Learning Mixtures of Gaussians Using the DDPM Objective

Kulin Shah
UT Austin
kulinshah@utexas.edu

Sitan Chen
Harvard University
sitan@seas.harvard.edu

Adam Klivans
UT Austin
klivans@cs.utexas.edu

Abstract

Recent works have shown that diffusion models can learn essentially any distribution provided one can perform score estimation. Yet it remains poorly understood under what settings score estimation is possible, let alone when practical gradient-based algorithms for this task can provably succeed.

In this work, we give the first provably efficient results along these lines for one of the most fundamental distribution families, Gaussian mixture models. We prove that gradient descent on the denoising diffusion probabilistic model (DDPM) objective can efficiently recover the ground truth parameters of the mixture model in the following two settings:

1. We show gradient descent with random initialization learns mixtures of two spherical Gaussians in d dimensions with $1/\mathrm{poly}(d)$-separated centers.
2. We show gradient descent with a warm start learns mixtures of K spherical Gaussians with $\Omega(\sqrt{\log(\min(K,d))})$-separated centers.

A key ingredient in our proofs is a new connection between score-based methods and two other approaches to distribution learning, the expectation-maximization (EM) algorithm and spectral methods.

1 Introduction

In recent years diffusion models [SSDK+20, SDWMG15, SE19] have emerged as a powerful framework for generative modeling and now form the backbone of notable image generation systems like DALL·E 2 [RDN+22], Imagen [SCS+22], and Stable Diffusion [RBL+22]. At the heart of this framework is a reduction from *distribution learning* to *denoising* or *score estimation*. That is, in order to generate new samples from a data distribution q given a collection of independent samples, it suffices to learn the score function, i.e., the gradient of the log-density of the data distribution when convolved with varying levels of noise (see Section 1.3). A popular and well-studied objective for score matching is the *denoising diffusion probabilistic model (DDPM) objective* due to [HJA20]. Optimizing this objective amounts to solving the following type of problem: given a noisy observation $\widetilde{x}$ of a sample x from q, estimate the mean of the posterior distribution over x.

While a number of theoretical works [DBTHD21, BMR22, CLL22, DB22, LLT22, LWYL22, Pid22, WY22, CCL+23b, CDD23, LLT23, CCL+23a, LWCC23, BDD23] have established rigorous convergence guarantees for diffusion models under mild assumptions on the data distribution, these works assume the existence of an oracle for score estimation and leave open whether one can actually provably implement such an oracle for interesting families of data distributions. In practice, the algorithm of choice for score estimation is simply to train a student network via gradient descent (GD) to fit a set of examples $(x, \widetilde{x})$. We thus ask:

Are there natural data distributions under which GD provably achieves accurate score estimation?

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

In this work, we consider the setting where q is given by a *mixture of Gaussians*. Concretely, we assume that there exist centers $\mu_1^*, \ldots, \mu_K^* \in \mathbb{R}^d$ such that

$$q = \frac{1}{K} \sum_{i=1}^{K} \mathcal{N}(\mu_i^*, \mathrm{Id}) \,.$$

We answer the above question in the affirmative for this class of distributions:

Theorem 1 (Informal, see Theorems 7 and 13)**.** *Gradient descent on the DDPM objective with random initialization efficiently learns the parameters of an unknown mixture of two spherical Gaussians with $1/\mathrm{poly}(d)$-separated centers.*

Theorem 2 (Informal, see Theorem 16)**.** *When there is a warm start of the centers, gradient descent on the DDPM objective efficiently learns the parameters an unknown mixture of K spherical Gaussians with $\Omega(\sqrt{\log(\min(K, d))})$-separated centers.*

The DDPM objective is described in Algorithm 1. The term "efficiently" above means that both the running time and sample complexity of our algorithm is polynomial in the dimension d, the inverse accuracy $1/\varepsilon$, and the number of components K. In the informal discussion, we often work with population gradients for simplicity, but in our proofs we show that empirical estimates of the gradient suffice (full details can be found in the Appendix).

Algorithm 1: GMMDENOISER$(t, \{\mu_i^{(0)}\}_{i=1}^K, H)$

Input: Noise scale t, initialization $\{\mu_i^{(0)}\}_{i=1}^K$, number of gradient descent steps H

1 Initialize the parameters for the score estimate at $\theta_t^{(0)} = \{\mu_{i,t}^{(0)}\}_{i=1}^K$ (see Eq. (9) for how the estimate s_θ depends on the parameters θ, and Eq. (8) for the definition of $\mu_{i,t}^{(0)}$)

2 Run gradient descent on the DDPM objective $L_t(s_{\theta_t})$ for H steps where

$$L_t(s_{\theta_t}) = \mathbb{E}\Big[\Big\| s_{\theta_t}(X_t) + \frac{Z_t}{\sqrt{1 - \exp(-2t)}} \Big\|^2\Big] \,,$$

3 **return** $\theta_t^{(H)} = \{\mu_{i,t}^{(H)}\}_{i=1}^K$ where $\theta_t^{(H)}$ denotes the parameters after H steps of GD.

We refer to Section 1.3 for a formal description of the quantities used in Algorithm 1. Note that there are by now a host of different algorithms for provably learning mixtures of Gaussians (see Section 1.1). For instance, it is already known that expectation-maximization (EM) achieves the quantitative guarantees of Theorems 1 and 2 [DTZ17, XHM16, KC20, SN21], and in fact even stronger guarantees are known via the method of moments. Unlike works based on the method of moments however, our algorithm is practical. And unlike works based on EM, it is based on an approach which is empirically successful for a wide range of realistic data distributions. Furthermore, as we discuss in Section 1.2, the analysis of Algorithm 1 leverages an intriguing and, to our knowledge, novel connection from score estimation to EM, as well as to another notable approach for learning mixture models, namely spectral methods. Roughly speaking, at large noise levels, the gradient updates in Algorithm 1 are essentially performing a type of power iteration, while at small noise levels, the gradient updates are performing the "M" step in the EM algorithm.

1.1 Related work

Theory for diffusion models. A number of works have given convergence guarantees for DDPMs and variants [DBTHD21, BMR22, CLL22, DB22, LLT22, LWYL22, Pid22, WY22, CCL+23b, CDD23, LLT23, LWCC23, BDD23, CCL+23a]. These results show that, given an oracle for accurate score estimation, diffusion models can learn essentially any distribution over $\mathbb{R}^d$ (e.g. [CCL+23b, LLT23, CLL22] show this for arbitrary compactly supported distributions). Additionally, two recent works [EAMS22, MW23] have used Eldan's stochastic localization [Eld13, Eld20], which is a reparametrization in time and space of the reverse SDE for DDPMs, to give sampling algorithms for certain distributions arising in statistical physics. As we discuss next, these works are end-to-end in that they also give provable algorithms for score estimation via approximate message passing, though the statistical task they address is not distribution learning.

Provable score estimation. There is a rich literature giving Bayes-optimal algorithms for various natural denoising problems via methods inspired by statistical physics, like approximate message passing (AMP) (e.g. [MV21, CFM21, BM11, Kab03, DMM09, DMM10]) and natural gradient descent (NGD) on the TAP free energy [CFM21, EAMS22, Cel22]. The abovementioned works [EAMS22, MW23] (see also [Cel22]) build on these techniques to give algorithms for the denoising problems that arise in their implementation of stochastic localization. These works on denoising via AMP or NGD are themselves part of a broader literature on variational inference, a suitable literature review would be beyond the scope of this work, see e.g. [BKM17, WJ+08, MM09].

We are not aware of any provable algorithms for score estimation explicitly in the context of *distribution learning*. That said, it may be possible to extract a distribution learning result from [EAMS22]. While their algorithm was for sampling from the Sherrington-Kirkpatrick (SK) model given the Hamiltonian rather than training examples as input, if one is instead given training examples drawn from the SK measure, then at sufficiently high temperature one can approximately recover the Hamiltonian [AG22]. In this case, a suitable modification [EAMS22] should be able to yield an algorithm for approximately generating fresh samples from the SK model given training examples.

Learning mixtures of Gaussians. The literature on provable algorithms for learning Gaussian mixture models is vast, dating back to the pioneering work of Pearson [Pea94], and we cannot do justice to it here. We mention only works whose quantitative guarantees are closest in spirit to ours and refer to the introduction of [LL22] for a comprehensive overview of recent works in this direction. For mixtures of identity-covariance Gaussians in high dimensions, the strongest existing guarantee is a polynomial-time algorithm [LL22] for learning the centers as long as their pairwise separation slightly exceeds $\Omega(\sqrt{\log K})$ based on a sophisticated instantiation of method of moments inspired by the quasipolynomial-time algorithms of [DKS18, HL18, KSS18]. By the lower bound in [RV17], this is essentially optimal. In contrast, our Theorem 2 only applies given one initializes in a neighborhood of the true parameters of the mixture. We also note the exponential-time spectral algorithm of [SOAJ14] and quasipolynomial-time tensor-based algorithm of [DK20], which achieve *density estimation* even in the regime where the centers are arbitrarily closely spaced and learning the centers is information-theoretically impossible.

A separate line of work has investigated the "textbook" algorithm for learning Gaussian mixtures, namely the EM algorithm [BWY17, DS07, DTZ17, XHM16, YYS17, ZLS20, KC20, SN21]. Notably, for balanced mixtures of two Gaussians with the same covariance, [DTZ17] showed that finite-sample EM with random initialization converges exponentially quickly to the true centers. For mixtures of K Gaussians with identity covariance, [KC20, SN21] showed that from an initialization sufficiently close to the true centers, finite-sample EM converges exponentially quickly to the true centers as long as their pairwise separation is $\Omega(\sqrt{\log K})$. In particular, [SN21] establish this local convergence as long as every center estimate is initialized at distance at most $\Delta/2$ away from the corresponding true center, where Δ is the minimum separation between any pair of true centers; this radius of convergence is provably best possible for EM.

Lastly, we note that there are many works giving parameter recovery algorithms mixtures of Gaussians with general mixing weights and covariances, all of which are based on method of moments [KMV10, HP15, Kan21, BS15, MV10, LM23, BDJ+22, DHKK20]. Unfortunately, for general mixtures of K Gaussians, these algorithms run in time at least $d^{O(K)}$, and there is strong evidence [DKS17, BRST21] that this is unavoidable for computationally efficient algorithms.

1.2 Technical overview

We begin by describing in greater detail the algorithm we analyze in this work. For the sake of intuition, in this overview we will focus on the case of mixtures of two Gaussians ($K = 2$) where the centers are well-separated and symmetric about the origin, that is, the data distribution is given by

$$q = \frac{1}{2}\mathcal{N}(\mu^*, \mathrm{Id}) + \frac{1}{2}\mathcal{N}(-\mu^*, \mathrm{Id})\,. \tag{1}$$

At the end of the overview, we briefly discuss the key challenges for handling smaller separation and general K.

Loss function, architecture of the score function and student network. The algorithmic task at the heart of score estimation is that of denoising. Formally, for some noise level $t > 0$, we are given

a noisy sample

$$X_t = \exp(-t)X_0 + \sqrt{1-\exp(-2t)}Z_t\,,$$

where X_0 is a clean sample drawn from the data distribution q, and $Z_t \sim \mathcal{N}(0, \mathrm{Id})$. Conditioning on X_t induces some posterior distribution over the noise Z_t, and our goal is to form an estimate s for the mean of this posterior which achieves small error on average over the randomness of X_0 and Z_t. That is, we would like to minimize the *DDPM objective*, which up to rescaling is given by[1]

$$L_t(s) = \mathbb{E}_{X_0, Z_t}\|s(X_t) - Z_t\|^2\,.$$

As discussed in the introduction, the algorithm of choice for minimizing this objective in practice is gradient descent on some student network. To motivate our choice of architecture, note that when the data distribution is given by (1), the true minimizer of L_t is, up to scaling,

$$\tanh(\langle \mu_t^*, x\rangle)\mu_t^* - x\,, \quad \text{where } \mu_t^* \triangleq \mu^*\exp(-t)\,. \tag{2}$$

See Appendix A for the derivation. Notably, Eq. (2) is exactly a two-layer neural network with $\tanh$ activation. As a result, we use the same architecture for our student network when running gradient descent. That is, given weights $\mu \in \mathbb{R}^d$, our student network is given by $s_\mu(x) \triangleq \tanh(\mu^\top x)\mu - x$. The exact gradient updates on μ are given in Lemma C.2.

As we discuss next, depending on whether the noise level t is large or small, this update closely approximates the update in one of two well-studied algorithms for learning mixtures of Gaussians: power method and EM respectively.

Learning mixtures of two Gaussians. We first provide a brief overview of the analysis and then go into the details of the analysis. We start with mixtures of two Gaussians of the form (1) where $\|\mu^*\|$ is $\Omega(1)$. In this case, we analyze the following two-stage algorithm. We first use gradient descent on the DDPM objective with large t starting from random initialization. We show that gradient descent in this "high noise" regime resembles a type of power iteration and gives μ that has a nontrivial correlation with μ_t^*. Starting from this μ, we then run gradient descent with small t. We show that the gradient descent in this "small noise" regime corresponds to the EM algorithm and converges exponentially quickly to the ground truth.

Large noise level: connection to power iteration. When t is large, we show that gradient descent on the DDPM objective is closely approximated by power iteration. More precisely, in this regime, the negative gradient of $L_t(s_\mu)$ is well-approximated by

$$-\nabla_\mu L_t(s_\mu) \approx (2\mu_t^*\mu_t^{*\top} - r\mathrm{Id})\,\mu\,,$$

where r is a scalar that depends on μ (See Lemma 8). So the result of a single gradient update with step size η starting from μ is given by

$$\mu' \triangleq \mu - \eta\nabla_\mu L_t(s_\mu) \approx ((1-\eta r)\,\mathrm{Id} + 2\eta\mu_t^*\mu_t^{*\top})\mu\,. \tag{3}$$

This shows us that each gradient step can be approximated by one step of power iteration (without normalization) on the matrix $(1-\eta r)\,\mathrm{Id} + 2\eta\mu_t^*\mu_t^{*\top}$. It is know that running enough iterations of the latter from a random initialization will converge in angular distance to the top eigenvector, which in this case is given by μ_t^*. This suggests that if we can keep the approximation error in (3) under control, then gradient descent on μ will also allow us to converge to a neighborhood of the ground truth. We implement this strategy in Lemma 10. Next, we argue that once we are in a neighborhood of the ground truth, we can run GD on the DDPM objective at *low noise level* to refine our estimate.

Low noise level: connection to the EM algorithm. When t is small, we show that gradient descent on the DDPM objective is closely approximated by EM. Here, our analysis uses the fact that μ^* is sufficiently large and requires that we initialize μ to have sufficiently large correlation with the true direction μ_t^*. We can achieve the latter using the large-t analysis in the previous section.

Provided we have this, when t is small it turns out that the negative gradient is well-approximated by

$$-\nabla_\mu L_t(s_\mu) \approx \mathbb{E}_{X\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\langle\mu, X\rangle)X] - \mu\,.$$

[1]The real DDPM objective is slightly different, see (5). The latter is what we actually consider in this paper, but this distinction is unimportant for the intuition in this overview.

Note that the expectation is precisely the "M"-step in the EM algorithm for learning mixtures of two Gaussians (see e.g. Eq. (2.2) of [DTZ17]). We conclude that a single gradient update with step size η starting from μ is given by mixing the old weights μ with the result of the "M"-step in EM:

$$\mu' \triangleq \mu - \eta \nabla_\mu L_t(s_\mu) \approx (1-\eta)\mu + \eta \underbrace{\mathbb{E}_{X \sim \mathcal{N}(\mu_t^*, \mathrm{Id})}[\tanh(\langle \mu, X \rangle) X]}_{\text{"M" step in the EM algorithm}} .$$

[XHM16] and [DTZ17] showed that EM converges exponentially quickly to the ground truth μ_t^* from a warm start, and we leverage ingredients from their analysis to prove the same guarantee for gradient descent on the DDPM objective at small noise level t (see Lemma 12).

Extending to small separation. Next, suppose we instead only assume that $\|\mu^*\|$ is $\Omega(1/\mathrm{poly}(d))$, i.e. the two components in the mixture may have small separation. The above analysis breaks down for the following reason: while it is always possible to show that gradient descent at large noise level converges in *angular distance* to the ground truth, if $\|\mu^*\|$ is small, then we cannot translate this to convergence in Euclidean distance.

We circumvent this as follows. Extending the connection between gradient descent at large t and power iteration, we show that a similar analysis where we instead run *projected* gradient descent over the ball of radius $\|\mu^*\|$ yields a solution arbitrarily close to the ground truth, even without the EM step.[2] The projection step can be thought of as mimicking the normalization step in power iteration.

It might appear to the reader that this projected gradient-based approach is strictly superior to the two-stage algorithm described at the outset. However, in addition to obviating the need for a projection step when separation is large, our analysis for the two-stage algorithm has the advantage of giving much more favorable statistical rates. Indeed, we can show that the sample complexity of the two-stage algorithm has optimal dependence on the target error ($1/\varepsilon^2$), whereas we can only show a suboptimal dependence ($1/\varepsilon^8$) for the single-stage algorithm.

Extending to general K. The connection between gradient descent on the DDPM objective at small t and the EM algorithm is sufficiently robust that for general K, our analysis for $K = 2$ can generalize once we replace the ingredients from [XHM16] and [DTZ17] with the analogous ingredients in existing analyses for EM with K Gaussians. For the latter, it is known that if the centers of the Gaussians have separation $\Omega(\sqrt{\log \min(K, d)})$, then EM will converge from a warm start [KC20, SN21]. By carefully tracking the error in approximating the negative gradient with the "M"-step in EM, we are able to show that gradient descent on the DDPM objective at small t achieves the same guarantee.

1.3 Preliminaries

Diffusion models. Throughout the paper, we use either q or q_0 to denote the data distribution and X or X_0 to denote the corresponding random variable on $\mathbb{R}^d$. The two main components in diffusion models are the *forward process* and the *reverse process*. The forward process transforms samples from the data distribution into noise, for instance via the *Ornstein-Uhlenbeck (OU) process*:

$$\mathrm{d}X_t = -X_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t \quad \text{with} \quad X_0 \sim q_0 \,,$$

where $(W_t)_{t \geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. We use q_t to denote the law of the OU process at time t. Note that for $X_t \sim q_t$,

$$X_t = \exp(-t) X_0 + \sqrt{1 - \exp(-2t)} Z_t \quad \text{with} \quad X_0 \sim q_0, \quad Z_t \sim \mathcal{N}(0, \mathrm{Id}) \,.$$

The reverse process then transforms noise into samples, thus performing generative modeling. Ideally, this could be achieved by running the following stochastic differential equation for some choice of terminal time T:

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla_x \ln q_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t \quad \text{with} \quad X_0^{\leftarrow} \sim q_T \,,$$

where now W_t is the reversed Brownian motion. In this reverse process, the iterate $X_t^{\leftarrow}$ is distributed acccording to q_{T-t} for every $t \in [0, T]$, so that the final iterate $X_T^{\leftarrow}$ is distributed according to the

[2] Note that although μ^* is unknown, we can estimate its norm from samples.

data distribution q_0. The function $\nabla_x \ln q_t$ is called the *score function*, and because it depends on q which is unknown, in practice one estimates it by minimizing the *score matching loss*

$$\min_{s_t} \ \mathbb{E}_{X_t \sim q_t}[\|s_t(X_t) - \nabla_x \ln q_t(X_t)\|^2] \,. \tag{4}$$

A standard calculation (see e.g. Appendix A of [CCL+23b]) shows that this is equivalent to minimizing the *DDPM objective* in which one wants to predict the noise Z_t from the noisy observation X_t, i.e.

$$\min_{s_t} \ L_t(s_t) = \mathbb{E}_{X_0, Z_t}\left[\left\| s_t(X_t) + \frac{Z_t}{\sqrt{1-\exp(-2t)}} \right\|^2\right] . \tag{5}$$

While we have provided background on diffusion models for context, in this work we focus specifically on the optimization problem (5).

Mixtures of Gaussians. We consider the case of learning mixtures of K equally weighted Gaussians:

$$q = q_0 = \frac{1}{K}\sum_{i=1}^{K} \mathcal{N}(\mu_i^*, \mathrm{Id}), \tag{6}$$

where μ_i^* denotes the mean of the i^{th} Gaussian component. We define $\theta^* = \{\mu_1^*, \mu_2^* \ldots, \mu_K^*\}$. For the mixtures of two Gaussians, we can simplify the data distribution as

$$q = q_0 = \frac{1}{2}\mathcal{N}(\mu^*, \mathrm{Id}) + \frac{1}{2}\mathcal{N}(-\mu^*, \mathrm{Id}). \tag{7}$$

Note that distribution in Eq. (7) is equivalent to the distribution Eq. (6) with $K = 2$ because shifting the latter by its mean will give the former distribution, and furthermore the necessary shift can be estimated from samples. The following is immediate:

Lemma 3. *If q_0 is a mixture of K Gaussians as in Eq. (6), then for any $t > 0$, q_t is the mixture of K Gaussians given by*

$$q_t = \frac{1}{K}\sum_{i=1}^{K} \mathcal{N}(\mu_{i,t}^*, \mathrm{Id}) \quad \textit{where} \ \ \mu_{i,t}^* \triangleq \mu_i^* \exp(-t) \,. \tag{8}$$

See Appendix A for a proof of this fact. We can see that the means of q_t get rescaled according to the noise level t. We also define $\theta_t^* = \{\mu_{1,t}^*, \mu_{2,t}^*, \ldots, \mu_{K,t}^*\}$.

Lemma 4. *The score function for distribution q_t, for any $t > 0$, is given by*

$$\nabla_x \ln q_t(x) = \sum_{i=1}^{K} w_{i,t}^*(x)\mu_{i,t}^* - x \,, \qquad \textit{where} \qquad w_{i,t}^*(x) = \frac{\exp(-\|x - \mu_{i,t}^*\|^2/2)}{\sum_{j=1}^{K}\exp(-\|x - \mu_{j,t}^*\|^2/2)}.$$

For a mixture of two Gaussians, the score function simplifies to

$$\nabla_x \log q_t(x) = \tanh(\mu_t^{*\top} x)\mu_t^* - x \,, \qquad \textit{where} \qquad \mu_t^* \triangleq \mu^* \exp(-t)$$

See Appendix A for the calculation.

Recall that $\nabla_x \log q_t(x)$ is the minimizer for the score-matching objective given in Eq. (4). Therefore, we parametrize our student network architecture similarly to the optimal score function. Our student architecture for mixtures of K Gaussians is

$$s_{\theta_t}(x) = \sum_{i=1}^{K} w_{i,t}(x)\mu_{i,t} - x \,, \qquad \text{where} \qquad w_{i,t}(x) \triangleq \frac{\exp(-\|x - \mu_{i,t}\|^2/2)}{\sum_{j=1}^{K}\exp(-\|x - \mu_{j,t}\|^2/2)} \tag{9}$$

$$\mu_{i,t} \triangleq \mu_i \exp(-t).$$

where $\theta_t = \{\mu_{1,t}, \mu_{2,t}, \ldots, \mu_{K,t}\}$ denotes the set of parameters at the noise scale t. For mixtures of two Gaussians, we simplify the student architecture as follows:

$$s_{\theta_t}(x) = \tanh(\mu_t^\top x)\mu_t - x \,, \quad \text{where} \quad \mu_t \triangleq \mu \exp(-t).$$

As θ_t only depends on μ_t in the case of mixtures of two Gaussians, we simplify the notation of the score function from $s_{\theta_t}(x)$ to $s_{\mu_t}(x)$ in that case. We use $\hat{\mu}_t$ and $\hat{\mu}_t^*$ to denote the unit vector along the direction of μ_t and μ_t^* respectively. Note that we often use μ_t (or θ_t) to denote the current iterate of gradient descent on the DDPM objective and μ_t' to denote the iterate after taking a gradient descent step from μ_t.

