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

Yilong Qin YILONGQ@ANDREW.CMU.EDU
Carnegie Mellon University

Andrej Risteski ARISTESK@ANDREW.CMU.EDU
Carnegie Mellon University

Editors: Shipra Agrawal and Aaron Roth

Abstract

Score matching is an approach to learning probability distributions parametrized up to a constant of proportionality (e.g., energy-based models). The idea is to fit the score of the distribution (i.e., $\nabla_x \log p(x)$), rather than the likelihood, thus avoiding the need to evaluate the constant of proportionality. While there's a clear algorithmic benefit, the statistical cost can be steep: recent work by Koehler et al. (2022) showed that for distributions that have poor isoperimetric properties (a large Poincaré or log-Sobolev constant), score matching is substantially statistically less efficient than maximum likelihood. However, many natural realistic distributions, e.g. multimodal distributions as simple as a mixture of two Gaussians in one dimension—have a poor Poincaré constant.

In this paper, 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 setting $\mathcal{O} = \nabla_x$, and $\mathcal{L}$ as 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. To show this we bound the mixing time of a "continuously tempered" version of Langevin diffusion for mixtures, which is of standalone interest.

1. Introduction

Score matching is a method introduced by Hyvärinen (2005) for learning the parameters of a distribution from data, useful for parametric families in which evaluating the likelihood is intractable. An illustrative example are energy-based models (EBMs), parametric families of the form $p_\theta(x) \propto \exp(E_\theta(x))$, for which evaluating and optimizing the likelihood is comutationally hard due to the partition function $Z_\theta = \int_x \exp(E_\theta(x))$ (Pabbaraju et al., 2023). Score matching obviates evaluating the partition function by instead fitting the score of the distribution (i.e., $\nabla_x \log p(x)$). 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.

In the generative models literature, 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 quantitative or formal seems very challenging.

In this paper, we show that there is a deep connection between the *mixing time* of broad classes 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 $\frac{\mathcal{O}p}{p}$. In the case that $\mathcal{L}$ is the generator of Langevin diffusion, and $\mathcal{O} = \nabla_x$, we recover the results of Koehler et al. (2022). This "dictionary" allows us to design score losses with better statistical behavior, by "translating" techniques for speeding up Markov chain convergence — e.g. preconditioning a diffusion and lifting the chain by introducing additional variables.

Our contributions are as follows:

1. **A general framework** for designing generalized score matching losses with good asymptotic 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 we provably show **statistical benefits of annealing for score matching**. Precisely, we exhibit continuously-tempered Langevin (CTLD), a Markov process which mixes in time $\mathrm{poly}(D, d, 1/\lambda_{\min}, \lambda_{\max})$ 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 the annealed score matching loss introduced in (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. Technically, this result involves bounding the mixing time of a "continuously tempered" version of Langevin dynamics for mixtures, using functional decomposition theorems.

On a conceptual level, 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. An in-depth review of related prior work is included in Appendix I.

2. Preliminaries

2.1. Generalized Score Matching

The conventional score-matching objective (Hyvärinen, 2005) is defined as

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

Note, the expression is asymmetric: p is the data distribution, p_θ 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 (under suitable decay conditions using integration by parts) in a form that is easy to fit given samples:

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

where K_p is some constant independent of q. To turn this into an algorithm given samples, one simply solves

$$\min_\theta \mathbb{E}_{X\sim \hat{p}}\left[\operatorname{Tr}\nabla_x^2 \log p_\theta + \frac{1}{2}\|\nabla_x \log p_\theta\|^2\right]$$

where $\hat{p}$ denotes the uniform distribution over the samples from p. This objective can be calculated efficiently given samples from p, so long as the gradient and Hessian of the log-pdf of p_θ can be efficiently calculated.[1]

Generalized Score Matching, first introduced in Lyu (2012), generalizes ∇_x to an arbitrary linear operator $\mathcal{O}$:

Definition 1 *The Generalized Score Matching (GSM) loss with a linear operator $\mathcal{O}$ acting on density functions is defined as* $D_{GSM}(p, p_\theta) = \frac{1}{2}\mathbb{E}_p \left\|\frac{\mathcal{O}p}{p} - \frac{\mathcal{O}p_\theta}{p_\theta}\right\|_2^2$.

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

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

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:

Lemma 1 (Integration by parts, Lyu (2012)) *The GSM loss satisfies*

$$D_{GSM}(p, p_\theta) = \frac{1}{2}\mathbb{E}_p\left[\left\|\frac{\mathcal{O}p_\theta}{p_\theta}\right\|_2^2 - 2\mathcal{O}^+\left(\frac{\mathcal{O}p_\theta}{p_\theta}\right)\right] + K_p \tag{4}$$

where $\mathcal{O}^+$ is the adjoint of $\mathcal{O}$ defined by $\langle \mathcal{O}f, g\rangle_{L^2} = \langle f, \mathcal{O}^+ g\rangle_{L^2}$.

1. In many score-based modeling approaches, e.g. (Song and Ermon, 2019; Song et al., 2020) one directly parametrizes the score $\nabla \log q$ instead of the distribution q.

Again, for the special case of the family of operators $\mathcal{O}$ in (3), the integration by parts form of the objective can be easily written down explicitly (the proof is provided in Appendix B):

Lemma 2 (Integration by parts for the GSM in (3)) *The generalized score matching objective in (3) satisfies the equality*

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

2.2. Continous-time Markov Processes

In this section, we introduce the key definitions related to continuous-time Markov chains and diffusion processes:

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 $s, t \geq 0$.*

Definition 3 (Time-homogeneous Markov processes) *A time-homogeneous Markov process $(X_t)_{t\geq 0}$ on state space Ω is defined by a Markov semigroup $\{P_t(x, y)\}_{t\geq 0}$ as follows: for any measurable $A \subseteq \Omega$, $\Pr(X_{s+t} \in A | X_s = x) = P_t(x, A) = \int_A P_t(x, dy)$. Moreover, P_t can be thought of as acting on a function g as $(P_t g)(x) = \mathbb{E}_{P_t(x,\cdot)}[g(y)] = \int_\Omega g(y)P_t(x, dy)$. 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 $dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$, where B_t is standard Brownian motion in $\mathbb{R}^d$. In fact, a completeness result due to Ma et al. (2015) states that we can characterize *all* Itô diffusions whose stationary distribution is $p(x) \propto \exp(-f(x))$ as "preconditioned" Langevin diffusion. Precisely:

Theorem 3 (Characterization of Itô diffusions w/ 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:*

$$dX_t = \left(-(D(X_t) + Q(X_t))\nabla f(X_t) + \Gamma(X_t)\right)dt + \sqrt{2D(X_t)}dB_t \tag{5}$$

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.

