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

Yilong Qin[1] **Andrej Risteski**[1]

Abstract

We show a close connection between the mixing time of an *arbitrary* Markov process with generator $\mathcal{L}$ and an appropriately chosen *generalized score matching loss* that tries to fit $\frac{\mathcal{O}p}{p}$. In the special case of $\mathcal{O} = \nabla_x$, and $\mathcal{L}$ being the generator of Langevin diffusion, this generalizes and recovers the results from (Koehler et al., 2022). If $\mathcal{L}$ corresponds to a Markov process corresponding to a continuous version of simulated tempering, we show the corresponding generalized score matching loss is a Gaussian-convolution annealed score matching loss, akin to the one proposed in (Song & Ermon, 2019). Moreover, we show that if the distribution being learned is a finite mixture of Gaussians in d dimensions with a shared covariance, the sample complexity of annealed score matching is polynomial in the ambient dimension, the diameter of the means, and the smallest and largest eigenvalues of the covariance—obviating the Poincaré constant-based lower bounds of the basic score matching loss shown in (Koehler et al., 2022). This is the first result characterizing the benefits of annealing for score matching—a crucial component in more sophisticated score-based approaches like (Song & Ermon, 2019; Song et al., 2020).

1. Introduction

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. As many distributions of interest (e.g. images) are multimodal in nature, the score-matching estimator is likely to be statistically untenable.

The seminal paper by (Song & Ermon, 2019) proposes a way to deal with multimodality and manifold structure in the data by annealing: namely, estimating the scores of convolutions of the data distribution with different levels of Gaussian noise. The intuitive explanation they propose is that the distribution smoothed with more Gaussian noise is easier to estimate (as there are no parts of the distribution that have low coverage by the training data), which should help estimate the score at lower levels of Gaussian noise. However, making this either quantitative or formal seems very challenging. Moreover, (Song & Ermon, 2019) propose annealing as a fix to another issue: using the score to sample from the distribution using Langevin dynamics is also problematic, as Langevin mixes slowly in the presence of multimodality and low-dimensional manifold structure.

In this paper, we show the following:

1. **A general framework** for designing generalized score matching losses with good sample complexity from fast-mixing Markov chains. Precisely, for every time-homogeneous Markov process with generator $\mathcal{L}$ with Poincaré constant C_P, we can choose a linear operator $\mathcal{O}$ (e.g. for self-adjoint $\mathcal{L}$, the choice $\mathcal{O} = (-\mathcal{L})^{1/2}$ works), 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. (We recall that C_P characterizes the mixing time of the Markov process with generator $\mathcal{L}$ in chi-squared distance.)

[1] Machine Learning Department, Carnegie Mellon University. Correspondence to: Yilong Qin <yilongq@cs.cmu.edu>, Andrej Risteski <aristeski@andrew.cmu.edu>.

Accepted to ICML workshop on Structured Probabilistic Inference Generative Modeling, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. Applying this framework to provide the first analysis of the **statistical benefits of annealing for score matching**. Precisely, we exhibit continuously-tempered Langevin, 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 annealed score matching loss (Song & Ermon, 2019; Song et al., 2020), with a particular choice of weighing for the different amounts of Gaussian convolution.

2. 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. In paricular, if $\mathcal{L}$ is self-adjoint, the choice of $\mathcal{O} = (-\mathcal{L})^{1/2}$ will be appropriate. Precisely, we will show:

Theorem 1 (Main, sample complexity bound)**.** *Consider the generalized score matching estimator, defined as in Definition 2 with a continuous operator $\mathcal{O}$, and suppose we are optimizing it over a parametric family $\{p_\theta : \theta \in \Theta\}$. Consider furthermore a Markov semigroup generator $\mathcal{L}$, whose stationary distribution is the data distribution p, with a (finite) Poincaré constant C_P, and such that $\mathcal{O}, \mathcal{L}$ satisfy:*

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

$$l_\theta(x) = \frac{1}{2} \left\| \frac{\mathcal{O} p_\theta(x)}{p_\theta(x)} \right\|_2^2 - 2\mathcal{O}^+ \left(\frac{\mathcal{O} p_\theta(x)}{p_\theta(x)} \right)$$

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

3. *(Compatibility of $\mathcal{O}$ and $\mathcal{L}$) For every vector w, the function $g(x) = \langle w, \nabla_\theta \log p_\theta(x)_{|\theta=\theta^*} \rangle$ satisfies $\mathbb{E}_p \|\mathcal{O} g\|^2 = -\langle g, \mathcal{L} g \rangle_p$. Note, in particular, if $\mathcal{L}$ is self-adjoint, that is $\mathcal{L}^+ = \mathcal{L}$, then $\mathcal{O} = (-\mathcal{L})^{1/2}$ satisfies this property.*

Then, we have:

$$\|\Gamma_{SM}\|_{OP} \leq 2C_P^2 \|\Gamma_{MLE}\|_{OP}^2 (\|\mathrm{Cov}(\mathcal{O}\nabla_\theta \log p_\theta)_{|\theta=\theta^*}\|_{OP} + \|\mathrm{Cov}((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta)_{|\theta=\theta^*}\|_{OP})$$

Remark 1. *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 2. *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 (Definition 7), 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 3. *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*

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

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 4 for completeness).From that lemma, when an estimator $\hat{\theta} = \arg\min \hat{\mathbb{E}} l_\theta(x)$ is asymptotically normal, we have $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla_\theta^2 L(\theta^*))^{-1} \mathrm{Cov}(\nabla_\theta \ell(x; \theta^*))(\nabla_\theta^2 L(\theta^*))^{-1}\right)$, 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 B.

