Proceedings of Machine Learning Research vol 291:1–62, 2025 38th Annual Conference on Learning Theory

Learning general Gaussian mixtures with efficient score matching

Sitan Chen SITAN@SEAS.HARVARD.EDU
Harvard SEAS

Vasilis Kontonis VASILIS@CS.UTEXAS.EDU
University of Texas at Austin

Kulin Shah KULINSHAH@UTEXAS.EDU
University of Texas at Austin

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the problem of learning mixtures of k Gaussians in d dimensions. We make no separation assumptions on the underlying mixture components: we only require that the covariance matrices have bounded condition number and that the means and covariances lie in a ball of bounded radius. We give an algorithm that draws $d^{\mathrm{poly}(k/\varepsilon)}$ samples from the target mixture, runs in sample-polynomial time, and constructs a sampler whose output distribution is ε-close from the unknown mixture in total variation. Prior works for this problem either (i) required exponential runtime in the dimension d, (ii) placed strong assumptions on the instance (e.g., spherical covariances or clusterability), or (iii) had doubly exponential dependence on the number of components k.

Our approach departs from commonly used techniques for this problem like the method of moments. Instead, we leverage a recently developed reduction, based on diffusion models, from distribution learning to a supervised learning task called score matching. We give an algorithm for the latter by proving a structural result showing that the score function of a Gaussian mixture can be approximated by a piecewise-polynomial function, and there is an efficient algorithm for finding it. To our knowledge, this is the first example of diffusion models achieving a state-of-the-art theoretical guarantee for an unsupervised learning task.

Keywords: Mixtures of Gaussians, Diffusion models, Learning and generalization.

1. Introduction

Gaussian mixture models (GMMs) are one of the most well-studied models in statistics, with a history going back to the work of Pearson (1894). Its computational study was initiated in the work of Dasgupta (1999a); since then, it has been one of the prototypical non-convex learning problems that has attracted significant attention from the theoretical computer science community (Vempala and Wang, 2002; Kannan et al., 2005; Brubaker and Vempala, 2008a; Kalai et al., 2010; Moitra and Valiant, 2010; Belkin and Sinha, 2015; Hopkins and Li, 2018; Kothari et al., 2018; Diakonikolas et al., 2020; Bakshi and Kothari, 2020; Diakonikolas and Kane, 2020; Liu and Li, 2022; Liu and Moitra, 2023; Bakshi et al., 2022; Buhai and Steurer, 2023).

Learning without separation We focus on learning even when parameter recovery is impossible, i.e., without assuming that the components of the mixture are separated. In this setting, the learner has to produce a hypothesis that is close to the target GMM in total variation distance (Feldman et al., 2008; Moitra and Valiant, 2010; Chan et al., 2013; Suresh et al., 2014; Daskalakis and Kamath, 2014; Diakonikolas et al., 2016; Acharya et al., 2017; Li and Schmidt, 2017; Ashtiani et al., 2018; Diakonikolas and Kane, 2020; Bakshi et al., 2022; Buhai and Steurer, 2023).

Statistically, this problem is essentially completely understood: in order to approximate the target mixture of k Gaussians in ε total variation distance, it is known that $\widetilde{\Theta}(kd^2/\varepsilon^2)$ samples are sufficient and also necessary (Ashtiani et al., 2018). Even though statistically almost optimal, the algorithm of Ashtiani et al. (2018) has a runtime scaling exponentially in $\tilde{O}(kd^2)$. This exponential dependence on the dimension is due to the fact that their algorithm is based on brute-force enumeration.

Despite significant efforts, the computational aspects of the problem are still far from well-understood. Suresh et al. (2014) provided an algorithm for learning mixtures of spherical (i.e., with covariance matrices that are multiples of the identity $\mathbf{Id}$) with $\mathrm{poly}(dk/\varepsilon)$ sample complexity and $\mathrm{poly}(d)(k/\varepsilon)^{\mathrm{poly}(k)}$ runtime. For spherical Gaussians, the runtime was more recently improved to quasi-polynomial in k: in Diakonikolas and Kane (2020), a runtime and sample complexity of $\mathrm{poly}(d)(k/\varepsilon)^{\log^2 k}$ was given.

For GMMs with general covariance matrices, the focus of the present work, the best-known runtime is due to Bakshi et al. (2022) and is doubly exponential in the number of components k, i.e., $(d/\varepsilon)^k(1/\varepsilon)^{k^{k^2}}$. To the best of our knowledge, this doubly exponential dependency on k is implicit in all works on learning general GMMs using the method of moments (Moitra and Valiant, 2010; Bakshi and Kothari, 2020; Diakonikolas et al., 2020; Liu and Moitra, 2023) (see Section 3.1 for intuition for where this comes from).

In particular, for any $k = \Omega(\sqrt{\log d})$, previously there was no algorithm that ran in time faster than the exponential-time algorithm of Ashtiani et al. (2018), even for constant ϵ!

On the negative side, there is strong evidence in the form of statistical query (SQ) (Diakonikolas et al., 2017) and lattice-based (Bruna et al., 2021; Gupte et al., 2022) hardness that runtime which scales super-polynomially in the number of components k is necessary. More precisely, the SQ lower bound of Diakonikolas et al. (2017) implies that even to learn within constant accuracy $\varepsilon > 0$, $d^{\Omega(k)}$ runtime is required. Our work aims to bridge the gaps between the best-known upper and lower bounds for learning GMMs – we ask the following fundamental question.

What is the best possible runtime for learning general Gaussian mixture models with k components? Can we improve over the doubly exponential runtime of moment-based methods?

We make significant progress towards answering this question. Under mild "condition number" bounds on the mixture components – and without assuming the components are separated – we give an algorithm that achieves runtime $d^{\mathrm{poly}(k)}$ for any constant accuracy $\varepsilon > 0$. Thus, for well-conditioned mixtures, our result improves *exponentially* over the best-known runtime of Bakshi et al. (2022) in the regime where $k = \Omega(\sqrt{\log d})$.

Diffusion models and learning Interestingly, our algorithm *does not rely on matching moments* with the target mixture. Instead, we draw inspiration from the recent literature on proving theoretical guarantees for diffusion models (De Bortoli et al., 2021; Block et al., 2022; Chen et al., 2022; De Bortoli, 2022; Lee et al., 2022; Liu et al., 2022; Pidstrigach, 2022; Wibisono and Yang, 2022; Chen et al., 2023c,d; Lee et al., 2023; Li et al., 2023; Benton et al., 2023b; Chen et al., 2023b; Benton et al., 2023a; Conforti et al., 2023; Wibisono et al., 2024), the state-of-the-art method in practice for audio and image generation (Sohl-Dickstein et al., 2015; Dhariwal and Nichol, 2021; Song et al., 2020; Ho et al., 2020). These works culminated in the key finding that for any distribution with bounded second moment, there is a reduction from distribution learning to a supervised learning task called *score matching*. Roughly speaking, this task is defined as follows: given a sample from the target distribution that has been corrupted by some Gaussian noise, predict the noise that was used to generate the sample (see Section 2.1 for an exposition of these concepts). Despite the striking level of generality with which this reduction holds, these works fell short of giving "end-to-end" learning guarantees as they didn't address how to actually perform score matching algorithmically.

Our main technical contribution is an algorithm for score matching for GMMs. This relies on a novel structural result showing that the score function of a GMM can be well-approximated by a piecewise polynomial, together with an efficient procedure to recover the polynomial pieces.