2.3. Dirichlet forms and Poincaré inequalities

Definition 4 *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*

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

respectively. We will use the shorthand $\mathcal{E}(g) := \mathcal{E}(g,g)$.

By Itô's Lemma, the generator of diffusions of the form (5) have the form:

$$(\mathcal{L}g)(x) = \langle -[D(x) + Q(x)]\nabla f(x) + \Gamma(x), \nabla g(x)\rangle + \mathrm{Tr}(D(x)\nabla^2 g(x)) \tag{6}$$

The Dirichlet form for diffusions of the form (5) also has a very convenient form:

Lemma 4 (Dirichlet form of continuous Markov Process) *An Itô diffusion of the form (5) has a Dirichlet form $\mathcal{E}(g) = \mathbb{E}_p\|\sqrt{D(x)}\nabla g(x)\|_2^2$. Notably, for Langevin diffusion, the Dirichlet form is just the l_2 norm of ∇g: $\mathcal{E}(g) = \mathbb{E}_p\|\nabla g\|_2^2$.*

For a general diffusion of the form (5), we can think of $D(x)$ as a (point-specific) preconditioner, specifying the norm with respect to which to measure ∇g. The proof of this lemma is given in Appendix B. Finally, we define the Poincaré constant:

Definition 5 (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* [2] *$\mathcal{E}(g) \geq \frac{1}{C}\mathrm{Var}_p(g)$. 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.

We will analyze mixing times using a decomposition technique similar to the ones employed in Ge et al. (2018); Moitra and Risteski (2020). Intuitively, 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. An example of such a result is Theorem 6.1 from Ge et al. (2018):

Theorem 5 (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 $\mathrm{Var}_{p_j}(g) \leq C\mathcal{E}_j(g,g)$.*

2. We will implicitly assume this condition whenever we discuss Poincaré inequalities.

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*

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

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

2.4. Asymptotic efficiency

We will also use some classical results about asymptotic normality of M-estimators, under standard identifiability and differentiability conditions. Namely, we will be considering estimators defined as $\min_{\theta\in\Theta} L(\theta)$, where $L(\theta) = \mathbb{E}_p[\ell_\theta(x)]$. In general, we will denote by n the number of samples, and $\hat{\mathbb{E}}$ will denote an empirical average, that is the expectation over the n training samples. Finally, $\hat{\theta}_n$ will denote $\min_{\theta\in\Theta} \hat{\mathbb{E}}[\ell_\theta(x)]$ when the number of samples is n. In Section A.2 we recall sufficient conditions for $\hat{\theta}_n$ to be asymptotically normal, and an expression for the asymptotic covariance.

3. Main Results: A Framework for Analyzing Generalized Score Matching

The goal of this section is to provide a general framework that provides a bound on the sample complexity of a generalized score matching objective with operator $\mathcal{O}$, under the assumption that some Markov process with generator $\mathcal{L}$ mixes fast. Precisely, we will show:

Theorem 6 (Main, sample complexity bound) *Consider an Itô diffusion of the form (5) 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 $\{p_\theta : \theta \in \Theta\}$ that satisfies:*

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:*

$$\begin{aligned} l_\theta(x) &:= \frac{1}{2}\left\|\frac{\mathcal{O}p_\theta(x)}{p_\theta(x)}\right\|_2^2 - 2\mathcal{O}^+\left(\frac{\mathcal{O}p_\theta(x)}{p_\theta(x)}\right) \\ &= \frac{1}{2}\left[\|\sqrt{D(x)}\nabla_x \log p_\theta(x)\|^2 + 2div\left(D(x)\nabla_x \log p_\theta(x)\right)\right] \end{aligned}$$

   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 $p_{\theta^*} = p$.*

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

Remark 7 *The two terms on the right hand sides qualitatively capture two intuitive properties necessary for a good sample complexity: the factor involving the covariances can be thought of as a smoothness term capturing how regular the score is as we change the parameters in the family we are fitting; the C_P term captures how the error compounds as we "extrapolate" the score into a probability density function.*

Remark 8 *This theorem generalizes Theorem 2 in Koehler et al. (2022), who show the above only in the case of $\mathcal{L}$ being the generator of Langevin diffusion and $\mathcal{O} = \nabla_x$, i.e. when D_{GSM} is the standard score matching loss. Furthermore, they only consider the case of p_θ being an exponential family, i.e. $p_\theta(x) \propto \exp(\langle \theta, T(x) \rangle)$ for some sufficient statistics $T(x)$. Finally, just as in Koehler et al. (2022), we can get a tighter bound by replacing C_P by the restricted Poincaré constant, which is the Poincaré constant when considering only the functions of the form $\langle w, \nabla_\theta \log p_\theta(x)_{|_{\theta=\theta^*}} \rangle$.*

Remark 9 *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 $\|\hat{\theta}_n - \theta^*\|_2^2 \leq \frac{\operatorname{Tr}(\Gamma_{SM})}{n}$.*

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 19 for completeness).From that lemma, when an estimator $\hat{\theta} = \arg\min \hat{\mathbb{E}} l_\theta(x)$ is asymptotically normal, we have $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla_\theta^2 L(\theta^*))^{-1} \operatorname{Cov}(\nabla_\theta \ell(x; \theta^*))(\nabla_\theta^2 L(\theta^*))^{-1}\right)$, 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 is a fairly straightforward calculation, which results in the following Lemma, proven in Appendix C.