Lemma 1 (Bound on smoothness)**.** *Let $l_\theta(x) =$*

$\frac{1}{2}\left[\left\|\frac{\mathcal{O}p_\theta(x)}{p_\theta(x)}\right\|_2^2 - 2\mathcal{O}^+\left(\frac{\mathcal{O}p_\theta(x)}{p_\theta(x)}\right)\right]$. *Then,*

$$\begin{aligned}\mathrm{Cov}(\nabla_\theta l_\theta(x)) \preceq 2\mathrm{Cov}\left((\mathcal{O}\nabla_\theta \log p_\theta)\frac{\mathcal{O}p_\theta}{p_\theta}\right)\\ + 2\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)\end{aligned}$$

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

Lemma 2 (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*

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

3. Instantiating the framework with Continuously Tempered Langevin Dynamics

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

Definition 1 (Continuously Tempered Langevin Dynamics (CTLD))**.** *We will consider an SDE over a temperature-augmented state space, that is a random variable $(X_t, \beta_t), X_t \in \mathbb{R}^d, \beta_t \in \mathbb{R}^+$, defined as*

$$\begin{cases} dX_t = \nabla_x \log p^\beta(X_t)dt + \sqrt{2}dB_t \\ d\beta_t = \nabla_\beta \log r(\beta_t)dt + \nabla_\beta \log p^\beta(X_t)dt \\ \qquad + \nu_t L(dt) + \sqrt{2}dB_t \end{cases}$$

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).

Remark 4. *The existence of the boundary measure is a standard result of reflecting diffusion processes via solutions to the Skorokhod problem (Saisho, 1987). If we ignore the boundary reflection term, the updates for CTLD are simply Langevin dynamics applied to the distribution $p(x, \beta)$. $r(\beta)$ specifies the distribution over the different levels of noise and is set up roughly so the Gaussians in the mixture have variance $\beta\Sigma$ with probability* $\exp(-\Theta(\beta))$.

Remark 5. *This chain has 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 1 (Dirichlet form for CTLD)**.** *The Dirichlet form corresponding to CTLD has the form*

$$\begin{aligned} \mathcal{E}(f(x,\beta)) &= \mathbb{E}_{p(x,\beta)}\|\nabla f(x,\beta)\|^2 & (1)\\ &= \mathbb{E}_{r(\beta)}\mathcal{E}_\beta(f(\cdot,\beta)) & (2) \end{aligned}$$

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

Proof. Equation 1 follows from the fact that CTLD is just a (reflected) Langevin diffusion with stationary distribution $p(x,\beta)$. Equation 2 follows from the tower rule of expectation and the definition of the Dirichlet form for Langevin from Proposition 4. □

Next, we derive the operator $\mathcal{O}$ that corresponds to the CTLD. We show:

Proposition 2. *The generalized score matching loss with $\mathcal{O} = (-\mathcal{L})^{1/2}$, where $\mathcal{L}$ is the generator of CTLD satisfies*

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

Moreover,

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

Proof. The operator $\mathcal{L}$ corresponding to CTLD is self-adjoint, so the first claim follows by Lemma 2.

For the second claim, the first equality follows since $(-\mathcal{L})^{1/2}$ simply gives the standard score matching loss for the temperature-augmented 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 10). □

This loss was derived from first principles from the Markov Chain-based framework in Section 2, however, it is readily seen that this loss is a "second-order" version of the annealed losses in (Song & Ermon, 2019; Song et al., 2020) — the weights being given by the distribution $r(\beta)$. Additionally, this loss has terms matching "second order" behavior of the distributions, namely $\operatorname{Tr}\nabla^2_x \log p(x|\beta)$ and $\|\nabla_x \log p(x|\beta)\|_2^2$ with a weighting of $\lambda_{\min}$.

Note this loss would be straightforward to train by the change of variables formula (Proposition 3)—and we also note that somewhat related "higher-order" analogues of score matching have appeared in the literature (without analysis or guarantees), for example, (Meng et al., 2021).

Proposition 3 (Integration-by-part Generalized Score Matching Loss for CTLD)**.** *The loss D_{GSM} in the integration by parts form (Lemma 3) as:*

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

where

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

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

The proof of this Lemma is a straightforward calculation, and is included in Appendix C. We remark that the last point of the proposition implies that this loss can be in principle fit by parametrizing the score $\nabla_x \log p_\theta(x|\beta)$ as an explicitly differentiable map (e.g. a neural network).

With this setup in mind, we will prove the following results.

Theorem 2 (Poincaré constant of CTLD)**.** *Under Assumption 1, the Poincaré constant of CTLD C_P enjoys the following upper bound:*

$$C_P \lesssim D^{22} d^2 \lambda_{\max}^9 \lambda_{\min}^{-2}$$

To get a bound on the asymptotic sample complexity of generalized score matching, according to the framework from Lemma 2, we also need to bound the smoothness terms as in Lemma 1. These terms of course depend on the choice of parametrization for the family of distributions we are fitting. To get a quantitative sense for how these terms might scale, we will consider the natural parametrization for a mixture:

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

Remark 6. *Note that in this parametrization, we assume that the weights $\{w_i\}_{i=1}^K$ and shared covariance matrix Σ are known, though the results can be straightforwardly generalized to the natural parametrization in which we are additionally fitting a vector $\{w_i\}_{i=1}^K$ and matrix Σ, at the expense of some calculational complexity.*

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

Theorem 3 (Smoothness under the natural parameterization)**.** *Under Assumptions 1 and 2, the smoothness defined in Theorem 1 enjoys the upper bound*

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

Finally, we show that the generalized score matching loss is asymptotically normal. The proof of this is in Appendix E, and proceeds by verifying the conditions of Lemma 4. Putting this together with the Poincaré inequality bound Theorem 2 and Theorem 1, we get a complete bound on the sample complexity of the generalized score matching loss with $\mathcal{O}$:

Theorem 4 (Main, Polynomial Sample Complexity Bound of CTLD)**.** *Let the data distribution p satisfy Assumption 1. Then, the generalized score matching loss defined in Proposition 3 with parametrization as in Assumption 2 satisfies:*

1. *The set of optima*

$$\begin{aligned}\Theta^* := \{\theta^* = (\mu_1, \mu_2, \dots, \mu_K) |\\ D_{GSM}(p, p_{\theta^*}) = \min_\theta D_{GSM}\left(p, p_\theta\right)\}\end{aligned}$$

satisfies $\theta^* = (\mu_1, \mu_2, \dots, \mu_K) \in \Theta^*$ 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*

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

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:*

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \Gamma_{SM}\right)$$

for a matrix Γ_{SM} satisfying

$$\|\Gamma_{SM}\|_{OP} \leq \mathrm{poly}\left(D, d, \lambda_{\max}, \lambda_{\min}^{-1}\right) \|\Gamma_{MLE}\|_{OP}^2$$

Remark 7. *Note that 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 $\lambda_{\max}$. This result thus applies to very general distributions, intuitively having an arbitrary number of modes that lie in a ball of radius D, and a bound on their "peakiness" and "spread".*

3.1. Bounding the Poincaré constant

In this section, we will sketch the proof of Theorem 2.