Expectation-Maximization (EM) algorithm. The EM algorithm is composed of two steps: the E-step and the M-step. For mixtures of Gaussians, the E-step computes the expected log-likelihood based on the current mean parameters and the M-step maximizes this expectation to find a new estimate of the parameters.

Fact 5 (See e.g., [DTZ17, YYS17, KC20] for more details)**.** *When X is the mixture of K Gaussian and $\{\mu_1, \mu_2, \ldots, \mu_K\}$ are current estimates of the means, the population EM update for all $i \in \{1, 2, \ldots, K\}$ is given by*

$$\mu_i' = \frac{\mathbb{E}_X[w_i(X)X]}{\mathbb{E}_X[w_i(X)]}, \quad \text{where} \quad w_i(X) = \frac{\exp(-\|X - \mu_i\|^2/2)}{\sum_{j=1}^K \exp(-\|X - \mu_j\|^2/2)}.$$

The EM update for mixtures of two Gaussians given in Eq. (7) simplifies to

$$\mu' = \mathbb{E}_{X \sim \mathcal{N}(\mu^*, \mathrm{Id})}[\tanh(\mu^\top X)X].$$

An analogous version of the EM algorithm, called the gradient EM algorithm, takes a gradient step in the direction of the M-step instead of optimizing the objective in the M-step fully.

Fact 6 (See e.g., [YYS17, SN21] for more details)**.** *For all $i \in \{1, 2, \ldots, K\}$, the gradient EM-update for mixtures of K Gaussian is given by*

$$\mu_i' = \mu_i + \eta\, \mathbb{E}_X[w_i(X)(X - \mu_i)],$$

where η is the learning rate.

2 Warmup: mixtures of two Gaussians with constant separation

In this section, we formally state our result for learning mixtures of two Gaussians with constant separation. This case highlights the main proof techniques, namely viewing gradient descent on the DDPM objective as power iteration and as the EM algorithm.

2.1 Result and algorithm

Theorem 7. *There is an absolute constant $c > 0$ such that the following holds. Suppose a mixture of two Gaussians with the mean parameter μ^* satisfies $\|\mu^*\| > c$. Then, for any $\varepsilon > 0$, there is a procedure that calls Algorithm 1 at two different noise scales t and outputs $\tilde{\mu}$ such that $\|\tilde{\mu} - \mu^*\| \le \varepsilon$ with high probability. Moreover, the algorithm has time and sample complexity $\mathrm{poly}(d)/\varepsilon^2$ (see Theorem C.1 for more precise quantitative bounds).*

Algorithm. The algorithm has two stages. In the first stage we run gradient descent on the DDPM objective described in Algorithm 1 from a random Gaussian initialization and noise scale t_1 for a fixed number of iterations H where $t_1 = \Theta(\log d)$ ("high noise") and $H = \mathrm{poly}(d, 1/\varepsilon)$. In the second stage, the procedure uses the output of the first step as initialization and runs Algorithm 1 at a "low noise" scale of $t_2 = \Theta(1)$.

2.2 Proof outline of Theorem 7

We provide a proof sketch of correctness of the above algorithm and summarize the main technical lemmas here. All proofs of the following lemmas can be found in Appendix C.

Part I: Analysis of high noise regime and connection to power iteration. We show that in the large noise regime, the negative gradient $-\nabla L_t(s_t)$ is well-approximated by $2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t$. Recall that this result is the key to showing the resemblance between gradient descent and power iteration. Concretely, we show the following lemma:

Lemma 8 (See Lemma C.3 for more details)**.** *For $t = \Theta(\log d)$, the gradient descent update on the DDPM objective $L_t(s_t)$ can be approximated with $2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t$:*

$$\left\| \left(-\nabla L_t(s_t)\right) - \left(2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t\right) \right\| \le \mathrm{poly}(1/d).$$

From Lemma 8, it immediately follows that $\mu' t$, the result of taking a single gradient step starting from μ_t, is well-approximated by the result of taking a single step of power iteration for a matrix whose leading eigenvector is μ_t^*:

$$\mu_t' = \mu_t - \eta \nabla L_t(s_\mu) \approx (\mathrm{Id}(1 - 3\eta\|\mu_t\|^2) + 2\mu_t^* \mu_t^{*\top})\mu_t \,.$$

The second key element is to show that as a consequence of the above power iteration update, the gradient descent converges in *angular distance* to the leading eigenvector. Concretely, we show the following lemma:

Lemma 9 (Informal, see Lemma C.5 for more details)**.** *Suppose μ_t' is the iterate after one step of gradient descent on the DDPM objective from μ_t. Denote the angle between μ_t and μ_t^* to be θ and between μ_t' and μ_t^* to be θ'. In this case, we show that*

$$\tan\theta' = \max\left(\kappa_1 \tan\theta, \kappa_2\right),$$

where $\kappa_1 < 1$ and $\kappa_2 \leq 1/\mathrm{poly}(d)$.

Note $\tan\theta' < \tan\theta$ implies that $\theta' < \theta$ or equivalently $\langle \hat{\mu}_t', \hat{\mu}_t^* \rangle > \langle \hat{\mu}_t, \hat{\mu}_t^* \rangle$. Thus, the above lemma shows that by taking a gradient step in the DDPM objective, the angle between μ_t and μ_t^* decreases. By iterating this, we obtain the following lemma:

Lemma 10 (Informal, see Lemma C.6 for more details)**.** *Running gradient descent from a random initialization on the DDPM objective $L_t(s_\mu)$ for $t = O(\log d)$ gives μ_t for which $\langle \hat{\mu}_t, \hat{\mu}_t^* \rangle$ is $\Omega(1)$.*

Note that we cannot keep running gradient descent at this high noise scale and hope to achieve μ such that $\|\mu - \mu^*\|$ is $O(\varepsilon)$. This is because Lemma 9 can only guarantee that the angle between μ_t and μ_t^* is $O(\varepsilon)$, but this does not imply $\|\mu - \mu^*\|$ is $O(\varepsilon)$. Instead, as described in Part II, we will proceed with a smaller noise scale.

Part II: Analysis of low noise regime and connection to EM. In the low noise regime, we run Algorithm 1 using the output from Part I as our initialization. Our analysis here shows that whenever the initialization μ_t satisfies the condition of $\langle \hat{\mu}_t, \hat{\mu}_t^* \rangle$ being $\Omega(1)$, $\|\mu_t - \mu_t^*\|$ contracts after every gradient step. To start with, we show that the result of a *population* gradient step on the DDPM objective $L_t(s_\mu)$ results in the following:

$$\mu_t' = (1-\eta)\mu_t + \eta\, \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^\top x)x] + \eta G(\mu_t, \mu_t^*),$$

where μ_t' is the parameter after a gradient step, η is the learning rate, and function G is given by

$$G(\mu,\mu^*) = \mathbb{E}_{x\sim\mathcal{N}(\mu^*,\mathrm{Id})}[-\frac{1}{2}\tanh''(\mu^\top x)\|\mu\|^2 x + \tanh'(\mu^\top x)\mu^\top x x - \tanh'(\mu^\top x)\mu].$$

Note we use the population gradient here only for simplicity; in the Appendix we show that empirical estimates of the gradient suffice. After some calculation, we can show that

$$\left\|\mu_t' - \mu_t^*\right\| \leq (1-\eta)\|\mu_t - \mu_t^*\| + \eta\|\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^\top x)x] - \mu_t^*\| + \eta\left\|G(\mu_t,\mu_t^*)\right\| \,. \quad (10)$$

Using Fact 5, we know that $\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^\top x)x]$ is precisely the result of one step of EM starting from μ_t, and it is known [DTZ17] that the EM update contracts the distance between μ_t and μ_t^* as follows:

$$\|\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^\top x)x] - \mu_t^*\| \leq \lambda_1\|\mu_t - \mu_t^*\| \quad \text{for some} \;\; \lambda_1 < 1 \quad (11)$$

It remains to control the second term in Eq. (10), for which we prove the following:

Lemma 11 (Informal, see Lemma C.9 for more details)**.** *When $\|\mu^*\| = \Omega(1)$ and the noise scale $t = \Theta(1)$, then for every μ with $\langle \hat{\mu}, \hat{\mu}^* \rangle$ being $\Omega(1)$, the following inequality holds:*

$$\|G(\mu_t, \mu_t*)\| \leq \lambda_2\|\mu_t - \mu_t^*\| \quad \textit{for some} \;\; \lambda_2 < 1 \,.$$

Combining Eq. (11) and Lemma 11 with Eq. (10), we have

$$\left\|\mu_t' - \mu_t^*\right\| \leq (1 - \eta(1 - \lambda_1 - \lambda_2))\|\mu_t - \mu_t^*\| \,. \quad (12)$$

We can set parameters to ensure that $\lambda_1 + \lambda_2 < 1$ and therefore that $\|\mu_t - \mu_t^*\|$ contracts with each gradient step. Applying Lemma 11 and Eq. (12), we obtain the following lemma summarizing the behavior of gradient descent on the DDPM objective in the low noise regime.

Lemma 12 (Informal). *For any $\varepsilon > 0$ and for the noise scale $t = \Theta(1)$, starting from an initialization μ_t for which $\langle \hat{\mu}_t, \hat{\mu}_t^* \rangle = \Omega(1)$, running gradient descent on the DDPM objective $L_t(s_\mu)$ will give us mean parameter $\tilde{\mu}$ such that $\|\tilde{\mu} - \mu^*\| \le O(\varepsilon)$.*

Combining Lemma 10 and Lemma 12, we obtain our first main result, Theorem 7, for learning mixtures of two Gaussians with constant separation. For the full technical details, see Appendix C.

3 Extensions: small separation and more components

3.1 Mixtures of two Gaussians with small separation

In this section, we briefly sketch how the ideas from Section 2 can be extended to give our second main result, namely on learning mixtures of two Gaussians even with *small separation*. We defer the full technical details to Appendix D.

Theorem 13. *Suppose a mixture of two Gaussians has mean parameter μ^* that satisfies $\|\mu^*\| = \Omega(\frac{1}{\mathrm{poly}(d)})$. Then, for any $\varepsilon > 0$, there exists a modification of Algorithm 1 that provides an estimate μ such that $\|\mu - \mu^*\| \le O(\varepsilon)$ with high probability. Moreover, the algorithm has time and sample complexity $\mathrm{poly}(d)/\varepsilon^8$ (see Theorem D.1 for more precise quantitative bounds).*

Algorithm modification. The algorithm that we analyze runs *projected* gradient descent on the DDPM objective but only in the high noise scale regime where $t = O(\log d)$. At each step, we project the iterate μ to the ball of radius R, where R is an empirical estimate for $\|\mu^*\|$ obtained by drawing samples $x_1, \ldots, x_n$ from the data distribution and forming $R \triangleq (\frac{1}{n}\sum_{i=1}^n \|x_i\|^2 - d)^{1/2}$.

Proof sketch. Lemma 9 and Lemma 10 apply even when the components of the mixture have small separation, and they show that running gradient descent on the DDPM objective results in μ_t and μ_t^* being $O(1)$ close in angular distance. Although our analysis can be extended to show that gradient descent can achieve $O(\varepsilon)$ angular distance, this does not guarantee that $\|\mu_t - \mu_t^*\|$ is $O(\varepsilon)$. If in addition to being $O(\varepsilon)$ close in angular distance, we also have that $\|\mu_t\| \approx \|\mu_t^*\|$, then it is easy to see that $\|\mu_t - \mu_t^*\|$ is indeed $O(\varepsilon)$.

Observe that if R is approximately equal to $\|\mu_t^*\|$, then the projection step in our algorithm ensures that our final estimate μ_t satisfies this additional condition of $\|\mu_t\| \approx \|\mu_t^*\|$. It is not hard to show that R^2 is an unbiased estimate of $\|\mu_t^*\|^2$, so standard concentration shows that taking $n = \mathrm{poly}(d, \frac{1}{\varepsilon})$ suffices to ensure that R is sufficiently close to $\|\mu_t^*\|$.

3.2 Mixtures of K Gaussians, from a warm start

In this section, we state our third main result, namely for learning mixtures of K Gaussians given by Eq. (6) from a warm start, and provide an overview of how the ideas from Section 2 can be extended to obtain this result.

Assumption 14. *(Separation) For a mixture of K Gaussians given by Eq. (6), for every pair of components $i, j \in \{1, 2, \ldots, K\}$ with $i \neq j$, we assume that the separation between their means $\|\mu_i^* - \mu_j^*\| \ge C\sqrt{\log(\min(K, d))}$ for sufficiently large absolute constant $C > 0$.*

Assumption 15. *(Initialization) For each component $i \in \{1, 2, \ldots, K\}$, we have an initialization $\mu_i^{(0)}$ with the property that $\|\mu_i^{(0)} - \mu_i^*\| \le C'\sqrt{\log(\min(K, d))}$ for sufficiently small absolute constant $C' > 0$.*

Theorem 16. *Suppose a mixture of K Gaussians satisfies Assumption 14. Then, for any $\varepsilon = \Theta(1/\mathrm{poly}(d))$, running gradient descent on the DDPM objective (Algorithm 1) at low noise scale $t = O(1)$ and with initialization satisfying Assumption 15 results in mean parameters $\{\mu_i\}_{i=1}^K$ such that with high probability, the mean parameters satisfy $\|\mu_i - \mu_i^*\| \le O(\varepsilon)$ for each $i \in \{1, 2, \ldots, K\}$. Additionally, the runtime and sample complexity of the algorithm is $\mathrm{poly}(d, 1/\varepsilon)$ (see Theorem E.1 for more precise quantitative bounds).*

We provide a brief overview of the proof here. The full proof can be found in Appendix E.

Proof sketch. For learning mixtures of two Gaussians, we have already established the connection between gradient descent on the DDPM objective and the EM algorithm. For mixtures of K Gaussians, however, in a local neighborhood around the ground truth parameters θ^*, we show an equivalence between *gradient* EM (recall gradient EM performs one-step of gradient descent on the "M" step objective) and gradient descent on the DDPM objective. In particular, our main technical lemma (Lemma E.4) shows that for noise scale $t = \Theta(1)$ and for any μ_i that satisfies $\|\mu_i - \mu_i^*\| \le O(\sqrt{\log(\min(K, d))})$, we have

$$-\nabla_{\mu_{i,t}} L_t(s_{\theta_t}) \approx \mathbb{E}_{X_t}[w_{i,t}(X_t)(X_t - \mu_{i,t})].$$

Therefore, the iterate $\mu'_{i,t}$ resulting from a single gradient step on the DDPM objective $L_t(s_{\theta_t})$ with learning rate η is given by

$$\mu'_{1,t} = \mu_{1,t} - \eta \nabla_{\mu_{1,t}} L_t(s_{\theta_t}) \approx \mu_{1,t} + \eta\, \mathbb{E}_{X_t}[w_{1,t}(X_t)(X_t - \mu_{1,t})]. \tag{13}$$

Comparing Fact 6 with Eq. (13), we see the correspondence in this regime between gradient descent on the DDPM objective to gradient EM. Using this connection and an existing local convergence guarantee from the gradient EM literature [SN21, KC20], we obtain our main theorem for mixtures of K Gaussians. Full details can be found in Appendix E.

Acknowledgments

SC would like to thank Sinho Chewi, Khashayar Gatmiry, Frederic Koehler, and Holden Lee for enlightening discussions on sampling and score estimation. KS and AK are supported by the NSF AI Institute for Foundations of Machine Learning (IFML). SC is supported by NSF Award 2103300.

References

[AG22] Ahmed El Alaoui and Jason Gaitonde. Bounds on the covariance matrix of the sherrington-kirkpatrick model. *arXiv preprint arXiv:2212.02445*, 2022.

[BDD23] Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

[BDJ+22] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022.

[BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[BM11] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[BMR22] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint 2002.00107*, 2022.

[BRST21] Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous lwe. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 694–707, 2021.

[BS15] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. *SIAM Journal on Computing*, 44(4):889–911, 2015.

[BWY17] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. 2017.

[CCL+23a] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023.

[CCL+23b] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023.

[CDD23] Sitan Chen, Giannis Daras, and Alexandros G Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *arXiv preprint arXiv:2303.03384*, 2023.

[Cel22] Michael Celentano. Sudakov-fernique post-amp, and a new proof of the local convexity of the tap free energy. *arXiv preprint arXiv:2208.09550*, 2022.

[CFM21] Michael Celentano, Zhou Fan, and Song Mei. Local convexity of the tap free energy and amp convergence for z2-synchronization. *arXiv preprint arXiv:2106.11428*, 2021.

[CLL22] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022.

[DB22] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

[DBTHD21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709. Curran Associates, Inc., 2021.

[DHKK20] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020.

[DK20] Ilias Diakonikolas and Daniel M Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195. IEEE, 2020.

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.

[DMM09] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[DMM10] David L Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing: I. motivation and construction. In *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*, pages 1–5. IEEE, 2010.

[DS07] Sanjoy Dasgupta and Leonard J Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[DTZ17] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710. PMLR, 2017.

[EAMS22] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.

[Eld13] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

[Eld20] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3-4):737–755, 2020.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

[HP15] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760, 2015.

[Kab03] Yoshiyuki Kabashima. A cdma multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111, 2003.

[Kan21] Daniel M Kane. Robust learning of mixtures of gaussians. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1246–1258. SIAM, 2021.

[KC20] Jeongyeol Kwon and Constantine Caramanis. The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020.

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

[LL22] Allen Liu and Jerry Li. Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261, 2022.

[LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

[LM23] Allen Liu and Ankur Moitra. Robustly learning general mixtures of gaussians. *Journal of the ACM*, 2023.

[LWCC23] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

[LWYL22] Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

[MM09] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

[MV21] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1), 2021.

[MW23] Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[Pid22] Jakiw Pidstrigach. Score-based generative models detect manifolds. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35852–35865. Curran Associates, Inc., 2022.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[RV17] Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.

[SCS+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[SN21] Nimrod Segol and Boaz Nadler. Improved convergence guarantees for learning gaussian mixture models by em and gradient em. *Electronic journal of statistics*, 15(2):4510–4544, 2021.

[SOAJ14] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. *Advances in Neural Information Processing Systems*, 27, 2014.

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Ver] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004. Special Issue on FOCS 2002.

[WJ+08] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[WY22] Andre Wibisono and Kaylee Y. Yang. Convergence in KL divergence of the inexact Langevin algorithm with application to score-based generative models. *arXiv preprint 2211.01512*, 2022.

[XHM16] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. *Advances in Neural Information Processing Systems*, 29, 2016.

[YYS17] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence analysis of gradient em for multi-component gaussian mixture. *arXiv preprint arXiv:1705.08530*, 2017.

[ZLS20] Ruofei Zhao, Yuanzhi Li, and Yuekai Sun. Statistical convergence of the em algorithm on gaussian mixture models. 2020.

Roadmap. In Appendix A, we provide proofs of some simple lemmas from Section 1.3 and some basic inequalities. In Appendix B we give additional notation and preliminaries. In Appendix C, we provide the proof details for Theorem 7, our result on learning mixtures of two Gaussians with constant separation. In Appendix D, we extend this analysis to give a proof of Theorem 13, our result on learning mixtures of two Gaussians with small separation. In Appendix E, we provide the proof details for Theorem 16, our result on learning mixtures of K Gaussians. In Appendix F, we give further deferred proofs. Finally, in Appendix G, we provide experiments to confirm our theoretical results.

A Proofs from Section 1.3

A.1 X_t is a mixture of Gaussians

Proof of Lemma 3. Suppose X_0 is mixture of K Gaussians with density function given by

$$q_0 = \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\mu^*_{i,0}, \mathrm{Id})$$

We know that $X_t = \exp(-t)X_0 + \sqrt{1-\exp(-2t)}Z_t$ where $Z_t \sim \mathcal{N}(0, \mathrm{Id})$. Then, by change of variable of probability density, we have

$$\text{pdf of } \exp(-t)X_0 = \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\mu^*_{i,0}\exp(-t), \exp(-2t)\cdot \mathrm{Id})$$
$$\text{pdf of } \sqrt{1-\exp(-2t)}Z_t = \mathcal{N}(0, (1-\exp(-2t))\cdot \mathrm{Id})\,.$$

Combining these, we have

$$q_t(X_t) = \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\mu^*_{i,t}, I) \qquad \text{where} \qquad \mu^*_{i,t} = \mu^*_{i,0}\exp(-t)\,,$$

as claimed. □

A.2 Derivation of score function

Proof of Lemma 4. For mixtures of K Gaussians in the form of Eq. (6), the score function at time t is given by

$$\begin{aligned}\nabla \log q_t(x) &= -\frac{\sum_{i=1}^{K} e^{-\frac{\|x-\mu^*_{i,t}\|^2}{2}}(x-\mu^*_{i,t})}{\sum_{j=1}^{K} e^{-\frac{\|x-\mu^*_{j,t}\|^2}{2}}} \\ &= \sum_{i=1}^{K} w^*_{i,t}(x)\mu^*_{i,t} - x \quad \text{where} \quad w^*_{i,t}(x) = \frac{e^{-\frac{\|x-\mu^*_{i,t}\|^2}{2}}}{\sum_{j=1}^{K} e^{-\frac{\|x-\mu^*_{j,t}\|^2}{2}}}.\end{aligned}$$

For mixtures of two Gaussians in the form of Eq. (7), the score function is given by

$$\begin{aligned}\nabla \log q_t(x) &= w^*_{1,t}(x)\mu^*_{1,t} + w^*_{2,t}(x)\mu^*_{2,t} - x \\ &= w^*_{1,t}(x)\mu^* - (1-w^*_{1,t}(x))\mu^* - x \\ &= (2w^*_{1,t}(x)-1)\mu^* - x \end{aligned} \tag{A.1}$$

By simplifying $w^*_{1,t}(x)$, we obtain

$$\begin{aligned}w^*_{1,t}(x) &= \frac{1}{1+\exp(\frac{\|x-\mu^*\|^2}{2} - \frac{\|x+\mu^*\|^2}{2})} \\ &= \frac{1}{1+\exp(-2\mu^{*\top}x)} \\ &= \sigma(2\mu^{*\top}x)\end{aligned} \tag{A.2}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Using Eq. (A.2) in Eq. (A.1), we obtain

$$\nabla \log q_t(x) = \tanh({\mu^*}^\top x)\mu^* - x.$$

□

B Additional notations and preliminaries

In this section, we provide additional notations and preliminaries for the proofs to follow. Recall that we use $L_t(s_{\theta_t})$ to denote the population denoising loss at noise scale t.

$$L_t(s_{\theta_t}) = \mathbb{E}\Big[\Big\|s_{\theta_t}(X_t) + \frac{Z_t}{\sqrt{1-\exp(-2t)}}\Big\|^2\Big].$$

We use $L_t(s_{\theta_t}(x_0, z_t))$ to denote the denoising loss at noise scale t on a sample x_0 from the data distribution and z_t from the standard Gaussian distribution:

$$L_t(s_{\theta_t}(x_0, z_t)) = \Big\|s_{\theta_t}(x_t) + \frac{z_t}{\sqrt{1-\exp(-2t)}}\Big\|^2,$$

where $x_t = \exp(-t)x_0 + \sqrt{1-\exp(-2t)}z_t$. We use α_t as shorthand notation for $\exp(-t)$ and β_t as shorthand notation for $\sqrt{1-\exp(-2t)}$.

