# GLOBAL CONVERGENCE OF GRADIENT DESCENT FOR SCORE MATCHING IN GAUSSIAN MIXTURES VIA REVERSE FISHER DIVERGENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The score matching problem is a central training objective in modern generative modeling, diffusion models, fitting unnormalized statistical models, and inverse problems. A standard approach is to minimize the *forward Fisher divergence*, where the expectation is taken with respect to the teacher distribution. However, recent results show that even in simple Gaussian mixture model settings, this objective can lead to undesirable and initialization-dependent convergence behavior. In this paper, we study an alternative objective: the *reverse Fisher divergence*, where the expectation is taken with respect to the student distribution. We analyze gradient descent (GD) for fitting Gaussian mixture models and show that this change in the objective leads to significantly better optimization properties. First, when the teacher distribution is a single Gaussian and the student is a Gaussian mixture model with fixed weights and identity covariances, we prove the global convergence of GD from *arbitrary initializations*. Second, we extend the analysis to the case where the teacher is also a Gaussian mixture model and prove global convergence guarantees under a global random initialization scheme and a $\widetilde{\Omega}(1)$-separation assumption on the target means. In particular, with high probability, each student component converges near its closest teacher component, and we provide conditions under which the student distribution converges in total variation distance. Our proofs rely on a new Lyapunov-based analysis of the gradient descent dynamics, showing that the *reverse Fisher divergence* has a much more favorable optimization landscape than the *forward Fisher divergence*.

## 1 INTRODUCTION

The score matching problem has become a central training objective in many applications, including modern generative modeling, diffusion models, fitting unnormalized statistical models, and inverse problems (Hyvärinen & Dayan, 2005; Song & Ermon, 2019; Chung et al., 2022). In this task, we have two distributions, $p_\mu(x)$ and $\bar{p}(x)$, called the student and the teacher, respectively, where the former is parameterized by $\mu$. The main goal is to find $\mu$ such that the score function $\nabla_x \log p_\mu(x)$ of the student is close to the score function $\nabla_x \log \bar{p}(x)$ of the teacher. Matching scores rather than densities directly is appealing, for instance, when the target density, or both densities, are known only up to normalizing constants, since the score functions do not depend on these constants.

Empirically, score-based methods have been highly successful and have led to a broad range of applications in image generation, text-to-image generation, and audio synthesis (Rombach et al., 2022; Kong et al., 2020; Ramesh et al., 2022). However, the *theoretical* properties of the score matching problem remain much less understood. *How should we match the score functions? How can we guarantee the convergence of algorithms? What theoretical convergence rates can we achieve?*

In both theory and practice, a typical approach to matching scores is to minimize the *forward Fisher divergence*, $\mathcal{L}_{\text{for}}(\mu) = \mathbb{E}_{x\sim\bar{p}}\left[\|\nabla_x \log p_\mu(x) - \nabla_x \log \bar{p}(x)\|^2\right]$, where the expectation is taken with respect to the teacher distribution. This objective is minimized using the gradient descent (GD) method $\mu^{k+1} = \mu^k - \gamma\nabla\mathcal{L}_{\text{for}}(\mu^k)$ (in practice, Adam (Kingma & Ba, 2014) and other adaptive stochastic methods are often used, whereas in this paper we consider GD for the theoretical analysis).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

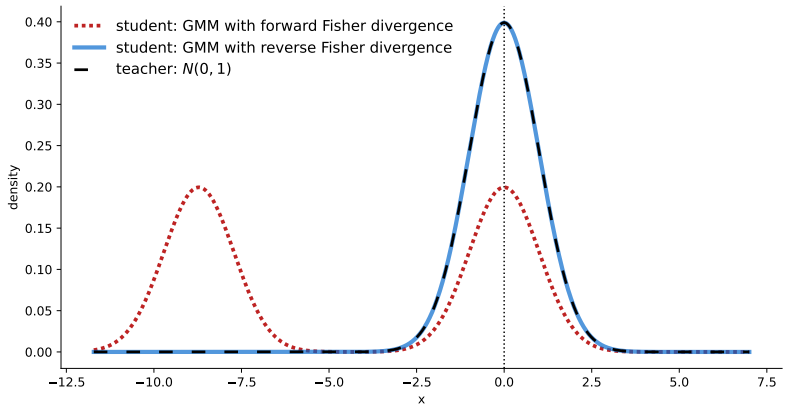



Figure 1: A result of running gradient descent (GD) on the *forward* and *reverse* Fisher divergences. The teacher/target distribution is a single Gaussian, while the student/model is a Gaussian mixture model with two components (see the setup in Sections 1.3 and 4). Even in this simple setting, GD on the *forward Fisher divergence* converges to an incorrect distribution, whereas GD on the *reverse Fisher divergence* almost perfectly recovers the teacher distribution.

### 1.1 Related work

Many theoretical works on diffusion and score-based models analyze the sampling process when an exact or sufficiently accurate score function is available (Chen et al., 2022; Lee et al., 2023; Benton et al., 2024; Gupta et al., 2025). These results provide important guarantees for the generative process, but they do not fully explain how and when gradient-based training can learn the score function. Another line of work studies the statistical and optimization properties of score estimation, including generalization and statistical complexity (Koehler et al., 2022; Li et al., 2023; Wibisono et al., 2024). Han et al. (2024) analyzed the convergence of gradient-based methods using the neural tangent kernel (NTK) approach. Wang et al. (2024) analyzed GD on a finite-sum empirical approximation of the score-matching objective under an overparameterized neural network. Gatmiry et al. (2024); Chen et al. (2024) provide statistical and computational guarantees for recovering Gaussian mixtures. In contrast, our study belongs to the optimization side of this literature: we analyze the behavior and properties of GD dynamics for the score-matching problem.

**Score Matching with Gaussian Mixture Models (GMMs).** Analyzing the convergence of GD applied to the forward Fisher divergence for arbitrary distributions is nontrivial. Recent studies focus on score matching problems where the student and teacher are *Gaussian mixture models* (GMMs), but even this setting leads to a complex analysis because the resulting optimization problem is non-convex. Before we discuss the most relevant works, it is also worth mentioning that fitting GMMs has been extensively studied in the context of the Expectation-Maximization (EM) and gradient EM algorithms (Xu et al., 2016; Daskalakis et al., 2017; Dwivedi et al., 2020; Wu & Zhou, 2021; Xu et al., 2024; Zhou et al., 2025); there are several parallels with score matching for GMMs, and some EM-type proof techniques are useful in this context.

Shah et al. (2023) considered the forward Fisher divergence with GD for fitting the mean parameters of a mixture of two symmetric Gaussians with constant separation. They study the DDPM/forward Fisher objective, where one can choose the noise level, and use this structure in a two-phase algorithm: first fitting the parameters in the large-noise regime and then refining them in the low-noise regime. Their random-initialization analysis is limited to the setting with two components. They also consider the problem of fitting GMMs with multiple components, but they achieve convergence only under the assumption that the initial student means are close to the ground truth. Zhang et al. (2026) analyzed the case where the teacher is a single Gaussian and the student is a GMM with multiple components, motivated by the fact that real models are often overparameterized and the target number of components may be unknown. Their theoretical results show that even in this simple setting, the forward Fisher divergence *exhibits undesirable and unstable convergence behavior*: depending on the noise level and initialization, GD may converge to the ground truth, become stuck at

a local maximum with nonzero loss, or drive only one student component toward the teacher while the remaining components diverge.

### 1.2 Motivation and Contributions

Zhang et al. (2026) proved that the *forward Fisher divergence* can lead to subtle and initialization-dependent optimization behavior. Ideally, a loss function should possess the following properties: *absence of local maxima or spurious stationary points* and *global convergence of GD*. The forward Fisher divergence clearly does *not* satisfy these properties, even in the simple setting of fitting the score function of a GMM to that of a single Gaussian, which motivates the search for alternative objectives with more favorable optimization landscapes.

♠ In this paper, we study an alternative objective: the *reverse Fisher divergence*,

$$\mathcal{L}_{\mathrm{rev}}(\mu) = \mathbb{E}_{x\sim p_\mu}\left[\left\|\nabla_x \log p_\mu(x) - \nabla_x \log \bar{p}(x)\right\|^2\right], \tag{1}$$

where the expectation is taken with respect to the student distribution instead of the teacher distribution. Although the difference between the forward Fisher divergence and (1) is only the distribution under the expectation, we show that this change leads to significantly better optimization properties.

♣ (**One target component**). We consider the *reverse Fisher divergence* and prove the convergence of GD for score matching a student GMM distribution to a single Gaussian distribution from *arbitrary initialization* (Theorem 2.1). In contrast to the previous results for the *forward Fisher divergence*, which study the same score-matching setting for GMMs, our analysis establishes *global convergence*. Moreover, the convergence rate is dimension-independent when $d$ is sufficiently large and depends only logarithmically on the initial distance to the target mean. These results provide strong theoretical evidence that the reverse Fisher divergence possesses a significantly more favorable optimization landscape and stronger convergence guarantees than the forward Fisher divergence.

♦ (**Multiple target components**). In Theorem 3.2, we extend this result to the case where the teacher is also a GMM. For arbitrary numbers $n$ and $m$ of student and teacher components, we establish *global convergence guarantees* for the student mean parameters to the target means under random initialization[1] and a $\tilde{\Omega}(1)$-separation assumption on the target means. In particular, we uniformly sample the initial means on a sufficiently large sphere of radius $R^0$ and show that, with high probability, each student mean converges to its closest teacher mean up to $\varepsilon$-accuracy. When the dimension satisfies $d \geq n$, the rate is $\tilde{\Theta}\left(n^2/\varepsilon\right)$, which depends only logarithmically on $R^0$ and is independent of $d$. In Section 3.2, we also analyze conditions under which the student distribution converges in total variation distance.

♥ (**Proof techniques**). In this work, we face mathematical difficulties similar to those in related work: the problem is nonstandard, with a nontrivial landscape and highly complex gradients. The fact that GD with the reverse Fisher divergence converges globally is somewhat surprising. To prove this, in Section 2, we develop a new proof technique. The main idea is to show that the method is equivalent to $\mathbf{Z}^{k+1} = \left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}^k\right)\mathbf{Z}^k$ and to prove that the Lyapunov function $\|\mathbf{Z}^k\|_F^2$ is non-increasing and converges at a sublinear rate. We show that $\mathbf{B}^k$ is symmetric, positive semidefinite, and has bounded norm. Most importantly, it satisfies $\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\rangle \geq \Omega(\min\{\|\mathbf{Z}^k\|_F^4, \|\mathbf{Z}^k\|_F^2\})$. In Section 3.1, we present new proof techniques to show convergence in the case of multiple target components, where we provide a two-stage analysis approach with careful control of errors.

### 1.3 Preliminaries

We consider the score matching problem, where $\bar{p}(x)$ and $p_\mu(x)$ denote the teacher and student distributions, respectively. Our goal is to find parameters $\mu$ such that the score function of the student distribution, $\nabla_x \log p_\mu(x)$, is as close as possible to the score function of the teacher distribution, $\nabla_x \log \bar{p}(x)$. Unlike most previous works, we consider the problem of minimizing the *reverse Fisher divergence* (1). In our setting, we follow (Shah et al., 2023; Zhang et al., 2026) and assume that $\bar{p}$ and $p_\mu$ are GMMs of the form

$$p_\mu(x) = \frac{1}{n}\sum_{i=1}^{n}\varphi(x-\mu_i) \quad \text{and} \quad \bar{p}(x) = \frac{1}{m}\sum_{a=1}^{m}\varphi(x-\bar{\mu}_a) \quad \forall x \in \mathbb{R}^d, \tag{2}$$

[1] As expected, convergence is not possible under deterministic initialization; see Theorem 3.1.

where $\varphi(x) := (2\pi)^{-d/2} \exp\left(-\|x\|^2/2\right)$ is the standard normal probability density function. Here, the target means $\bar{\mu}_1, \ldots, \bar{\mu}_m \in \mathbb{R}^d$ and the number of components $m \geq 1$ are fixed but unknown parameters, and $n \geq 1$ is the number of Gaussian components in the student distribution, parameterized by $\mu \equiv [\mu_1; \ldots; \mu_n] \in \mathbb{R}^{nd}$.

**Gradient descent.** To minimize (1), we consider the gradient descent (GD) method:

$$\mu^{k+1} = \mu^k - \gamma \nabla_\mu \mathcal{L}_{\text{rev}}(\mu^k), \tag{3}$$

where $\gamma > 0$ is the step size, $\mu^k \equiv [\mu_1^k; \ldots; \mu_n^k]$, and $\{\mu_i^0\}_{i\in[n]}$ are the initial means. Before we derive the gradient, it is convenient to introduce the following notation. Define $r_i(x) := \frac{\exp\left(-\frac{1}{2}\|x-\mu_i\|^2\right)}{\sum_{j=1}^n \exp\left(-\frac{1}{2}\|x-\mu_j\|^2\right)}$, $m_\mu(x) := \sum_{i=1}^n r_i(x)\mu_i$, and $\mathbf{C}_\mu(x) := \sum_{i=1}^n r_i(x)\big(\mu_i - m_\mu(x)\big)\big(\mu_i - m_\mu(x)\big)^\top$. Similarly, $\bar{r}_a(x) := \frac{\exp\left(-\frac{1}{2}\|x-\bar{\mu}_a\|^2\right)}{\sum_{b=1}^m \exp\left(-\frac{1}{2}\|x-\bar{\mu}_b\|^2\right)}$, $\bar{m}(x) := \sum_{a=1}^m \bar{r}_a(x)\bar{\mu}_a$, and $\bar{\mathbf{C}}(x) := \sum_{a=1}^m \bar{r}_a(x)\big(\bar{\mu}_a - \bar{m}(x)\big)\big(\bar{\mu}_a - \bar{m}(x)\big)^\top$. In Section C, we derive the gradient of $\mathcal{L}_{\text{rev}}$.

**Proposition 1.1** (Gradient of (1) with GMMs). *Let $\nabla_\mu \mathcal{L}_{\text{rev}}(\mu) \equiv [\nabla_{\mu_1}\mathcal{L}_{\text{rev}}(\mu); \ldots; \nabla_{\mu_n}\mathcal{L}_{\text{rev}}(\mu)]$. Then, for all $\mu \in \mathbb{R}^{nd}$ and $\ell \in [n]$, the gradient with respect to the $\ell^{th}$ mean component is*

$$\nabla_{\mu_\ell}\mathcal{L}_{\text{rev}}(\mu) = \frac{2}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[m_\mu(x) - \bar{m}(x) + \mathbf{C}_\mu(x)\big(\mu_\ell - m_\mu(x)\big) - \bar{\mathbf{C}}(x)\big(\mu_\ell - \bar{m}(x)\big)\right].$$

## 2 Learning Single Gaussian Component

We present our first theoretical result for the case where the teacher distribution consists of a single component, i.e., $m = 1$. Equivalently, the teacher distribution is a Gaussian distribution. We consider the setting with multiple Gaussian components in the teacher in Section 3. Unlike the case of the forward Fisher divergence, where it is infeasible to establish global convergence guarantees (Zhang et al., 2026), we establish the *global convergence* of GD for the reverse Fisher divergence from an *arbitrary initialization* to the target mean $\bar{\mu}_1$ of the Gaussian distribution $\bar{p}$.

**Theorem 2.1** (Global Convergence of GD). *If $\gamma \leq n/(4+512\log(n))$, then for all $\mu_1^0, \ldots, \mu_n^0 \in \mathbb{R}^d$, GD (3) with GMMs (2) and $m = 1$ guarantees that*

$$\left\|\mu_i^k - \bar{\mu}_1\right\|^2 \leq \max\left\{\left(1 - \tfrac{\gamma}{2n}\right)^{k/2} n(R^0)^2, \tfrac{8n^2\min\{n,d\}}{\gamma k}\right\}$$

*for all $i \in [n], k \geq 1$, where $R^0 := \max_{i\in[n]}\left\|\mu_i^0 - \bar{\mu}_1\right\|$.*

We now present the proof of this result and highlight some interesting properties of the problem that enable us to establish global convergence.

*Proof.* We first rewrite the update (3) minus $\bar{\mu}_1$ in matrix form to make the proof more compact and intuitive. We defer the derivations of Lemma 2.2 to Section D.

**Lemma 2.2.** *Let $m = 1$ and $\mathbf{Z}^k := \left[(\mu_1^k - \bar{\mu}_1)^\top; \ldots; (\mu_n^k - \bar{\mu}_1)^\top\right] \in \mathbb{R}^{n\times d}$, stacking the vectors vertically. Using this notation and (3), the matrix sequence $\{\mathbf{Z}^k\}_{k\geq 0}$ evolves as*

$$\mathbf{Z}^{k+1} = \left(\mathbf{I} - \tfrac{2\gamma}{n}\mathbf{B}^k\right)\mathbf{Z}^k,$$

*where $\mathbf{B}^k \in \mathbb{R}^{n\times n}$ and $\mathbf{B}_{\ell i}^k := \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[r_i^k(x) + r_i^k(x)\big(\mu_i^k - m_{\mu^k}(x)\big)^\top\big(\mu_\ell^k - m_{\mu^k}(x)\big)\right]$ is the $(\ell,i)^{th}$ entry of $\mathbf{B}^k$ and $r_i^k(x) := \exp\left(-\frac{1}{2}\|x-\mu_i^k\|^2\right) / \sum_{j=1}^n \exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)$.*

Using Lemma 2.2, we can write the distance between the student and teacher means in matrix form:

$$\left\|\mathbf{Z}^{k+1}\right\|_F^2 = \left\|\left(\mathbf{I} - \tfrac{2\gamma}{n}\mathbf{B}^k\right)\mathbf{Z}^k\right\|_F^2 = \left\|\mathbf{Z}^k\right\|_F^2 - \tfrac{4\gamma}{n}\left\langle\mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle + \tfrac{4\gamma^2}{n^2}\left\|\mathbf{B}^k\mathbf{Z}^k\right\|_F^2, \tag{4}$$

where $\|\cdot\|_F$ is the standard Frobenius norm and $\langle\cdot,\cdot\rangle$ is the standard matrix inner product. The matrix $\mathbf{B}^k$ has the following useful properties, proved in Section D:

**Lemma 2.3.** *For all $k \geq 0$, matrix $\mathbf{B}^k$ is symmetric and positive semidefinite, and $\left\|\mathbf{B}^k\right\|_{\text{op}} \leq 1 + 128\log(n)$.*

Due to Lemma 2.3, (4), and standard inequalities, $\left\|\mathbf{B}^k\mathbf{Z}^k\right\|_F^2 \leq \left\|\mathbf{B}^k\right\|_{\text{op}} \left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k \right\rangle$ and

$$\left\|\mathbf{Z}^{k+1}\right\|_F^2 \leq \left\|\mathbf{Z}^k\right\|_F^2 - \tfrac{4\gamma}{n} \left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k \right\rangle + \tfrac{(4+512\log(n))\gamma^2}{n^2} \left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k \right\rangle.$$

Since $\gamma \leq n/(4+512\log(n))$ and $\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k \right\rangle \geq 0$, we get

$$\left\|\mathbf{Z}^{k+1}\right\|_F^2 \leq \left\|\mathbf{Z}^k\right\|_F^2 - \tfrac{\gamma}{n} \left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k \right\rangle. \tag{5}$$

We conclude that $\{\|\mathbf{Z}^k\|_F^2\}_{k\geq 0}$ is non-increasing. However, this is not sufficient to prove the convergence rate, and we still need one more property of $\mathbf{B}^k$ and $\mathbf{Z}^k$, proved in Section D:

**Lemma 2.4.** *For all $k \geq 0$, $\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k \right\rangle \geq \min\left\{ \frac{1}{4n\min\{n,d\}} \left\|\mathbf{Z}^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}^k\right\|_F^2 \right\}$.*

Substituting this inequality into (5),

$$\left\|\mathbf{Z}^{k+1}\right\|_F^2 \leq \left\|\mathbf{Z}^k\right\|_F^2 - \tfrac{\gamma}{n} \min\left\{ \tfrac{1}{4n\min\{n,d\}} \left\|\mathbf{Z}^k\right\|_F^4, \tfrac{1}{2}\left\|\mathbf{Z}^k\right\|_F^2 \right\},$$

and it remains to use Lemma B.11 and $n \max\limits_{i\in[n]} \left\|\mu_i^k - \bar{\mu}_1\right\|^2 \geq \left\|\mathbf{Z}^k\right\|_F^2 \geq \max\limits_{i\in[n]} \left\|\mu_i^k - \bar{\mu}_1\right\|^2$. □

**Corollary 2.5.** *In view of Theorem 2.1, with $\gamma = n/(4+512\log(n))$, GD finds $\{\mu_i^k\}_{i\in[n]}$ such that $\left\|\mu_i^k - \bar{\mu}_1\right\|^2 \leq \varepsilon$ for all $i \in [n]$ and $\mathcal{L}_{\text{rev}}(\mu^k) \leq \varepsilon$ after at most*

$$4096 \max\left\{ (1+\log(n)) \log\left(\tfrac{n(R^0)^2}{\varepsilon}\right), \tfrac{(1+\log(n))n\min\{n,d\}}{\varepsilon} \right\} \tag{6}$$

*iterations.*

**Discussion.** Theorem 2.1 and Corollary 2.5 provide global convergence guarantees for GD with the reverse Fisher divergence in the score-matching problem. This highlights the importance of this loss compared to the forward Fisher divergence, since the latter does not guarantee convergence in the same setting. With the forward Fisher divergence, one can guarantee convergence only when the initial means are close to the target mean (the large-noise regime) (Zhang et al., 2026), whereas the reverse Fisher divergence guarantees convergence in all regimes. Moreover, all $\{\mu_i^k\}_{i\in[n]}$ converge to $\bar{\mu}_1$, which is an important property in the score matching problem.

Intuitively, the reverse Fisher divergence, due to its averaging with respect to $p_\mu$, emphasizes the regions where $p_\mu$ places mass and tends to move such mass towards the target distribution $\bar{p}$, thereby yielding the global convergence of GD. On the other hand, the forward Fisher divergence places little weight on regions where $\bar{p}$ has small mass, and therefore may not penalize far-away modes of $p_\mu$; in some settings, it can even push them away, as formalized in (Zhang et al., 2026).

The convergence rate (6) has many nice properties: first, it does not depend on $d$ when $d$ is large; second, it has a logarithmic dependence on the initial distance $R^0$; and third, the dependence on $\varepsilon$ is $\tilde{\mathcal{O}}\left(1/\varepsilon\right)$. The proof technique is arguably compact and intuitive (which we believe is a major advantage of this theory): we use the Lyapunov function $\|\mathbf{Z}^k\|_F^2$ and show its monotonicity. The only technical parts are Lemmas 2.3 and 2.4, which establish the properties of the matrix $\mathbf{B}^k$. We show that $\mathbf{B}^k$ is symmetric and positive semidefinite, and $\|\mathbf{B}^k\|_{\text{op}} \leq \tilde{\mathcal{O}}(1)$. Moreover, in Lemma 2.4, we prove that $\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k \rangle$ is larger than zero, with the gap $\Omega(\min\{\|\mathbf{Z}^k\|_F^4, \|\mathbf{Z}^k\|_F^2\})$, which allows us to obtain the final two-regime convergence rate.

## 3 Learning Multiple Gaussian Components

We are ready to extend our result to the setting with multiple Gaussian components in the teacher distribution, i.e., $m > 1$ in $\bar{p}$. We will show that global convergence up to a small $\varepsilon$-accuracy is possible even in this setting. However, as expected, it can only be guaranteed with high probability

and under a (non-local) random initialization. Indeed, before presenting our main result, let us first demonstrate a simple counterexample with a "bad" initialization.

**Theorem 3.1** (Folklore result). *For $m = 2$ and $n \geq 1$, let $\bar{\mu}_1 = -1$, $\bar{\mu}_2 = 1$, and $\mu_i^0 = 0$ for all $i \in [n]$. Then $\mu^0$ is a stationary point of both the forward and reverse Fisher divergences. Consequently, GD initialized at $\mu^0$ does not move for all $\gamma > 0$, i.e., $\mu^k = \mu^0$ for all $k \geq 0$, and both losses remain unchanged.*

This result and counterexample likely extend to other "reasonable" losses. Nevertheless, we can show that this example is unstable and establish global convergence under random initialization with the reverse Fisher divergence.

**Theorem 3.2** (Theorem E.1 when $d \geq n$). *Let $\varepsilon \in (0, 1]$ denote the target accuracy. We define the smallest and largest gaps in the target distribution as $\delta_{\min} := \min_{a \neq b} \|\bar{\mu}_a - \bar{\mu}_b\| > 0$ and $\delta_{\max} := \max_{a \neq b} \|\bar{\mu}_a - \bar{\mu}_b\|$, choose the initialization radius as any $R^0 \geq \tilde{\Theta}\left(\max\left\{\frac{nm^2\sqrt{d}\delta_{\max}^2}{\rho}, \sqrt{n}\bar{R}\right\}\right)$, where $\bar{R} := \max_{b \in [m]} \|\bar{\mu}_b\|$ and $\rho \in (0, 1]$. We sample the initial parameters as $\mu_\ell^0 = R^0 \eta_\ell / \|\eta_\ell\|$, where $\eta_\ell \sim \mathcal{N}(0, \mathbf{I})$ for all $\ell \in [n]$ (i.i.d.). Assume that $\delta_{\min} \geq \tilde{\Omega}(1)$ and $d \geq \tilde{\Omega}(n)$. If $\gamma \leq \tilde{\Theta}(n)$ and GD (3) with GMMs (2) is run for*

$$\tilde{K} := \tilde{\Theta}\left(\frac{n^3}{\gamma\varepsilon}\right)$$

*iterations, then, with probability at least $1 - 4\rho$, we have that $\left\|\mu_\ell^k - \bar{\mu}_{a_\ell^*}\right\|^2 \leq \varepsilon$, for all $\ell \in [n]$ and $k \geq \tilde{K}$, where $a_\ell^* = \arg\min_{a \in [m]} \left\|\mu_\ell^0 - \bar{\mu}_a\right\|$ is the index of the target mode closest to the initialization point $\mu_\ell^0$. In particular, $\tilde{K} = \tilde{\Theta}\left(\frac{n^2}{\varepsilon}\right)$ if $\gamma = \tilde{\Theta}(n)$.*

**Discussion.** This theorem states that, if the initial means $\{\mu_\ell^0\}_{\ell \in [n]}$ are sampled uniformly from the sphere of radius $R^0$, then, for every $\ell \in [n]$, the sequence $\{\mu_\ell^k\}_{k \geq 0}$ converges to a small-$\varepsilon$ neighborhood of the target mean $\bar{\mu}_{a_\ell^*}$, where $\bar{\mu}_{a_\ell^*}$ is the target mean closest to $\mu_\ell^0$. Notice that the initial means are sampled uniformly from the sphere of large radius $R^0$, and that the initialization algorithm does not depend on $\{\bar{\mu}_a\}_{a \in [m]}$. Nevertheless, under the random initialization scheme, we obtain global convergence of the means to small neighborhoods of their closest target modes. The fact that $R^0$ is large is not a problem, since the dependence on $R^0$ is only logarithmic. Similarly to the single-target case, up to logarithmic factors, the convergence rate does not depend on $d$ when $d$ is large, and the dependence on $\varepsilon$ is $\tilde{\mathcal{O}}(1/\varepsilon)$.

The theorem relies on two assumptions: $d \geq \tilde{\Omega}(n)$ and $\delta_{\min} \geq \tilde{\Omega}(1)$. The former is arguably mild; moreover, in Theorem E.1 we only require $d \geq \tilde{\Omega}(1)$ and include $d \geq \tilde{\Omega}(n)$ here only for simplicity. The latter assumption $\delta_{\min} \geq \tilde{\Omega}(1)$ requires a logarithmic separation between the target means and is standard in the literature (Shah et al., 2023; Zhou et al., 2025); analyzing GD using the reverse Fisher divergence under a near-zero separation is an important direction for future work.

*Remark.* The full dependencies on all the parameters are presented in Theorem E.1. Notice the requirement on $\delta_{\min}$ with explicit logarithmic dependencies when $d \geq \tilde{\Omega}(n)$ is $\delta_{\min} \geq \Omega\left(\sqrt{\log\left(n^3 \max\{n, m\} (R^0)^4/\varepsilon^2\right)}\sqrt{\log(2d/\rho)}\right)$. The requirement depends on $\varepsilon$, which is expected since, intuitively, when $\mu_\ell^k$ iterates towards $\bar{\mu}_{a_\ell^*}$, all other target means $\{\bar{\mu}_b\}_{b \neq a_\ell^*}$ also "attract" $\mu_\ell^k$, and in reality it converges to some point not belonging to the set $\{\bar{\mu}_b\}_{b \in [m]}$. However, when $\delta_{\min} \geq \tilde{\Omega}(1)$, this point is extremely close to one of the targets. A similar observation about $\delta_{\min}$ and $\varepsilon$ appears in (Shah et al., 2023), where the authors obtain convergence under only *local initialization* close to the targets, whereas our approach allows *global randomized initialization*.

### 3.1 Proof technique

While the proof techniques of Theorem 3.2 use ideas from Section 2, the proof as a whole is much more challenging and technical. Unlike Theorem 3.2, we divide the proof into two stages.

**Global convergence stage.** The fact that we initialize on the sphere with a large radius $R^0$ helps us significantly, since the initial means $\mu_i^0$ are far away from each other with high probability (w.h.p.). Since they are far apart, we can show that $\nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k) \approx \frac{2}{n}\left(\mu_\ell^k - \bar{\mu}_{a_\ell^*}\right)$. Thus, $\mu_\ell^{k+1} \approx \bar{\mu}_{a_\ell^*} + (1 - 2\gamma/n)^{k+1}\left(\mu_\ell^0 - \bar{\mu}_{a_\ell^*}\right)$ during the first iterations, and each $\mu_\ell^k$ converges linearly towards the corresponding closest target mean. However, these approximations contain errors that require careful control: we derive auxiliary lemmas to control how these errors evolve and show that they do not introduce significant fluctuations, and that each $\mu_\ell^k$ indeed converges linearly towards its target.

**Local convergence stage.** After $\bar{K}$ iterations, each $\mu_\ell^{\bar{K}}$ is close to its target mean $\bar{\mu}_{a_\ell^*}$. This is the final stage, where the gradients of $\mu_\ell^{\bar{K}}$ corresponding to the same group $I_a := \{i \in [n] : a_i^* = a\}$ start "feeling" each other, and we can no longer ignore the fact that the means from $I_a$ are close to each other. In this stage, for each group $I_a$, we use almost the same analysis as in Section 2. However, the main difficulty here is to show that $\mu_i^k$ and $\mu_j^k$ corresponding to different groups, i.e., $i \in I_a$, $j \in I_b$, and $a \neq b$, do not "interfere with each other significantly," and that the means corresponding to $I_a$ converge towards $\bar{\mu}_a$ as if there were no other target means $\{\bar{\mu}_b\}_{b \neq a}$ and $\{\mu_i^k\}_{i \notin I_a}$, which also requires careful control of errors.