While diffusion models have achieved remarkable empirical successes (Betker et al., 2023), to our knowledge our guarantee marks the first example of an unsupervised learning problem where diffusion models can even yield improved *theoretical* guarantees. Our techniques are a synthesis of this modern algorithmic technique on the one hand and classic ideas from theoretical computer science like low-degree approximation on the other. We leave it as an intriguing open question to identify other problems for which this marriage of toolkits could prove useful.

1.1. Our results and techniques

We first give the formal definition of the well-conditioned GMMs that we consider in this work. Roughly, we require that the covariance matrices of the components are well-conditioned in the sense that their eigenvalues are upper and lower bounded and that the means and covariances lie within an ℓ_2 ball of bounded radius.

Definition 1 (Well-Conditioned Gaussian Mixture) *Let $\mathcal{N}_1, \ldots, \mathcal{N}_k$ be d-dimensional Gaussian distributions with means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and covariances $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$. We denote by $\mathcal{M}$ the mixture of these distributions with weights $\lambda_1, \ldots, \lambda_k$. We will say that $\mathcal{M}$ is τ-well-conditioned if for some $\alpha \leq 1 \leq \beta$ and $R > 0$ with $(\beta/\alpha) \log R \leq \tau$, it holds that: for all i, $\alpha\, \mathbf{Id} \preceq \mathbf{Q}_i \preceq \beta\, \mathbf{Id}$ and $\|\boldsymbol{\mu}_i\|_2 + \|\mathbf{Q}_i - \mathbf{Id}\|_F \leq R$. When we want to distinguish between parameters we will also say that $\mathcal{M}$ is (α, β, R)-well-conditioned. Moreover, we denote by $\lambda_{\min}$ the minimum weight $\min_{i \in [k]} \lambda_i$.*

We now present our main result: an efficient algorithm for learning well-conditioned GMMs.

Theorem 2 (Informal – Learning Gaussian mixtures, see Theorem 11) *Let $\mathcal{M}$ be a τ-well-conditioned mixture of k Gaussians in d dimensions, and suppose $\lambda_{\min} \geq 1/\mathrm{poly}(k)$. There exists an algorithm that draws $N = d^{\mathrm{poly}(k\tau/\varepsilon)}$ samples from $\mathcal{M}$, runs in sample-polynomial time, and constructs a sampling oracle whose output distribution is ε-close to $\mathcal{M}$ in total variation. To generate a new sample the oracle requires $\mathrm{poly}(N, d)$ time.*

To our knowledge, this is the first example of an unsupervised learning problem for which a diffusion-based sampler outperforms existing state-of-the-art theoretical approaches (Moitra and Valiant, 2010; Bakshi et al., 2022). In particular, when the number of components k is super-constant, i.e., $k = \Omega(\sqrt{\log d})$, we obtain a quasipolynomial $2^{\mathrm{poly}(\log d)}$ runtime, improving over the exponential $2^{\mathrm{poly}(d)}$ runtime following from Bakshi et al. (2022). Moreover, we remark that using moment methods for Gaussian mixtures, e.g., Bakshi et al. (2022), results in a doubly exponential runtime in k even for well-conditioned mixtures, see Section 3.1. Finally, our improvements hold for any $\epsilon = 1/\mathrm{polylog}(d)$. In fact, prior to our work, nothing better than doubly exponential in k was known even for constant accuracy $\epsilon = \Omega(1)$. We leave investigating whether the dependency on $1/\epsilon$ can be improved as an interesting question for future work.

Learning mixtures of degenerate Gaussians. As stated, Theorem 2 does not appear to give anything for mixtures with covariances that are not full rank. This includes, for instance, mixtures of linear regressions and mixtures of linear subspaces (Chen et al., 2020; Diakonikolas and Kane, 2020). It turns out that we can still give a learning guarantees in this case, though in *Wasserstein distance* rather than total variation, see Theorem 4.

On the condition number assumption. The aforementioned works on general Gaussian mixtures, e.g. Bakshi et al. (2022), do not need to assume a condition number or radius bound like in Theorem 1, and we leave as an important open question whether we can similarly do away with this assumption using our techniques. Nevertheless, we view the assumption as relatively mild, and to our knowledge, it is unclear how to exploit it to improve upon the doubly exponential runtimes of existing moment-based methods. In fact, the reason why those methods incur this dependence appears to be present even when $d = 1$ and the components have unit variance.

Additionally, note that the construction in the aforementioned statistical query lower bound Diakonikolas et al. (2017), often referred to as the "parallel pancakes" instance, is $\mathrm{poly}(k)$-well-conditioned. In that example, the covariances are $\mathbf{Id} - (1-\eta)\mathbf{v}\mathbf{v}^\intercal$ for some parameter $\eta = 1/\mathrm{poly}(k)$, while the means have norm at most $\mathrm{poly}(k)$, so we can take $\alpha = \eta = 1/\mathrm{poly}(k)$, $\beta = 1$, and $R = \mathrm{poly}(k)$ in Theorem 1, resulting in a learning algorithm with runtime $d^{\mathrm{poly}(k/\epsilon)}$. As the statistical query lower bound itself is for learning to within constant error, our guarantee is qualitatively tight, up to the precise polynomial dependence on k in the exponent. See Section 3.2 for further discussion.

1.2. Related work

Learning mixtures of Gaussians A thorough literature review on learning Gaussian mixtures is outside the scope of this work. In addition to the works (Dasgupta, 1999b; Feldman et al., 2008; Moitra and Valiant, 2010; Belkin and Sinha, 2015; Chan et al., 2013; Suresh et al., 2014; Daskalakis and Kamath, 2014; Diakonikolas et al., 2016; Acharya et al., 2017; Li and Schmidt, 2017; Ashtiani et al., 2018; Diakonikolas and Kane, 2020; Bakshi and Kothari, 2020; Diakonikolas et al., 2020; Bakshi et al., 2022; Buhai and Steurer, 2023) mentioned in the introduction which deal with parameter estimation or distribution learning, we also mention a related line of work on *clustering* Gaussian mixtures. This is a setting where there is a large enough separation between components that one can reliably identify which component generated a given sample. Some representative works in this line include (Vempala and Wang, 2004; Brubaker and Vempala, 2008b; Regev and Vijayaraghavan, 2017; Hopkins and Li, 2018; Diakonikolas et al., 2018; Kothari et al., 2018; Liu and Li, 2022).

Similar in spirit to the present work is the interesting work of Yan et al. (2023) which also eschews the method of moments in favor of a variational method. Whereas we use diffusion models, they use a certain interacting particle system that approximates a Wasserstein gradient flow. They focus on the case of Gaussian mixtures with identity covariance components. While they prove that the gradient flow itself converges in an asymptotic sense and numerically demonstrate the effectiveness of their approach, they do not prove non-asymptotic, end-to-end learning guarantees like in the present work.

General theory for diffusion models Several works have provided convergence guarantees for DDPMs and variants (De Bortoli et al., 2021; Block et al., 2022; Chen et al., 2022; De Bortoli, 2022; Lee et al., 2022; Liu et al., 2022; Pidstrigach, 2022; Wibisono and Yang, 2022; Chen et al., 2023c,d; Lee et al., 2023; Li et al., 2023; Benton et al., 2023b; Chen et al., 2023b; Benton et al., 2023a). These works assume the existence of an oracle for accurate score estimation and show that diffusion models can learn essentially any distribution over $\mathbb{R}^d$ (e.g. Chen et al. (2023c); Lee et al. (2023) show this for arbitrary compactly supported distributions, and Chen et al. (2022); Benton et al. (2023a) extended this to arbitrary distributions with finite second moment).