For mixtures of two Gaussians, we use B to denote the upper bound on $\|\mu^*\|^2$, that is,

$$\|\mu^*\|^2 \le B\,.$$

Throughout, we assume that $B = \mathrm{poly}(d)$.

For any vector v, we use $\hat{v}$ to denote the unit vector along the direction of v. For a vector v, we use $[v]_i$ to denote the i^{th} coordinate of v. Similarly, for a matrix X, we use $[X]_i$ to denote the i^{th} row of the matrix. For any positive integer n, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We use $\mathcal{N}(\mu, \sigma^2 \cdot \mathrm{Id})$ to denote the standard Gaussian with mean μ and covariance $\sigma^2 \cdot \mathrm{Id}$. Sometimes, we use a shorter notation $\mathcal{N}_\mu$ to denote $\mathcal{N}(\mu, \mathrm{Id})$. For any two quantities X and Y that are both implicitly functions of some parameter a over $\mathbb{R}_{\ge 0}$, we use the shorthand $X \lesssim Y$ and $X = O(Y)$ interchangeably to denote that there exists absolute constant $C > 0$ such that for all a sufficiently large, $X(a) \le CY(a)$. We also use the shorthand $X \gtrsim Y$ and $X = \Omega(Y)$, defined in the obvious way.

Finally, we will use the following standard bounds.

Lemma B.1 (Sub-Gaussian norm, see e.g. [Ver])**.** *The sub-Gaussian norm of a random variable $X \in \mathbb{R}$, denoted by $\|X\|_{\psi_2}$ is defined as*

$$\|X\|_{\psi_2} = \inf\{t > 0 \,:\, \mathbb{E}[\exp(X^2/t^2)] \le 2\}.$$

The sub-Gaussian norm has the following properties:

1. *(Bounded): Any bounded random variable X (i.e., there is a finite A for which $|X| \le A$ with probability 1) is sub-Gaussian:*

$$\|X\|_{\psi_2} \le \frac{A}{\sqrt{\ln 2}}$$

2. *(Centering): If X is a sub-Gaussian random variable, then $X - \mathbb{E}[X]$ is also a sub-Gaussian random variable. Specifically, the following holds for some absolute constant C.*

$$\|X - \mathbb{E}[X]\|_{\psi_2} \le C\|X\|_{\psi_2}$$

3. *(Moment generating function bound): If X is a sub-Gaussian random variable with $E[X] = 0$, then*

$$\mathbb{E}[\exp(\lambda X)] \le \exp(C\lambda^2\|X\|_{\psi_2}^2) \quad \text{for all } \lambda \in \mathbb{R},$$

where C is some absolute constant.

4. *(Sum of sub-Gaussian random variables): If X_1 and X_2 are mean zero sub-Gaussian random variables, then*

$$\|X_1 + X_2\|_{\psi_2} \leq \|X_1\|_{\psi_2} + \|X_2\|_{\psi_2} .$$

5. *(Product with a bounded random variable): If X is a sub-Gaussian random variable and Y is a bounded random variable $Y \in [0, 1]$, then*

$$\|XY\|_{\psi_2} \leq \|X\|_{\psi_2} .$$

Lemma B.2 (Sub-exponential norm, see e.g. [Ver])**.** *The sub-exponential norm of a random variable $X \in \mathbb{R}$, denoted by $\|X\|_{\psi_1}$ is defined as*

$$\|X\|_{\psi_1} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|/t)] \leq 2\}.$$

The sub-exponential norm has the following properties:

1. *(Sum of sub-exponential distributions): If X_1 and X_2 are mean-zero sub-exponential random variables, then $X_1 + X_2$ is also a mean-zero sub-exponential variable. Specifically,*

$$\|X_1 + X_2\|_{\psi_1} \leq \sqrt{2}(\|X_1\|_{\psi_1} + \|X_2\|_{\psi_1}) .$$

2. *(Centering) If X is a sub-exponential random variable, then $X - \mathbb{E}[X]$ is sub-exponential with*

$$\|X - \mathbb{E}[X]\|_{\psi_1} \leq C\|X\|_{\psi_1},$$

where C is some absolute constant.

Proof. The proof follows from following the equivalent definition of a sub-exponential random variable: If any random variable X satisfies

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(C\|X\|_{\psi_1}^2 \lambda^2) \ \text{ for all } \lambda \text{ such that} |\lambda| \leq \frac{1}{C\|X\|_{\psi_1}^2},$$

for some constant C, then X is sub-exponential random variable with sub-exponential norm $\|X\|_{\psi_1}$. Then, for any $|\lambda| \leq \frac{1}{2C\max(\|X_1\|_{\psi_1}^2, \|X_2\|_{\psi_1}^2)}$, the MGF of $X_1 + X_2$ is given by

$$\begin{aligned}
\mathbb{E}[\exp(\lambda(X_1 + X_2))] &\leq \mathbb{E}[\exp(2\lambda X_1)]^{1/2}\mathbb{E}[\exp(2\lambda X_2)]^{1/2} \\
&\leq \exp(C\|X_1\|_{\psi_1}^2 2\lambda^2)\exp(C\|X_2\|_{\psi_1}^2 2\lambda^2) \\
&\leq \exp(C\lambda^2(2\|X_1\|_{\psi_1}^2 + 2\|X_2\|_{\psi_1}^2)) .
\end{aligned}$$

□

Using $\|X_1\|_{\psi_1} + \|X_2\|_{\psi_1} \geq \max(\|X_1\|_{\psi_1}, \|X_2\|_{\psi_1})$, we know that above inequality is true for any λ with $|\lambda| \leq \frac{1}{2C(\|X_1\|_{\psi_1} + \|X_2\|_{\psi_1})^2} \leq \frac{1}{2C\max(\|X_1\|_{\psi_1}^2, \|X_2\|_{\psi_1}^2)}$. This completes the proof.

Lemma B.3 (Corollary 2.8.4 in [Ver])**.** *(Bernstein's inequality for sub-exponential random variable) Let $X_1, X_2, \ldots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $\varepsilon \geq 0$, we have*

$$\Pr\left[\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right| \geq \varepsilon\right] \leq 2\exp\left[-cN\min\Big(\frac{\varepsilon}{\max_i \|X_i\|_{\psi_1}}, \frac{\varepsilon^2}{(\max_i \|X_i\|_{\psi_1})^2}\Big)\right]$$

where $c > 0$ is some absolute constant.

C Learning mixtures of two Gaussians with constant separation

In this section, we provide the details and proofs for learning mixtures of two Gaussians with constant separation. Our results in this section can be summarized in the following theorem statement.

Theorem C.1 (Formal version of Theorem 7)**.** *Let q be a mixture of two Gaussians (in the form of Eq. (7)) with mean parameter μ^* satisfying $\|\mu^*\| > c$ for some absolute constant $c > 0$. Recalling that B denotes an* a priori *upper bound on $\|\mu^*\|$, we have that for any $\varepsilon \le \varepsilon'$ where $\varepsilon' \lesssim \frac{1}{d^2 B^9}$, there exists a procedure satisfying the following. If the procedure is run for at least $\Omega(B^6 \log(d/\varepsilon))$ iterations with at least $\mathrm{poly}(d, B)/\varepsilon^2$ samples from q, then it outputs $\tilde{\mu}$ such that $\|\tilde{\mu} - \mu^*\| \le \varepsilon$ with high probability.*

As described earlier, the procedure first runs gradient descent on the DDPM objective described in Algorithm 1 from a random Gaussian initialization in a high noise scale regime with noise scale $t_1 = O(\log d)$. It then uses the output of the first step as initialization and runs the Algorithm 1 in a low noise scale regime with noise scale $t_2 = O(1)$.

We begin by calculating the form of the gradient updates:

Lemma C.2. *For any noise scale $t > 0$, the gradient update for the mixture of two Gaussians on the DDPM objective is given by*

$$-\nabla_{\mu_t} L_t(s_{\mu_t}) = \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\Big[\big(\tanh(\mu_t^\top x) - \frac{1}{2}\tanh''(\mu_t^\top x)\|\mu_t\|^2 + \tanh'(\mu_t^\top x)\mu_t^\top x\big)x\Big] - \mu_t - \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\Big[\tanh'(\mu_t^\top x)\mu_t\Big]\,.$$

The proof of Lemma C.2 is given in Appendix F.1.

C.1 High noise regime–connection to power iteration

Here we show that running population gradient descent on the DDPM objective at *high* noise scale behaves like power iteration on the covariance matrix of the data and thus reaches an iterate μ with constant correlation with μ^*.

Lemma C.3. *For any noise scale $t > t'$ and number of samples $n > n'$ where $t' \lesssim \log d$ and $n' = \Theta\big(\frac{d^4 B^3}{\varepsilon^2}\big)$, with high probability, the negative gradient of the diffusion model objective $L_t(s_t)$ can be approximated by $2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t$. More precisely, given independent samples $\{x_{i,t}\}_{i=1,\dots,n}$ from q_t generated using noise vectors $\{z_{i,t}\}_{i=1,\dots,n}$ sampled from $\mathcal{N}(0,\mathrm{Id})$, we have*

$$\Big\| -\nabla\Big(\frac{1}{n}\sum_{i=1}^n L_t(s_{\mu_t}(x_{i,t}, z_{i,t}))\Big) - \Big(2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t\Big)\Big\| \le 250\sqrt{d}\|\mu_t\|^5 + 10\|\mu_t\|^3\|\mu_t^*\|^2 + \varepsilon\,.$$

Proof. Recall that the population gradient update on the DDPM objective is given by

$$\begin{aligned}
-\nabla L_t(s_{\mu_t}) &= \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\big[\tanh(\mu_t^\top x)x - \frac{1}{2}\tanh''(\mu_t^\top x)\|\mu_t\|^2 x + \tanh'(\mu_t^\top x)\mu_t^\top x x\big] \\
&\quad - \mu_t - \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh'(\mu_t^\top x)\mu_t] \\
&= \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\big[\tanh(\mu_t^\top x)x - \frac{1}{2}\tanh''(\mu_t^\top x)\|\mu_t\|^2 x + \tanh'(\mu_t^\top x)\mu_t^\top x\mu_t^* \\
&\quad + \tanh''(\mu_t^\top x)\mu_t^\top x\mu_t\big] - \mu_t\,,
\end{aligned}$$

where the last equality follows from the Stein's lemma on $\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh'(\mu_t^\top x)\mu_t^\top x x]$, as

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh'(\mu_t^\top x)\mu_t^\top x x] = \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh'(\mu_t^\top x)\mu_t^\top x\mu_t^* + \tanh'(\mu_t^\top x)\mu_t + \tanh''(\mu_t^\top x)\mu_t^\top x\mu_t]\,.$$

Using Taylor's theorem, we know that

$$\begin{aligned}
&\tanh(\mu_t^\top x) = \mu_t^\top x - \frac{2}{3}(\mu_t^\top x)^3 + O(\xi(x)^5) \qquad \text{where } \xi(x) \in [0, \mu_t^\top x] \\
\implies &\tanh(\mu^\top x)x = \mu^\top x x - \frac{2}{3}(\mu_t^\top x)^3 x + O(\xi(x)^5 x) \\
\implies &\Big\|\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^\top x)x] - \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\big[\mu_t^\top x x - \frac{2}{3}(\mu_t^\top x)^3 x\big]\Big\| \le \|\mathbb{E}[\xi(x)^5 x]\| \lesssim \sqrt{d}\|\mu_t\|^5
\end{aligned}$$

where the last inequality follows from $\left\|\mathbb{E}[\xi(x)^5 x]\right\| \leq \mathbb{E}[|\mu_t^\top x|^5\|x\|] \leq \left(\mathbb{E}[|\mu_t^\top x|^{10}]\right)^{1/2}\left(\mathbb{E}[\,\|x\|^2]\right)^{1/2} \lesssim \|\mu_t\|^5\sqrt{d+\|\mu_t^*\|^2} \lesssim \sqrt{d}\|\mu_t\|^5$. Similarly, using Taylor's theorem, we get

$$
\begin{aligned}
&\tanh''(\mu_t^\top x) = -2\mu_t^\top x + O(\xi(x)^3) \qquad \text{where } \xi(x) \in [0, \mu_t^\top x] \\
\implies &\tanh''(\mu_t^\top x)\left(-\frac{1}{2}\|\mu_t\|^2 x + \mu_t^\top x\mu_t\right) = \left(-2\mu_t^\top x + O(\xi(x)^3)\right)\left(-\frac{1}{2}\|\mu_t\|^2 x + \mu_t^\top x\mu_t\right) \\
\implies &\left\|\mathbb{E}[\tanh''(\mu_t^\top x)\big(-\frac{1}{2}\|\mu_t\|^2 x + \mu_t^\top x\mu_t)] - \mathbb{E}\Big[-2\mu_t^\top x\left(-\frac{1}{2}\|\mu_t\|^2 x + \mu_t^\top x\mu_t\right)\Big]\right\| \\
&\quad \leq \left\|-\frac{1}{2}\|\mu_t\|^2\,\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,I)}[O(\xi(x)^3)x] + \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,I)}[O(\xi(x)^3)\mu_t^\top x\mu_t]\right\| \\
&\quad \leq \frac{1}{2}\|\mu_t\|^2\,\mathbb{E}[|\mu_t^\top x|^3\|x\|] + \|\mu_t\|\,\mathbb{E}[|\mu_t^\top x|^4] \\
&\quad \leq \frac{1}{2}\|\mu_t\|^2\sqrt{\mathbb{E}[|\mu_t^\top x|^6]\mathbb{E}\|x\|^2]} + \|\mu_t\|\,\mathbb{E}[|\mu_t^\top x|^4] \\
&\quad \leq 10\|\mu_t\|^5\sqrt{d} + 6\|\mu_t\|^5
\end{aligned}
$$

Using Taylor's theorem for $\tanh'$, we get

$$
\begin{aligned}
&\tanh'(\mu_t^\top x) = 1 - (\mu_t^\top x)^2 + O(\xi(x)^4) \qquad \text{where } \xi(x) \in [0, \mu_t^\top x] \\
\implies &\tanh'(\mu_t^\top x)\mu_t^\top x\mu_t^* = \mu_t^\top x\mu_t^* - (\mu_t^\top x)^3\mu_t^* + O(\xi(x)^4\mu_t^\top x\mu_t^*) \qquad \text{where } \xi(x) \in [0, \mu_t^\top x] \\
\implies &\left\|\mathbb{E}[\tanh'(\mu_t^\top x)\mu_t^\top x\mu_t^*] - \mathbb{E}[\mu_t^\top x\mu_t^* - (\mu_t^\top x)^3\mu_t^*]\right\| \leq \left\|\mathbb{E}[\xi(x)^4(\mu_t^\top x)\mu_t^*]\right\| \\
&\qquad\qquad \leq \mathbb{E}[|\mu_t^\top x|^5]\|\mu_t^*\| \lesssim \|\mu_t^*\|\|\mu_t\|^5
\end{aligned}
$$

Additionally, we have

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[xx^\top\mu_t(1+\|\mu_t\|^2) - \frac{2}{3}(\mu_t^\top x)^3 x - 2\mu_t(\mu_t^\top x)^2 + \mu_t^\top x\mu_t^* - (\mu_t^\top x)^3\mu_t^*] \\
&= (I + \mu_t^*\mu_t^{*\top})\mu_t(1+\|\mu_t\|^2) - \frac{5}{3}\mathbb{E}[(\mu_t^\top x)^3\mu_t^*] + \mu_t^*\mu_t^{*\top}\mu_t - 4\mathbb{E}[\mu_t(\mu_t^\top x)^2] \\
&= (I + \mu_t^*\mu_t^{*\top})\mu_t(1+\|\mu_t\|^2) - \frac{5\mu_t^*}{3}((\mu_t^\top\mu_t^*)^3 + 3(\mu_t^\top\mu_t^*)\|\mu_t\|^2) \\
&\qquad + \mu_t^*\mu_t^{*\top}\mu_t - 4\mu_t(\|\mu_t\|^2 + (\mu_t^\top\mu_t^*)^2) \\
&= \mu_t^*\mu_t^{*\top}\mu_t(2 - 4\|\mu_t\|^2) + \mu_t(1 - 3\|\mu_t\|^2) - \frac{5\mu_t^*(\mu_t^\top\mu_t^*)^3}{3} - 4\mu_t(\mu_t^\top\mu_t^*)^2
\end{aligned}
$$

where the second equality uses Stein's lemma on $\mathbb{E}[(\mu_t^\top x)^3 x]$ and $\mathbb{E}[xx^\top] = \mathrm{Id} + \mu_t^*\mu_t^{*\top}$ and the third equality uses Gaussian moments for $\mathbb{E}[(\mu_t^\top x)^2]$ and $\mathbb{E}[(\mu_t^\top x)^3]$. Putting it all together and using triangle inequality, we obtain the desired bound on $\|-\nabla L_t(s_{\mu_t}) - (2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t)\|$.

$$
\begin{aligned}
&\|-\nabla L_t(s_{\mu_t}) - (2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t)\| \\
&\leq \left\|-\nabla L_t(s_{\mu_t}) - \mathbb{E}[xx^\top\mu_t(1+\|\mu_t\|^2) - \frac{2}{3}(\mu_t^\top x)^3 x - 2\mu_t(\mu_t^\top x)^2 + \mu_t^\top x\mu_t^* - (\mu_t^\top x)^3\mu_t^* - \mu_t]\right\| \\
&\quad + \Big\|\mathbb{E}[xx^\top\mu_t(1+\|\mu_t\|^2) - \frac{2}{3}(\mu_t^\top x)^3 x - 2\mu_t(\mu_t^\top x)^2 + \mu_t^\top x\mu_t^* - (\mu_t^\top x)^3\mu_t^* - \mu_t] \\
&\qquad - \left(2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t\right)\Big\| \\
&\leq \left(200\sqrt{d}\|\mu_t\|^5 + 10\|\mu_t\|^5\sqrt{d} + 6\|\mu_t\|^5 + 20\|\mu_t^*\|\|\mu_t\|^5\right) + 10\|\mu_t\|^3\|\mu_t^*\|^2 \\
&\leq 250\sqrt{d}\|\mu_t\|^5 + 10\|\mu_t\|^3\|\mu_t^*\|^2
\end{aligned}
$$

Using Lemma E.7 and triangle inequality, we obtain the result. □

We will use the following simple bound on the correlation between the ground truth and a random initialization:

Lemma C.4. *A randomly initialized $\mu_0 \sim \mathcal{N}(0, \mathrm{Id})$ satisfies that $\left|\langle \hat{\mu}_0, \hat{\mu}^* \rangle\right| \geq \frac{1}{2d}$ with probability at least $1 - O(d^{-1/2})$.*

Proof. For $\mu_0 \sim \mathcal{N}(0, I)$, we know that $\langle \mu_0, \hat{\mu}^* \rangle \sim \mathcal{N}(0, I)$. Using Gaussian anti-concentration, with probability at least $1 - 1/\sqrt{d}$, we have $\left|\langle \mu_0, \hat{\mu}^* \rangle\right| \geq 1/\sqrt{d}$. Because the L_2 norm of a Gaussian vector is sub-exponential, with probability at least $1 - \exp(-\Omega(d))$, we have $\|\mu_0\| \leq 2\sqrt{d}$. Using the norm bound, with probability at least $1 - 1/\sqrt{d} - \exp(-O(d)) = 1 - O(d^{-1/2})$, we obtain the claimed bound on $\left|\langle \hat{\mu_0}, \hat{\mu}^* \rangle\right|$. □

We can now track the correlation between the iterates of gradient descent and the ground truth:

Lemma C.5. *Suppose that the vector μ_t satisfies $|\langle \hat{\mu}_t, \hat{\mu}_t^* \rangle| \geq \frac{1}{2d}$, and let μ_t' denote the iterate resulting from a single empirical gradient step with learning rate η starting from μ_t. Suppose that the empirical gradient and the population gradient differ by at most ε. Denote the angle between μ_t (resp. μ_t') and μ_t^* by θ (resp. θ'). Then*

$$\tan\theta' = \max\left(\kappa_1 \tan\theta, \kappa_2\right)$$

for

$$\kappa_1 = \frac{1 - 3\eta\|\mu_t\|^2}{1 - 3\eta\|\mu_t\|^2 + \eta(\|\mu_t^*\|^2 - 500\sqrt{d^3}\|\mu_t\|^4 - 20d\|\mu_t\|^2\|\mu_t^*\|^2 - \eta\tilde{\varepsilon})} \ ,$$
$$\kappa_2 = \frac{500\eta\sqrt{d^3}\|\mu_t\|^4 + 20\eta d\|\mu_t\|^2\|\mu_t^*\|^2 + \eta\tilde{\varepsilon}}{\|\mu_t^*\|^2} \ \text{ and } \ \tilde{\varepsilon} \lesssim \frac{d\varepsilon}{\|\mu_t\|} \, .$$

Proof. Define $\hat{\mu}_t^{*\perp}$ as the orthogonal vector to μ_t^* in the plane of μ_t and μ_t^*. Note that μ_t' still lies in this plane, so the orthogonal vector to μ_t^* in the plane of μ_t' and μ_t^* is also given by $\hat{\mu}_t^{*\perp}$.

