Hessian)**.** *The loss D_{GSM} defined in Theorem 2 satisfies*

$$\left[\nabla_\theta^2 D_{GSM}(p, p_{\theta^*})\right]^{-1} \preceq C_P \Gamma_{\boldsymbol{MLE}}.$$

Proof. To reduce notational clutter, we will drop $|_{\theta=\theta^*}$ since all the functions of θ are evaluated at θ^*. Consider an arbitrary direction w. We have:

$$\begin{aligned}
&\left\langle w, \nabla_\theta^2 D_{GSM}(p,p_\theta) w\right\rangle \overset{1}{=} \mathbb{E}_p\|\sqrt{D(x)}\nabla_x\nabla_\theta \log p_\theta(x) w\|_2^2 \\
&\overset{2}{\geq} \frac{1}{C_P}\text{Var}_p(\langle w, \nabla_\theta \log p_\theta(x)\rangle) \overset{3}{=} \frac{1}{C_P} w^T \Gamma_{MLE}^{-1} w
\end{aligned}$$

1 follows from a straightforward calculation (in Lemma 11), 2 follows from the definition of Poincaré inequality of a diffusion process with Dirichlet form derived in Lemma 8, applied to the function $\langle w, \nabla_\theta \log p_\theta\rangle$, and 3 follows since $\Gamma_{MLE} = \left[\mathbb{E}_p \nabla_\theta \log p_\theta \nabla_\theta \log p_\theta^\top\right]^{-1}$ (i.e. the inverse Fisher matrix [Van der Vaart, 2000]). Since this holds for every vector w, we have $\nabla_\theta^2 D_{GSM} \succeq \frac{1}{C_P}\Gamma_{MLE}^{-1}$. By monotonicity of the matrix inverse operator [Toda, 2011], the claim of the lemma follows. $\square$

Lemma 11 (Hessian of GSM loss)**.** *The Hessian of D_{GSM} defined in Theorem 2 satisfies*

$$\nabla_\theta^2 D_{GSM}(p, p_{\theta^*}) = \mathbb{E}_p\left[\nabla_\theta\nabla_x \log p_{\theta^*}(x)^\top D(x)\nabla_\theta\nabla_x \log p_{\theta^*}(x)\right]$$

Proof. By a straightforward calculation, we have:

$$\nabla_\theta D_{GSM}(p, p_\theta) = \mathbb{E}_p \nabla_\theta \left(\frac{\sqrt{D(x)} \nabla_x p_\theta(x)}{p_\theta(x)} \right) \left(\frac{\sqrt{D(x)} \nabla_x p_\theta(x)}{p_\theta(x)} - \frac{\sqrt{D(x)} \nabla_x p(x)}{p(x)} \right)$$

$$\begin{aligned} \nabla_\theta^2 D_{GSM}(p, p_\theta) &= \mathbb{E}_p \nabla_\theta \left(\frac{\sqrt{D(x)} \nabla_x p_\theta(x)}{p_\theta(x)} \right)^\top \nabla_\theta \left(\frac{\sqrt{D(x)} \nabla_x p_\theta(x)}{p_\theta(x)} \right) \\ &\quad - \left(\frac{\sqrt{D(x)} \nabla_x p_\theta(x)}{p_\theta(x)} - \frac{\sqrt{D(x)} \nabla_x p(x)}{p(x)} \right)^\top \nabla_\theta^2 \left(\frac{\sqrt{D(x)} \nabla_x p_\theta(x)}{p_\theta(x)} \right) \end{aligned}$$

Since $\frac{\sqrt{D(x)} \nabla_x p_{\theta^*}(x)}{p_{\theta^*}(x)} = \frac{\sqrt{D(x)} \nabla_x p(x)}{p(x)}$, the second term vanishes at $\theta = \theta^*$.

$$\nabla_\theta^2 D_{GSM}(p, p_{\theta^*}) = \mathbb{E}_p \left[\nabla_\theta \left(\frac{\sqrt{D(x)} \nabla_x p_{\theta^*}(x)}{p_{\theta^*}(x)} \right)^\top \nabla_\theta \left(\frac{\sqrt{D(x)} \nabla_x p_{\theta^*}(x)}{p_{\theta^*}(x)} \right) \right]$$

□

Lemma 12 (Bound on smoothness)**.** *For $l_\theta(x)$ defined in Theorem 2,*

$$\begin{aligned} \mathit{cov}(\nabla_\theta l_\theta(x)) \lesssim\ & \mathit{cov}\left(\nabla_\theta \nabla_x \log p_\theta(x) D(x) \nabla_x \log p_\theta(x) \right) \\ &+ \mathit{cov}\left(\nabla_\theta \nabla_x \log p_\theta(x)^\top \mathit{div}(D(x)) \right) \\ &+ \mathit{cov}\left(\nabla_\theta \operatorname{Tr}[D(x) \Delta \log p_\theta(x) \right) \end{aligned}$$

Proof. We have

$$\begin{aligned} \nabla_\theta l_\theta(x) &= \frac{1}{2} \nabla_\theta \left[\| \sqrt{D(x)} \nabla_x \log p_\theta(x) \|^2 + 2 \mathrm{div} \left(D(x) \nabla_x \log p_\theta(x) \right) \right] \\ &= \nabla_\theta \nabla_x \log p_\theta(x) D(x) \nabla_x \log p_\theta(x) + \nabla_\theta \nabla_x \log p_\theta(x)^\top \mathrm{div}(D(x)) + \nabla_\theta \operatorname{Tr}[D(x) \Delta \log p_\theta(x)] \end{aligned}$$

By Lemma 2 in Koehler et al. [2022], we also have

$$\begin{aligned} \mathrm{cov}(\nabla_\theta l_\theta(x)) \lesssim\ & \mathrm{cov}\left(\nabla_\theta \nabla_x \log p_\theta(x) D(x) \nabla_x \log p_\theta(x) \right) \\ &+ \mathrm{cov}\left(\nabla_\theta \nabla_x \log p_\theta(x)^\top \mathrm{div}(D(x)) \right) \\ &+ \mathrm{cov}\left(\nabla_\theta \operatorname{Tr}[D(x) \Delta \log p_\theta(x)] \right) \end{aligned}$$

which completes the proof. □

D Benefits of Annealing: Continuously Tempered Langevin Dynamics

In this section, we will flesh out the results on how speed-ups in mixing due to annealing can be translated to score losses with improved sample complexity.

First, we recall a nice property of mixture of Gaussians that facilitates our analysis: a convolution of a Gaussian mixture with a Gaussian produces another Gaussian mixture. Namely, the following holds from the distributivity property of the convolution operator, which is due to the linearity of an integral:

Proposition 3 (Convolution with Gaussian)**.** *Under Assumption 1, the distribution $p * \mathcal{N}(x; 0, \sigma^2 I)$ satisfies $p * \mathcal{N}(x; 0, \sigma^2 I) = \sum_i w_i \left(p_0(x - \mu_i) * \mathcal{N}(x; 0, \sigma^2 I) \right)$ and $(p_0(x - \mu_i) * \mathcal{N}(x; 0, \sigma^2 I))$ is a multivariate Gaussian with mean μ_i and covariance $\Sigma + \sigma^2 I$.*

Next, we make several remarks on the CTLD process we introduced in Defition 5:

Remark 1. *CTLD can be readily seen as a "continuous-time" analogue of the usual simulated tempering chain [Lee et al., 2018, Ge et al., 2018], which either evolves x according to a Markov chain with probability p^β, or changes β (which has a discrete number of possible values), and applies an appropriate Metropolis-Hastings filter. The stationary distribution is $p(x,\beta) = r(\beta)p^\beta(x)$, since the updates amount to performing (reflected) Langevin dynamics corresponding to this stationary distribution.*

Remark 2. *The existence of the boundary measure is a standard result of reflecting diffusion processes via solutions to the Skorokhod problem [Saisho, 1987]. If we ignore the boundary reflection term, the updates for CTLD are simply Langevin dynamics applied to the distribution $p(x,\beta)$. $r(\beta)$ specifies the distribution over the different levels of noise and is set up roughly so the Gaussians in the mixture have variance $\beta\Sigma$ with probability $\exp(-\Theta(\beta))$.*

Since CTLD amounts to performing (reflected) Langevin dynamics on the appropriate joint distribution $p(x,\beta)$, the corresponding generator $\mathcal{L}$ for CTLD is also readily written down:

Proposition 4 (Dirichlet form for CTLD)**.** *The Dirichlet form corresponding to CTLD has the form*

$$\mathcal{E}(f(x,\beta)) = \mathbb{E}_{p(x,\beta)}\|\nabla f(x,\beta)\|^2 = \mathbb{E}_{r(\beta)}\mathcal{E}_\beta(f(\cdot,\beta)) \tag{17}$$

where $\mathcal{E}_\beta$ is the Dirichlet form corresponding to the Langevin diffusion (Lemma 8) with stationary distribution $p(x|\beta)$.