Recently, Koehler and Vuong (2023) showed that Langevin diffusion with data-dependent initialization can also learn multimodal distributions like mixtures of Gaussians, provided one can perform score matching but with a doubly exponential overhead in the number of components. Several months after our results were obtained, Koehler et al. (2024); Huang et al. (2024a) refined the bound in Koehler and Vuong (2023) to only a linear overhead. As such, our algorithm for score estimation could also be plugged into this reduction, rather into the diffusion-based reduction, to achieve the same guarantee as in the present work. This difference is not so important for us however as the main contribution of our work is our algorithm for score estimation.

In another sampling context, Anari et al. (2023, 2024) gave fast parallel algorithms based on a similar diffusion-style sampler for various problems like Eulerian tours and determinantal point processes.

End-to-end applications of diffusions In this work, we use a diffusion process as a tool to obtain *end-to-end efficient learning algorithms* and we are not making "black-box" assumptions about the computational or the statistical complexity of learning the score function. The recent works of Shah et al. (2023); Cui et al. (2023) also consider learning Gaussian mixtures, specifically with well-separated identity covariance components, using diffusions and show in different settings that gradient descent can provably perform score matching. The results of Shah et al. (2023); Cui et al. (2023) only apply to the special case of learning spherical Gaussian mixtures — a setting that is already known to admit efficient learning algorithms. The focus of those works is mainly in understanding why gradient descent for score matching can achieve guarantees similar to the prior known results while our goal in this work is to provide new efficient algorithms for general mixtures that are not captured by prior works.

Several recent results use diffusion models to obtain new *sampling* algorithms with a focus on graphical models. This is a different setting than the one considered in the present work: instead of being given samples from the target distribution, one is given a Hamiltonian describing some graphical model, or some combinatorial object such that one would like to sample certain structures defined on it. For example, El Alaoui et al. (2022); Montanari and Wu (2023); Alaoui et al. (2023); Montanari (2023); Huang et al. (2024b) have used Eldan's stochastic localization (Eldan, 2013, 2020) method to give sampling algorithms for certain distributions arising in statistical physics. These works provide an algorithmic implementation for the drift in the diffusion process, which is defined by the score, using approximate message passing and natural gradient descent (see Celentano (2022)).

Finally, in a concurrent and independent work Gatmiry et al. (2024) the authors give diffusion-based algorithms for the special case of learning spherical (identity covariance) Gaussian mixtures, qualitatively matching the best-known results by Diakonikolas and Kane (2020). Our focus here is different: we learn Gaussian mixtures with *general, well-conditioned* covariance matrices and improve over the prior works (Moitra and Valiant, 2010; Bakshi et al., 2022) yielding exponential savings in runtime when the number of components k is not constant, i.e., $k = \Omega(\text{polylog} d)$.

Statistical guarantees for score matching Several recent works have investigated the *statistical* complexity of score matching. Koehler et al. (2023) showed a connection between the statistical efficiency of score matching and functional inequalities satisfied by the data distribution. Pabbaraju et al. (2024) studied score matching for learning log-polynomial distributions. Like in Koehler et al. (2023), they focus on the score function of the base distribution and not noisy versions thereof; as the authors note, in this case, score matching is computationally tractable as it is exactly an instance of polynomial regression, and their focus was on proving that the statistical efficiency of score matching here is comparable to that of maximum likelihood estimation.

Recently, Wibisono et al. (2024) established the optimal rate for score estimation of nonparametric distributions in high dimensions. Chen et al. (2023a); Oko et al. (2023) studied the sample complexity of score matching for nonparametric distributions specifically using a neural network. Mei and Wu (2023) bounded the sample complexity of learning certain graphical models using diffusion models by arguing that neural network layers can implement iterations of certain variational inference algorithms. Again, these guarantees are all statistical in nature rather than algorithmic.

2. Technical overview

In this section, we provide an overview of our approach, sketches for the main arguments, and pointers to the relevant sections for more details.

2.1. Learning via DDPM

Our algorithm is based on a denoising diffusion probabilistic model (DDPM) (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020). Here we give a self-contained exposition of the basic tools from this literature (see Appendix A.2 for details); readers who are familiar with diffusion models may safely skip to Theorem 3 below.

The most common (Song et al., 2020; Montanari, 2023) approach is to consider the Ornstein-Uhlenbeck process, which given some distribution q_0 corresponds to the SDE $d\mathbf{x}_t = -\mathbf{x}_t dt + \sqrt{2} d\mathbf{w}_t$, with $\mathbf{x}_0 \sim q_0$ and $\mathbf{w}_t$ being a standard Brownian motion. The distribution q_0 here corresponds to the target distribution that we want to learn to generate samples from. In what follows, we use q_t to denote the law of the OU process at time t. It holds that q_t converges to the standard normal distribution and in particular at time t we have that

$$\mathbf{x}_t = e^{-t}\mathbf{x}_0 + \sqrt{1 - e^{-2t}}\, \mathbf{z}_t, \quad \text{for} \quad \mathbf{x}_0 \sim q_0, \;\; \mathbf{z}_t \sim \mathcal{N}\,. \tag{1}$$

Given some terminal timestep T of the forward process with distribution q_T, the following reverse process transforms noisy distribution q_T (which is close to standard Gaussian) to the data distribution:

$$\mathrm{d}\mathbf{x}_t^{\leftarrow} = \{\mathbf{x}_t^{\leftarrow} + 2\nabla_{\mathbf{x}} \log q_{T-t}(\mathbf{x}_t^{\leftarrow})\}\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}\mathbf{w}_t \quad \text{with} \quad \mathbf{x}_0^{\leftarrow} \sim q_T\,.$$

In this reverse process, the iterate $\mathbf{x}_t^{\leftarrow}$ is distributed according to q_{T-t} for every $t \in [0, T]$, so that the final iterate $\mathbf{x}_T^{\leftarrow}$ is distributed according to the data distribution q_0. To be able to generate samples using the reverse SDE we need access to the *score function* $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$. Given approximate oracle access to the score function of the target density q_0 (for us this is the mixture of Gaussians) at close enough noise levels, we can discretize the reverse SDE that starts with a sample from the Gaussian noise and generates a sample whose distribution is close to the target density. In particular, for timesteps $t_0, \ldots, t_N$, given estimates $\widehat{\mathbf{s}}(\mathbf{x}, T - t_\ell)$ we will be using the following update rule to generate a sample (sometimes called the exponential integrator scheme as it replaces the time-dependent score term in the reverse SDE with the score approximation at time-step $T - t_\ell$). More precisely, at the ℓ-th iteration, we sample $\mathbf{z}_\ell \sim \mathcal{N}(\mathbf{0}, \mathbf{Id})$ and update our guess as follows:

$$\mathbf{y}_{\ell+1} \leftarrow \rho_\ell \, \mathbf{y}_\ell + 2(\rho_\ell - 1) \, \widehat{\mathbf{s}}(\mathbf{y}_\ell, T - t_\ell) + \sqrt{\rho_\ell^2 - 1} \, \mathbf{z}_\ell, \tag{2}$$

where ρ_ℓ is an appropriately chosen "step-size" parameter, see Algorithm 1 for more details. Several recent works (see, e.g., Chen et al. (2023c); Lee et al. (2023); Chen et al. (2022); Benton et al. (2023a)) have studied the convergence of the above (discretized) reverse SDE to the data distribution under black-box assumptions on the quality of the score estimates $\widehat{\mathbf{s}}(\cdot, \cdot)$. We will be using a recent result from Benton et al. (2023a) (see Theorem 12) that places minimal assumptions on the data distribution and gives fast convergence rates. More precisely, for the case of well-conditioned Gaussian mixtures, it implies that if the score functions are approximated within L_2 error roughly $\mathrm{poly}(\varepsilon/\tau)$, then iterating Equation (2) will produce a sample within total variation distance ε from the target Gaussian mixture after $\mathrm{poly}(d\tau/\varepsilon)$ iterations.

