The Large and Small Noise Regimes in Diffusion Models

Yi Zhang

2024/10/10

HKU IDS

Shah, K., Chen, S., & Klivans, A. (2023). Learning mixtures of gaussians using the DDPM objective. *Advances in Neural Information Processing Systems*, *36*, 19636-19649.

Outline

- Denoising for Score Matching
  - Preliminary
  - Data settings MoG and network structure
- Large Noise Regime: Connection to power iteration
  - Approximate DDPM loss gradient using Taylor's Expansion
  - One-step DDPM loss (population) gradient decent can be approximated by a matrix by Taylor's Expansion.
- Small Noise Regime: Connection to EM algorithm
  - One-step DDPM loss (population) gradient decent can be (partially) seen as a gradient EM step.
  - The remaining terms are very small.
- A trade-off between learning in large and small noise regime
  - Collapse error induced by determinant sampler
  - Analysis

- Denoising for score matching

Forward diffusion process

Density Evolution

Reverse diffusion process

Score Matching

Denosing

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 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*

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

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.

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

MoG Settings

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.

Two gaussians with symmetric means
and identity covariance matrix

Construct a single-layer “MLP” with Tanh Activation

Calculate Score Function

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

Construct network structure

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

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

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:

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

μ is the learnable parameters

Derivation

$$\tanh(\mu^{*T}x) = w^*_{1,t}(x) + w^*_{2,t}(x)$$

Tanh comes from the summation of the belongingness of x to the 2 Gaussians.

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

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

□

Recall: EM step for MoG Modeling

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*

Latent Variable

M step → $\mu_i' = \dfrac{\mathbb{E}_X[w_i(X)X]}{\mathbb{E}_X[w_i(X)]}$, *where* $w_i(X) = \dfrac{\exp(-\|X - \mu_i\|^2/2)}{\sum_{j=1}^{K} \exp(-\|X - \mu_j\|^2/2)}$. ← E step

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

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.

Derivation from the cited paper DTZ17

Analysis of Population EM for Mixtures of Two Gaussians. To elucidate the optimization features of the algorithm and avoid analytical distractions arising due to sampling error, it has been standard practice in the literature of theoretical analyses of EM to consider the "population version" of the algorithm, where the EM iterations are performed assuming access to infinitely many samples from a distribution p_{μ_1,μ_2} as above. With infinitely many samples, we can identify the mean, $\frac{\mu_1+\mu_2}{2}$, of p_{μ_1,μ_2}, and re-parametrize the density around the mean as follows:

$$p_{\boldsymbol{\mu}}(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \Sigma). \tag{1.1}$$

We first study the convergence of EM when we perform iterations with respect to the parameter $\boldsymbol{\mu}$ of $p_{\boldsymbol{\mu}}(\boldsymbol{x})$ in (1.1). Starting with an initial guess $\boldsymbol{\lambda}^{(0)}$ for the unknown mean vector $\boldsymbol{\mu}$, the t-th iteration of EM amounts to the following update:

$$\boldsymbol{\lambda}^{(t+1)} = M(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}) \triangleq \frac{\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}\boldsymbol{x}\right]}{\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}\right]}, \tag{1.2}$$

where we have compacted both the E- and M-step of EM into one update.

The intuition behind the EM update formula is as follows. First, we take expectations with respect to $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ because we are studying the population version of EM, hence we assume access to infinitely many samples from $p_{\boldsymbol{\mu}}$. For each sample $\boldsymbol{x}$, the ratio $\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}$ is our belief, at step t, that $\boldsymbol{x}$ was sampled from the first Gaussian component of $p_{\boldsymbol{\mu}}$, namely the one for which our current estimate of its mean vector is $\boldsymbol{\lambda}^{(t)}$. (The complementary probability is our present belief that $\boldsymbol{x}$ was sampled from the other Gaussian component.) Given these beliefs for all vectors $\boldsymbol{x}$, the update (1.2) is the result of the M-step of EM. Intuitively, our next guess $\boldsymbol{\lambda}^{(t+1)}$ for the mean vector of the first Gaussian component is a weighted combination over all samples $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ where the weight of every $\boldsymbol{x}$ is our belief that it came from the first Gaussian component.