Notation: 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 1 and 4, it's easy to see that the generator for CTLD satisfies $\mathcal{E} = \sum_i w_i \mathcal{E}_i$. This verifies condition (1) in Theorem 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). The details of this are provided in Section D.1. To verify condition (3), we will show the projected chain "between" the components (as defined in Theorem 5) mixes fast. The details of this are provided in Section D.2.

3.2. Smoothness under the natural parametrization

To obtain the polynomial upper bound in Theorem 3, 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 10). The polynomial bound requires three ingredients: In Lemma 9, we relate the derivatives of the mixture to derivatives of components by recognizing the higher-order score functions (Janzamin et al., 2014) of the form $\frac{Dp}{p}$ is closely related to the convex perspective map. In Lemma 6, we derive a new result in mixed derivatives of Gaussian components based on Hermite polynomials. In Corollary 1, we handle log derivatives with higher-order versions of the Faá di Bruno formula (Constantine & Savits, 1996), which is a combinatorial formula characterizing higher-order analogues of the chain rule. See Appendix F for details.

Acknowledgements

This work was in part supported by NSF awards IIS-2211907, CCF-2238523, and Amazon Research Award.

References

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

Bakry, D. and Émery, M. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pp. 177–206. Springer, 2006.

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Bakshi, A., Diakonikolas, I., Jia, H., Kane, D. M., Kothari, P. K., and Vempala, S. S. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1234–1247, 2022.

Balakrishnan, S., Wainwright, M. J., and Yu, B. Statistical guarantees for the em algorithm: From population to sample-based analysis. 2017.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

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

Chen, H.-B., Chewi, S., and Niles-Weed, J. Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

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

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

Daskalakis, C., Tzamos, C., and Zampetakis, M. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pp. 704–710. PMLR, 2017.

Diaconis, P. and Stroock, D. Geometric bounds for eigenvalues of markov chains. *The annals of applied probability*, pp. 36–61, 1991.

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

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

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

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

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

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

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

Janzamin, M., Sedghi, H., and Anandkumar, A. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

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

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

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985. PMLR, 2023.

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

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

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

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

Meng, C., Song, Y., Li, W., and Ermon, S. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34: 25359–25369, 2021.

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

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

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

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

Pabbaraju, C., Rohatgi, D., Sevekari, A., Lee, H., Moitra, A., and Risteski, A. Provable benefits of score matching. *arXiv preprint arXiv:2306.01993*, 2023.

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

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

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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

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

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

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

Woodard, D., Schmidler, S., and Huber, M. Sufficient conditions for torpid mixing of parallel and simulated tempering. 2009a.

Woodard, D. B., Schmidler, S. C., and Huber, M. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. 2009b.

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

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

Appendix

Table of Contents

A. Preliminaries

A.1. Generalized Score Matching

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

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

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

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

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

$$\min_{q\in\mathcal{Q}} \mathbb{E}_{X\sim\hat{p}}\left[\operatorname{Tr}\nabla^2 \log q + \frac{1}{2}\|\nabla \log q\|^2\right]$$

for some parametrized family of distributions $\mathcal{Q}$, 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 q can be efficiently calculated.[1]

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

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

$$D_{GSM}\left(p, q\right) = \frac{1}{2}\mathbb{E}_p \left\|\frac{\mathcal{O}p}{p} - \frac{\mathcal{O}q}{q}\right\|_2^2$$

This 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 3 (Integration by parts, (Lyu, 2012))**.** *The GSM loss satisfies*

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

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

A.2. Dirichlet forms and Poincaré inequalities

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

Definition 3 (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 $s, t \geq 0$.

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

Definition 4 (Continuous time Markov processes)**.** *A continuous time 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.

Definition 5. *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 $\mathcal{E}(g) := \mathcal{E}(g,g)$.

Next, we define the Poincaré constant, which captures the mixing time of the process in the χ^2-sense:

Definition 6 (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),*[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 heavily use Langevin diffusion in our paper, for which the Dirichlet form has a particularly simple form:

Definition 7 (Langevin diffusion)**.** *Langevin diffusion is the following stochastic process:*

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$$

where $f : \mathbb{R}^d \to \mathbb{R}$, dB_t is Brownian motion in $\mathbb{R}^d$ with covariance matrix I. Under mild regularity conditions on f, the stationary distribution of this process is $p(x) : \mathbb{R}^N \to \mathbb{R}$, s.t. $p(x) \propto e^{-f(x)}$.

Proposition 4 ((Bakry et al., 2014))**.** *The Dirichlet form corresponding to Langevin has the form $\mathcal{E}(f) = \mathbb{E}_p\|\nabla f\|_2^2$.*

We will analyze mixing times using a decomposition technique similar to the ones employed in (Ge et al., 2018; Moitra & 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.*

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

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*

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

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*

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

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

Then M satisfies the Poincaré inequality

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

A.3. Asymptotic efficiency

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

Lemma 4 ((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:

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

A.4. 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 6 (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.5. Hermite Polynomials

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

Definition 8 (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 5 (Integral representation of Hermite polynomial, (Holmquist, 1996))**.** *The Hermite polynomial H_k defined in Definition 8 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:

$$H_k(-x; \Sigma) = (-1)^k H_k(x; \Sigma) \tag{3}$$

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 F.

Lemma 5. *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 8, 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:

$$\frac{\nabla_x^k \phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (-1)^k (\Sigma^{-1})^{\otimes k} H_k(x-\mu;\Sigma) \tag{4}$$

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 (3), we get:

$$\frac{\nabla_\mu^k \phi(x-\mu;\Sigma)}{\phi(x-\mu;\Sigma)} = (\Sigma^{-1})^{\otimes k} H_k(x-\mu;\Sigma) \tag{5}$$

Combining (4) and (5), 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 5, 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 6. *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 5 and use the convexity of the norm

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

Bounding the right-hand side, we have:

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

Applying Lemma 26 yields the desired result. □

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

Lemma 7. *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 $f:\mathbb{R}^d \to \mathbb{R}$

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

Thus, we have

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

Applying Lemma 6, the result follows.