### 3.2 Convergence of the total variation distance

In general, Theorem 3.2 guarantees the convergence of each $\mu_\ell^k$ only to a small neighborhood of the closest target mean. As an example, with $m = n = 2$, it might be possible that $\mu_1^k$ and $\mu_2^k$ converge towards $\bar{\mu}_1$, while none of the student's means converge near $\bar{\mu}_2$. Nevertheless, intuitively, when $n$ is large, $\{\mu_i^k\}_{i \in [n]}$ should be distributed among $\{\bar{\mu}_a\}_{a \in [m]}$ according to how "open" each $\bar{\mu}_a$ is in space. Let us formalize this.

**Theorem 3.3.** *Using the notation and assumptions of Theorem E.1, we additionally assume that* $n \geq \frac{m^2}{2\varepsilon} \log \frac{2m}{\rho}$. *Let* $n_a$ *be the number of student's components that converge to* $\bar{\mu}_a$, *i.e.,* $n_a := |\{\ell \in [n] : a = a_\ell^*\}|$. *Then, with probability at least* $1 - 5\rho$, *we have* $\max_{a \in [m]} \left|\frac{n_a}{n} - p_a\right| \leq \sqrt{\varepsilon}/m$ *and*

$$\text{TV}(p_{\mu^k}, \bar{p}) := \frac{1}{2}\int_{\mathbb{R}^d} \left|p_{\mu^k}(x) - \bar{p}(x)\right| dx \leq \sqrt{\varepsilon} + \frac{1}{2}\sum_{a=1}^m \left|p_a - \frac{1}{m}\right| \quad \forall k \geq \tilde{K},$$

*where* $\eta \sim \text{Unif}(\mathbb{S}^{d-1})$ *and* $p_a := \mathbb{P}\left(\left\|R^0\eta - \bar{\mu}_a\right\| < \min_{b \in [m], b \neq a} \left\|R^0\eta - \bar{\mu}_b\right\|\right)$ *is the probability that* $\bar{\mu}_a$ *is the closest vector to* $R^0\eta$ *among* $\{\bar{\mu}_b\}_{b \in [m]}$.

The theorem extends Theorem E.1 and shows that the total variation distance between the student's and teacher's distributions is less than or equal to $\sqrt{\varepsilon}$ up to $\frac{1}{2}\sum_{a=1}^m \left|p_a - \frac{1}{m}\right|$. The latter depends on the probabilities $p_a$, which characterize the probability that $\bar{\mu}_a$ is the closest target mean to a uniformly sampled random variable. Moreover, the student's components are approximately distributed between the target components according to probabilities $p_a$ since $\max_{a \in [m]} |n_a/n - p_a|$ is small.

**Corollary 3.4.** *Using the notation and assumptions of Theorem 3.3, if the vectors* $\{\bar{\mu}_b\}_{b \in [m]}$ *are symmetric, i.e.,* $p_a = \frac{1}{m}$ *for all* $a \in [m]$, *then the total variation distance satisfies* $\text{TV}(p_{\mu^k}, \bar{p}) \leq \sqrt{\varepsilon}$ *for all* $k \geq \tilde{K}$. *In particular, if the vectors* $\{\bar{\mu}_b\}_{b \in [m]}$ *have the same norm and are pairwise orthogonal, then we have* $p_a = \frac{1}{m}$ *for all* $a \in [m]$, *and hence* $\text{TV}(p_{\mu^k}, \bar{p}) \leq \sqrt{\varepsilon}$ *for all* $k \geq \tilde{K}$.

Thus, if the target means are symmetric and $n \gg m$, then it is possible to show the convergence of GD with the reverse Fisher divergence in terms of TV distance even when fitting one GMM to another, and the student's components are distributed between the target components almost evenly since $n_a \approx n/m$. A modified bound also holds when $p_a \neq 1/m$. For instance, if $p_a \in [1-\delta/m, 1+\delta/m]$, then $\text{TV}(p_{\mu^k}, \bar{p}) \leq \sqrt{\varepsilon} + \delta/2$.

### 3.3 One-step convergence and non-isotropic generalization

While proving Theorem E.1, we noticed an interesting convergence property of the problem. When $\mu_i^0$ is initialized on the sphere of radius $R^0$, the method can converge "very close" to the closest target mean in one step. To achieve this, one should choose the step size $\gamma$ to be exactly equal to $n/2$. This is not surprising in view of our discussion in Section 3.1, where we explain that $\mu_\ell^1 \approx \bar{\mu}_{a_\ell^*} + (1 - 2\gamma/n)(\mu_\ell^0 - \bar{\mu}_{a_\ell^*})$. Exactly the same behavior is well known for GD and the optimization problem $\min_{x\in\mathbb{R}^d}\{f(x) := \|x\|^2\}$, where GD can also converge in one step because $f$ is *isotropic*. We now consider a *non-isotropic* score matching problem and extend our theoretical results to this setting.

Let $\mu_1 \equiv [\mu_{1,1}; \ldots; \mu_{1,n_{\min}}] \in \mathbb{R}^{n_{\min}\cdot d}$ and $\mu_2 \equiv [\mu_{2,1}; \ldots; \mu_{2,n_{\max}}] \in \mathbb{R}^{n_{\max}\cdot d}$ be two collections of parameters with $n_1 \equiv n_{\min} < n_2 \equiv n_{\max}$. Define the joint model on $\mathbb{R}^{2d}$ by $q(x,y) := p_{\mu_1}(x)p_{\mu_2}(y)$, and define the joint target distribution by $\bar{q}(x,y) := \bar{p}_1(x)\bar{p}_2(y)$, where $\bar{p}_1$ and $\bar{p}_2$ are arbitrary GMMs of the form (2) with the same number $m$ of target components $\bar{\mu}_{1,1}, \ldots, \bar{\mu}_{1,m}$ and $\bar{\mu}_{2,1}, \ldots, \bar{\mu}_{2,m}$, respectively. We consider the reverse Fisher loss $\mathcal{L}_{\text{rev}}^{\text{joint}}(\mu_1,\mu_2) := \mathbb{E}_{(x,y)\sim q}[\|\nabla_{(x,y)} \log q(x,y) - \nabla_{(x,y)} \log \bar{q}(x,y)\|^2]$. One can show that

$$\mathcal{L}_{\text{rev}}^{\text{joint}}(\mu_1,\mu_2) = \mathcal{L}_{\text{rev}}^{(1)}(\mu_1) + \mathcal{L}_{\text{rev}}^{(2)}(\mu_2), \tag{7}$$

where $\mathcal{L}_{\text{rev}}^{(1)}$ is equal to (1) with $p_\mu \to p_{\mu_1}$ and $\bar{p} \to \bar{p}_1$ ($\mathcal{L}_{\text{rev}}^{(2)}$ is similar). For this separable problem, we can directly apply Theorem 3.2 and get the following result.

**Corollary 3.5.** *Under the assumptions and parameters of Theorem 3.2, we sample the initial parameters as $\mu_{j,\ell}^0 = R^0\eta_{j,\ell}/\|\eta_{j,\ell}\|$, where $\eta_{j,\ell} \sim \mathcal{N}(0,\mathbf{I})$ for all $j \in \{1,2\}, \ell \in [n_j]$ (i.i.d.) and $R^0 = \tilde{\Theta}\left(\max\left\{n_{\max}\sqrt{d}\delta_{\max}^2/\rho, \sqrt{n_{\max}}\bar{R}\right\}\right)$ Assume that the smallest target gap $\min_{j\in\{1,2\}}\min_{a\neq b}\|\bar{\mu}_{j,a} - \bar{\mu}_{j,b}\| \geq \tilde{\Omega}(1)$ and $d \geq \tilde{\Omega}(n_{\max})$. If $\gamma \leq \tilde{\Theta}(n_{\min})$ and GD is run for $\tilde{K} := \tilde{\Theta}\left(n_{\max}^3/\gamma\varepsilon\right)$ iterations, then, with probability at least $1 - 8\rho$, we have that $\|\mu_{j,\ell}^k - \bar{\mu}_{j,a_{j,\ell}^*}\|^2 \leq \varepsilon$, for all $j \in \{1,2\}, \ell \in [n_j]$ and $k \geq \tilde{K}$, where $a_{j,\ell}^* = \arg\min_{a\in[m]}\|\mu_{j,\ell}^0 - \bar{\mu}_{j,a}\|$. In particular, $\tilde{K} = \tilde{\Theta}\left(n_{\max}^3/n_{\min}\varepsilon\right)$ if $\gamma = \tilde{\Theta}(n_{\min})$.*

In this more general case, the convergence rate depends on the ratio (condition number) between $n_{\max}$ and $n_{\min}$. Moreover, in order to apply Theorem 3.2, we have to take $\gamma \leq \tilde{\Theta}(\min\{n_{\min}, n_{\max}\}) = \tilde{\Theta}(n_{\min})$, and the one-step convergence phenomenon is unlikely to be possible for this problem. We can formalize this and prove a lower bound showing that we cannot take $\gamma \gg \tilde{\Theta}(n_{\min})$ and the method requires at least $\tilde{\Omega}(n_{\max}/n_{\min})$ steps.

**Theorem 3.6** (Lower bound; Simplified version of Theorem F.1)**.** *We run GD on the problem (7) with $m = 1$, $\bar{\mu}_{1,1} = \bar{\mu}_{2,1}$, $d \geq \tilde{\Omega}(n_{\max})$, and $n_{\min} < n_{\max}/32$. Assume that $\varepsilon \in (0,1]$, $\rho \in (0,1]$, $R^0 = \tilde{\Theta}(\max\{1, \sqrt{n_{\max}}\bar{R}\})$, where $\bar{R} := \|\bar{\mu}_{1,1}\|$, and we sample the initial parameters as $\mu_{i,\ell}^0 = R^0\eta_{i,\ell}/\|\eta_{i,\ell}\|$, where $\eta_{i,\ell} \sim \mathcal{N}(0,\mathbf{I})$ for all $i \in \{1,2\}, \ell \in [n_i]$ (i.i.d.). Under these assumptions, if the step size $\gamma \leq 2n_{\min}$, then, with probability at least $1 - \rho$, for all $k < \tilde{K}_1 := \tilde{\Theta}(n_{\max}/n_{\min})$, there exists $i \in \{1,2\}$ and $\ell \in [n_i]$ such that $\|\mu_{i,\ell}^k - \bar{\mu}_{i,1}\|^2 > \varepsilon$. If instead $\gamma > 2n_{\min}$, then, with probability at least $1 - \rho$, the method diverges.*

## 4 Experiments

We compare the forward and reverse Fisher divergences numerically for the cases of a single target component, $m = 1$, and multiple target components, $m > 1$. In all experiments, we run GD with a step size of 0.01, use Gauss–Hermite quadrature to evaluate expectations numerically, and compare the convergence trajectories of GD under the different losses. GD starts from the same initial means for both losses. In Figure 2, we visualize the convergence trajectories, which agree with the theory: the student parameters under the forward Fisher divergence may converge in an arbitrary direction, while the reverse Fisher divergence guarantees convergence of all parameters to a small neighborhood of one of the target components.

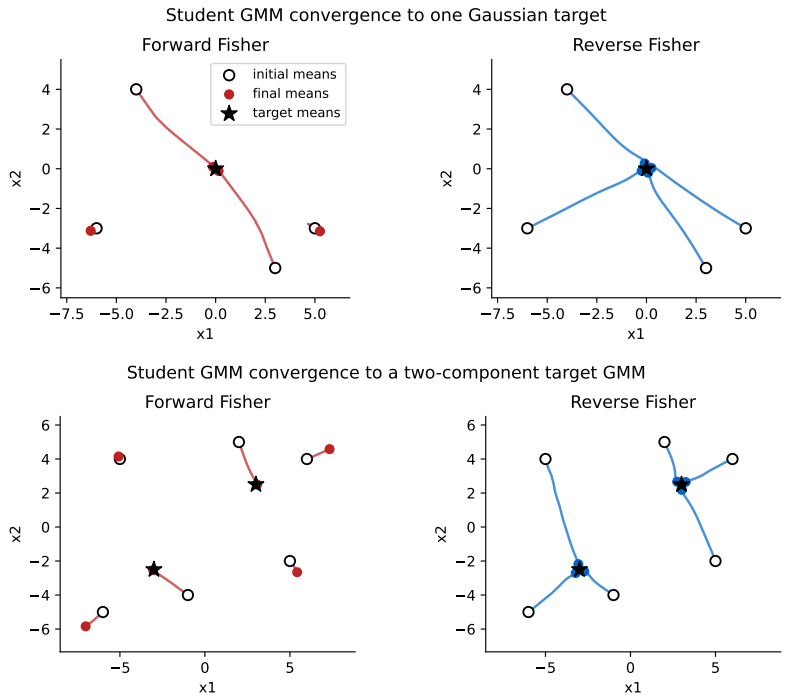



Figure 2: Convergence trajectories of GD with the forward and reverse Fisher divergences in two dimensions. 1st row: a single target component ($m = 1$). 2nd row: two target components ($m = 2$).

### 4.1 Computing gradients in practice

A natural question concerns the computational complexity and feasibility of computing the gradient of the reverse Fisher divergence. Notice that this loss depends on the parameters through the expectation. For GMMs, computing the gradient is as feasible as in the forward loss case, where almost all formulas can be computed analytically, except for the outer expectation, which can be approximated, for instance, using sampling or numerical integration techniques. For more general distributions, depending on the domain, there may be different ways to find or approximate the gradient of (1). One generic way is to use a zeroth-order approximation. The idea is to approximate $\nabla\mathcal{L}_{\text{rev}}(\mu)$ by ${}^{\dim(\mu)}/_{B}\sum_{j=1}^{B}(\mathcal{L}_{\text{rev}}(\mu+\alpha u_j)-\mathcal{L}_{\text{rev}}(\mu-\alpha u_j))(u_j/2\alpha)$, where $B \geq 1, \alpha > 0$ and $u_j$ are i.i.d. uniformly random directions from the sphere of radius 1. The function values can be approximated using Monte Carlo sampling, where samples from $p_{\mu+\alpha u_j}$ and $p_{\mu-\alpha u_j}$ can be approximately drawn using Langevin-type samplers. Developing effective algorithms and adapting current approaches to reverse-divergence methods is an important present and future research direction, and development has already begun in many related topics (Yang et al., 2019; Yu & Zhang, 2023; Cai et al., 2024b;a).

## 5 Conclusion

The comparison of the forward and reverse Fisher divergences is not new and has been carried out in other contexts (e.g., (Margossian et al., 2025; Lai et al., 2026)). Our works analyzes the reverse Fisher divergence and provides strong theoretical evidence that *it has a much more favorable optimization landscape and better convergence properties formalized by our theory.* Even for the score matching problem for GMMs, we believe that this result is interesting and may motivate new research directions for designing algorithms and developing approaches based on the reverse Fisher divergence. In future work, it would be interesting to extend the result to more general families of distributions and to consider more realistic settings with stochastic gradients instead of exact gradients.

### AI USE STATEMENT

In this work, we used generative AI tools to check for and correct grammatical mistakes, assist in implementing methods, and, during the research phase, AI tools to accelerate technical calculations and quickly check different research directions before obtaining the final result. The main observations, ideas, proofs, and research directions were proposed by the authors. AI tools helped us significantly in proving one of the key lemmas, Lemma 2.4, by suggesting an interesting way to bound the trace of a square matrix and derive the lower bound (18) for the dot product. We have not used generative AI tools for other tasks that require disclosure. All results were manually checked and written in LaTeX by the authors. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## REFERENCES

S. S. Barsov and V. V. Ulyanov. Estimates of the proximity of gaussian measures. In *Doklady Mathematics*, volume 34, pp. 462–466, 1987.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly $d$-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, volume 2024, pp. 36916–36936, 2024.

Diana Cai, Chirag Modi, Charles C Margossian, Robert M Gower, David M Blei, and Lawrence K Saul. Eigenvi: score-based variational inference with orthogonal function expansions. *Advances in Neural Information Processing Systems*, 37:132691–132721, 2024a.

Diana Cai, Chirag Modi, Loucas Pillaud-Vivien, Charles Margossian, Robert M Gower, David Blei, and Lawrence K Saul. Batch and match: black-box variational inference with a score-based divergence. In *Forty-first International Conference on Machine Learning*, 2024b.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

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

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin Wainwright, Michael Jordan, and Bin Yu. Sharp analysis of expectation-maximization for weakly identifiable models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1866–1876. PMLR, 2020.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Shivam Gupta, Linda Cai, and Sitan Chen. Faster diffusion sampling with randomized midpoints: Sequential and parallel. In *International Conference on Learning Representations*, volume 2025, pp. 97663–97698, 2025.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *International Conference on Learning Representations*, volume 2024, pp. 42520–42558, 2024.

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

V. I. Khokhlov. The uniform distribution on a sphere in rs. properties of projections. i. *Theory of Probability & Its Applications*, 50(3):386–399, 2006. doi: 10.1137/S0040585X97981846.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Chieh-Hsin Lai, Bac Nguyen, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon, and Molei Tao. A unified view of score-based and drifting models. *arXiv preprint arXiv:2603.07514*, 2026.

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

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 36:2097–2127, 2023.

Charles C Margossian, Loucas Pillaud-Vivien, and Lawrence K Saul. Variational inference for uncertainty quantification: An analysis of trade-offs. *Journal of Machine Learning Research*, 26 (202):1–41, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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*, pp. 10684–10695, 2022.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

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

Marcus Spruill. Asymptotic distribution of coordinates on high dimensional spheres. 2007.

Roman Vershynin. High-dimensional probability. 2019. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.

Yuqing Wang, Ye He, and Molei Tao. Evaluating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 37:19307–19352, 2024.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4958–4991. PMLR, 2024.

Yihong Wu and Harrison H Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $o(\sqrt{n})$ iterations. *Mathematical Statistics & Learning*, 4, 2021.

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.

Weihang Xu, Maryam Fazel, and Simon S Du. Toward global convergence of gradient em for over-parameterized gaussian mixture models. *Advances in Neural Information Processing Systems*, 37:10770–10800, 2024.

Yue Yang, Ryan Martin, and Howard Bondell. Variational approximations using fisher divergence. *arXiv preprint arXiv:1905.05284*, 2019.

Longlin Yu and Cheng Zhang. Semi-implicit variational inference via score matching. In *The Eleventh International Conference on Learning Representations*, 2023.

Yiran Zhang, Weihang Xu, Mo Zhou, Maryam Fazel, and Simon Shaolei Du. Convergence dynamics of over-parameterized score matching for a single gaussian. In *International Conference on Learning Representations*, 2026.

Mo Zhou, Weihang Xu, Maryam Fazel, and Simon S Du. Global convergence of gradient em for over-parameterized gaussian mixtures. *arXiv preprint arXiv:2506.06584*, 2025.

# CONTENTS

**1 Introduction** **1**
1.1 Related work . . . 2
1.2 Motivation and contributions . . . 3
1.3 Preliminaries . . . 3

**2 Learning Single Gaussian Component** **4**

**3 Learning Multiple Gaussian Components** **5**
3.1 Proof technique . . . 6
3.2 Convergence of the total variation distance . . . 7
3.3 One-step convergence and non-isotropic generalization . . . 8

**4 Experiments** **8**
4.1 Computing gradients in practice . . . 9

**5 Conclusion** **9**

**A Table of Notations** **14**

**B Generic Lemmas** **15**

**C Gradient of the Reverse Fisher Divergence for Gaussian Mixture Models** **26**

**D Missing Steps in the Proof of Theorem 2.1** **29**

**E Proof of Theorem E.1** **32**
E.1 Auxiliary lemmas . . . 43
E.2 Corollaries . . . 50

**F Proof of Theorem F.1** **52**

# A TABLE OF NOTATIONS

| Notation | Description |
|---|---|
| $d$ | Dimension |
| $[A_1, \ldots, A_n]$ | Horizontal concatenation of matrices or vectors |
| $[A_1; \ldots; A_n]$ | Vertical concatenation of matrices or vectors |
| $\mathrm{dist}(\mathcal{A}, \mathcal{B})$ | Euclidean distance between sets $\mathcal{A}$ and $\mathcal{B}$, i.e., $\mathrm{dist}(\mathcal{A}, \mathcal{B}) = \inf_{x\in\mathcal{A},\, y\in\mathcal{B}} \|x - y\|$ |
| $[x, y]$ | Closed line segment joining $x$ and $y$, i.e., $\{(1-\lambda)x + \lambda y : \lambda \in [0, 1]\}$ |
| $\bar{\mu}_a$ | Target mode $a$ |
| $\bar{R}$ | Maximum norm of a target mode, $\bar{R} = \max_{a\in[m]} \|\bar{\mu}_a\|$ |
| $\delta_{\min}$ | Minimum distance between distinct target modes |
| $\delta_{\max}$ | Maximum distance between distinct target modes |
| $\mu_\ell^k$ | Parameter $\ell$ at iteration $k$ |
| $\mu_\ell^0$ | Initial parameter $\ell$ |
| $a_\ell^*$ | Index of the target mode closest to $\mu_\ell^0$ |
| $R^0$ | Initialization radius |
| $\gamma$ | Gradient descent step size |
| $\rho$ | Failure probability parameter |
| $\varepsilon$ | Target accuracy |
| $\bar{p}$ | Teacher probability density function |
| $p_\mu$ | Student probability density function parameterized by $\mu$ |
| $\varphi(x)$ | Standard Gaussian density, $\varphi(x) = (2\pi)^{-d/2} \exp(-\|x\|^2/2)$ |
| $\mathbb{E}_{x\sim p}$ | Expectation with respect to the distribution $p$ |
| $\Gamma(t)$ | Gamma function, $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx$ |
| $\mathbf{I}$ | Identity matrix |
| $\|\cdot\|$ | Euclidean norm of a vector |
| $\|\cdot\|_F$ | Frobenius norm of a matrix |
| $\|\cdot\|_{\mathrm{op}}$ | Operator (spectral) norm of a matrix |
| $\mathrm{Tr}(A)$ | Trace of a square matrix $A$ |
| $\langle A, B\rangle$ | Frobenius inner product, $\mathrm{Tr}(A^\top B)$ |
| $\dim(\mu)$ | Dimension of the parameter space containing $\mu$ |
| $g = \mathcal{O}(f)$ | There exists $C > 0$ such that $g(z) \le C\, f(z)$ for all $z \in \mathcal{Z}$. |
| $g = \Omega(f)$ | There exists $C > 0$ such that $g(z) \ge C\, f(z)$ for all $z \in \mathcal{Z}$. |
| $g = \Theta(f)$ | There exist $C_1, C_2 > 0$ such that $C_1 f(z) \le g(z) \le C_2 f(z)$ for all $z \in \mathcal{Z}$. |
| $\tilde{\mathcal{O}}, \tilde{\Omega}$, and $\tilde{\Theta}$ | The same as $\mathcal{O}$, $\Omega$, and $\Theta$, but up to logarithmic factors. |

Table 1: Frequently used notation.

## B GENERIC LEMMAS

In this section, we state several useful generic lemmas that will be used to prove the convergence of GD.

**Lemma B.1.** *Let the Gamma function be defined by*

$$\Gamma(t) := \int_0^\infty x^{t-1} e^{-x}\, dx, \qquad t > 0.$$

*Then, for any integer $d \geq 2$,*

$$\Gamma\left(\frac{d}{2}\right) \leq \sqrt{\frac{d-1}{2}}\, \Gamma\left(\frac{d-1}{2}\right).$$

*Proof.* Since $\Gamma$ is log-convex, $\log\left(\Gamma\left(\frac{x}{2}+\frac{y}{2}\right)\right) \leq \frac{1}{2}\log\left(\Gamma(x)\right) + \frac{1}{2}\log\left(\Gamma(y)\right)$ for all $x, y \geq 0$. Taking $x = \frac{d-1}{2}$ and $y = \frac{d+1}{2}$, we get $\log\left(\Gamma\left(\frac{d}{2}\right)\right) \leq \frac{1}{2}\log\left(\Gamma\left(\frac{d-1}{2}\right)\right) + \frac{1}{2}\log\left(\Gamma\left(\frac{d+1}{2}\right)\right)$. Due to $\Gamma\left(\frac{d+1}{2}\right) = \frac{d-1}{2}\Gamma\left(\frac{d-1}{2}\right)$, we obtain $\Gamma\left(\frac{d}{2}\right) \leq \sqrt{\frac{d-1}{2}}\Gamma\left(\frac{d-1}{2}\right)$. □

**Lemma B.2.** *Let $\rho \in (0,1]$ and $\mu_\ell^0 = R^0 \eta_\ell / \|\eta_\ell\|$, where $R^0 > 0$ and $\eta_\ell \sim \mathcal{N}(0, \mathbf{I}_d)$ for all $\ell \in [n]$, independently. The vectors $\{\bar{\mu}_a\}_{a\in[m]}$ are deterministic. Assume that $d > 3$ and define*

$$\delta_{\min} := \min_{a \neq b} \|\bar{\mu}_a - \bar{\mu}_b\| > 0.$$

*Let*

$$a_i^* := \arg\min_{a\in[m]} \left\|\mu_i^0 - \bar{\mu}_a\right\|$$

*with ties broken arbitrarily, and let*

$$\Omega_1 := \left\{\forall i \in [n] : \left\|\mu_i^0 - \bar{\mu}_{a_i^*}\right\|^2 \leq \min_{b \neq a_i^*} \left\|\mu_i^0 - \bar{\mu}_b\right\|^2 - \Delta\right\}.$$

*Then choosing*

$$\Delta = \frac{2R^0 \delta_{\min} \rho}{nm^2\sqrt{d}}$$

*guarantees that*

$$\mathbb{P}(\Omega_1) \geq 1 - \rho.$$

*Proof.* For fixed $i \in [n]$, define $g_i := \min_{b\neq a_i^*} \left\|\mu_i^0 - \bar{\mu}_b\right\|^2 - \left\|\mu_i^0 - \bar{\mu}_{a_i^*}\right\|^2$. Then $\Omega_1 = \{g_i \geq \Delta\ \forall i \in [n]\}$, and hence

$$\mathbb{P}(\Omega_1^c) \leq \sum_{i=1}^n \mathbb{P}(g_i < \Delta) = n\mathbb{P}(g_1 < \Delta) = n\mathbb{P}\left(\left\|\mu_1^0 - \bar{\mu}_{b_1^*}\right\|^2 - \left\|\mu_1^0 - \bar{\mu}_{a_1^*}\right\|^2 < \Delta\right),$$

where $\Omega_1^c$ is the complementary event of $\Omega_1$ and $b_1^* := \arg\min_{a\in[m], a\neq a_1^*} \left\|\mu_1^0 - \bar{\mu}_a\right\|$. Thus

$$\mathbb{P}(\Omega_1^c) \leq n \sum_{a,b\in[m], a\neq b} \mathbb{P}\left(\left|\left\|\mu_1^0 - \bar{\mu}_b\right\|^2 - \left\|\mu_1^0 - \bar{\mu}_a\right\|^2\right| < \Delta\right), \tag{8}$$

Let $u = \eta_1 / \|\eta_1\|$, so that $\mu_1^0 = R^0 u$ and $u$ is uniform on the unit sphere. For fixed $a \neq b$,

$$\left\|R^0 u - \bar{\mu}_b\right\|^2 - \left\|R^0 u - \bar{\mu}_a\right\|^2 = \|\bar{\mu}_b\|^2 - \|\bar{\mu}_a\|^2 - 2R^0 \langle u, \bar{\mu}_b - \bar{\mu}_a\rangle.$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\|\mu_1^0 - \bar{\mu}_b\right\|^2 - \left\|\mu_1^0 - \bar{\mu}_a\right\|^2\right| < \Delta\right) \\
&= \mathbb{P}\left(\left|\left\langle u, \frac{\bar{\mu}_b - \bar{\mu}_a}{\|\bar{\mu}_b - \bar{\mu}_a\|}\right\rangle - \frac{\|\bar{\mu}_b\|^2 - \|\bar{\mu}_a\|^2}{2R^0\|\bar{\mu}_b - \bar{\mu}_a\|}\right| < \frac{\Delta}{2R^0\|\bar{\mu}_b - \bar{\mu}_a\|}\right),
\end{aligned}$$

$$= \mathbb{P}\left(\left|\left\langle u, \frac{\bar{\mu}_b - \bar{\mu}_a}{\|\bar{\mu}_b - \bar{\mu}_a\|}\right\rangle - \bar{t}\right| < \bar{b}\right),$$

where $\bar{t} := \frac{\|\bar{\mu}_b\|^2 - \|\bar{\mu}_a\|^2}{2R^0\|\bar{\mu}_b - \bar{\mu}_a\|}$ and $\bar{b} := \frac{\Delta}{2R^0\|\bar{\mu}_b - \bar{\mu}_a\|}$. It is well-known (e.g. (Spruill, 2007)) that $Z := \left\langle u, \frac{\bar{\mu}_b - \bar{\mu}_a}{\|\bar{\mu}_b - \bar{\mu}_a\|}\right\rangle$ has the density

$$p(x) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}(1-x^2)^{(d-3)/2} \tag{9}$$

for all $x \in [-1,1]$, and $p(x) = 0$ otherwise, where $\Gamma$ is the gamma function. Since $d > 3$, the density $p$ is symmetric and uniquely maximized at 0. Hence

$$\mathbb{P}\left(\left|\left\|\mu_1^0 - \bar{\mu}_b\right\|^2 - \left\|\mu_1^0 - \bar{\mu}_a\right\|^2\right| < \Delta\right)$$
$$= \mathbb{P}\left(|Z - \bar{t}| < \bar{b}\right) \le \mathbb{P}\left(|Z| < \bar{b}\right) \le 2\bar{b}\,p(0) = 2\bar{b}\frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}.$$

Using Lemma B.1,

$$\mathbb{P}\left(\left|\left\|\mu_1^0 - \bar{\mu}_b\right\|^2 - \left\|\mu_1^0 - \bar{\mu}_a\right\|^2\right| < \Delta\right) \le \frac{1}{\sqrt{\pi}}\sqrt{\frac{d-1}{2}}\frac{\Delta}{R^0\|\bar{\mu}_b - \bar{\mu}_a\|} \le \frac{\sqrt{d}\Delta}{2R^0\|\bar{\mu}_b - \bar{\mu}_a\|}.$$

Substituting into (8) and using the definition of $\delta_{\min}$,

$$\mathbb{P}(\Omega_1^c) \le \frac{nm^2\sqrt{d}\Delta}{2R^0\delta_{\min}}.$$

It remains to use our choice of $\Delta$. □

**Lemma B.3.** *Let $\rho \in (0,1]$ and $\mu_\ell^0 = R^0\eta_\ell/\|\eta_\ell\|$, where $R^0 > 0$ and $\eta_\ell \sim \mathcal{N}(0, \mathbf{I}_d)$ for all $\ell \in [n]$, independently. Assume $d \ge 3$. Let*

$$\Omega_2 := \left\{\forall i \ne j \in [n] : \left\|\mu_i^0 - \mu_j^0\right\|^2 \ge \Delta_2\right\}.$$