In fact, we can derive the explicit score loss corresponding to CTLD:

Proposition 5. *The generalized score matching loss with $\mathcal{O} = \nabla_{x,\beta}$ satisfies*

$$\left[\nabla_\theta^2 D_{GSM}(p, p_{\theta^*})\right]^{-1} \preceq C_P \Gamma_{MLE}$$

Moreover,

$$\begin{aligned}
D_{GSM}(p,p_\theta) &= \mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p^\beta}(\|\nabla_x \log p(x,\beta) - \nabla_x \log p_\theta(x,\beta)\|^2 + \|\nabla_\beta \log p(x,\beta) - \nabla_\beta \log p_\theta(x,\beta)\|^2) \\
&= \mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p^\beta}\|\nabla_x \log p(x|\beta) - \nabla_x \log p_\theta(x|\beta)\|^2 \\
&+ \lambda_{\min}\mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p^\beta}\left(\left(\operatorname{Tr}\nabla_x^2 \log p(x|\beta) - \operatorname{Tr}\nabla_x^2 \log p_\theta(x|\beta)\right) + \left(\|\nabla_x \log p(x|\beta)\|_2^2 - \|\nabla_x \log p_\theta(x|\beta)\|_2^2\right)\right)^2
\end{aligned}$$

Proof. The first claim follows by Lemma 10 as a special case of Langevin on the lifted distribution. The second claim follows by writing $\nabla_\beta \log p(x|\beta)$ and $\nabla_\beta \log p_\theta(x|\beta)$ through the Fokker-Planck equation for $p(x|\beta)$ (see Lemma 13). □

Remark 3. *This loss was derived from first principles from the Markov Chain-based framework we propose, however, it is readily seen that this loss is a "second-order" version of the annealed losses in Song and Ermon [2019], Song et al. [2020] — the weights being given by the distribution $r(\beta)$. Additionally, this loss has terms matching "second order" behavior of the distributions, namely $\operatorname{Tr}\nabla_x^2 \log p(x|\beta)$ and $\|\nabla_x \log p(x|\beta)\|_2^2$ with a weighting of $\lambda_{\min}$. Note this loss would be straightforward to train by the change of variables formula (Proposition 6, Appendix E)—and we also note that somewhat related "higher-order" analogues of score matching have appeared in the literature (without analysis or guarantees), for example, Meng et al. [2021].*

To get a bound on the asymptotic sample complexity of generalized score matching, according to the framework from Theorem 2, we also need to bound the smoothness terms (Lemma 12 in the general framework). These terms of course depend on the choice of parametrization for the family of distributions we are fitting. To get a quantitative sense for how these terms might scale, we will consider the natural parametrization for a mixture:

Assumption 2. *Consider the case of learning unknown means, such that the parameters to be learned are a vector $\theta = (\mu_1, \mu_2, \dots, \mu_K) \in \mathbb{R}^{dK}$.*

Note that in this parametrization, we assume that the weights $\{w_i\}_{i=1}^K$ and shared covariance matrix Σ are known, though the results can be straightforwardly generalized to the natural parametrization in which we are additionally fitting a vector $\{w_i\}_{i=1}^K$ and matrix Σ, at the expense of some calculational complexity. With this parametrization, the smoothness term can be bounded as follows:

Theorem 6 (Smoothness under the natural parameterization)**.** *Under Assumptions 1 and 2, the following upper bound obtains:*

$$\|cov\left(\nabla_\theta \nabla_{x,\beta} \log p_\theta^\top \nabla_{x,\beta} \log p_\theta\right)\|_{OP} + \|cov\left(\nabla_\theta \Delta_{x,\beta} \log p_\theta\right)\|_{OP} \lesssim \operatorname{poly}\left(D, d, \lambda_{\min}^{-1}\right)$$

Note the above result *also has no dependence* on the number of components, or on the smallest component weight $w_{\min}$. Finally, we show that the generalized score matching loss is asymptotically normal. The proof of this is in Appendix G, and proceeds by verifying standard technical conditions for asymptotic behavior of M-estimators (Lemma 2), along with the Poincaré inequality bound in Theorem 3 and the framework in Theorem 2. As in Theorem 2, n will denote the number of samples, and $\hat{\mathbb{E}}$ will denote an empirical average, that is the expectation over the n training samples. We show:

Theorem 7 (Main, Polynomial Sample Complexity Bound of CTLD)**.** *Let the data distribution p satisfy Assumption 1. Then, the generalized score matching loss defined in Proposition 6 with parametrization as in Assumption 2 satisfies:*

1. *The set of optima $\Theta^* := \{\theta^* = (\mu_1, \mu_2, \ldots, \mu_K) | D_{GSM}(p, p_{\theta^*}) = \min_\theta D_{GSM}(p, p_\theta)\}$ satisfies:*

   $\theta^* = (\mu_1, \mu_2, \ldots, \mu_K) \in \Theta^*$ *if and only if* $\exists \pi : [K] \to [K]$ *satisfying* $\forall i \in [K], \mu_{\pi(i)} = \mu_i^*, w_{\pi(i)} = w_i\}$

2. *Let $\theta^* \in \Theta^*$ and let C be any compact set containing θ^*. Denote $C_0 = \{\theta \in C : p_\theta(x) = p(x)$ almost everywhere $\}$. Finally, let D be any closed subset of C not intersecting C_0. Then, we have $\lim_{n\to\infty} \Pr\left[\inf_{\theta \in D} \widehat{D_{GSM}}(\theta) < \widehat{D_{GSM}}(\theta^*)\right] \to 0$.*

3. *For every $\theta^* \in \Theta^*$ and every sufficiently small neighborhood S of θ^*, there exists a sufficiently large n, such that there is a unique minimizer $\hat{\theta}_n$ of $\hat{\mathbb{E}} l_\theta(x)$ in S. Furthermore, $\hat{\theta}_n$ satisfies: $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Gamma_{SM})$ for a matrix Γ_{SM} satisfying $\|\Gamma_{SM}\|_{OP} \leq \mathrm{poly}\left(D, d, \lambda_{\max}, \lambda_{\min}^{-1}\right) \|\Gamma_{MLE}\|_{OP}^2$.*

We provide some brief comments on each parts of this theorem. The first condition is the standard identifiability condition [Yakowitz and Spragins, 1968] for mixtures of Gaussians: the means are identifiable up to "renaming" the components. This is of course, inevitable if some of the weights are equal; if all the weights are distinct, Θ^* would in fact only consist of one point, s.t. $\forall i \in [K], \mu_i = \mu_i^*$. The second condition says that asymptotically, the empirical minimizers of D_{GSM} are the points in Θ^*. It can be viewed as (and follows from) a uniform law of large numbers. Finally, the third point characterizes the sample complexity of minimizers in the neirhborhood of each of the points in Θ^*, and is a consequence of the CTLD Poincaré inequality estimate (Theorem 3) and the smoothness estimate (Theorem 6). Note that in fact the RHS of point 3 has *no dependence* on the number of components. This makes the result extremely general: the loss compared to MLE is very mild even for distributions with a large number of modes. [2]

Bounding the Poincaré constant: We will first sketch the proof of Theorem 3. By slight abuse of notation, we will define the distribution of the "individual components" of the mixture at a particular temperature, namely for $i \in [K]$, define $p(x, \beta, i) = r(\beta) w_i \mathcal{N}(x; \mu_i, \Sigma + \beta\lambda_{\min} I_d)$. Correspondingly, we will denote the conditional distribution for the i-th component by $p(x, \beta | i) \propto r(\beta)\mathcal{N}(x; \mu_i, \Sigma + \beta\lambda_{\min} I_d)$.The proof proceeds by applying the decomposition Theorem 4 to CTLD. Towards that, we denote by $\mathcal{E}_i$ the Dirichlet form corresponding to Langevin with stationary distribution $p(x, \beta|i)$. By Proposition 4, it's easy to see that the generator for CTLD satisfies $\mathcal{E} = \sum_i w_i \mathcal{E}_i$. This verifies condition (1) in Theorem 4. To verify condition (2), we will show Langevin for each of the distributions $p(x, \beta|i)$ mixes fast (i.e. the Poincaré constant is bounded). The details of this are provided in Section F.1. To verify condition (3), we will show the projected chain "between" the components (as defined in Theorem 4) mixes fast. The details of this are provided in Section F.2.

Smoothness under the natural parametrization: To obtain the polynomial upper bound in Theorem 6, we note the two terms $\|\mathrm{cov}\left(\nabla_\theta \nabla_{x,\beta} \log p_\theta^\top \nabla_{x,\beta} \log p_\theta\right)\|_{OP}$ and $\|\mathrm{cov}\left(\nabla_\theta \Delta_{x,\beta} \log p_\theta\right)\|_{OP}$ can be completely characterized by bounds on the higher-order derivatives with respect to x and μ_i of the log-pdf since derivatives with respect to β can be related to derivatives with respect to x via the Fokker-Planck equation (Lemma 13). The polynomial bound requires three ingredients: In Lemma 7, we relate the derivatives of the mixture to derivatives of components by recognizing the higher-order score functions [Janzamin et al., 2014] of the form $\frac{D p}{p}$ is closely related to the convex perspective

[2] Of course, in the parametrization in Assumption 2, $\|\Gamma_{MLE}\|_{OP}$ itself will generally have dependence on K, which has to be the case since we are fitting $\Omega(K)$ parameters.

map. In Lemma 4, we derive a new result in mixed derivatives of Gaussian components based on Hermite polynomials. In Corollary 1, we handle log derivatives with higher-order versions of the Faá di Bruno formula [Constantine and Savits, 1996], which is a combinatorial formula characterizing higher-order analogues of the chain rule. See Appendix H for details.

E Technical calculations related to CTLD

In this section, we provide several calculations around the score matching losses associated with Continuously Tempered Langevin Dynamics.

Lemma 13 (β derivatives via Fokker Planck)**.** *For any distribution p^β such that $p^\beta = p * \mathcal{N}(0, \lambda_{\min}\beta I)$ for some p, we have the following PDE for its log-density:*

$$\nabla_\beta \log p^\beta(x) = \lambda_{\min} \left(\operatorname{Tr}\left(\nabla_x^2 \log p^\beta(x)\right) + \|\nabla_x \log p^\beta(x)\|_2^2\right)$$

As a consequence, both $p(x|\beta, i)$ and $p(x|\beta)$ follow the above PDE.