Lemma 10 (Bound on smoothness) *For $l_\theta(x)$ defined in Theorem 6,*

$$\textit{cov}(\nabla_\theta l_\theta(x)) \precsim \textit{cov}\left(\nabla_\theta \nabla_x \log p_\theta(x) D(x) \nabla_x \log p_\theta(x)\right) + \textit{cov}\left(\nabla_\theta \nabla_x \log p_\theta(x)^\top \textit{div}(D(x))\right) + \textit{cov}\left(\nabla_\theta \operatorname{Tr}[D(x)\Delta \log p_\theta(x)\right)$$

The bound on the Hessian is where the connection to the Poincaré constant manifests. Namely, we show:

Lemma 11 (Bounding Hessian) *Let the operators $\mathcal{O}, \mathcal{L}$ be such that for every vector w, the function $g(x) = \langle w, \nabla_\theta \log p_\theta(x)_{|_{\theta=\theta^*}} \rangle$ satisfies $\mathbb{E}_p \|\mathcal{O} g\|^2 = -\langle g, \mathcal{L} g \rangle_p$. Then it holds that*

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

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:

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

① follows from a straightforward calculation (in Lemma 5), ② follows from the definition of Poincaré inequality of a diffusion process with Dirichlet form derived in Lemma 4, 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. ■

4. Main Results: Benefits of Annealing

In this section, we instantiate the framework from the previous section to the specific case of a Markov process, Continuously Tempered Langevin Dynamics, which is a close relative of simulated tempering (Marinari and Parisi, 1992), where the number of "temperatures" is infinite, and we temper by convolving with Gaussian noise. We show that the generalized score matching loss corresponding to this Markov process mixes in time $\mathrm{poly}(D, d)$ for a mixture of K Gaussians (with identical covariance) in d dimensions, and means in a ball of radius D. More precisely, in this section, we will consider the following family of distributions:

Assumption 1 *Let p_0 be a d-dimensional Gaussian distribution with mean 0 and covariance Σ. We will assume the data distribution p is a K-Gaussian mixture, namely $p = \sum_{i=1}^K w_i p_i$, where $p_i(x) = p_0(x - \mu_i)$, i.e. a shift of the distribution p_0 so its mean is μ_i. We will assume the means μ_i lie within a ball with diameter D. We will denote the min and max eigenvalues of covariance with $\lambda_{\min}(\Sigma) = \lambda_{\min}$ and $\lambda_{\max}(\Sigma) = \lambda_{\max}$. We will denote the min and max mixture proportion with $\min_i w_i = w_{\min}$ and $\max_i w_i = w_{\max}$. Let $\Sigma_\beta = \Sigma + \beta \lambda_{\min} I_d$ be the shorthand notation of the covariance of individual Gaussian at temperature β.*

Mixtures of Gaussians are one of the most classical distributions in statistics—and they have very rich modeling properties. For instance, they are universal approximators in the sense that any distribution can be approximated (to any desired accuracy), if we consider a mixture with sufficiently many components (Alspach and Sorenson, 1972). A mixture of K Gaussians is also the prototypical example of a distribution with K modes — the shape of which is determined by the covariance of the components.

Note at this point 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 need not necessarily fit the natural unknown parameters (the means, covariances and weights). The primary reason this family of distributions is convenient for technical analysis is a closure property under convolutions: a convolution of a Gaussian mixture with a Gaussian produces another Gaussian mixture. Namely, the distributivity property of the convolution operator implies:

Proposition 1 (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* $\Sigma + \sigma^2 I$.

The Markov process we will be analyzing (and the corresponding score matching loss) is a continuous-time analog of the Simulated Tempering Langevin chain introduced in Ge et al. (2018):

Definition 6 (Continuously Tempered Langevin Dynamics (CTLD)) *We will consider an SDE over a temperature-augmented state space, that is a random variable* (X_t, β_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 $r : [0, \beta_{\max}] \to \mathbb{R}$ *denotes the distribution over* β, $r(\beta) \propto \exp\left(-\frac{7D^2}{\lambda_{\min}(1+\beta)}\right)$ *and* $\beta_{\max} = \frac{14D^2}{\lambda_{\min}} - 1$. *Let* $p^\beta := p * \mathcal{N}(0, \beta\lambda_{\min} I_d)$ *denotes the distribution* p *convolved with a Gaussian of covariance* $\beta\lambda_{\min} I_d$. *Furthermore,* $L(dt)$ *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, such that the stationary distribution of this SDE is* $p(x, \beta) = r(\beta)p^\beta(x)$ *(Saisho, 1987)*[3].

If we ignore the boundary reflection term, the updates for CTLD are simply Langevin dynamics applied to the distribution $p(x, \beta)$, and $r(\beta)$ specifies the distribution over the different levels of noise. CTLD can be readily seen as a "continuous-time" analogue of the usual simulated tempering chain. Namely, in the usual (discrete-time) simulated tempering (Lee et al., 2018; Ge et al., 2018), the tempering chain has two types of moves: one which evolves the position in the current temperature, and one which tries to change the temperatures, followed by a Metropolis Hastings filtering step.

We point out several similarities and crucial differences with the chain proposed in Ge et al. (2018). The chain in Ge et al. (2018) has a finite number of temperatures and the distribution in each temperature is defined as scaling the log-pdf, rather than convolution with a Gaussian—this is because the mode of access in Ge et al. (2018) is the gradient of the log-pdf, whereas in score matching, we have samples from the distribution. The distributions in Ge et al. (2018) are geometrically spaced out—so β being distributed as $\exp(-\Theta(\beta))$ in our case can be thought of as a natural continuous analogue.

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 6). The operator $\mathcal{O}$ that corresponds to the CTLD is also easy to derive:

Proposition 2 *The generalized score matching loss with* $\mathcal{O} = \nabla_{x,\beta}$ *verifies* $\left[\nabla^2_\theta D_{GSM}(p, p_{\theta^*})\right]^{-1} \preceq C_P \Gamma_{MLE}$. *Moreover,*

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

3. The existence of the boundary measure is a standard result of reflecting diffusion processes via solutions to the Skorokhod problem (Saisho, 1987).

Proof The first equality follows as a special case of Langevin on the lifted distribution. The second equality 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 28). ■

This loss was derived from first principles from the Markov Chain-based framework in Section 3, 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 $r(\beta)$.

With this setup in mind, we can proceed to the main technical results of this section.

Theorem 12 (Poincaré constant of CTLD) *Under Assumption 1, the Poincaré constant of CTLD C_P enjoys the upper bound $C_P \lesssim D^{22} d^2 \lambda_{\max}^9 \lambda_{\min}^{-2}$*

Note that perhaps surprisingly, the above result has *no dependence* on the number of components, or on the smallest component weight $w_{\min}$—only on the diameter D, the ambient dimension d, and $\lambda_{\min}$ and $\lambda_{\max}$. The results in Ge et al. (2018) have a dependence on $w_{\min}$, but in their model, it's not possible to query convolutions of the pdf with a Gaussian. This result can be seen as a "time-homogenous" analogue of recent results (Lee et al., 2023) that the reverse SDE (which is time-inhomogenous) converges to the data distribution in polynomial time under minimal assumptions (Lipschitzness of the score). This result is of independent technical interest as it illustrates the power of having an oracle for convolutions of the target distribution.