*Then choosing*

$$\Delta_2 = (R^0)^2\left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\,n^2}\right)^{\frac{2}{d-1}}$$

*guarantees that*

$$\mathbb{P}(\Omega_2) \ge 1 - \rho.$$

*Proof.* Let $u_i := \eta_i/\|\eta_i\|$. Then $u_1, \ldots, u_n$ are independent and uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$, and $\mu_i^0 = R^0 u_i$. For any fixed pair $i \ne j$,

$$\left\|\mu_i^0 - \mu_j^0\right\|^2 = (R^0)^2\left\|u_i - u_j\right\|^2.$$

Therefore,

$$\mathbb{P}\left(\left\|\mu_i^0 - \mu_j^0\right\|^2 \le \Delta_2\right) = \mathbb{P}\left(-2(R^0)^2\langle u_i, u_j\rangle \le \Delta_2 - 2(R^0)^2\right)$$
$$= \mathbb{P}\left(\langle u_i, u_j\rangle \ge t\right),$$

where $t := 1 - \frac{\Delta_2}{2(R^0)^2}$. Conditioning on $u_i$, $Z := \langle u_i, u_j\rangle$ has the density (9). Thus,

$$\mathbb{P}\left(\left\|\mu_i^0 - \mu_j^0\right\|^2 \le \Delta_2\right) \le \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}\int_t^1 (1-x^2)^{(d-3)/2}\,dx$$
$$\le \frac{\Gamma\left(\frac{d}{2}\right)(1-t)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}(1-t^2)^{(d-3)/2}$$
$$\le \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}(1-t^2)^{(d-1)/2}.$$

Due to Lemma B.1,

$$\mathbb{P}\left(\left\|\mu_i^0 - \mu_j^0\right\|^2 \le \Delta_2\right) \le \frac{\sqrt{d}}{\sqrt{2\pi}}(1-t^2)^{(d-1)/2}.$$

Moreover,

$$1 - t^2 = 1 - \left(1 - \frac{\Delta_2}{2(R^0)^2}\right)^2 = \frac{\Delta_2}{(R^0)^2} - \frac{\Delta_2^2}{4(R^0)^4} \le \frac{\Delta_2}{(R^0)^2}.$$

Therefore, by the union bound,

$$\mathbb{P}\left(\Omega_2^c\right) \le \sum_{i<j} \mathbb{P}\left(\left\|\mu_i^0 - \mu_j^0\right\|^2 \le \Delta_2\right) \le \frac{\sqrt{d}n^2}{\sqrt{2\pi}}\left(\frac{\Delta_2}{(R^0)^2}\right)^{(d-1)/2}.$$

Choosing

$$\Delta_2 = (R^0)^2\left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\,n^2}\right)^{\frac{2}{d-1}}$$

gives $\mathbb{P}\left(\Omega_2^c\right) \le \rho$, and hence $\mathbb{P}\left(\Omega_2\right) \ge 1-\rho$. □

**Lemma B.4.** *Let $\rho \in (0,1]$, $d \ge 3$, $n \ge 2$, and $m \ge 2$. Let $\mu_\ell^0 = R^0\eta_\ell / \|\eta_\ell\|$, where $R^0 > 0$ and $\eta_\ell \sim \mathcal{N}(0, \mathbf{I}_d)$ for all $\ell \in [n]$, independently. Moreover, $\{\bar{\mu}_a\}_{a\in[m]}$ are deterministic vectors such that $\delta_{\min} := \min_{a\neq b}\|\bar{\mu}_a - \bar{\mu}_b\| > 0$ and $\|\bar{\mu}_a\| \le R^0$ for all $a \in [m]$. Let*

$$\Omega_3 := \left\{\forall i \neq j \in [n] \text{ s.t. } a_i^* \neq a_j^* : \text{dist}([\mu_i^0, \bar{\mu}_{a_i^*}], [\mu_j^0, \bar{\mu}_{a_j^*}]) \ge \Delta_3\right\},$$

*where $a_i^* = \arg\min_{a\in[m]}\left\|\mu_i^0 - \bar{\mu}_a\right\|$ for all $i \in [n]$. Then choosing*

$$\Delta_3 = \frac{\delta_{\min}}{8}\left(\frac{\rho}{mn^2}\right)^{\frac{1}{d-2}}$$

*guarantees that*

$$\mathbb{P}\left(\Omega_3\right) \ge 1-\rho.$$

*Proof.* Clearly,

$$\mathbb{P}\left(\Omega_3^c\right) \le \sum_{i\neq j\in[n]} \mathbb{P}\left(\left\{a_i^* \neq a_j^*\right\} \cap \left\{\text{dist}([\mu_i^0, \bar{\mu}_{a_i^*}], [\mu_j^0, \bar{\mu}_{a_j^*}]) < \Delta_3\right\}\right).$$

Fix $i \neq j$, and assume that $a_i^* \neq a_j^*$. Next, $Q := \left\{x \in \mathbb{R}^d : \left\|x - \bar{\mu}_{a_i^*}\right\| \le \left\|x - \bar{\mu}_{a_j^*}\right\|\right\}$. This set is convex; thus, $[\mu_i^0, \bar{\mu}_{a_i^*}] \subseteq Q$, and

$$\frac{\delta_{\min}}{2} \le \frac{1}{2}\left\|\bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*}\right\| \le \frac{1}{2}\left\|a - \bar{\mu}_{a_i^*}\right\| + \frac{1}{2}\left\|a - \bar{\mu}_{a_j^*}\right\| \le \left\|a - \bar{\mu}_{a_j^*}\right\| \tag{10}$$

for all $a \in [\mu_i^0, \bar{\mu}_{a_i^*}]$. Clearly,

$$\begin{aligned}
&\mathbb{P}\left(\left\{a_i^* \neq a_j^*\right\} \cap \text{dist}([\mu_i^0, \bar{\mu}_{a_i^*}], [\mu_j^0, \bar{\mu}_{a_j^*}]) < \Delta_3\right) \\
&\le \underbrace{\mathbb{P}\left(\left\{a_i^* \neq a_j^*\right\} \cap \exists b \in [\mu_j^0, \bar{\mu}_{a_j^*}] \text{ such that } \text{dist}([\mu_i^0, \bar{\mu}_{a_i^*}], b) < \Delta_3 \text{ and } \left\|b - \bar{\mu}_{a_j^*}\right\| < \frac{\delta_{\min}}{4}\right)}_{P_1} \\
&+ \underbrace{\mathbb{P}\left(\left\{a_i^* \neq a_j^*\right\} \cap \exists b \in [\mu_j^0, \bar{\mu}_{a_j^*}] \text{ such that } \text{dist}([\mu_i^0, \bar{\mu}_{a_i^*}], b) < \Delta_3 \text{ and } \left\|b - \bar{\mu}_{a_j^*}\right\| \ge \frac{\delta_{\min}}{4}\right)}_{P_2}.
\end{aligned}$$

Consider any point $b \in [\mu_j^0, \bar{\mu}_{a_j^*}]$. If $\left\|b - \bar{\mu}_{a_j^*}\right\| < \frac{\delta_{\min}}{4}$, then

$$\|a - b\| = \left\|(a - \bar{\mu}_{a_j^*}) - (b - \bar{\mu}_{a_j^*})\right\| \ge \left\|a - \bar{\mu}_{a_j^*}\right\| - \left\|b - \bar{\mu}_{a_j^*}\right\| \ge \frac{\delta_{\min}}{4} > \Delta_3$$

for all $a \in [\mu_i^0, \bar{\mu}_{a_i^*}]$ due to (10). Thus,

$$P_1 = 0.$$

Otherwise, consider the case $\left\|b - \bar{\mu}_{a_j^*}\right\| \geq \frac{\delta_{\min}}{4}$. Let $\mathcal{A} = \{x \in \mathbb{R}^d \,:\, x = \bar{\mu}_{a_j^*} + \alpha(\mu_i^0 - \bar{\mu}_{a_j^*}) + \beta(\bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*}), \alpha, \beta \in \mathbb{R}\}$ be the affine space. Notice that if $\left\|b - \bar{\mu}_{a_j^*}\right\| \geq \frac{\delta_{\min}}{4}$ and $b \in [\mu_j^0, \bar{\mu}_{a_j^*}]$, then

$$b = \bar{\mu}_{a_j^*} + t(\mu_j^0 - \bar{\mu}_{a_j^*})$$

for some $t \geq \frac{\delta_{\min}}{4\left\|\mu_j^0 - \bar{\mu}_{a_j^*}\right\|} \geq \frac{\delta_{\min}}{8R^0}$. We get

$$\begin{aligned}
P_2 &\leq \mathbb{P}\left(\cup_{t\in[\delta_{\min}/8R^0,1]}\text{dist}(\mathcal{A}, \bar{\mu}_{a_j^*} + t(\mu_j^0 - \bar{\mu}_{a_j^*})) < \Delta_3\right) \\
&= \mathbb{P}\left(\cup_{t\in[\delta_{\min}/8R^0,1]}\text{dist}(\mathcal{B}_{a_j^*}, t(\mu_j^0 - \bar{\mu}_{a_j^*})) < \Delta_3\right) \\
&= \mathbb{P}\left(\cup_{t\in[\delta_{\min}/8R^0,1]}\text{dist}(\mathcal{B}_{a_j^*}, \mu_j^0 - \bar{\mu}_{a_j^*}) < \frac{\Delta_3}{t}\right) \\
&\leq \mathbb{P}\left(\text{dist}(\mathcal{B}_{a_j^*}, \mu_j^0 - \bar{\mu}_{a_j^*}) < \frac{8\Delta_3 R^0}{\delta_{\min}}\right),
\end{aligned}$$

where $\mathcal{B}_{a_j^*} := \{x \in \mathbb{R}^d \,:\, x = \alpha(\mu_i^0 - \bar{\mu}_{a_j^*}) + \beta(\bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*}), \alpha, \beta \in \mathbb{R}\}$ is the linear space. Since $a_j^*$ is random, we further bound

$$P_2 \leq \sum_{a\in[m]} \mathbb{P}\left(\text{dist}(\mathcal{B}_a, \mu_j^0 - \bar{\mu}_a) < \frac{8\Delta_3 R^0}{\delta_{\min}}\right),$$

where

$$\mathcal{B}_a := \{x \in \mathbb{R}^d \,:\, x = \alpha(\mu_i^0 - \bar{\mu}_a) + \beta(\bar{\mu}_{a_i^*} - \bar{\mu}_a), \alpha, \beta \in \mathbb{R}\}.$$

The space $\mathcal{B}_a$ has dimension at most 2 and is independent of $\mu_j^0$. Let $\mathcal{L}_a$ be a linear subspace of dimension exactly 2 such that $\mathcal{B}_a \subseteq \mathcal{L}_a$. Then, for all $r > 0$,

$$\mathbb{P}\left(\text{dist}(\mathcal{B}_a, \mu_j^0 - \bar{\mu}_a) < r\right) \leq \mathbb{P}\left(\text{dist}(\mathcal{L}_a, \mu_j^0 - \bar{\mu}_a) < r\right).$$

Therefore, for $r > 0$,

$$\mathbb{P}\left(\text{dist}(\mathcal{B}_a, \mu_j^0 - \bar{\mu}_a) < r\right) \leq \mathbb{P}\left(\left\|\mathbf{P}\mu_j^0 - \mathbf{P}\bar{\mu}_a\right\| < r\right),$$

where $\mathbf{P}$ is the orthogonal projection onto $\mathcal{L}_a^\perp$. Since $\mathbf{P} = \mathbf{U}\mathbf{L}\mathbf{U}^\top$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{L}$ is a diagonal matrix with diagonal entries $[1, \ldots, 1, 0, 0] \in \mathbb{R}^d$,

$$\mathbb{P}\left(\text{dist}(\mathcal{B}_a, \mu_j^0 - \bar{\mu}_a) < r\right) \leq \mathbb{P}\left(\left\|\mathbf{L}\mathbf{U}^\top\mu_j^0 - \mathbf{L}\mathbf{U}^\top\bar{\mu}_a\right\| < r\right),$$

It is well known that the first $d-2$ coordinates of $\mathbf{L}\mathbf{U}^\top\mu_j^0$ are uniformly distributed in the ball of radius $R^0$ (e.g., (Khokhlov, 2006, formula (6), with $k = s - 2$)). Thus,

$$\mathbb{P}\left(\text{dist}(\mathcal{B}_a, \mu_j^0 - \bar{\mu}_a) < r\right) \leq \frac{\text{Vol}_{d-2}(B(0, R^0) \bigcap B(v, r))}{\text{Vol}_{d-2}(B(0, R^0))} \leq \frac{\text{Vol}_{d-2}(B(0, r))}{\text{Vol}_{d-2}(B(0, R^0))} = \left(\frac{r}{R^0}\right)^{d-2}.$$

Here, $B(a, r) := \{x \in \mathbb{R}^{d-2} \,:\, \|x - a\| \leq r\}$, $\text{Vol}_{d-2}(\cdot)$ is the volume of a set, and $v$ is the vector of the first $d-2$ coordinates of $\mathbf{L}\mathbf{U}^\top\bar{\mu}_a$. Applying this with $r = 8\Delta_3 R^0/\delta_{\min}$, we get

$$P_2 \leq m\left(\frac{8\Delta_3}{\delta_{\min}}\right)^{d-2}.$$

Therefore,

$$\mathbb{P}\left(\Omega_3^c\right) \leq mn^2\left(\frac{8\Delta_3}{\delta_{\min}}\right)^{d-2} \leq \rho$$

by the choice of $\Delta_3$. Hence,

$$\mathbb{P}\left(\Omega_3\right) \geq 1 - \rho.$$

□

**Lemma B.5.** *Let $\rho \in (0,1]$ and $\mu_\ell^0 = R^0 \eta_\ell / \|\eta_\ell\|$, where $R^0 > 0$ and $\eta_\ell \sim \mathcal{N}(0, \mathbf{I}_d)$ for all $\ell \in [n]$, independently. Let*

$$\Omega_4^{\text{full}} := \left\{ \left\| \begin{bmatrix} (\mu_1^0)^\top \\ \vdots \\ (\mu_n^0)^\top \end{bmatrix} \right\|_{\text{op}} \leq \Delta_4 \right\}.$$

*Then choosing*

$$\Delta_4 = R^0 \sqrt{\frac{n}{d} + \max\left\{ \frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)} \right\}}$$

*guarantees that*

$$\mathbb{P}\left(\Omega_4^{\text{full}}\right) \geq 1 - \rho.$$

*Proof.* Without loss of generality, we assume that $R^0 = 1$. It is sufficient to rescale the final result. For all $v \in \mathbb{R}^d$,

$$\left\| \begin{bmatrix} (\mu_1^0)^\top \\ \vdots \\ (\mu_n^0)^\top \end{bmatrix} v \right\|^2 = \left\| \begin{bmatrix} (\mu_1^0)^\top v \\ \vdots \\ (\mu_n^0)^\top v \end{bmatrix} \right\|^2 = \sum_{i=1}^n \left( (\mu_i^0)^\top v \right)^2 = v^\top \left( \sum_{i=1}^n \mu_i^0 (\mu_i^0)^\top \right) v$$

Let $\mathbf{A}_i := \mu_i^0 (\mu_i^0)^\top$ for all $i \in [n]$. Thus,

$$\left\| \begin{bmatrix} (\mu_1^0)^\top \\ \vdots \\ (\mu_n^0)^\top \end{bmatrix} \right\|_{\text{op}}^2 = \left\| \sum_{i=1}^n \mathbf{A}_i \right\|_{\text{op}} \tag{11}$$

Let us also temporarily define $\xi = [\xi_1, \ldots, \xi_d]^\top$, where $\{\xi_i\}_{i=1}^d$ are i.i.d. normal random variables, then, for all $i \neq j$, $\mathbb{E}\left[\frac{\xi_i \xi_j}{\|\xi\|^2}\right] = c$ for some $c \in \mathbb{R}$. At the same time, $\mathbb{E}\left[\frac{-\xi_i \xi_j}{\|\xi\|^2}\right] = c$ since $\xi_i$ and $-\xi_i$ have the same distribution. Thus, $c = 0$. Moreover, $\mathbb{E}\left[\frac{\xi_i^2}{\|\xi\|^2}\right] = \bar{c}$ for all $i \in [d]$ and some $\bar{c} \in \mathbb{R}$. Therefore, $1 = \sum_{i=1}^d \mathbb{E}\left[\frac{\xi_i^2}{\|\xi\|^2}\right] = d\bar{c}$ and $\bar{c} = \frac{1}{d}$. We can conclude that

$$\mathbb{E}\left[\mathbf{A}_i\right] = \frac{1}{d}\mathbf{I}_d$$

for all $i \in [n]$. Next,

$$\left\|\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right]\right\|_{\text{op}} = \left\| \mu_i^0 (\mu_i^0)^\top - \frac{1}{d}\mathbf{I}_d \right\|_{\text{op}} \leq 1$$

for all $i \in [n]$ since $v^\top(\mu_i^0(\mu_i^0)^\top - \frac{1}{d}\mathbf{I}_d)v = (v^\top \mu_i^0)^2 - \frac{1}{d}\|v\|^2 \leq \left(1 - \frac{1}{d}\right)\|v\|^2 \leq \|v\|^2$ and $v^\top(\mu_i^0(\mu_i^0)^\top - \frac{1}{d}\mathbf{I}_d)v \geq -\frac{1}{d}\|v\|^2$ for all $v \in \mathbb{R}^d$. Finally,

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right])^2\right] &= \mathbb{E}\left[\left(\mu_i^0(\mu_i^0)^\top - \frac{1}{d}\mathbf{I}_d\right)^2\right] \\
&= \mathbb{E}\left[\left(\mu_i^0(\mu_i^0)^\top \mu_i^0(\mu_i^0)^\top - \frac{2}{d}\mu_i^0(\mu_i^0)^\top + \frac{1}{d^2}\mathbf{I}_d\right)\right] \\
&= \left(\mathbb{E}\left[\mu_i^0(\mu_i^0)^\top\right] - \frac{2}{d^2}\mathbf{I}_d + \frac{1}{d^2}\mathbf{I}_d\right) = \frac{1}{d}\left(1 - \frac{1}{d}\right)\mathbf{I}_d
\end{aligned}$$

and

$$\sigma^2 := \left\| \sum_{i=1}^n \mathbb{E}\left[(\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right])^2\right] \right\|_{\text{op}} \leq \frac{n}{d}.$$

Using (11),

$$\left\|\begin{bmatrix}(\mu_1^0)^\top\\ \vdots \\ (\mu_n^0)^\top\end{bmatrix}\right\|_{\text{op}}^2 \le \left\|\sum_{i=1}^n(\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right])\right\|_{\text{op}} + \left\|\sum_{i=1}^n \mathbb{E}\left[\mathbf{A}_i\right]\right\|_{\text{op}} = \left\|\sum_{i=1}^n(\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right])\right\|_{\text{op}} + \frac{n}{d},$$

where $\{\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right]\}_{i=1}^n$ are i.i.d. symmetric random matrices with mean zero, $\|\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right]\|_{\text{op}} \le 1$ for all $i \in [n]$, and $\sigma^2 \le \frac{n}{d}$. We can use the matrix Bernstein's inequality (Theorem 5.4.1 in (Vershynin, 2019)) to get

$$\mathbb{P}\left(\left\|\sum_{i=1}^n(\mathbf{A}_i - \mathbb{E}\left[\mathbf{A}_i\right])\right\|_{\text{op}} \ge t\right) \le 2d\exp\left(-\frac{t^2}{2\sigma^2 + \frac{2t}{3}}\right) \le 2d\exp\left(-\min\left\{\frac{t^2}{4\sigma^2}, \frac{3t}{4}\right\}\right).$$

Thus, with probability at least $1-\rho$,

$$\left\|\begin{bmatrix}(\mu_1^0)^\top\\ \vdots \\ (\mu_n^0)^\top\end{bmatrix}\right\|_{\text{op}}^2 \le t + \frac{n}{d},$$

for

$$t = \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}.$$

□

**Lemma B.6.** *Let $\rho \in (0,1]$ and $\mu_\ell^0 = R^0\eta_\ell / \|\eta_\ell\|$, where $R^0 > 0$ and $\eta_\ell \sim \mathcal{N}(0, \mathbf{I}_d)$ for all $\ell \in [n]$, independently. Moreover, let $\{I_1, \dots, I_k\}$ be a (potentially random) partition of $[n]$ of (potentially random) size $k \in [n]$. Let*

$$\Omega_4 := \left\{\forall a \in [k] : \left\|\begin{bmatrix}\left(\mu_{i_{a,1}}^0\right)^\top\\ \vdots \\ \left(\mu_{i_{a,n_a}}^0\right)^\top\end{bmatrix}\right\|_{\text{op}} \le \Delta_4\right\}, \qquad I_a \equiv \{i_{a,1}, \dots, i_{a,n_a}\}.$$

*Then choosing*

$$\Delta_4 = R^0\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}$$

*guarantees that*

$$\mathbb{P}\left(\Omega_4\right) \ge 1 - \rho.$$

*Proof.* Clearly, if the event

$$\left\|\begin{bmatrix}(\mu_1^0)^\top\\ \vdots \\ (\mu_n^0)^\top\end{bmatrix}\right\|_{\text{op}} \le \Delta_4$$

holds, then event $\Omega_4$ holds since

$$\sup_{\|v\|=1}\left\|\begin{bmatrix}(\mu_1^0)^\top\\ \vdots \\ (\mu_n^0)^\top\end{bmatrix} v\right\|^2 \ge \sup_{\|v\|=1}\left\|\begin{bmatrix}\left(\mu_{i_{a,1}}^0\right)^\top\\ \vdots \\ \left(\mu_{i_{a,n_a}}^0\right)^\top\end{bmatrix} v\right\|^2$$

for all $a \in [k]$. It remains to use Lemma B.5. □

**Lemma B.7.** *Let $\rho \in (0,1]$. Let $\mu_\ell^0 = R^0 \eta_\ell / \|\eta_\ell\|$, where $R^0 > 0$ and $\eta_\ell \sim \mathcal{N}(0, \mathbf{I}_d)$ for all $\ell \in [n]$, independently. Moreover, $\{\bar{\mu}_a\}_{a\in[m]}$ are non-equal deterministic vectors. Let*

$$\Omega_5 := \left\{ \max_{a\in[m]} \left| \frac{n_a}{n} - p_a \right| \le \Delta_5 \right\}, \tag{12}$$

*where $I_a := \{i \in [n] : a_i^* = a\}$, $a_i^* = \arg\min_{a\in[m]} \left\|\mu_i^0 - \bar{\mu}_a\right\|$, $n_a := |I_a|$, and $p_a := \mathbb{P}(a_1^* = a)$ for all $a \in [m]$. Then choosing*

$$\Delta_5 = \sqrt{\frac{1}{2n} \log \frac{2m}{\rho}}$$

*guarantees that*

$$\mathbb{P}(\Omega_5) \ge 1 - \rho.$$

*Proof.* Clearly,

$$n_a = \sum_{i=1}^n \mathbf{1}[a_i^* = a].$$

Since $\mu_1^0, \dots, \mu_n^0$ are independent and identically distributed, the random variables $\mathbf{1}[a_i^* = a]$ are independent Bernoulli random variables with mean $p_a$. By Hoeffding's inequality, for every $t > 0$,

$$\mathbb{P}\left( \left| \frac{n_a}{n} - p_a \right| \ge t \right) \le 2\exp(-2nt^2).$$

Taking the union bound over $a \in [m]$, we obtain

$$\mathbb{P}\left( \max_{a\in[m]} \left| \frac{n_a}{n} - p_a \right| \ge t \right) \le \sum_{a=1}^m \mathbb{P}\left( \left| \frac{n_a}{n} - p_a \right| \ge t \right) \le 2m\exp(-2nt^2).$$

Choosing

$$t = \Delta_5 := \sqrt{\frac{1}{2n} \log \frac{2m}{\rho}}$$

gives

$$2m\exp(-2n\Delta_5^2) = \rho.$$

Hence,

$$\mathbb{P}\left( \max_{a\in[m]} \left| \frac{n_a}{n} - p_a \right| \le \Delta_5 \right) \ge 1 - \rho.$$

□

**Lemma B.8.** *For all $a, b \in \mathbb{R}^d$,*

$$\mathbb{E}_{x\sim\mathcal{N}(a,\mathbf{I})} \left[ \min\left\{ \exp\left( \frac{1}{2}\|x-a\|^2 - \frac{1}{2}\|x-b\|^2 \right), 1 \right\} \right] \le 2\exp\left( -\frac{1}{32}\|b-a\|^2 \right).$$

*Proof.* If $a = b$, then the bound is true. Assume that $a \ne b$. Then

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(a,\mathbf{I})} \left[ \min\left\{ \exp\left( \frac{1}{2}\|x-a\|^2 - \frac{1}{2}\|x-b\|^2 \right), 1 \right\} \right] \\
&= \mathbb{E}_{\xi\sim\mathcal{N}(0,\mathbf{I})} \left[ \min\left\{ \exp\left( \frac{1}{2}\|\xi\|^2 - \frac{1}{2}\|\xi-(b-a)\|^2 \right), 1 \right\} \right] \\
&= \mathbb{E}_{\xi\sim\mathcal{N}(0,\mathbf{I})} \left[ \min\left\{ \exp\left( \langle \xi, b-a\rangle - \frac{1}{2}\|b-a\|^2 \right), 1 \right\} \right].
\end{aligned}$$

Note that $\left\langle \xi, \frac{b-a}{\|b-a\|} \right\rangle \sim \mathcal{N}(0,1)$. Thus

$$\mathbb{P}(\langle \xi, b-a\rangle \ge c) = \mathbb{P}\left( \left\langle \xi, \frac{b-a}{\|b-a\|} \right\rangle \ge \frac{c}{\|b-a\|} \right) \le \exp\left( -\frac{c^2}{2\|b-a\|^2} \right)$$

for all $c \geq 0$ and

$$\mathbb{P}\left(\langle \xi, b-a\rangle \geq \frac{\|b-a\|^2}{4}\right) \leq \exp\left(-\frac{\|b-a\|^2}{32}\right)$$

for $c = \frac{\|b-a\|^2}{4}$. Using this bound, we get

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(a,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x-a\|^2 - \frac{1}{2}\|x-b\|^2\right), 1\right\}\right] \\
&\leq \mathbb{E}_{\xi\sim\mathcal{N}(0,\mathbf{I})}\left[\min\left\{\exp\left(\langle \xi, b-a\rangle - \frac{1}{2}\|b-a\|^2\right), 1\right\}\mathbf{1}\left[\langle \xi, b-a\rangle < \frac{\|b-a\|^2}{4}\right]\right] \\
&\quad + \mathbb{E}_{\xi\sim\mathcal{N}(0,\mathbf{I})}\left[\min\left\{\exp\left(\langle \xi, b-a\rangle - \frac{1}{2}\|b-a\|^2\right), 1\right\}\mathbf{1}\left[\langle \xi, b-a\rangle \geq \frac{\|b-a\|^2}{4}\right]\right] \\
&\leq \exp\left(-\frac{1}{4}\|b-a\|^2\right) + \exp\left(-\frac{\|b-a\|^2}{32}\right) \leq 2\exp\left(-\frac{1}{32}\|b-a\|^2\right).
\end{aligned}$$

□

**Lemma B.9.** *For all $a, b, c \in \mathbb{R}^d$ such that $\|b-a\| < \|c-a\|$, we have*

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(a,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x-b\|^2 - \frac{1}{2}\|x-c\|^2\right), 1\right\}\right] \\
&\leq \exp\left(-\frac{1}{2}\left(\frac{1}{2}\|c-a\|^2 - \frac{1}{2}\|b-a\|^2\right)\right) + 2\exp\left(-\frac{\left(\frac{1}{2}\|c-a\|^2 - \frac{1}{2}\|b-a\|^2\right)^2}{2\|c-b\|^2}\right).
\end{aligned}$$

*Proof.* If $b = c$, then the bound is true. Assume that $b \neq c$. Then

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(a,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x-b\|^2 - \frac{1}{2}\|x-c\|^2\right), 1\right\}\right] \\
&= \mathbb{E}_{\xi\sim\mathcal{N}(0,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|\xi-(b-a)\|^2 - \frac{1}{2}\|\xi-(c-a)\|^2\right), 1\right\}\right] \\
&= \mathbb{E}_{\xi\sim\mathcal{N}(0,\mathbf{I})}\left[\min\left\{\exp\left(\langle \xi, c-b\rangle - \left(\frac{1}{2}\|c-a\|^2 - \frac{1}{2}\|b-a\|^2\right)\right), 1\right\}\right].
\end{aligned}$$

Note that $Z := \left\langle \xi, \frac{c-b}{\|c-b\|}\right\rangle \sim \mathcal{N}(0,1)$. We also define $\delta := \|c-b\|$ and $\kappa := \frac{1}{2}\|c-a\|^2 - \frac{1}{2}\|b-a\|^2 \geq 0$. Then,

$$\mathbb{E}_{x\sim\mathcal{N}(a,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x-b\|^2 - \frac{1}{2}\|x-c\|^2\right), 1\right\}\right] = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\min\left\{\exp\left(\delta Z - \kappa\right), 1\right\}\right].$$

Next,

$$\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\min\left\{\exp\left(\delta Z - \kappa\right), 1\right\}\right] \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\exp\left(\delta Z - \kappa\right)\mathbf{1}\left[\delta Z \leq \kappa\right]\right] + \mathbb{P}\left(\delta Z \geq \kappa\right) \\
&= e^{-\kappa}\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[e^{\delta Z}\mathbf{1}\left[Z \leq \frac{\kappa}{\delta}\right]\right] + \mathbb{P}\left(Z \geq \frac{\kappa}{\delta}\right) \\
&= e^{-\kappa+\delta^2/2}\mathbb{P}\left(Z \leq \frac{\kappa}{\delta} - \delta\right) + \mathbb{P}\left(Z \geq \frac{\kappa}{\delta}\right),
\end{aligned}$$

since

$$e^{-\kappa}\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[e^{\delta Z}\mathbf{1}\left[Z \leq \frac{\kappa}{\delta}\right]\right] = e^{-\kappa}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\kappa/\delta} e^{\delta z}e^{-z^2/2}\,dz = e^{-\kappa}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\kappa/\delta}\exp\left(-\frac{1}{2}(z-\delta)^2 + \frac{\delta^2}{2}\right)dz$$

$$= e^{-\kappa+\delta^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\kappa/\delta-\delta} e^{-u^2/2}\,du = e^{-\kappa+\delta^2/2}\mathbb{P}\left(Z \leq \frac{\kappa}{\delta} - \delta\right).$$

If $\kappa \geq \delta^2$, then the Gaussian tail bound gives

$$\begin{aligned}&\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\min\left\{\exp\left(\delta Z-\kappa\right),1\right\}\right]\\&=e^{-\kappa+\delta^2/2}\mathbb{P}\left(Z\leq\frac{\kappa}{\delta}-\delta\right)+\mathbb{P}\left(Z\geq\frac{\kappa}{\delta}\right)\leq e^{-\kappa/2}+e^{-\frac{\kappa^2}{2\delta^2}}.\end{aligned}$$

Otherwise,

$$\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\min\left\{\exp\left(\delta Z-\kappa\right),1\right\}\right]\leq e^{-\kappa+\delta^2/2-\frac{1}{2}\left(\frac{\kappa}{\delta}-\delta\right)^2}+e^{-\frac{\kappa^2}{2\delta^2}}=2e^{-\frac{\kappa^2}{2\delta^2}}.$$

Therefore,

$$\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\min\left\{\exp\left(\delta Z-\kappa\right),1\right\}\right]\leq e^{-\kappa/2}+2e^{-\frac{\kappa^2}{2\delta^2}}.$$

□

**Lemma B.10.** *Let $\alpha>0$ and let $\{z_k\}_{k=0}^{\bar{k}+1}$ be a non-negative sequence such that*

$$z_{k+1}\leq z_k-\alpha z_k^2$$

*for all $k\in\{0,\ldots,\bar{k}\}$. Then, for all $k\in\{1,\ldots,\bar{k}+1\}$,*

$$z_k\leq\frac{1}{\alpha k}.$$

*Proof.* Since

$$z_{k+1}\leq z_k-\alpha z_k^2,$$

and $z_{k+1}\geq 0$, we have

$$z_k-z_{k+1}\geq\alpha z_k^2.$$

In particular, $z_{k+1}\leq z_k$. Therefore, if $z_k=0$ for some $k$, then $z_{k+1}=0$, and the claim is trivial for all subsequent iterates. Let $\hat{k}+1$ be the first index such that $z_{\hat{k}+1}=0$. If it does not exist, then we take $\hat{k}=\infty$. If $\hat{k}+1=0$, then the claim is true. We now assume that $\hat{k}+1\geq 1$. For all $k\in\{0,\ldots,\min\{\hat{k}-1,\bar{k}\}\}$, dividing by $z_kz_{k+1}$, we get

$$\frac{1}{z_{k+1}}-\frac{1}{z_k}=\frac{z_k-z_{k+1}}{z_kz_{k+1}}\geq\alpha\frac{z_k}{z_{k+1}}\geq\alpha,$$

where we used $z_{k+1}\leq z_k$. Summing this inequality, we obtain

$$\frac{1}{z_{k+1}}\geq\frac{1}{z_0}+\alpha(k+1)$$

and

$$z_{k+1}\leq\frac{1}{\alpha(k+1)}$$

for all $k\in\{0,\ldots,\min\{\hat{k}-1,\bar{k}\}\}$. □

**Lemma B.11.** *Let $\alpha>0$, $\beta\in(0,1)$, and let $\{z_k\}_{k=0}^{\bar{k}+1}$ be a non-negative sequence such that*

$$z_{k+1}\leq z_k-\min\{\alpha z_k^2,\beta z_k\}$$

*for all $k\in\{0,\ldots,\bar{k}\}$. Then, for all $k\in\{1,\ldots,\bar{k}+1\}$,*

$$z_k\leq\max\left\{(1-\beta)^{k/2}z_0,\frac{2}{\alpha k}\right\}.$$

*Proof.* Since

$$z_{k+1} \leq z_k - \min\{\alpha z_k^2, \beta z_k\},$$

and $\min\{\alpha z_k^2, \beta z_k\} \geq 0$, we have

$$z_{k+1} \leq z_k.$$

Therefore, if $z_k = 0$ for some $k$, then $z_{k+1} = 0$, and the claim is trivial for all subsequent iterates. Fix $k \in \{1, \ldots, \bar{k} + 1\}$ and define

$$\tau := \frac{\beta}{\alpha}.$$

We distinguish two cases. First, suppose that $z_j \geq \tau$ for all $j \in \{0, \ldots, k-1\}$. Then $\alpha z_j^2 \geq \beta z_j$, and hence $\min\{\alpha z_j^2, \beta z_j\} = \beta z_j$. Therefore,

$$z_{j+1} \leq z_j - \beta z_j = (1-\beta) z_j.$$

Unrolling the recursion, we get

$$z_k \leq (1-\beta)^k z_0 \leq (1-\beta)^{k/2} z_0.$$

We now consider the case where there exists $j \in \{0, \ldots, k-1\}$ such that $z_j < \tau$. Let $m$ be the first such index. Then, for all $j \in \{0, \ldots, m-1\}$, we have $z_j \geq \tau$, and therefore, arguing as above,

$$z_m \leq (1-\beta)^m z_0.$$

Since the sequence is non-increasing, for all $j \geq m$, we have

$$z_j \leq z_m < \tau.$$

Hence, for all $j \in \{m, \ldots, k-1\}$, $\min\{\alpha z_j^2, \beta z_j\} = \alpha z_j^2$.