Proof. Consider the SDE $dX_t = \sqrt{2\lambda_{\min}} dB_t$. Let q_t be the law of X_t. Then, $q_t = q_0 * N(0, \lambda_{\min} tI)$. On the other hand, by the Fokker-Planck equation, $\frac{d}{dt} q_t(x) = \lambda_{\min}\Delta_x q_t(x)$. From this, it follows that

$$\begin{aligned}\nabla_\beta p^\beta(x) &= \lambda_{\min}\Delta_x p^\beta(x)\\ &= \lambda_{\min}\operatorname{Tr}(\nabla_x^2 p^\beta(x))\end{aligned}$$

Hence, by the chain rule,

$$\nabla_\beta \log p^\beta(x) = \frac{\lambda_{\min}\operatorname{Tr}(\nabla_x^2 p^\beta(x))}{p^\beta(x)} \tag{18}$$

Furthermore, by a straightforward calculation, we have

$$\nabla_x^2 \log p^\beta(x) = \frac{\nabla_x^2 p^\beta(x)}{p^\beta(x)} - \left(\nabla_x \log p^\beta(x)\right)\left(\nabla_x \log p^\beta(x)\right)^\top$$

Plugging this in (18), we have

$$\begin{aligned}\frac{\lambda_{\min}\operatorname{Tr}(\nabla_x^2 p^\beta(x))}{p^\beta(x)} &= \lambda_{\min}\left(\operatorname{Tr}\left(\nabla_x^2 \log p^\beta(x)\right) + \operatorname{Tr}\left(\left(\nabla_x \log p^\beta(x)\right)\left(\nabla_x \log p^\beta(x)\right)^\top\right)\right)\\ &= \lambda_{\min}\left(\operatorname{Tr}\left(\nabla_x^2 \log p^\beta(x)\right) + \operatorname{Tr}\left(\left(\nabla_x \log p^\beta(x)\right)^\top\left(\nabla_x \log p^\beta(x)\right)\right)\right)\\ &= \lambda_{\min}\left(\operatorname{Tr}\left(\nabla_x^2 \log p^\beta(x)\right) + \|\nabla_x \log p^\beta(x)\|_2^2\right)\end{aligned}$$

as we needed. □

Proposition 6 (Integration-by-part Generalized Score Matching Loss for CTLD)**.** *The loss D_{GSM} can be written in the integration by parts form as $D_{GSM}(p, p_\theta) = \mathbb{E}_p l_\theta(x, \beta) + K_p$, where*

$$l_\theta(x,\beta) := l_\theta^1(x,\beta) + l_\theta^2(x,\beta), \text{ and } l_\theta^1(x,\beta) := \frac{1}{2}\|\nabla_x \log p_\theta(x|\beta)\|_2^2 + \Delta_x \log p_\theta(x|\beta), \text{ and}$$

$$l_\theta^2(x,\beta) := \frac{1}{2}(\nabla_\beta \log p_\theta(x|\beta))^2 + \nabla_\beta \log r(\beta)\nabla_\beta \log p_\theta(x|\beta) + \Delta_\beta \log p_\theta(x|\beta)$$

Moreover, all the terms in the definition of $l_\theta^1(x,\beta)$ and $l_\theta^2(x,\beta)$ can be written as a sum of powers of partial derivatives of $\nabla_x \log p_\theta(x|\beta)$.

Proof of Lemma 6.

$$
\begin{aligned}
&D_{GSM}\left(p, p_\theta\right) \\
&= \frac{1}{2}\mathbb{E}_p[\left\|\nabla_{(x,\beta)} \log p_\theta(x,\beta)\right\|_2^2 + 2\Delta_{(x,\beta)} \log p_\theta(x,\beta)] \\
&= \frac{1}{2}\mathbb{E}_p[\|\nabla_x \log p_\theta(x,\beta)\|_2^2 + 2\Delta_x \log p_\theta(x,\beta) + \|\nabla_\beta \log p_\theta(x,\beta)\|_2^2 + 2\Delta_\beta \log p_\theta(x,\beta)] \\
&= \frac{1}{2}\mathbb{E}_p[\|\nabla_x \log p_\theta(x|\beta) + \nabla_x \log r(\beta)\|_2^2 + 2\Delta_x \log p_\theta(x|\beta) + 2\Delta_x \log r(\beta) \\
&+ \|\nabla_\beta \log p_\theta(x|\beta) + \nabla_\beta \log r(\beta)\|_2^2 + 2\Delta_\beta \log p_\theta(x|\beta) + 2\Delta_\beta \log r(\beta)] \\
&= \mathbb{E}_p[\frac{1}{2}\left\|\nabla_x \log p_\theta(x|\beta)\right\|_2^2 + \Delta_x \log p_\theta(x|\beta) \\
&+ \frac{1}{2}\left\|\nabla_\beta \log p_\theta(x|\beta)\right\|_2^2 + \nabla_\beta \log r(\beta)\nabla_\beta \log p_\theta(x|\beta) + \Delta_\beta \log p_\theta(x|\beta)] + C
\end{aligned}
$$

By Lemma 13, $\nabla_\beta \log p_\theta(x|\beta)$ is a function of partial derivatives of the score $\nabla_x \log p_\theta(x|\beta)$. Similarly, $\nabla_\beta^2 \log p_\theta(x|\beta)$ can be shown to be a function of partial derivatives of the score $\nabla_x \log p_\theta(x|\beta)$ as well:

$$
\begin{aligned}
\Delta_\beta \log p_\theta(x|\beta) &= \nabla_\beta \lambda_{\min}(\mathrm{Tr}(\nabla_x^2 \log p_\theta(x|\beta)) + \|\nabla_x \log p_\theta(x|\beta)\|_2^2) \\
&= \lambda_{\min}(\mathrm{Tr}(\nabla_x^2 \nabla_\beta \log p_\theta(x|\beta)) + 2\nabla_x \nabla_\beta \log p_\theta(x|\beta)^\top \nabla_x \log p_\theta(x|\beta))
\end{aligned}
$$

□

F Polynomial mixing time bound: proof of Theorem 3

The proof of Theorem 3 will follow by applying Theorem 4. Towards that, we need to verify the three conditions of the theorem:

1. (Decomposition of Dirichlet form) The Dirichlet energy of CTLD for $p(x,\beta)$, by the tower rule of expectation, decomposes into a linear combination of the Dirichlet forms of Langevin with stationary distribution $p(x,\beta|i)$. Precisely, we have

$$\mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\nabla f(x,\beta)\|^2 = \sum_i w_i \mathbb{E}_{(x,\beta)\sim p(x,\beta|i)}\|\nabla f(x,\beta)\|^2$$

2. (Polynomial mixing for individual modes) By Lemma 14, for all $i \in [K]$ the distribution $p(x,\beta|i)$ has Poincaré constant $C_{x,\beta|i}$ with respect to the Langevin generator that satisfies:

$$C_{x,\beta|i} \lesssim D^{20} d^2 \lambda_{\max}^9 \lambda_{\min}^{-1}$$

3. (Polynomial mixing for projected chain) To bound the Poincaré constant of the projected chain, by Lemma 17 we have

$$\bar{C} \lesssim D^2 \lambda_{\min}^{-1}$$

Putting the above together, by Theorem 6.1 in Ge et al. [2018] we have:

$$
\begin{aligned}
C_P &\le C_{x,\beta|i}\left(1 + \frac{\bar{C}}{2}\right) \\
&\le C_{x,\beta|i}\bar{C} \\
&\lesssim D^{22} d^2 \lambda_{\max}^9 \lambda_{\min}^{-2}
\end{aligned}
$$

F.1 Fast Mixing Within a Component

The first claim we will show is that we have fast mixing "inside" each of the components of the mixture. Formally, we show:

Lemma 14. *For $i \in [K]$, let $C_{x,\beta|i}$ be the Poincaré constant of $p(x,\beta|i)$. Then, we have $C_{x,\beta|i} \lesssim D^{20} d^2 \lambda_{\max}^9 \lambda_{\min}^{-1}$.*

The proof of this lemma proceeds via another (continuous) decomposition theorem. Intuitively, what we show is that for every β, $p(x|\beta,i)$ has a good Poincaré constant; moreover, the marginal distribution of β, which is $r(\beta)$, is log-concave and supported over a convex set (an interval), so has a good Poincaré constant. Putting these two facts together via a continuous decomposition theorem (Theorem D.3 in Ge et al. [2018], repeated as Theorem 5), we get the claim of the lemma.