We have

$$\begin{aligned}
\tan\theta' &= \frac{\langle \hat{\mu}^{*\perp}, \hat{\mu}_t' \rangle}{\langle \hat{\mu}_t^*, \hat{\mu}_t' \rangle} = \frac{\langle \hat{\mu}_t^{*\perp}, \mu_t' \rangle}{\langle \hat{\mu}_t^*, \mu_t' \rangle} \\
&= \frac{\langle \hat{\mu}_t^{*\perp}, \mu_t + \eta F(\mu_t, \mu_t^*) \rangle + \langle \hat{\mu}_t^{*\perp}, -\eta\nabla L_t(s_t) - \eta F(\mu_t, \mu_t^*) \rangle + \eta\varepsilon}{\langle \hat{\mu}_t^*, \mu_t + \eta F(\mu_t, \mu_t^*) \rangle + \langle \hat{\mu}_t^{*\perp}, -\eta\nabla L_t(s_t) - \eta F(\mu_t, \mu_t^*) \rangle - \eta\varepsilon} \\
&\qquad\qquad \text{where} \quad F(\mu, \mu^*) = \left(2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t\right) \\
&\leq \frac{\sigma_2\langle \hat{\mu}_t^{*\perp}, \mu_t \rangle + \eta\left\|\nabla L_t(s_t) + F(\mu_t, \mu_t^*)\right\| + \eta\varepsilon}{\sigma_1\langle \hat{\mu}_t^*, \mu_t \rangle - \eta\left\|\nabla L_t(s_t) + F(\mu_t, \mu_t^*)\right\| - \eta\varepsilon} \qquad \text{(C.1)}
\end{aligned}$$

where σ_1 and σ_2 are the first and second eigenvalues of $\mathrm{Id} + F(\mu_t, \mu_t^*) = (1 - 3\eta\|\mu_t\|^2)\mathrm{Id} + 2\eta\mu_t^*\mu_t^{*\top}$, given by

$$\sigma_1 = 1 + \eta(2\|\mu_t^*\|^2 - 3\|\mu_t\|^2)$$
$$\sigma_2 = 1 - 3\eta\|\mu_t\|^2 \ .$$

The last inequality (C.1) follows from the fact that

$$\begin{aligned}
\langle \hat{\mu}_t^*, \mu_t + \eta F(\mu_t, \mu_t^*) \rangle &= \hat{\mu}_t^{*\top}((1 - 3\eta\|\mu_t\|^2)\mathrm{Id} + 2\eta\mu_t^*\mu_t^{*\top})\mu_t \\
&= \mu_t^\top((1 - 3\eta\|\mu_t\|^2)\mathrm{Id} + 2\eta\mu_t^*\mu_t^{*\top})\hat{\mu}_t^* = \sigma_1\mu_t^\top\hat{\mu}_t^*
\end{aligned}$$

because $\hat{\mu}^*$ is the first eigenvector of $(1 - 3\eta\|\mu_t\|^2)\mathrm{Id} + 2\eta\mu_t^*\mu_t^{*\top}$. Recall from Lemma C.3 that the deviation between the negative population gradient and the power iteration update $F(\mu_t, \mu_t^*)$ is bounded by

$$\frac{\left\|\nabla L_t(s_t) + F(\mu_t, \mu_t^*)\right\|}{\langle \mu_t, \hat{\mu}_t^* \rangle} \leq \frac{250\eta\sqrt{d}\|\mu_t\|^4 + 10\eta\|\mu_t\|^2\|\mu_t^*\|^2}{\langle \hat{\mu}_t, \hat{\mu}_t^* \rangle} \leq 500\eta\sqrt{d^3}\|\mu_t\|^4 + 20d\eta\|\mu_t\|^2\|\mu_t^*\|^2 \ .$$

Substituting this into Eq. (C.1), we get

$$\begin{aligned}
\tan\theta' &\le \frac{\sigma_2\langle \hat{\mu}_t^{*\perp}, \mu_t\rangle + \eta\|\nabla L_t(s_t) + F(\mu_t,\mu_t^*)\| + \eta\varepsilon}{\langle \hat{\mu}_t^*, \mu_t\rangle(\sigma_1 - 500\eta\sqrt{d^3}\|\mu_t\|^4 - 20d\eta\|\mu_t\|^2\|\mu_t^*\|^2 - \eta\tilde{\varepsilon})} \quad \text{where} \quad \tilde{\varepsilon} \lesssim \frac{d\varepsilon}{\|\mu\|} \\
&\le \frac{\sigma_2}{\tilde{\sigma}_1}\tan\theta + \frac{1}{\tilde{\sigma}_1}\left(500\eta\sqrt{d^3}\|\mu\|^4 + 20d\eta\|\mu\|^2\|\mu_t^*\|^2 + \eta\tilde{\varepsilon}\right) \\
&\qquad\qquad \text{where} \quad \tilde{\sigma}_1 \triangleq \sigma_1 - 500\eta\sqrt{d^3}\|\mu\|^4 - 20d\eta\|\mu\|^2\|\mu_t^*\|^2 - \eta\tilde{\varepsilon} \\
&\le \left(1 - \frac{\eta\|\mu_t^*\|^2}{\tilde{\sigma}_1}\right)\frac{\sigma_2}{\tilde{\sigma}_1 - \eta\|\mu_t^*\|^2}\tan\theta + \left(\frac{\eta\|\mu_t^*\|^2}{\tilde{\sigma}_1}\right)\frac{500\eta\sqrt{d^3}\|\mu_t\|^4 + 20d\eta\|\mu_t\|^2\|\mu_t^*\|^2 + \eta\tilde{\varepsilon}}{\eta\|\mu_t^*\|^2} \\
&\le \max\left(\frac{\sigma_2}{\tilde{\sigma}_1 - \eta\|\mu_t^*\|^2}\tan\theta, \frac{500\eta\sqrt{d^3}\|\mu_t\|^4 + 20\eta d\,\|\mu_t\|^2\|\mu_t^*\|^2 + \eta\tilde{\varepsilon}}{\|\mu_t^*\|^2}\right)
\end{aligned}$$

where the last inequality uses the fact that convex combinations of two values is less than the maximum of two values. □

Finally, we obtain the following bound on the correlation between the ground truth and the final iterate of gradient descent:

Lemma C.6. *For any $h \in \mathbb{N}$, let $\mu_t^{(h)}$ denote the iterate after h empirical gradient steps with learning rate $\eta = 1/20$ starting from random initialization, where the empirical gradients are estimated from at least $\Theta(\frac{d^4B^3}{\varepsilon^2})$ samples. Let $\theta^{(h)}$ denote the angle between $\mu_t^{(h)}$ and μ_t^*. For any $\varepsilon \lesssim \frac{1}{d^2B^9}$, there exists $H' \lesssim B^6\log d$ such that for any $H \ge H'$, if $\frac{1}{B^3} \le \|\mu_t^*\| \le \frac{1}{B^2}$, we have*

$$\tan\theta^{(H)} \lesssim 1\,.$$

Proof. Denote the h-th iterate of gradient descent by $\mu_t^{(h)}$. In Lemma C.7 we show that $\left\|\mu_t^{(h)}\right\| \le \frac{1}{B^2}$ for all h. We would like to apply the bound in Lemma C.5 to argue that the angle with μ_t^* decreases when going from $\mu_t^{(h)}$ to $\mu_t^{(h+1)}$. Using that $\frac{1}{B^3} \le \|\mu_t^*\| \le \frac{1}{B^2}$ and $\|\mu_t\| \le \frac{1}{B^2}$, we can bound the quantity κ_1 that appears in Lemma C.5 by

$$\begin{aligned}
\kappa_1 &\le \frac{1 - 3\eta\|\mu_t\|^2}{1 - 3\eta\|\mu_t\|^2 + \frac{\eta}{B^6}(1 - \frac{500\sqrt{d^3}}{B^2} - \frac{20d}{B^2} - \varepsilon dB^9)} \\
&\le \frac{1}{1 + \frac{\eta}{B^6}(1 - \frac{500\sqrt{d^3}}{B^2} - \frac{20d}{B^2} - \varepsilon dB^9)} \le \frac{1}{1 + \frac{\eta}{2B^6}}\,.
\end{aligned}$$

On the other hand, for B a sufficiently large polynomial in d, we can again use that $\frac{1}{B^3} \le \|\mu_t^*\| \le \frac{1}{B^2}$ and $\|\mu_t\| \le \frac{1}{B^2}$ to bound the quantity κ_2 that appears in Lemma C.5 by

$$\kappa_2 \le \frac{500\eta\sqrt{d^3}}{B^2} + \frac{20\eta d}{B^4} + B^9\eta d\varepsilon \lesssim \frac{\eta}{d}\,.$$

As $\left|\langle\hat{\mu}, \hat{\mu}^*\rangle\right| \ge \frac{1}{2d}$, this implies $|\tan\theta^{(h)}| \le 2d$. Without loss of generality assume that $\tan\theta^{(h)} \le 2d$.

By Lemma C.5, for any h we either have $\tan\theta^{(h)} \lesssim \eta/d \ll 1$, in which case we are done as this bound will also hold for subsequent iterates, or $\tan\theta^{(h)} \lesssim (1 + \frac{\eta}{2B^6})^{-1}\tan\theta^{(h-1)}$. If the latter happens consecutively for $H \ge \frac{\log d}{\log(1+\frac{\eta}{2B^6})}$ steps, then because $(1 + \frac{\eta}{2B^6})^{-H} = \frac{1}{d}$, the angle θ will satisfy $\tan\theta \le 2d \cdot (1/d) \lesssim 1$. The proof is complete because, by hypothesis, $H \ge \frac{4B^6\log d}{\eta} \ge \frac{\log d}{\log(1+\frac{\eta}{2B^6})}$ (the last inequality follows from $\log(1+x) \ge \frac{x}{2}$ for any $0 < x < 1$). □

Lemma C.7. *When parameter μ_t satisfies $\|\mu_t\| \le \frac{1}{B^2}$ for the noise scale $t = O(\log d)$ and μ_t' is the new parameter after performing a gradient descent update on the DDPM objective at noise scale $t = O(\log d)$, then parameter μ_t' satisfies $\|\mu_t'\| \le \frac{1}{B^2}$.*

Proof. When $\|\mu_t\| \le 0.9\|\mu_t^*\| \le \frac{0.9}{B^2}$, we have

$$\|\mu_t'\| \le \|\mu_t + \eta F(\mu_t, \mu_t^*)\| + \eta\|(-\nabla L_t(s_{\mu_t}) - F(\mu, \mu^*))\| + \eta\varepsilon \le (1 + 2\eta\|\mu_t^*\|^2)\|\mu_t\| + \frac{1}{dB^9}$$
$$\le 1.05\|\mu_t\| + \frac{1}{dB^9} \le \frac{1}{B^2}.$$

When $\|\mu_t\| \ge 0.9\|\mu_t^*\|$, then maximum eigenvalue of $F(\mu_t, \mu_t^*)$ is negative. Therefore, $\|\mu_t'\|$ is less than $\frac{1}{B^2}$. Specifically, we have

$$\|\mu_t'\| \le \|\mu_t + \eta F(\mu_t, \mu_t^*)\| + \eta\|(-\nabla L_t(s_{\mu_t}) - F(\mu, \mu^*))\| + \eta\varepsilon$$
$$\le (1 + \eta(2\|\mu_t^*\|^2 - 3\|\mu_t\|^2))\|\mu_t\| + \frac{1}{dB^9} \le (1 - 0.01\|\mu_t^*\|^2)\|\mu_t\| + \frac{1}{dB^9} \le \frac{1}{B^2}. \square$$

C.2 Low noise regime - connection to EM algorithm

In the previous section we showed how to obtain a warm start by running gradient descent on the DDPM objective at high noise. We now focus on proving the contraction of $\|\mu_t - \mu_t^*\|$ starting from this warm start, by running gradient descent at *low* noise. We first prove the contraction for population gradient descent and then, we argue that the empirical gradient descent concentrates well around the population gradient descent.

As before, we denote μ_t as the current iterate and μ_t' as the next iterate obtained by performing (population) gradient descent on the DDPM objective with step size η. We upper bound $\|\mu_t' - \mu_t^*\|$ as follows:

$$\begin{aligned}\|\mu_t' - \mu_t^*\| &= \|\mu_t - \eta\nabla_{\mu_t} L_t(s_{\mu_t}) - \mu_t^*\| \\ &= \Big\|(1-\eta)(\mu_t - \mu_t^*) + \eta\,\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,1)}\big[\big(\tanh(\mu_t^\top x) - \frac{1}{2}\tanh''(\mu_t^\top x)\|\mu_t\|^2 \\ &\quad + \tanh'(\mu_t^\top x)\mu_t^\top x\big)x\big] - \eta\,\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,1)}[\tanh'(\mu_t^\top x)\mu_t] - \eta\mu_t^*\Big\| \\ &\le (1-\eta)\,\|\mu_t - \mu_t^*\| + \eta\big\|\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,1)}[\tanh(\mu_t^\top x)x] - \mu_t^*\big\| + \eta\,\|G(\mu_t, \mu_t^*)\|,\end{aligned}$$

where

$$G(\mu_t, \mu_t^*) \triangleq \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\Big[-\frac{1}{2}\tanh''(\mu_t^\top x)\|\mu_t\|^2 x + (\tanh'(\mu_t^\top x)\mu_t^\top x)x - \tanh'(\mu_t^\top x)\mu_t\Big].$$

Recall that $\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,1)}[\tanh(\mu_t^\top x)x]$ is the EM update for mixtures of two Gaussians (See Fact 5). If we can show that the $G(\mu_t, \mu_t^*)$ term above is "contractive" in the sense that it is decreasing in $\|\mu_t - \mu_t^*\|$, then we can invoke existing results on convergence of EM to show that the distance between the current iterate and μ_t^* contracts in a single gradient step [DTZ17, XHM16]. Our goal is thus to control $G(\mu_t, \mu_t^*)$.

For this, we start with the 1D case in Lemma C.8. We then extend to the multi-dimensional case in Lemma C.9.

Lemma C.8 (One-dimensional version)**.** *Let $\mu, \mu^* > 0$, and consider $\mu \in [c, \frac{4\mu^*}{3}]$ for some constant c. In this one-dimensional case, the function G specializes to*

$$G(\mu, \mu^*) = \mathbb{E}_{x\sim\mathcal{N}(\mu^*,1)}\Big[-\frac{1}{2}\tanh''(\mu x)\mu^2 x + \tanh'(\mu x)\mu x^2 - \tanh'(\mu x)\mu\Big], \tag{C.2}$$

and we have

$$G(\mu, \mu^*) \le 0.01|\mu - \mu^*|$$

The proof uses the fact that the function G only contains first or higher-order derivatives of the $\tanh$ function and all the derivatives of $\tanh$ decay exponential quickly as μ increases. Therefore, when μ is at least a constant, we obtain the result. The complete proof of lemma C.8 is given in Appendix F.2.

Lemma C.9 (Multi-dimensional version)**.** *For any noise scale t, when the current parameter at noise scale t, μ_t, satisfies $\|\mu_t\| \in [c, \frac{4\langle\hat{\mu}_t, \mu_t^*\rangle}{3}]$ for some sufficiently large constant c, then the following inequality holds:*

$$\big\|G(\mu_t, \mu_t^*)\big\| \le 0.01\|\mu_t - \mu_t^*\|$$

Proof. Suppose $\{v_1, v_2, \ldots, v_d\}$ are d orthonormal directions such that $v_1 = \hat{\mu}_t$ and v_2 is either of the two unit vectors $\hat{\mu}_t^\perp$ which are orthogonal to $\hat{\mu}_t$ in the plane of μ_t and μ_t^*. Recall that

$$\begin{aligned}
G(\mu_t, \mu_t^*) &= \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\Big[-\frac{1}{2}\tanh''(\mu_t^\top x)\|\mu_t\|^2 x + (\tanh'(\mu_t^\top x)\mu_t^\top x)x - \tanh'(\mu_t^\top x)\mu_t\Big] \\
&= \mathbb{E}_{x\sim\mathcal{N}(0,I)}\Big[-\frac{1}{2}\tanh''(\mu_t^\top (x+\mu_t^*))\|\mu_t\|^2 (x+\mu_t^*) \\
&\qquad + \tanh'(\mu_t^\top (x+\mu_t^*))(\mu_t^\top (x+\mu_t^*))(x+\mu_t^*) - \tanh'(\mu_t^\top (x+\mu_t^*))\mu_t\Big] \\
&= \mathbb{E}_{\alpha_1,\alpha_2,\ldots,\alpha_d\sim\mathcal{N}(0,1)}\Big[-\frac{1}{2}\tanh''(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*))\|\mu_t\|^2 \big(\textstyle\sum_i \alpha_i v_i + \mu_t^*\big) \\
&\qquad + \tanh'(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*))\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*)\big(\textstyle\sum_i \alpha_i v_i + \mu_t^*\big) \\
&\qquad - \tanh'(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*))\mu_t\Big],
\end{aligned}$$

where in the last equality we rewrote $x \sim \mathcal{N}(0, I)$ as $\sum_{i=1}^d \alpha_i v_i$ for $\alpha_i \sim \mathcal{N}(0, 1)$. Therefore, we have

$$\begin{aligned}
&\langle \hat{\mu}_t, G(\mu_t, \mu_t^*)\rangle \\
&= \mathbb{E}_{\alpha_1,\alpha_2,\ldots,\alpha_d\sim\mathcal{N}(0,I)}\Big[-\frac{1}{2}\tanh''(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*))\|\mu_t\|^2 (\alpha_1 + \hat{\mu}_t{}^\top \mu_t^*) \\
&\qquad + \tanh'(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t{}^\top \mu_t^*))\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*)^2 - \tanh'(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t{}^\top \mu_t^*))\|\mu_t\|\Big] \\
&= \mathbb{E}_{\alpha_1\sim\mathcal{N}(\hat{\mu}_t^\top \mu_t^*,1)}\Big[-\frac{1}{2}\tanh''(\|\mu_t\|\alpha_1)\|\mu_t\|^2 \alpha_1 + \tanh'(\,\|\mu_t\|\,\alpha_1)\|\mu_t\|\,\alpha_1^2 - \tanh'(\,\|\mu_t\|\,\alpha_1)\|\mu_t\|\Big].
\end{aligned}$$

By taking $\|\mu_t\|$ to be μ and $\langle \hat{\mu}_t, \mu_t^*\rangle$ to be μ^*, we observe the similarity between the right side of the above equation and the one-dimensional definition of G defined in Eq. (C.2). Using Lemma C.8 and if $\|\mu_t\| \in [c, \frac{4\langle \hat{\mu}_t, \mu_t^*\rangle}{3}]$, we have

$$\langle \hat{\mu}_t, G(\mu_t, \mu_t^*)\rangle \le 0.01\big|\langle \hat{\mu}_t, \mu_t\rangle - \langle \hat{\mu}_t, \mu_t^*\rangle\big|$$

Taking the dot product of $G(\mu_t, \mu_t^*)$ with $v_2 = \hat{\mu}_t^\perp$, we have

$$\begin{aligned}
\langle \hat{\mu}_t^\perp, G(\mu_t, \mu_t^*)\rangle &= \mathbb{E}_{\alpha_1,\alpha_2,\ldots,\alpha_d\sim\mathcal{N}(0,1)}\Big[-\frac{1}{2}\tanh''(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*))\|\mu_t\|^2 (\alpha_2 + \langle \hat{\mu}_t^\perp, \mu_t^*\rangle) \\
&\qquad + \tanh'(\,\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*))\|\mu_t\|\,(\alpha_1 + \hat{\mu}_t^\top \mu_t^*)(\alpha_2 + \langle \hat{\mu}_t^\perp, \mu_t^*\rangle)\Big] \\
&= \mathbb{E}_{\alpha_1\sim\mathcal{N}(\hat{\mu}_t^\top \mu_t^*,1)}\Big[-\frac{1}{2}\tanh''(\,\|\mu_t\|\,\alpha_1)\|\mu_t\|^2 \langle \hat{\mu}_t^\perp, \mu_t^*\rangle \\
&\qquad + \tanh'(\,\|\mu_t\|\,\alpha_1)\|\mu_t\|\,\alpha_1 \langle \hat{\mu}_t^\perp, \mu_t^*\rangle\Big] \\
&= \langle \hat{\mu}_t^\perp, \mu_t^*\rangle\, \mathbb{E}_{\alpha_1\sim\mathcal{N}(\hat{\mu}_t^\top \mu_t^*,1)}\Big[-\frac{1}{2}\tanh''(\,\|\mu_t\|\,\alpha_1)\|\mu_t\|^2 + \tanh'(\,\|\mu_t\|\,\alpha_1)\|\mu_t\|\,\alpha_1\Big].
\end{aligned}$$

In Lemma F.5 below, we show that when $\|\mu_t\| \in [c, \frac{4\langle \hat{\mu}_t, \mu_t^*\rangle}{3}]$, the expectation in the last expression is upper bounded by 0.01. Therefore, we have

$$\big|\langle \hat{\mu}_t^\perp, G(\mu_t, \mu_t^*)\rangle\big| \le 0.01|\langle \hat{\mu}_t^\perp, \mu_t^*\rangle| \implies \big|\langle \hat{\mu}_t^\perp, G(\mu_t, \mu_t^*)\rangle\big| \le 0.01\big|\langle \hat{\mu}_t^\perp, \mu_t - \mu_t^*\rangle\big|$$

Observe that for $i = 3, \ldots, d$, $\langle G(\mu_t, \mu_t^*), v_i\rangle = 0$. Therefore, we have

$$\big\|G(\mu_t, \mu_t^*)\big\|^2 = \sum_{i=1}^d \langle v_i, G(\mu_t, \mu_t^*)\rangle^2 \le 0.01^2 \|\mu_t - \mu_t^*\|^2 .$$ □

The next Lemma ensures that the parameter μ_t after a few steps of gradient descent on the DDPM objective stays in the region where the function G satisfies $\big\|G(\mu_t, \mu_t^*)\big\| \le 0.01\|\mu_t - \mu_t^*\|$. Recall that the condition of the Lemma is satisfied because we initialize at the warm start obtained by gradient descent in the high noise regime.

Lemma C.10. *Suppose the angle between initialization $\hat{\mu}_t^{(0)}$ and optimal parameter μ_t^* is $\Theta(1)$, then for any h, we have $\|\mu_t^{(h)}\| \in [c, \frac{4\langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle}{3}]$.*

The proof of Lemma C.10 is given in Appendix F.3. Finally, we are ready to prove the main result of this section:

Proof of Theorem C.1. To obtain the contraction of $\|\mu_t^{(h)} - \mu_t^*\|$ after a gradient descent step on the DDPM objective, we write $\|\mu_t^{(h+1)} - \mu_t^*\|$ in terms of $\|\mu_t^{(h)} - \mu_t^*\|$ as follows:

$$\left\|\mu_t^{(h+1)} - \mu_t^*\right\| = \left\|\mu_t^{(h)} - \eta \nabla L_t(s_{\mu_t^{(h)}}) - \mu_t^*\right\| + \eta \left\| \left(\frac{1}{n}\sum_{i=1}^n \nabla L_t(s_{\mu_t^{(h)}}(x_i, z_i))\right) - \nabla L_t(s_{\mu_t^{(h)}}) \right\|$$

$$\leq (1-\eta)\|\mu_t^{(h)} - \mu_t^*\| + \eta \left\|\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,1)}[(\tanh({\mu_t^{(h)}}^\top x))x] - \mu_t^*\right\| + \eta\|G(\mu_t^{(h)}, \mu_t^*)\| + \eta\varepsilon\,,$$

where in the last step we used Lemma E.7 below to bound the distance between the population and empirical gradient.

Recall that gradient descent in the low noise regime was initialized using the output of the gradient descent in the high noise regime. Therefore, $\langle \hat{\mu}_t^{(0)}, \hat{\mu}_t^* \rangle \gtrsim 1$. Using Lemma C.10, we know that the condition on Lemma C.8 is always satisfied. Using the contractivity of G established in Lemma C.8 combined with [DTZ17, Theorem 2], and choosing $\eta = 0.05$, we conclude that the distance to the ground truth contracts:

$$\begin{aligned}\left\|\mu_t^{(h+1)} - \mu_t^*\right\| &\leq (1-0.05)\left\|\mu_t^{(h)} - \mu_t^*\right\| + 0.01\left\|\mu_t^{(h)} - \mu_t^*\right\| + 0.01\left\|\mu_t^{(h)} - \mu_t^*\right\| + \eta\varepsilon \\ &\leq 0.97\left\|\mu_t^{(h)} - \mu_t^*\right\| + \eta\varepsilon.\end{aligned}$$

Applying the above for all $h \in [H]$, we obtain

$$\|\mu_t^{(H)} - \mu_t^*\| \leq 0.97^H \|\mu_t^{(0)} - \mu_t^*\| + 50\varepsilon.$$

The choice of H given in the Theorem statement proves the result. □

D Learning mixtures of two Gaussians with small separation

In this section, we extend the analysis for learning mixtures of two Gaussians with constant separation, provided in Section C, to the low-separation regime and prove the following:

Theorem D.1 (Formal version of Theorem 13)**.** *For any $\mathcal{L} > 0$, let q be a mixture of two Gaussians (in the form of Eq. (7)) with mean parameter μ^* satisfying $\|\mu^*\| > \mathcal{L}$. Recalling that B denotes an* a priori *upper bound on $\|\mu^*\|$, we have that for any $\varepsilon \leq \varepsilon'$, where $\varepsilon' \lesssim \frac{1}{d^2 B^9}$, there exists a procedure satisfying the following. If the procedure is run for at least $\mathrm{poly}(d, B, \frac{1}{\mathcal{L}})\frac{1}{\varepsilon^3}$ iterations with at least $\mathrm{poly}(d, B, \frac{1}{\mathcal{L}}) * \frac{1}{\varepsilon^8}$ samples from q, then it outputs $\tilde{\mu}$ such that $\|\tilde{\mu} - \mu^*\| \leq \varepsilon$ with high probability.*

As described in Section 1.2, the algorithm is a simple modification of Algorithm 1 in which gradient descent is replaced by projected gradient descent. We start in Lemma D.2 by showing that the projection step in the algorithm ensures that the norm of the current iterate μ_t is approximately that of μ_t^*. Then in Lemma D.3, we extend the analysis of Lemma C.5 to show that every projected gradient step contracts the distance to the ground truth. Combined with Lemma D.2, this allows us to conclude the proof of Theorem 13.