If $m > k/2$, then

$$z_k \leq z_m \leq (1-\beta)^m z_0 \leq (1-\beta)^{k/2} z_0.$$

It remains to consider the case $m \leq k/2$. If $z_j = 0$ for some $j \in \{m, \ldots, k\}$, then the claim is trivial. Thus, without loss of generality, we assume that $z_j > 0$ for all $j \in \{m, \ldots, k\}$. For all $j \in \{m, \ldots, k-1\}$, we have

$$z_{j+1} \leq z_j - \alpha z_j^2.$$

Hence,

$$z_j - z_{j+1} \geq \alpha z_j^2.$$

Dividing by $z_j z_{j+1}$, we get

$$\frac{1}{z_{j+1}} - \frac{1}{z_j} = \frac{z_j - z_{j+1}}{z_j z_{j+1}} \geq \alpha \frac{z_j}{z_{j+1}} \geq \alpha,$$

where we used $z_{j+1} \leq z_j$. Summing this inequality from $j = m$ to $j = k-1$, we obtain

$$\frac{1}{z_k} \geq \frac{1}{z_m} + \alpha(k-m) \geq \alpha(k-m).$$

Since $m \leq k/2$, we have $k - m \geq k/2$. Thus,

$$\frac{1}{z_k} \geq \frac{\alpha k}{2},$$

and therefore

$$z_k \leq \frac{2}{\alpha k}.$$

Combining the two cases, we obtain

$$z_k \leq \max\left\{(1-\beta)^{k/2} z_0, \frac{2}{\alpha k}\right\}.$$

□

**Lemma B.12.** *Let $a > 1$, $b > 1$, and*

$$x \geq 4a\sqrt{\max\{1, \log(a^4 b)\}}.$$

*Then*

$$x \geq a\sqrt{\log(bx^4)}.$$

*Proof.* Let

$$L := \max\{1, \log(a^4 b)\}.$$

Since $x \geq 4a\sqrt{L}$, we have

$$\frac{x^2}{a^2} \geq 16L.$$

Denote $y := x^2/a^2$. Then $y \geq 16L$ and

$$\log(bx^4) = \log(a^4 b y^2) = \log(a^4 b) + 2\log y \leq L + 2\log y.$$

Since $y \geq 16L \geq 16$, we have

$$2\log y \leq \frac{15}{16} y.$$

Moreover, $L \leq y/16$. Therefore,

$$\log(bx^4) \leq L + 2\log y \leq \frac{y}{16} + \frac{15y}{16} = y.$$

Thus

$$a\sqrt{\log(bx^4)} \leq a\sqrt{y} = x.$$

□

**Lemma B.13.** *Let $a \geq 1$, $b \geq 1$, and*

$$x \geq 4\sqrt{a\max\{1, \log(ab)\}}.$$

*Then*

$$x \geq \sqrt{a\log(bx^3)}.$$

*Proof.* Let

$$L := \max\{1, \log(ab)\}.$$

Since $x \geq 4\sqrt{aL}$, we have

$$\frac{x^2}{a} \geq 16L.$$

Denote $y := x^2/a$. Then $y \geq 16L$ and

$$\log(bx^3) = \log\left(ba^{3/2}y^{3/2}\right) = \log(ab) + \frac{1}{2}\log a + \frac{3}{2}\log y.$$

Since $b \geq 1$, we have $\log a \leq \log(ab) \leq L$, and therefore

$$\log(bx^3) \leq \frac{3}{2}L + \frac{3}{2}\log y.$$

Since $y \geq 16L \geq 16$, we have

$$\log y \leq \frac{1}{2} y.$$

Moreover, $L \leq y/16$. Therefore,

$$\log(bx^3) \leq \frac{3y}{32} + \frac{3y}{4} = \frac{27y}{32} \leq y.$$

Thus

$$\sqrt{a\log(bx^3)} \leq \sqrt{ay} = x.$$

□

# C GRADIENT OF THE REVERSE FISHER DIVERGENCE FOR GAUSSIAN MIXTURE MODELS

**Proposition 1.1** (Gradient of (1) with GMMs). *Let* $\nabla_\mu \mathcal{L}_{\text{rev}}(\mu) \equiv [\nabla_{\mu_1}\mathcal{L}_{\text{rev}}(\mu); \ldots; \nabla_{\mu_n}\mathcal{L}_{\text{rev}}(\mu)]$. *Then, for all* $\mu \in \mathbb{R}^{nd}$ *and* $\ell \in [n]$, *the gradient with respect to the* $\ell^{th}$ *mean component is*

$$\nabla_{\mu_\ell}\mathcal{L}_{\text{rev}}(\mu) = \frac{2}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[m_\mu(x) - \bar{m}(x) + \mathbf{C}_\mu(x)\big(\mu_\ell - m_\mu(x)\big) - \bar{\mathbf{C}}(x)\big(\mu_\ell - \bar{m}(x)\big)\right].$$

*Proof.* The loss (1) is differentiable. Thus, using standard differential calculus, it is sufficient to find a vector $a$ such that $d\mathcal{L}_{\text{rev}}(\mu) = \langle a, d\mu_\ell\rangle$, where $d\mathcal{L}_{\text{rev}}(\mu)$ is the differential with respect to $\mu_\ell$. Direct calculations yield

$$\mathcal{L}_{\text{rev}}(\mu) = \mathbb{E}_{x\sim p_\mu}\left[\|m_\mu(x) - \bar{m}(x)\|^2\right].$$

Thus

$$d\mathcal{L}_{\text{rev}}(\mu) = d\mathbb{E}_{x\sim p_\mu}\left[\|m_\mu(x) - \bar{m}(x)\|^2\right] = d\left(\int_{x\in\mathbb{R}^d}\|m_\mu(x) - \bar{m}(x)\|^2 p_\mu(x)\,dx\right).$$

Interchanging the order of integration and differentiation,

$$d\mathcal{L}_{\text{rev}}(\mu) = \int_{x\in\mathbb{R}^d}\left(d\|m_\mu(x) - \bar{m}(x)\|^2\right)p_\mu(x)\,dx + \int_{x\in\mathbb{R}^d}\|m_\mu(x) - \bar{m}(x)\|^2(dp_\mu(x))\,dx. \tag{13}$$

We now consider the terms separately. Clearly,

$$d\|m_\mu(x) - \bar{m}(x)\|^2 = 2\langle m_\mu(x) - \bar{m}(x), d(m_\mu(x) - \bar{m}(x))\rangle = 2\langle m_\mu(x) - \bar{m}(x), dm_\mu(x)\rangle \tag{14}$$

because $\bar{m}(x)$ does not depend on $\mu_\ell$. Next,

$$dm_\mu(x) = \sum_{i=1}^n (dr_i(x))\mu_i + \sum_{i=1}^n r_i(x)d\mu_i = \sum_{i=1}^n (dr_i(x))\mu_i + r_\ell(x)d\mu_\ell.$$

Since

$$dr_\ell(x) = d\left(\frac{\exp\left(-\frac{1}{2}\|x-\mu_\ell\|^2\right)}{\sum_{k=1}^n \exp\left(-\frac{1}{2}\|x-\mu_k\|^2\right)}\right) = r_\ell(x)(1-r_\ell(x))\langle x-\mu_\ell, d\mu_\ell\rangle.$$

and

$$dr_i(x) = d\left(\frac{\exp\left(-\frac{1}{2}\|x-\mu_i\|^2\right)}{\sum_{k=1}^n \exp\left(-\frac{1}{2}\|x-\mu_k\|^2\right)}\right) = -r_i(x)r_\ell(x)\langle x-\mu_\ell, d\mu_\ell\rangle$$

for all $i \neq \ell$, we get

$$\begin{aligned} dm_\mu(x) &= -\sum_{i=1}^n r_i(x)r_\ell(x)\langle x-\mu_\ell, d\mu_\ell\rangle\mu_i + r_\ell(x)\langle x-\mu_\ell, d\mu_\ell\rangle\mu_\ell + r_\ell(x)d\mu_\ell \\ &= -r_\ell(x)\langle x-\mu_\ell, d\mu_\ell\rangle m_\mu(x) + r_\ell(x)\langle x-\mu_\ell, d\mu_\ell\rangle\mu_\ell + r_\ell(x)d\mu_\ell \\ &= r_\ell(x)(\mu_\ell - m_\mu(x))\langle x-\mu_\ell, d\mu_\ell\rangle + r_\ell(x)d\mu_\ell \end{aligned}$$

due to the definition of $m_\mu(x)$. We now consider $dp_\mu(x)$ :

$$dp_\mu(x) = d\left(\frac{1}{n}\sum_{i=1}^n \varphi(x-\mu_i)\right) = \frac{1}{n}\varphi(x-\mu_\ell)\langle x-\mu_\ell, d\mu_\ell\rangle.$$

Substituting $dm_\mu(x)$ and $dp_\mu(x)$ into (14) and (13), we get

$$d\mathcal{L}_{\text{rev}}(\mu) = 2\int_{x\in\mathbb{R}^d}\langle m_\mu(x) - \bar{m}(x), r_\ell(x)(\mu_\ell - m_\mu(x))\langle x-\mu_\ell, d\mu_\ell\rangle + r_\ell(x)d\mu_\ell\rangle\, p_\mu(x)\,dx$$

$$+\frac{1}{n}\int_{x\in\mathbb{R}^d}\|m_\mu(x)-\bar{m}(x)\|^2\varphi(x-\mu_\ell)\,\langle x-\mu_\ell,d\mu_\ell\rangle\;dx.$$

Since $r_\ell(x)p_\mu(x)=\frac{\exp\left(-\frac{1}{2}\|x-\mu_\ell\|^2\right)}{\sum_{k=1}^n\exp\left(-\frac{1}{2}\|x-\mu_k\|^2\right)}\frac{1}{n}\sum_{i=1}^n\varphi(x-\mu_i)=\frac{1}{n}\varphi(x-\mu_\ell)$,

$$\begin{aligned}
d\mathcal{L}_{\text{rev}}(\mu)&=\frac{2}{n}\int_{x\in\mathbb{R}^d}\langle m_\mu(x)-\bar{m}(x),(\mu_\ell-m_\mu(x))\,\langle x-\mu_\ell,d\mu_\ell\rangle+d\mu_\ell\rangle\,\varphi(x-\mu_\ell)\,dx\\
&\quad+\frac{1}{n}\int_{x\in\mathbb{R}^d}\|m_\mu(x)-\bar{m}(x)\|^2\,\langle x-\mu_\ell,d\mu_\ell\rangle\,\varphi(x-\mu_\ell)\,dx\\
&=\frac{2}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[\langle m_\mu(x)-\bar{m}(x),(\mu_\ell-m_\mu(x))\,\langle x-\mu_\ell,d\mu_\ell\rangle+d\mu_\ell\rangle\right]\\
&\quad+\frac{1}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[\|m_\mu(x)-\bar{m}(x)\|^2\,\langle x-\mu_\ell,d\mu_\ell\rangle\right]\\
&=\left\langle\frac{2}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[m_\mu(x)-\bar{m}(x)\right],d\mu_\ell\right\rangle\\
&\quad+\left\langle\frac{1}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[\left(-\|m_\mu(x)\|^2+\|\bar{m}(x)\|^2+2\,\langle m_\mu(x)-\bar{m}(x),\mu_\ell\rangle\right)(x-\mu_\ell)\right],d\mu_\ell\right\rangle.
\end{aligned}$$

Using Stein's identity,

$$\begin{aligned}
d\mathcal{L}_{\text{rev}}(\mu)&=\left\langle\frac{2}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[m_\mu(x)-\bar{m}(x)\right],d\mu_\ell\right\rangle\\
&\quad+\left\langle\frac{1}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[\nabla_x\left(-\|m_\mu(x)\|^2+\|\bar{m}(x)\|^2+2\,\langle m_\mu(x)-\bar{m}(x),\mu_\ell\rangle\right)\right],d\mu_\ell\right\rangle.
\end{aligned}\tag{15}$$

Let $d_x$ be the differential with respect to $x$. Then,

$$\begin{aligned}
&d_x\left(-\|m_\mu(x)\|^2+\|\bar{m}(x)\|^2+2\,\langle m_\mu(x)-\bar{m}(x),\mu_\ell\rangle\right)\\
&=\left(-2\,\langle m_\mu(x),d_xm_\mu(x)\rangle+2\,\langle\bar{m}(x),d_x\bar{m}(x)\rangle+2\,\langle d_xm_\mu(x)-d_x\bar{m}(x),\mu_\ell\rangle\right).
\end{aligned}\tag{16}$$

Next,

$$\begin{aligned}
d_xm_\mu(x)&=\sum_{i=1}^n d_xr_i(x)\mu_i=\sum_{i=1}^n\left(d_x\left(\frac{\exp\left(-\frac{1}{2}\|x-\mu_i\|^2\right)}{\sum_{k=1}^n\exp\left(-\frac{1}{2}\|x-\mu_k\|^2\right)}\right)\right)\mu_i\\
&=\sum_{i=1}^n\left(-r_i(x)\,\langle x-\mu_i,dx\rangle+r_i(x)\left(\sum_{k=1}^n r_k(x)\,\langle x-\mu_k,dx\rangle\right)\right)\mu_i\\
&=\sum_{i=1}^n\left(-r_i(x)\,\langle x-\mu_i,dx\rangle+r_i(x)\,\langle x-m_\mu(x),dx\rangle\right)\mu_i\\
&=\sum_{i=1}^n\left(r_i(x)\,\langle\mu_i-m_\mu(x),dx\rangle\right)\mu_i\\
&=\sum_{i=1}^n r_i(x)(\mu_i-m_\mu(x))(\mu_i-m_\mu(x))^\top dx=\mathbf{C}_\mu(x)dx
\end{aligned}$$

because $\sum_{i=1}^n r_i(x)m_\mu(x)(\mu_i-m_\mu(x))=0$. Similarly,

$$d_x\bar{m}(x)=\sum_{i=1}^m d_x\bar{r}_i(x)\bar{\mu}_i=\bar{\mathbf{C}}(x)dx.$$

Substituting into (16),

$$\begin{aligned}
&d_x\left(-\|m_\mu(x)\|^2+\|\bar{m}(x)\|^2+2\,\langle m_\mu(x)-\bar{m}(x),\mu_\ell\rangle\right)\\
&=-2\,\langle m_\mu(x),\mathbf{C}_\mu(x)dx\rangle+2\,\langle\bar{m}(x),\bar{\mathbf{C}}(x)dx\rangle+2\,\langle\mathbf{C}_\mu(x)dx-\bar{\mathbf{C}}(x)dx,\mu_\ell\rangle
\end{aligned}$$

$$= \left\langle 2\mathbf{C}_\mu(x)(\mu_\ell - m_\mu(x)) - 2\bar{\mathbf{C}}(x)(\mu_\ell - \bar{m}(x)), dx \right\rangle .$$

We have found the gradient that we substitute into (15):

$$d\mathcal{L}_{\text{rev}}(\mu) = \left\langle \frac{2}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[m_\mu(x) - \bar{m}(x) + \mathbf{C}_\mu(x)(\mu_\ell - m_\mu(x)) - \bar{\mathbf{C}}(x)(\mu_\ell - \bar{m}(x))\right], d\mu_\ell \right\rangle,$$

which yields the result. □

## D MISSING STEPS IN THE PROOF OF THEOREM 2.1

The full proof of Theorem 2.1 is provided in Section 2. Here, we provide auxiliary lemmas. The following lemma rewrites the GD steps in matrix format.

**Lemma 2.2.** *Let $m = 1$ and $\mathbf{Z}^k := \left[(\mu_1^k - \bar{\mu}_1)^\top; \ldots; (\mu_n^k - \bar{\mu}_1)^\top\right] \in \mathbb{R}^{n\times d}$, stacking the vectors vertically. Using this notation and (3), the matrix sequence $\{\mathbf{Z}^k\}_{k\geq 0}$ evolves as*

$$\mathbf{Z}^{k+1} = \left(\mathbf{I} - \tfrac{2\gamma}{n}\mathbf{B}^k\right)\mathbf{Z}^k,$$

*where $\mathbf{B}^k \in \mathbb{R}^{n\times n}$ and $\mathbf{B}_{\ell i}^k := \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[r_i^k(x) + r_i^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right]$ is the $(\ell,i)^{th}$ entry of $\mathbf{B}^k$ and $r_i^k(x) := \exp\left(-\frac{1}{2}\|x-\mu_i^k\|^2\right) / \sum_{j=1}^n \exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)$.*

*Proof.* Using Proposition 1.1, in the case when $m = 1$,

$$\nabla_{\mu_\ell}\mathcal{L}_{\text{rev}}(\mu) = \frac{2}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell,\mathbf{I})}\left[m_\mu(x) - \bar{\mu}_1 + \mathbf{C}_\mu(x)\left(\mu_\ell - m_\mu(x)\right)\right]$$

since $\bar{\mathbf{C}}(x) = 0$ and $\bar{m}(x) = \bar{\mu}_1$. Then, using (3), for all $\ell \in [n]$,

$$\begin{aligned}
&\mu_\ell^{k+1} - \bar{\mu}_1 \\
&= \mu_\ell^k - \bar{\mu}_1 - \gamma\nabla_{\mu_\ell^k}\mathcal{L}_{\text{rev}}(\mu^k) \\
&= \mu_\ell^k - \bar{\mu}_1 - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[m_{\mu^k}(x) - \bar{\mu}_1 + \mathbf{C}^k(x)\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right] \\
&= \mu_\ell^k - \bar{\mu}_1 - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\sum_{i=1}^n r_i^k(x)(\mu_i^k - \bar{\mu}_1) + \sum_{i=1}^n r_i^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right].
\end{aligned}$$

Since $\sum_{i=1}^n r_i^k(x) b\left(\mu_i^k - m_{\mu^k}(x)\right)^\top = 0$ for all $b \in \mathbb{R}^d$,

$$\begin{aligned}
\mu_\ell^{k+1} - \bar{\mu}_1 &= \mu_\ell^k - \bar{\mu}_1 - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\sum_{i=1}^n r_i^k(x)(\mu_i^k - \bar{\mu}_1) + \sum_{i=1}^n r_i^k(x)\left(\mu_i^k - \bar{\mu}_1\right)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right] \\
&= \mu_\ell^k - \bar{\mu}_1 - \frac{2\gamma}{n}\sum_{i=1}^n \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[r_i^k(x) + r_i^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right](\mu_i^k - \bar{\mu}_1).
\end{aligned}$$

We can conclude that

$$\mu_\ell^{k+1} - \bar{\mu}_1 = \mu_\ell^k - \bar{\mu}_1 - \frac{2\gamma}{n}\sum_{i=1}^n \mathbf{B}_{\ell i}^k(\mu_i^k - \bar{\mu}_1).$$

□

**Lemma 2.3.** *For all $k \geq 0$, matrix $\mathbf{B}^k$ is symmetric and positive semidefinite, and $\left\|\mathbf{B}^k\right\|_{\text{op}} \leq 1 + 128\log(n)$.*

*Proof.* Let us define $p_\mu^k(x) := \frac{1}{n}\sum_{i=1}^n \varphi(x - \mu_i^k)$. Then $np_\mu^k(x)r_j^k(x) = \varphi(x - \mu_j^k)$ for all $j \in [n]$ and

$$\begin{aligned}
\mathbf{B}_{\ell i}^k &:= \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[r_i^k(x) + r_i^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right] \\
&= \int_{x\in\mathbb{R}^d}\left[r_i^k(x) + r_i^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right]\varphi(x - \mu_\ell^k)\,dx \\
&= n\int_{x\in\mathbb{R}^d}\left[r_i^k(x)r_\ell^k(x) + r_i^k(x)r_\ell^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right]p_\mu^k(x)\,dx \\
&= n\mathbb{E}_{x\sim p_\mu^k}\left[r_i^k(x)r_\ell^k(x) + r_i^k(x)r_\ell^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_\ell^k - m_{\mu^k}(x)\right)\right],
\end{aligned} \tag{17}$$

where $p_\mu^k(x) := \frac{1}{n}\sum_{i=1}^n \varphi(x-\mu_i^k)$. Thus, $\mathbf{B}^k$ is symmetric. For all $v \in \mathbb{R}^n$,

$$v^\top \mathbf{B}^k v = n\int_{x\in\mathbb{R}^d}\left[\left(\sum_{j=1}^n r_j^k(x)v_j\right)^2 + \left\|\sum_{j=1}^n r_j^k(x)v_j(\mu_j^k - m_{\mu^k}(x))\right\|^2\right] p_\mu^k(x)\,dx,$$

so it is positive semidefinite. Using Jensen's inequality,

$$\begin{aligned}
v^\top \mathbf{B}^k v &\le n\int_{x\in\mathbb{R}^d}\left[\sum_{j=1}^n r_j^k(x)v_j^2 + \sum_{j=1}^n r_j^k(x)v_j^2\left\|\mu_j^k - m_{\mu^k}(x)\right\|^2\right] p_\mu^k(x)\,dx \\
&= \sum_{j=1}^n\left[\int_{x\in\mathbb{R}^d} np_\mu^k(x)r_j^k(x)\,dx\right]v_j^2 + \sum_{j=1}^n\left[\int_{x\in\mathbb{R}^d}\left\|\mu_j^k - m_{\mu^k}(x)\right\|^2 np_\mu^k(x)r_j^k(x)\,dx\right]v_j^2 \\
&= \sum_{j=1}^n\left[\int_{x\in\mathbb{R}^d} np_\mu^k(x)r_j^k(x)\,dx\right]v_j^2 + \sum_{j=1}^n \mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[\left\|\mu_j^k - m_{\mu^k}(x)\right\|^2\right]v_j^2 \\
&\le \sum_{j=1}^n v_j^2 + 128\log(n)\sum_{j=1}^n v_j^2 = (1+128\log(n))\|v\|^2
\end{aligned}$$

since $np_\mu^k(x)r_j^k(x) = \varphi(x-\mu_j^k)$. The last inequality follows from the following steps. Let $S := \{i\in[n] : \left\|\mu_j^k - \mu_i^k\right\|^2 > 64\log(n)\}$. Assume $n \ge 2$; the case $n = 1$ is trivial. Then

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[\left\|\mu_j^k - m_{\mu^k}(x)\right\|^2\right] = \mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[\left\|\mu_j^k - \sum_{i=1}^n r_i^k(x)\mu_i^k\right\|^2\right] \\
&\le \sum_{i=1}^n \mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[r_i^k(x)\right]\left\|\mu_j^k - \mu_i^k\right\|^2 \\
&= \sum_{i\in S}\mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[r_i^k(x)\right]\left\|\mu_j^k - \mu_i^k\right\|^2 + \sum_{i\notin S}\mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[r_i^k(x)\right]\left\|\mu_j^k - \mu_i^k\right\|^2 \\
&\le \sum_{i\in S}\mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x-\mu_j^k\|^2 - \frac{1}{2}\|x-\mu_i^k\|^2\right), 1\right\}\right]\left\|\mu_j^k - \mu_i^k\right\|^2 + 64\log(n) \\
&\overset{\text{L.B.8}}{\le} 2\sum_{i\in S}\exp\left(-\frac{1}{32}\left\|\mu_j^k - \mu_i^k\right\|^2\right)\left\|\mu_j^k - \mu_i^k\right\|^2 + 64\log(n) \\
&\le 128n\log(n)\exp\left(-2\log(n)\right) + 64\log(n),
\end{aligned}$$

where in the second inequality we use the definition of $r_i^k(x)$, and in the last inequality we use that $\exp\left(-\frac{x}{32}\right)x$ is decreasing for all $x \ge 32$. Thus, $\mathbb{E}_{x\sim\mathcal{N}(\mu_j^k,\mathbf{I})}\left[\left\|\mu_j^k - m_{\mu^k}(x)\right\|^2\right] \le 128\log(n)$. □

**Lemma 2.4.** *For all* $k \ge 0$, $\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle \ge \min\left\{\frac{1}{4n\min\{n,d\}}\left\|\mathbf{Z}^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}^k\right\|_F^2\right\}$.