□

A.6. 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 6 ((Constantine & 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. $|I| \le N$, 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: $(a_1,a_2,\dots,a_d) := a \prec b := (b_1,b_2,\dots,b_d)$ if:

1. $|a| < |b|$
2. $|a| = |b|$ *and* $a_1 < b_1$.
3. $|a| = |b|$ *and* $\exists k >= 1$, *s.t.* $\forall j \le k, a_j = b_j$ *and* $a_{k+1} < b_{k+1}$.

As a straightforward corollary, we have the following:

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

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

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

A.7. 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 8. *(Convexity of perspective, Boyd & 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 9. *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 8. Then, we proceed as follows:

$$\begin{aligned}
\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left\| \frac{(Dp_\theta)(x|\beta)}{p_\theta(x|\beta)} \right\|^k &= \mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p(x|\beta)} \left\| \frac{(Dp_\theta)(x|\beta)}{p_\theta(x|\beta)} \right\|^k \\
&= \mathbb{E}_{\beta\sim r(\beta)} \int g((Dp_\theta)(x|\beta), p_\theta(x|\beta))dx \\
&= \mathbb{E}_{\beta\sim r(\beta)} \int g\left(\sum_{i=1}^K w_i(Dp_\theta)(x|\beta,i), \sum_{i=1}^K w_i p_\theta(x|\beta,i)\right) dx \qquad (6) \\
&\leq \mathbb{E}_{\beta\sim r(\beta)} \int \sum_{i=1}^K w_i g((Dp_\theta)(x|\beta,i), p_\theta(x|\beta,i))dx \qquad (7) \\
&= \mathbb{E}_{\beta\sim r(\beta)} \sum_{i=1}^K w_i \mathbb{E}_{x\sim p(x|\beta,i)} \left\| \frac{(Dp_\theta)(x|\beta,i)}{p_\theta(x|\beta,i)} \right\|^k \\
&\leq \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\| \frac{(Dp_\theta)(x|\beta,i)}{p_\theta(x|\beta,i)} \right\|^k
\end{aligned}$$

where (6) follows by linearity of D, and (7) by convexity of the function g. □

B. A Framework for Analyzing Generalized Score Matching

Proposition 7 (Hessian of GSM loss). *The Hessian of D_{GSM} satisfies*

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

Proof. By a straightforward calculation, we have:

$$\begin{aligned}
\nabla_\theta D_{GSM}(p, p_\theta) &= \mathbb{E}_p \nabla_\theta \left(\frac{\mathcal{O}p_\theta}{p_\theta} \right) \left(\frac{\mathcal{O}p_\theta}{p_\theta} - \frac{\mathcal{O}p}{p} \right) \\
\nabla_\theta^2 D_{GSM}(p, p_\theta) &= \mathbb{E}_p \left[\nabla_\theta \left(\frac{\mathcal{O}p_\theta}{p_\theta} \right)^\top \nabla_\theta \left(\frac{\mathcal{O}p_\theta}{p_\theta} \right) - \sum_{i=1}^m \left(\frac{\mathcal{O}p_\theta}{p_\theta} - \frac{\mathcal{O}p}{p} \right)_i \nabla_\theta^2 \left(\frac{\mathcal{O}p_\theta}{p_\theta} \right)_i \right]
\end{aligned}$$

Since $\frac{\mathcal{O}p_{\theta^*}}{p_{\theta^*}} = \frac{\mathcal{O}p}{p}$, the second term vanishes at $\theta = \theta^*$, which proves the statement.

□

Proof of Lemma 1. We have

$$\begin{aligned} \nabla_\theta l_\theta(x) &= \nabla_\theta \left(\frac{\mathcal{O}p_\theta(x)}{p_\theta(x)} \right) \frac{\mathcal{O}p_\theta(x)}{p_\theta(x)} - \nabla_\theta \mathcal{O}^+ \left(\frac{\mathcal{O}p_\theta(x)}{p_\theta(x)} \right) \\ &= \left((\mathcal{O}\nabla_\theta \log p_\theta) \frac{\mathcal{O}p_\theta}{p_\theta} \right) - (\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta \end{aligned}$$

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

$$\operatorname{cov}(\nabla_\theta l_\theta(x)) \preceq 2\operatorname{cov}\left((\mathcal{O}\nabla_\theta \log p_\theta) \frac{\mathcal{O}p_\theta}{p_\theta} \right) + 2\operatorname{cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta \right)$$

which completes the proof. □

Proof of Lemma 2. To reduce notational clutter, we will drop $_{|\theta=\theta^*}$ since all the functions of θ are evaluated at θ^*. Consider an arbitrary direction w. We have:

$$\begin{aligned} \left\langle w, \nabla_\theta^2 D_{GSM}(p, p_\theta) w \right\rangle &= \left\langle w, \mathbb{E}_p \left[\mathcal{O}\left(\nabla_\theta \log p_\theta\right)^\top \mathcal{O}\left(\nabla_\theta \log p_\theta\right) \right] w \right\rangle \\ &= \mathbb{E}_p \left\| \mathcal{O}\left(\nabla_\theta \log p_\theta\right) w \right\|_2^2 \\ &\overset{\textcircled{1}}{=} \mathbb{E}_p \left\| \sum_i w_i \mathcal{O}\left(\nabla_\theta \log p_\theta\right)_i \right\|_2^2 \\ &\overset{\textcircled{2}}{=} \mathbb{E}_p \left\| \mathcal{O}\left(\sum_{i=1}^{d_\theta} w_i \frac{\partial}{\partial \theta_i} \log p_\theta \right) \right\|_2^2 \\ &= \mathbb{E}_p \| \mathcal{O}\left(\langle w, \nabla_\theta \log p_\theta \rangle \right) \|_2^2 \\ &\overset{\textcircled{3}}{=} -\langle \langle w, \nabla_\theta \log p_\theta \rangle, \mathcal{L}\langle w, \nabla_\theta \log p_\theta \rangle \rangle_p \\ &\overset{\textcircled{4}}{\geq} \frac{1}{C_P} \operatorname{Var}_p(\langle w, \nabla_\theta \log p_\theta \rangle) \\ &\overset{\textcircled{5}}{=} \frac{1}{C_P} w^T \Gamma_{MLE}^{-1} w \end{aligned}$$