To get a complete bound on the asymptotic sample complexity of generalized score matching, according to the framework from Lemma 3, we also need to bound the smoothness terms as in Lemma 10. These terms of course depend on the choice of parametrization for the family of distributions we are fitting — in particular, there is no "canonical" parametrization for multimodal distributions. To get a quantitative sense for how the smoothness might scale, we will consider one natural parametrization for a mixture:

Assumption 2 *Consider the case of learning unknown means*[4]*, such that the parameters to be learned are a vector $\theta = (\mu_1, \mu_2, \ldots, \mu_K) \in \mathbb{R}^{dK}$.*

With this parametrization, the smoothness term can be bounded as follows:

Theorem 13 (Smoothness under the natural parameterization) *Under Assumptions 1 and 2, the smoothness defined in Theorem 6 enjoys the upper bound*

$$\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)_{|_{\theta=\theta^*}}\|_{OP} + \|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)_{|_{\theta=\theta^*}}\|_{OP} \lesssim \mathrm{poly}\left(D, d, \lambda_{\min}^{-1}\right)$$

Finally, we show that the generalized score matching loss is asymptotically normal. The proof of this is in Appendix F, and proceeds by verifying the conditions of Lemma 19. Putting this together with the Poincaré inequality bound Theorem 12 and Theorem 6, we get a complete bound on the sample complexity of the generalized score matching loss with $\mathcal{O}$:

Theorem 14 (Main, Polynomial Sample Complexity Bound of CTLD)
Let the data distribution p satisfy Assumption 1. Then, the generalized score matching loss defined in Proposition 7 with parametrization as in Assumption 2 satisfies:

4. 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.

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:*

   $\theta^* = (\mu_1, \mu_2, \ldots, \mu_K) \in \Theta^*$ *iff* $\exists \pi : [K] \to [K]$ *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*

   $$C_0 = \{\theta \in C : p_\theta(x) = p(x) \text{ almost everywhere } \}$$

   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 a matrix Γ_{SM} *satisfying* $\|\Gamma_{SM}\|_{OP} \leq \text{poly}\left(D, d, \lambda_{\max}, \lambda_{\min}^{-1}\right) \|\Gamma_{\boldsymbol{MLE}}\|_{OP}^2$.

We provide some brief comments the theorem:

- The goal of this result is *not* to provide a new algorithm for learning mixtures of Gaussians, but to provide a paradigm case in which annealing improves the statistical behavior of score matching. In fact, the generalized score matching loss is not necessarily convex, and our result is only statistical in nature. Moreover, from the point of view of score matching, the statistical complexity relative to MLE worsens when the modes are well-separated — which is exactly the case when many algorithms for learning mixtures of Gaussians tend to work. (Appendix I).

- Condition (1) is the standard identifiability condition (Yakowitz and Spragins, 1968) for mixtures of Gaussians: the means are identifiable up to "renaming" the components. This is 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. $\forall i \in [K], \mu_i = \mu_i^*$. Condition (2) 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.

- Condition (3) characterizes the sample complexity of minimizers in the neighborhood of each of the points in Θ^*, and is a consequence of the CTLD Poincaré inequality estimate (Theorem 12) and the smoothness estimate (Theorem 13). 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. [5]

5. 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.

4.1. Proof Sketch: Bounding the Poincaré Constant of CTLD

In this section, we will sketch the proof of Theorem 12. 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 will proceed by applying the decomposition Theorem 5 to CTLD. Towards that, we denote by $\mathcal{E}_i$ the Dirichlet form corresponding to Langevin with stationary distribution $p(x, \beta|i)$. By Propositions 6, 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 5. 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). To verify condition (3), we will show the projected chain "between" the components (as defined in Theorem 5) mixes fast. We will expand on each of these parts in turn.

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 15 *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)), we get the claim of the lemma. The details are in Appendix E.1.

Fast mixing between components: Next, we show the "projected" chain between the components mixes fast (full proofs are in Appendix E.2):

Lemma 16 (Poincaré constant of projected chain) *Define the projected chain $\bar{M}$ over $[K]$ with transition probability*

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

where $\chi^2_{\max}(p, q) = \max\{\chi^2(p, q), \chi^2(q, p)\}$. 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:

Lemma 17 *For every $i, j \in [K]$, we have $\chi^2(p(x, \beta|i), p(x, \beta|j)) \leq 14 D^2 \lambda_{\min}^{-1}$.*

4.2. Proof Sketch: Bounding the Smoothness Terms

To obtain Theorem 13, we note $\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)\|_{OP}$ and $\|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)\|_{OP}$ 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 28). The main technical tools involved are: (1) the convexity of the *perspective map* to relate derivatives of the mixture to derivatives of the components (Lemma 37); (2) bounds on derivatives of the components via Hermite polynomial machinery (Lemma 38); (3) bounds on logarithmic derivatives via higher-order versions of the Faá di Bruno formula (Constantine and Savits, 1996). The complete proofs are in Appendix G.

5. Conclusion

In this paper, we provide a general framework about designing statistically efficient generalized score matching losses from fast-mixing Markov Chains. As a demonstration of the power of the framework, we provide the first formal analysis of the statistical benefits of annealing for score matching for multimodal distributions. A core technical result of this part is bounding the mixing time for a continuously tempered version of Langevin diffusion. The framework can be likely used to analyze other common continuous and discrete Markov Chains (and corresponding generalized score losses), like underdamped Langevin dynamics and Gibbs samplers.

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.

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.

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.

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.

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.