Lemma D.2. *Let $x_1, \ldots, x_n$ be independent samples from q, and define radius parameter R by $R^2 \triangleq \frac{1}{n}\sum_{i=1}^n \|x_i\|^2 - d$. For any $\varepsilon > 0$, provided that $n \gtrsim \frac{B^4 + d^2}{\varepsilon^2 \mathcal{L}^2}$, we have $|R - \|\mu^*\|| \leq \varepsilon$ with high probability.*

Proof. Observe that we can write the random variable corresponding to the mixture of two Gaussians $X_0 = X = Z + p\mu^*$ where $Z \sim \mathcal{N}(0, I)$ and p is a Rademacher random variable. Using Theorem 3.1.1 (concentration of norms) from [Ver], we know that $\|\|Z\| - \sqrt{d}\|_{\psi_2} \lesssim 1$. Therefore, sub-Gaussian norm $\left\|\|X_0\|\right\|_{\psi_2} \lesssim \left\|\|Z\|\right\|_{\psi_2} + \left\|\|p\mu^*\|\right\|_{\psi_2} \lesssim B + \sqrt{d}$. Using Lemma 2.7.4 from [Ver],

we have $\left\|\|X_0\|^2\right\|_{\psi_1} \lesssim \left\|\|X_0\|\right\|_{\psi_2}^2 \lesssim B^2 + d$. Therefore, using number of samples n specified in the Lemma statement, with high probability, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2 - \mathbb{E}[\|X_0\|^2]\right| \le \varepsilon\mathcal{L} \implies \left|\|\mu\|^2 - \|\mu^*\|^2\right| \le \varepsilon\mathcal{L} \implies \left|\|\mu\| - \|\mu^*\|\right| \le \varepsilon$$

where the penultimate implication uses the fact that $\mathbb{E}_{X_0}[\|X_0\|^2] = \mathbb{E}[\|Z\|^2 + \|\mu^*\|^2] = d + \|\mu^*\|^2$. □

Lemma D.3. *Assume that $\mathcal{L} \le \|\mu^*\| \le B$. Then, for any small $\varepsilon > 0$, running projected GD on diffusion models with step size $\eta = \frac{1}{20}$ at noise scale $t = \log\frac{d}{\varepsilon}$ for number of steps $H > H'$ and number of samples $n > n'$ steps will achieve*

$$\left\|\mu^{(H)} - \mu^*\right\| \lesssim d^2B^4\varepsilon,$$

where $H' = \frac{d^2}{\mathcal{L}^2\varepsilon^3}$ and $n' = \frac{d^{10}B^3}{\varepsilon^8\mathcal{L}^6}$.

Proof. Recalling that $\mu_t^* = \mu_0^*\exp(-t)$, note that for $t = \log\frac{d}{\varepsilon}$, $\frac{\varepsilon\mathcal{L}}{d} \le \|\mu_t^*\| \le \frac{\varepsilon B}{d}$. We would like to apply Lemma C.5. Note that we may apply this even though it is only stated for gradient descent (without projection). The reason is that it bounds the change in angle between the iterate and the ground truth after a single gradient step, and this angle is unaffected by projection.

Suppose we take one projected gradient step with learning rate η starting from an iterate μ_t. As μ_t was the result of a projection, by Lemma D.2 we have $\frac{\varepsilon\mathcal{L}}{d} \lesssim \left\|\mu_t^{(h)}\right\| \lesssim \frac{\varepsilon B}{d}$.

We now bound κ_2 in Lemma C.5:

$$\begin{aligned}
\kappa_2 &= \frac{500\eta\sqrt{d^3}\|\mu_t\|^4 + 20\eta d\|\mu_t\|^2\|\mu_t^*\|^2 + \eta\tilde{\varepsilon}}{\|\mu_t^*\|^2} \\
&\lesssim 500\eta\sqrt{d^7}\|\mu_t\|^2 + 20\eta d\|\mu_t\|^2 + \frac{d^2\varepsilon}{\|\mu_t^*\|^3} \\
&\le 550d^{7/2}B^2\exp(-2t) + \frac{d^5\varepsilon}{\varepsilon^3\mathcal{L}^3} \\
&\lesssim d^2B^2\varepsilon,
\end{aligned}$$

where the last inequality follows by choosing population gradient estimation error parameter $\varepsilon = \frac{\varepsilon^4\mathcal{L}^3}{d^3}$ with the number of samples $n' = \frac{d^{11}B^6}{\varepsilon^8\mathcal{L}^6}$. Additionally, κ_1 in Lemma C.5 is given by

$$\begin{aligned}
\kappa_1 &= \frac{1 - 3\eta\|\mu_t\|^2}{(1 - 3\eta\|\mu_t\|^2) + \eta(\|\mu_t^*\|^2 - 500\sqrt{d^3}\|\mu_t\|^4 - 20d\|\mu_t\|^2\|\mu_t^*\|^2 - \tilde{\varepsilon})} \\
&= \frac{1 - 3\eta\|\mu_t\|^2}{(1 - 3\eta\|\mu_t\|^2) + \eta\|\mu_t^*\|^2(1 - \kappa_2)} \\
&\lesssim \frac{1 - 3\eta\|\mu_t^{(h)}\|^2}{(1 - 3\eta\|\mu_t^{(h)}\|^2) + \eta\|\mu_t^*\|^2(1 - d^2B^2\varepsilon)} \\
&\le \frac{1}{1 + \frac{\mathcal{L}^2\varepsilon^2}{20d^2}(1 - d^2B^2\varepsilon)}.
\end{aligned}$$

Using bounds on κ_1 and κ_2 and Lemma C.5, we conclude that if θ (resp. θ') is the angle between μ_t (resp. the next iterate of projected gradient descent after μ_t) and μ_t^*

$$\tan\theta' \le \max\left(\frac{1}{1 + \frac{\mathcal{L}^2\varepsilon^2}{20d^2}(1 - B^2\varepsilon)}\tan\theta, d^2B^2\varepsilon\right).$$

Doing projected gradient descent for $H = \frac{20d^2}{\mathcal{L}^2\varepsilon^3}$ steps, if $\theta^{(h)}$ denotes the angle between the h-th iterate and μ_t^*, we obtain

$$\tan\theta^{(H)} \le \tan\theta^{(h+1)} \le \max\left(\left(\frac{1}{1+\frac{\mathcal{L}^2\varepsilon^2}{20d^2}(1-d^2B^2\varepsilon)}\right)^H \tan\theta^{(0)}, d^2B^2\varepsilon\right)$$
$$\le \max\left(\frac{\tan\theta^{(0)}}{1+\frac{H\mathcal{L}^2\varepsilon^2}{20d^2}(1-B^2\varepsilon)}, d^2B^2\varepsilon\right) \le d^2B^2\varepsilon\,,$$

where the last inequality uses $1+\frac{H\mathcal{L}^2\varepsilon^2}{20d^2}(1-B^2\varepsilon) \ge \frac{1}{\varepsilon}$ for $\varepsilon \lesssim \frac{1}{B^3}$. Additionally, for a random initialization, Lemma C.4 shows that $\cos\theta^{(0)} \ge \frac{1}{2d}$ which implies $\tan\theta^{(0)} \le \sqrt{\sec^2\theta^{(0)}-1} \lesssim d$. Using Lemma D.2, we have $\|\mu^{(H)}\| \ge \|\mu^*\| - \varepsilon$ which implies $-2\|\mu^{(H)}\|\|\mu^*\|\cos\theta^{(H)} \le -2\|\mu^*\|^2\cos\theta^{(H)} + 2B\varepsilon$ and $\|\mu^{(H)}\|^2 \le \|\mu^*\|^2 + 3B\varepsilon$. Using this result, we obtain

$$\|\mu^{(H)}-\mu^*\|^2 = \|\mu^{(H)}\|^2 + \|\mu^*\|^2 - 2\|\mu^{(H)}\|\|\mu^*\|\cos\theta^{(H)}$$
$$\lesssim 2\|\mu^*\|^2 - 2\|\mu^*\|^2\cos\theta^{(H)} + 5B\varepsilon \lesssim 2B^2\Big(1-\frac{1}{\sqrt{1+d^4B^4\varepsilon^2}}\Big) + 5B\varepsilon \lesssim d^2B^4\varepsilon,$$

where the last inequality follows from the fact that $\sqrt{1+x} \le 1+\sqrt{x}$ for any $x > 0$. □

E Learning mixtures of K Gaussians from a warm start

In this section, we provide details about our main result on learning mixtures of K Gaussians. We start by describing our main theorem in this case.

Theorem E.1 (Formal version of Theorem 16)**.** *Let q be a mixture of Gaussians (in the form of Eq. (6)) with center parameters $\theta^* = \{\mu_1^*, \mu_2^*, \dots, \mu_K^*\} \in \mathbb{R}^d$ satisfying the separation Assumption 14, and suppose we have estimates θ for the centers such that the warm initialization Assumption 15 is satisfied. For any $\varepsilon > \varepsilon_0$ and noise scale t where*

$$\varepsilon_0 = 1/\mathrm{poly}(d) \quad \textit{and} \quad t = \Theta(\varepsilon)\,,$$

gradient descent on the DDPM objective at noise scale t' (Algorithm 1) outputs $\tilde\theta = \{\tilde\mu_1, \tilde\mu_2, \dots, \tilde\mu_K\}$ such that $\min_i \|\tilde\mu_i - \mu_i^*\| \le \varepsilon$ with high probability. The algorithm runs for $H \ge H'$ iterations and uses $n \ge n'$ number of samples where

$$H' = \Theta(\log(\varepsilon^{-1}\log d)) \quad \textit{and} \quad n' = \Theta(K^4d^5B^6/\varepsilon^2)\,.$$

We first give an overview of the proof for population gradient descent, and then show that the empirical gradients concentrate well around the population gradients. We start by simplifying the population gradient update for mixtures of K Gaussians using Stein's lemma in Lemma E.2, which yields

$$-\nabla_{\mu_{1,t}} L_t(s_{\theta_t}) = \mathbb{E}[w_{1,t}(X_t)(X_t-\mu_{1,t})] + [\text{extra terms}]\,,$$

recalling the notation of Eq. (9). As discussed in the body of the paper, $\mathbb{E}[w_{1,t}(X_t)(X_t-\mu_{1,t})]$ is precisely the update for the gradient EM algorithm (see Fact 6) and known results for the latter [KC20, SN21] can be used to show that the distance $\|\mu_{1,t}-\mu_{1,t}^*\|$ contracts in each step when the separation Assumption 14 and the warm initialization Assumption 15 are satisfied. Therefore, showing that the "extra terms" do not disturb the progress coming from the gradient EM update is sufficient. We prove that the "extra terms" are $1/\mathrm{poly}(d)$ in Lemma E.4 when the separation Assumption 14 and warm initialization Assumption 15 hold.

The intuition behind Lemma E.4 is as follows: We start with a key observation that each of the "extra terms" either contains $w_{1,t}(X_t)(1-w_{1,t}(X_t))$ or $w_{1,t}(X_t)w_{j,t}(X_t)$ where $j \ne 1$. Note that the $w_{1,t}(X_t)$ can be interpreted as the conditional probability of the underlying component being $\mathcal{N}(\mu_{1,t}, I)$ given X_t. When Assumption 14 and Assumption 15 are satisfied, Proposition 4.1 of [SN21] shows that

$$\mathbb{E}_{X_t\sim\mathcal{N}(\mu_{1,t}^*,I)}[w_{j,t}(X_t)] \lesssim 1/\mathrm{poly}(d) \quad \text{for any } j \ne 1\,.$$

This result can be extended to show both $\mathbb{E}_{X_t}[w_{1,t}(X_t)(1-w_{1,t}(X_t))] \lesssim 1/\mathrm{poly}(d)$ as well as $\mathbb{E}_{X_t}[w_{1,t}(X_t)w_{j,t}(X_t)] \lesssim 1/\mathrm{poly}(d)$ for any $j \ne 1$ (see Lemma E.5 for the proof). Using these bounds, we conclude that ["extra terms"] $\lesssim 1/\mathrm{poly}(d)$ in Lemma E.4.

E.1 EM and population gradient descent on DDPM objective

We begin by writing out the gradient update explicitly:

Lemma E.2. *For any noise scale $t > 0$, the gradient of the population DDPM objective $\mathbb{E}[L_t(s_{\theta_t}(X_t))]$ with respect to parameter $\mu_{1,t}$ is given by*

$$\nabla_{\mu_{1,t}} L_t(s_{\theta_t}) = \mathbb{E}\Big[- w_{1,t}(X_t)(X_t - \mu_{1,t}) + w_{1,t}(X_t)(X_t - \mu_{1,t}) \sum_{i=1}^{K} w_{i,t}(X_t)\mu_{i,t}^\top (X_t - \mu_{1,t}) + w_{1,t}(X_t)\mu_{1,t} - w_{1,t}(X_t)(X_t - \mu_{1,t})^\top \mu_{1,t}(X_t - \mu_{1,t}) - w_{1,t}(X_t)\sum_{i=1}^{K} w_{i,t}(X_t)\mu_{i,t} - w_{1,t}(X_t)\sum_{i=1}^{K} \nabla_x w_{i,t}(X_t)^\top \mu_{i,t}(X_t - \mu_{1,t})\Big]$$

where $w_{1,t}(x)$ and $\mu_{1,t}$ are defined in Eq. (9).

Proof. Recall that the score function of mixture of Gaussians is given by

$$s_{\theta_t}(X_t) = \sum_i w_{i,t}(X_t)\mu_{i,t} - X_t$$

Finding the gradient $\nabla_{\mu_{1,t}} w_{i,t}(X_t)$, we have

$$\nabla_{\mu_{1,t}} w_{i,t}(X_t) = \begin{cases} w_{1,t}(X_t)(1 - w_{1,t}(X_t))(X_t - \mu_{1,t}) & \text{if } i = 1 \\ -w_{1,t}(X_t) w_{i,t}(X_t)(X_t - \mu_{1,t}) & \text{otherwise.} \end{cases}$$

The gradient of the score function is given by

$$\nabla_{\mu_{1,t}} s_{\theta_t}(X_t) = \nabla_{\mu_{1,t}} \left(w_{1,t}(X_t)\mu_{1,t}\right) + \sum_{i=2}^{K} \nabla_{\mu_{1,t}} \left(w_{i,t}(X_t)\mu_{i,t}\right)$$

$$= w_{1,t}(X_t)(1 - w_{1,t}(X_t))\mu_{1,t}(X_t - \mu_{1,t})^\top + w_{1,t}(X_t) I - w_{1,t}(X_t)\sum_{i=2}^{K} w_{i,t}(X_t)\mu_{i,t}(X_t - \mu_{1,t})^\top$$

$$= w_{1,t}(X_t)\mu_{1,t}(X_t - \mu_{1,t})^\top + w_{1,t}(X_t) I - w_{1,t}(X_t)\sum_{i=1}^{K} w_{i,t}(X_t)\mu_{i,t}(X_t - \mu_{1,t})^\top .$$

The gradient of $\frac{1}{2}\|s_{\theta_t}\|^2$ is given by

$$\frac{1}{2}\nabla \left\| s_{\theta_t}(X_t) \right\|^2 = \sum_{j=1}^{d} [s_{\theta_t}(X_t)]_j [\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)]_j = \nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top s_{\theta_t}(X_t)$$

where $[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)]_j$ is j^{th} row of $\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)$.

The gradient of this is given by

$$\frac{\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t}{\beta_t} = \frac{1}{\beta_t}\Big(w_{1,t}(X_t)(X_t - \mu_{1,t})\mu_{1,t}^\top Z_t + w_{1,t}(X_t) Z_t - w_{1,t}(X_t)\sum_{i=1}^{K} w_{i,t}(X_t)(X_t - \mu_{1,t})\mu_{i,t}^\top Z_t \Big) \tag{E.1}$$

Applying Stein's lemma to the expectation of the first term in Eq. (E.1), we have

$$\begin{aligned} \mathbb{E}_{X_0, Z_t}[w_{1,t}(X_t)(X_t - \mu_{1,t})\mu_{1,t}^\top Z_t] &= \sum_{j=1}^{d} \mathbb{E}_{X_0, Z_t}[w_{1,t}(X_t)(X_t - \mu_{1,t})\mu_{1,t,j} Z_{t,j}] \\ &= \sum_{j=1}^{d} \mathbb{E}_{X_0, Z_t}[w_{1,t}(X_t)\beta_t e_j \mu_{1,t,j} + \beta_t \nabla_x w_{1,t}(X_t)^\top e_j (X_t - \mu_{1,t})\mu_{1,t,j}] \\ &= \mathbb{E}_{X_0, Z_t}[w_{1,t}(X_t)\beta_t \mu_{1,t} + \beta_t \nabla_x w_{1,t}(X_t)^\top \mu_{1,t}(X_t - \mu_{1,t})] \end{aligned}$$

The expectation of the second term in Eq. (E.1) simplifies to $\beta_t \mathbb{E}_{X_t}[\nabla_x w_{1,t}(X_t)]$ by Stein's Lemma. Each summand in the third term in Eq. (E.1) simplifies as following:

$$\begin{aligned}
&\mathbb{E}_{X_0,Z_t}\left[w_{1,t}(X_t)w_{i,t}(X_t)(X_t-\mu_{1,t})\mu_{i,t}^\top Z_t\right] \\
&\quad=\sum_{j=1}^{d}\mathbb{E}_{X_0,Z_t}\left[w_{1,t}(X_t)w_{i,t}(X_t)(X_t-\mu_{1,t})\mu_{i,t,j}Z_{t,j}\right] \\
&\quad=\sum_{j}\mu_{i,t,j}\mathbb{E}_{X_0,Z_t}\Big[w_{1,t}(X_t)w_{i,t}(X_t)\beta_t e_j+\beta_t w_{1,t}(X_t)\nabla_x w_{i,t}(X_t)^\top e_j(X_t-\mu_{1,t}) \\
&\qquad\qquad+\beta_t\nabla_x w_{1,t}(X_t)^\top e_j w_{i,t}(X_t)(X_t-\mu_{1,t})\Big] \\
&\quad=\beta_t\,\mathbb{E}_{X_0,Z_t}\Big[w_{1,t}(X_t)w_{i,t}(X_t)\mu_{i,t}+w_{1,t}(X_t)\nabla_x w_{i,t}(X_t)^\top\mu_{i,t}(X_t-\mu_{1,t}) \\
&\qquad\qquad+\nabla_x w_{1,t}(X_t)^\top\mu_{i,t}w_{i,t}(X_t)(X_t-\mu_{1,t})\Big] \qquad\qquad \text{(E.2)}
\end{aligned}$$

Combining the gradients of all the terms of Eq. (E.2), we have

$$\begin{aligned}
&\nabla_{\mu_{1,t}}L_t(s_{\theta_t}) \\
&\quad=\mathbb{E}\Big[w_{1,t}(X_t)(X_t-\mu_{1,t})\mu_{1,t}^\top s_{\theta_t}(X_t)+w_{1,t}(X_t)s_{\theta_t}(X_t)-w_{1,t}(X_t)(X_t-\mu_{1,t})\sum_i w_{i,t}(X_t)\mu_{i,t}^\top s_{\theta_t}(X_t) \\
&\qquad+\nabla_x w_{1,t}(X_t)+w_{1,t}(X_t)\mu_{1,t}+\nabla_x w_{1,t}(X_t)^\top\mu_{1,t}(X_t-\mu_{1,t})-w_{1,t}(X_t)\sum_i w_{i,t}(X_t)\mu_{i,t} \\
&\qquad-w_{1,t}(X_t)\sum_i\nabla_x w_{i,t}(X_t)^\top\mu_{i,t}(X_t-\mu_{1,t})-\sum_i\nabla_x w_{1,t}(X_t)^\top\mu_{i,t}w_{i,t}(X_t)(X_t-\mu_{1,t})\Big] \\
&\quad=\mathbb{E}\Big[-w_{1,t}(X_t)(X_t-\mu_{1,t})+w_{1,t}(X_t)(X_t-\mu_{1,t})\sum_i w_{i,t}(X_t)\mu_{i,t}^\top(X_t-\mu_{1,t}) \\
&\qquad+w_{1,t}(X_t)\mu_{1,t}-w_{1,t}(X_t)(X_t-\mu_{1,t})^\top\mu_{1,t}(X_t-\mu_{1,t})-w_{1,t}(X_t)\sum_i w_{i,t}(X_t)\mu_{i,t} \\
&\qquad-w_{1,t}(X_t)\sum_i\nabla_x w_{i,t}(X_t)^\top\mu_{i,t}(X_t-\mu_{1,t})\Big],
\end{aligned}$$

where the last equality uses Lemma E.3. Specifically, it uses

$$\begin{aligned}
\nabla_x w_{1,t}(X_t)+w_{1,t}(X_t)s_{\theta_t}(X_t)&=-w_{1,t}(X_t)(X_t-\mu_{1,t}) \\
(\nabla_x w_{1,t}(X_t)+w_{1,t}(X_t)s_{\theta_t}(X_t))^\top\mu_{1,t}(X_t-\mu_{1,t})&=-w_{1,t}(X_t)(X_t-\mu_{1,t})^\top\mu_{1,t}(X_t-\mu_{1,t})\,.\ \square
\end{aligned}$$

We will also need the following intermediate calculation:

Lemma E.3. *For any $i\in[K]$, the gradient of $w_{i,t}(X_t)$ with respect to X_t is given by*

$$\begin{aligned}
\nabla_x w_{i,t}(X_t)&=-w_{i,t}(X_t)(X_t-\mu_{i,t})-w_{i,t}(X_t)s_{\theta_t}(X_t) \\
&=-w_{i,t}(X_t)(1-w_{i,t}(X_t))(X_t-\mu_{i,t})+w_{i,t}(X_t)\cdot\sum_{j\in[K]:j\neq i}w_{j,t}(X_t)(X_t-\mu_{j,t})\,.
\end{aligned}$$

Proof. By taking the gradient of $w_{i,t}(X_t)$ and simplifying it, we get the result:

$$\begin{aligned}
\nabla_x w_{i,t}(X_t) &= -\frac{\exp\Big(-\frac{\|X_t-\mu_{i,t}\|^2}{2}\Big)(X_t-\mu_{i,t})}{\sum_{j=1}^K \exp\Big(-\frac{\|X_t-\mu_{j,t}\|^2}{2\sigma^2}\Big)} \\
&\qquad + \frac{\exp\big(-\frac{\|X_t-\mu_{i,t}\|^2}{2}\big)\cdot\sum_{j=1}^K \exp\Big(-\frac{\|X_t-\mu_{j,t}\|^2}{2}\Big)(X_t-\mu_{j,t})}{\Big(\sum_{j=1}^K \exp\Big(-\frac{\|X_t-\mu_{j,t}\|^2}{2}\Big)\Big)^2} \\
&= -w_{i,t}(X_t)(X_t-\mu_{i,t}) + w_{i,t}(X_t)\left(\sum_{j=1}^K w_{j,t}(X_t)(X_t-\mu_{j,t})\right) \\
&= -w_{i,t}(X_t)(1-w_{i,t}(X_t))(X_t-\mu_{i,t}) + w_{i,t}(X_t)\left(\sum_{j=1,j\neq i}^K w_{j,t}(X_t)(X_t-\mu_{j,t})\right). \square
\end{aligned}$$

We are now ready to establish the connection between gradient descent on the DDPM objective and the gradient EM update, for mixtures of K Gaussians:

Lemma E.4. *Suppose the centers of the mixture of K Gaussians are well-separated according to Assumption 14, and the parameters $\theta = \{\mu_1, \mu_2, \ldots, \mu_K\}$ that the student network is initialized to satisfy the warm start Assumption 15. Then, for noise scale $t = O(1)$, gradient descent on the DDPM objective is close to the gradient EM update:*

$$\big\|\nabla_{\mu_{1,t}} L_t(s_{\theta_t}) + \mathbb{E}[w_{1,t}(X_t)(X_t-\mu_{1,t})]\big\| \lesssim \frac{K^2B^2}{d^{c_r^2/4000}} = \frac{1}{\mathrm{poly}(d)},$$

where c_r is a large constant.