Learning the score We have now reduced the original sampling problem to roughly $N = \mathrm{poly}(d\tau/\varepsilon)$ regression problems to get the approximate score functions at times $t_1, \ldots, t_N$. More precisely for every $t \in \{t_1, \ldots, t_N\}$ we would like to use some expressive enough class of functions $\mathcal{G}$ and solve the following minimization (score-matching) problem: $\min_{\mathbf{g} \in \mathcal{G}} \quad \mathbb{E}_{\mathbf{x}_0, \mathbf{z}_t}[\|\mathbf{g}(\mathbf{x}_t) - \nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t)\|_2^2]$ where $\mathbf{x}_t$ is generated by adding the Gaussian noise $\mathbf{z}_t$ to the sample $\mathbf{x}_0 \sim \mathcal{M}$, $\mathbf{x}_t = e^{-t}\mathbf{x}_0 + \sqrt{1 - e^{-2t}}\mathbf{z}_t$. Since we have sample access to the unknown mixture $\mathcal{M}$, we can generate i.i.d. copies of $\mathbf{x}_t$ to solve the regression task. However, the target score function at noise-level t is not available (as it depends on the density of the unknown mixture). A standard workaround (Hyvärinen, 2005; Vincent, 2011; Ho et al., 2020; Song et al., 2020) is the denoising approach where conditional on the observed $\mathbf{x}_t$ we try to predict the added noise $\mathbf{z}_t$. It is a well-known consequence of Gaussian integration by parts (see e.g. Appendix A of Chen et al. (2023c) for a proof) that the following regression task is equivalent to the original score-matching problem with the benefit that it does not require knowledge of the score function of the distribution q_t (that corresponds to the distribution of $\mathbf{x}_t$): $\min_{\mathbf{g} \in \mathcal{G}} \; L_t(\mathbf{g}) = \min_{\mathbf{g} \in \mathcal{G}} \quad \mathbb{E}_{\mathbf{x}_0, \mathbf{z}_t}\left[\left\|\mathbf{g}(\mathbf{x}_t) + \frac{\mathbf{z}_t}{\sqrt{1 - \exp(-2t)}}\right\|_2^2\right]$.

Our main technical contribution is an efficient algorithm that uses the above denoising formulation of the score-matching problem and yields an approximation to the score function $\widehat{\mathbf{s}}(\mathbf{x}_t)$.

Proposition 3 (Informal - Efficiently Learning the Score - Theorem 49) *Let $\mathcal{M}$ be a τ-well-conditioned mixture. Then, for any $\varepsilon > 0$ and noise scale $t \geq \mathrm{poly}(\varepsilon/\tau)$, there exists an algorithm that draws $d^{\mathrm{poly}(k\tau/\varepsilon)}$ samples from $\mathcal{M}$, runs in sample-polynomial time, and returns a score function $\widehat{\mathbf{s}}(\cdot)$ such that with high probability it holds $\mathbb{E}_{\mathbf{x}_t \sim \mathcal{M}_t}\left[\|\widehat{\mathbf{s}}(\mathbf{x}_t) - \nabla_{\mathbf{x}} \log q_t(\mathbf{x})\|^2\right] \leq \varepsilon$.*

A detailed theorem statement and the details of the algorithm can be found in Theorem 49. The details of the proof of Theorem 3 can be found in Appendix E. Combining the above efficient algorithm with the convergence rate of the reverse SDE we are able to get our end-to-end efficient algorithm for sampling from the mixture $\mathcal{M}$. Our efficient algorithm in Theorem 3 relies on a structural result showing that the score function of the mixture $\mathcal{M}$ can be approximated by a piecewise-polynomial function, and an efficient algorithm to recover the partition of the piecewise polynomial approximation. In the following sections, we describe the main ideas of each part.

Remark 4 (Learning mixtures of low-dimensional (degenerate) Gaussians) *Here we briefly discuss how our techniques can also give a learning guarantee even when the covariances of the components are degenerate. The reason is that we can simply stop the reverse diffusion δ time steps early. Instead of approximately sampling from the original mixture $\mathcal{M}$, this would approximately sample in total variation from a slightly noisy version of $\mathcal{M}$, namely the distribution $\mathcal{M}_\delta$ given by starting at $\mathcal{M}$ and running the forward process for a small amount of time δ. Given a component $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i)$ of $\mathcal{N}$, the corresponding component of $\mathcal{M}_\delta$ is given by $\mathcal{N}(e^{-\delta}\boldsymbol{\mu}_i, e^{-2\delta}\mathbf{Q}_i + (1-e^{-2\delta})\mathbf{Id})$. In particular, the minimum singular value of the covariance is at least $1 - e^{-2\delta} = \Omega(\delta)$, and we can thus apply Theorem 2 to $\mathcal{M}_\delta$ instead of $\mathcal{M}$, incurring exponential dependence on $\mathrm{poly}(1/\delta)$. Moreover, the Wasserstein distance between $\mathcal{M}$ and $\mathcal{M}_\delta$ scales with $\delta(R + \mathrm{poly}(\beta/\alpha))$. Altogether, we find that we can sample from a distribution that is TV-close to a distribution which is Wasserstein-close to $\mathcal{M}$, even if $\mathcal{M}$ might have degenerate covariances.*

2.2. Approximating the score function using piecewise polynomials

We now present the key ideas behind our main technical result showing that a piecewise polynomial approximation of the score function exists. In the following discussion, we will be focusing on estimating the score function of the Gaussian mixture at a specific noise level t. At noise level t, each component of the mixture is rescaled by e^{-t} and convolved with a mean-zero Gaussian with covariance $(1-e^{-2t})\mathbf{Id}$ (see Equation (1)). Therefore, the score function at every noise level corresponds to the score function of a Gaussian mixture with means $e^{-t}\boldsymbol{\mu}_i$ and covariances $e^{-2t}\mathbf{Q}_i + (1-e^{-2t})\,\mathbf{Id}$, where $\boldsymbol{\mu}_i$ and $\mathbf{Q}_i$ denote the parameters of i^{th} component of the original target mixture $\mathcal{M}$. For simplicity, we assume that the minimum mixing weight of the mixture $\mathcal{M}$ is at least $\mathrm{poly}(1/k)$ in the following discussion. It turns out that the bottleneck is to approximate the score function of the original mixture $\mathcal{M}$ and therefore, to keep the notation simple, for this presentation we will focus on this problem. We will denote the score function (i.e., the gradient of the log-density) of a mixture of Gaussians by $\mathbf{s}(\mathbf{x}; \mathcal{M})$:

$$\mathbf{s}(\mathbf{x}; \mathcal{M}) = -\sum_{i=1}^{k} w_i(\mathbf{x}) \underbrace{\mathbf{Q}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)}_{\mathbf{g}_i(\mathbf{x})} \quad \text{where} \quad w_i(\mathbf{x}) = \frac{\lambda_i \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i; \mathbf{x})}{\sum_{j=1}^{k} \lambda_j \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{Q}_j; \mathbf{x})} \tag{3}$$