Contents

1 Introduction **1**

2 Preliminaries **3**
2.1 Generalized Score Matching . . . 3
2.2 Continous-time Markov Processes . . . 4
2.3 Dirichlet forms and Poincaré inequalities . . . 5
2.4 Asymptotic efficiency . . . 6

3 Main Results: A Framework for Analyzing Generalized Score Matching **6**

4 Main Results: Benefits of Annealing **8**
4.1 Proof Sketch: Bounding the Poincaré Constant of CTLD . . . 12
4.2 Proof Sketch: Bounding the Smoothness Terms . . . 13

5 Conclusion **13**

A Preliminaries **18**
A.1 Continuous Markov Chain Decomposition . . . 18
A.2 Asymptotic normality of M-estimators . . . 18
A.3 Hermite Polynomials . . . 19
A.4 Logarithmic derivatives . . . 21
A.5 Moments of mixtures and the perspective map . . . 22

B Derivations of generators and score losses for diffusions **23**

C A Framework for Analyzing Generalized Score Matching **26**

D Overview of Continuously Tempered Langevin Dynamics **27**

E Polynomial mixing time bound: proof of Theorem 12 **29**
E.1 Mixing inside components: proof of Lemma 15 . . . 30
E.2 Mixing between components: proof of Lemma 16 . . . 32

F Asymptotic normality of generalized score matching for CTLD **34**

G Polynomial smoothness bound: proof of Theorem 13 **39**
G.1 Overview . . . 39
G.2 Detailed proofs . . . 40

H Technical Lemmas **43**
H.1 Moments of a chi-squared random variable . . . 43

I Related work **43**

Appendix A. Preliminaries

A.1. Continuous Markov Chain Decomposition

The Poincaré constant bounds we will prove will also use a "continuous" version of the decomposition Theorem 5, which also appeared in Ge et al. (2018):

Theorem 18 (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 normality of M-estimators

The following Theorem recalls classical sufficient conditions for asymptotic normality of M-estimators, and the expression for the covariance matrix of the resulting normal distribution:

Lemma 19 (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 7 (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 3 (Integral representation of Hermite polynomial, (Holmquist, 1996)) *The Hermite polynomial H_k defined in Definition 7 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:

$$H_k(-x;\Sigma) = (-1)^k H_k(x;\Sigma) \tag{7}$$

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 G.

Lemma 20 *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 7, 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:

$$\frac{\nabla_x^k \phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (-1)^k (\Sigma^{-1})^{\otimes k} H_k(x-\mu;\Sigma) \tag{8}$$

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 (7), we get:

$$\frac{\nabla_\mu^k \phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (\Sigma^{-1})^{\otimes k} H_k(x-\mu;\Sigma) \tag{9}$$

Combining (8) and (9), 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 3, 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 21 (Lemma 38 restated) *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 20 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 46 yields the desired result. ■

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

Lemma 22 *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 21, the result follows.

■

A.4. Logarithmic derivatives

Finally, we will need similar bounds for logarithic 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 4 (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 23 (Restatement of Lemma 39) *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 24 *(Convexity of perspective, Boyd and Vandenberghe (2004)) Let* f *be a convex function. Then, its corresponding* perspective map $g(u, v) := vf\left(\frac{u}{v}\right)$ *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 25 (Restatement of Lemma 37) *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 24. Then, we proceed as follows:

$$\begin{aligned}
\mathbb{E}_{(x,\beta)\sim p(x,\beta)}\left\|\frac{(Dp_\theta)(x|\beta)}{p_\theta(x|\beta)}\right\|^k &= \mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p(x|\beta)}\left\|\frac{(Dp_\theta)(x|\beta)}{p_\theta(x|\beta)}\right\|^k \\
&= \mathbb{E}_{\beta\sim r(\beta)}\int g((Dp_\theta)(x|\beta), p_\theta(x|\beta))dx \\
&= \mathbb{E}_{\beta\sim r(\beta)}\int g\left(\sum_{i=1}^K w_i(Dp_\theta)(x|\beta,i), \sum_{i=1}^K w_i p_\theta(x|\beta,i)\right)dx \quad (10)\\
&\leq \mathbb{E}_{\beta\sim r(\beta)}\int \sum_{i=1}^K w_i g((Dp_\theta)(x|\beta,i), p_\theta(x|\beta,i))dx \quad (11)\\
&= \mathbb{E}_{\beta\sim r(\beta)}\sum_{i=1}^K w_i \mathbb{E}_{x\sim p(x|\beta,i)}\left\|\frac{(Dp_\theta)(x|\beta,i)}{p_\theta(x|\beta,i)}\right\|^k \\
&\leq \max_{\beta,i}\mathbb{E}_{x\sim p(x|\beta,i)}\left\|\frac{(Dp_\theta)(x|\beta,i)}{p_\theta(x|\beta,i)}\right\|^k
\end{aligned}$$

where (10) follows by linearity of D, and (11) by convexity of the function g. ■

Appendix B. Derivations of generators and score losses for diffusions

First, we derive the Dirichlet form of Itô diffusions of the form (5). Namely, we show:

Lemma 26 (Dirichlet form of continuous Markov Process) *For an Itô diffusion of the form (5), 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 (5) is:

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

The Dirichlet form is given by

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

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

$$\begin{aligned}
\mathrm{I} &= \int p(x)\langle D(x)\nabla f(x), \nabla g(x)\rangle g(x)dx \quad (12)\\
&+ \int p(x)\langle Q(x)\nabla_x f(x), \nabla g(x)\rangle g(x)dx \quad (13)\\
&- \int p(x)\sum_{i,j}\partial_j D_{ij}(x)\partial_i g(x) g(x)dx \quad (14)\\
&- \int p(x)\sum_{i,j}\partial_j Q_{ij}(x)\partial_i g(x) g(x)dx \quad (15)
\end{aligned}$$

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

$$
\begin{aligned}
\text{II} &= -\int p(x)\,\mathrm{Tr}(D(x)\nabla^2 g(x))g(x)dx \\
&= -\int p(x)\left(\sum_{i,j} D_{ij}(x)\partial_{ij}g(x)\right)g(x)dx \\
&= -\sum_{i,j}\int p(x)D_{ij}(x)g(x)\partial_{ij}g(x)dx \\
&= -\sum_{i,j}\left(p(x)D_{ij}(x)g(x)\partial_i g(x)\Big|_{x=-\infty}^{\infty} - \int \partial_j[p(x)D_{ij}(x)g(x)]\partial_i g(x)dx\right) \\
&= \sum_{i,j}\int \partial_j[p(x)D_{ij}(x)g(x)]\partial_i g(x)dx \\
&= \sum_{i,j}\int \partial_j p(x)D_{ij}(x)g(x)\partial_i g(x)dx \qquad (16) \\
&+ \sum_{i,j}\int p(x)\partial_j D_{ij}(x)g(x)\partial_i g(x)dx \qquad (17) \\
&+ \sum_{i,j}\int p(x)D_{ij}(x)\partial_j g(x)\partial_i g(x)dx \qquad (18)
\end{aligned}
$$

The term 16 cancels out with term 12,

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

The term 17 cancels out with the term 14.