linearity of expectation for mixture distributions

2 Preliminary Observations

In this section we illustrate some simple properties of the EM update (1.2) and simplify the formul
First, it is easy to see that plugging in the values $\boldsymbol{\lambda} \in \{-\boldsymbol{\mu}, \mathbf{0}, \boldsymbol{\mu}\}$ into $M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ results into

$$M(-\boldsymbol{\mu}, \boldsymbol{\mu}) = -\boldsymbol{\mu} \quad ; \quad M(\mathbf{0}, \boldsymbol{\mu}) = \mathbf{0} \quad ; \quad M(\boldsymbol{\mu}, \boldsymbol{\mu}) = \boldsymbol{\mu}. \tag{2.}$$

In particular, for all $\boldsymbol{\mu}$, these values are certainly fixed points of the EM iteration. Next, we rewri
$M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ as follows:

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\boldsymbol{x}\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(-\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\boldsymbol{x}\right]}{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(-\boldsymbol{\mu},\Sigma)}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)}{p_{\boldsymbol{\lambda}}(\boldsymbol{x})}\right]}.$$

It is easy to observe that by symmetry this simplifies to

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)-\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\boldsymbol{x}\right]}{\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\frac{1}{2}\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\frac{1}{2}\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\right]} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)-\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}{\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)+\mathcal{N}(\boldsymbol{x};-\boldsymbol{\lambda},\Sigma)}\boldsymbol{x}\right].$$

Simplifying common terms in the density functions $\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda},\Sigma)$, we get that

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\frac{\exp\left(\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right) - \exp\left(-\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right)}{\exp\left(\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right) + \exp\left(-\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x}\right)}\boldsymbol{x}\right].$$

We thus get the following expression for the EM iteration

$$M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}\left[\tanh(\boldsymbol{\lambda}^T\Sigma^{-1}\boldsymbol{x})\boldsymbol{x}\right]. \tag{2.}$$

$\tanh(\mu^{*T}x) = w^*_{1,t}(x) + w^*_{2,t}(x)$ Tanh also comes from the summation of the belief (belongingness of x to the 2 Gaussians

Large Noise Regime: Power Iteration

Gradient Approximation

↓

Power Iteration

↓

Angle Contraction

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 $t = O(\log d)$, the gradient descent update on the DDPM objective $L_t(s_t)$ can be approximated with $2\mu_t^*\mu_t^{*\top}\mu_t - 3\|\mu_t\|^2\mu_t$:*

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

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^*:

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

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 $\kappa_2 \le 1/\text{poly}(d)$.

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:

μ_t^* is the first eigenvector, because it is the first eigenvector of Id and $\mu_t^*\mu_t^{*T}$

Calculate Loss Gradient by standard Calculus

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.

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

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

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

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

Take derivative on μ_t and simply expand

Simplify Loss Gradient by Stein's Lemma

Because Z_t and X_t are dependent random variable, analyzing them at the same time is difficult.

After Stein's Lemma, the 2^{th} order and Z_t are canceled

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

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

Observe that $\Big(\tanh(\mu^\top x) - \frac{1}{2}\tanh''(\mu^\top x)\|\mu\|^2 + \tanh'(\mu^\top x)\mu^\top x\Big)x$ and $\tanh'(\mu^\top x)$ are even functions and X_t is a symmetric distribution, therefore, for any even function f, we can write $\mathbb{E}_{X_t}[f(X_t)] = \frac{1}{2}\mathbb{E}_{X_t\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[f(X_t)] + \frac{1}{2}\mathbb{E}_{X_t\sim\mathcal{N}(-\mu_t^*,I)}[f(X_t)] = \mathbb{E}_{X_t\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[f(X_t)]$. 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

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

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

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

I don't know why summation here

□

Approximate loss gradient by Taylor Expansion

Lemma C.3. *For any noise scale $t > t'$ and number of samples $n > n'$ where $t' \lesssim \log d$ and $n' = \Theta(\frac{d^4 B^3}{\varepsilon^2})$, 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*

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

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

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

Stein's Lemma

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

$$\begin{aligned} &\tanh(\mu_t^\top x) = \mu_t^\top x - \frac{2}{3}(\mu_t^\top x)^3 + O(\xi(x)^5) \qquad \text{where } \xi(x) \in [0, \mu_t^\top x] \\ \implies &\tanh(\mu^\top x)x = \mu^\top x x - \frac{2}{3}(\mu_t^\top x)^3 x + O(\xi(x)^5 x) \qquad \text{It does not follow big O definition?} \\ \implies &\left\|\mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}[\tanh(\mu_t^\top x)x] - \mathbb{E}_{x\sim\mathcal{N}(\mu_t^*,\mathrm{Id})}\big[\mu_t^\top x x - \frac{2}{3}(\mu_t^\top x)^3 x\big]\right\| \le \|\mathbb{E}[\xi(x)^5 x]\| \lesssim \sqrt{d}\|\mu_t\|^5 \end{aligned}$$

Cauchy Holder

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

Holds for large noise regime (μ_t^* is small)

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

Holds for large noise regime (x almost follows a standard gaussian)

Why it is 10 and 6.