Proof. The proof will follow by an application of a continuous decomposition result (Theorem D.3 in Ge et al. [2018], repeated as Theorem 5) , which requires three bounds:

1. A bound on the Poincaré constants of the distributions $p(\beta|i)$: since β is independent of i, we have $p(\beta|i) = r(\beta)$. Since $r(\beta)$ is a log-concave distribution over a convex set (an interval), we can bound its Poincaré constant by standard results [Bebendorf, 2003]. The details are in Lemma 15, $C_\beta \leq \frac{14D^2}{\pi\lambda_{\min}}$.
2. A bound on the Poincaré constant $C_{x|\beta,i}$ of the conditional distribution $p(x|\beta,i)$: We claim $C_{x|\beta,i} \leq \lambda_{\max} + \beta\lambda_{\min}$. This follows from standard results on Poincaré inequalities for strongly log-concave distributions. Namely, by the Bakry-Emery criterion, an α-strongly log-concave distribution has Poincaré constant $\frac{1}{\alpha}$ [Bakry and Émery, 2006]. Since $p(x|\beta,i)$ is a Gaussian whose covariance matrix has smallest eigenvalue lower bounded by $\lambda_{\max} + \beta\lambda_{\min}$, it is $(\lambda_{\max} + \beta\lambda_{\min})^{-1}$-strongly log-concave. Since $\beta \in [0, \beta_{\max}]$, we have $C_{x|\beta,i} \leq \lambda_{\max} + \beta_{\max}\lambda_{\min} \leq \lambda_{\max} + 14D^2$.
3. A bound on the "rate of change" of the density $p(x|\beta,i)$, i.e. $\left\| \int \frac{\|\nabla_\beta p(x|\beta,i)\|_2^2}{p(x|\beta,i)} dx \right\|_{L^\infty}$: This is done via an explicit calculation, the details of which are in Lemma 16.

By Theorem D.3 in Ge et al. [2018], the Poincaré constant $C_{x,\beta|i}$ of $p(x,\beta|i)$ enjoys the upper bound:

$$\begin{aligned}
C_{x,\beta|i} &\leq \max\left\{ C_{x|\beta_{\max},i} \left(1 + C_\beta \left\| \int \frac{\|\nabla_\beta p(x|\beta,i)\|_2^2}{p(x|\beta,i)} dx \right\|_{L^\infty(\beta)} \right), 2C_\beta \right\} \\
&\lesssim \max\left\{ \left(\lambda_{\max} + 14D^2\right) \left(1 + \frac{14D^2}{\pi\lambda_{\min}} d^2 \max\{\lambda_{\max}^8, D^{16}\} \right), \frac{28D^2}{\pi\lambda_{\min}} \right\} \\
&\lesssim \frac{D^{20} d^2 \lambda_{\max}^9}{\lambda_{\min}}
\end{aligned}$$

which completes the proof. □

Lemma 15 (Bound on the Poincaré constant of $r(\beta)$)**.** *Let C_β be the Poincaré constant of the distribution $r(\beta)$ with respect to reflected Langevin diffusion. Then,*

$$C_\beta \leq \frac{14D^2}{\pi\lambda_{\min}}$$

Proof. We first show that $r(\beta)$ is a log-concave distribution. By a direct calculation, the second derivative in β satisfies:

$$\nabla_\beta^2 \log r(\beta) = -\frac{14D^2}{\lambda_{\min}(1+\beta)^3} \leq 0$$

Since the interval is a convex set, with diameter $\beta_{\max}$, by Bebendorf [2003] we have

$$C_\beta \leq \frac{\beta_{\max}}{\pi} = \frac{14D^2}{\pi\lambda_{\min}} - \frac{1}{\pi}$$

from which the Lemma immediately follows. □

Lemma 16 (Bound on "rate of change" of the density $p(x|\beta, i)$)**.**

$$\left\| \int \frac{\|\nabla_\beta p(x|\beta,i)\|_2^2}{p(x|\beta,i)} dx \right\|_{L^\infty(\beta)} \lesssim d^2 \max\{\lambda_{\max}^8, D^{16}\}$$

Proof.

$$\begin{aligned}
\left\| \int \frac{\|\nabla_\beta p(x|\beta,i)\|_2^2}{p(x|\beta,i)} dx \right\|_{L^\infty(\beta)} &= \left\| \int \left(\nabla_\beta \log p(x|\beta,i)\right)^2 p(x|\beta,i) dx \right\|_{L^\infty(\beta)} \\
&= \sup_\beta \mathbb{E}_{x\sim p(x|\beta,i)} \left(\nabla_\beta \log p(x|\beta,i)\right)^2
\end{aligned}$$

We can apply Lemma 13 to derive explicit expressions for the right-hand side:

$$\begin{aligned}
\left\| \int \frac{\|\nabla_\beta p(x|\beta,i)\|_2^2}{p(x|\beta,i)} dx \right\|_{L^\infty(\beta)} &= \sup_\beta \mathbb{E}_{x\sim p(x|\beta,i)} \lambda_{\min}^2 \left[\mathrm{Tr}(\Sigma_\beta^{-1}) + \|\Sigma_\beta(x-\mu_i)\|_2^2\right]^2 \\
&\overset{①}{\leq} 2\lambda_{\min}^2 \sup_\beta \left[\mathrm{Tr}(\Sigma_\beta^{-1})^2 + \mathbb{E}_{x\sim p(x|\beta,i)} \|\Sigma_\beta(x-\mu_i)\|_2^4\right] \\
&\leq 2\lambda_{\min}^2 \sup_\beta \left[d^2((1+\beta)\lambda_{\min})^{-2} + \mathbb{E}_{z\sim\mathcal{N}(0,I)} \|\Sigma_\beta^{\frac{3}{2}} z \Sigma_\beta^{\frac{1}{2}}\|_2^4\right] \\
&\leq 2\lambda_{\min}^2 \sup_\beta \left[d^2((1+\beta)\lambda_{\min})^{-2} + \|\Sigma_\beta^{\frac{3}{2}}\|_{OP}^4 \|\Sigma_\beta^{\frac{1}{2}}\|_{OP}^4 \mathbb{E}_{z\sim\mathcal{N}(0,I)} \|z\|_2^4\right] \\
&\overset{②}{\leq} 4 \sup_\beta \left[d^2(1+\beta)^{-2} + \lambda_{\min}^2 \|\Sigma_\beta\|_{OP}^8 d^2\right] \\
&= 4 \sup_\beta \left[d^2(1+\beta)^{-2} + \lambda_{\min}^2 (\lambda_{\max} + \beta\lambda_{\min})^8 d^2\right] \\
&= 4\left(d^2 + \lambda_{\min}^2(\lambda_{\max} + \beta_{\max}\lambda_{\min})^8 d^2\right) \\
&\overset{③}{\leq} 4d^2 + 4d^2\lambda_{\min}^2(\lambda_{\max} + 14D^2)^8 \\
&\leq 16d^2 \max\{\lambda_{\max}^8, 14^8 D^{16}\}
\end{aligned}$$

In ①, we use $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \geq 0$; in ② we apply the moment bound for the Chi-Squared distribution of degree-of-freedom d in Lemma 28; and in ③ we plug in the bound on $\beta_{\max}$. ☐

F.2 Mixing between components

Next, we show the "projected" chain between the components mixes fast:

Lemma 17 (Poincaré constant of projected chain)**.** *Define the projected chain $\bar{M}$ over $[K]$ with transition probability*

$$T(i,j) = \frac{w_j}{\max\{\chi_{\max}^2(p(x,\beta|i), p(x,\beta|j)), 1\}}$$

where $\chi_{\max}^2(p,q) = \max\{\chi^2(p,q), \chi^2(q,p)\}$. If $\sum_{j\neq i} T(i,j) < 1$, the remaining mass is assigned to the self-loop $T(i,i)$. The stationary distribution $\bar{p}$ of this chain satisfies $\bar{p}(i) = w_i$. Furthermore, the projected chain has Poincaré constant

$$\bar{C} \lesssim D^2 \lambda_{\min}^{-1}.$$

The intuition for this claim is that the transition probability graph is complete, i.e. $T(i,j) \neq 0$ for every pair $i, j \in [K]$. Moreover, the transition probabilities are lower bounded, since the χ^2 distances between any pair of "annealed" distributions $p(x, \beta|i)$ and $p(x, \beta|j)$ can be upper bounded. The reason for this is that at large β, the Gaussians with mean μ_i and μ_j are smoothed enough so that they have substantial overlap; moreover, the distribution $r(\beta)$ is set up so that enough mass is placed on the large β. The precise lemma bounding the χ^2 divergence between the components is stated as Lemma 18.

Proof. The stationary distribution follows from the detailed balance condition $w_i T(i, j) = w_j T(j, i)$.

We upper bound the Poincaré constant using the method of canonical paths [Diaconis and Stroock, 1991]. For all $i, j \in [K]$, we set $\gamma_{ij} = \{(i, j)\}$ to be the canonical path. Define the weighted length of the path

$$\begin{aligned}
\|\gamma_{ij}\|_T &= \sum_{(k,l)\in\gamma_{ij}, k,l\in[K]} T(k, l)^{-1} \\
&= T(i, j)^{-1} \\
&= \frac{\max\{\chi^2_{\max}(p(x, \beta|i), p(x, \beta|j)), 1\}}{w_j} \\
&\leq \frac{14D^2}{\lambda_{\min} w_j}
\end{aligned}$$

where the inequality comes from Lemma 18 which provides an upper bound for the chi-squared divergence. Since D is an upper bound and $\lambda_{\min}$ is a lower bound, we may assume without loss of generality that $\chi^2_{\max} \geq 1$.

Finally, we can upper bound the Poincaré constant using Proposition 1 in Diaconis and Stroock [1991]

$$\begin{aligned}
\bar{C} &\leq \max_{k,l\in[K]} \sum_{\gamma_{ij}\ni(k,l)} \|\gamma_{ij}\|_T w_i w_j \\
&= \max_{k,l\in[K]} \|\gamma_{kl}\|_T w_k w_l \\
&\leq \frac{14D^2 w_{\max}}{\lambda_{\min}} \\
&\leq \frac{14D^2}{\lambda_{\min}}
\end{aligned}$$

□

Next, we will prove a bound on the chi-square distance between the joint distributions $p(x, \beta|i)$ and $p(x, \beta|j)$. Intuitively, this bound is proven by showing bounds on the chi-square distances between $p(x|\beta, i)$ and $p(x|\beta, j)$ (Lemma 19) — which can be explicitly calculated since they are Gaussian, along with tracking how much weight $r(\beta)$ places on each of the β. Moreover, the Gaussians are flatter for larger β, so they overlap more — making the chi-square distance smaller.