Proposition 5 (Informal - Efficient Piecewise Polynomial Approximation - Theorem 48) *Let $\mathcal{M}$ be a τ-well-conditioned mixture of k Gaussians. There exists a function $\mathsf{c}(\cdot) : \mathbb{R}^d \mapsto [n_\mathrm{c}]$ and polynomials $p_1, \ldots, p_{n_\mathrm{c}}$ of degree at most $\ell = \mathrm{poly}(k\tau/\varepsilon)$ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{M}}[\|s(\mathbf{x}; \mathcal{M}) - p_{\mathsf{c}(\mathbf{x})}(\mathbf{x})\|^2] \le \varepsilon$, Moreover, there exists an efficient algorithm that with high-probability finds this piecewise polynomial approximation with $d^{\mathrm{poly}(\ell)}$ samples and runtime.*

Why piecewise polynomials? We first give some intuition behind the structure of the score function of a Gaussian mixture, and its piecewise polynomial approximation. We observe that the score function (see Equation (3)) is a weighted combination of linear functions. For example, for a mixture of two standard one-dimensional Gaussians with means at $-R$ and R, it behaves (approximately) like the function $-\mathbb{1}\{x \le 0\}(x + R) - \mathbb{1}\{x \ge 0\}(x - R)$, see the left figure in Figure 1. In this case, the total length of support is roughly an interval of length $O(R)$ and the slope of the score function is approximately $O(R)$ close to the origin. We would like to have a polynomial approximation of degree $\mathrm{poly}(\log R/\varepsilon)$ for this instance but naively applying polynomial approximation results (see, e.g., Jackson's theorem, Theorem 41) would yield a degree $\mathrm{poly}(R/\varepsilon)$ even for 1-dimensional mixtures.[1] Therefore, as we observe in Figure 1, two reasons prohibit us from applying polynomial approximation results in a black-box manner: (1) the total support is of radius R and (2) there are regions (far from the mixture means) where the slope of the score function is also large (also R).

For the case of two Gaussians, we see that the "effective" support is much smaller (intervals of size roughly $\sqrt{\log(1/\varepsilon)}$ around the means). Moreover, by focusing on the "effective" support we also avoid the area where the derivative of the score function is large (close to the origin). Thus one could hope to solve both issues discussed above by creating an interpolating polynomial by concentrating the nodes on the effective support. Such an approach would work when the support consisted of actual "hard" intervals (and not "approximate" intervals with Gaussian tails). The main issue is a race condition between the value of the interpolating polynomial far from the interpolation nodes (roughly exponentially large in the degree) and the decay of the Gaussian density. While this race condition can be solved in some special cases (such as for mixtures of two Gaussians with very well-separated means on $-R$ and $+R$), in general when more Gaussians are present in the mixture, the mental image of a union of "hard" intervals is incorrect and it is not clear that the tails will always be able to cancel out the large error of the polynomial far from the interpolation intervals.

The above structure of the score function naturally leads to a piecewise polynomial approximation approach. For the symmetric mixture of two Gaussians discussed above there is an obvious candidate for the partition: we should perform polynomial approximation in $\mathrm{poly}(\log(R/\varepsilon))$ sized intervals around $\pm R$ and output zero in the rest of the space. That would lead to the desired degree of $\mathrm{poly}(\log(R/\varepsilon)/\varepsilon)$. For the more complicated example of the right figure of Figure 1 we could similarly try to split the instance in an interval containing almost all the mass of two left components and one interval containing the three right components and perform polynomial approximation (and output zero out of those two intervals). In both examples, by using the piecewise polynomial approximation we avoided both issues discussed earlier, i.e., using polynomial approximation over large intervals or approximating over intervals where the derivative of the score is large.

Clustering and polynomial approximation: a win-win analysis Piecewise polynomial regression is a computationally hard, non-convex problem when we search both for the polynomials and for the partition of the space. Therefore, we have to make sure that we have an efficient algorithm to find the partition of the space and then apply polynomial regression inside each cell of the partition. Our main algorithm is enabled by a win-win argument in the sense that the areas where polynomial approximation requires high degree (i.e., $\mathrm{poly}(R)$) can be easily avoided by a crude clustering algorithm and the areas where the clustering algorithm fails to separate between a set of components of the mixture are those where the polynomial approximation is effective.

2.3. Approximating the score given a crude partition

As we observed in the previous examples, the main difficulty in providing a polynomial approximation of the score function arises when it involves multiple Gaussians that are far apart. We first make more precise the notion of "crude" clustering [2] that we require.

1. When dealing with d dimensional mixtures things are even worse since the effective support has a radius depending on the dimension d.
2. We use the terms "clustering" and "partition" function interchangeably.

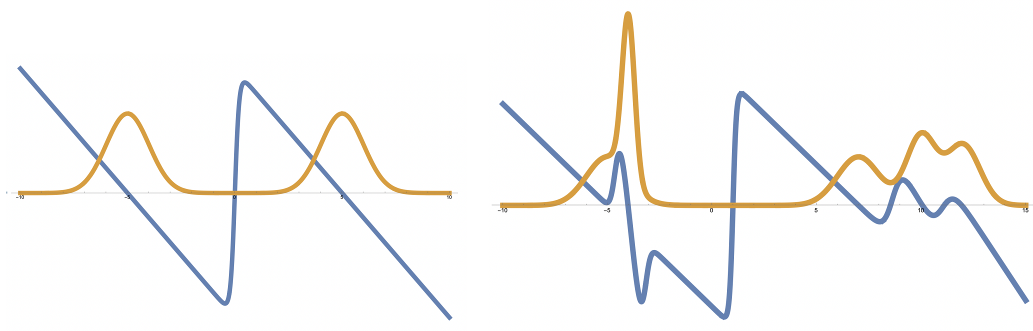

Figure 1: When approximation is hard, clustering is easy. On the left figure, we plot the density (gold) and score function (blue) of mixtures of two standard Gaussians with well-separated means (their distance is R). We observe that in that case, the score function is (almost) a piecewise linear function with a large slope, i.e., roughly R, close to the origin. In the right image, we have a mixture of 5 Gaussians with different means and variances that can be split into two clusters: a group of 2 on the left and 3 on the right. Again the area where the derivative of the score function (blue) is high, falls in between the two clusters (where the Gaussian density is exponentially small). In both cases, a piecewise polynomial approximation yields the correct degree that scaling with $(\log R)/\varepsilon$ instead of R/ε. Moreover, we expect that it is easy to cluster the points in the corresponding sub-mixtures that have much smaller effective support than the original mixture.