where $\textcircled{1}$ follows by commuting $\mathcal{O}$ and ∇_θ, which holds by Lemma 24, $\textcircled{2}$ from linearity of $\mathcal{O}$, and $\textcircled{3}$ since by our assumption for any function g, we have $\mathbb{E}_p \|\mathcal{O}g\|^2 = -\langle g, \mathcal{L}g \rangle_p$ and we apply this condition to $g = \langle w, \nabla_\theta \log p_\theta \rangle$. $\textcircled{4}$ follows from the definition of Poincaré inequality, applied to the function $\langle w, \nabla_\theta \log p_\theta \rangle$, and $\textcircled{5}$ 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.

□

C. Overview of Continuously Tempered Langevin Dynamics

Lemma 10 (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:*

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

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

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

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

Hence, by the chain rule,

$$\nabla_\beta \log p^\beta(x) = \frac{\lambda_{\min}\,\mathrm{Tr}(\nabla_x^2 p^\beta(x))}{p^\beta(x)} \tag{8}$$

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 (8), we have

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

as we needed. □

We also provide the proof of Lemma 3:

Proof of Lemma 3.

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

By Lemma 10, $\nabla_\beta \log p_\theta(x|\beta)$ 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:

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

□

D. Polynomial mixing time bound: proof of Theorem 2

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 11, for all $i \in [K]$ the distribution $p(x, \beta|i)$ has Poincaré constant $C_{x,\beta|i}$ with respect to the Langevin generator that satisfies:

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

3. (Polynomial mixing for projected chain) To bound the Poincaré constant of the projected chain, by Lemma 14 we have

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

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

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

□

D.1. Mixing inside components

In this section, we will show is that we have fast mixing "inside" each of the components of the mixture. Formally, we show:

Lemma 11. *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 D.1.

Proof. The proof will follow by an application of a continuous decomposition result (Theorem D.3 in (Ge et al., 2018), repeated as Theorem 6) , 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 12, $C_\beta \leq \frac{14D^2}{\pi\lambda_{\min}}$.
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 & 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 13.

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

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

which completes the proof. □

Lemma 12 (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 13 (Bound on "rate of change" of the density $p(x|\beta,i)$)**.**

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

Proof.

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

We can apply Lemma 10 to derive explicit expressions for the right-hand side:

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

In $\textcircled{1}$, we use $(a+b)^2 \le 2(a^2+b^2)$ for $a, b \ge 0$; in $\textcircled{2}$ we apply the moment bound for the Chi-Squared distribution of degree-of-freedom d in Lemma 26; and in $\textcircled{3}$ we plug in the bound on $\beta_{\max}$. □

D.2. Mixing between components

In this section, we show the "projected" chain between the components mixes fast:

Lemma 14 (Poincaré constant of projected chain)**.** *Define the projected chain $\bar{M}$ over $[K]$ with transition probability*

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

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

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

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

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 & 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 15 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 & 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 16) — 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 15 (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:

$$\begin{aligned}\chi^2(p(x,\beta|i), p(x,\beta|j)) &= \int \left(\frac{p(x,\beta|i)}{p(x,\beta|j)} - 1\right)^2 p(x,\beta|i) dx d\beta \\ &= \int_0^{\beta_{\max}} \int_{-\infty}^{+\infty} \left(\frac{p(x|\beta,i)r(\beta)}{p(x|\beta,j)r(\beta)} - 1\right)^2 p(x|\beta,i) r(\beta) dx d\beta \\ &= \int_0^{\beta_{\max}} \chi^2(p(x|\beta,i), p(x|\beta,j)) r(\beta) d\beta \\ &\leq \int_0^{\beta_{\max}} \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) r(\beta) d\beta \qquad (9) \\ &= \int_0^{\beta_{\max}} \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) \frac{1}{Z(D,\lambda_{\min})} \exp\left(-\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) d\beta \\ &= \frac{\beta_{\max}}{Z(D,\lambda_{\min})}\end{aligned}$$

where in Line 9, we apply our Lemma 16 to bound the χ^2-divergence between two Gaussians with identical covariance. By a change of variable $\tilde{\beta} := \frac{7D^2}{\lambda_{\min}(1+\beta)}$, $\beta = \frac{7D^2}{\lambda_{\min}\tilde{\beta}} - 1$, $d\beta = -\frac{7D^2}{\lambda_{\min}}\frac{1}{\tilde{\beta}^2}d\tilde{\beta}$, we can rewrite the integral as:

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

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

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

Finally, we get the desired bound

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

□

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

Lemma 16 (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:

$$\begin{aligned}
&\chi^2(p(x|\beta,i),p(x|\beta,j)) \\
&\leq \frac{\det(\Sigma_\beta)^{\frac{1}{2}}}{\det(\Sigma_\beta)} \det\left(\Sigma_\beta^{-1}\right)^{-\frac{1}{2}} \\
&\exp\left(\frac{1}{2}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)^\top (\Sigma_\beta^{-1})^{-1}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)+\frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i-\mu_j^\top\Sigma_\beta^{-1}\mu_j\right) \qquad (10)\\
&= \exp\left(\frac{1}{2}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)^\top (\Sigma_\beta^{-1})^{-1}\left(\Sigma_\beta^{-1}(2\mu_j-\mu_i)\right)+\frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i\right) \\
&\exp\left(-\mu_j^\top\Sigma_\beta^{-1}\mu_j\right) \\
&\leq \exp\left(\frac{1}{2}(2\mu_j-\mu_i)^\top\Sigma_\beta^{-1}(2\mu_j-\mu_i)+\frac{1}{2}\mu_i^\top\Sigma_\beta^{-1}\mu_i\right) \qquad (11)\\
&\leq \exp\left(\frac{\|2\mu_j-\mu_i\|_2^2+\|2\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right) \\
&\leq \exp\left(\frac{(\|2\mu_j\|_2+\|\mu_i\|_2)^2+4\|\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right) \\
&\leq \exp\left(\frac{2\|2\mu_j\|_2^2+2\|\mu_i\|_2^2+4\|\mu_i\|_2^2}{2\lambda_{\min}(1+\beta)}\right) \\
&\leq \exp\left(\frac{7D^2}{\lambda_{\min}(1+\beta)}\right)
\end{aligned}$$

In Equation 10, we apply Lemma G.7 from (Ge et al., 2018) for the chi-square divergence between two Gaussian distributions. In Equation 11, we use the fact that Σ_β^{-1} is PSD.