Lemma 18 (χ^2-divergence between joint "annealed" Gaussians)**.**

$$\chi^2(p(x, \beta|i), p(x, \beta|j)) \leq \frac{14D^2}{\lambda_{\min}}$$

Proof. Expanding the definition of χ^2-divergence, we have:

$$\begin{aligned}
\chi^2(p(x,\beta|i),p(x,\beta|j)) &= \int \left(\frac{p(x,\beta|i)}{p(x,\beta|j)}-1\right)^2 p(x,\beta|i)dxd\beta \\
&= \int_0^{\beta_{\max}} \int_{-\infty}^{+\infty} \left(\frac{p(x|\beta,i)r(\beta)}{p(x|\beta,j)r(\beta)}-1\right)^2 p(x|\beta,i)r(\beta)dxd\beta \\
&= \int_0^{\beta_{\max}} \chi^2(p(x|\beta,i),p(x|\beta,j))r(\beta)d\beta \\
&\leq \int_0^{\beta_{\max}} \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) r(\beta)d\beta \qquad (19)\\
&= \int_0^{\beta_{\max}} \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) \frac{1}{Z(D,\lambda_{\min})} \exp\left(-\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) d\beta \\
&= \frac{\beta_{\max}}{Z(D,\lambda_{\min})}
\end{aligned}$$

where in Line 19, we apply our Lemma 19 to bound the χ^2-divergence between two Gaussians with identical covariance. By a change of variable $\tilde{\beta} := \frac{7D^2}{\lambda_{\min}(1+\beta)}$, $\beta = \frac{7D^2}{\lambda_{\min}\tilde{\beta}} - 1$, $d\beta = -\frac{7D^2}{\lambda_{\min}}\frac{1}{\tilde{\beta}^2}d\tilde{\beta}$, we can rewrite the integral as:

$$\begin{aligned}
Z(D,\lambda_{\min}) &= \int_0^{\beta_{\max}} \exp\left(-\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) d\beta \\
&= -\frac{7D^2}{\lambda_{\min}} \int_{\frac{7D^2}{\lambda_{\min}}}^{\frac{7D^2}{\lambda_{\min}(1+\beta_{\max})}} \exp\left(-\tilde{\beta}\right) \frac{1}{\tilde{\beta}^2} d\tilde{\beta} \\
&= \frac{7D^2}{\lambda_{\min}} \int_{\frac{7D^2}{\lambda_{\min}(1+\beta_{\max})}}^{\frac{7D^2}{\lambda_{\min}}} \exp\left(-\tilde{\beta}\right) \frac{1}{\tilde{\beta}^2} d\tilde{\beta} \\
&\geq \frac{7D^2}{\lambda_{\min}} \int_{\frac{7D^2}{\lambda_{\min}(1+\beta_{\max})}}^{\frac{7D^2}{\lambda_{\min}}} \exp\left(-2\tilde{\beta}\right) d\tilde{\beta} \\
&= \frac{7D^2}{2\lambda_{\min}} \left(\exp\left(-\frac{14D^2}{\lambda_{\min}(1+\beta_{\max})}\right) - \exp\left(-\frac{14D^2}{\lambda_{\min}}\right)\right)
\end{aligned}$$

Since D is an upper bound and $\lambda_{\min}$ is a lower bound, we can assume $\frac{D^2}{\lambda_{\min}} \geq 1$ without loss of generality. Plugging in $\beta_{\max} = \frac{14D^2}{\lambda_{\min}} - 1$, we get

$$Z(D,\lambda_{\min}) \geq \frac{7}{2}\left(\exp\left(-1\right) - \exp\left(-14\right)\right) \geq 1$$

Finally, we get the desired bound

$$\chi^2(p(x,\beta|i),p(x,\beta|j)) \leq \beta_{\max} = \frac{14D^2}{\lambda_{\min}} - 1$$

□

The next lemma bounds the χ^2-divergence between two Gaussians with the same covariance.

Lemma 19 (χ^2-divergence between Gaussians with same covariance)**.**

$$\chi^2(p(x|\beta,i),p(x|\beta,j)) \leq \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right)$$

Proof. Plugging in the definition of χ^2-distance for Gaussians, we have:

$$\begin{aligned}
&\chi^2(p(x|\beta,i),p(x|\beta,j)) \\
&\le \frac{\det(\Sigma_\beta)^{\frac{1}{2}}}{\det(\Sigma_\beta)} \det\left(\Sigma_\beta^{-1}\right)^{-\frac{1}{2}} \\
&\exp\left(\frac{1}{2}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)^\top (\Sigma_\beta^{-1})^{-1}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)+\frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i-\mu_j^\top\Sigma_\beta^{-1}\mu_j\right) && (20)\\
&= \exp\left(\frac{1}{2}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)^\top (\Sigma_\beta^{-1})^{-1}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)+\frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i\right)\\
&\exp\left(-\mu_j^\top\Sigma_\beta^{-1}\mu_j\right)\\
&\le \exp\left(\frac{1}{2}(2\mu_j-\mu_i)^\top\Sigma_\beta^{-1}(2\mu_j-\mu_i)+\frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i\right) && (21)\\
&\le \exp\left(\frac{\|2\mu_j-\mu_i\|_2^2+\|2\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right)\\
&\le \exp\left(\frac{(\|2\mu_j\|_2+\|\mu_i\|_2)^2+4\|\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right)\\
&\le \exp\left(\frac{2\|2\mu_j\|_2^2+2\|\mu_i\|_2^2+4\|\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right)\\
&\le \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right)
\end{aligned}$$

In Equation 20, we apply Lemma G.7 from Ge et al. [2018] for the chi-square divergence between two Gaussian distributions. In Equation 21, we use the fact that Σ_β^{-1} is PSD.

□

G Asymptotic normality of generalized score matching for CTLD

The main theorem of this section is proving asymptotic normality for the generalized score matching loss corresponding to CTLD. Precisely, we show:

Theorem 8 (Asymptotic normality of generalized score matching for CTLD)**.** *Let the data distribution p satisfy Assumption 1. Then, the generalized score matching loss defined in Proposition 6 satisfies:*

1. *The set of optima*
$$\Theta^* := \{\theta^* = (\mu_1,\mu_2,\dots,\mu_K) | D_{GSM}(p,p_{\theta^*}) = \min_\theta D_{GSM}(p,p_\theta)\}$$
satisfies
$$\theta^* = (\mu_1,\mu_2,\dots,\mu_K)\in\Theta^* \text{ if and only if } \exists\pi:[K]\to[K] \text{ satisfying } \forall i\in[K], \mu_{\pi(i)}=\mu_i^*, w_{\pi(i)}=w_i\}$$
2. *Let $\theta^*\in\Theta^*$ and let C be any compact set containing θ^*. Denote*
$$C_0=\{\theta\in C: p_\theta(x)=p(x) \text{ almost everywhere }\}$$
Finally, let D be any closed subset of C not intersecting C_0. Then, we have:
$$\lim_{n\to\infty} Pr\left[\inf_{\theta\in D}\widehat{D_{GSM}}(\theta)<\widehat{D_{GSM}}(\theta^*)\right]\to 0$$
3. *For every $\theta^*\in\Theta^*$ and every sufficiently small neighborhood S of θ^*, there exists a sufficiently large n, such that there is a unique minimizer $\hat\theta_n$ of $\hat{\mathbb{E}}l_\theta(x)$ in S. Furthermore, $\hat\theta_n$ satisfies:*
$$\sqrt{n}(\hat\theta_n-\theta^*)\xrightarrow{d}\mathcal{N}(0,\Gamma_{SM})$$
for some matrix Γ_{SM}.

Proof. Part 1 is shown in Lemma 20: the claim roughly follows by classic results on the identifiability of the parameters of a mixture (up to permutations of the components) [Yakowitz and Spragins, 1968].

Part 2 is shown in Lemma 22: it follows from a uniform law of large numbers.

Finally, Part 3 follows from an application of Lemma 2—so we verify the conditions of the lemma are satisfied. The gradient bounds on l_θ are verified Lemma 21—and it largely follows by moment bounds on gradients of the score derived in Section H. Uniform law of large numbers is shown in Lemma 22, and the the existence of Hessian of $L = D_{GSM}$ is trivially verified. $\square$

For the sake of notational brevity, in this section, we will slightly abuse notation and denote $D_{GSM}(\theta) := D_{GSM}(p, p_\theta)$.

Lemma 20 (Uniqueness of optima)**.** *Suppose for $\theta := (\mu_1, \mu_2, \dots, \mu_K)$ there is no permutation $\pi : [K] \to [K]$, such that $\mu_{\pi(i)} = \mu_i^*$ and $w_{\pi(i)} = w_i, \forall i \in [K]$. Then, $D_{GSM}(\theta) > D_{GSM}(\theta^*)$*

Proof. For notational convenience, let D_{SM} denote the standard score matching loss, and let us denote $D_{SM}(\theta) := D_{SM}(p, p_\theta)$. For any distributions p_θ, by Proposition 1 in Koehler et al. [2022], it holds that

$$D_{SM}(\theta) - D_{SM}(\theta^*) \geq \frac{1}{LSI(p_\theta)} \mathrm{KL}(p_{\theta^*}, p_\theta)$$

where $LSI(q)$ denotes the Log-Sobolev constant of the distribution q. If $\theta = (\mu_1, \mu_2, \dots, \mu_K)$ is such that there is no permutation $\pi : [K] \to [K]$ satisfying $\mu_{\pi(i)} = \mu_i^*$ and $w_{\pi(i)} = w_i, \forall i \in [K]$, by Yakowitz and Spragins [1968] we have $\mathrm{KL}(p_{\theta^*}, p_\theta) > 0$. Furthermore, the distribution p_θ, by virtue of being a mixture of Gaussians, has a finite log-Sobolev constant (Theorem 1 in Chen et al. [2021]). Therefore, $D_{SM}(\theta) > D_{SM}(\theta^*)$.