*Proof.* Let $p_\mu^k(x) := \frac{1}{n}\sum_{i=1}^n \varphi(x-\mu_i^k)$. Using Lemma 2.2 and (17),

$$\begin{aligned}
&\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle \\
&= n\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}_{x\sim p_\mu^k}\left[r_i^k(x)r_j^k(x) + r_i^k(x)r_j^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\mu_j^k - m_{\mu^k}(x)\right)\right](\mu_i^k - \bar\mu_1)^\top(\mu_j^k - \bar\mu_1) \\
&= n\mathbb{E}_{x\sim p_\mu^k}\left[\sum_{i=1}^n r_i^k(x)(\mu_i^k - \bar\mu_1)^\top\left(\sum_{j=1}^n r_j^k(x)(\mu_j^k - \bar\mu_1)\right)\right] \\
&\quad + n\mathbb{E}_{x\sim p_\mu^k}\left[\sum_{i=1}^n r_i^k(x)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\left(\sum_{j=1}^n r_j^k(x)\left(\mu_j^k - m_{\mu^k}(x)\right)(\mu_j^k - \bar\mu_1)^\top\right)(\mu_i^k - \bar\mu_1)\right]
\end{aligned}$$

$$= n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|\sum_{i=1}^n r_i^k(x)(\mu_i^k-\bar{\mu}_1)\right\|^2\right] + n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|\sum_{i=1}^n r_i^k(x)\big(\mu_i^k - m_{\mu^k}(x)\big)(\mu_i^k-\bar{\mu}_1)^\top\right\|_F^2\right].$$

Using the definition of $m_{\mu^k}(x)$ and $\sum_{i=1}^n r_i^k(x)\big(\mu_i^k - m_{\mu^k}(x)\big)b^\top = 0$ for all $b\in\mathbb{R}^d$,

$$\begin{aligned}
&\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle \\
&= n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right] + n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|\sum_{i=1}^n r_i^k(x)\big(\mu_i^k - m_{\mu^k}(x)\big)\big(\mu_i^k - m_{\mu^k}(x)\big)^\top\right\|_F^2\right] \\
&= n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right] + n\mathbb{E}_{x\sim p_\mu^k}\mathrm{tr}\left(\left(\sum_{i=1}^n r_i^k(x)\big(\mu_i^k - m_{\mu^k}(x)\big)\big(\mu_i^k - m_{\mu^k}(x)\big)^\top\right)^2\right).
\end{aligned}$$

Let $\kappa := \frac{n}{\min\{n,d\}}$. Due to $(\mathrm{tr}(\mathbf{A}))^2 \le \mathrm{rank}(\mathbf{A})\mathrm{tr}(\mathbf{A}^2)$ for all symmetric $\mathbf{A}\in\mathbb{R}^{d\times d}$, we get

$$\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle \ge n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right] + \kappa\mathbb{E}_{x\sim p_\mu^k}\left[\left(\sum_{i=1}^n r_i^k(x)\left\|\mu_i^k - m_{\mu^k}(x)\right\|^2\right)^2\right],$$

where the rank of our matrix under the trace is $\le \min\{n,d\}$. Using Jensen's inequality,

$$\begin{aligned}
\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle &\ge n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right] + \kappa\mathbb{E}_{x\sim p_\mu^k}\left[\left(\sum_{i=1}^n r_i^k(x)\left\|\mu_i^k-\bar{\mu}_1\right\|^2 - \left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right)^2\right] \\
&\ge n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right] + \kappa\left(\mathbb{E}_{x\sim p_\mu^k}\left[\sum_{i=1}^n r_i^k(x)\left\|\mu_i^k-\bar{\mu}_1\right\|^2\right] - \mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right]\right)^2 \\
&= n\mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right] + \kappa\left(\frac{1}{n}\left\|\mathbf{Z}^k\right\|_F^2 - \mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right]\right)^2
\end{aligned}$$

since

$$\begin{aligned}
&\mathbb{E}_{x\sim p_\mu^k}\left[\sum_{i=1}^n r_i^k(x)\left\|\mu_i^k-\bar{\mu}_1\right\|^2\right] = \sum_{i=1}^n\int_{x\in\mathbb{R}^d}\left[r_i^k(x)p_\mu^k(x)\left\|\mu_i^k-\bar{\mu}_1\right\|^2\right]dx \\
&= \frac{1}{n}\sum_{i=1}^n\int_{x\in\mathbb{R}^d}\left[\left\|\mu_i^k-\bar{\mu}_1\right\|^2\right]\varphi(x-\mu_i^k)\,dx = \frac{1}{n}\sum_{i=1}^n\left\|\mu_i^k-\bar{\mu}_1\right\|^2 = \frac{1}{n}\left\|\mathbf{Z}^k\right\|_F^2
\end{aligned}$$

Notice also that

$$\mathbb{E}_{x\sim p_\mu^k}\left[\left\|m_{\mu^k}(x)-\bar{\mu}_1\right\|^2\right] \le \mathbb{E}_{x\sim p_\mu^k}\left[\sum_{i=1}^n r_i^k(x)\left\|\mu_i^k-\bar{\mu}_1\right\|^2\right] = \frac{1}{n}\left\|\mathbf{Z}^k\right\|_F^2.$$

Thus,

$$\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle \ge \min_{0\le x\le\bar{x}}\left[nx + \kappa\left(\bar{x}-x\right)^2\right], \tag{18}$$

where $\bar{x} := \frac{1}{n}\left\|\mathbf{Z}^k\right\|_F^2$. Since

$$\min_{0\le x\le\bar{x}}\left[nx+\kappa\left(\bar{x}-x\right)^2\right] \ge \min\left\{\min_{0\le x\le\frac{\bar{x}}{2}}\kappa\left(\bar{x}-x\right)^2, \min_{\frac{\bar{x}}{2}\le x\le\bar{x}} nx\right\} = \min\left\{\frac{\kappa}{4}\bar{x}^2, \frac{n}{2}\bar{x}\right\},$$

we obtain

$$\left\langle \mathbf{Z}^k, \mathbf{B}^k\mathbf{Z}^k\right\rangle \ge \min\left\{\frac{\kappa}{4n^2}\left\|\mathbf{Z}^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}^k\right\|_F^2\right\} = \min\left\{\frac{1}{4n\min\{n,d\}}\left\|\mathbf{Z}^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}^k\right\|_F^2\right\}.$$

□

# E PROOF OF THEOREM E.1

**Theorem E.1** (Main Result for Multiple Target Modes). *Let $\varepsilon \in (0,1]$ denote the target accuracy. We define the smallest and largest gaps in the target distribution as*

$$\delta_{\min} := \min_{a \neq b} \|\bar{\mu}_a - \bar{\mu}_b\| > 0 \quad \text{and} \quad \delta_{\max} := \max_{a \neq b} \|\bar{\mu}_a - \bar{\mu}_b\|,$$

*and the initialization radius as*

$$R^0 \geq \max\left\{50 + \frac{2^{20} n m^2 \sqrt{d} \delta_{\max}^2}{\rho}, 2^{12}\sqrt{n \max\{\log\left(2^{66} n^3 \bar{R}^4\right), 1\}}\bar{R}\right\},$$

*where $\bar{R} := \max_{b\in[m]} \|\bar{\mu}_b\|$ and $\rho \in (0,1]$. We sample the initial parameters as $\mu_\ell^0 = R^0 \eta_\ell / \|\eta_\ell\|$, where $\eta_\ell \sim \mathcal{N}(0, \mathbf{I})$ for all $\ell \in [n]$, and the random vectors $\{\eta_\ell\}_{\ell=1}^n$ are i.i.d. Assume that the minimal gap satisfies*

$$\delta_{\min} \geq 2^{10}\sqrt{\log\left(\frac{2^{20} n^3 \max\{n,m\} (R^0)^4}{\varepsilon^2}\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}$$

*and $d \geq 8 + \max\left\{2\log\left(\frac{mn^2}{\rho}\right), 8\log\left(\frac{n^4}{2\pi\rho^2}\right)\right\}$. If $\gamma \leq \frac{n}{36\tilde{z}^2}$, where $\tilde{z} := 65\sqrt{\log\left(2^{16} n (R^0)^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}$, and GD (3) is run for*

$$\tilde{K} := \frac{n \log\left(R^0\right)}{2\gamma} + \frac{2n \min\{n,d\} \max\left\{4n, 8\tilde{z}^2\right\}}{\gamma\varepsilon}$$

*iterations, then, with probability at least $1 - 4\rho$, we have that*

$$\left\|\mu_\ell^k - \bar{\mu}_{a_\ell^*}\right\|^2 \leq \varepsilon,$$

*for all $\ell \in [n]$ and $k \geq \tilde{K}$, where $a_\ell^* = \arg\min_{a\in[m]} \left\|\mu_\ell^0 - \bar{\mu}_a\right\|$ is the index of the target mode closest to the initialization point $\mu_\ell^0$.*

*Proof.* Let

$$I_a := \{i \in [n] : a_i^* = a\},$$

where $a_i^* = \arg\min_{a\in[m]} \left\|\mu_i^0 - \bar{\mu}_a\right\|$, $n_a := |I_a|$, and $A^*$ is the set of $a \in [m]$ such that $I_a$ is non-empty. We have the initialization $\mu_\ell^0 = R^0 \eta_\ell / \|\eta_\ell\|$, where $\eta_\ell \sim \mathcal{N}(0, \mathbf{I})$ for all $\ell \in [n]$ and $\{\eta_\ell\}$ are i.i.d. Consider an event $\Omega = \Omega_1 \bigcap \Omega_2 \bigcap \Omega_3 \bigcap \Omega_4$, where

$$\Omega_1 := \left\{\forall i \in [n] : \left\|\mu_i^0 - \bar{\mu}_{a_i^*}\right\|^2 \leq \min_{b \neq a_i^*} \left\|\mu_i^0 - \bar{\mu}_b\right\|^2 - \Delta_1\right\}, \tag{19}$$

$$\Omega_2 := \left\{\forall i \neq j \in [n] : \left\|\mu_i^0 - \mu_j^0\right\|^2 \geq \Delta_2\right\}, \tag{20}$$

$$\Omega_3 := \left\{\forall i \neq j \in [n] \text{ s.t. } a_i^* \neq a_j^* : \text{dist}([\mu_i^0, \bar{\mu}_{a_i^*}], [\mu_j^0, \bar{\mu}_{a_j^*}]) \geq \Delta_3\right\}, \tag{21}$$

$$\Omega_4 := \left\{\forall a \in A^* : \left\|\begin{bmatrix} \left(\mu_{i_{a,1}}^0\right)^\top \\ \vdots \\ \left(\mu_{i_{a,n_a}}^0\right)^\top \end{bmatrix}\right\|_{\text{op}} \leq \Delta_4\right\}, \qquad I_a \equiv \{i_{a,1}, \ldots, i_{a,n_a}\}. \tag{22}$$

Using Lemmas B.2, B.3, B.4, and B.6, we get $\mathbb{P}(\Omega) \geq 1 - 4\rho$ with $\Delta_1 = \frac{2R^0 \delta_{\min} \rho}{nm^2\sqrt{d}}$, $\Delta_2 = (R^0)^2 \left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\, n^2}\right)^{\frac{2}{d-1}}$, $\Delta_3 = \frac{\delta_{\min}}{8}\left(\frac{\rho}{mn^2}\right)^{\frac{1}{d-2}}$, and

$$\Delta_4 = R^0\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}.$$

Starting from here, we assume that the event $\Omega$ holds. We now fix any $\ell \in [n]$. Using Proposition 1.1, we have

$$\nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k) = \frac{2}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ m_{\mu^k}(x) - \bar{m}(x) + \mathbf{C}_{\mu^k}(x)\big(\mu_\ell^k - m_{\mu^k}(x)\big) - \bar{\mathbf{C}}(x)\big(\mu_\ell^k - \bar{m}(x)\big) \right].$$

We define

$$r_i^k(x) := \frac{\exp\left(-\frac{1}{2}\|x - \mu_i^k\|^2\right)}{\sum_{j \in [n]} \exp\left(-\frac{1}{2}\|x - \mu_j^k\|^2\right)}.$$

For all $i \in I_a$, we also define

$$r_i^{k,a}(x) := \frac{\exp\left(-\frac{1}{2}\|x - \mu_i^k\|^2\right)}{\sum_{j \in I_a} \exp\left(-\frac{1}{2}\|x - \mu_j^k\|^2\right)}, \qquad m_a^k(x) := \sum_{i \in I_a} r_i^{k,a}(x) \mu_i^k,$$

and

$$\mathbf{C}_a^k(x) := \sum_{i \in I_a} r_i^{k,a}(x)\big(\mu_i^k - m_a^k(x)\big)\big(\mu_i^k - m_a^k(x)\big)^\top.$$

We now add and subtract $m_{a_\ell^*}^k(x) - \bar{\mu}_{a_\ell^*} + \mathbf{C}_{a_\ell^*}^k(x)\big(\mu_\ell^k - m_{a_\ell^*}^k(x)\big)$, to get

$$\nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k) = \frac{2}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ m_{a_\ell^*}^k(x) - \bar{\mu}_{a_\ell^*} + \mathbf{C}_{a_\ell^*}^k(x)\big(\mu_\ell^k - m_{a_\ell^*}^k(x)\big) + \mathcal{E}_\ell^k(x) \right], \tag{23}$$

where we define

$$\begin{aligned}\mathcal{E}_\ell^k(x) := {} & (m_{\mu^k}(x) - m_{a_\ell^*}^k(x)) - (\bar{m}(x) - \bar{\mu}_{a_\ell^*}) + (\mathbf{C}_{\mu^k}(x) - \mathbf{C}_{a_\ell^*}^k(x))(\mu_\ell^k - m_{a_\ell^*}^k(x)) \\ & + \mathbf{C}_{\mu^k}(x)(m_{a_\ell^*}^k(x) - m_{\mu^k}(x)) - \bar{\mathbf{C}}(x)\big(\mu_\ell^k - \bar{m}(x)\big).\end{aligned}$$

Let $\hat{\tau}_a^k(x) := \sum_{i \in [n], i \notin I_a} r_i^k(x)$ and $\bar{\tau}_a(x) := \sum_{b \in [m], b \neq a} \bar{r}_b(x)$. In Section E.1, we prove the following Lemma E.2.

**Lemma E.2.** *Let* $R_{\max} := \max\{\max_{i \in [n]} \left\|\mu_i^k\right\|, \max_{a \in [m]} \|\bar{\mu}_a\|\}$ *for some* $k \geq 0$. *Then,*

$$\left\|\mathcal{E}_\ell^k(x)\right\| \leq 52 \max\{R_{\max}, R_{\max}^3\} \max\{\hat{\tau}_{a_\ell^*}^k(x), \bar{\tau}_{a_\ell^*}(x)\}.$$

*for all* $x \in \mathbb{R}^d$ *and* $\ell \in [n]$.

Next, we add and subtract $\mu_\ell^k - \bar{\mu}_{a_\ell^*}$ once again to get

$$\nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k) = \frac{2}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ \mu_\ell^k - \bar{\mu}_{a_\ell^*} + \bar{\mathcal{E}}_\ell^k(x) + \mathcal{E}_\ell^k(x) \right], \tag{24}$$

where

$$\bar{\mathcal{E}}_\ell^k(x) := m_{a_\ell^*}^k(x) - \mu_\ell^k + \mathbf{C}_{a_\ell^*}^k(x)\big(\mu_\ell^k - m_{a_\ell^*}^k(x)\big).$$

Let $\tau_\ell^k(x) := \sum_{i \in I_{a_\ell^*}, i \neq \ell} r_i^{k,a_\ell^*}(x)$. In Section E.1, we prove the following Lemma E.3.

**Lemma E.3.** *Let* $R_{\max} := \max\{\max_{i \in [n]} \left\|\mu_i^k\right\|, \max_{a \in [m]} \|\bar{\mu}_a\|\}$ *for some* $k \geq 0$. *Then,*

$$\left\|\bar{\mathcal{E}}_\ell^k(x)\right\| \leq 10 \max\{R_{\max}, R_{\max}^3\} \tau_\ell^k(x).$$

*for all* $x \in \mathbb{R}^d$ *and* $\ell \in [n]$.

**(Global convergence stage).** Using (24) and the update rule $\mu_\ell^{k+1} = \mu_\ell^k - \gamma \nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k)$,

$$\mu_\ell^{k+1} = \bar{\mu}_{a_\ell^*} + \left(1 - \frac{2\gamma}{n}\right)^{k+1} (\mu_\ell^0 - \bar{\mu}_{a_\ell^*}) - E_\ell^k, \tag{25}$$

where

$$E_\ell^k := \frac{2\gamma}{n} \sum_{j=0}^{k} \left(1 - \frac{2\gamma}{n}\right)^{k-j} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^j, \mathbf{I})} \left[ \bar{\mathcal{E}}_\ell^j(x) + \mathcal{E}_\ell^j(x) \right]$$

and

$$\left\|E_\ell^k\right\| \leq \max_{j\in\{0,\ldots,k\}} \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^j,\mathbf{I})}\left[\left\|\bar{\mathcal{E}}_\ell^j(x)+\mathcal{E}_\ell^j(x)\right\|\right]. \tag{26}$$

Let

$$\bar{K} := \left\lfloor \log\left(\frac{R^0}{4\sqrt{\bar{\Delta}}}\right) / \left(-\log\left(1-\frac{2\gamma}{n}\right)\right)\right\rfloor, \tag{27}$$

where $\bar{\Delta} > 0$ is a free parameter. Then

$$\left(1-\frac{2\gamma}{n}\right)^k \left\|\mu_\ell^0 - \bar{\mu}_{a_\ell^*}\right\| \geq \left(1-\frac{2\gamma}{n}\right)^k \frac{R^0}{2} \geq 2\sqrt{\bar{\Delta}} \tag{28}$$

for all $0 \leq k \leq \bar{K}$. Using mathematical induction, for all $0 \leq k \leq \bar{K}$, we assume that

$$\left\|\mu_\ell^k\right\| \leq R, \tag{29}$$

$$\left\|\mu_\ell^k - \bar{\mu}_{a_\ell^*}\right\|^2 \leq \min_{b\neq a_\ell^*}\left[\left\|\mu_\ell^k-\bar{\mu}_b\right\|^2 - \frac{\left\|\bar{\mu}_{a_\ell^*}-\bar{\mu}_b\right\|^2}{2}\right], \tag{30}$$

$$\left\|\mu_\ell^k - \mu_j^k\right\|^2 \geq \hat{\Delta}_2 \tag{31}$$

for all $\ell \neq j \in [n]$, where

$$R := 2R^0, \quad E_1 := \frac{\delta_{\min}}{8R}, \quad \hat{\Delta}_2 := 32\log\left(3472nR^4\right), \quad \bar{\Delta} := \frac{\hat{\Delta}_2}{2}. \tag{32}$$

For our choice of parameters, we prove the following useful lemma in Section E.1.

**Lemma E.4.** *For our choice of parameters,*

$$\begin{aligned} &\Delta_1 \geq \delta_{\max}^2, \quad \Delta_2 \geq \hat{\Delta}_2 \geq 32\log(434nR^3/E_1), \quad \delta_{\min} \geq \sqrt{32\log\left(434R^3m/E_1\right)}, \\ &E_1 \leq \frac{R}{2}, \quad \frac{4\sqrt{\bar{\Delta}\Delta_2}}{R^0} \geq 2\sqrt{\hat{\Delta}_2}, \quad \Delta_3 \geq 2\sqrt{\hat{\Delta}_2}, \quad E_1 \leq \frac{\sqrt{\hat{\Delta}_2}}{2}. \end{aligned} \tag{33}$$

The case $k = 0$ of the mathematical induction is true: i) (29) follows from $\left\|\mu_\ell^0\right\| = R^0$; ii) (30) follows from (19) and $\Delta_1 \geq \delta_{\max}^2$ (see (33)); iii) (31) follows from (20) and $\hat{\Delta}_2 \leq \Delta_2$ (see (33)). We now prove the next step of the induction $k \to k+1$. For all $\ell \in [n]$, using Lemmas E.2 and E.3,

$$\begin{aligned} &\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\left\|\bar{\mathcal{E}}_\ell^k(x)+\mathcal{E}_\ell^k(x)\right\|\right] \leq \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left(\left\|\bar{\mathcal{E}}_\ell^k(x)\right\| + \left\|\mathcal{E}_\ell^k(x)\right\|\right) \\ &\leq 62\max\{R,R^3\}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}(\hat{\tau}_{a_\ell^*}^k(x) + \bar{\tau}_{a_\ell^*}(x) + \tau_\ell^k(x)) \end{aligned} \tag{34}$$

First,

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\tau_\ell^k(x) = \sum_{i\in I_{a_\ell^*}, i\neq\ell} \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})} r_i^{k,a_\ell^*}(x), \tag{35}$$

where

$$\begin{aligned} \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})} r_i^{k,a_\ell^*}(x) &\leq \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x-\mu_\ell^k\|^2 - \frac{1}{2}\|x-\mu_i^k\|^2\right), 1\right\}\right] \\ &\leq 2\exp\left(-\frac{1}{32}\left\|\mu_\ell^k-\mu_i^k\right\|^2\right) \end{aligned}$$

for all $i \in I_{a_\ell^*}$, $i \neq \ell$, due to Lemma B.8. Using (31),

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})} r_i^{k,a_\ell^*}(x) \leq 2\exp\left(-\frac{\hat{\Delta}_2}{32}\right).$$

Substituting into (35),

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\tau_\ell^k(x) \le 2n\exp\left(-\frac{\hat{\Delta}_2}{32}\right). \tag{36}$$

Similarly,

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\hat{\tau}_{a_\ell^*}^k(x) \le 2n\exp\left(-\frac{\hat{\Delta}_2}{32}\right)$$

because

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}r_i^k(x) &\le \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x-\mu_\ell^k\|^2-\frac{1}{2}\|x-\mu_i^k\|^2\right),1\right\}\right]\\ &\le 2\exp\left(-\frac{1}{32}\left\|\mu_\ell^k-\mu_i^k\right\|^2\right)\end{aligned}$$

for all $i\notin I_{a_\ell^*}$. Finally,

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\bar{\tau}_{a_\ell^*}(x) = \sum_{b\in[m],b\neq a_\ell^*}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\bar{r}_b(x).$$

Due to Lemma B.9,

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\bar{r}_b(x) &\le \mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\min\left\{\exp\left(-\frac{1}{2}\left(\|x-\bar{\mu}_b\|^2-\|x-\bar{\mu}_{a_\ell^*}\|^2\right)\right),1\right\}\right]\\ &\le \exp\left(-\frac{\delta_{ab}^k}{2}\right)+2\exp\left(-\frac{(\delta_{ab}^k)^2}{2\Delta_{ab}^2}\right),\end{aligned}$$

where $\delta_{ab}^k := \frac{1}{2}\|\mu_\ell^k-\bar{\mu}_b\|^2-\frac{1}{2}\|\mu_\ell^k-\bar{\mu}_{a_\ell^*}\|^2$ and $\Delta_{ab}:=\|\bar{\mu}_{a_\ell^*}-\bar{\mu}_b\|$. Using (30),

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\bar{r}_b(x) \le \exp\left(-\frac{\Delta_{ab}^2}{8}\right)+2\exp\left(-\frac{\Delta_{ab}^2}{32}\right)\le 3\exp\left(-\frac{\Delta_{ab}^2}{32}\right)\le 3\exp\left(-\frac{\delta_{\min}^2}{32}\right)$$

and

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\bar{\tau}_{a_\ell^*}(x) \le 3m\exp\left(-\frac{\delta_{\min}^2}{32}\right).$$

Notice that the derived bounds are also valid for $k-1,\dots,0$. Substituting the bounds into (34), we get

$$\begin{aligned}\left\|E_\ell^k\right\| &\le \max_{j\in\{0,\dots,k\}}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^j,\mathbf{I})}\left[\left\|\bar{\mathcal{E}}_\ell^j(x)+\mathcal{E}_\ell^j(x)\right\|\right]\\ &\le 62\max\{R,R^3\}\left(4n\exp\left(-\hat{\Delta}_2/32\right)+3m\exp\left(-\delta_{\min}^2/32\right)\right)\\ &\le 434\max\{R,R^3\}\max\left\{n\exp\left(-\hat{\Delta}_2/32\right),m\exp\left(-\delta_{\min}^2/32\right)\right\}.\end{aligned}$$

Due to $R\ge 1$, $\hat{\Delta}_2\ge 32\log(434nR^3/E_1)$ (see (33)), and $\delta_{\min}\ge\sqrt{32\log\left(434R^3m/E_1\right)}$ (see (33)). Thus

$$\left\|E_\ell^k\right\| \le \max_{j\in\{0,\dots,k\}}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^j,\mathbf{I})}\left[\left\|\bar{\mathcal{E}}_\ell^j(x)+\mathcal{E}_\ell^j(x)\right\|\right]\le E_1. \tag{37}$$

Using (25) and the choice of $\gamma$ and $R$, $\left\|\bar{\mu}_{a_\ell^*}\right\|\le\bar{R}\le\frac{R}{2}$, $\left\|\mu_\ell^0\right\|=R^0\le\frac{R}{2}$, we get

$$\left\|\mu_\ell^{k+1}\right\| \le \max\left\{\left\|\bar{\mu}_{a_\ell^*}\right\|,\left\|\mu_\ell^0\right\|\right\}+\left\|E_\ell^k\right\|\le\frac{R}{2}+E_1\le R$$

since $E_1\le\frac{R}{2}$ (see (33)). For all $i,j\in[n]$, using (25) and (37),

$$\left\|\mu_i^{k+1}-\mu_j^{k+1}\right\|$$

$$= \left\| \bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*} + \left(1 - \frac{2\gamma}{n}\right)^{k+1} (\mu_i^0 - \mu_j^0 - (\bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*})) - E_i^k + E_j^k \right\|$$
$$\geq \left\| \bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*} + \left(1 - \frac{2\gamma}{n}\right)^{k+1} (\mu_i^0 - \mu_j^0 - (\bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*})) \right\| - 2E_1.$$

If $a_i^* = a_j^*$, then

$$\left\|\mu_i^{k+1} - \mu_j^{k+1}\right\| \geq \left(1 - \frac{2\gamma}{n}\right)^{k+1} \left\|\mu_i^0 - \mu_j^0\right\| - 2E_1 \geq \frac{4\sqrt{\bar{\Delta}\Delta_2}}{R^0} - 2E_1,$$

where we use (28) and (20). If $a_i^* \neq a_j^*$, then

$$\left\|\mu_i^{k+1} - \mu_j^{k+1}\right\|$$
$$\geq \left\| \bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*} + \left(1 - \frac{2\gamma}{n}\right)^{k+1} (\mu_i^0 - \mu_j^0 - (\bar{\mu}_{a_i^*} - \bar{\mu}_{a_j^*})) \right\| - 2E_1.$$

Notice that $\bar{\mu}_{a_i^*} + \left(1 - \frac{2\gamma}{n}\right)^{k+1} (\mu_i^0 - \bar{\mu}_{a_i^*}) \in [\mu_i^0, \bar{\mu}_{a_i^*}]$ for all $i \in [n]$. Using (21), we get

$$\left\|\mu_i^{k+1} - \mu_j^{k+1}\right\| \geq \Delta_3 - 2E_1.$$

In total,

$$\left\|\mu_i^{k+1} - \mu_j^{k+1}\right\| \geq \min\left\{\frac{4\sqrt{\bar{\Delta}\Delta_2}}{R^0}, \Delta_3\right\} - 2E_1$$

and

$$\left\|\mu_i^{k+1} - \mu_j^{k+1}\right\|^2 \geq \hat{\Delta}_2$$

due to $\frac{4\sqrt{\bar{\Delta}\Delta_2}}{R^0} \geq 2\sqrt{\hat{\Delta}_2}, \Delta_3 \geq 2\sqrt{\hat{\Delta}_2}, E_1 \leq \frac{\sqrt{\hat{\Delta}_2}}{2}$ (see (33)). It remains to prove (30) for the step $k \to k+1$. Indeed, for all $b \neq a_\ell^*$,

$$\begin{aligned}
D_{\ell,ab}^{k+1} &:= \|\mu_\ell^{k+1} - \bar{\mu}_b\|^2 - \|\mu_\ell^{k+1} - \bar{\mu}_{a_\ell^*}\|^2 \\
&= \|\mu_\ell^k - \gamma \nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k) - \bar{\mu}_b\|^2 - \|\mu_\ell^k - \gamma \nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k) - \bar{\mu}_{a_\ell^*}\|^2 \\
&= D_{\ell,ab}^k + 2\gamma \left\langle \nabla_{\mu_\ell^k} \mathcal{L}_{\text{rev}}(\mu^k), \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle \\
&= D_{\ell,ab}^k + \frac{4\gamma}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ \left\langle \mu_\ell^k - \bar{\mu}_{a_\ell^*} + \bar{\mathcal{E}}_\ell^k(x) + \mathcal{E}_\ell^k(x), \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle \right] \\
&= D_{\ell,ab}^k + \frac{4\gamma}{n} \left\langle \mu_\ell^k - \bar{\mu}_{a_\ell^*}, \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle + \frac{4\gamma}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ \left\langle \bar{\mathcal{E}}_\ell^k(x) + \mathcal{E}_\ell^k(x), \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle \right] \\
&= D_{\ell,ab}^k + \frac{2\gamma}{n} \left( \|\bar{\mu}_b - \bar{\mu}_{a_\ell^*}\|^2 - D_{\ell,ab}^k \right) + \frac{4\gamma}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ \left\langle \bar{\mathcal{E}}_\ell^k(x) + \mathcal{E}_\ell^k(x), \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle \right] \\
&= \left(1 - \frac{2\gamma}{n}\right) D_{\ell,ab}^k + \frac{2\gamma}{n} \|\bar{\mu}_b - \bar{\mu}_{a_\ell^*}\|^2 + \frac{4\gamma}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ \left\langle \bar{\mathcal{E}}_\ell^k(x) + \mathcal{E}_\ell^k(x), \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle \right].
\end{aligned}$$

Thus

$$D_{\ell,ab}^{k+1} - \|\bar{\mu}_b - \bar{\mu}_{a_\ell^*}\|^2 = \left(1 - \frac{2\gamma}{n}\right) \left(D_{\ell,ab}^k - \|\bar{\mu}_b - \bar{\mu}_{a_\ell^*}\|^2\right) + \frac{4\gamma}{n} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \left[ \left\langle \bar{\mathcal{E}}_\ell^k(x) + \mathcal{E}_\ell^k(x), \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle \right]$$

and

$$D_{\ell,ab}^{k+1} - \|\bar{\mu}_b - \bar{\mu}_{a_\ell^*}\|^2 = \left(1 - \frac{2\gamma}{n}\right)^{k+1} \left(D_{\ell,ab}^0 - \|\bar{\mu}_b - \bar{\mu}_{a_\ell^*}\|^2\right)$$
$$+ \frac{4\gamma}{n} \sum_{j=0}^{k} \left(1 - \frac{2\gamma}{n}\right)^{k-j} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^j, \mathbf{I})} \left[ \left\langle \bar{\mathcal{E}}_\ell^j(x) + \mathcal{E}_\ell^j(x), \bar{\mu}_b - \bar{\mu}_{a_\ell^*} \right\rangle \right].$$

Notice that

$$D^0_{\ell,ab} := \|\mu^0_\ell - \bar{\mu}_b\|^2 - \|\mu^0_\ell - \bar{\mu}_{a^*_\ell}\|^2 \geq \Delta_1 \geq \delta^2_{\max} \geq \|\bar{\mu}_b - \bar{\mu}_{a^*_\ell}\|^2$$

due to (19) and $\Delta_1 \geq \delta^2_{\max}$ (see (33)). Using this inequality,

$$\begin{aligned}
D^{k+1}_{\ell,ab} - \|\bar{\mu}_b - \bar{\mu}_{a^*_\ell}\|^2 &\geq \frac{4\gamma}{n}\sum_{j=0}^{k}\left(1-\frac{2\gamma}{n}\right)^{k-j}\mathbb{E}_{x\sim\mathcal{N}(\mu^j_\ell,\mathbf{I})}\left[\left\langle \bar{\mathcal{E}}^j_\ell(x)+\mathcal{E}^j_\ell(x), \bar{\mu}_b - \bar{\mu}_{a^*_\ell}\right\rangle\right] \\
&\geq -\frac{4\gamma}{n}\sum_{j=0}^{k}\left(1-\frac{2\gamma}{n}\right)^{k-j} 2R\mathbb{E}_{x\sim\mathcal{N}(\mu^j_\ell,\mathbf{I})}\left(\left\|\bar{\mathcal{E}}^j_\ell(x)+\mathcal{E}^j_\ell(x)\right\|\right) \\
&\geq -4R\max_{j\in\{0,\ldots,k\}}\mathbb{E}_{x\sim\mathcal{N}(\mu^j_\ell,\mathbf{I})}\left(\left\|\bar{\mathcal{E}}^j_\ell(x)+\mathcal{E}^j_\ell(x)\right\|\right) \\
&\overset{(37)}{\geq} -4RE_1.
\end{aligned}$$

Due to (32) and $\delta_{\min} \geq 1$, $4RE_1 \leq \frac{\delta^2_{\min}}{2}$ and

$$\|\mu^{k+1}_\ell - \bar{\mu}_b\|^2 - \|\mu^{k+1}_\ell - \bar{\mu}_{a^*_\ell}\|^2 \geq \frac{\|\bar{\mu}_b - \bar{\mu}_{a^*_\ell}\|^2}{2}$$

for all $b \neq a^*_\ell$, and we have proved the mathematical induction for $k \to k+1$.