For term 13,

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

Combining term 13 and term 15,

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

In the end, we are only left with term 18:

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

■

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

Lemma 27 (Integration by parts for the GSM in (3)) *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 (under suitable decay at infinity):

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

■

Appendix C. A Framework for Analyzing Generalized Score Matching

Proposition 5 (Hessian of GSM loss) *The Hessian of D_{GSM} satisfies*

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

Proof By a straightforward calculation, we have:

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

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

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 $\theta = \theta^*$.

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

■

Proof We have

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

By Lemma 2 in Koehler et al. (2022), we also have

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

which completes the proof. ■

Appendix D. Overview of Continuously Tempered Langevin Dynamics

Proposition 6 (Dirichlet form for CTLD) *The Dirichlet form corresponding to CTLD has the form*

$$\begin{aligned}\mathcal{E}(f(x,\beta)) &= \mathbb{E}_{p(x,\beta)}\|\nabla f(x,\beta)\|^2 &(19)\\ &= \mathbb{E}_{r(\beta)}\mathcal{E}_\beta(f(\cdot,\beta)) &(20)\end{aligned}$$

where $\mathcal{E}_\beta$ is the Dirichlet form corresponding to the Langevin diffusion (Lemma 4) with stationary distribution $p(x|\beta)$.

Proof Equation 19 follows from the fact that CTLD is just a (reflected) Langevin diffusion with stationary distribution $p(x,\beta)$. Equation 20 follows from the tower rule of expectation and the definition of the Dirichlet form for Langevin from Proposition 4. ■

Lemma 28 (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, $q_t = q_0 * N(0, \lambda_{\min} t I)$. 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,

$$\nabla_\beta \log p^\beta(x) = \frac{\lambda_{\min} \operatorname{Tr}(\nabla_x^2 p^\beta(x))}{p^\beta(x)} \tag{21}$$

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 (21), 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 7 (Integration-by-part Generalized Score Matching Loss for CTLD) *The loss D_{GSM} in the integration by parts form (Lemma 1) as:*

$$D_{GSM}\left(p, p_\theta\right) = \mathbb{E}_p l_\theta(x, \beta) + K_p$$

where

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

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 [Proof of Proposition 7]

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

By Lemma 28, $\nabla_\beta \log p_\theta(x|\beta)$ is a function of partial derivatives of the score $\nabla_x \log p_\theta(x|\beta)$. Similarly, $\nabla^2_\beta \log p_\theta(x|\beta)$ can be shown to be a function of partial derivatives of the score $\nabla_x \log p_\theta(x|\beta)$ as well:

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

■

Appendix E. Polynomial mixing time bound: proof of Theorem 12

Proof The proof will follow by applying Theorem 5. 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 15, 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:

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

3. (Polynomial mixing for projected chain) To bound the Poincaré constant of the projected chain, by Lemma 16 we have

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

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

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

■

E.1. Mixing inside components: proof of Lemma 15

Proof The proof will follow by an application of a continuous decomposition result (Theorem D.3 in Ge et al. (2018), repeated as Theorem 18) , 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 29, $C_\beta \leq \frac{14D^2}{\pi\lambda_{\min}}$.

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 30.

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:

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

which completes the proof. ■

Lemma 29 (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 30 (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

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

We can apply Lemma 28 to derive explicit expressions for the right-hand side:

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

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 46; and in ③ we plug in the bound on $\beta_{\max}$. ■

E.2. Mixing between components: proof of Lemma 16

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 17 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 32) — 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 31 (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:

$$\begin{aligned}
\chi^2(p(x,\beta|i),p(x,\beta|j)) &= \int \left(\frac{p(x,\beta|i)}{p(x,\beta|j)}-1\right)^2 p(x,\beta|i)dxd\beta \\
&= \int_0^{\beta_{\max}}\int_{-\infty}^{+\infty}\left(\frac{p(x|\beta,i)r(\beta)}{p(x|\beta,j)r(\beta)}-1\right)^2 p(x|\beta,i)r(\beta)dxd\beta \\
&= \int_0^{\beta_{\max}} \chi^2(p(x|\beta,i),p(x|\beta,j))r(\beta)d\beta \\
&\le \int_0^{\beta_{\max}} \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) r(\beta)d\beta \qquad (22)\\
&= \int_0^{\beta_{\max}} \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right)\frac{1}{Z(D,\lambda_{\min})}\exp\left(-\frac{7D^2}{\lambda_{\min}(1+\beta)}\right)d\beta \\
&= \frac{\beta_{\max}}{Z(D,\lambda_{\min})}
\end{aligned}$$

where in Line 22, we apply our Lemma 32 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:

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

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

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

Finally, we get the desired bound

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

■

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

Lemma 32 (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:

$$\begin{aligned}
&\chi^2(p(x|\beta,i),p(x|\beta,j)) \\
&\leq \frac{\det(\Sigma_\beta)^{\frac{1}{2}}}{\det(\Sigma_\beta)} \det\left(\Sigma_\beta^{-1}\right)^{-\frac{1}{2}} \\
&\exp\left(\frac{1}{2}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)^\top (\Sigma_\beta^{-1})^{-1}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right) + \frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i - \mu_j^\top\Sigma_\beta^{-1}\mu_j\right) \qquad (23)\\
&= \exp\left(\frac{1}{2}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)^\top (\Sigma_\beta^{-1})^{-1}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right) + \frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i\right) \\
&\exp\left(-\mu_j^\top\Sigma_\beta^{-1}\mu_j\right) \\
&\leq \exp\left(\frac{1}{2}(2\mu_j-\mu_i)^\top\Sigma_\beta^{-1}(2\mu_j-\mu_i) + \frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i\right) \qquad (24)\\
&\leq \exp\left(\frac{\|2\mu_j-\mu_i\|_2^2 + \|2\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right) \\
&\leq \exp\left(\frac{(\|2\mu_j\|_2 + \|\mu_i\|_2)^2 + 4\|\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right) \\
&\leq \exp\left(\frac{2\|2\mu_j\|_2^2 + 2\|\mu_i\|_2^2 + 4\|\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right) \\
&\leq \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right)
\end{aligned}$$

In Equation 23, we apply Lemma G.7 from Ge et al. (2018) for the chi-square divergence between two Gaussian distributions. In Equation 24, we use the fact that Σ_β^{-1} is PSD. ■

Appendix F. 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 33 (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 7 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*

$$C_0 = \{\theta \in C : p_\theta(x) = p(x) \text{ almost everywhere } \}$$

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 34: 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 36: it follows from a uniform law of large numbers.