However, note that $D_{GSM}(p_\theta)$ is a (weighted) average of D_{SM} losses, treating the data distribution as $p_{\theta^*}^\beta$, a convolution of p_{θ^*} with a Gaussian with covariance $\beta\lambda_{\min} I_d$; and the distribution being fitted as p_θ^β. Thus, the above argument implies that if $\theta \neq \theta^*$, we have $D_{GSM}(\theta) > D_{GSM}(\theta^*)$, as we need. $\square$

Lemma 21 (Gradient bounds of l_θ)**.** *Let $l_\theta(x, \beta)$ be as defined in Proposition 6. Then, there exists a constant $C(d, D, \frac{1}{\lambda_{\min}})$ (depending on d, D, $\frac{1}{\lambda_{\min}}$), such that*

$$\mathbb{E}\|\nabla_\theta l(x, \beta)\|^2 \leq C\left(d, D, \frac{1}{\lambda_{\min}}\right)$$

Proof. By Proposition 6,

$$\begin{aligned}
l_\theta(x, \beta) &= l_\theta^1(x, \beta) + l_\theta^2(x, \beta), \text{ and} \\
l_\theta^1(x, \beta) &:= \frac{1}{2}\|\nabla_x \log p_\theta(x|\beta)\|_2^2 + \Delta_x \log p_\theta(x|\beta) \\
l_\theta^2(x, \beta) &:= \frac{1}{2}(\nabla_\beta \log p_\theta(x|\beta))^2 + \nabla_\beta \log r(\beta) \nabla_\beta \log p_\theta(x|\beta) + \Delta_\beta \log p_\theta(x|\beta)
\end{aligned}$$

Using repeatedly the fact that $\|a + b\|^2 \leq 2\left(\|a\|^2 + \|b\|^2\right)$, we have:

$$\begin{aligned}
\mathbb{E}\,\|l_\theta(x, \beta)\|_2^2 &\lesssim \mathbb{E}\left\|l_\theta^2(x, \beta)\right\|_2^2 + \mathbb{E}\left\|l_\theta^2(x, \beta)\right\|_2^2 \\
\mathbb{E}\left\|l_\theta^1(x, \beta)\right\|_2^2 &\lesssim \mathbb{E}\,\|\nabla_x \log p_\theta(x, \beta)\|_2^4 + \mathbb{E}\left(\Delta_x \log p_\theta(x, \beta)\right)^2 \\
\mathbb{E}\left\|l_\theta^2(x, \beta)\right\|_2^2 &\lesssim \mathbb{E}\left(\nabla_\beta \log p_\theta(x|\beta)\right)^4 + \mathbb{E}\left(\nabla_\beta \log r(\beta) \nabla_\beta \log p_\theta(x|\beta)\right)^2 + \mathbb{E}\left(\Delta_\beta \log p_\theta(x|\beta)\right)^2
\end{aligned}$$

We proceed to bound the right hand sides above. We have:

$$\begin{aligned}
\mathbb{E}\left\|l_\theta^1(x, \beta)\right\|_2^2 &\lesssim \mathbb{E}\,\|\nabla_x \log p_\theta(x, \beta)\|_2^4 + \mathbb{E}\left(\Delta_x \log p_\theta(x, \beta)\right)^2 \\
&\lesssim \max_{\beta, i} \mathbb{E}_{x \sim p(x|\beta, i)} \|\nabla_x \log p_\theta(x|\beta, i)\|_2^4 + \max_{\beta, i} \mathbb{E}_{x \sim p(x|\beta, i)} \left(\Delta_x \log p_\theta(x|\beta, i)\right)^2 &(22) \\
&\leq \mathrm{poly}\left(d, \frac{1}{\lambda_{\min}}\right) &(23)
\end{aligned}$$

Where (22) follows by Lemma 7, and (23) follows by combining Corollaries 2 and 1.

The same argument, along with Lemma 13, and the fact that $\max_\beta(\nabla_\beta \log r(\beta))^4 \lesssim D^8\lambda_{\min}^{-4}$ by a direct calculation shows that

$$\mathbb{E}\left\|l_\theta^2(x,\beta)\right\|_2^2 \lesssim \mathbb{E}\left(\nabla_\beta \log p_\theta(x|\beta)\right)^4 + \mathbb{E}\left(\nabla_\beta \log r(\beta)\nabla_\beta \log p_\theta(x|\beta)\right)^2 + \mathbb{E}\left(\Delta_\beta \log p_\theta(x|\beta)\right)^2$$
$$\leq \text{poly}\left(d, D, \frac{1}{\lambda_{\min}}\right)$$

□

Lemma 22 (Uniform convergence)**.** *The generalized score matching loss satisfies a uniform law of large numbers:*

$$\sup_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| \xrightarrow{p} 0$$

Proof. The proof will proceed by a fairly standard argument, using symmetrization and covering number bounds. Precisely, let $T = \{(x_i,\beta_i)\}_{i=1}^n$ be the training data. We will denote by $\hat{\mathbb{E}}_T$ the empirical expectation (i.e. the average over) a training set T.

We will first show that

$$\mathbb{E}_T \text{sup}_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| \leq \frac{C\left(K, d, D, \frac{1}{\lambda_{\min}}\right)}{\sqrt{n}} \tag{24}$$

from which the claim will follow. First, we will apply the symmetrization trick, by introducing a "ghost training set" $T' = \{(x_i',\beta_i')\}_{i=1}^n$. Precisely, we have:

$$\begin{aligned}\mathbb{E}_T \text{sup}_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| &= \mathbb{E}_T \text{sup}_{\theta\in\Theta}\left|\hat{\mathbb{E}}_T l_\theta(x,\beta) - D_{GSM}(\theta)\right| \\ &= \mathbb{E}_T \text{sup}_{\theta\in\Theta}\left|\hat{\mathbb{E}}_T l_\theta(x,\beta) - \mathbb{E}_{T'}\hat{\mathbb{E}}_{T'} l_\theta(x,\beta)\right| \qquad (25) \\ &\leq \mathbb{E}_{T,T'} \text{sup}_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \left(l_\theta(x_i,\beta_i) - l_\theta(x_i',\beta_i')\right)\right| \qquad (26)\end{aligned}$$

where (25) follows by noting the population expectation can be expressed as the expectation over a choice of a (fresh) training set T', (26) follows by applying Jensen's inequality. Next, consider Rademacher variables $\{\varepsilon_i\}_{i=1}^n$. Since a Rademacher random variable is symmetric about 0, we have

$$\begin{aligned}\mathbb{E}_{T,T'} \text{sup}_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \left(l_\theta(x_i,\beta_i) - l_\theta(x_i',\beta_i')\right)\right| &= \mathbb{E}_{T,T'} \text{sup}_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(l_\theta(x_i,\beta_i) - l_\theta(x_i',\beta_i')\right)\right| \\ &\leq 2\mathbb{E}_T \text{sup}_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i l_\theta(x_i,\beta_i)\right|\end{aligned}$$

For notational convenience, let us denote by

$$R := \sqrt{\frac{1}{n}\sum_{i=1}^n \|\nabla_\theta l_\theta(x_i,\beta_i)\|^2}$$

We will bound this supremum by a Dudley integral, along with covering number bounds. Considering T as fixed, with respect to the randomness in $\{\varepsilon_i\}$, the process $\frac{1}{n}\sum_{i=1}^n \varepsilon_i l_\theta(x_i,\beta_i)$ is subgaussian with respect to the metric

$$d(\theta,\theta') := \frac{1}{\sqrt{n}} R\|\theta - \theta'\|_2$$

In other words, we have

$$\mathbb{E}_{\{\varepsilon_i\}} \exp\left(\lambda\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(l_\theta(x_i,\beta_i) - l_{\theta'}(x_i,\beta_i)\right)\right) \leq \exp\left(\lambda^2 d(\theta,\theta')\right) \tag{27}$$

The proof of this is as follows: since ε_i is 1-subgaussian, and

$$|l_\theta(x_i,\beta_i) - l_{\theta'}(x_i,\beta_i)| \le \|\nabla_\theta l_\theta(x_i,\beta_i)\|\|\theta-\theta'\|$$

we have that $\varepsilon_i\left(l_\theta(x_i,\beta_i) - l_{\theta'}(x_i,\beta_i)\right)$ is subgaussian with variance proxy $\|\nabla_\theta(x_i,\beta_i)\|^2\|\theta-\theta'\|^2$. Thus, $\frac{1}{n}\sum_{i=1}^n \varepsilon_i l_\theta(x_i,\beta_i)$ is subgaussian with variance proxy $\frac{1}{n^2}\sum_{i=1}^n \|\nabla_\theta l_\theta(x_i,\beta_i)\|^2\|\theta-\theta'\|_2^2$, which is equivalent to (27).