□

E. 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 7 (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 3 satisfies:*

1. *The set of optima*
$$\Theta^* := \{\theta^* = (\mu_1,\mu_2,\dots,\mu_K)|D_{GSM}(p,p_{\theta^*}) = \min_\theta D_{GSM}\left(p,p_\theta\right)\}$$
satisfies
$$\theta^* = (\mu_1,\mu_2,\dots,\mu_K)\in\Theta^* \textit{ if and only if } \exists\pi:[K]\to[K] \textit{ satisfying } \forall i\in[K], \mu_{\pi(i)}=\mu_i^*, w_{\pi(i)}=w_i\}$$
2. *Let $\theta^*\in\Theta^*$ and let C be any compact set containing θ^*. Denote*
$$C_0 = \{\theta\in C : p_\theta(x)=p(x) \textit{ almost everywhere }\}$$
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:*
$$\sqrt{n}(\hat{\theta}_n-\theta^*)\xrightarrow{d}\mathcal{N}\left(0,\Gamma_{SM}\right)$$
for some matrix Γ_{SM}.

Proof. Part 1 is shown in Lemma 17: the claim roughly follows by classic results on the identifiability of the parameters of a mixture (up to permutations of the components) (Yakowitz & Spragins, 1968).

Part 2 is shown in Lemma 19: it follows from a uniform law of large numbers.

Finally, Part 3 follows from an application of Lemma 4—so we verify the conditions of the lemma are satisfied. The gradient bounds on l_θ are verified Lemma 18—and it largely follows by moment bounds on gradients of the score derived in Section F. Uniform law of large numbers is shown in Lemma 19, 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 17 (Uniqueness of optima)**.** *Suppose for* $\theta := (\mu_1, \mu_2, \dots, \mu_K)$ *there is no permutation* $\pi : [K] \to [K]$*, such that* $\mu_{\pi(i)} = \mu_i^*$ *and* $w_{\pi(i)} = w_i, \forall i \in [K]$*. Then,* $D_{GSM}(\theta) > D_{GSM}(\theta^*)$

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

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

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

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

Lemma 18 (Gradient bounds of l_θ)**.** *Let* $l_\theta(x, \beta)$ *be as defined in Proposition 3. 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 3,

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

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

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

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

$$\begin{aligned} \mathbb{E}\left\|l_\theta^1(x, \beta)\right\|_2^2 &\lesssim \mathbb{E}\|\nabla_x \log p_\theta(x, \beta)\|_2^4 + \mathbb{E}\left(\Delta_x \log p_\theta(x, \beta)\right)^2 \\ &\lesssim \max_{\beta, i} \mathbb{E}_{x \sim p(x|\beta, i)} \|\nabla_x \log p_\theta(x|\beta, i)\|_2^4 + \max_{\beta, i} \mathbb{E}_{x \sim p(x|\beta, i)} \left(\Delta_x \log p_\theta(x|\beta, i)\right)^2 \quad (12) \\ &\leq \mathrm{poly}\left(d, \frac{1}{\lambda_{\min}}\right) \quad (13) \end{aligned}$$

Where (12) follows by Lemma 9, and (13) follows by combining Corollaries 2 and 1.

The same argument, along with Lemma 10, and the fact that $\max_\beta(\nabla_\beta \log r(\beta))^4 \lesssim D^8\lambda_{\min}^{-4}$ by a direct calculation shows that

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

□

Lemma 19 (Uniform convergence)**.** *The generalized score matching loss satisfies a uniform law of large numbers:*

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

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

We will first show that

$$\mathbb{E}_T\text{sup}_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| \leq \frac{C\left(K, d, D, \frac{1}{\lambda_{\min}}\right)}{\sqrt{n}} \tag{14}$$

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

$$\begin{aligned}\mathbb{E}_T\text{sup}_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| &= \mathbb{E}_T\text{sup}_{\theta\in\Theta}\left|\hat{\mathbb{E}}_T l_\theta(x,\beta) - D_{GSM}(\theta)\right| \\ &= \mathbb{E}_T\text{sup}_{\theta\in\Theta}\left|\hat{\mathbb{E}}_T l_\theta(x,\beta) - \mathbb{E}_{T'}\hat{\mathbb{E}}_{T'} l_\theta(x,\beta)\right| \quad (15)\\ &\leq \mathbb{E}_{T,T'}\text{sup}_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\left(l_\theta(x_i,\beta_i) - l_\theta(x_i',\beta_i')\right)\right| \quad (16)\end{aligned}$$

where (15) follows by noting the population expectation can be expressed as the expectation over a choice of a (fresh) training set T', (16) follows by applying Jensen's inequality. Next, consider Rademacher variables $\{\varepsilon_i\}_{i=1}^n$. Since a Rademacher random variable is symmetric about 0, we have

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

For notational convenience, let us denote by

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

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

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

In other words, we have

$$\mathbb{E}_{\{\varepsilon_i\}} \exp\left(\lambda \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left(l_\theta(x_i,\beta_i) - l_{\theta'}(x_i,\beta_i)\right)\right) \leq \exp\left(\lambda^2 d(\theta,\theta')\right) \tag{17}$$

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 (17).

The Dudley entropy integral then gives

$$\text{sup}_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i l_\theta(x_i,\beta_i)\right| \lesssim \int_0^\infty \sqrt{\log N(\epsilon,\Theta,d)}d\epsilon \tag{18}$$

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 (18), we get

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

Taking expectations over the set T (keeping in mind that R is a function of T), by Lemma 18 we get

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

This completes the proof of (14). By Markov's inequality, (14) implies that for every $\epsilon > 0$,

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

Thus, for every $\epsilon > 0$,

$$\lim_{n\to\infty}\text{Pr}_T\left[\text{sup}_{\theta\in\Theta}\left|\widehat{D_{GSM}}(\theta) - D_{GSM}(\theta)\right| > \epsilon\right] = 0$$

Thus,

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

as we need. □

F. Polynomial smoothness bound: proof of Theorem 3

First, we need several easy consequences of the machinery developed in Section A.5, specialized to Gaussians appearing in CTLD.