Using (25), (27), and (37),

$$\begin{aligned}
\left\|\mu^{\bar{K}}_\ell - \bar{\mu}_{a^*_\ell}\right\| &\leq \left(1-\frac{2\gamma}{n}\right)^{\bar{K}}\left\|\mu^0_\ell - \bar{\mu}_{a^*_\ell}\right\| + \left\|E^{\bar{K}-1}_\ell\right\| \\
&\leq \frac{8\sqrt{\bar{\Delta}}}{R^0}\left\|\mu^0_\ell - \bar{\mu}_{a^*_\ell}\right\| + E_1 \leq 16\sqrt{\bar{\Delta}} + E_1
\end{aligned}$$

because $\left\|\mu^0_\ell\right\| = R^0$ and $\left\|\bar{\mu}_{a^*_\ell}\right\| \leq R^0$. Due to (33) and (32),

$$\left\|\mu^{\bar{K}}_\ell - \bar{\mu}_{a^*_\ell}\right\| \leq 16\sqrt{\bar{\Delta}} + E_1 \leq 17\sqrt{\bar{\Delta}} \tag{38}$$

for all $\ell \in [n]$. Let $I_a = \{i_{a,1}, \ldots, i_{a,n_a}\}$. The inequality ensures that after $\bar{K}$ steps, each component converges to a small-$\bar{\Delta}$ neighborhood of the corresponding target component.

**(Local convergence stage).** The proof now would be similar to that of Theorem 2.1, but with careful and technical error control. For all $a \in A^*$, we define

$$\mathbf{Z}^k_a := \begin{bmatrix} (\mu^k_{i_{a,1}} - \bar{\mu}_a)^\top \\ \vdots \\ (\mu^k_{i_{a,n_a}} - \bar{\mu}_a)^\top \end{bmatrix}.$$

Then,

$$\begin{aligned}
\mathbf{Z}^{\bar{K}}_a &= \begin{bmatrix} (\mu^{\bar{K}}_{i_{a,1}} - \bar{\mu}_a)^\top \\ \vdots \\ (\mu^{\bar{K}}_{i_{a,n_a}} - \bar{\mu}_a)^\top \end{bmatrix} = \begin{bmatrix} \left(\left(1-\frac{2\gamma}{n}\right)^{\bar{K}}(\mu^0_{i_{a,1}} - \bar{\mu}_a) - E^{\bar{K}-1}_{i_{a,1}}\right)^\top \\ \vdots \\ \left(\left(1-\frac{2\gamma}{n}\right)^{\bar{K}}(\mu^0_{i_{a,n_a}} - \bar{\mu}_a) - E^{\bar{K}-1}_{i_{a,n_a}}\right)^\top \end{bmatrix} \\
&= \left(1-\frac{2\gamma}{n}\right)^{\bar{K}}\begin{bmatrix} \left(\mu^0_{i_{a,1}}\right)^\top \\ \vdots \\ \left(\mu^0_{i_{a,n_a}}\right)^\top \end{bmatrix} - \left(1-\frac{2\gamma}{n}\right)^{\bar{K}}\begin{bmatrix} (\bar{\mu}_a)^\top \\ \vdots \\ (\bar{\mu}_a)^\top \end{bmatrix} - \begin{bmatrix} \left(E^{\bar{K}-1}_{i_{a,1}}\right)^\top \\ \vdots \\ \left(E^{\bar{K}-1}_{i_{a,n_a}}\right)^\top \end{bmatrix}.
\end{aligned}$$

Thus,

$$\left\|\mathbf{Z}_a^{\bar{K}}\right\|_{\text{op}} \leq \left(1-\frac{2\gamma}{n}\right)^{\bar{K}} \left\|\begin{bmatrix} \left(\mu_{i_{a,1}}^0\right)^\top \\ \vdots \\ \left(\mu_{i_{a,n_a}}^0\right)^\top \end{bmatrix}\right\|_{\text{op}} + \left(1-\frac{2\gamma}{n}\right)^{\bar{K}} \left\|\begin{bmatrix} (\bar{\mu}_a)^\top \\ \vdots \\ (\bar{\mu}_a)^\top \end{bmatrix}\right\|_{\text{op}} + \left\|\begin{bmatrix} \left(E_{i_{a,1}}^{\bar{K}-1}\right)^\top \\ \vdots \\ \left(E_{i_{a,n_a}}^{\bar{K}-1}\right)^\top \end{bmatrix}\right\|_{\text{op}}$$
$$\leq \left(1-\frac{2\gamma}{n}\right)^{\bar{K}} \left\|\begin{bmatrix} \left(\mu_{i_{a,1}}^0\right)^\top \\ \vdots \\ \left(\mu_{i_{a,n_a}}^0\right)^\top \end{bmatrix}\right\|_{\text{op}} + \sqrt{n_a}\left(1-\frac{2\gamma}{n}\right)^{\bar{K}} \|\bar{\mu}_a\| + \sqrt{n_a}E_1$$

Using the choice of $\bar{K}$, $\|\bar{\mu}_a\| \leq \bar{R}$, and (22),

$$\begin{aligned} \left\|\mathbf{Z}_a^{\bar{K}}\right\|_{\text{op}} &\leq \frac{8\sqrt{\bar{\Delta}}\Delta_4}{R^0} + \frac{8\sqrt{n_a\bar{\Delta}}\bar{R}}{R^0} + \sqrt{n_a}E_1 \\ &\leq \bar{z} := \frac{8\sqrt{\bar{\Delta}}\Delta_4}{R^0} + \frac{8\sqrt{n\bar{\Delta}}\bar{R}}{R^0} + \sqrt{n}E_1 \end{aligned} \tag{39}$$

for all $a \in A^*$. For all $a \in A^*$, consider the subgroup $I_a$. Due to (23), for all $\ell \in I_a$,

$$\begin{aligned} \mu_\ell^{k+1} - \bar{\mu}_a &= \mu_\ell^k - \bar{\mu}_a - \gamma\nabla_{\mu_\ell^k}\mathcal{L}_{\text{rev}}(\mu^k) \\ &= \mu_\ell^k - \bar{\mu}_a - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[m_a^k(x) - \bar{\mu}_a + \mathbf{C}_a^k(x)\left(\mu_\ell^k - m_a^k(x)\right)\right] - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\mathcal{E}_\ell^k(x)\right] \\ &= \mu_\ell^k - \bar{\mu}_a \\ &\quad - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\sum_{i\in I_a} r_i^{k,a}(x)(\mu_i^k - \bar{\mu}_a) + \sum_{i\in I_a} r_i^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)\left(\mu_i^k - m_a^k(x)\right)^\top\left(\mu_\ell^k - m_a^k(x)\right)\right] \\ &\quad - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\mathcal{E}_\ell^k(x)\right]. \end{aligned}$$

Since $\sum_{i\in I_a} r_i^{k,a}(x) b\left(\mu_i^k - m_a^k(x)\right)^\top = 0$ for all $b \in \mathbb{R}^d$,

$$\begin{aligned} \mu_\ell^{k+1} - \bar{\mu}_a &= \mu_\ell^k - \bar{\mu}_a \\ &\quad - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\sum_{i\in I_a} r_i^{k,a}(x)(\mu_i^k - \bar{\mu}_a) + \sum_{i\in I_a} r_i^{k,a}(x)\left(\mu_i^k - \bar{\mu}_a\right)\left(\mu_i^k - m_a^k(x)\right)^\top\left(\mu_\ell^k - m_a^k(x)\right)\right] \\ &\quad - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\mathcal{E}_\ell^k(x)\right] \\ &= \mu_\ell^k - \bar{\mu}_a \\ &\quad - \frac{2\gamma}{n}\sum_{i\in I_a}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[r_i^{k,a}(x) + r_i^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)^\top\left(\mu_\ell^k - m_a^k(x)\right)\right]\left(\mu_i^k - \bar{\mu}_a\right) \\ &\quad - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\mathcal{E}_\ell^k(x)\right]. \end{aligned}$$

We can conclude that

$$\mu_\ell^{k+1} - \bar{\mu}_a = \mu_\ell^k - \bar{\mu}_a - \frac{2\gamma}{n}\sum_{i\in I_a}\mathbf{B}_{a,\ell i}^k(\mu_i^k - \bar{\mu}_a) - \frac{2\gamma}{n}\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\left[\mathcal{E}_\ell^k(x)\right],$$

and

$$\mathbf{Z}_a^{k+1} = \left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k - \frac{2\gamma}{n}\mathbf{E}_a^k, \tag{40}$$

where $\mathbf{E}_a^k = \left[\mathbb{E}_{x\sim\mathcal{N}(\mu^k_{i_{a,1}},\mathbf{I})}\left[\mathcal{E}^k_{i_{a,1}}(x)\right]^\top;\ldots;\mathbb{E}_{x\sim\mathcal{N}(\mu^k_{i_{a,n_a}},\mathbf{I})}\left[\mathcal{E}^k_{i_{a,n_a}}(x)\right]^\top\right]$ and $\mathbf{B}_a^k \in \mathbb{R}^{n_a\times n_a}$ such that

$$\begin{aligned}
\mathbf{B}^k_{a,\ell j} &:= \mathbb{E}_{x\sim\mathcal{N}(\mu^k_{i_{a,\ell}},\mathbf{I})}\left[r^{k,a}_{i_{a,j}}(x) + r^{k,a}_{i_{a,j}}(x)\big(\mu^k_{i_{a,j}} - m^k_a(x)\big)^\top\big(\mu^k_{i_{a,\ell}} - m^k_a(x)\big)\right]\\
&= \int_{x\in\mathbb{R}^d}\left[r^{k,a}_{i_{a,j}}(x) + r^{k,a}_{i_{a,j}}(x)\big(\mu^k_{i_{a,j}} - m^k_a(x)\big)^\top\big(\mu^k_{i_{a,\ell}} - m^k_a(x)\big)\right]\varphi(x-\mu^k_{i_{a,\ell}})\,dx\\
&= n_a\int_{x\in\mathbb{R}^d}\left[r^{k,a}_{i_{a,j}}(x)r^{k,a}_{i_{a,\ell}}(x) + r^{k,a}_{i_{a,j}}(x)r^{k,a}_{i_{a,\ell}}(x)\big(\mu^k_{i_{a,j}} - m^k_a(x)\big)^\top\big(\mu^k_{i_{a,\ell}} - m^k_a(x)\big)\right]p^{k,a}_\mu(x)\,dx\\
&= n_a\mathbb{E}_{x\sim p^{k,a}_\mu}\left[r^{k,a}_{i_{a,j}}(x)r^{k,a}_{i_{a,\ell}}(x) + r^{k,a}_{i_{a,j}}(x)r^{k,a}_{i_{a,\ell}}(x)\big(\mu^k_{i_{a,j}} - m^k_a(x)\big)^\top\big(\mu^k_{i_{a,\ell}} - m^k_a(x)\big)\right],
\end{aligned} \tag{41}$$

where $p^{k,a}_\mu(x) := \frac{1}{n_a}\sum_{i\in I_a}\varphi(x-\mu^k_i)$. Thus, $\mathbf{B}^k_a$ is symmetric and non-negative definite. Let us define $R^k_a := \max_{i\in I_a}\left\|\mu^k_i - \bar\mu_a\right\|$ for all $a\in A^*$. For all $v\in\mathbb{R}^{n_a}$,

$$\begin{aligned}
v^\top\mathbf{B}^k_a v &= n_a\int_{x\in\mathbb{R}^d}\left[\left(\sum_{j=1}^{n_a} r^{k,a}_{i_{a,j}}(x)v_j\right)^2 + \left\|\sum_{j=1}^{n_a} r^{k,a}_{i_{a,j}}(x)v_j(\mu^k_{i_{a,j}} - m^k_a(x))\right\|^2\right]p^{k,a}_\mu(x)\,dx\\
&\le n_a\int_{x\in\mathbb{R}^d}\left[\sum_{j=1}^{n_a} r^{k,a}_{i_{a,j}}(x)v_j^2 + \sum_{j=1}^{n_a} r^{k,a}_{i_{a,j}}(x)v_j^2\left\|\mu^k_{i_{a,j}} - m^k_a(x)\right\|^2\right]p^{k,a}_\mu(x)\,dx\\
&= \sum_{j=1}^{n_a}\left[\int_{x\in\mathbb{R}^d} n_a p^{k,a}_\mu(x)r^{k,a}_{i_{a,j}}(x)\,dx\right]v_j^2 + \sum_{j=1}^{n_a}\left[\int_{x\in\mathbb{R}^d}\left\|\mu^k_{i_{a,j}} - m^k_a(x)\right\|^2 n_a p^{k,a}_\mu(x)r^{k,a}_{i_{a,j}}(x)\,dx\right]v_j^2\\
&\le \sum_{j=1}^{n_a} v_j^2 + 4(R^k_a)^2\sum_{j=1}^{n_a} v_j^2 = (1+4(R^k_a)^2)\,\|v\|^2
\end{aligned}$$

since $n_a p^{k,a}_\mu(x)r^{k,a}_{i_{a,j}}(x) = \varphi(x-\mu^k_{i_{a,j}})$, and

$$\left\|\mu^k_{i_{a,j}} - m^k_a(x)\right\|^2 = \left\|\mu^k_{i_{a,j}} - \bar\mu_a - (m^k_a(x) - \bar\mu_a)\right\|^2 \le 4(R^k_a)^2.$$

Thus,

$$\left\|\mathbf{B}^k_a\right\|_{\text{op}} \le 1 + 4(R^k_a)^2$$

for all $k\ge 0$ and $a\in A^*$. Using mathematical induction, for all $\bar K\le k\le\hat K$, we assume that

$$\left\|\mathbf{Z}^k_a\right\|_{\text{op}} \le 2\bar z, \tag{42}$$

for all $a\in A^*$, where

$$\hat K := \bar K + \left\lceil\frac{n\min\{n,d\}\max\left\{4n, 8\bar z^2\right\}}{\gamma\varepsilon}\right\rceil. \tag{43}$$

We also define

$$E_2 := \min\left\{\frac{n\bar z}{2\gamma\hat K}, \frac{\varepsilon^2}{16\min\{n,d\}\sqrt{\min\{n,d\}}\bar z}\min\left\{\frac{1}{4n},\frac{1}{8\bar z^2}\right\}, \sqrt{\frac{n\varepsilon^2}{8\gamma\min\{n,d\}}\min\left\{\frac{1}{4n},\frac{1}{8\bar z^2}\right\}}\right\}. \tag{44}$$

We consider the following lemma, proved in Section E.1.

**Lemma E.5.** *For our choice of parameters,*

$$\begin{aligned}
&\delta_{\min} \geq 8\bar{z}, \quad \bar{z} \geq 1, \\
&\bar{z} \leq \tilde{z} := 65\sqrt{\log\left(2^{16} n (R^0)^4\right)} \sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}, \\
&\delta_{\min} \geq \sqrt{128 \log\left(\frac{156\sqrt{n} R^3 \max\{n, m\}}{E_2}\right)}, \\
&\hat{K} \leq \tilde{K} := \frac{n \log\left(R^0\right)}{2\gamma} + \frac{2n \min\{n, d\} \max\left\{4n, 8\tilde{z}^2\right\}}{\gamma\varepsilon}, \\
&R \geq 2\bar{z} + \bar{R}.
\end{aligned} \tag{45}$$

For $k = \bar{K}$, the induction hypothesis holds due to (39). We now prove the next step of the induction. Due to (42) and $R \geq 2\bar{z} + \bar{R}$ (see (45)),

$$\left\|\mu_\ell^k - \bar{\mu}_a\right\| \leq 2\bar{z}$$

and

$$\left\|\mu_\ell^k\right\| \leq R \tag{46}$$

for all $a \in A^*$ and $\ell \in I_a$. Thus, for all $i \notin I_a$,

$$\begin{aligned}
&\left\|\mu_\ell^k - \mu_i^k\right\| = \left\|(\bar{\mu}_{a_\ell^*} - \bar{\mu}_{a_i^*}) + (\mu_\ell^k - \bar{\mu}_{a_\ell^*}) - (\mu_i^k - \bar{\mu}_{a_i^*})\right\| \\
&\geq \left\|\bar{\mu}_{a_\ell^*} - \bar{\mu}_{a_i^*}\right\| - \left\|\mu_\ell^k - \bar{\mu}_{a_\ell^*}\right\| - \left\|\mu_i^k - \bar{\mu}_{a_i^*}\right\| \geq \delta_{\min} - 4\bar{z} \geq \frac{\delta_{\min}}{2}
\end{aligned} \tag{47}$$

since $\delta_{\min} \geq 8\bar{z}$ (see (45)). Similarly, for all $a \neq b \in [m]$ and for all $\ell \in I_a$,

$$\begin{aligned}
&\frac{1}{2}\|\mu_\ell^k - \bar{\mu}_b\|^2 - \frac{1}{2}\|\mu_\ell^k - \bar{\mu}_a\|^2 = \frac{1}{2}\|\mu_\ell^k - \bar{\mu}_a - (\bar{\mu}_b - \bar{\mu}_a)\|^2 - \frac{1}{2}\|\mu_\ell^k - \bar{\mu}_a\|^2 \\
&= -\left\langle \mu_\ell^k - \bar{\mu}_a, \bar{\mu}_b - \bar{\mu}_a \right\rangle + \frac{1}{2}\|\bar{\mu}_b - \bar{\mu}_a\|^2 \geq \frac{1}{2}\|\bar{\mu}_b - \bar{\mu}_a\|^2 - \left\|\mu_\ell^k - \bar{\mu}_a\right\| \|\bar{\mu}_b - \bar{\mu}_a\| \\
&\geq \frac{1}{4}\|\bar{\mu}_b - \bar{\mu}_a\|^2.
\end{aligned} \tag{48}$$

because $\frac{1}{4}\|\bar{\mu}_b - \bar{\mu}_a\| \geq \frac{1}{4}\delta_{\min} \geq 2\bar{z} \geq \left\|\mu_\ell^k - \bar{\mu}_a\right\|$. For all $a \in A^*$,

$$\begin{aligned}
&\left\|\mathbf{E}_a^k\right\|_F \leq \sqrt{n_a} \max_{\ell \in I_a} \left\|\mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})}\left[\mathcal{E}_\ell^k(x)\right]\right\| \leq \sqrt{n_a} \max_{\ell \in I_a} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})}\left[\left\|\mathcal{E}_\ell^k(x)\right\|\right] \\
&\leq 52\sqrt{n} \max\{R, R^3\} \max_{\ell \in I_a}\left(\mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \hat{\tau}_a^k(x) + \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \bar{\tau}_a(x)\right)
\end{aligned} \tag{49}$$

where we use Lemma E.2, (46), and $n_a \leq n$. Using Lemma B.8,

$$\begin{aligned}
&\mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \hat{\tau}_a^k(x) = \sum_{i \in [n], i \notin I_a} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} r_i^k(x) \\
&\leq \sum_{i \in [n], i \notin I_a} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})}\left[\min\left\{\exp\left(\frac{1}{2}\|x - \mu_\ell^k\|^2 - \frac{1}{2}\|x - \mu_i^k\|^2\right), 1\right\}\right] \\
&\leq 2 \sum_{i \in [n], i \notin I_a} \exp\left(-\frac{1}{32}\left\|\mu_\ell^k - \mu_i^k\right\|^2\right) \overset{(47)}{\leq} 2n \exp\left(-\frac{\delta_{\min}^2}{128}\right).
\end{aligned}$$

Next,

$$\mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \bar{\tau}_a(x) = \sum_{b \in [m], b \neq a} \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \bar{r}_b(x).$$

Due to Lemma B.9,

$$\mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})} \bar{r}_b(x) \leq \mathbb{E}_{x \sim \mathcal{N}(\mu_\ell^k, \mathbf{I})}\left[\min\left\{\exp\left(-\frac{1}{2}\left(\|x - \bar{\mu}_b\|^2 - \|x - \bar{\mu}_a\|^2\right)\right), 1\right\}\right]$$

$$\leq \exp\left(-\frac{\delta_{ab}^k}{2}\right) + 2\exp\left(-\frac{(\delta_{ab}^k)^2}{2\Delta_{ab}^2}\right),$$

where $\delta_{ab}^k := \frac{1}{2}\|\mu_\ell^k - \bar{\mu}_b\|^2 - \frac{1}{2}\|\mu_\ell^k - \bar{\mu}_a\|^2$ and $\Delta_{ab} := \|\bar{\mu}_a - \bar{\mu}_b\|$. Using (48), $\delta_{ab}^k \geq \frac{1}{4}\Delta_{ab}^2$ and

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\bar{r}_b(x) \leq \exp\left(-\frac{\Delta_{ab}^2}{8}\right) + 2\exp\left(-\frac{\Delta_{ab}^2}{32}\right) \leq 3\exp\left(-\frac{\Delta_{ab}^2}{32}\right) \leq 3\exp\left(-\frac{\delta_{\min}^2}{32}\right).$$

Therefore,

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_\ell^k,\mathbf{I})}\bar{r}_a(x) \leq 3m\exp\left(-\frac{\delta_{\min}^2}{32}\right).$$

Substituting the derived bounds into (49),

$$\begin{aligned}\left\|\mathbf{E}_a^k\right\|_F &\leq 52\sqrt{n}\max\{R,R^3\}\left(2n\exp\left(-\frac{\delta_{\min}^2}{128}\right) + 3m\exp\left(-\frac{\delta_{\min}^2}{32}\right)\right)\\ &\leq 156\sqrt{n}R^3\max\{n,m\}\exp\left(-\frac{\delta_{\min}^2}{128}\right).\end{aligned} \tag{50}$$

Due to $R \geq 1$ and $\delta_{\min} \geq \sqrt{128\log\left(\frac{156\sqrt{n}R^3\max\{n,m\}}{E_2}\right)}$ (see (45)),

$$\left\|\mathbf{E}_a^k\right\|_{\text{op}} \leq \left\|\mathbf{E}_a^k\right\|_F \leq E_2 \tag{51}$$

for all $a \in A^*$. Note that (51) is true for $k-1,\ldots,\bar{K}$. Due to (42), for all $a \in A^*$ and $i \in I_a$, $\left\|\mu_i^k - \bar{\mu}_a\right\| \leq 2\bar{z}$, $R_a^k \leq 2\bar{z}$, and $\left\|\mathbf{B}_a^k\right\|_{\text{op}} \leq 1 + 16\bar{z}^2$. Using this and (40),

$$\left\|\mathbf{Z}_a^{k+1}\right\|_{\text{op}} = \left\|\left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k - \frac{2\gamma}{n}\mathbf{E}_a^k\right\|_{\text{op}} \leq \left\|\mathbf{Z}_a^k\right\|_{\text{op}} + \frac{2\gamma}{n}\left\|\mathbf{E}_a^k\right\|_{\text{op}}$$

since $\gamma \leq \frac{n}{1+16\bar{z}^2}$. Unrolling the recursion and using (39) and (51),

$$\left\|\mathbf{Z}_a^{k+1}\right\|_{\text{op}} \leq \bar{z} + \frac{2\gamma\hat{K}}{n}E_2 \leq 2\bar{z}$$

since $k+1 \leq \hat{K}$ and $E_2 \leq \frac{n\bar{z}}{2\gamma\hat{K}}$ (see (44)). Thus, we have proved the next step of the mathematical induction (42).

We now fix $a \in A^*$. For all $\bar{K} \leq k \leq \hat{K}-1$, consider the update under the squared Frobenius norm:

$$\begin{aligned}\left\|\mathbf{Z}_a^{k+1}\right\|_F^2 &= \left\|\left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k - \frac{2\gamma}{n}\mathbf{E}_a^k\right\|_F^2\\ &= \left\|\left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k\right\|_F^2 - \frac{4\gamma}{n}\left\langle\left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k, \mathbf{E}_a^k\right\rangle + \frac{4\gamma^2}{n^2}\left\|\mathbf{E}_a^k\right\|_F^2.\end{aligned} \tag{52}$$

Consider the first norm separately:

$$\begin{aligned}\left\|\left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k\right\|_F^2 &= \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{4\gamma}{n}\left\langle\mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle + \frac{4\gamma^2}{n^2}\left\|\mathbf{B}_a^k\mathbf{Z}_a^k\right\|_F^2\\ &\leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{4\gamma}{n}\left\langle\mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle + \frac{4\gamma^2}{n^2}\left\|\mathbf{B}_a^k\right\|_{\text{op}}\left\langle\mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle\\ &\leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{4\gamma}{n}\left\langle\mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle + \frac{2\gamma}{n}\left\langle\mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle\\ &= \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n}\left\langle\mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle.\end{aligned} \tag{53}$$

since $\frac{\gamma}{n} \leq \frac{1}{2(1+16\bar{z}^2)} \leq \frac{1}{2\left\|\mathbf{B}_a^k\right\|_{\text{op}}}$. We now use the following lemma separately proved in Section E.1.

**Lemma E.6.** *For all $k \geq \bar{K}$ and $a \in A^*$,*

$$\left\langle \mathbf{Z}_a^k, \mathbf{B}_a^k \mathbf{Z}_a^k \right\rangle \geq \min\left\{ \frac{\kappa_a}{4n_a^2} \left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\},$$

*where $\kappa_a := \frac{n_a}{\min\{n_a, d\}}$.*

Substituting this bound into (52) and (53),

$$\begin{aligned}
\left\|\mathbf{Z}_a^{k+1}\right\|_F^2 &\leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n}\min\left\{ \frac{\kappa_a}{4n_a^2} \left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\} - \frac{4\gamma}{n}\left\langle \left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k, \mathbf{E}_a^k \right\rangle + \frac{4\gamma^2}{n^2}\left\|\mathbf{E}_a^k\right\|_F^2 \\
&\leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n}\min\left\{ \frac{\kappa_a}{4n_a^2} \left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\} + \frac{4\gamma}{n}\left\| \left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\mathbf{Z}_a^k \right\|_F \left\|\mathbf{E}_a^k\right\|_F + \frac{4\gamma^2}{n^2}\left\|\mathbf{E}_a^k\right\|_F^2 \\
&\leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n}\min\left\{ \frac{\kappa_a}{4n_a^2} \left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\} + \frac{4\gamma}{n}\left\|\mathbf{Z}_a^k\right\|_F \left\|\mathbf{E}_a^k\right\|_F + \frac{4\gamma^2}{n^2}\left\|\mathbf{E}_a^k\right\|_F^2,
\end{aligned}$$

since $\left\|\left(\mathbf{I} - \frac{2\gamma}{n}\mathbf{B}_a^k\right)\right\|_{\text{op}} \leq 1$. Using (42), $\left\|\mathbf{Z}_a^k\right\|_F \leq 2\sqrt{\min\{n_a, d\}}\bar{z}$ and

$$\begin{aligned}
\left\|\mathbf{Z}_a^{k+1}\right\|_F^2 &\leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n}\min\left\{ \frac{\kappa_a}{4n_a^2} \left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\} + \frac{8\gamma\sqrt{\min\{n_a, d\}}\bar{z}}{n}\left\|\mathbf{E}_a^k\right\|_F + \frac{4\gamma^2}{n^2}\left\|\mathbf{E}_a^k\right\|_F^2 \\
&\overset{(51)}{\leq} \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n}\min\left\{ \frac{\kappa_a}{4n_a^2} \left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\} + \frac{8\gamma\sqrt{\min\{n, d\}}\bar{z}E_2}{n} + \frac{4\gamma^2 E_2^2}{n^2}.
\end{aligned}$$

Let us temporarily define $E_3 := 8\sqrt{\min\{n, d\}}\bar{z}E_2 + \frac{4\gamma E_2^2}{n}$ to get

$$\left\|\mathbf{Z}_a^{k+1}\right\|_F^2 \leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n}\min\left\{ \frac{\kappa_a}{4n_a^2} \left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\} + \frac{\gamma E_3}{n}.$$

Due to $\left\|\mathbf{Z}_a^k\right\|_F \leq 2\sqrt{\min\{n, d\}}\bar{z}$ and $\frac{\kappa_a}{n_a^2} = \frac{1}{n_a \min\{n_a, d\}} \geq \frac{1}{n\min\{n, d\}}$,

$$\left\|\mathbf{Z}_a^{k+1}\right\|_F^2 \leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{2\gamma}{n\min\{n, d\}}\min\left\{ \frac{1}{4n}, \frac{1}{8\bar{z}^2} \right\}\left\|\mathbf{Z}_a^k\right\|_F^4 + \frac{\gamma E_3}{n},$$

for all $\bar{K} \leq k \leq \hat{K} - 1$. Using proof by contradiction, assume that $\left\|\mathbf{Z}_a^k\right\|_F^2 > \varepsilon$ for all $\bar{K} \leq k \leq \hat{K}$. Since $E_3 \leq \frac{\varepsilon^2}{\min\{n, d\}}\min\left\{\frac{1}{4n}, \frac{1}{8\bar{z}^2}\right\}$ (see (44)), we get

$$\left\|\mathbf{Z}_a^{k+1}\right\|_F^2 \leq \left\|\mathbf{Z}_a^k\right\|_F^2 - \frac{\gamma}{n\min\{n, d\}}\min\left\{ \frac{1}{4n}, \frac{1}{8\bar{z}^2} \right\}\left\|\mathbf{Z}_a^k\right\|_F^4.$$

Applying Lemma B.10, we get

$$\left\|\mu_\ell^{\hat{K}} - \bar{\mu}_a\right\|^2 \leq \left\|\mathbf{Z}_a^{\hat{K}}\right\|_F^2 \leq \frac{n\min\{n, d\}\max\left\{4n, 8\bar{z}^2\right\}}{\gamma}\frac{1}{\hat{K} - \bar{K}} \leq \varepsilon,$$

for all $\ell \in I_a$ due to the definition of $\hat{K}$ in (43). We have proved a contradiction. Thus, $\left\|\mathbf{Z}_a^{\bar{k}}\right\|_F^2 \leq \varepsilon$ for some $\bar{k} \in \{\bar{K}, \ldots, \hat{K}\}$ and for all $\ell \in I_a$. It remains to prove that this inequality holds for all $k > \bar{k}$. Notice that

$$\left\|\mathbf{Z}_a^{\bar{k}}\right\|_{\text{op}}^2 \leq \left\|\mathbf{Z}_a^{\bar{k}}\right\|_F^2 \leq \varepsilon \leq 2\bar{z}$$

since $\bar{z} \geq 1$ (see (45)). Similarly to the previous steps, (51) holds for $k = \bar{k}$ and

$$\begin{aligned}
\left\|\mathbf{Z}_a^{\bar{k}+1}\right\|_F^2 &\leq \left\|\mathbf{Z}_a^{\bar{k}}\right\|_F^2 - \frac{2\gamma}{n\min\{n, d\}}\min\left\{ \frac{1}{4n}, \frac{1}{8\bar{z}^2} \right\}\left\|\mathbf{Z}_a^{\bar{k}}\right\|_F^4 + \frac{\gamma E_3}{n} \\
&\leq \left\|\mathbf{Z}_a^{\bar{k}}\right\|_F^2 - \frac{2\gamma}{n\min\{n, d\}}\min\left\{ \frac{1}{4n}, \frac{1}{8\bar{z}^2} \right\}\left\|\mathbf{Z}_a^{\bar{k}}\right\|_F^4 + \frac{\gamma\varepsilon^2}{n\min\{n, d\}}\min\left\{ \frac{1}{4n}, \frac{1}{8\bar{z}^2} \right\}.
\end{aligned}$$

Since

$$\max_{0\le x\le\varepsilon}\left\{g(x) := x - \frac{2\gamma}{n\min\{n,d\}}\min\left\{\frac{1}{4n},\frac{1}{8\bar{z}^2}\right\}x^2 + \frac{\gamma\varepsilon^2}{n\min\{n,d\}}\min\left\{\frac{1}{4n},\frac{1}{8\bar{z}^2}\right\}\right\}$$
$$\overset{x^*=\varepsilon}{=} \varepsilon - \frac{2\gamma}{n\min\{n,d\}}\min\left\{\frac{1}{4n},\frac{1}{8\bar{z}^2}\right\}\varepsilon^2 + \frac{\gamma\varepsilon^2}{n\min\{n,d\}}\min\left\{\frac{1}{4n},\frac{1}{8\bar{z}^2}\right\} \le \varepsilon$$

because the (unconstrained) optimal point of $g$ on $\mathbb{R}$ is $\frac{n\min\{n,d\}}{4\gamma\min\left\{\frac{1}{4n},\frac{1}{8\bar{z}^2}\right\}} \ge \frac{n}{4\gamma} \ge 9\bar{z}^2 \ge 1 \ge \varepsilon$, we can conclude that

$$\left\|\mathbf{Z}_a^{\bar{k}+1}\right\|_F^2 \le \varepsilon. \tag{54}$$

Inductively, repeating these steps, we get $\left\|\mu_\ell^k - \bar{\mu}_a\right\|^2 \le \varepsilon$ for all $k \in \{\bar{k},\dots,\hat{K}\}$ and $\ell \in I_a$. Notice that

$$\left\|\mathbf{Z}_a^{\hat{K}}\right\|_F^2 \le \varepsilon \le 2\bar{z} \tag{55}$$

for all $a \in A^*$. Repeating the previous steps again, we can get $\left\|\mu_\ell^k - \bar{\mu}_a\right\|^2 \le \left\|\mathbf{Z}_a^k\right\|_F^2 \le \varepsilon$ for all $k \ge \hat{K}$, $a \in A^*$, and $\ell \in I_a$. □

### E.1 Auxiliary lemmas

The following lemmas are used in the proof of Theorem E.1 and rely on the definitions and notation introduced therein.