Finally, Part 3 follows from an application of Lemma 19—so we verify the conditions of the lemma are satisfied. The gradient bounds on l_θ are verified Lemma 35—and it largely follows by moment bounds on gradients of the score derived in Section G. Uniform law of large numbers is shown in Lemma 36, and the the existence of Hessian of $L = D_{GSM}$ is trivially verified. ■

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 34 (Uniqueness of optima) *Suppose for* $\theta := (\mu_1, \mu_2, \ldots, \mu_K)$ *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 $\theta = (\mu_1, \mu_2, \ldots, \mu_K)$ 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 $p^\beta_{\theta^*}$, a convolution of p_{θ^*} with a Gaussian with covariance $\beta\lambda_{\min} I_d$; and the distribution being fitted as p^β_θ. Thus, the above argument implies that if $\theta \neq \theta^*$, we have $D_{GSM}(\theta) > D_{GSM}(\theta^*)$, as we need. ■

Lemma 35 (Gradient bounds of l_θ) *Let $l_\theta(x,\beta)$ be as defined in Proposition 7. 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 7,

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left\|l^1_\theta(x,\beta)\right\|_2^2 &\lesssim \mathbb{E}\left\|\nabla_x \log p_\theta(x,\beta)\right\|_2^4 + \mathbb{E}\left(\Delta_x \log p_\theta(x,\beta)\right)^2 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)}\left\|\nabla_x \log p_\theta(x|\beta,i)\right\|_2^4 + \max_{\beta,i}\mathbb{E}_{x\sim p(x|\beta,i)}\left(\Delta_x \log p_\theta(x|\beta,i)\right)^2 \qquad (25) \\
&\leq \text{poly}\left(d, \frac{1}{\lambda_{\min}}\right) \qquad (26)
\end{aligned}$$

Where (25) follows by Lemma 25, and (26) follows by combining Corollaries 41 and 23.

The same argument, along with Lemma 28, and the fact that $\max_\beta(\nabla_\beta \log r(\beta))^4 \lesssim D^8\lambda_{\min}^{-4}$ by a direct calculation shows that

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

■

Lemma 36 (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

$$\mathbb{E}_T \sup_{\theta \in \Theta} \left| \widehat{D_{GSM}}(\theta) - D_{GSM}(\theta) \right| \leq \frac{C\left(K, d, D, \frac{1}{\lambda_{\min}}\right)}{\sqrt{n}} \tag{27}$$

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:

$$\begin{aligned}
\mathbb{E}_T \sup_{\theta \in \Theta} \left| \widehat{D_{GSM}}(\theta) - D_{GSM}(\theta) \right| &= \mathbb{E}_T \sup_{\theta \in \Theta} \left| \hat{\mathbb{E}}_T l_\theta(x, \beta) - D_{GSM}(\theta) \right| \\
&= \mathbb{E}_T \sup_{\theta \in \Theta} \left| \hat{\mathbb{E}}_T l_\theta(x, \beta) - \mathbb{E}_{T'} \hat{\mathbb{E}}_{T'} l_\theta(x, \beta) \right| \qquad (28) \\
&\leq \mathbb{E}_{T, T'} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \left(l_\theta(x_i, \beta_i) - l_\theta(x_i', \beta_i') \right) \right| \qquad (29)
\end{aligned}$$

where (28) follows by noting the population expectation can be expressed as the expectation over a choice of a (fresh) training set T', (29) 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

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

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

$$\mathbb{E}_{\{\varepsilon_i\}} \exp\left(\lambda \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(l_\theta(x_i, \beta_i) - l_{\theta'}(x_i, \beta_i) \right) \right) \leq \exp\left(\lambda^2 d(\theta, \theta') \right) \tag{30}$$

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

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

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 (30).

The Dudley entropy integral then gives

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i l_\theta(x_i, \beta_i)\right| \lesssim \int_0^\infty \sqrt{\log N(\epsilon, \Theta, d)}d\epsilon \tag{31}$$

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

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

Plugging this estimate back in (31), 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 35 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 (27). By Markov's inequality, (27) implies that for every $\epsilon > 0$,

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

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. ■

Appendix G. Polynomial smoothness bound: proof of Theorem 13

G.1. Overview

To obtain the polynomial upper bound in Theorem 13, we note the two terms $\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)\|_{OP}$ and $\|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)\|_{OP}$ 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 28). We first provide a high-level overview, then in Section G.2, we provide the full proofs.

The derivatives of x and μ_i are handled by a combination of several techniques. First, we use the convexity of the so-called *perspective map* to relate derivatives of the mixture to derivatives of the components. For example, we show:

Lemma 37 *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*

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

For proof, and more details, see Section A.5. By applying this for various differentiation operators D, this reduces showing bounds for the mixture to showing bounds for the individual components.

Proceeding to the individual components, we can use machinery from Hermite polynomials to get bounds on terms that look like $\frac{Dp(x)}{p(x)}$ for various differentiation operators D. (These quantities are also sometimes called higher-order score functions (Janzamin et al., 2014).) For example, we can show the following:

Lemma 38 *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 $\mathbb{R}^{d^{k_1+k_2}}$.

For more details and proof, see Appendix A.3.

Finally, to get bounds on derivatives of the log-pdf, we use machinery commonly used in analyzing logarithmic derivatives: 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. For example, we can show:

Lemma 39 *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 $\forall i \in d, J_i \leq I_i$.

For more details, and proof, see Appendix A.4.

G.2. Detailed proofs

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

Lemma 40 *For all $k \in \mathbb{N}$, we have:*

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

Proof

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

where the last inequality follows by Lemma 46. ■

Combining this Lemma with Lemmas 21 and 22, we get the following corollary:

Corollary 41

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

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

Finally, we will need the following simple technical lemma:

Lemma 42 *Let X be a vector-valued random variable with finite Var(X). Then, we have*

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

Proof We have

$$\begin{aligned}\|\mathrm{Var}(X)\|_{OP} &= \left\|\mathbb{E}\left[(X-\mathbb{E}[X])\,(X-\mathbb{E}[X])^\top\right]\right\|_{OP} \\ &\le \mathbb{E}\,\|X-\mathbb{E}[X]\|_2^2 \qquad (32)\\ &\le 6\mathbb{E}\|X\|_2^2 \qquad (33)\end{aligned}$$

where (32) follows from the subadditivity of the spectral norm, (33) 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, $\|\mathbb{E}[X]\|_2^2 \le \mathbb{E}\|X\|_2^2$. ■

Given this lemma, it suffices to bound $\mathbb{E}\|(\mathcal{O}\nabla_\theta \log p_\theta)\frac{\mathcal{O}p_\theta}{p_\theta}\|_2^2$ and $\mathbb{E}\|(\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\|_2^2$, which are given by Lemma 43 and Lemma 44, respectively.