Proof. Observe that the first term in the expression for the population gradient of the DDPM objective in Lemma E.2 is exactly the gradient EM update for the mixture of K Gaussian in Fact 6. To prove the closeness between the GD update and the gradient EM update, we will show that the additional terms in Lemma E.2 are small.

Note that when the ground truth parameters $\theta^* = \{\mu_1^*, \mu_2^*, \ldots, \mu_K^*\}$ satisfy Assumption 14, θ_t^* also satisfies Assumption 14 for $t = O(1)$. Similarly, it is straightforward to show that when the parameters θ satisfy Assumption 15, $\theta_t = \{\mu_{1,t}, \mu_{2,t}, \ldots, \mu_{K,t}\}$ also satisfies the assumption.

We focus on the $d \le K$ case for this proof. A similar calculation with projection onto $O(K)$ dimensional subspace of $\mu_{i,t}^*$ will give the result for $d \ge K$ case [VW04, YYS17].

Using Lemma E.6 below, we have

$$\big\|\mathbb{E}\big[w_{1,t}(X_t)(1-w_{1,t}(X_t))(X_t-\mu_{1,t})(X_t-\mu_{1,t})^\top\big]\mu_{1,t}\big\| \le \frac{d^2c_r^2B}{d^{c_r^2/1000}},$$

for any $i \in [K]$. We can simplify additional terms as

$$\begin{aligned}
&\Big\|\sum_{i=2}^K \mathbb{E}[w_{1,t}(X_t)w_{i,t}(X_t)(X_t-\mu_{1,t})(X_t-\mu_{1,t})^\top\mu_{i,t}]\Big\| \\
&\quad\le \sum_{i=2}^K \mathbb{E}[\|w_{1,t}(X_t)w_{i,t}(X_t)(X_t-\mu_{1,t})(X_t-\mu_{1,t})^\top\mu_{i,t}\|] \\
&\quad\le \sum_{i=2}^K \sqrt{\mathbb{E}\big[|w_{1,t}(X_t)w_{i,t}(X_t)|^2\big]\cdot\mathbb{E}\big[\|(X_t-\mu_{1,t})(X_t-\mu_{1,t})^\top\mu_{i,t}\|^2\big]} \\
&\quad\le \frac{KB^2}{d^{c_r^2/2000}},
\end{aligned}$$

where in the last step we used the second part of Lemma E.5. This will allow us to prove that $\|\mathbb{E}[w_{1,t}(X_t)(X_t-\mu_{1,t})\sum_{i=1}^K w_{i,t}(X_t)\mu_{i,t}^\top(X_t-\mu_{1,t}) - w_{1,t}(X_t)(X_t-\mu_{1,t})^\top\mu_{1,t}(X_t-\mu_{1,t})]\|$ is small.

Using the expression for $\nabla_x w_{i,t}(X_t)$ from Lemma E.3, we have

$$\begin{aligned}
&\sum_{i=1}^K w_{1,t}(X_t)\nabla_x w_{i,t}(X_t)^\top\mu_{i,t}(X_t-\mu_{1,t}) \\
&\quad= -\sum_{i=1}^K w_{1,t}(X_t)w_{i,t}(X_t)(1-w_{i,t}(X_t))(X_t-\mu_{1,t})(X_t-\mu_{i,t})^\top\mu_{i,t} \\
&\qquad+\sum_{i=1}^K\sum_{j=1,j\neq i}^K w_{1,t}(X_t)w_{i,t}(X_t)w_{j,t}(X_t)(X_t-\mu_{1,t})(X_t-\mu_{j,t})^\top\mu_{i,t}\,.
\end{aligned}$$

The first term can be simplified as follows:

$$\begin{aligned}
&\Big\|\sum_{i=1}^K\mathbb{E}\Big[w_{1,t}(X_t)w_{i,t}(X_t)(1-w_{i,t}(X_t))(X_t-\mu_{1,t})(X_t-\mu_{i,t})^\top\mu_{i,t}\Big]\Big\| \\
&\quad\le\sum_{i=1}^K\mathbb{E}\big[\big\|w_{1,t}(X_t)w_{i,t}(X_t)(1-w_{i,t}(X_t))(X_t-\mu_{1,t})(X_t-\mu_{i,t})^\top\mu_{i,t}\big\|\big] \\
&\quad\le\sum_{i=2}^K\sqrt{\mathbb{E}[w_{1,t}(X_t)^2w_{i,t}(X_t)^2]\cdot\mathbb{E}\big[(1-w_{i,t}(X_t))^2\cdot\|X_t-\mu_{1,t}\|^2\cdot\|X_t-\mu_{i,t}\|^2\cdot\|\mu_{i,t}\|^2\big]} \\
&\quad\lesssim\frac{KB^2}{d^{c_r^2/4000}}\,,
\end{aligned}$$

where the last inequality follows from

$$\mathbb{E}\big[\|X_t-\mu_{1,t}\|^2\|X_t-\mu_{i,t}\|^2\big]\le\sqrt{\mathbb{E}\big[\|X_t-\mu_{1,t}\|^4\big]\mathbb{E}\big[\|X_t-\mu_{i,t}\|^4\big]}\lesssim B^2\,.$$

Similarly, by simplifying the second term, we get

$$\begin{aligned}
&\sum_{i=1}^K\sum_{j=1,j\neq i}^K\mathbb{E}\big[\big\|w_{1,t}(X_t)w_{i,t}(X_t)w_{j,t}(X_t)(X_t-\mu_{1,t})(X_t-\mu_{j,t})^\top\mu_{i,t}\big\|\big] \\
&\quad\le\sum_{i=1}^K\sum_{j=1,j\neq i}^K\sqrt{\mathbb{E}\big[w_{i,t}^2(X_t)w_{j,t}^2(X_t)\big]\mathbb{E}\big[w_{1,t}^2(X_t)\|(X_t-\mu_{1,t})(X_t-\mu_{j,t})\mu_{i,t}\|^2\big]}\lesssim\frac{K^2B^2}{d^{c_r^2/4000}}\,,
\end{aligned}$$

where the last inequality uses Lemma E.5. Simplifying the following term using Lemma E.5, we have

$$\begin{aligned}
&\Big\|\mathbb{E}[w_{1,t}(X_t)\mu_{1,t}-w_{1,t}(X_t)\sum_{i=1}^K w_{i,t}(X_t)\mu_{i,t}]\Big\| \\
&\quad\le\sum_{i=2}^K\mathbb{E}\big[\big\|w_{1,t}(X_t)w_{i,t}(X_t)\mu_{i,t}\big\|\big]+\sum_{i=2}^K\mathbb{E}\big[\big\|w_{1,t}(X_t)w_{i,t}(X_t)\mu_{1,t}\big\|\big]\le\frac{2KB}{d^{c_r^2/200}}\,.
\end{aligned}$$

Combining all the results, we obtain the theorem statement. □

The above proof made use of the following two helper lemmas which follow from prior work analyzing EM for learning mixtures of Gaussians:

Lemma E.5. *There is some absolute constant $c_r>0$ for which the following holds. For any $\theta=\{\mu_1,\mu_2,\ldots,\mu_K\}$ such that $\|\mu_i-\mu_i^*\|\le\frac{c_r}{4}\sqrt{\log d}$ for all $i\in[K]$ and any j such that $j\neq i$, we have*

$$\mathbb{E}_{X_t\sim\mathcal{N}(\mu_{i,t}^*,I)}[w_{j,t}(X_t)]\le\frac{1}{d^{c_r^2/100}}\,.$$

Additionally, for any $j\neq k$ such that $j\in[K]$ and $k\in[K]$, we have

$$\mathbb{E}_{X_t}[w_{j,t}(X_t)w_{k,t}(X_t)]\le\frac{1}{d^{c_r^2/200}}\,.$$

Proof. Using Proposition 4.1 from [SN21], for any $\theta = \{\mu_1, \mu_2, \ldots, \mu_K\}$ such that $\|\mu_i - \mu_i^*\| \le \frac{c_r}{4}\sqrt{\log d}$ for all $i \in [K]$ and $j \neq i$, we have

$$\mathbb{E}_{X_t \sim \mathcal{N}(\mu_{i,t}^*, I)}[w_{j,t}(X_t)] \le \frac{1}{d^{c_r^2/100}}.$$

Computing the expectation of the product of the weights $w_{j,t}$ and $w_{k,t}$ for any distinct j, k, we have

$$\begin{aligned}
\mathbb{E}_{X_t}[w_{j,t}(X_t)w_{k,t}(X_t)] &= \sum_{i=1}^{K} \frac{1}{K}\mathbb{E}_{x \sim \mathcal{N}(\mu_i^*, I)}[w_{j,t}(x)w_{k,t}(x)] \\
&\le \frac{1}{K}\sum_{i=1}^{K} \sqrt{\mathbb{E}_{x \sim \mathcal{N}(\mu_i^*, I)}[w_{j,t}(x)^2]\mathbb{E}_{x \sim \mathcal{N}(\mu_i^*, I)}[w_{k,t}(x)^2]} \\
&\le \frac{1}{d^{c_r^2/200}}
\end{aligned}$$

where the last inequality uses the fact that either $i \neq j$ or $i \neq k$ and $w_{j,t}(x)^2 \le w_{j,t}(x) \le 1$. □

Lemma E.6 (Lemma 4.3 of [SN21])**.** *Suppose X is distributed according to a mixture of K Gaussians with centers $\theta^* = \{\mu_1^*, \ldots, \mu_K^*\}$ as in Eq. (6). For any $\theta = \{\mu_1, \mu_2, \ldots, \mu_K\}$ such that $\|\mu_i - \mu_i^*\| \le \frac{c_r}{4}\sqrt{\log d}$ for all $i \in [K]$, then for any distinct $i, j \in [K]$, we have*

$$\begin{aligned}
\left\|\mathbb{E}_X[w_i(X,\mu)(1 - w_i(X,\mu))(X - \mu_i)(X - \mu_i)^\top]\right\|_{\mathsf{op}} &\le \frac{d^2 c_r^2}{d^{c_r^2/1000}} \\
\left\|\mathbb{E}_X[w_i(X,\theta)w_j(x,\theta)(X - \mu_i)(X - \mu_j)^\top]\right\|_{\mathsf{op}} &\le \frac{d^2 c_r^2}{d^{c_r^2/1000}}
\end{aligned}$$

E.2 Closeness between population gradient descent and empirical gradient descent

In this section, we show that the population gradient descent on the DDPM objective is close to the empirical gradient descent for mixtures of K Gaussians.

Lemma E.7. *For any ε that is $\Theta(\frac{1}{\mathrm{poly}(d)})$ and noise scale $t > t'$ where $t' \lesssim 1$, the empirical estimate of gradient descent update on the DDPM objective with the number of samples $n > n'$ concentrates well to the population gradient descent update where $n' = O(\frac{K^4 d^5 B^6}{\varepsilon^2})$. More specifically, the following inequality holds with probability at least $1 - \exp(-d^{0.99})$:*

$$\left\|\nabla_{\mu_{1,t}}\Big(\frac{1}{n}\sum_{i=1}^{n} L_t(s_{\theta_t}(x_{i,0}, z_{i,t}))\Big) - \nabla_{\mu_{1,t}} L_t(s_{\theta_t})\right\| \le \varepsilon.$$

Proof. Recall that the population gradient is given by

$$\nabla_{\mu_{1,t}} L_t(s_{\theta_t}) = \mathbb{E}\Big[\frac{1}{2}\nabla_{\mu_{1,t}}\big\|s_{\theta_t}(X_t)\big\|^2 + \frac{\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t}{\beta_t}\Big],$$

where

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{2}\nabla_{\mu_{1,t}}\big\|s_{\theta_t}(X_t)\big\|^2\Big] = \mathbb{E}\Big[&\Big(w_{1,t}(X_t)(X_t - \mu_{1,t})\mu_{1,t}^\top + w_{1,t}(X_t)\cdot \mathrm{Id} \\
&- w_{1,t}(X_t)\sum_{i=1}^{K} w_{i,t}(X_t)(X_t - \mu_{1,t})\mu_{i,t}^\top\Big)\cdot \sum_{i=1}^{K}\big(w_{i,t}(X_t)\mu_{i,t} - X_t\big)\Big],
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\Big[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t\Big] = \mathbb{E}\Big[&\Big(w_{1,t}(X_t)(X_t - \mu_{1,t})\mu_{1,t}^\top Z_t \\
&+ w_{1,t}(X_t)Z_t - w_{1,t}(X_t)\sum_{i=1}^{K} w_{i,t}(X_t)(X_t - \mu_{1,t})\mu_{i,t}^\top Z_t\Big)\Big]
\end{aligned} \tag{E.3}$$

We will prove that the sample estimate of each coordinate in Eq. (E.3) concentrates well around the expectation. We will prove the concentration of the first coordinate and a similar analysis holds for other coordinates. For the rest of the proof, we use $\tilde{x}_t$ to denote the first coordinate of X_t and $\tilde{\mu}_{i,t}$ to indicate the first coordinate $\mu_{i,t}$. For any random variable $Y \in \mathbb{R}$, we use $\|Y\|_{\psi_1}$ to denote the sub-exponential norm of Y and $\|Y\|_{\psi_2}$ to denote the sub-gaussian norm of Y (See lemma B.1 for details). Using properties of a sub-Gaussian random variable from Lemma B.1, we get

$$\begin{aligned}
&\Big\| \sum_{j=1}^{K} w_{1,t}(X_t) w_{j,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t}) \mu_{1,t}^\top \mu_{j,t} \Big\|_{\psi_2} \\
&\lesssim \sum_{j=1}^{K} \Big\| w_{1,t}(X_t) w_{j,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t}) \mu_{1,t}^\top \mu_{j,t} \Big\|_{\psi_2} \\
&\qquad\qquad \text{(Using sum of sub-Gaussian random variables property in Lemma B.1)} \\
&\lesssim \sum_{j=1}^{K} \Big\| w_{1,t}(X_t) w_{j,t}(X_t) \mu_{1,t}^\top \mu_{j,t} z \Big\|_{\psi_2} + \Big\| w_{1,t}(X_t) w_{j,t}(X_t) \mu_{1,t}^\top \mu_{j,t} (\tau - \tilde{\mu}_{1,t}) \Big\|_{\psi_2} \\
&\lesssim KB^2 + KB^3 \lesssim KB^3, \qquad \text{(E.4)}
\end{aligned}$$

where the third inequality follows by writing $\tilde{x}_t = z + \tau$ where $z \sim \mathcal{N}(0,1)$ and τ is a random variable that takes $\tilde{\mu}_{i,t}^*$ for every $i \in [K]$ with probability $\frac{1}{K}$. The fourth inequality follows from the sub-Gaussian property of a bounded random variable and the product of a sub-Gaussian random variable with bounded random variable property in Lemma B.1. Using the sum of sub-Gaussian random variable property in Lemma B.1, we have

$$\Big\| \sum_{i=1}^{K} w_{1,t}(X_t) w_{i,t}(X) \tilde{\mu}_{i,t} \Big\|_{\psi_2} \lesssim \sum_{i=1}^{K} \| w_{1,t}(X_t) w_{i,t}(X) \tilde{\mu}_{i,t} \|_{\psi_2} \lesssim KB. \qquad \text{(E.5)}$$

Using properties of the sub-Gaussian random variable from Lemma B.1 in a similar way of Eq. (E.4), we have

$$\begin{aligned}
&\Big\| \sum_{i=1}^{K} \sum_{j=1}^{K} w_{1,t}(X_t) w_{i,t}(X_t) w_{j,t}(X_t) \mu_{i,t}^\top \mu_{j,t} (\tilde{x}_t - \tilde{\mu}_{1,t}) \Big\|_{\psi_2} \\
&\leq \sum_{i=1}^{K} \sum_{j=1}^{K} \Big\| w_{1,t}(X_t) w_{i,t}(X_t) w_{j,t}(X_t) \mu_{i,t}^\top \mu_{j,t} (\tilde{x}_t - \tilde{\mu}_{1,t}) \Big\|_{\psi_2} \\
&\leq \sum_{i=1}^{K} \sum_{j=1}^{K} \Big\| w_{1,t}(X_t) w_{i,t}(X_t) w_{j,t}(X_t) \mu_{i,t}^\top \mu_{j,t} z \Big\|_{\psi_2} + \Big\| w_{1,t}(X_t) w_{i,t}(X_t) w_{j,t}(X_t) \mu_{i,t}^\top \mu_{j,t} (\tau - \tilde{\mu}_{i,t}) \Big\|_{\psi_2} \\
&\leq K^2 B^2 + K^2 B^3 \lesssim K^2 B^3 \qquad \text{(E.6)}
\end{aligned}$$

We know that $\|w_{1,t}(X_t) \mu_{1,t}^\top X_t\|_{\psi_2} \leq \| \sum_{i=1}^{d} \mu_{1,t}(i) X_t(i) \|_{\psi_2} \lesssim dB^2$ and $\|\tilde{x}_t - \tilde{\mu}_{1,t}\|_{\psi_2} \lesssim B$. Using the fact that the product of two sub-Gaussian random variables is a sub-exponential random variable, we have

$$\|w_{1,t}(X_t) \mu_{1,t}^\top X_t (\tilde{x}_t - \tilde{\mu}_{1,t})\|_{\psi_1} \leq \|\tilde{x}_t - \tilde{\mu}_{1,t}\|_{\psi_2} \|w_{1,t}(X_t) \mu_{1,t}^\top X_t\|_{\psi_2} \lesssim dB^3 \qquad \text{(E.7)}$$

The sub-gaussian norm of $w_{1,t}(X_t)\tilde{x}_t$ term in the gradient is given by

$$\|w_{1,t}(X_t)\tilde{x}_t\|_{\psi_2} \leq \|X_t\|_{\psi_2} \lesssim \|Z\|_{\psi_2} + \|\tau\|_{\psi_2} \lesssim B \qquad \text{(E.8)}$$

Using the property that the product of two sub-Gaussian random variables is a sub-exponential random variable, we obtain

$$\begin{aligned}
&\Big\| w_{1,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t}) \Big(\sum_{i=1}^{K} w_{i,t}(X_t) \mu_{i,t}^\top X_t \Big) \Big\|_{\psi_1} \\
&\lesssim \|w_{1,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t})\|_{\psi_2} \Big\| \Big(\sum_{i=1}^{K} w_{i,t}(X_t) \mu_{i,t}^\top X_t \Big) \Big\|_{\psi_2} \\
&\lesssim KdB^3 \qquad \text{(E.9)}
\end{aligned}$$

For any random variable Y, we know that $\|X\|_{\psi_1} \leq \|X\|_{\psi_2}$. Therefore, combining Eq. (E.4), (E.5), (E.6), (E.7), (E.8) and (E.9), we have

$$\begin{aligned}\|[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top s_{\theta_t}(X_t)]_1 - \mathbb{E}[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top s_{\theta_t}(X_t)]_1\|_{\psi_1} &\lesssim \|[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top s_{\theta_t}(X_t)]_1\|_{\psi_1} \\ &\lesssim K^2 dB^3 \end{aligned} \tag{E.10}$$

Now, we shift our focus on obtaining the sub-exponential norm of $\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t$. Using $\|w_{1,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t})\|_{\psi_2} \lesssim B$ and $\|\mu_{1,t}^\top Z_t\|_{\psi_2} \lesssim dB$, we obtain

$$\|w_{1,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t})\mu_{1,t}^\top Z_t\|_{\psi_1} \leq \|w_{1,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t})\|_{\psi_2} \|\mu_{1,t}^\top Z_t\|_{\psi_2} \lesssim dB^2 \tag{E.11}$$

Using Lemma B.1, we have $\|w_{1,t}(X_t) z_t\|_{\psi_2} \leq \|z_t\|_{\psi_2} \lesssim 1$. For the last term, we have

$$\begin{aligned}\Big\|w_{1,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t}) \sum_{i=1}^K w_{i,t}(X_t)\mu_{i,t}^\top Z_t\Big\|_{\psi_1} &\leq \|w_{1,t}(X_t)(\tilde{x}_t - \tilde{\mu}_{1,t})\|_{\psi_2} \Big\| \sum_{i=1}^K w_{i,t}(X_t)\mu_{i,t}^\top Z_t\Big\|_{\psi_2} \\ &\lesssim KdB^2 \end{aligned} \tag{E.12}$$

Combining Eq. (E.11), (E.12), we have

$$\Big\|\frac{[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t]_1}{\beta_t} - \frac{\mathbb{E}[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t]_1}{\beta_t}\Big\|_{\psi_1} \lesssim \Big\|\frac{[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t]_1}{\beta_t}\Big\|_{\psi_1} \lesssim \frac{KdB^2}{\beta_t} \tag{E.13}$$

where $[\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t]_1$ denotes the first coordinate of $\nabla_{\mu_{1,t}} s_{\theta_t}(X_t)^\top Z_t$. Combining Eq. (E.10) and Eq. (E.13), we have

$$\Big\|[\nabla_{\mu_{1,t}} L_t(s_{\theta_t}(X_t))]_1 - [\nabla_{\mu_{1,t}} L_t(s_{\theta_t})]_1\Big\|_{\psi_1} \lesssim \frac{K^2 dB^3}{\beta_t}$$

For each i.i.d. sample $x_{i,t}$, the term $[\nabla_{\mu_{1,t}} L_t(s_{\theta_t}(x_{i,t}))]_1 - [\nabla_{\mu_{1,t}} L_t(s_{\theta_t})]_1$ is also independent and identically distributed. Therefore, using Lemma B.3, for any ε that is $\Theta(\frac{1}{\text{poly}(d)})$, we have

$$\Pr\Big[\Big|\frac{1}{n}\sum_{i=1}^n [\nabla_{\mu_{1,t}} L_t(s_{\theta_t}(x_{i,t}))]_1 - [\nabla_{\mu_{1,t}} L_t(s_{\theta_t})]_1\Big| \geq \varepsilon\Big] \leq 2\exp\Big(-\frac{n\varepsilon^2\beta_t^2}{K^4 d^2 B^6}\Big).$$