**Lemma E.2.** *Let* $R_{\max} := \max\{\max_{i\in[n]}\left\|\mu_i^k\right\|, \max_{a\in[m]}\|\bar{\mu}_a\|\}$ *for some* $k \ge 0$. *Then,*

$$\left\|\mathcal{E}_\ell^k(x)\right\| \le 52\max\{R_{\max}, R_{\max}^3\}\max\{\hat{\tau}_{a_\ell^*}^k(x), \bar{\tau}_{a_\ell^*}(x)\}.$$

*for all* $x \in \mathbb{R}^d$ *and* $\ell \in [n]$.

*Proof.* Clearly,

$$\left\|m_{\mu^k}(x) - m_{a_\ell^*}^k(x)\right\| = \left\|\sum_{i=1}^n r_i^k(x)\mu_i^k - \sum_{i\in I_{a_\ell^*}} r_i^{k,a_\ell^*}(x)\mu_i^k\right\| = \left\|\sum_{i\in I_{a_\ell^*}}(r_i^k(x) - r_i^{k,a_\ell^*}(x))\mu_i^k + \sum_{i\notin I_{a_\ell^*}} r_i^k(x)\mu_i^k\right\|.$$

Since

$$\begin{aligned} r_i^k(x) - r_i^{k,a_\ell^*}(x) &= \frac{\exp\left(-\frac{1}{2}\|x-\mu_i^k\|^2\right)}{\sum_{j=1}^n\exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)} - \frac{\exp\left(-\frac{1}{2}\|x-\mu_i^k\|^2\right)}{\sum_{j\in I_{a_\ell^*}}\exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)} \\ &= r_i^{k,a_\ell^*}(x)\left(\frac{\sum_{j\in I_{a_\ell^*}}\exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)}{\sum_{j=1}^n\exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)} - 1\right) \\ &= -r_i^{k,a_\ell^*}(x)\left(\frac{\sum_{j\notin I_{a_\ell^*}}\exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)}{\sum_{j=1}^n\exp\left(-\frac{1}{2}\|x-\mu_j^k\|^2\right)}\right) = -r_i^{k,a_\ell^*}(x)\sum_{j\notin I_{a_\ell^*}} r_j^k(x), \end{aligned}$$

for all $i \in I_{a_\ell^*}$,

$$\left\|m_{\mu^k}(x) - m_{a_\ell^*}^k(x)\right\| = \left\|\sum_{i\notin I_{a_\ell^*}} r_i^k(x)\left(\mu_i^k - \sum_{i\in I_{a_\ell^*}} r_i^{k,a_\ell^*}(x)\mu_i^k\right)\right\| \le 2\hat{\tau}_{a_\ell^*}^k(x)R_{\max}. \tag{56}$$

Next,

$$\left\|\bar{m}(x) - \bar{\mu}_{a_\ell^*}\right\| = \left\|\sum_{b=1}^m \bar{r}_b(x)(\bar{\mu}_b - \bar{\mu}_{a_\ell^*})\right\| \le \bar{\tau}_{a_\ell^*}(x)\max_{b\in[m]}\left\|\bar{\mu}_b - \bar{\mu}_{a_\ell^*}\right\| \le 2\bar{\tau}_{a_\ell^*}(x)R_{\max}, \tag{57}$$

$$
\begin{aligned}
&\left\|\mathbf{C}_{\mu^k}(x)-\mathbf{C}^k_{a^*_\ell}(x)\right\|_{\mathrm{op}}\\
&=\left\|\sum_{i=1}^{n} r_i^k(x)\big(\mu_i^k-m_{\mu^k}(x)\big)\big(\mu_i^k-m_{\mu^k}(x)\big)^\top-\sum_{i\in I_{a^*_\ell}} r_i^{k,a^*_\ell}(x)\big(\mu_i^k-m^k_{a^*_\ell}(x)\big)\big(\mu_i^k-m^k_{a^*_\ell}(x)\big)^\top\right\|_{\mathrm{op}}\\
&\le\left\|\sum_{i\notin I_{a^*_\ell}} r_i^k(x)\big(\mu_i^k-m_{\mu^k}(x)\big)\big(\mu_i^k-m_{\mu^k}(x)\big)^\top\right\|_{\mathrm{op}}\\
&\quad+\left\|\sum_{i\in I_{a^*_\ell}}\Big(r_i^k(x)-r_i^{k,a^*_\ell}(x)\Big)\big(\mu_i^k-m_{\mu^k}(x)\big)\big(\mu_i^k-m_{\mu^k}(x)\big)^\top\right\|_{\mathrm{op}}\\
&\quad+\left\|\sum_{i\in I_{a^*_\ell}} r_i^{k,a^*_\ell}(x)\left[\big(\mu_i^k-m_{\mu^k}(x)\big)\big(\mu_i^k-m_{\mu^k}(x)\big)^\top-\big(\mu_i^k-m^k_{a^*_\ell}(x)\big)\big(\mu_i^k-m^k_{a^*_\ell}(x)\big)^\top\right]\right\|_{\mathrm{op}}.
\end{aligned}
$$

First,

$$
\left\|\sum_{i\notin I_{a^*_\ell}} r_i^k(x)\big(\mu_i^k-m_{\mu^k}(x)\big)\big(\mu_i^k-m_{\mu^k}(x)\big)^\top\right\|_{\mathrm{op}}\le\sum_{i\notin I_{a^*_\ell}} r_i^k(x)\left\|\mu_i^k-m_{\mu^k}(x)\right\|^2\le 4R^2_{\max}\sum_{i\notin I_{a^*_\ell}} r_i^k(x)=4R^2_{\max}\hat\tau^k_{a^*_\ell}(x).
$$

Second, using

$$
r_i^k(x)-r_i^{k,a^*_\ell}(x)=-r_i^{k,a^*_\ell}(x)\hat\tau^k_{a^*_\ell}(x),\qquad i\in I_{a^*_\ell},
$$

we get

$$
\begin{aligned}
&\left\|\sum_{i\in I_{a^*_\ell}}\Big(r_i^k(x)-r_i^{k,a^*_\ell}(x)\Big)\big(\mu_i^k-m_{\mu^k}(x)\big)\big(\mu_i^k-m_{\mu^k}(x)\big)^\top\right\|_{\mathrm{op}}\\
&\le\sum_{i\in I_{a^*_\ell}}\left|r_i^k(x)-r_i^{k,a^*_\ell}(x)\right|\left\|\mu_i^k-m_{\mu^k}(x)\right\|^2\\
&=\hat\tau^k_{a^*_\ell}(x)\sum_{i\in I_{a^*_\ell}} r_i^{k,a^*_\ell}(x)\left\|\mu_i^k-m_{\mu^k}(x)\right\|^2.
\end{aligned}
$$

Finally, since

$$
\mu_i^k-m_{\mu^k}(x)=\mu_i^k-m^k_{a^*_\ell}(x)+m^k_{a^*_\ell}(x)-m_{\mu^k}(x),
$$

and

$$
\sum_{i\in I_{a^*_\ell}} r_i^{k,a^*_\ell}(x)\big(\mu_i^k-m^k_{a^*_\ell}(x)\big)=0,
$$

the cross terms vanish after averaging. Therefore,

$$
\begin{aligned}
&\left\|\sum_{i\in I_{a^*_\ell}} r_i^{k,a^*_\ell}(x)\left[\big(\mu_i^k-m_{\mu^k}(x)\big)\big(\mu_i^k-m_{\mu^k}(x)\big)^\top-\big(\mu_i^k-m^k_{a^*_\ell}(x)\big)\big(\mu_i^k-m^k_{a^*_\ell}(x)\big)^\top\right]\right\|_{\mathrm{op}}\\
&=\left\|\big(m^k_{a^*_\ell}(x)-m_{\mu^k}(x)\big)\big(m^k_{a^*_\ell}(x)-m_{\mu^k}(x)\big)^\top\right\|_{\mathrm{op}}.
\end{aligned}
$$

Using these bounds,

$$
\begin{aligned}
&\left\|\mathbf{C}_{\mu^k}(x)-\mathbf{C}^k_{a^*_\ell}(x)\right\|_{\mathrm{op}}\\
&\le 4R^2_{\max}\hat\tau^k_{a^*_\ell}(x)+\hat\tau^k_{a^*_\ell}(x)\sum_{i\in I_{a^*_\ell}} r_i^{k,a^*_\ell}(x)\left\|\mu_i^k-m_{\mu^k}(x)\right\|^2+\left\|\Big(m^k_{a^*_\ell}(x)-m_{\mu^k}(x)\Big)\Big(m^k_{a^*_\ell}(x)-m_{\mu^k}(x)\Big)^\top\right\|_{\mathrm{op}}
\end{aligned}
$$

$$
\begin{aligned}
&\le 4R_{\max}^2\hat{\tau}_{a_\ell^*}^k(x) + 4R_{\max}^2\hat{\tau}_{a_\ell^*}^k(x) + \left\|m_{a_\ell^*}^k(x) - m_{\mu^k}(x)\right\|^2 \\
&\le 8R_{\max}^2\hat{\tau}_{a_\ell^*}^k(x) + 4R_{\max}^2\left(\hat{\tau}_{a_\ell^*}^k(x)\right)^2 \\
&\le 12R_{\max}^2\hat{\tau}_{a_\ell^*}^k(x),
\end{aligned}
$$

where we used (56). Using (57),

$$
\left\|\bar{\mathbf{C}}(x)\right\|_{\text{op}} = \left\|\sum_{b=1}^m \bar{r}_b(x)\left(\bar{\mu}_b - \bar{m}(x)\right)\left(\bar{\mu}_b - \bar{m}(x)\right)^\top\right\|_{\text{op}} \le 4(\bar{\tau}_{a_\ell^*}(x))^2R_{\max}^2 + 4\bar{\tau}_{a_\ell^*}(x)R_{\max}^2 \le 8\bar{\tau}_{a_\ell^*}(x)R_{\max}^2
$$

because $\left\|m_{\mu^k}(x)\right\| = \left\|\sum_{i=1}^n r_i^k(x)\mu_i^k\right\| \le R_{\max}$ and $\|\bar{m}(x)\| \le \|\sum_{a=1}^m \bar{r}_a(x)\bar{\mu}_a\| \le R_{\max}$. Combining all bounds together and using the triangle inequality,

$$
\begin{aligned}
&\left\|\mathcal{E}_\ell^k(x)\right\| \\
&\le 2\hat{\tau}_{a_\ell^*}^k(x)R_{\max} + 2\bar{\tau}_{a_\ell^*}(x)R_{\max} + 12R_{\max}^2\hat{\tau}_{a_\ell^*}^k(x)\left\|\mu_\ell^k - m_{a_\ell^*}^k(x)\right\| + 2\hat{\tau}_{a_\ell^*}^k(x)R_{\max}\left\|\mathbf{C}_{\mu^k}(x)\right\|_{\text{op}} \\
&\quad + 8\bar{\tau}_{a_\ell^*}(x)R_{\max}^2\left\|\mu_\ell^k - \bar{m}(x)\right\| \\
&\le 2\hat{\tau}_{a_\ell^*}^k(x)R_{\max} + 2\bar{\tau}_{a_\ell^*}(x)R_{\max} + 24R_{\max}^3\hat{\tau}_{a_\ell^*}^k(x) \\
&\quad + 8\hat{\tau}_{a_\ell^*}^k(x)R_{\max}^3 + 16\bar{\tau}_{a_\ell^*}(x)R_{\max}^3 \le 52\max\{R_{\max}, R_{\max}^3\}\max\{\hat{\tau}_{a_\ell^*}^k(x), \bar{\tau}_{a_\ell^*}(x)\}
\end{aligned}
$$

because

$$
\left\|\mathbf{C}_{\mu^k}(x)\right\|_{\text{op}} \le \sum_{i=1}^n r_i^k(x)\left\|\left(\mu_i^k - m_{\mu^k}(x)\right)\left(\mu_i^k - m_{\mu^k}(x)\right)^\top\right\|_{\text{op}} \le 4R_{\max}^2.
$$

□

**Lemma E.3.** *Let* $R_{\max} := \max\{\max_{i\in[n]}\left\|\mu_i^k\right\|, \max_{a\in[m]}\|\bar{\mu}_a\|\}$ *for some* $k \ge 0$. *Then,*

$$
\left\|\bar{\mathcal{E}}_\ell^k(x)\right\| \le 10\max\{R_{\max}, R_{\max}^3\}\tau_\ell^k(x).
$$

*for all* $x \in \mathbb{R}^d$ *and* $\ell \in [n]$.

*Proof.* Indeed,

$$
\begin{aligned}
\left\|m_{a_\ell^*}^k(x) - \mu_\ell^k\right\| &= \left\|\sum_{i\in I_{a_\ell^*}} r_i^{k,a_\ell^*}(x)\mu_i^k - \mu_\ell^k\right\| = \left\|\sum_{i\in I_{a_\ell^*}, i\ne\ell} r_i^{k,a_\ell^*}(x)(\mu_i^k - \mu_\ell^k)\right\| \\
&\le \sum_{i\in I_{a_\ell^*}, i\ne\ell} r_i^{k,a_\ell^*}(x)\left\|\mu_i^k - \mu_\ell^k\right\| \le 2R_{\max}\tau_\ell^k(x).
\end{aligned}
$$

Thus,

$$
\left\|\bar{\mathcal{E}}_\ell^k(x)\right\| \le 2R_{\max}\tau_\ell^k(x) + 2R_{\max}\tau_\ell^k(x)\left\|\mathbf{C}_{a_\ell^*}^k(x)\right\|_{\text{op}} \le 10\max\{R_{\max}, R_{\max}^3\}\tau_\ell^k(x)
$$

because

$$
\left\|\mathbf{C}_{a_\ell^*}^k(x)\right\|_{\text{op}} \le \sum_{i\in I_{a_\ell^*}} r_i^{k,a_\ell^*}(x)\left\|\left(\mu_i^k - m_{a_\ell^*}^k(x)\right)\left(\mu_i^k - m_{a_\ell^*}^k(x)\right)^\top\right\|_{\text{op}} \le 4R_{\max}^2.
$$

□

**Lemma E.4.** *For our choice of parameters,*

$$
\begin{aligned}
&\Delta_1 \ge \delta_{\max}^2, \quad \Delta_2 \ge \hat{\Delta}_2 \ge 32\log(434nR^3/E_1), \quad \delta_{\min} \ge \sqrt{32\log\left(434R^3m/E_1\right)}, \\
&E_1 \le \frac{R}{2}, \quad \frac{4\sqrt{\bar{\Delta}\Delta_2}}{R^0} \ge 2\sqrt{\hat{\Delta}_2}, \quad \Delta_3 \ge 2\sqrt{\hat{\Delta}_2}, \quad E_1 \le \frac{\sqrt{\hat{\Delta}_2}}{2}.
\end{aligned} \tag{33}
$$

*Proof.* Indeed,

$$\Delta_1 = \frac{2R^0\delta_{\min}\rho}{nm^2\sqrt{d}} \geq \delta_{\max}^2$$

due to the choice of $R^0$ and $\delta_{\min} \geq 1$. Next,

$$\Delta_2 = (R^0)^2 \left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\,n^2}\right)^{\frac{2}{d-1}} \geq \frac{(R^0)^2}{2}.$$

due to the condition on $d$. Since $(R^0)^2 \geq 256\log(55552 \cdot n)$, we get

$$\Delta_2 \geq \frac{(R^0)^2}{2} \geq 32\log\left(55552n(R^0)^4\right) = \hat{\Delta}_2.$$

Next,

$$32\log(434nR^3/E_1) = 32\log(55552n(R^0)^4/\delta_{\min}) \leq 32\log\left(55552n(R^0)^4\right) = \hat{\Delta}_2$$

because $\delta_{\min} \geq 1$. Notice that

$$\sqrt{32\log\left(434R^3m/E_1\right)} = \sqrt{32\log\left(55552(R^0)^4m/\delta_{\min}\right)} \leq \delta_{\min}$$

because $\delta_{\min} \geq 2^{10}\sqrt{\log\left(2^{19}m(R^0)^4\right)}$. Since

$$R^0 \geq \frac{nm^2\sqrt{d}\delta_{\max}^2}{2\rho} \geq \delta_{\min}, \tag{58}$$

we get

$$E_1 = \frac{\delta_{\min}}{16R^0} \leq R^0 = \frac{R}{2}.$$

Due to the definitions of $\bar{\Delta}$ and $\Delta_2$,

$$\frac{4\sqrt{\bar{\Delta}\Delta_2}}{R^0} = 2\sqrt{\hat{\Delta}_2} \times \frac{2\sqrt{\left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\,n^2}\right)^{\frac{2}{d-1}}}}{\sqrt{2}} \geq 2\sqrt{\hat{\Delta}_2},$$

where we use the condition on $d$. Also, due to the condition on $d$ and $\delta_{\min}$,

$$\Delta_3 = \frac{\delta_{\min}}{8}\left(\frac{\rho}{mn^2}\right)^{\frac{1}{d-2}} \geq \frac{\delta_{\min}}{16} \geq 2\sqrt{\hat{\Delta}_2}.$$

The next inequality

$$E_1 = \frac{\delta_{\min}}{8R} \leq \frac{\sqrt{\hat{\Delta}_2}}{2}$$

holds because $R = 2R^0 \overset{(58)}{\geq} \frac{\delta_{\min}}{4}$ and $\hat{\Delta}_2 \geq 1$. □

**Lemma E.5.** *For our choice of parameters,*

$$\begin{aligned}
&\delta_{\min} \geq 8\bar{z}, \quad \bar{z} \geq 1,\\
&\bar{z} \leq \tilde{z} := 65\sqrt{\log\left(2^{16}n(R^0)^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}},\\
&\delta_{\min} \geq \sqrt{128\log\left(\frac{156\sqrt{n}R^3\max\{n,m\}}{E_2}\right)},\\
&\hat{K} \leq \tilde{K} := \frac{n\log\left(R^0\right)}{2\gamma} + \frac{2n\min\{n,d\}\max\left\{4n, 8\tilde{z}^2\right\}}{\gamma\varepsilon},\\
&R \geq 2\bar{z} + \bar{R}.
\end{aligned} \tag{45}$$

*Proof.* Due to the definition of $\bar{z}$,

$$\begin{aligned}
\bar{z} &:= \frac{8\sqrt{\bar{\Delta}}\Delta_4}{R^0} + \frac{8\sqrt{n\bar{\Delta}}\bar{R}}{R^0} + \sqrt{n}E_1 \\
&\overset{(32)}{\leq} \frac{8\sqrt{16\log\left(3472nR^4\right)}\Delta_4}{R^0} + \frac{8\sqrt{16n\log\left(3472nR^4\right)}\bar{R}}{R^0} + \frac{\sqrt{n}\delta_{\min}}{16R^0} \\
&\overset{\text{Def. of } \Delta_4}{\leq} 8\sqrt{16\log\left(3472nR^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}} \\
&\quad + \frac{8\sqrt{16n\log\left(3472nR^4\right)}\bar{R}}{R^0} + \frac{\sqrt{n}\delta_{\min}}{16R^0}.
\end{aligned} \tag{59}$$

Notice that

$$\frac{8\sqrt{16n\log\left(3472nR^4\right)}\bar{R}}{R^0} \overset{(32)}{=} \frac{8\sqrt{16n\log\left(55552n(R^0)^4\right)}\bar{R}}{R^0} \leq \frac{\delta_{\min}}{16}$$

since

$$R^0 \geq 1024\sqrt{16n\max\{\log\left(2^{66}n^3\bar{R}^4\right), 1\}}\bar{R}$$

and $\delta_{\min} \geq 1$, where we use Lemma B.12 and the definition of $R^0$. Moreover, $R^0 \geq \frac{2nm^2\sqrt{d}\delta_{\max}^2}{\rho} \geq 2\sqrt{n}$ due to the definition of $R^0$, and

$$\begin{aligned}
\delta_{\min} &\geq 256\sqrt{16\log\left(55552n(R^0)^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}} \\
&= 256\sqrt{16\log\left(3472nR^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}.
\end{aligned}$$

Thus, $\delta_{\min} \geq 8\bar{z}$. Notice that

$$\begin{aligned}
\bar{z} &:= \frac{8\sqrt{\bar{\Delta}}\Delta_4}{R^0} + \frac{8\sqrt{n\bar{\Delta}}\bar{R}}{R^0} + \sqrt{n}E_1 \\
&\geq 8\sqrt{16\log\left(55552n(R^0)^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}} \\
&\geq 8\sqrt{16\log\left(55552n(R^0)^4\right)}\sqrt{\frac{4}{3}\log\left(\frac{2d}{\rho}\right)} \geq 1.
\end{aligned}$$

Next, we bound $\bar{z}$ using (59):

$$\begin{aligned}
\bar{z} &\leq 8\sqrt{16\log\left(3472nR^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}} \\
&\quad + \frac{8\sqrt{16n\log\left(3472nR^4\right)}\bar{R}}{R^0} + \frac{\sqrt{n}\delta_{\min}}{16R^0}.
\end{aligned}$$

Since $R = 2R^0$, $R^0 \geq \sqrt{n}\bar{R}$, and $R^0 \geq \frac{2nm^2\sqrt{d}\delta_{\max}}{\rho} \geq 2\sqrt{n}\delta_{\max}$,

$$\bar{z} \leq 65\sqrt{\log\left(2^{16}n(R^0)^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}.$$

Next,

$$\frac{1}{E_2} = \max\left\{\frac{2\gamma\hat{K}}{n\bar{z}}, \frac{16\min\{n,d\}\sqrt{\min\{n,d\}}\bar{z}}{\varepsilon^2}\max\left\{4n, 8\bar{z}^2\right\}, \sqrt{\frac{8\gamma\min\{n,d\}}{n\varepsilon^2}\max\left\{4n, 8\bar{z}^2\right\}}\right\}$$

$$\overset{\text{Def. of } \gamma}{\leq} \max\left\{\frac{2\gamma\hat{K}}{n\bar{z}}, \frac{16\min\{n,d\}\sqrt{\min\{n,d\}}\bar{z}}{\varepsilon^2}\max\left\{4n, 8\bar{z}^2\right\}, \sqrt{\frac{2\min\{n,d\}}{9\bar{z}^2\varepsilon^2}\max\left\{4n, 8\bar{z}^2\right\}}\right\}$$
$$\overset{\bar{z}\geq 1}{\leq} \max\left\{\frac{2\gamma\hat{K}}{n\bar{z}}, \frac{16\min\{n,d\}\sqrt{\min\{n,d\}}\bar{z}}{\varepsilon^2}\max\left\{4n, 8\bar{z}^2\right\}, \sqrt{\frac{2n\min\{n,d\}}{\varepsilon^2}}\right\}$$
$$\overset{\bar{z}\geq 1}{\leq} \max\left\{\frac{2\gamma\hat{K}}{n\bar{z}}, \frac{128n^{5/2}\bar{z}^3}{\varepsilon^2}, \sqrt{\frac{2n^2}{\varepsilon^2}}\right\}.$$

Using (27) and (43),

$$\begin{aligned}\hat{K} &\leq \frac{n\log\left(\frac{R^0}{4\sqrt{16\log(55552n(R^0)^4)}}\right)}{2\gamma} + \frac{2n\min\{n,d\}\max\left\{4n, 8\bar{z}^2\right\}}{\gamma\varepsilon}\\ &\leq \frac{n\log\left(R^0\right)}{2\gamma} + \frac{2n\min\{n,d\}\max\left\{4n, 8\bar{z}^2\right\}}{\gamma\varepsilon}.\end{aligned}$$

Thus

$$\begin{aligned}\frac{1}{E_2} &\overset{R^0\geq 1}{\leq} \max\left\{\frac{2\gamma}{n\bar{z}}\left(\frac{n\log\left(R^0\right)}{2\gamma} + \frac{2n^2\max\left\{4n, 8\bar{z}^2\right\}}{\gamma\varepsilon}\right), \frac{128n^{5/2}\bar{z}^3}{\varepsilon^2}, \sqrt{\frac{2n^2}{\varepsilon^2}}\right\}\\ &\leq \max\left\{\frac{2\log\left(R^0\right)}{\bar{z}}, \frac{8n\max\left\{4n, 8\bar{z}^2\right\}}{\varepsilon\bar{z}}, \frac{128n^{5/2}\bar{z}^3}{\varepsilon^2}, \sqrt{\frac{2n^2}{\varepsilon^2}}\right\}\\ &\overset{\bar{z}\geq 1}{\leq} \max\left\{2\log\left(R^0\right), \frac{8n\max\left\{4n, 8\bar{z}\right\}}{\varepsilon}, \frac{128n^{5/2}\bar{z}^3}{\varepsilon^2}, \sqrt{\frac{2n^2}{\varepsilon^2}}\right\}\\ &\leq \frac{128n^{5/2}\log\left(R^0\right)}{\min\{\varepsilon^2,1\}}\bar{z}^3.\end{aligned}$$

Next,

$$\begin{aligned}&\sqrt{128\log\left(\frac{156\sqrt{n}R^3\max\{n,m\}}{E_2}\right)}\\ &\leq \sqrt{128\log\left(1248\sqrt{n}(R^0)^3\max\{n,m\}\frac{128n^{5/2}\log\left(R^0\right)}{\min\{\varepsilon^2,1\}}\bar{z}^3\right)}\\ &\leq \sqrt{128\log\left(\frac{159744n^3(R^0)^4\max\{n,m\}}{\min\{\varepsilon^2,1\}}\bar{z}^3\right)}\\ &\overset{\bar{z}\leq\frac{\delta_{\min}}{8}}{\leq} \sqrt{128\log\left(\frac{312n^3(R^0)^4\max\{n,m\}}{\min\{\varepsilon^2,1\}}\delta_{\min}^3\right)}.\end{aligned}$$

Using Lemma B.13, it sufficient for $\delta_{\min}$ to satisfy

$$\delta_{\min} \geq 46\sqrt{\log\left(\frac{2^{20}n^3(R^0)^4\max\{n,m\}}{\min\{\varepsilon^2,1\}}\right)}$$

and ensure that the fourth inequality of the lemma holds, which indeed holds under our assumption about $\delta_{\min}$. Finally,

$$R \geq 2\bar{z} + \bar{R}$$

since $R \geq 2\bar{R}$ and

$$2\bar{z} \leq 130\sqrt{\log\left(2^{16}n(R^0)^4\right)}\sqrt{\frac{n}{d} + \max\left\{\frac{4}{3}\log\left(\frac{2d}{\rho}\right), \sqrt{\frac{4n}{d}\log\left(\frac{2d}{\rho}\right)}\right\}}$$

$$\leq 130\sqrt{5\log\left(2^{16}n(R^0)^4\right)}\sqrt{n\log\left(\frac{2d}{\rho}\right)}$$

$$\leq 130\sqrt{5\cdot 16\log(2)+5n+5\log\left((R^0)^4\right)}\sqrt{n\log\left(\frac{2d}{\rho}\right)}$$

$$\leq \left(1300\sqrt{n}+260\sqrt{5}\sqrt{\log\left(R^0\right)}\right)\sqrt{n\log\left(\frac{2d}{\rho}\right)} \leq R^0 = \frac{R}{2}$$

because $R^0 \geq \frac{2^{20}n\sqrt{d}}{\rho} \geq \frac{2^{19}n}{\sqrt{2}}\log\left(\frac{2d}{\rho}\right)$ and $R^0 \geq \sqrt{\frac{2^{19}}{\sqrt{2}}}\sqrt{\log(R^0)}\sqrt{n\log\left(\frac{2d}{\rho}\right)}$. □

**Lemma E.6.** *For all $k \geq \bar{K}$ and $a \in A^*$,*

$$\left\langle \mathbf{Z}_a^k, \mathbf{B}_a^k \mathbf{Z}_a^k \right\rangle \geq \min\left\{ \frac{\kappa_a}{4n_a^2}\left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2 \right\},$$