The Dudley entropy integral then gives

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i l_\theta(x_i,\beta_i)\right| \lesssim \int_0^\infty \sqrt{\log N(\epsilon,\Theta,d)}d\epsilon \tag{28}$$

where $N(\epsilon,\Theta,d)$ denotes the size of the smallest possible ϵ-cover of the set of parameters Θ in the metric d.

Note that the ϵ in the integral bigger than the diameter of Θ in the metric d does not contribute to the integral, so we may assume the integral has an upper limit

$$M = \frac{2}{\sqrt{n}}RD$$

Moreover, Θ is a product of K d-dimensional balls of (Euclidean) radius D, so

$$\begin{aligned}\log N(\epsilon,\Theta,d) &\le \log\left(\left(1+\frac{RD}{\sqrt{n}\epsilon}\right)^{Kd}\right)\\ &\le \frac{KdRD}{\sqrt{n}\epsilon}\end{aligned}$$

Plugging this estimate back in (28), we get

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i l_\theta(x_i,\beta_i)\right| &\lesssim \sqrt{KdRD/\sqrt{n}}\int_0^M \frac{1}{\sqrt{\epsilon}}d\epsilon\\ &\lesssim \sqrt{MKdRD/\sqrt{n}}\\ &\lesssim RD\sqrt{\frac{Kd}{n}}\end{aligned}$$

Taking expectations over the set T (keeping in mind that R is a function of T), by Lemma 21 we get

$$\begin{aligned}\mathbb{E}_T\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i l_\theta(x_i,\beta_i)\right| &\lesssim \mathbb{E}_T[R]D\sqrt{\frac{Kd}{n}}\\ &\lesssim \frac{C\left(K,d,D,\frac{1}{\lambda_{\min}}\right)}{\sqrt{n}}\end{aligned}$$

This completes the proof of (24). By Markov's inequality, (24) implies that for every $\epsilon > 0$,

$$\Pr_T\left[\sup_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| > \epsilon\right] \le \frac{C\left(K,d,D,\frac{1}{\lambda_{\min}}\right)}{\sqrt{n}\epsilon}$$

Thus, for every $\epsilon > 0$,

$$\lim_{n\to\infty}\Pr_T\left[\sup_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| > \epsilon\right] = 0$$

Thus,

$$\sup_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| \xrightarrow{p} 0$$

as we need. $\square$

H Polynomial smoothness bound: proof of Theorem 6

First, we need several easy consequences of the machinery developed in Section A.3, specialized to Gaussians appearing in CTLD.

Lemma 23. *For all $k \in \mathbb{N}$, we have:*

$$\max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\Sigma_\beta^{-1}(x-\mu_i)\|_2^{2k} \le d^k \lambda_{\min}^{-k}$$

Proof.

$$\begin{aligned}
\mathbb{E}_{x\sim p(x|\beta,i)} \|\Sigma_\beta^{-1}(x-\mu_i)\|_2^{2k} &= \mathbb{E}_{z\sim\mathcal{N}(0,I_d)} \|\Sigma_\beta^{-\frac{1}{2}} z\|_2^{2k} \\
&\le \mathbb{E}_{z\sim\mathcal{N}(0,I_d)} \|\Sigma_\beta^{-1}\|_{OP}^k \|z\|_2^{2k} \\
&\le \lambda_{\min}^{-k} \mathbb{E}_{z\sim\mathcal{N}(0,I_d)} \|z\|_2^{2k} \\
&\le d^k \lambda_{\min}^{-k}
\end{aligned}$$

where the last inequality follows by Lemma 28. $\square$

Combining this Lemma with Lemmas 4 and 5, we get the following corollary:

Corollary 2.

$$\max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\| \frac{\nabla_{\mu_i}^{k_1} \nabla_x^{k_2} p(x|\beta,i)}{p(x|\beta,i)} \right\|^{2k} \lesssim d^{(k_1+k_2)k} \lambda_{\min}^{-(k_1+k_2)k}$$

$$\max_{\beta,i} \mathbb{E}_{(x,\beta)\sim p(x|\beta,i)} \left\| \frac{\nabla_{\mu_i}^{k_1} \Delta_x^{k_2} p(x|\beta,i)}{p(x|\beta,i)} \right\|^{2k} \lesssim d^{(k_1+3k_2)k} \lambda_{\min}^{-(k_1+3k_2)k}$$

Finally, we will need the following simple technical lemma:

Lemma 24. *Let X be a vector-valued random variable with finite Var(X). Then, we have*

$$\|\mathrm{Var}(X)\|_{OP} \le 6\mathbb{E}\|X\|_2^2$$

Proof. We have

$$\begin{aligned}
\|\mathrm{Var}(X)\|_{OP} &= \left\| \mathbb{E}\left[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^\top \right] \right\|_{OP} \\
&\le \mathbb{E}\,\|X-\mathbb{E}[X]\|_2^2 \qquad (29) \\
&\le 6\mathbb{E}\|X\|_2^2 \qquad (30)
\end{aligned}$$

where (29) follows from the subadditivity of the spectral norm, (30) follows from the fact that

$$\|x+y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 + 2\langle x,y\rangle \le 3(\|x\|_2^2 + \|y\|_2^2)$$

for any two vectors x, y, as well as the fact that by Jensen's inequality, $\|\mathbb{E}[X]\|_2^2 \le \mathbb{E}\|X\|_2^2$. $\square$

Given this lemma, it suffices to bound $\mathbb{E}\|(\nabla_\theta \nabla_{x,\beta} \log p_\theta^\top \nabla_{x,\beta} \log p_\theta\|_2^2$ and $\mathbb{E}\|\nabla_\theta \Delta_{x,\beta} \log p_\theta\|_2^2$, which are given by Lemma 25 and Lemma 26, respectively.

Lemma 25.

$$\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left\| \nabla_\theta \nabla_{x,\beta} \log p_\theta^\top \nabla_{x,\beta} \log p_\theta \right\|_2^2 \le \mathrm{poly}\left(D, d, \frac{1}{\lambda_{\min}}\right)$$

Proof. Recall that $\theta = (\mu_1, \mu_2, \dots, \mu_K)$, where each μ_i is a d-dimensional vector, and we are viewing θ as a dK-dimensional vector.

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left\|\nabla_\theta \nabla_{x,\beta} \log p_\theta^\top \nabla_{x,\beta} \log p_\theta\right\|_2^2 \\
&\leq \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left[\|\nabla_\theta \nabla_{x,\beta} \log p_\theta\|_{OP}^2 \|\nabla_{x,\beta} \log p_\theta\|_2^2\right] \\
&\leq \sqrt{\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\theta \nabla_{x,\beta} \log p_\theta\|_{OP}^4} \sqrt{\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_{x,\beta} \log p_\theta\|_2^4}
\end{aligned}$$

where the last step follows by Cauchy-Schwartz. To bound both factors above, we will essentially first use Lemma 7 to relate moments over the mixture, with moments over the components of the mixture. Subsequently, we will use estimates for a single Gaussian, i.e. Corollaries 2 and 1.

Proceeding to the first factor, we have:

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_{x,\beta} \nabla_\theta \log p_\theta(x,\beta)\|_{OP}^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \nabla_\theta \log p_\theta(x,\beta)\|_{OP}^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\beta \nabla_\theta \log p_\theta(x,\beta)\|_2^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \nabla_\theta \log p_\theta(x|\beta)\|_{OP}^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\beta \nabla_\theta \log p_\theta(x|\beta)\|_2^4 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\nabla_x \nabla_\theta \log p_\theta(x|\beta,i)\|_{OP}^4 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\nabla_\beta \nabla_\theta \log p_\theta(x|\beta,i)\|_2^4 \quad (31) \\
&\leq \mathrm{poly}(d, 1/\lambda_{\min}) \quad (32)
\end{aligned}$$

where (31) follows from Lemma 7, and (32) follows by combining Corollaries 2 and 1 and Lemma 13.

The second factor is handled similarly[3]. We have:

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_{x,\beta} \log p_\theta(x,\beta)\|_2^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \log p_\theta(x,\beta)\|_2^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\nabla_\beta \log p_\theta(x,\beta)\right)^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \log p_\theta(x|\beta)\|_2^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\nabla_\beta \log p_\theta(x|\beta)\right)^4 + \mathbb{E}_{\beta\sim r(\beta)} \left(\nabla_\beta \log r(\beta)\right)^4 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\nabla_x \log p_\theta(x|\beta,i)\|_2^4 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} (\nabla_\beta \log p_\theta(x|\beta,i))^4 + \max_\beta (\nabla_\beta \log r(\beta))^4 \\
&\qquad (33) \\
&\leq \mathrm{poly}(d, D, 1/\lambda_{\min}) \quad (34)
\end{aligned}$$

where (33) follows from Lemma 7, and (34) follows by combining Corollaries 2 and 1 and Lemma 13, as well as the fact that $\max_\beta (\nabla_\beta \log r(\beta))^4 \lesssim D^8 \lambda_{\min}^{-4}$ by a direct calculation.

Together the estimates (32) and (34) complete the proof of the lemma. □

Lemma 26.

$$\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\theta \Delta_{x,\beta} \log p_\theta(x,\beta)\|_2^2 \leq \mathrm{poly}\left(d, \frac{1}{\lambda_{\min}}\right)$$

Proof.

$$\begin{aligned}
&\nabla_\theta \Delta_{(x,\beta)} \log p_\theta(x,\beta) && (35) \\
&= \nabla_\theta \Delta_x \log p_\theta(x,\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x,\beta) && (36) \\
&= \nabla_\theta \Delta_x \log p_\theta(x|\beta) + \nabla_\theta \Delta_x \log r(\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta) + \nabla_\theta \nabla_\beta^2 \log r(\beta) \\
&= \nabla_\theta \Delta_x \log p_\theta(x|\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta) && (37)
\end{aligned}$$

where (35) follows by exchanging the order of derivatives, (36) since β is a scalar, so the Laplacian just equals to the Hessian, (37) by dropping the derivatives that are zero in the prior expression.