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

Holds for large noise regime (x almost follows a standard gaussian)

Approximate loss gradient by Taylor Expansion

Sum up all Taylor approximations

Stein's Lemma

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

Gaussian moments

Approximate Taylor approximation by another polynomials

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

Angle Contraction

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

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

$$\kappa_2 = \frac{500\eta\sqrt{d^3}\|\mu_t\|^4 + 20\eta d\|\mu_t\|^2\|\mu_t^*\|^2 + \eta\tilde{\varepsilon}}{\|\mu_t^*\|^2} \quad \textit{and} \quad \tilde{\varepsilon} \lesssim \frac{d\varepsilon}{\|\mu_t\|} \ .$$

Angle Contraction

This because in power iteration, μ', μ, μ^* lie in the same plane.

Because $Id + F$ is symmetric, the eigenvectors are orthogonal to each other. Therefore, if μ^* is the first eigenvector, $(\mathrm{Id} + \mathrm{F})\mu^{*T} \leq \sigma_2 \mu^{*T}$

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

$$\tan\theta' = \frac{\langle \hat{\mu}^{*\perp}, \hat{\mu}_t' \rangle}{\langle \hat{\mu}_t^*, \hat{\mu}_t' \rangle} = \frac{\langle \hat{\mu}_t^{*\perp}, \mu_t' \rangle}{\langle \hat{\mu}_t^*, \mu_t' \rangle}$$

$$= \frac{\langle \hat{\mu}_t^{*\perp}, \mu_t + \eta F(\mu_t, \mu_t^*) \rangle + \langle \hat{\mu}_t^{*\perp}, -\eta \nabla L_t(s_t) - \eta F(\mu_t, \mu_t^*) \rangle + \eta\varepsilon}{\langle \hat{\mu}_t^*, \mu_t + \eta F(\mu_t, \mu_t^*) \rangle + \langle \hat{\mu}_t^{*\perp}, -\eta \nabla L_t(s_t) - \eta F(\mu_t, \mu_t^*) \rangle - \eta\varepsilon}$$

where $F(\mu, \mu^*) = \left(2\mu_t^* \mu_t^{*\top} \mu_t - 3\|\mu_t\|^2 \mu_t\right)$

$$\leq \frac{\sigma_2 \langle \hat{\mu}_t^{*\perp}, \mu_t \rangle + \eta \left\| \nabla L_t(s_t) + F(\mu_t, \mu_t^*) \right\| + \eta\varepsilon}{\sigma_1 \langle \hat{\mu}_t^*, \mu_t \rangle - \eta \left\| \nabla L_t(s_t) + F(\mu_t, \mu_t^*) \right\| - \eta\varepsilon} \tag{C.1}$$

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

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

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

Angle Contraction

This is because of assumption that

$$\mu_t \ \textit{satisfies} \ |\langle \hat{\mu}_t, \hat{\mu}_t^* \rangle| \geq \frac{1}{2d},$$

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

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

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

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

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

Low Noise Regime: EM algorithm

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

One-step of EM

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

It remains to show G contracts.
Intuition: The first and high order derivative of Tanh decays exponentially with μ.
And different from the large noise regime, μ is large in small noise regime.

Contraction of 1-D $G(\mu, \mu^*)$

In one 1-D case, G contracts.

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*

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

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.

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

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

Mean Value Thm

Exponentially decay with a (μ)

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

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

Conclusion

- If we perfectly build a network to fit MoG score,
  - In large noise regime (small μ_t), the one-step DDPM loss gradient decent can be regarded as power iteration. (angle contraction)
    - Proved by Taylor's expansions and remaining terms are high orders of μ_t.
  - In small noise regime (large μ_t), the gradient decent can be regarded as gradient EM algorithm. (amplitude contraction)
    - Proved by derivatives of tanh decays exponentially with μ_t