Lemma 43

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

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

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

where the last step follows by Cauchy-Schwartz. To bound both factors above, we will essentially first use Lemma 25 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 41 and 23.

Proceeding to the first factor, we have:

$$\begin{aligned}&\mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\mathcal{O}\nabla_\theta \log p_\theta(x,\beta)\|_{OP}^4 \\ &\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\nabla_x\nabla_\theta \log p_\theta(x,\beta)\|_{OP}^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\nabla_\beta\nabla_\theta \log p_\theta(x,\beta)\|_2^4 \qquad (34)\\ &\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\nabla_x\nabla_\theta \log p_\theta(x|\beta)\|_{OP}^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\nabla_\beta\nabla_\theta \log p_\theta(x|\beta)\|_2^4 \\ &\lesssim \max_{\beta,i}\mathbb{E}_{x\sim p(x|\beta,i)}\|\nabla_x\nabla_\theta \log p_\theta(x|\beta,i)\|_{OP}^4 + \max_{\beta,i}\mathbb{E}_{x\sim p(x|\beta,i)}\|\nabla_\beta\nabla_\theta \log p_\theta(x|\beta,i)\|_2^4 \qquad (35)\\ &\le \mathrm{poly}(d, 1/\lambda_{\min}) \qquad (36)\end{aligned}$$

where (34) follows from the fact that $\mathcal{O}f = (\nabla_x f, \nabla_\beta f)^T$, (35) follows from Lemma 25, and (36) follows by combining Corollaries 41 and 23 and Lemma 28.

The second factor is handled similarly[6]. We have:

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\frac{\|\mathcal{O}p_\theta(x,\beta)\|_2}{p_\theta(x,\beta)} \right)^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\frac{\|\nabla_x p_\theta(x,\beta)\|_2}{p_\theta(x,\beta)} \right)^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\frac{\nabla_\beta p_\theta(x,\beta)}{p_\theta(x,\beta)} \right)^4 \\
&= \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \log p_\theta(x,\beta)\|_2^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\nabla_\beta \log p_\theta(x,\beta)\right)^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \log p_\theta(x|\beta)\|_2^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\nabla_\beta \log p_\theta(x|\beta)\right)^4 + \mathbb{E}_{\beta\sim r(\beta)} \left(\nabla_\beta \log r(\beta)\right)^4 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\nabla_x \log p_\theta(x|\beta,i)\|_2^4 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} (\nabla_\beta \log p_\theta(x|\beta,i))^4 + \max_{\beta} (\nabla_\beta \log r(\beta))^4 \qquad (37) \\
&\leq \mathrm{poly}(d, D, 1/\lambda_{\min}) \qquad (38)
\end{aligned}$$

where (37) follows from Lemma 25, and (38) follows by combining Corollaries 41 and 23 and Lemma 28, 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 (36) and (38) complete the proof of the lemma. ■

Lemma 44

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

Proof Since $\mathcal{O}^+\mathcal{O} = \Delta_{(x,\beta)}$, we have

$$\begin{aligned}
&(\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta(x,\beta) \\
&= \nabla_\theta \Delta_{(x,\beta)} \log p_\theta(x,\beta) \qquad (39) \\
&= \nabla_\theta \Delta_x \log p_\theta(x,\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x,\beta) \qquad (40) \\
&= \nabla_\theta \Delta_x \log p_\theta(x|\beta) + \nabla_\theta \Delta_x \log r(\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta) + \nabla_\theta \nabla_\beta^2 \log r(\beta) \\
&= \nabla_\theta \Delta_x \log p_\theta(x|\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta) \qquad (41)
\end{aligned}$$

where (39) follows by exchanging the order of derivatives, (40) since β is a scalar, so the Laplacian just equals to the Hessian, (41) by dropping the derivatives that are zero in the prior expression.

To bound both summands above, we will essentially first use Lemma 25 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 23 and 41. Precisely, we have:

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|(\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\|_2^2 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\theta \operatorname{Tr}(\nabla_x^2 \log p_\theta(x|\beta))\|_2^2 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta)\|_2^2 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\| \frac{\nabla_\theta \Delta_x p_\theta(x|\beta,i)}{p_\theta(x|\beta,i)} \right\|_2^2 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\| \frac{\nabla_\theta \nabla_x p_\theta(x|\beta,i)}{p_\theta(x|\beta,i)} \right\|_{OP}^4 \qquad (42) \\
&\leq \mathrm{poly}(d, 1/\lambda_{\min}) \qquad (43)
\end{aligned}$$

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

where (42) follows from Lemma 25 and Lemma 28, and (43) follows by combining Corollaries 23 and 41.

■

Appendix H. Technical Lemmas

H.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 45 (Norm of Gaussian) *The random variable z enjoys the bound*

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

Proof

$$\begin{aligned}
\left(\mathbb{E}\|z\|_2\right)^2 &\leq \mathbb{E}\|z\|_2^2 && (44)\\
&= \mathbb{E}\sum_{i=1}^{d} z_i^2 \\
&= d && (45)
\end{aligned}$$

where (44) follows from Jensen, and (45) by plugging in the mean of a chi-squared distribution with d degree of freedom. ■

Lemma 46 (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.

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

■

Appendix I. 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" form of the Poincaré inequality suffices for good sample complexity.

On the empirical side, Song and Ermon (2019) proposed an annealed version of score matching, in which they proposed fitting the scores of the distribution convolved with multiple levels of Gaussian noise. They also proposed using the learned scores to sample via annealed Langevin dynamics, which uses samples from Langevin at higher levels of Gaussian convolution as a warm start for running a Langevin at lower levels of Gaussian convolution. Subsequently, this line of work developed into score-based diffusion models (Song et al., 2020), which can be viewed as a "continuously annealed" version of (Song and Ermon, 2019).

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) mixes in time $\text{poly}(K)$.

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 (5). 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 4. 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).