Lemma 20. *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 26. □

Combining this Lemma with Lemmas 6 and 7, we get the following corollary:

Corollary 2.

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

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

Finally, we will need the following simple technical lemma:

Lemma 21. *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

$$\begin{aligned}
\|\mathrm{Var}(X)\|_{OP} &= \left\| \mathbb{E}\left[(X-\mathbb{E}[X]) (X-\mathbb{E}[X])^\top \right] \right\|_{OP} \\
&\leq \mathbb{E}\left\| X-\mathbb{E}[X] \right\|_2^2 && (19) \\
&\leq 6\mathbb{E}\|X\|_2^2 && (20)
\end{aligned}$$

where (19) follows from the subadditivity of the spectral norm, (20) follows from the fact that

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

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

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 22 and Lemma 23, respectively.

Lemma 22.

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

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

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

where the last step follows by Cauchy-Schwartz. To bound both factors above, we will essentially first use Lemma 9 to relate moments over the mixture, with moments over the components of the mixture. Subsequently, we will use estimates for a single Gaussian, i.e. Corollaries 2 and 1.

Proceeding to the first factor, we have:

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\mathcal{O}\nabla_\theta \log p_\theta(x,\beta)\|_{OP}^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \nabla_\theta \log p_\theta(x,\beta)\|_{OP}^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\beta \nabla_\theta \log p_\theta(x,\beta)\|_2^4 && (21) \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \nabla_\theta \log p_\theta(x|\beta)\|_{OP}^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_\beta \nabla_\theta \log p_\theta(x|\beta)\|_2^4 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\nabla_x \nabla_\theta \log p_\theta(x|\beta,i)\|_{OP}^4 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\nabla_\beta \nabla_\theta \log p_\theta(x|\beta,i)\|_2^4 && (22) \\
&\leq \text{poly}(d, 1/\lambda_{\min}) && (23)
\end{aligned}$$

where (21) follows from the fact that $\mathcal{O}f = (\nabla_x f, \nabla_\beta f)^T$, (22) follows from Lemma 9, and (23) follows by combining Corollaries 2 and 1 and Lemma 10.

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

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\frac{\|\mathcal{O}p_\theta(x,\beta)\|_2}{p_\theta(x,\beta)} \right)^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\frac{\|\nabla_x p_\theta(x,\beta)\|_2}{p_\theta(x,\beta)} \right)^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \left(\frac{\nabla_\beta p_\theta(x,\beta)}{p_\theta(x,\beta)} \right)^4 \\
&= \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \log p_\theta(x,\beta)\|_2^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} (\nabla_\beta \log p_\theta(x,\beta))^4 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)} \|\nabla_x \log p_\theta(x|\beta)\|_2^4 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)} (\nabla_\beta \log p_\theta(x|\beta))^4 + \mathbb{E}_{\beta\sim r(\beta)} (\nabla_\beta \log r(\beta))^4 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \|\nabla_x \log p_\theta(x|\beta,i)\|_2^4 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} (\nabla_\beta \log p_\theta(x|\beta,i))^4 + \max_\beta (\nabla_\beta \log r(\beta))^4 && (24) \\
&\leq \text{poly}(d, D, 1/\lambda_{\min}) && (25)
\end{aligned}$$

where (24) follows from Lemma 9, and (25) follows by combining Corollaries 2 and 1 and Lemma 10, 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 (23) and (25) complete the proof of the lemma. □

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

Lemma 23.

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

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

$$\begin{aligned}
&(\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta(x,\beta) \\
&= \nabla_\theta \Delta_{(x,\beta)} \log p_\theta(x,\beta) && (26)\\
&= \nabla_\theta \Delta_x \log p_\theta(x,\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x,\beta) && (27)\\
&= \nabla_\theta \Delta_x \log p_\theta(x|\beta) + \nabla_\theta \Delta_x \log r(\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta) + \nabla_\theta \nabla_\beta^2 \log r(\beta)\\
&= \nabla_\theta \Delta_x \log p_\theta(x|\beta) + \nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta) && (28)
\end{aligned}$$

where (26) follows by exchanging the order of derivatives, (27) since β is a scalar, so the Laplacian just equals to the Hessian, (28) by dropping the derivatives that are zero in the prior expression.

To bound both summands above, we will essentially first use Lemma 9 to relate moments over the mixture, with moments over the components of the mixture. Subsequently, we will use estimates for a single Gaussian, i.e. Corollaries 1 and 2. Precisely, we have:

$$\begin{aligned}
&\mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|(\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\|_2^2 \\
&\lesssim \mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\nabla_\theta \operatorname{Tr}(\nabla_x^2 \log p_\theta(x|\beta))\|_2^2 + \mathbb{E}_{(x,\beta)\sim p(x,\beta)}\|\nabla_\theta \nabla_\beta^2 \log p_\theta(x|\beta)\|_2^2 \\
&\lesssim \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\|\frac{\nabla_\theta \Delta_x p_\theta(x|\beta,i)}{p_\theta(x|\beta,i)}\right\|_2^2 + \max_{\beta,i} \mathbb{E}_{x\sim p(x|\beta,i)} \left\|\frac{\nabla_\theta \nabla_x p_\theta(x|\beta,i)}{p_\theta(x|\beta,i)}\right\|_{OP}^4 && (29)\\
&\leq \text{poly}(d, 1/\lambda_{\min}) && (30)
\end{aligned}$$

where (29) follows from Lemma 9 and Lemma 10, and (30) follows by combining Corollaries 1 and 2.

□

G. Technical Lemmas

G.1. $\mathcal{O}$ and ∇_θ commute

The proof of Lemma 2 requires that we commute the application of ∇_θ and $\mathcal{O}$. This is obviously the case in the standard score matching ($\mathcal{O} = \nabla_x$) by Clairaut's Theorem (or equality of mixed partials) — even in the case of the operator $\mathcal{O}$ corresponding to CTLD, since it just requires exchanging partial derivatives. This section shows this property holds for much more general $\mathcal{O}$.