Definition 6 ($(\Delta_{\text{in}}, \Delta_{\text{out}})$-separated partition) *Given a mixture of Gaussians $\mathcal{N}_1 = \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{Q}_1), \ldots, \mathcal{N}_k = \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{Q}_k)$, we require that the clustering function $\mathsf{c}(\mathbf{x})$ assigns each $\mathbf{x} \in \mathbb{R}^d$ to one of n_c subsets $U_1, \ldots, U_{n_c}$ of $[k]$ that form a partition of the original k components such that:*

1. *If $\mathcal{N}_i, \mathcal{N}_j$ belong in different subsets U_t and $U_{t'}$, they have to be* at least $\Delta_{out} = \text{poly}(\tau k \log(1/\varepsilon))$ *far in parameter distance, i.e., $D_p(\mathcal{N}_i, \mathcal{N}_j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 + \|\mathbf{Q}_i - \mathbf{Q}_j\|_F \geq \Delta_{\text{out}}$.*

2. *If $\mathcal{N}_i, \mathcal{N}_j$ belong in the same subset U_t, they have to be* at most $\Delta_{\text{in}} = \text{poly}(\tau k \log(1/\varepsilon))$ *far in parameter distance, i.e., $D_p(\mathcal{N}_i, \mathcal{N}_j) \leq \Delta_{\text{in}}$.*

3. *$\mathsf{c}(\mathbf{x})$ is consistent with the partition $U_1, \ldots, U_t$ with high-probability, i.e., for any $i \in U_t$, $\mathbb{P}_{\mathbf{x} \sim \mathcal{N}_i}[\mathsf{c}(\mathbf{x}) \neq t] \leq \varepsilon_{\text{part}}$, where $\varepsilon_{\text{part}}$ is a small error parameter.*

Given the above $(\Delta_{\text{in}}, \Delta_{out})$-partition, our proof consists of two steps: (i) show that we can reduce the original problem of approximating the score function of the whole mixture to approximating the score function of the sub-mixtures U_t and (ii) providing low-degree approximations of the sub-mixture score functions. We describe these steps in the next two paragraphs.

Simplifying the score As we discussed, the first obstacle in approximating the score function is that it is a function over a domain of radius $\text{poly}(R)$ (inducing a $\text{poly}(R)$ dependency on the degree). Fortunately, there is an additional structure connecting the weights $w_i(\mathbf{x})$ and the linear terms $\mathbf{g}_i(\mathbf{x})$. We use this structure to prove that when $\mathbf{x}$ is sampled from some component $\mathcal{N}_i$ then on expectation over the component $\mathcal{N}_i$ we can remove a term in the score function corresponding to a component

$\mathcal{N}_j$ that is far from $\mathcal{N}_i$ without introducing large error, see Theorem 34. More precisely, we show that given a partition function $\mathsf{c}(\cdot)$ that satisfies Theorem 6, for all $\mathbf{x}$ where $\mathsf{c}(\mathbf{x}) = t$, we can "simplify" the score function by removing the contribution of all components $\mathcal{N}_j$ that do not belong in U_t.

Given a subset U_t of indices of $[k]$, we denote by $\mathcal{M}(U_t)$ the submixture containing the components $\mathcal{N}_i$ for $i \in U_t$ and by $\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))$ the score function containing only the contribution of components from U_t, i.e.,

$$\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) = -\sum_{i \in U_t} \mathbf{g}_i(\mathbf{x}) \frac{\lambda_i \mathcal{N}_i(\mathbf{x})}{\sum_{j \in U_t} \lambda_j \mathcal{N}_j(\mathbf{x})}$$

We prove the following proposition showing that, inside each cell t of the partition given by $\mathsf{c}(\cdot)$, we can replace the original score function $\mathbf{s}(\mathbf{x}; \mathcal{M})$ by the score function of the sub-mixture $\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))$. Each sub-mixture score function corresponding to U_t contains components that are all Δ_{in}-close to each other, thus reducing the effective radius of the approximation domain to $\text{poly}(\log R)$.

Proposition 7 (Informal – Score Simplification, see Theorem 32) *Fix $\varepsilon > 0$. let $\mathcal{M}$ be a τ-well-conditioned mixture of k Gaussian distributions and satisfies $\|\boldsymbol{\mu}_i\|_2 + \|\mathbf{Q}_i - \mathbf{Id}\|_F \le R$ for all the components. Moreover, assume that c satisfies Theorem 6. Define the following piecewise approximation to the score function $s(\mathbf{x}; \mathsf{c}(\cdot)) = \sum_{t=1}^{n_{\mathsf{c}}} s(\mathbf{x}; \mathcal{M}(U_t)) \, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}$. It holds that $\mathbb{E}_{\mathbf{x} \sim \mathcal{M}}[\|s(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}; \mathsf{c}(\cdot))\|_2^2] \le \text{poly}(k\tau R)\sqrt{\varepsilon}$.*

Polynomial approximation of the simplified score Recall from Eq. (3) that the score function for any Gaussian mixture is a sum of the softmax function $w_i(\mathbf{x})$ multiplied by a linear function $\mathbf{Q}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$. A polynomial approximation of the softmax will provide a polynomial approximation for the simplified score. Note that we want to approximate the simplified score with the degree at most $\text{poly}(k\tau/\varepsilon)$ to obtain runtime of polynomial regression of $O(d^{\text{poly}(k\tau/\varepsilon)})$.

The degree of a polynomial approximation of a function generally depends on the domain of the approximation and smoothness of the function (in terms of the norm of its gradient), see Theorem 41. The softmax function is smooth and has a bounded gradient but the input to the softmax is $\{\|\mathbf{x} - \boldsymbol{\mu}_i\|_{\mathbf{Q}_i^{-1}}^2\}_{i=1}^{|U_t|}$ which can be as large as $\text{poly}(d)$ and hence, the degree of the naive polynomial approximation could be $\text{poly}(d/\varepsilon)$.

To overcome this issue, we show that even though each input $\|\mathbf{x} - \boldsymbol{\mu}_i\|_{\mathbf{Q}_i^{-1}}^2$ is large, there exists a normalization of the softmax for which the inputs to the softmax are $\text{poly}(\tau\Delta_{\text{in}})$. More precisely, we normalize the softmax such that $\{\|\mathbf{x} - \boldsymbol{\mu}_i\|_{\mathbf{Q}_i^{-1}}^2 - \|\mathbf{x} - \boldsymbol{\mu}_1\|_{\mathbf{Q}_1^{-1}}^2 - \langle \mathbf{Q}_1, \mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1} \rangle\}_{i=1}^{|U_t|}$ are the inputs to the softmax function and show that its norm is $\text{poly}(\tau\Delta_{\text{in}})$ with high probability. Therefore, using multivariate Jackson's theorem (Theorem 41), we obtain the polynomial approximation for the softmax function and hence, for the simplified score function.

Lemma 8 (Informal - See Theorem 45) *Let $\mathcal{M}(U)$ be a τ-well-conditioned mixture of k Gaussian distributions restricted to the subset of components in U. Then, there exist a polynomial $p(\mathbf{x}; \mathcal{M}(U))$ of degree $\text{poly}(\tau\Delta_{\text{in}}/\varepsilon)$ and coefficients bounded in magnitude by $dR\exp(\text{poly}(\tau\Delta_{\text{in}}/\varepsilon))$ such that for $\mathbf{x} \sim \mathcal{M}(U)$, with high probability, the polynomial satisfies $\|s(\mathbf{x}; \mathcal{M}(U)) - p(\mathbf{x}; \mathcal{M}(U))\| \le \varepsilon$.*

2.4. Crude clustering via PCA

We now describe our crude clustering algorithm for obtaining the partition satisfying the assumptions of Theorem 6. Our approach consists of two main steps: (1) approximately recover the span of the means and covariances using PCA on the second and fourth-order moment tensors of the mixture and (2) recover estimates of the parameters by brute forcing over the k-dimensional subspace recovered in the first step and using pairwise log-likelihood tests to create the final partition function.

Obtaining estimates of means and covariances The algorithm operates in two phases. First, we obtain a crude estimate for the subspace spanned by the means, after which we brute-force within this low-dimensional subspace to find points close to each of the means. Second, we use these mean estimates to form an estimator for the subspace spanned by the covariances, after which we can similarly brute-force to find points close to each of the covariances. With roughly $d^{O(k)}$ runtime, we can construct a list of candidate parameters for the means and covariances of the mixture containing crude (in the sense that they can be $\mathrm{poly}(k\tau)$-far) of the target parameters.