A similar analysis will give the concentration for each coordinate. Using the union bound and rescaling ε as $\frac{\varepsilon}{d}$, with probability at least $1 - 2d\exp\Big(-\frac{n\varepsilon^2\beta_t^2}{K^4 d^4 B^6}\Big)$, we have

$$\Big\|\nabla_{\mu_{1,t}}\Big(\frac{1}{n}\sum_{i=1}^n L_t(s_{\theta_t}(x_{i,t}))\Big) - \nabla_{\mu_{1,t}} L_t(s_{\theta_t})\Big\| \leq \varepsilon$$

Note that for any $t = \Omega(1)$, $\beta_t \geq c$ for some constant c. Therefore, choosing n provided in the Lemma E.7 statement, we obtain the result. □

E.3 Proof of Theorem E.1

Proof of Theorem E.1. For any training iteration h, assume that parameters $\theta_t^{(h)}$ are such that $\Big\|\mu_{i,t}^{(h)} - \mu_{i,t}^*\Big\| \leq \frac{c_r}{4}\sqrt{\log d}$ we can write the update on the DDPM objective as follows:

$$\begin{aligned}\|\mu_{1,t}^{(h+1)} - \mu_{1,t}^*\| &= \Big\|\mu_{1,t}^{(h)} - \eta\nabla\Big(\frac{1}{n}\sum_{i=1}^n L_t(s_{\theta_t^{(h)}}(x_{i,0}, z_{i,t}))\Big) - \mu_{1,t}^*\Big\| \\ &\leq \big\|\mu_{1,t}^{(h)} + \eta\,\mathbb{E}[w_{1,t}(X_t)(X_t - \mu_{1,t}^{(h)})] - \mu_{1,t}^*\big\| \\ &\quad + \eta\Big\|\big(-\nabla_{\mu_{1,t}} L_t(s_{\theta_t})\big) - \mathbb{E}[w_{1,t}(X_t)(X_t - \mu_{1,t}^{(h)})]\Big\| \\ &\quad + \eta\Big\|\big(\nabla_{\mu_{1,t}} L_t(s_{\theta_t})\big) - \nabla_{\mu_{1,t}}\Big(\frac{1}{n}\sum_{i=1}^n L_t(s_{\theta_t^{(h)}}(x_{i,0}, z_{i,t}))\Big)\Big\|. \end{aligned}$$

Using Lemma E.4, Lemma E.7 and Theorem 3.2 from [SN21], for any $\eta \in (0, K)$, we have

$$\|\mu_{1,t}^{(h+1)} - \mu_{1,t}^*\| \leq \left(1 - \frac{3\eta}{8K}\right)\|\mu_{1,t}^{(h)} - \mu_{1,t}^*\| + \frac{\eta K^2 B^2}{d^{\frac{c_r^2}{4000}}} + \eta\varepsilon.$$

Choosing $\eta = \frac{2K}{3}$, c_r to be sufficiently large constant and ε to be $\Theta(\frac{1}{\text{poly}(d)})$, we have

$$\|\mu_{1,t}^{(h+1)} - \mu_{1,t}^*\| \leq \frac{3}{4}\|\mu_{1,t}^{(h)} - \mu_{1,t}^*\| + \varepsilon$$

By assumption 15, $\|\mu_{1,t}^{(0)} - \mu_{1,t}^*\| \leq O(\sqrt{\log d})$ and therefore, choosing H to be $\Omega(\log(\frac{\log d}{\varepsilon}))$, we obtain the result. □

F Additional proofs

F.1 Proof of Lemma C.2

Proof of Lemma C.2. By calculating the negative gradient of the DDPM objective in Eq. (5), we obtain

$$\begin{aligned}
-\nabla_{\mu_t} L_t(s_{\mu_t}) &= -\mathbb{E}_{X_0,Z_t}[(\tanh(\mu_t^\top X_t)I + \tanh'(\mu_t^\top X_t)X_t\mu_t^\top)(s_{\mu_t}(X_t) + \frac{Z_t}{\beta_t})] \\
&= -\mathbb{E}[(\tanh(\mu_t^\top X_t)I + \tanh'(\mu_t^\top X_t)X_t\mu_t^\top)(\tanh(\mu_t^\top X_t)\mu_t - X_t + \frac{Z_t}{\beta_t})] \\
&= \mathbb{E}[-\tanh^2(\mu_t^\top X_t)\mu_t - \tanh(\mu_t^\top X_t)\tanh'(\mu_t^\top X_t)X_t\|\mu_t\|^2 + \tanh(\mu_t^\top X_t)X_t \\
&\quad + \tanh'(\mu_t^\top X_t)\mu_t^\top X_t X_t - \tanh(\mu_t^\top X_t)\frac{Z_t}{\beta_t} - \tanh'(\mu_t^\top X_t)X_t\mu_t^\top\frac{Z_t}{\beta_t}]
\end{aligned}$$

By simplifying the gradient terms involving Z_t by the Stein's identity as in Lemma F.1 and plugging it back in the gradient, we obtain

$$\begin{aligned}
-\nabla_{\mu_t} L_t(s_{\mu_t}) &= \mathbb{E}\Big[\left(\tanh(\mu_t^\top X_t) - \tanh(\mu_t^\top X_t)\tanh'(\mu_t^\top X_t)\|\mu_t\|^2 + \tanh'(\mu_t^\top X_t)\mu_t^\top X_t\right)X_t\Big] \\
&\quad - \mu_t - \mathbb{E}\left[\tanh''(\mu_t^\top X_t)\|\mu_t\|^2 X_t\right] - \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t\right] \\
&= \mathbb{E}\Big[\left(\tanh(\mu_t^\top X_t) - 0.5\tanh''(\mu_t^\top X_t)\|\mu_t\|^2 + \tanh'(\mu_t^\top X_t)\mu_t^\top X_t\right)X_t\Big] \\
&\quad - \mu_t - \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t\right]
\end{aligned}$$

Observe that $\left(\tanh(\mu^\top x) - \frac{1}{2}\tanh''(\mu^\top x)\|\mu\|^2 + \tanh'(\mu^\top x)\mu^\top x\right)x$ and $\tanh'(\mu^\top x)$ are even functions and X_t is a symmetric distribution, therefore, for any even function f, we can write $\mathbb{E}_{X_t}[f(X_t)] = \frac{1}{2}\mathbb{E}_{X_t\sim\mathcal{N}(\mu_t^*,\text{Id})}[f(X_t)] + \frac{1}{2}\mathbb{E}_{X_t\sim\mathcal{N}(-\mu_t^*,I)}[f(X_t)] = \mathbb{E}_{X_t\sim\mathcal{N}(\mu_t^*,\text{Id})}[f(X_t)]$. Applying this property of the even function on the gradient update, we obtain the result. □

Lemma F.1. *When random variable $X_t = \alpha_t X_0 + \beta_t Z_t$ where $Z_t \sim \mathcal{N}(0, I)$, $\alpha_t = \exp(-t)$ and $\beta_t = \sqrt{1 - \exp(-2t)}$, then for any $t > 0$, the following two equations hold.*

$$\mathbb{E}_{X_0,Z_t}\left[\tanh(\mu_t^\top X_t)\frac{Z_t}{\beta_t} + \tanh^2(\mu_t^\top X_t)\mu_t\right] = \mu_t$$

$$\mathbb{E}_{X_0,Z_t}\left[\tanh'(\mu_t^\top X_t)\frac{\mu_t^\top Z_t}{\beta_t}X_t\right] = \mathbb{E}_{X_0,Z_t}\left[\tanh''(\mu_t^\top X_t)\|\mu_t\|^2 X_t + \tanh'(\mu_t^\top X_t)\mu_t\right]$$

Proof. Applying Stein's lemma on the first term, we get the first equation of the statement in the Lemma.

$$\begin{aligned}
\mathbb{E}_{X_0,Z_t}\left[\tanh(\mu_t^\top X_t)\frac{Z_t}{\beta_t}\right] &= \mathbb{E}_{X_0,Z_t}\left[\tanh(\mu_t^\top(\alpha_t X_0 + \beta_t Z_t))\frac{Z_t}{\beta_t}\right] = \mathbb{E}_{X_0,Z_t}\left[\tanh'(\mu_t^\top X_t)\mu_t\right] \\
&= \mathbb{E}_{X_0,Z_t}\left[\left(1 - \tanh^2(\mu_t^\top X_t)\right)\mu_t\right]
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
\mathbb{E}\Big[\tanh'(\mu_t^\top X_t)\frac{\mu_t^\top Z_t}{\beta_t}X_t\Big] &= \mathbb{E}\Big[\tanh'(\mu_t^\top X_t)\frac{\mu_t^\top Z_t}{\beta_t}\alpha_t X_0\Big] + \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t^\top Z_t Z_t\right] \\
&= \sum_{i=1}^{d}\mathbb{E}\Big[\alpha_t X_0 \tanh'(\mu_t^\top X_t)\frac{\mu_t(i)Z_t(i)}{\beta_t}\Big] + \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t\right] + \mathbb{E}\left[\tanh''(\mu_t^\top X_t)\mu_t^\top Z_t\beta_t\mu_t\right] \\
&= \sum_{i=1}^{d}\mathbb{E}\Big[\alpha_t X_0 \tanh''(\mu_t^\top X_t)\mu_t(i)\mu_t(i)\Big] + \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t\right] + \mathbb{E}\left[\tanh''(\mu_t^\top X_t)\mu_t^\top Z_t\beta_t\mu_t\right]
\end{aligned}$$

where the second equality follows from the Stein's lemma on the $\mathbb{E}[\tanh'(\mu_t^\top X_t)\mu_t^\top Z_t Z_t]$ and the last equality follows from the Stein's lemma on $\mathbb{E}[\alpha_t X_0 \tanh''(\mu_t^\top X_t)\mu_t(i)Z_t(i)]$. Applying Stein's inequality on the $\mathbb{E}\left[\tanh''(\mu_t^\top X_t)\mu_t^\top Z_t\beta_t\mu_t\right]$, we obtain

$$\begin{aligned}
&= \mathbb{E}\left[\alpha_t X_0 \tanh''(\mu_t^\top X_t)\|\mu_t\|^2\right] + \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t\right] + \sum_{i=1}^{d}\beta_t\mu_t\mathbb{E}\left[\tanh'''(\mu_t^\top X_t)\mu_t(i)\beta_t\mu_t(i)\right] \\
&= \mathbb{E}\left[X_t \tanh''(\mu_t^\top X_t)\|\mu_t\|^2\right] - \mathbb{E}\left[\beta_t Z_t \tanh''(\mu_t^\top X_t)\|\mu_t\|^2\right] + \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t\right] \\
&\quad + \beta_t^2\|\mu_t\|^2\mu_t\mathbb{E}\left[\tanh'''(\mu_t^\top X_t)\right] \\
&= \mathbb{E}\left[X_t \tanh''(\mu_t^\top X_t)\|\mu_t\|^2\right] + \mathbb{E}\left[\tanh'(\mu_t^\top X_t)\mu_t\right].
\end{aligned}$$

□

F.2 Proof of Lemma C.8

Proof of Lemma C.8. Recall that the gradient update for any μ_t^* is given by

$$-\nabla_{\mu_t^*}L_t(s_{\mu_t^*}) = G(\mu_t^*, \mu_t^*) + \eta\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^{*\top}x)x] - \eta\mu_t^* \tag{F.1}$$

We know that $\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^{*\top}x)x] = \mu_t^*$ (Eq.(2.1) of [DTZ17]) and $\nabla_{\mu_t^*}L_t(s_{\mu_t^*}) = 0$ because μ_t^* is a stationary point of the regression objective of diffusion model. This implies that $G(\mu_t^*, \mu_t^*) = 0$ for any μ_t^*.

Note that this proof only talks about 1D case therefore, for the purpose of this proof, we use a to denote μ and b to denote μ^*. In 1D, using Mean value theorem, we have

$$\frac{G(a,b) - G(a,a)}{b-a} = \frac{dG(a,\xi)}{d\xi} \text{ for some } \xi \in [a,b] \text{ (if } a < b) \tag{F.2}$$

Using the fact that $G(a,a) = 0$ in Eq. (F.2), we have

$$\big|G(a,b)\big| = \left|\frac{dG(a,\xi)}{d\xi}\right||b-a|$$

Observe that it suffices to prove $\left|\frac{dG(a,\xi)}{d\xi}\right| \le 0.01$ to obtain the lemma. By computing the gradient of G, we obtain

$$\frac{dG(a,\xi)}{d\xi} = \eta\mathbb{E}_{x\sim\mathcal{N}(\xi,1)}\Big[2\tanh'(ax)ax + \tanh''(ax)\left(\frac{-3a^2}{2} + a^2x^2\right) - \frac{1}{2}a^3x\tanh'''(ax)\Big]$$

For the first term, we have

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(\xi,I)}[\tanh'(ax)ax] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tanh'(ax)axe^{-\frac{(x-\xi)^2}{2}}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}\tanh'(ax)ax\left(e^{-\frac{(x-\xi)^2}{2}}-e^{-\frac{(x+\xi)^2}{2}}\right)dx \\
&\le \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}e^{-ax}axe^{-\frac{(x-\xi)^2}{2}}dx \\
&\le \frac{ae^{\frac{a^2-2a\xi}{2}}}{\sqrt{2\pi}}\int_{0}^{\infty}xe^{-\frac{(x-\xi+a)^2}{2}}dx \\
&\le ae^{\frac{a^2-2a\xi}{2}}(\sqrt{\frac{2}{\pi}}e^{-\frac{(\xi-a)^2}{2}}+(\xi-a)\text{erf}\left(\frac{\xi-a}{\sqrt{2}}\right)) \\
&\le ae^{-\frac{\xi^2}{2}}+a|\xi-a|\,e^{\frac{-2a(\xi-a)-a^2}{2}}
\end{aligned}$$

Using Lemma 1 of [DTZ17], we know that $\mathbb{E}_{x\sim\mathcal{N}(\xi,I)}[\tanh'(ax)ax]>0$. Therefore, we have

$$\left|\mathbb{E}_{x\sim\mathcal{N}(\xi,I)}[\tanh'(ax)ax]\right| \le ae^{-\frac{\xi^2}{2}}+a|\xi-a|\,e^{\frac{-2a(\xi-a)-a^2}{2}}$$

For the second term, we have

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(\xi,1)}[\tanh''(ax)(-\frac{3a^2}{2}+a^2x^2)] \\
&= \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}a^2\tanh''(ax)(-\frac{3}{2}+x^2)\left(\exp(-\frac{(x-\xi)^2}{2})-\exp(-\frac{(x+\xi)^2}{2})\right)dx \\
&\le \frac{1}{\sqrt{2\pi}}\int_{0}^{\sqrt{\frac{3}{2}}}a^2e^{-2ax}(\frac{3}{2}-x^2)\exp(-\frac{(x-\xi)^2}{2})dx \\
&\le \frac{3}{\sqrt{2\pi}}a^2\exp(-\frac{a^2}{16})
\end{aligned}$$

Assuming $a\ge\sqrt{6}$, then when $\xi\ge a\ge\sqrt{6}$, we have $\exp(-\frac{(x-\xi)^2}{2})\le\exp(-\frac{a^2}{4})$ and when $\xi\le a$, using $\xi\ge\frac{3a}{4}$, we have $\exp(-\frac{(x-\xi)^2}{2})\le\exp(-\frac{a^2}{16})$. For the lower bound, we have

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(\xi,1)}[\tanh''(ax)(-\frac{3a^2}{2}+a^2x^2)] \\
&= \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}\tanh''(ax)(-\frac{3a^2}{2}+a^2x^2)\left(\exp(-\frac{(x-\xi)^2}{2})-\exp(-\frac{(x+\xi)^2}{2})\right)dx \\
&\ge \frac{1}{\sqrt{2\pi}}\int_{\sqrt{\frac{3}{2}}}^{\infty}\tanh''(ax)(-\frac{3a^2}{2}+a^2x^2)\left(\exp(-\frac{(x-\xi)^2}{2})-\exp(-\frac{(x+\xi)^2}{2})\right)dx \\
&\ge \frac{1}{\sqrt{2\pi}}\int_{\sqrt{\frac{3}{2}}}^{\infty}\tanh''(ax)a^2x^2\left(\exp(-\frac{(x-\xi)^2}{2})-\exp(-\frac{(x+\xi)^2}{2})\right)dx \\
&\ge -\frac{8a^2}{\sqrt{2\pi}}\int_{\sqrt{\frac{3}{2}}}^{\infty}e^{-2ax}x^2\left(\exp(-\frac{(x-\xi)^2}{2})-\exp(-\frac{(x+\xi)^2}{2})\right)dx \\
&\ge -\frac{8a^2e^{-\sqrt{6}a}}{\sqrt{2\pi}}\int_{\sqrt{\frac{3}{2}}}^{\infty}x^2\exp(-\frac{(x-\xi)^2}{2})dx \ge -8a^2e^{-\sqrt{6}a}
\end{aligned}$$

Using upper bound and lower bound, we have

$$\left|\mathbb{E}_{x\sim\mathcal{N}(\xi,1)}[\tanh''(ax)a^2(-\frac{3}{2}+x^2)]\right| \le 8a^2e^{-\sqrt{6}a}$$

For the third term, we have

$$\begin{aligned}
&\left|\mathbb{E}_{x\sim\mathcal{N}(\xi,1)}[\frac{a^3x}{2}\tanh'''(ax)]\right| \\
&= \left|\frac{1}{32\sqrt{2\pi}}\int_0^\infty a^3x\sigma(2ax)(1-\sigma(2ax))\left(1-6\sigma(2ax)(1-\sigma(2ax))\right)\left(\exp\left(-\frac{(x-\xi)^2}{2}\right)\right.\right. \\
&\qquad \left.\left.-\exp\left(-\frac{(x+\xi)^2}{2}\right)\right)dx\right| \\
&\le \left|\frac{3a^3}{16\sqrt{2\pi}}\int_0^\infty x\sigma^2(2ax)(1-\sigma(2ax))^2\left(\exp\left(-\frac{(x-\xi)^2}{2}\right)-\exp\left(-\frac{(x+\xi)^2}{2}\right)\right)dx\right| \\
&\le \frac{3a^3}{16\sqrt{2\pi}}\int_0^\infty xe^{-ax}\exp\left(-\frac{(x-\xi)^2}{2}\right)dx \\
&\le \frac{a^3}{10}e^{-\frac{\xi^2}{2}}+\frac{a^3}{10}|\xi-a|\,e^{\frac{-2a(\xi-a)-a^2}{2}}.
\end{aligned}$$

We can lower bound the third term as follows:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(\xi,1)}[\frac{a^3x}{2}\tanh'''(ax)] \\
&\ge \frac{1}{2\sqrt{2\pi}}\int_0^c a^3x\tanh'''(ax)\left(\exp\left(-\frac{(x+\xi)^2}{2}\right)-\exp\left(-\frac{(x-\xi)^2}{2}\right)\right)dx \\
&\ge \frac{a^3}{2\sqrt{2\pi}}\int_0^c x\exp\left(-\frac{(x-\xi)^2}{2}\right)\left(\exp\left(-2\xi x\right)-1\right)dx \\
&\ge -\frac{a^3\xi}{\sqrt{2\pi}}\int_0^c x^2\exp\left(-\frac{(x-\xi)^2}{2}\right)dx \ge -\frac{\xi\exp(-\frac{\xi^2}{4})}{\sqrt{2\pi}}
\end{aligned}$$

Using all the bounds, we have

$$\left|\frac{dG(a,\xi)}{d\xi}\right| \le \frac{a^3}{10}e^{-\frac{\xi^2}{2}}+\frac{a^3}{10}|\xi-a|\,e^{\frac{-2a(\xi-a)-a^2}{2}}+8a^2e^{-\sqrt{6}a}+ae^{-\frac{\xi^2}{2}}+a|\xi-a|\,e^{\frac{-2a(\xi-a)-a^2}{2}}$$

When $\xi \ge a$ and $a \ge c$ for some sufficiently large constant c (for example, $c = 25$), then, we have

$$\left|\frac{dG(a,\xi)}{d\xi}\right| \le \frac{a^3}{10}e^{-\frac{a^2}{2}}+\frac{a^3}{10}|\xi-a|\,e^{\frac{-a^2}{2}}+8a^2e^{-\sqrt{6}a}+ae^{-\frac{a^2}{2}}+a|\xi-a|\,e^{\frac{-a^2}{2}} \le 0.01$$

When $\frac{3a}{4} \le \xi \le a$ and $a > c$ for sufficiently large constant c (for example, $c = 25$), we have

$$\left|\frac{dG(a,\xi)}{d\xi}\right| \le \frac{a^3}{10}e^{-\frac{9a^2}{32}}+\frac{a^4}{40}e^{\frac{-a^2}{4}}+8a^2e^{-\sqrt{6}a}+ae^{-\frac{a^2}{2}}+\frac{a^2}{4}e^{\frac{-a^2}{4}} \le 0.01$$

Pluggint the bound on $|\frac{dG(a,\xi)}{d\xi}|$ in Eq. (F.1), we obtain the final result. □

F.3 Proof of Lemma C.10

Proof of Lemma C.10. We will prove this by induction. For $h = 0$, this is true because the algorithm initializes the gradient descent on the low noise regime with the output of gradient descent on the high noise regime, and the output is guaranteed to have $\langle\hat{\mu}_t^{(0)}, \hat{\mu}_t^*\rangle$ to be $\Omega(1)$ and by assumption $\|\mu_t^*\| > c'$, therefore $\|\mu_t^{(0)}\| \in [c, \frac{4\langle\hat{\mu}_t^{(0)},\mu_t^*\rangle}{3}]$.