Lemma 24. *Let $\mathcal{O} : \mathcal{F} \to \mathcal{R}$, be a continuous linear operator, where $\mathcal{F}$ is a space of univariate functions, and $\mathcal{R}$ a space of (possibly multivariate) functions. Let $\{p_\theta\}_{\theta\in\Theta}$ be a space of parametrized functions by a vector of parameters $\theta \in \mathbb{R}^m$.*

Then $\mathcal{O}$ and ∇_θ commute, that is for every $p_\theta \in \mathcal{F}$, such that $\partial_{\theta_i} p_\theta(x)$ exists for every x and $\partial_{\theta_i} p_\theta \in \mathcal{F}, \forall i \in [m]$, we have:

$$\nabla_\theta \mathcal{O} p_\theta(x) = \tilde{\mathcal{O}} \nabla_\theta p_\theta(x)$$

where $\tilde{\mathcal{O}}$ is defined to be the element-wise extension of $\mathcal{O}$ to vector valued functions (i.e. $\mathcal{O}$ is applied to each coordinate).

Proof. We start with the left-hand side and rewrite it as follows:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j}(\mathcal{O}p)_i(x) &= \lim_{h\to 0} \frac{(\mathcal{O}p)_i(x;\theta + he_j) - (\mathcal{O}p)_i(x;\theta)}{h} \\
&= \lim_{h\to 0} \mathcal{O}\left(\frac{p(\cdot;\theta + he_j) - p(\cdot;\theta)}{h}\right)_i (x) \\
&= \mathcal{O}\left(\lim_{h\to 0}\frac{p(\cdot;\theta + he_j) - p(\cdot;\theta)}{h}\right)_i (x) \qquad (31) \\
&= \mathcal{O}\left(\frac{\partial p}{\partial \theta_j}\right)_i (x)
\end{aligned}
$$

In equation 31, we used the fact that the operator $\mathcal{O}$ is continuous. With this assumption in place, we have

$$
\begin{aligned}
\nabla_\theta \mathcal{O}p &= \begin{bmatrix} \frac{\partial}{\partial\theta_1}(\mathcal{O}p)_1 & \cdots & \frac{\partial}{\partial\theta_{d_\theta}}(\mathcal{O}p)_1 \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\theta_1}(\mathcal{O}p)_{d_{out}} & \cdots & \frac{\partial}{\partial\theta_{d_\theta}}(\mathcal{O}p)_{d_{out}} \end{bmatrix} \\
&= \begin{bmatrix} \mathcal{O}\left(\frac{\partial p}{\partial\theta_1}\right)_1 & \cdots & \mathcal{O}\left(\frac{\partial p}{\partial\theta_{d_\theta}}\right)_1 \\ \vdots & \ddots & \vdots \\ \mathcal{O}\left(\frac{\partial p}{\partial\theta_1}\right)_{d_{out}} & \cdots & \mathcal{O}\left(\frac{\partial p}{\partial\theta_{d_\theta}}\right)_{d_{out}} \end{bmatrix} \\
&= \begin{bmatrix} \mathcal{O}\left(\frac{\partial p}{\partial\theta_1}\right) & \cdots & \mathcal{O}\left(\frac{\partial p}{\partial\theta_{d_\theta}}\right) \end{bmatrix} \\
&= \tilde{\mathcal{O}}\nabla_\theta p
\end{aligned}
$$

This completes the proof. □

By way of remarks, note that the operator $\mathcal{O} = \nabla_x$ (which corresponds to standard score matching) is bounded when viewed as an operator $\nabla_x : H^1(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$. Namely, for a function $f \in H^1(\mathbb{R}^d)$, we have

$$\|f\|_{H^1(\mathbb{R}^d)} = \|f\|_{L^2(\mathbb{R}^d)} + \|\nabla f\|_{L^2(\mathbb{R}^d)}$$

Thus, trivially,

$$\|\nabla_f\|_{L^2(\mathbb{R}^d)} \leq \|f\|_{H^1(\mathbb{R}^d)}$$

i.e. the operator H^1 is bounded.

G.2. 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 $\Sigma \preceq \sigma^2_{\max} I$.

Lemma 25 (Norm of Gaussian)**.** *The random variable z enjoys the bound*

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

Proof.

$$
\begin{aligned}
(\mathbb{E}\|z\|_2)^2 &\leq \mathbb{E}\|z\|_2^2 \qquad (32) \\
&= \mathbb{E}\sum_{i=1}^d z_i^2 \\
&= d \qquad (33)
\end{aligned}
$$

where (32) follows from Jensen, and (33) by plugging in the mean of a chi-squared distribution with d degree of freedom. □

Lemma 26 (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.

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

□

H. Related work

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.

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

On the empirical side, breakthrough work by (Song & 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 proposed this as a mechanism to alleviate the poor estimate of the score inbetween modes for multimodal distributions, as well in the presence of low-dimensional manifold structure in the data. 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 the approach in (Song & 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 & Parisi, 1992; Neal, 1996; Earl & 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 & Parisi, 1992) with temperature annealing (i.e. scaling the log-pdf of the distribution), along with Metropolis-Hastings to swap the temperature in the chain mixes in time poly(K). In subsequent work (Moitra & Risteski, 2020), the authors show that for distributions sufficiently concentrated near a manifold of positive Ricci curvature, Langevin mixes fast.

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 & Risteski, 2020; Madras & 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 & 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 & Spragins, 1968). Early work in the theoretical computer science community provided guarantees for clustering-based algorithms (Dasgupta, 1999; Sanjeev & Kannan, 2001); subsequent work provided polynomial-time algorithms down to the information theoretic threshold for identifiability based on the method of moments (Moitra & Valiant, 2010; Belkin & Sinha, 2010); even more recent work tackles robust algorithms for learning mixtures in the presence of outliers (Hopkins & 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).

Speeding up mixing via tempering techniques 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 & Parisi, 1992). In parallel tempering (Swendsen & Wang, 1986; Hukushima & 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 & 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 3. 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.