Lemma 9 (Informal – Recovering crude estimates of the parameters, see Theorem 13) *There is an algorithm that returns a list $\mathcal{W}$ such that for every $i \in [k]$, there exists $(\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i) \in \mathcal{W}$ for which $\|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i\|^2 \lesssim \beta/\lambda_{\min}$ and $\|\mathbf{Q}_i - \widehat{\mathbf{Q}}_i\|_F \lesssim k^{3/2}\beta/\lambda_{\min} + k^2\alpha \log R$. Furthermore, $|\mathcal{W}| \leq (R/\sqrt{\beta})^{O(k^2)} \cdot d^{O(k)}$, and the algorithm runs in time $(R/\sqrt{\beta})^{O(k^2)} \cdot (\mathrm{poly}(dR/\beta) + d^{O(k)})$ and draws $\mathrm{poly}(dR/\beta)$ samples.*

We use PCA on the covariance $\mathbf{M} = \mathbb{E}_{\mathbf{x}\sim\mathcal{M}}[\mathbf{x}\mathbf{x}^\intercal]$ to obtain the subspace spanned by the means. We observe that $\mathbf{M} = \sum_{i=1}^k \lambda_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\intercal + \sum_{i=1}^k \lambda_i \mathbf{Q}_i$. The main idea here is to think of $\mathbf{M}$ as approximately low-rank and treat the contribution of the covariances as an error $\mathcal{E} = \sum_{i=1}^k \lambda_i \mathbf{Q}_i$. Since the covariances $\mathbf{Q}_i$ are well-conditioned (i.e., their eigenvalues are not bigger than β – see Theorem 1) we can show that if some $\boldsymbol{\mu}_i$ is larger than $\beta/\lambda_{\min}$ then its contribution in $\mathbf{M}$ cannot be "hidden" by the error term $\mathcal{E}$ and will have a large projection onto the subspace spanned by the top eigenvectors of $\mathbf{M}$. The proof of this is standard and can be found in Appendix B.1.

Finding estimates for the covariances is more complicated but similarly relies on recovering the subspace spanned by the low-rank components of the (flattened) fourth-order tensor

$$\boldsymbol{\Psi} = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}[\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)] \,.$$

The intuition is that if the means of the mixture were all sufficiently close to zero, then the top-k singular subspace of the matrix $\boldsymbol{\Psi}$ can be shown to contain points close to $\mathrm{vec}(\mathbf{Q}_1), \ldots, \mathrm{vec}(\mathbf{Q}_k)$. In general, if the means are arbitrary, then we can use the estimates $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$ from the previous section to approximately "recenter" the components near zero. Since the means recovered in the previous step were already crude $\mathrm{poly}(k)$ approximations of the true means a careful error analysis must be done so that this recentering does not introduce significantly more (i.e. dimension-dependent) error in the covariance estimates. We refer to Appendix B and Algorithm 3 for more details.

Clustering using the log-likelihood ratios We now present our main clustering guarantee, which leverages the estimates for the parameters we obtained previously. As those estimates are only crude approximations to the true parameters, we will obtain a commensurately crude clustering.

Our algorithm starts by brute-forcing over mean-based and covariance-based partitions $\mathcal{S}$ (resp. $\mathcal{T}$). $\mathcal{S}$ (resp. $\mathcal{T}$) partitions the mixture components into groups such that any two components in the same group have means (resp. covariances) that are not far, and any two components from two different groups have means (resp. covariances) that are not close. Their common refinement is a partition $\mathcal{U}$ satisfying the assumptions of Theorem 6: any two components in the same group have both means and covariances not too far, and any two components from two different groups either have means not too close or covariances not too close.

By brute-forcing over pairs of partitions of $[k]$ (of which there are at most k^{2k}) we may assume we have access to $\mathcal{S}$ and $\mathcal{T}$, and thus to $\mathcal{U}$. Our goal is then to assign to every $\mathbf{x} \in \mathbb{R}^d$ an index into the partition $\mathcal{U}$. For $\mathbf{x}$ which is sampled from the i-th component of the mixture which belongs to the t-th group in $\mathcal{U}$, we would like our assignment to be t with high probability. At a high level, the idea is as follows. It is not too hard to determine which group in $\mathcal{S}$ a given point $\mathbf{x}$ should belong to, simply by checking which mean estimate $\widehat{\boldsymbol{\mu}}_i$ is closest to $\mathbf{x}$ after projecting to the subspace spanned by $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$. For each group in $\mathcal{S}$, we can then effectively restrict our attention to components within that group and focus on clustering them according to their covariances. Roughly speaking, we accomplish this by comparing log-likelihoods of sampling $\mathbf{x}$ under $\mathcal{N}(\widehat{\boldsymbol{\mu}}_1, \widehat{\mathbf{Q}}_1), \ldots, \mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{Q}}_k)$ and choosing the group in $\mathcal{T}$ containing the component maximizing log-likelihood. For more details, we refer to Appendix C and to Theorem 25 for the formal clustering statement that we prove.

3. Discussion on qualitative aspects of our bounds

3.1. Avoiding the doubly exponential dependency on k

Here we provide some intuition for the origin of the doubly exponential dependence on k which is implicit in existing works on learning mixtures of general Gaussians with the method of moments, and how our technique outlined above avoids this issue. Our starting point is the algorithm of Moitra and Valiant (2010); in fact, for this discussion, it will suffice to consider the case of $d = 1$ and components of variance 1.

Specialized to this case, in the analysis in Moitra and Valiant (2010), the authors first proved that if all of the components have means with nonnegligible separation, say η, from each other, then one can learn the means by brute-forcing over a grid with sufficiently small granularity and finding a setting of parameters in this grid for which the corresponding mixture matches the first $O(k)$ moments with the target mixture to error η^k (here we ignore constants in the exponent for simplicity).

Now what happens if the minimal separation η is arbitrarily small? The authors noted that for means that are particularly close, one can simply "merge them": they are statistically close to a single component, and in a bounded number of samples one would not be able to tell the difference. Because the number of samples used by the algorithm outlined above is $(1/\eta)^k$, this implies that if there is some scale η at which there is a *gap* in the sense that all means are either η^k-close or η-far apart, then one can learn in the same amount of time/samples as in the η-separated case.

The last question that remains is how to ensure such a scale exists. The idea is that if one looks at $k^2 + 1$ consecutive windows $\{[\eta^{k^i}, \eta^{k^{i-1}}]\}_{i=1,\ldots,k^2+1}$, by pigeonhole principle there must exist some window such that the separation between any pair of means lies outside this window. At that scale, one can apply the above reasoning to learn the means. This is the origin of the doubly exponential scaling in k that is present in all existing algorithms for learning mixtures of general Gaussians, including the state-of-the-art guarantee of Bakshi et al. (2022).

It is instructive to contrast this with our approach. The main reason for the doubly exponential dependence in the above windowing argument was that one needed a scale at which the components break up into "gapped clusters" such that the separation within clusters is significantly smaller than the separation across clusters. For this clustering structure to exist, we need to go down potentially to a doubly exponentially small scale. In contrast, in our work, we make do with a very crude clustering for the purposes of our piecewise regression. We simply require that for components from different clusters, their parameter distance is sufficiently large, while for components from the same cluster, their parameter distance is *not too large*. Crucially, we don't need to make any assumption about a gap between the intra- versus inter-cluster separations, ensuring we avoid the doubly exponential dependence on k.