*where $\kappa_a := \frac{n_a}{\min\{n_a,d\}}$.*

*Proof.* Using (41),

$$\begin{aligned}
&\left\langle \mathbf{Z}_a^k, \mathbf{B}_a^k \mathbf{Z}_a^k \right\rangle \\
&= n_a \sum_{i\in I_a}\sum_{j\in I_a} \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ r_i^{k,a}(x) r_j^{k,a}(x) + r_i^{k,a}(x) r_j^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)^\top\left(\mu_j^k - m_a^k(x)\right)\right](\mu_i^k - \bar{\mu}_a)^\top(\mu_j^k - \bar{\mu}_a) \\
&= n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \sum_{i\in I_a} r_i^{k,a}(x)(\mu_i^k - \bar{\mu}_a)^\top \left( \sum_{j\in I_a} r_j^{k,a}(x)(\mu_j^k - \bar{\mu}_a)\right)\right] \\
&\quad + n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \sum_{i\in I_a} r_i^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)^\top \left( \sum_{j\in I_a} r_j^{k,a}(x)\left(\mu_j^k - m_a^k(x)\right)(\mu_j^k - \bar{\mu}_a)^\top\right)(\mu_i^k - \bar{\mu}_a)\right] \\
&= n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| \sum_{i\in I_a} r_i^{k,a}(x)(\mu_i^k - \bar{\mu}_a)\right\|^2\right] + n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| \sum_{i\in I_a} r_i^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)(\mu_i^k - \bar{\mu}_a)^\top\right\|_F^2\right].
\end{aligned}$$

Using the definition of $m_a^k(x)$ and $\sum_{i\in I_a} r_i^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)b^\top = 0$ for all $b \in \mathbb{R}^d$,

$$\begin{aligned}
&\left\langle \mathbf{Z}_a^k, \mathbf{B}_a^k \mathbf{Z}_a^k \right\rangle \\
&= n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| m_a^k(x) - \bar{\mu}_a\right\|^2\right] + n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| \sum_{i\in I_a} r_i^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)\left(\mu_i^k - m_a^k(x)\right)^\top\right\|_F^2\right] \\
&= n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| m_a^k(x) - \bar{\mu}_a\right\|^2\right] + n_a \mathbb{E}_{x\sim p_\mu^{k,a}} \mathrm{tr}\left( \left( \sum_{i\in I_a} r_i^{k,a}(x)\left(\mu_i^k - m_a^k(x)\right)\left(\mu_i^k - m_a^k(x)\right)^\top\right)^2\right)
\end{aligned}$$

Let $\kappa_a := \frac{n_a}{\min\{n_a,d\}}$. Due to $(\mathrm{tr}(\mathbf{A}))^2 \leq \mathrm{rank}(\mathbf{A})\mathrm{tr}(\mathbf{A}^2)$ for all symmetric $\mathbf{A} \in \mathbb{R}^{d\times d}$, we get

$$\begin{aligned}
\left\langle \mathbf{Z}_a^k, \mathbf{B}_a^k \mathbf{Z}_a^k \right\rangle &\geq n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| m_a^k(x) - \bar{\mu}_a\right\|^2\right] + \kappa_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left( \sum_{i\in I_a} r_i^{k,a}(x)\left\|\mu_i^k - m_a^k(x)\right\|^2\right)^2\right] \\
&= n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| m_a^k(x) - \bar{\mu}_a\right\|^2\right] + \kappa_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left( \sum_{i\in I_a} r_i^{k,a}(x)\left\|\mu_i^k - \bar{\mu}_a\right\|^2 - \left\| m_a^k(x) - \bar{\mu}_a\right\|^2\right)^2\right] \\
&\geq n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| m_a^k(x) - \bar{\mu}_a\right\|^2\right] + \kappa_a \left( \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \sum_{i\in I_a} r_i^{k,a}(x)\left\|\mu_i^k - \bar{\mu}_a\right\|^2\right] - \mathbb{E}_{x\sim p_\mu^{k,a}}\left[ \left\| m_a^k(x) - \bar{\mu}_a\right\|^2\right]\right)^2
\end{aligned}$$

$$= n_a \mathbb{E}_{x\sim p_\mu^{k,a}}\left[\left\|m_a^k(x)-\bar\mu_a\right\|^2\right] + \kappa_a\left(\frac{1}{n_a}\left\|\mathbf{Z}_a^k\right\|_F^2 - \mathbb{E}_{x\sim p_\mu^{k,a}}\left[\left\|m_a^k(x)-\bar\mu_a\right\|^2\right]\right)^2$$

since

$$\mathbb{E}_{x\sim p_\mu^{k,a}}\left[\sum_{i\in I_a} r_i^{k,a}(x)\left\|\mu_i^k-\bar\mu_a\right\|^2\right] = \sum_{i\in I_a}\int_{x\in\mathbb{R}^d}\left[r_i^{k,a}(x)p_\mu^{k,a}(x)\left\|\mu_i^k-\bar\mu_a\right\|^2\right]dx$$
$$= \frac{1}{n_a}\sum_{i\in I_a}\int_{x\in\mathbb{R}^d}\left[\left\|\mu_i^k-\bar\mu_a\right\|^2\right]\varphi(x-\mu_i^k)\,dx = \frac{1}{n_a}\sum_{i\in I_a}\left\|\mu_i^k-\bar\mu_a\right\|^2 = \frac{1}{n_a}\left\|\mathbf{Z}_a^k\right\|_F^2$$

Notice also that

$$\mathbb{E}_{x\sim p_\mu^{k,a}}\left[\left\|m_a^k(x)-\bar\mu_a\right\|^2\right] \le \mathbb{E}_{x\sim p_\mu^{k,a}}\left[\sum_{i\in I_a} r_i^{k,a}(x)\left\|\mu_i^k-\bar\mu_a\right\|^2\right] = \frac{1}{n_a}\left\|\mathbf{Z}_a^k\right\|_F^2.$$

Thus,

$$\left\langle \mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle \ge \min_{0\le x\le\bar x}\left[n_a x + \kappa_a(\bar x - x)^2\right],$$

where $\bar x := \frac{1}{n_a}\left\|\mathbf{Z}_a^k\right\|_F^2$. Since

$$\min_{0\le x\le\bar x}\left[n_a x+\kappa_a(\bar x-x)^2\right] \ge \min\left\{\min_{0\le x\le\frac{\bar x}{2}}\kappa_a(\bar x-x)^2, \min_{\frac{\bar x}{2}\le x\le\bar x} n_a x\right\} = \min\left\{\frac{\kappa_a}{4}\bar x^2, \frac{n_a}{2}\bar x\right\},$$

we obtain

$$\left\langle \mathbf{Z}_a^k, \mathbf{B}_a^k\mathbf{Z}_a^k\right\rangle \ge \min\left\{\frac{\kappa_a}{4n_a^2}\left\|\mathbf{Z}_a^k\right\|_F^4, \frac{1}{2}\left\|\mathbf{Z}_a^k\right\|_F^2\right\}.$$

□

### E.2 Corollaries

The following corollary states and simplifies the result of Theorem E.1 assuming that $d \ge n$ and ignoring logarithmic factors.

**Theorem 3.2** (Theorem E.1 when $d \ge n$)**.** *Let $\varepsilon \in (0,1]$ denote the target accuracy. We define the smallest and largest gaps in the target distribution as $\delta_{\min} := \min_{a\ne b}\|\bar\mu_a - \bar\mu_b\| > 0$ and $\delta_{\max} := \max_{a\ne b}\|\bar\mu_a-\bar\mu_b\|$, choose the initialization radius as any $R^0 \ge \tilde\Theta\left(\max\left\{\frac{nm^2\sqrt{d}\delta_{\max}^2}{\rho}, \sqrt{n}\bar R\right\}\right)$, where $\bar R := \max_{b\in[m]}\|\bar\mu_b\|$ and $\rho\in(0,1]$. We sample the initial parameters as $\mu_\ell^0 = R^0\eta_\ell/\|\eta_\ell\|$, where $\eta_\ell \sim \mathcal{N}(0,\mathbf{I})$ for all $\ell\in[n]$ (i.i.d.). Assume that $\delta_{\min} \ge \tilde\Omega(1)$ and $d \ge \tilde\Omega(n)$. If $\gamma \le \tilde\Theta(n)$ and GD (3) with GMMs (2) is run for*

$$\tilde K := \tilde\Theta\left(\frac{n^3}{\gamma\varepsilon}\right)$$

*iterations, then, with probability at least $1-4\rho$, we have that $\left\|\mu_\ell^k - \bar\mu_{a_\ell^*}\right\|^2 \le \varepsilon$, for all $\ell\in[n]$ and $k \ge \tilde K$, where $a_\ell^* = \arg\min_{a\in[m]}\left\|\mu_\ell^0-\bar\mu_a\right\|$ is the index of the target mode closest to the initialization point $\mu_\ell^0$. In particular, $\tilde K = \tilde\Theta\left(\frac{n^2}{\varepsilon}\right)$ if $\gamma = \tilde\Theta(n)$.*

*Proof.* It is sufficient to use $d \ge n$ and simplify the terms from the theorem. □

**Theorem 3.3.** *Using the notation and assumptions of Theorem E.1, we additionally assume that $n \ge \frac{m^2}{2\varepsilon}\log\frac{2m}{\rho}$. Let $n_a$ be the number of student's components that converge to $\bar\mu_a$, i.e., $n_a := |\{\ell\in[n] : a = a_\ell^*\}|$. Then, with probability at least $1-5\rho$, we have $\max_{a\in[m]}\left|\frac{n_a}{n} - p_a\right| \le \sqrt{\varepsilon}/m$ and*

$$\mathrm{TV}(p_{\mu^k}, \bar p) := \frac{1}{2}\int_{\mathbb{R}^d}\left|p_{\mu^k}(x) - \bar p(x)\right|dx \le \sqrt{\varepsilon} + \frac{1}{2}\sum_{a=1}^m\left|p_a - \frac{1}{m}\right| \quad \forall k \ge \tilde K,$$

*where $\eta\sim\mathrm{Unif}(\mathbb{S}^{d-1})$ and $p_a := \mathbb{P}\left(\left\|R^0\eta-\bar\mu_a\right\| < \min_{b\in[m],b\ne a}\left\|R^0\eta-\bar\mu_b\right\|\right)$ is the probability that $\bar\mu_a$ is the closest vector to $R^0\eta$ among $\{\bar\mu_b\}_{b\in[m]}$.*

*Proof.* In addition to the events $\Omega_1, \ldots, \Omega_4$, we also assume that the event $\Omega_5$ from Lemma B.7 holds. Then, the result of Theorem E.1 holds with probability at least $1 - 5\rho$. Using the notations from the proof of Theorem E.1,

$$\begin{aligned}
\mathrm{TV}(p_{\mu^k}, \bar{p}) &:= \frac{1}{2}\int_{\mathbb{R}^d} \left|p_{\mu^k}(x) - \bar{p}(x)\right| dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \left|\frac{1}{n}\sum_{i=1}^{n}\varphi(x-\mu_i^k) - \frac{1}{m}\sum_{a=1}^{m}\varphi(x-\bar{\mu}_a)\right| dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \left|\frac{1}{n}\sum_{a=1}^{m}\sum_{\ell\in I_a}\varphi(x-\mu_\ell^k) - \frac{1}{m}\sum_{a=1}^{m}\varphi(x-\bar{\mu}_a)\right| dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \left|\sum_{a=1}^{m}\left(\frac{1}{n}\sum_{\ell\in I_a}\varphi(x-\mu_\ell^k) - \frac{1}{m}\varphi(x-\bar{\mu}_a)\right)\right| dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} \left|\sum_{a=1}^{m}\left(\frac{1}{n}\sum_{\ell\in I_a}(\varphi(x-\mu_\ell^k) - \varphi(x-\bar{\mu}_a)) + \left(\frac{n_a}{n} - \frac{1}{m}\right)\varphi(x-\bar{\mu}_a)\right)\right| dx \\
&\le \frac{1}{2n}\sum_{a=1}^{m}\sum_{\ell\in I_a}\int_{\mathbb{R}^d}\left|\varphi(x-\mu_\ell^k) - \varphi(x-\bar{\mu}_a)\right| dx + \frac{1}{2}\sum_{a=1}^{m}\left|\frac{n_a}{n} - \frac{1}{m}\right|.
\end{aligned}$$

It is well-known (Barsov & Ulyanov, 1987; Devroye et al., 2018) that for all $a, b \in \mathbb{R}^d$,

$$\frac{1}{2}\int_{\mathbb{R}^d}|\varphi(x-a) - \varphi(x-b)|\, dx = \mathbb{P}\left(-\frac{\|a-b\|}{2} \le \xi \le \frac{\|a-b\|}{2}\right),$$

where $\xi$ is a standard normal random variable. Thus

$$\frac{1}{2}\int_{\mathbb{R}^d}|\varphi(x-a) - \varphi(x-b)|\, dx \le \frac{1}{\sqrt{2\pi}}\|a-b\|$$

and

$$\mathrm{TV}(p_{\mu^k}, \bar{p}) \le \frac{1}{\sqrt{2\pi}n}\sum_{a=1}^{m}\sum_{\ell\in I_a}\left\|\mu_\ell^k - \bar{\mu}_a\right\| + \frac{1}{2}\sum_{a=1}^{m}\left|\frac{n_a}{n} - \frac{1}{m}\right|.$$

Using the result of Theorem E.1,

$$\mathrm{TV}(p_{\mu^k}, \bar{p}) \le \frac{\sqrt{\varepsilon}}{\sqrt{2\pi}} + \frac{1}{2}\sum_{a=1}^{m}\left|\frac{n_a}{n} - \frac{1}{m}\right|,$$

for all $k \ge \tilde{K}$. Next,

$$\mathrm{TV}(p_{\mu^k}, \bar{p}) \le \frac{\sqrt{\varepsilon}}{\sqrt{2\pi}} + \frac{1}{2}\sum_{a=1}^{m}\left|\frac{n_a}{n} - p_a\right| + \frac{1}{2}\sum_{a=1}^{m}\left|p_a - \frac{1}{m}\right|,$$

where $p_a$ is the probability that $\bar{\mu}_a$ is the closest vector to $\mu_1^0$ among $\{\bar{\mu}_b\}_{b\in[m]}$. Using (12),

$$\mathrm{TV}(p_{\mu^k}, \bar{p}) \le \frac{\sqrt{\varepsilon}}{2} + \sqrt{\frac{m^2}{8n}\log\frac{2m}{\rho}} + \frac{1}{2}\sum_{a=1}^{m}\left|p_a - \frac{1}{m}\right|.$$

Since $\frac{m^2}{8n}\log\frac{2m}{\rho} \le \frac{\varepsilon}{4}$, we get the result of the theorem. □

**Corollary 3.4.** *Using the notation and assumptions of Theorem 3.3, if the vectors $\{\bar{\mu}_b\}_{b\in[m]}$ are symmetric, i.e., $p_a = \frac{1}{m}$ for all $a \in [m]$, then the total variation distance satisfies $\mathrm{TV}(p_{\mu^k}, \bar{p}) \le \sqrt{\varepsilon}$ for all $k \ge \tilde{K}$. In particular, if the vectors $\{\bar{\mu}_b\}_{b\in[m]}$ have the same norm and are pairwise orthogonal, then we have $p_a = \frac{1}{m}$ for all $a \in [m]$, and hence $\mathrm{TV}(p_{\mu^k}, \bar{p}) \le \sqrt{\varepsilon}$ for all $k \ge \tilde{K}$.*

*Proof.* In the symmetric case, $p_a = \frac{1}{m}$ for all $a \in [m]$, so

$$\frac{1}{2}\sum_{a=1}^{m}\left|p_a - \frac{1}{m}\right| = 0.$$

By Theorem 3.3, this gives $\mathrm{TV}(p_{\mu^k}, \bar{p}) \le \sqrt{\varepsilon}$.

For orthogonal equal-norm modes, write $\bar{\mu}_a = re_a$ in some orthonormal basis. Then

$$a^*(\eta) = \arg\min_{a\in[m]} \left\|R^0\eta - \bar{\mu}_a\right\|^2 = \arg\max_{a\in[m]} \langle \eta, \bar{\mu}_a\rangle = \arg\max_{a\in[m]} \eta_a,$$

where $\eta \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. The coordinates $\eta_1, \dots, \eta_m$ are exchangeable. Thus

$$p_a = \mathbb{P}\left(a^*(\eta) = a\right) = \frac{1}{m}$$

for all $a \in [m]$.

□

## F Proof of Theorem F.1

**Theorem F.1** (Lower bound). *Let $\varepsilon \in (0,1]$ and $\rho \in (0,1]$. We consider the problem (7) with $m = 1$, $\bar{\mu}_{1,1} = \bar{\mu}_{2,1}$, $d \ge 8 + \max\left\{n_{\max}, 2\log\left(n_{\max}^2/\rho\right), 8\log\left(n_{\max}^4/2\pi\rho^2\right)\right\}$, and $n_{\min} < n_{\max}/32$. Assume that the initialization radius satisfies $R^0 = \max\left\{\sqrt{2048\log(55552\cdot n_{\max})}, 1024\sqrt{16 n_{\max}\max\{\log\left(2^{66}n_{\max}^3\bar{R}^4\right), 1\}}\bar{R}\right\}$, where $\bar{R} := \max\{\|\bar{\mu}_{1,1}\|, \|\bar{\mu}_{2,1}\|\}$. We sample the initial parameters as $\mu_{i,\ell}^0 = R^0\eta_{i,\ell}/\|\eta_{i,\ell}\|$, where $\eta_{i,\ell} \sim \mathcal{N}(0,\mathbf{I})$ for all $i \in \{1,2\}, \ell \in [n_i]$, and the random vectors $\{\eta_{i,\ell}\}_{i\in\{1,2\},\ell\in[n_i]}$ are i.i.d. If the step size $\gamma \le 2n_{\min}$, then, with probability at least $1-\rho$, for all $k < \tilde{K}_1 := \frac{n_{\max}}{16 n_{\min}} \times \log\left(\frac{R^0}{16\sqrt{\log(55552 n_{\max}(R^0)^4)}}\right)$, there exists $i \in \{1,2\}$ and $\ell \in [n_i]$ such that $\left\|\mu_{i,\ell}^k - \bar{\mu}_{i,1}\right\|^2 > \varepsilon$. If instead $\gamma > 2n_{\min}$, then, with probability at least $1-\rho$, the method diverges.*

*Proof.* We define $\bar{\mu}_1 := \bar{\mu}_{1,1} = \bar{\mu}_{2,1}$. If the step size $\gamma \le 2n_{\min}$, then the proof is the same as in Theorem E.1. Consider an event $\Omega = \Omega_2$, where

$$\Omega_2 := \left\{\forall i \ne j \in [n_{\max}] : \left\|\mu_{2,i}^0 - \mu_{2,j}^0\right\|^2 \ge \Delta_2\right\}.$$

Using Lemma B.3, we get $\mathbb{P}(\Omega) \ge 1-\rho$ with $\Delta_2 = (R^0)^2\left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\, n_{\max}^2}\right)^{\frac{2}{d-1}}$. Similarly to the proof of Theorem E.1, we obtain

$$\mu_{2,i}^{k+1} = \bar{\mu}_1 + \left(1 - \frac{2\gamma}{n_{\max}}\right)^{k+1}(\mu_{2,i}^0 - \bar{\mu}_1) - E_{2,i}^k, \tag{60}$$

where

$$E_{2,\ell}^k := \frac{2\gamma}{n_{\max}}\sum_{j=0}^{k}\left(1 - \frac{2\gamma}{n_{\max}}\right)^{k-j}\mathbb{E}_{x\sim\mathcal{N}(\mu_{2,\ell}^j,\mathbf{I})}\left[\bar{\mathcal{E}}_{2,\ell}^j(x)\right],$$

and

$$\bar{\mathcal{E}}_{2,\ell}^k(x) := m_{\mu_2^k}(x) - \mu_{2,\ell}^k + \mathbf{C}_{\mu_2^k}(x)\left(\mu_{2,\ell}^k - m_{\mu_2^k}(x)\right).$$

Moreover

$$\left\|E_{2,\ell}^k\right\| \le \max_{j\in\{0,\dots,k\}}\mathbb{E}_{x\sim\mathcal{N}(\mu_{2,\ell}^j,\mathbf{I})}\left[\left\|\bar{\mathcal{E}}_{2,\ell}^j(x)\right\|\right]. \tag{61}$$

This holds since $\gamma < \frac{n_{\max}}{2}$. Similarly, for all $0 \le k \le \bar{K}$, we can prove that

$$\left\|\mu_{2,\ell}^k\right\| \le R, \tag{62}$$

$$\left\|\mu_{2,\ell}^k - \mu_{2,j}^k\right\|^2 \ge \hat{\Delta}_2 \tag{63}$$

for all $\ell \ne j \in [n_{\max}]$, where

$$R := 2R^0, \quad \hat{\Delta}_2 := 32 \log\left(3472 n_{\max} R^4\right), \quad \bar{\Delta} := \frac{\hat{\Delta}_2}{2}, \tag{64}$$

and

$$\bar{K} := \left\lfloor \log\left(\frac{R^0}{4\sqrt{\bar{\Delta}}}\right) / \left(-\log\left(1 - \frac{2\gamma}{n_{\max}}\right)\right) \right\rfloor.$$

Moreover,

$$\left\|E_{2,\ell}^k\right\| \le 248 n_{\max} \max\{R, R^3\} \exp\left(-\hat{\Delta}_2/32\right) \le \frac{1}{4}$$

for all $0 \le k \le \bar{K} - 1$. Using (60), for all $1 \le k \le \bar{K}$,

$$\begin{aligned}\left\|\mu_{2,i}^k - \bar{\mu}_1\right\| &\ge \left(1 - \frac{2\gamma}{n_{\max}}\right)^k \left\|(\mu_{2,i}^0 - \bar{\mu}_1)\right\| - \left\|E_{2,i}^{k-1}\right\| \\ &\ge 2\sqrt{\bar{\Delta}} - \frac{1}{4} \ge 2 > \sqrt{\varepsilon}\end{aligned}$$

due to the choice of $\bar{K}$. Thus, the method required at least

$$\left\lfloor \log\left(\frac{R^0}{4\sqrt{\bar{\Delta}}}\right) / \left(-\log\left(1 - \frac{2\gamma}{n_{\max}}\right)\right) \right\rfloor \ge \frac{n_{\max}}{16 n_{\min}} \log\left(\frac{R^0}{16\sqrt{\log\left(55552 n_{\max} (R^0)^4\right)}}\right)$$

iterations.

We now consider the case when $\gamma > 2n_{\min}$. Consider an event $\Omega = \Omega_1$, where

$$\Omega_1 := \left\{\forall i \ne j \in [n_{\min}] : \left\|\mu_{1,i}^0 - \mu_{1,j}^0\right\|^2 \ge \Delta_1\right\} \tag{65}$$

and $\Delta_1 = (R^0)^2 \left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\, n_{\min}^2}\right)^{\frac{2}{d-1}}$. Using Lemma B.3, we get $\mathbb{P}(\Omega_1) \ge 1 - \rho$. Using mathematical induction, we assume that

$$\left\|\mu_{1,i}^k\right\| \le 2R\left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^k, \tag{66}$$

$$\left\|\mu_{1,i}^k - \mu_{1,j}^k\right\|^2 \ge \left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{2k} \frac{\Delta_1}{2} \tag{67}$$

for all $k \ge 0$, and for all $i \ne j \in [n_{\min}]$, where

$$R := 2R^0.$$

The base case holds due to $\left\|\mu_{1,i}^0\right\| = R^0$ and (65). Similarly to the proof of Theorem E.1, we get

$$\mu_{1,i}^{k+1} = \bar{\mu}_1 + \left(1 - \frac{2\gamma}{n_{\min}}\right)^{k+1} (\mu_{1,i}^0 - \bar{\mu}_1) - E_{1,i}^k, \tag{68}$$

where

$$E_{1,\ell}^k := \frac{2\gamma}{n_{\min}} \sum_{j=0}^{k} \left(1 - \frac{2\gamma}{n_{\min}}\right)^{k-j} \mathbb{E}_{x \sim \mathcal{N}(\mu_{1,\ell}^j, \mathbf{I})}\left[\bar{\mathcal{E}}_{1,\ell}^j(x)\right],$$

and

$$\bar{\mathcal{E}}_{1,\ell}^k(x) := m_{\mu_1^k}(x) - \mu_{1,\ell}^k + \mathbf{C}_{\mu_1^k}(x)\left(\mu_{1,\ell}^k - m_{\mu_1^k}(x)\right).$$

Let us define $r_{1,i}^k(x) := \frac{\exp\left(-\frac{1}{2}\|x-\mu_{1,i}^k\|^2\right)}{\sum_{j\in[n_{\min}]}\exp\left(-\frac{1}{2}\|x-\mu_{1,j}^k\|^2\right)}$ and $\tau_{1,\ell}^k(x) := \sum_{i\in[n_{\min}],i\neq\ell} r_{1,i}^k(x)$. Using the same arguments as in Lemma E.3 and (36),

$$\left\|\bar{\mathcal{E}}_{1,\ell}^k(x)\right\| \leq 80R^3\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{3k}\tau_{1,\ell}^k(x)$$

with

$$\mathbb{E}_{x\sim\mathcal{N}(\mu_{1,\ell}^k,\mathbf{I})}\tau_{1,\ell}^k(x) \leq 2n_{\min}\exp\left(-\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{2k}\frac{\Delta_1}{64}\right),$$

where we use (67). Therefore,

$$\begin{aligned}
\left\|E_{1,\ell}^k\right\| &\leq \left\|\frac{2\gamma}{n_{\min}}\sum_{j=0}^{k}\left(1-\frac{2\gamma}{n_{\min}}\right)^{k-j}\mathbb{E}_{x\sim\mathcal{N}(\mu_{1,\ell}^j,\mathbf{I})}\left[\bar{\mathcal{E}}_{1,\ell}^j(x)\right]\right\| \\
&\leq \frac{2\gamma}{n_{\min}}\sum_{j=0}^{k}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k-j}\mathbb{E}_{x\sim\mathcal{N}(\mu_{1,\ell}^j,\mathbf{I})}\left\|\bar{\mathcal{E}}_{1,\ell}^j(x)\right\| \\
&\leq 320R^3\gamma\sum_{j=0}^{k}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k-j}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{3j}\exp\left(-\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{2j}\frac{\Delta_1}{64}\right).
\end{aligned}$$

Let $x := \left|1-\frac{2\gamma}{n_{\min}}\right| > 2$, then

$$\left\|E_{1,\ell}^k\right\| \leq 320R^3\gamma\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k}\sum_{j=0}^{k}x^{2j}\exp\left(-x^{2j}\frac{\Delta_1}{64}\right).$$

Due to $x \geq 2$,

$$\left\|E_{1,\ell}^k\right\| \leq 320R^3\gamma\exp\left(-\frac{\Delta_1}{128}\right)\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k}\sum_{j=0}^{k}x^{2j}\exp\left(-x^{2j}\frac{\Delta_1}{128}\right).$$

Since $\Delta_1 \geq 128$ for our choice of $R^0$,

$$\sum_{j=0}^{k}x^{2j}\exp\left(-x^{2j}\frac{\Delta_1}{128}\right) \leq \sum_{j=0}^{k}x^{2j}\exp\left(-x^{2j}\right) \leq e^{-1}+\int_1^\infty te^{-t}\,dt \leq 2,$$

and

$$\begin{aligned}
\left\|E_{1,\ell}^k\right\| &\leq \frac{640R^3\gamma}{\exp\left(\frac{\Delta_1}{128}\right)}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k} = \frac{640R^3n_{\min}}{2\exp\left(\frac{\Delta_1}{128}\right)}\times\frac{2\gamma}{n_{\min}}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k} \\
&\leq \frac{640R^3n_{\min}}{\exp\left(\frac{\Delta_1}{128}\right)}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k+1},
\end{aligned}$$

where we use that $\frac{2\gamma}{n_{\min}} \leq 2\left|1-\frac{2\gamma}{n_{\min}}\right|$. Recall that

$$\Delta_1 = (R^0)^2\left(\frac{\sqrt{2\pi}\rho}{\sqrt{d}\,n_{\min}^2}\right)^{\frac{2}{d-1}} \geq \frac{(R^0)^2}{2}$$

due to the condition on $d$. Thus

$$\left\|E_{1,\ell}^k\right\| \leq \frac{5120(R^0)^3n_{\min}}{\exp\left(\frac{\Delta_1}{128}\right)}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k+1} \leq \frac{5120(R^0)^3n_{\min}}{\exp\left(\frac{(R^0)^2}{512}\right)}\left(\left|1-\frac{2\gamma}{n_{\min}}\right|\right)^{k+1}.$$

Since $R^0 \geq \sqrt{2048 \log(55552 \cdot n_{\max})}$, we get

$$\left\|E_{1,\ell}^k\right\| \leq \frac{1}{4}\left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1}. \tag{69}$$

Using (68),

$$\begin{aligned}\left\|\mu_{1,i}^{k+1} - \mu_{1,j}^{k+1}\right\| &\geq \left|\left(1 - \frac{2\gamma}{n_{\min}}\right)^{k+1}\right| \left\|\mu_{1,i}^0 - \mu_{1,j}^0\right\| - \left\|E_{1,i}^k\right\| - \left\|E_{1,j}^k\right\| \\ &\geq \left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} \sqrt{\Delta_1} - \left\|E_{1,i}^k\right\| - \left\|E_{1,j}^k\right\| \geq \left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} \frac{\sqrt{\Delta_1}}{\sqrt{2}}\end{aligned}$$

since $\Delta_1 \geq 16$, and

$$\begin{aligned}\left\|\mu_{1,\ell}^{k+1}\right\| &= \left\|\bar{\mu}_1 + \left(1 - \frac{2\gamma}{n_{\min}}\right)^{k+1} (\mu_{1,\ell}^0 - \bar{\mu}_1) - E_{1,\ell}^k\right\| \\ &\leq \|\bar{\mu}_1\| + \left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} (\left\|\mu_{1,\ell}^0\right\| + \|\bar{\mu}_1\|) + \left\|E_{1,\ell}^k\right\|.\end{aligned}$$

Since $\|\bar{\mu}_1\| \leq \frac{R}{2}$, $\left\|\mu_{1,\ell}^0\right\| = R^0 \leq \frac{R}{2}$, and $\left|1 - \frac{2\gamma}{n_{\min}}\right| \geq 1$,

$$\left\|\mu_{1,\ell}^{k+1}\right\| \leq \frac{3R}{2}\left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} + \left\|E_{1,\ell}^k\right\| \leq 2R\left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1}$$

because $R \geq 1$. We have proved the next step of the mathematical induction. For all $k \geq 0$,

$$\begin{aligned}\left\|\mu_{1,\ell}^{k+1}\right\| &\geq \left\|\bar{\mu}_1 + \left(1 - \frac{2\gamma}{n_{\min}}\right)^{k+1} (\mu_{1,\ell}^0 - \bar{\mu}_1) - E_{1,\ell}^k\right\| \\ &\geq \left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} (\left\|\mu_{1,\ell}^0\right\| - \|\bar{\mu}_1\|) - \|\bar{\mu}_1\| - \left\|E_{1,\ell}^k\right\| \\ &\geq \left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} (R^0 - \bar{R}) - \bar{R} - \frac{1}{4}\left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} \\ &\geq \left(\left|1 - \frac{2\gamma}{n_{\min}}\right|\right)^{k+1} \frac{R^0}{8} \geq \frac{2^{k+1} R^0}{8}\end{aligned}$$

since $R^0 \geq 16\bar{R}$ and $R^0 \geq 2$. The last inequality proves that $\{\mu_{1,\ell}^k\}_{k \geq 0}$ diverges. □