[3]Note, $\nabla_\beta f(\beta)$ for $f : \mathbb{R} \to \mathbb{R}$ is a scalar, since β is scalar.

To bound both summands above, we will essentially first use Lemma 7 to relate moments over the mixture, with moments over the components of the mixture. Subsequently, we will use estimates for a single Gaussian, i.e. Corollaries 1 and 2. Precisely, we have:

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\theta \Delta_{x,\beta} \log p_\theta\|_2^2 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\theta \operatorname{Tr}(\nabla_x^2 \log p_\theta(x|\beta))\|_2^2 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta)\|_2^2 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\| \frac{\nabla_\theta \Delta_x p_\theta(x|\beta,i)}{p_\theta(x|\beta,i)} \right\|_2^2 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\| \frac{\nabla_\theta \nabla_x p_\theta(x|\beta,i)}{p_\theta(x|\beta,i)} \right\|_{OP}^4 &(38)\\
&\leq \operatorname{poly}(d, 1/\lambda_{\min}) &(39)
\end{aligned}$$

where (38) follows from Lemma 7 and Lemma 13, and (39) follows by combining Corollaries 1 and 2.

□

I Technical Lemmas

I.1 Moments of a chi-squared random variable

For the lemmas in this subsection, we consider a random variable $z \sim \mathcal{N}(0, I_d)$ and random variable $x \sim \mathcal{N}(\mu, \Sigma)$ where $\|\mu\| \leq D$ and $\Sigma \preceq \sigma_{\max}^2 I$.

Lemma 27 (Norm of Gaussian)**.** *The random variable z enjoys the bound*

$$\mathbb{E}\|z\|_2 \leq \sqrt{d}$$

Proof.

$$\begin{aligned}
(\mathbb{E}\|z\|_2)^2 &\leq \mathbb{E}\|z\|_2^2 &(40)\\
&= \mathbb{E}\sum_{i=1}^d z_i^2 \\
&= d &(41)
\end{aligned}$$

where (40) follows from Jensen, and (41) by plugging in the mean of a chi-squared distribution with d degree of freedom. □

Lemma 28 (Moments of Gaussian)**.** *Let $z \sim \mathcal{N}(0, I_d)$. For $l \in \mathbb{Z}^+$, $\mathbb{E}\|z\|_2^{2l} \lesssim d^l$.*

Proof. The key observation required is $\|z\|_2^2 = \sum_{i=1}^d z_i^2$ is a Chi-Squared distribution of degree d.

$$\begin{aligned}
\mathbb{E}\|z\|_2^{2l} &= \mathbb{E}\left(\|z\|_2^2\right)^l = \mathbb{E}_{q\sim\chi^2(d)} q^l \\
&= \frac{(d+2l-2)!!}{(d-2)!!} \leq (d+2l-2)^l \\
&\lesssim d^l
\end{aligned}$$

□

J Related work

Score matching: Score matching was originally proposed by Hyvärinen [2005], who also provided some conditions under which the estimator is consistent and asymptotically normal. Asymptotic normality is also proven for various kernelized variants of score matching in Barp et al. [2019]. Recent work by Koehler et al. [2022] proves that when the family of distributions being fit is rich enough, the statistical sample complexity of score matching is comparable to the sample complexity of maximum likelihood *only* when the distribution satisfies a Poincaré inequality. In particular, even simple bimodal distributions in 1 dimension (like a mixture of 2 Gaussians) can significantly worsen the sample complexity of score matching (*exponential* with respect to mode separation). For restricted parametric families (e.g. exponential families with sufficient statistics consisting of bounded-degree polynomials), recent work [Pabbaraju et al., 2023] showed that score matching can be comparably efficient to maximum likelihood, by leveraging the fact that a restricted version of the Poincaré inequality suffices for good sample complexity.

Theoretical understanding of annealed versions of score matching is still very impoverished. A recent line of work [Lee et al., 2022, 2023, Chen et al., 2022] explores how accurately one can sample using a learned (annealed) score, *if the (population) score loss is successfully minimized*. This line of work can be viewed as a kind of "error propagation" analysis: namely, how much larger the sampling error with a score learned up to some tolerance. It does not provide insight on when the score can be efficiently learned, either in terms of sample complexity or computational complexity.

Sampling by annealing: There are a plethora of methods proposed in the literature that use temperature heuristics [Marinari and Parisi, 1992, Neal, 1996, Earl and Deem, 2005] to alleviate the slow mixing of various Markov Chains in the presence of multimodal structure or data lying close to a low-dimensional manifold. A precise understanding of when such strategies have provable benefits, however, is fairly nascent. Most related to our work, in Ge et al. [2018], Lee et al. [2018], the authors show that when a distribution is (close to) a mixture of K Gaussians with identical covariances, the classical simulated tempering chain [Marinari and Parisi, 1992] with temperature annealing (i.e. scaling the log-pdf of the distribution), along with Metropolis-Hastings to swap the temperature in the chain mixes in time poly(K).

Decomposition theorems and mixing times The mixing time bounds we prove for CTLD rely on decomposition techniques. At the level of the state space of a Markov Chain, these techniques "decompose" the Markov chain by partitioning the state space into sets, such that: (1) the mixing time of the Markov chain inside the sets is good; (2) the "projected" chain, which transitions between sets with probability equal to the probability flow between sets, also mixes fast. These techniques also can be thought of through the lens of functional inequalities, like Poincaré and Log-Sobolev inequalities. Namely, these inequalities relate the variance or entropy of functions to the Dirichlet energy of the Markov Chain: the decomposition can be thought of as decomposing the variance/entropy inside the sets of the partition, as well as between the sets.

Most related to our work are Ge et al. [2018], Moitra and Risteski [2020], Madras and Randall [2002], who largely focus on decomposition techniques for bounding the Poincaré constant. Related "multiscale" techniques for bounding the log-Sobolev constant have also appeared in the literature Otto and Reznikoff [2007], Lelièvre [2009], Grunewald et al. [2009].

Learning mixtures of Gaussians Even though not the focus of our work, the annealed score-matching estimator with the natural parametrization (i.e. the unknown means) can be used to learn the parameters of a mixture from data. This is a rich line of work with a long history. Identifiability of the parameters from data has been known since the works of Teicher [1963], Yakowitz and Spragins [1968]. Early work in the theoretical computer science community provided guarantees for clustering-based algorithms [Dasgupta, 1999, Sanjeev and Kannan, 2001]; subsequent work provided polynomial-time algorithms down to the information theoretic threshold for identifiability based on the method of moments [Moitra and Valiant, 2010, Belkin and Sinha, 2010]; even more recent work tackles robust algorithms for learning mixtures in the presence of outliers [Hopkins and Li, 2018, Bakshi et al., 2022]; finally, there has been a lot of interest in understanding the success and failure modes of practical heuristics like expectation-maximization [Balakrishnan et al., 2017, Daskalakis et al., 2017].

Techniques to speed up mixing time of Markov chains SDEs with different choices of the drift and covariance term are common when designing faster mixing Markov chains. A lot of such schemas "precondition" by a judiciously chosen $D(x)$ in the formalism of equation (4). A particularly common choice is a Newton-like method, which amounts to preconditioning by the Fisher matrix [Girolami and Calderhead, 2011, Li et al., 2016, Simsekli et al., 2016], or some cheaper approximation thereof. More generally, non-reversible SDEs by judicious choice of D, Q have been shown to be quite helpful practically [Ma et al., 2015]

"Lifting" the Markov chain by introducing new variables is also a very rich and useful paradigms. There are many related techniques for constructing Markov Chains by introducing an annealing parameter (typically called a "temperature"). Our chain is augmented by a temperature random variable, akin to the simulated tempering chain proposed by Marinari and Parisi [1992]. In parallel tempering [Swendsen and Wang, 1986, Hukushima and Nemoto, 1996], one maintains multiple particles (replicas), each evolving according to the Markov Chain at some particular temperature, along with allowing swapping moves. Sequential Monte Carlo [Yang and Dunson, 2013] is a related technique available when gradients of the log-likelihood can be evaluated.

Analyses of such techniques are few and far between. Most related to our work, Ge et al. [2018] analyze a variant of simulated tempering when the data distribution looks like a mixture of (unknown) Gaussians with identical covariance, and can be accessed via gradients to the log-pdf. We compare in more detail to this work in Section D. In the discrete case (i.e. for Ising models), Woodard et al. [2009b,a] provide some cases in which simulated and parallel tempering provide some benefits to mixing time.

Another way to "lift" the Markov chain is to introduce a velocity variable, and come up with "momentum-like" variants of Langevin. The two most widely known ones are underdamped Langevin and Hamiltonian Monte Carlo. There are many recent results showing (both theoretically and practically) the benefit of such variants of Langevin, e.g. [Chen and Vempala, 2019, Cao et al., 2023]. The proofs of convergence times of these chains is unfortunately more involved than merely a bound on a Poincaré constant (in fact, one can prove that they don't satisfy a Poincaré constant) — and it's not so clear how to "translate" them into a statistical complexity analysis using the toolkit we provide in this paper. This is fertile ground for future work, as score losses including a velocity term have already shown useful in training score-based models [Dockhorn et al., 2021].