Suppose $\|\mu_t^{(h)}\| \in [c, \frac{4\langle\hat{\mu}_t^{(h)},\mu_t^*\rangle}{3}]$, then we know that $\|\mu_t^{(h+1)} - \mu_t^*\| < \|\mu_t^{(h)} - \mu_t^*\|$. To prove $\|\mu_t^{(h+1)}\| \in [c, \frac{4\langle\hat{\mu}_t^{(h+1)},\mu_t^*\rangle}{3}]$, first we will prove that $\langle\hat{\mu}_t^{(h)}, \mu_t^{(r+1)}\rangle \in [c, \frac{6\langle\hat{\mu}_t^{(h)},\mu_t^*\rangle}{5}]$. Note that the update in the direction of $\langle\hat{\mu}_t, \mu_t\rangle$ works like 1D. Therefore, we have a contraction for it as follows.

$$\left|\langle\hat{\mu}_t^{(h)}, \mu_t^{(h+1)}\rangle - \langle\hat{\mu}_t^{(h)}, \mu_t^*\rangle\right| < \left|\langle\hat{\mu}_t^{(h)}, \mu_t^{(h)}\rangle - \langle\hat{\mu}_t, \mu_t^*\rangle\right|$$

If $\|\mu_t^{(h)}\| \leq \langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle$, then using Lemma F.4, we know $\langle \hat{\mu}_t^{(h)}, \mu_t^{(h+1)} \rangle \leq \frac{6\langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle}{5}$ and $\langle \hat{\mu}_t^{(h)}, \mu_t^{(h+1)} \rangle \geq \|\mu_t^{(h)}\| \geq c$ because of the contraction. If $\|\mu_t^{(h)}\| \geq \langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle$ and $\langle \hat{\mu}_t^{(h)}, \mu_t^{(h+1)} \rangle \geq \langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle$, then $\langle \hat{\mu}_t^{(h)}, \mu_t^{(h+1)} \rangle \leq \|\mu_t^{(h)}\|$ because of the contraction. If $\|\mu_t^{(h)}\| \geq \langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle$ and $\langle \hat{\mu}_t^{(h+1)}, \mu_t^{(h)} \rangle \leq \langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle$, then using $\langle \hat{\mu}_t^{(h+1)}, \mu_t^{(h)} \rangle \geq \|\mu_t^{(h)}\| - \Big| U(\langle \hat{\mu}_t^{(h)}, \mu_t^{(h)} \rangle, \langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle) \Big| \geq \frac{4\langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle}{5} \geq \frac{4\langle \hat{\mu}_t^{(0)}, \mu_t^* \rangle}{5} \geq c$ from Lemma F.2, we get the result that $\langle \hat{\mu}_t^{(h)}, \mu_t^{(h+1)} \rangle \in [c, \frac{6\langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle}{5}]$. Now, using Lemma F.2, we get

$$
\begin{aligned}
\langle \hat{\mu}_t^{(h)}, \mu_t^{(h+1)} \rangle \in [c, \frac{6\langle \hat{\mu}_t^{(h)}, \mu_t^* \rangle}{5}] &\implies \|\mu_t^{(h+1)}\| \in \big[\frac{c}{\cos\alpha_h}, \frac{6\|\mu_t^*\|\cos\beta_h}{5\cos\alpha_h}\big] \\
&\implies \|\mu_t^{(h+1)}\| \in \big[c, \frac{4\|\mu_t^*\|\cos\beta_{h+1}}{3}\big] \\
&\implies \|\mu_t^{(h+1)}\| \in \big[c, \frac{4\langle \hat{\mu}_t^{(h+1)}, \mu_t^* \rangle}{3}\big] \qquad \square
\end{aligned}
$$

Lemma F.2. *Suppose the angle between $\mu^{(r)}$ and μ^* is β_r and α_r is the angle between $\mu^{(r)}$ and $\mu^{(r+1)}$ and assume the contraction is true at time r. Assume that $\beta_0 \in (0, \frac{\pi}{2})$. Then:*

$$
\alpha_r \in (0, \pi/2) \ \ \forall r \qquad \textit{and} \qquad \cos\beta_r \leq \cos\beta_{r+1}
$$

which implies that

$$
\cos\beta_r \leq \cos\beta_{r+1} \ \ \forall r \implies \langle \hat{\mu}^{(r)}, \mu^* \rangle \geq \langle \hat{\mu}^{(0)}, \mu^* \rangle
$$

Proof. First, we will prove that if $\beta_r \in (0, \frac{\pi}{2})$ and $\|\mu^{(r)}\| \in [c, \frac{4\langle \hat{\mu}_t^{(r)}, \mu_t^* \rangle}{3}]$, then $\alpha_r \in (0, \beta_r)$ for any r. We denote $\alpha_r > 0$ if $\mu^{(r)}$ moves towards $\mu^{(r)\perp}$ and hence towards μ^*. The following simple observation of $\langle \hat{\mu}^{(r)\perp}, \mu^{(r+1)} \rangle \geq 0$ proves that $\alpha_r > 0$.

$$
\begin{aligned}
&\langle \hat{\mu}^{(r)\perp}, \mu^{(r+1)} \rangle \\
&\quad = \mathbb{E}_{x\sim\mathcal{N}(\mu^*,1)}\Big[\eta\Big(\tanh(\mu^{(r)\top}x) - \frac{1}{2}\tanh''(\mu^{(r)\top}x)\|\mu^{(r)}\|^2 + \tanh'(\mu^{(r)\top}x)\mu^{(r)\top}x\Big) \cdot \langle \hat{\mu}^{(r)\perp}, x \rangle\Big] \\
&\quad = \mathbb{E}_{x\sim\mathcal{N}(0,1)}\Big[\eta\Big(\tanh(\mu^{(r)\top}(x+\mu^*)) - \frac{1}{2}\tanh''(\mu^{(r)\top}(x+\mu^*))\|\mu^{(r)}\|^2 \\
&\qquad\quad + \tanh'(\mu^{(r)\top}(x+\mu^*))\mu^{(r)\top}(x+\mu^*)\Big) \cdot \langle \hat{\mu}^{(r)\perp}, (x+\mu^*) \rangle\Big] \\
&\quad = \mathbb{E}_{\alpha_1,\alpha_2\sim\mathcal{N}(\langle \hat{\mu}^{(r)}, \mu^* \rangle, 1)}\Big[\eta\Big(\tanh(\|\mu^{(r)}\|\alpha_1) - \frac{1}{2}\tanh''(\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|^2 \\
&\qquad\quad + \tanh'(\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|\alpha_1\Big)(\alpha_2 + \langle \hat{\mu}^{(r)\perp}, \mu^* \rangle)\Big] \\
&\quad = \mathbb{E}_{\alpha_1,\alpha_2\sim\mathcal{N}(\langle \hat{\mu}^{(r)}, \mu^* \rangle, 1)}\Big[\eta\Big(\tanh(\|\mu^{(r)}\|\alpha_1) - \frac{1}{2}\tanh''(\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|^2 \\
&\qquad\qquad + \tanh'(\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|\alpha_1\Big) \cdot \langle \hat{\mu}^{(r)\perp}, \mu^* \rangle\Big] > 0\,,
\end{aligned}
$$

where in the last step we used the fact that $\langle \hat{\mu}^{(r)}, \mu^* \rangle > 0$ and $\langle \hat{\mu}^{(r)\perp}, \mu^* \rangle > 0$.

Now, we will prove that $\cot\alpha_r > \cot\beta_r$ which will prove that $\alpha_r \in (0, \beta_r)$. Note that

$$\begin{aligned}
\cot\alpha_r &= \frac{\langle \hat{\mu}^{(r)}, \mu^{(r+1)}\rangle}{\langle \hat{\mu}^{(r)\perp}, \mu^{(r+1)}\rangle} \qquad \text{where} \\
\langle \hat{\mu}^{(r)}, \mu^{(r+1)}\rangle &= (1-\eta)\|\mu^{(r)}\| + \eta\mathbb{E}_{\alpha_1\sim\mathcal{N}(\hat{\mu}^{(r)\top}\mu^*,1)}[\tanh(\|\mu^{(r)}\|\alpha_1)\alpha_1] \\
&\quad + \eta\mathbb{E}_{\alpha_1\sim\mathcal{N}(\hat{\mu}^{(r)\top}\mu^*,1)}[-\frac{1}{2}\tanh''(\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|^2\alpha_1 + \tanh'(\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|\alpha_1^2 \\
&\quad - \tanh'(\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|] \\
\langle \hat{\mu}^{(r)\perp}, \mu^{(r+1)}\rangle &= \eta\langle \hat{\mu}^{(r)\perp}, \mu^*\rangle\mathbb{E}_{\alpha_1\sim\mathcal{N}(\hat{\mu}^{(r)\top}\mu^*,1)}[\tanh(\|\mu^{(r)}\|\alpha_1) - \frac{1}{2}\tanh''(\,\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|^2 \\
&\quad + \tanh'(\,\|\mu^{(r)}\|\alpha_1)\|\mu^{(r)}\|\alpha_1] \\
\text{and} \quad \cot\beta_r &= \frac{\langle \hat{\mu}^{(r)}, \mu^*\rangle}{\langle \hat{\mu}^{(r)\perp}, \mu^*\rangle}
\end{aligned}$$

Observe the fact that to prove $\frac{a+c'}{b+c} - \frac{a}{b} > 0$, it is sufficient to prove $c' > \frac{ac}{b}$ for $b, c > 0$. Using this observation, to prove $\cot\alpha_r > \cot\beta_r$, it is sufficient to prove

$$\begin{aligned}
&\Big(1 - \eta - \eta\mathbb{E}[\tanh'(\|\mu^{(r)}\|x)]\Big)\|\mu^{(r)}\| + \eta\mathbb{E}_x\Big[-\frac{1}{2}\tanh''(\|\mu^{(r)}\|x)\|\mu^{(r)}\|^2(x - \langle\hat{\mu}^{(r)}, \mu^*\rangle) \\
&+ \tanh'(\|\mu^{(r)}\|x)(x^2 - \langle\hat{\mu}^{(r)}, \mu^*\rangle x) + \tanh(\|\mu^{(r)}\|x)(x - \langle\hat{\mu}^{(r)}, \mu^*\rangle)\Big] > 0,
\end{aligned}$$

where the expectation is wrt $\mathcal{N}(\langle\mu^{(r)}, \mu^*\rangle, 1)$. Lemma F.3 shows that this is indeed true. □

Lemma F.3. *For any $\eta = \frac{1}{20}$, assuming $a \in [30, \frac{4b}{3}]$, we have*

$$\begin{aligned}
&(1 - \eta - \eta\mathbb{E}_{x\sim\mathcal{N}(b,1)}[\tanh'(ax)])a \\
&\quad + \eta\,\mathbb{E}_{x\sim\mathcal{N}(b,1)}\Big[-\frac{1}{2}\tanh''(ax)a^2(x-b)\tanh'(ax)(x^2 - bx) + \tanh(ax)(x-b)\Big] > 0\,.
\end{aligned}$$

Proof. First, we will find the upper bound on $\mathbb{E}[\tanh''(ax)(x-b)]$.

$$\begin{aligned}
\mathbb{E}[\tanh''(ax)(x-b)] &= \int_{-\infty}^{\infty}\tanh''(ax)(x-b)\exp\Big(-\frac{(x-b)^2}{2}\Big)dx \\
&\le \int_{0}^{b}\tanh''(ax)(x-b)\exp\Big(-\frac{(x-b)^2}{2}\Big)dx \\
&\le \int_{0}^{b}\tanh''(ax)x\exp\Big(-\frac{(x-b)^2}{2}\Big)dx \\
&\le \int_{0}^{b}\exp(-ax)x\exp\Big(-\frac{(x-b)^2}{2}\Big)dx \\
&\le \exp\Big(\frac{a^2-2ab}{2}\Big)\int_{0}^{b}x\exp\Big(-\frac{(x-b)^2 + 2a(x-b) + a^2}{2}\Big)dx \\
&\le \exp(\frac{a^2-2ab}{2})\int_{0}^{\infty}x\Big[\exp\Big(-\frac{(x-b+a)^2}{2}\Big) + \exp\Big(-\frac{(x+b-a)^2}{2}\Big)\Big]dx \\
&\le \exp(-b^2/2) + |a-b|\cdot\exp\Big(\frac{a^2-2ab}{2}\Big)\,.
\end{aligned}$$

Now, for the second term, we have

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(b,1)}[\tanh'(ax)(x^2-bx)] \\
&\quad= \int_{-\infty}^{\infty} \tanh'(ax)x(x-b)\exp\Big(-\frac{(x-b)^2}{2}\Big)dx \\
&\quad\geq -b\int_0^b xe^{-ax}\exp\Big(-\frac{(x-b)^2}{2}\Big)dx \\
&\quad\geq -b\exp\Big(\frac{a^2-2ab}{2}\Big)\int_0^{\infty} x\Big[\exp\Big(-\frac{(x-b+a)^2}{2}\Big)+\exp\Big(-\frac{(x+b-a)^2}{2}\Big)\Big]dx \\
&\quad\geq -b\exp(-b^2/2)-b|a-b|\cdot\exp\Big(\frac{a^2-2ab}{2}\Big)
\end{aligned}$$

We can rewrite the last term as $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\tanh(a(x+b))x]$. Using the fact that $\tanh(a(x+b)) > \tanh(a(-x+b))$, we get that $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\tanh(a(x+b))x] > 0$. Finally, using the upper bound on $\mathbb{E}[\tanh'(ax)]$, we get the following lower bound.

$$\begin{aligned}
&(1-\eta-\eta\,\mathbb{E}_{x\sim\mathcal{N}(b,1)}[\tanh'(ax)])\,a+\eta\mathbb{E}_{x\sim\mathcal{N}(b,1)}\Big[-\frac{1}{2}\tanh''(ax)a^2(x-b)+\tanh'(ax)(x^2-bx)\Big] \\
&\quad\geq \frac{a}{20}(19-4e^{\frac{a^2-2ab}{2}})+\frac{1}{20}\Big(-\frac{a^2}{2}\Big[\exp(-b^2/2)+|a-b|\exp\Big(\frac{a^2-2ab}{2}\Big)\Big] \\
&\qquad\qquad -b\exp(-b^2/2)-b|a-b|\exp(\frac{a^2-2ab}{2})\Big)\geq 1\,. \qquad\square
\end{aligned}$$

Lemma F.4. *For any $a, b > 0$ and $a \in [30, \frac{4b}{3}]$, the following holds. Define*

$$U(a,b)\triangleq \eta\mathbb{E}_{x\sim\mathcal{N}(b,1)}\Big[\Big(\tanh(ax)-\frac{1}{2}\tanh''(ax)a^2+\tanh'(ax)ax\Big)x\Big]-\eta\mathbb{E}_{x\sim\mathcal{N}(b,1)}\left[\tanh'(ax)a\right]-\eta a\,.$$

When the learning rate $\eta = \frac{1}{20}$, is given by, we have

$$\big|U(a,b)\big|\leq\frac{a+b}{10}$$

Proof. We upper bound each term in $U(a,b)$ and they apply triangle inequality to get the result. We start with $|\mathbb{E}_{x\sim\mathcal{N}(b,1)}\left[\tanh''(ax)a^2x\right]|$:

$$\begin{aligned}
-\mathbb{E}_{x\sim\mathcal{N}(b,1)}\left[\tanh''(ax)a^2x\right] &= \frac{a^2}{8\sqrt{2\pi}}\int_0^{\infty} x\sigma(2ax)(1-\sigma(2ax))(2\sigma(2ax)-1)\left(e^{-\frac{(x-b)^2}{2}}+e^{-\frac{(x+b)^2}{2}}\right)dx \\
&\leq \frac{a^2}{4\sqrt{2\pi}}\int_0^{\infty} xe^{-2ax}e^{-\frac{(x-b)^2}{2}}dx \\
&\leq \frac{a^2}{4\sqrt{2\pi}}\int_0^{\infty} e^{-ax}xe^{-\frac{(x-b)^2}{2}}dx \\
&\leq \frac{a^2}{2}e^{-\frac{b^2}{2}}+\frac{a^2}{2}|b-a|\,e^{-\frac{-2a(b-a)-a^2}{2}}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(b,1)}[\tanh'(ax)ax^2] &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty}\tanh'(ax)ax^2\left(e^{-\frac{(x-b)^2}{2}}+e^{-\frac{(x+b)^2}{2}}\right)dx \\
&\leq a\int_0^{\infty} e^{-ax}x^2e^{-\frac{(x-b)^2}{2}}dx \\
&\leq ae^{\frac{a^2-2ab}{2}}\int_0^{\infty} x^2e^{-\frac{(x-b+a)^2}{2}}dx \\
&\leq 2a(a-b)^2e^{\frac{a^2-2ab}{2}}
\end{aligned}$$

$$
\begin{aligned}
-\mathbb{E}_{x\sim\mathcal{N}(b,1)}[a\tanh'(ax)] &= -\frac{a}{\sqrt{2\pi}}\int_0^\infty \tanh'(ax)\left(e^{-\frac{(x-b)^2}{2}}+e^{-\frac{(x+b)^2}{2}}\right)dx \\
&\geq -a\int_0^\infty e^{-ax}e^{-\frac{(x-b)^2}{2}}dx \\
&\geq -ae^{\frac{a^2-2ab}{2}}\int_0^\infty e^{-\frac{(x-b+a)^2}{2}}dx \\
&\geq -4ae^{\frac{a^2-2ab}{2}}.
\end{aligned}
$$

Now, using the fact that $\tanh'(x)$ and $-\tanh''(x)x$ are always positive, we have the following upper bound.

$$
\begin{aligned}
\big|U(a,b)\big| &\leq \eta\left|\mathbb{E}_{x\sim\mathcal{N}(b,1)}\Big[\big(\tanh(ax)-\frac{1}{2}\tanh''(ax)a^2+\tanh'(ax)ax\big)\cdot x\Big]\right| \\
&\quad +\eta|a|+\eta\,\big|-\mathbb{E}_{x\sim\mathcal{N}(b,I)}\left[\tanh'(ax)a\right]\big| \\
&\leq \eta\Big(2b+a+\frac{a^2}{2}e^{-\frac{b^2}{2}}+\frac{a^2}{2}|b-a|\,e^{\frac{-2a(b-a)-a^2}{2}}+2a(b-a)^2e^{\frac{a^2-2ab}{2}}+2ae^{\frac{a^2-2ab}{2}}\Big)
\end{aligned}
$$

If $b\geq a$ and $a\geq 30$, then we have

$$
\big|U(a,b)\big| \leq \eta\,(2b+a+0.1)
$$

If $b\leq a\leq \frac{4b}{3}$ and $a\geq 30$, then

$$
\big|U(a,b)\big| \leq \eta\,(2b+a+0.1)
$$

Using $\eta=1/20$ and for any $a>30$, we have

$$
\big|U(a,b)\big| \leq \frac{a+b}{10}.
$$

□

F.4 Additional proofs for mixtures of two Gaussians

Lemma F.5. *Suppose $a,b>0$ satisfy $a\in[30,\frac{4b}{3}]$, then the following inequality holds:*

$$
|\mathbb{E}_{x\sim\mathcal{N}(b,1)}[-0.5\tanh''(ax)a^2+\tanh'(ax)ax]|\leq 0.01
$$

Proof. We first show that $\mathbb{E}_{x\sim\mathcal{N}(b,1)}[-0.5\tanh''(ax)a^2]>0$ for any $a,b>0$.

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(b,1)}[-0.5\tanh''(ax)a^2] &= -0.5a^2\int_{-\infty}^{\infty}\tanh''(ax)\exp(-0.5(x-b)^2)dx \\
&= -0.5a^2\int_0^\infty \tanh''(ax)(\exp(-0.5(x-b)^2)-\exp(-0.5(x+b)^2))dx>0
\end{aligned}
$$

where the last inequality follows from $\exp(-0.5(x-b)^2)>\exp(-0.5(x+b)^2)$ and $\tanh''(ax)<0$ for $x>0$. We can upper bound $\mathbb{E}_{x\sim\mathcal{N}(b,1)}[-0.5\tanh''(ax)a^2]$ as follows:

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(b,1)}[-\frac{1}{2}\tanh''(ax)a^2] &\leq -\frac{1}{2}a^2\int_0^\infty \tanh''(ax)\exp(-\frac{1}{2}(x-b)^2)dx \\
&\leq a^2\int_0^\infty \exp(-ax)\exp(-\frac{1}{2}(x-b)^2)dx \\
&\leq a^2\exp(\frac{1}{2}(a^2-2ab))\int_0^\infty \exp(-\frac{1}{2}(x-b+a)^2)dx \\
&\leq a^2\exp(\frac{1}{2}(a^2-2ab))
\end{aligned}
$$

When $a\leq b$, by writing $a^2-2ab=-2a(b-a)-a^2\leq -a^2$, we have $\mathbb{E}[-\frac{1}{2}\tanh''(ax)a^2]\leq 0.005$ for $a\geq 30$. When $a\in[b,\frac{4b}{3}]$, $a^2-2ab=\leq -\frac{2b^2}{9}$, we have $|\mathbb{E}[-\frac{1}{2}\tanh''(ax)a^2]|\leq 0.005$. Similar to the $\mathbb{E}_{x\sim\mathcal{N}(b,1)}[-\frac{1}{2}\tanh''(ax)a^2]$, we prove $\mathbb{E}_{x\sim\mathcal{N}(b,1)}[\tanh'(ax)ax]>0$ and $\mathbb{E}_{x\sim\mathcal{N}(b,1)}[\tanh'(ax)ax]<0.005$. Combining bounds for $|\mathbb{E}[\tanh'(ax)ax]|$ and $|\mathbb{E}[-\frac{1}{2}\tanh''(ax)a^2]|$ using triangle inequality, we obtain the result. □

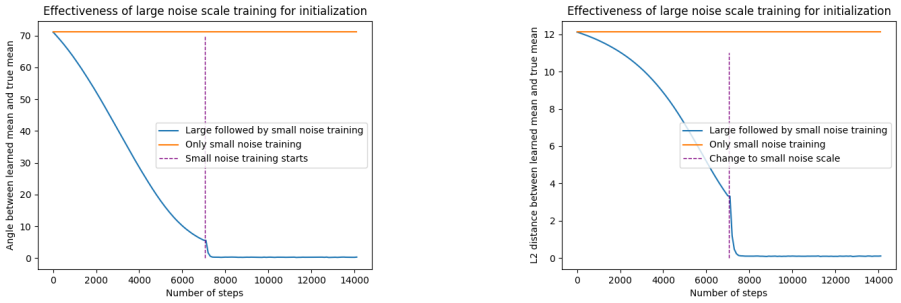
G Experiments

In this section, we perform two sets of experiments to understand the role of large and small noise regimes in the training of mixtures of two Gaussians. Mainly, we want to answer the following questions:

1. Does the large noise regime helps in achieving the warm start required for the small noise regime (as predicted by theory)? *Answer:* Yes
2. Does the large noise scale regime learn the direction of the true mean vector despite having a high amount of noise? *Answer:* Yes

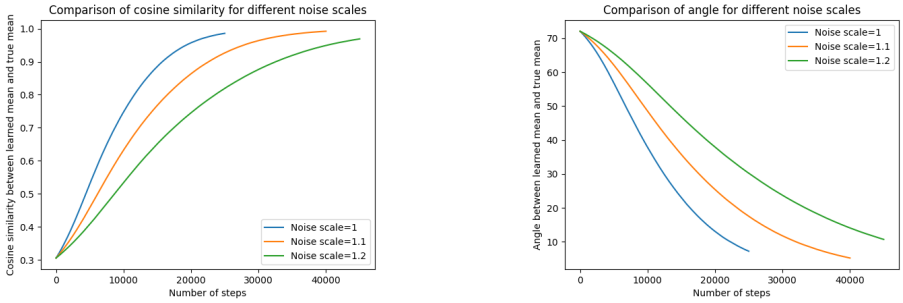
Setup. The task in both experiments is to learn the true parameters of zero-centered mixtures of two Gaussians in 100 dimensions. We use μ^* and $-\mu^*$ to denote the mean vectors of two mixtures. Each element of the μ^* vector is sampled uniformly from $[0, 1]$. We use stochastic gradient descent (SGD) with batch size 128 and learning rate 0.001 for the training. We use $t = 0.01$ for the small noise scale training and $t \in \{1, 1.1, 1.2\}$ for the large noise scale. All results are averaged over 5 independent runs.

Results. To answer the first question, we plot the angle and L_2 distance between the iterate and the ground truth in Figure 1. From the figure, it is evident that the large noise scale training brings the iterate near the ground truth μ^* and then, training with smaller noise scale reduces the L_2 distance quickly. In contrast, only small noise scale training does not make any progress. For the second question, even for large noise scale t, we show that the angle between the learned mean and true mean is decreasing.



(a) Angle between the iterate and the ground truth. (b) L_2 distance between the iterate and the ground truth.

Figure 1: For the blue curve, we initialize randomly, first train in the large t regime for 7000 steps, and then train in the small t regime for 7000 steps. For the orange curve, we initialize randomly and only train in the small t regime for 14000 steps. We see that large t training helps get in a neighborhood of the ground truth, at which point small t training decreases L_2 distance much more quickly, as our theory predicts. In contrast, if we only train with small t, we do not make any noticeable progress.



(a) Cosine similarity between the iterate and the ground truth.

(b) Angle between the iterate and the ground truth.

Figure 2: We show that for some large noise scale, the cosine similarity of learned mean and true mean is increasing (or equivalently, angle is decreasing) as we run for more steps.