3.2. Further discussion on the condition number dependence

Recall the parallel pancakes instance of Diakonikolas et al. (2017) discussed at the end of Section 1 consists of components with covariance $\mathbf{Id} - (1-\eta)\mathbf{v}\mathbf{v}^\intercal$ for parameter $\eta = 1/\mathrm{poly}(k)$. The expert reader might note that the $d^{\Omega(k)}$ statistical query lower bound of Diakonikolas et al. (2017) still persists for any $\exp(-\Theta(k)) \le \eta \ll 1/k$. Note however that even in the highly ill-conditioned regime where $\delta = \exp(-\Theta(k))$, the exponent in our runtime is polynomial in $1/\delta$ so that our runtime is doubly exponential in k, meaning that we still qualitatively match the best existing guarantees based on the moment of moments Bakshi et al. (2022); Moitra and Valiant (2010) even in this challenging regime. Furthermore, for any $\delta = \exp(-o(k))$, our runtime is better. In other words, while the advantage of our algorithmic guarantee over prior guarantees is most pronounced when the covariances are well-conditioned, we achieve a meaningful improvement even in highly ill-conditioned settings.

Acknowledgments

We thank Adam Klivans for many illuminating discussions about score estimation, polynomial regression, and diffusion models throughout the preparation of this work. We also thank the authors of Gatmiry et al. (2024) for coordinating the submission of manuscripts with us. VK and KS are supported by the NSF AI Institute for Foundations of Machine Learning (IFML).

References

Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from mean-field gibbs measures via diffusion processes. *arXiv preprint arXiv:2310.08912*, 2023.

Nima Anari, Yizhi Huang, Tianyu Liu, Thuy-Duong Vuong, Brian Xu, and Katherine Yu. Parallel discrete sampling via continuous walks. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 103–116, 2023.

Nima Anari, Sinho Chewi, and Thuy-Duong Vuong. Fast parallel sampling under isoperimetry. *arXiv preprint arXiv:2401.09016*, 2024.

Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. *Advances in Neural Information Processing Systems*, 31, 2018.

Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *arXiv preprint arXiv:2005.02970*, 2020.

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

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

Shalev Ben-David, Adam Bouland, Ankit Garg, and Robin Kothari. Classical lower bounds from quantum upper bounds. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 339–349. IEEE, 2018.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023a.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

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

S. C. Brubaker and S. Vempala. Isotropic PCA and Affine-Invariant Clustering. In *Proc. 49th IEEE Symposium on Foundations of Computer Science*, pages 551–560, 2008a.

S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. *Building Bridges: Between Mathematics and Computer Science*, pages 241–281, 2008b.

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

Rares-Darius Buhai and David Steurer. Beyond parallel pancakes: Quasi-polynomial time guarantees for non-spherical gaussian mixtures. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 548–611. PMLR, 2023.

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

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

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

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023a.

Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–600, 2020.

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

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023c. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

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

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. Analysis of learning a flow-based generative model from limited sample complexity. *arXiv preprint arXiv:2310.03575*, 2023.

S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40^{th} Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999a.

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

C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27^{th} Conference on Learning Theory, COLT 2014*, pages 1183–1213, 2014.

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

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, pages 655–664, 2016.

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

Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 11–20, 2010. doi: 10.1109/FOCS.2010.8.

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

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

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

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

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

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

Jon Feldman, Ryan O'Donnell, and Rocco A Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

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

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1162–1173. IEEE, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

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

Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X Wu. Weak poincar\'e inequalities, simulated annealing, and sampling from spherical spin glasses. *arXiv preprint arXiv:2411.09075*, 2024a.

Brice Huang, Andrea Montanari, and Huy Tuan Pham. Sampling from spherical spin glasses in total variation via algorithmic stochastic localization. *arXiv preprint arXiv:2404.15651*, 2024b.

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

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

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT)*, pages 444–457, 2005.

Frederic Koehler and Thuy-Duong Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. *arXiv preprint arXiv:2310.01762*, 2023.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *International Conference on Learning Representations*, 2023.

Frederic Koehler, Holden Lee, and Thuy-Duong Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. *arXiv preprint arXiv:2411.09117*, 2024.

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

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

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

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

Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, pages 1302–1382. PMLR, 2017.

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

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

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

Song Mei and Yuchen Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*, 2023.

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

Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

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

D. J. Newman and H. S. Shapiro. *Jackson's Theorem in Higher Dimensions*, pages 208–219. Springer Basel, Basel, 1964. ISBN 978-3-0348-4131-3. doi: 10.1007/978-3-0348-4131-3_20. URL https://doi.org/10.1007/978-3-0348-4131-3_20.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

Chirag Pabbaraju, Dhruv Rohatgi, Anish Prasad Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

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

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the DDPM objective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aig7sgdRfI.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

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

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

A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1395–1403, 2014.

S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 113–122, 2002.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

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

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*, 2024.

Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning gaussian mixtures using the wasserstein-fisher-rao gradient flow. *arXiv preprint arXiv:2301.01766*, 2023.

Appendix A. Diffusion models and other technical preliminaries

In this section, we collect various technical ingredients. The bulk of this section is dedicated to an exposition of diffusion models in Appendix A.2.

A.1. Notation for mixture models

Throughout the paper, we use either q or q_0 to denote the data distribution on $\mathbb{R}^d$, i.e., the mixture of Gaussians with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$, covariances $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_k$, and mixing weights $\lambda_1, \ldots, \lambda_k$ respectively. We will use $\mathcal{N}_i$ to denote the distribution for its i-th component, i.e. $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i)$. We use p or p_T to denote the learned distribution.

Definition 10 *Let $\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k} \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i)$ be a (α, β, R)-well-conditioned Gaussian mixture. We say that a partition of $[k]$ into subsets $S_1, \ldots, S_m$ is $(\Delta_{\text{in}}, \Delta_{\text{out}})$-separated if for all $i, j \in S_\ell$ it holds that $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| + \|\mathbf{Q}_i - \mathbf{Q}_j\|_F \leq \Delta_{\text{in}}$ and for all $i \in S_\ell$, $j \in S_{\ell'}$ for $\ell \neq \ell'$ it holds $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| + \|\mathbf{Q}_i - \mathbf{Q}_j\|_F \geq \Delta_{\text{out}}$. We denote by $\mathcal{M}(S_i)$ the mixture distribution corresponding to the components of S_i, i.e., $\mathcal{M}(S_i) = \frac{1}{|S_i|}\sum_{j \in S_i} \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{Q}_j)$.*

Moreover, given a mixture $\mathcal{M} = \sum_{i=1}^{k} \lambda_i \mathcal{D}_i$ we denote by $\mathcal{M}^J$ the joint distribution over tuples $(j, \mathbf{x})$ where $j = i$ with probability λ_i and, conditional on $j = i$, $\mathbf{x}$ is drawn from $\mathcal{D}_i$.

A.2. Learning Gaussian mixtures via a denoising diffusion process

We start by introducing some standard terminology and notation on diffusion models. We will be using the diffusion algorithmic template in a more or less black box manner and therefore we try to keep the presentation short but still self-contained. Throughout the paper, we use either q or q_0 to denote the data distribution on $\mathbb{R}^d$. The two main components in diffusion models are the *forward process* and the *reverse process*. The forward process transforms samples from the data distribution into noise, for instance via the *Ornstein-Uhlenbeck (OU) process*:

$$\mathrm{d}\mathbf{x}_t = -\mathbf{x}_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{w}_t \quad \text{with} \quad \mathbf{x}_0 \sim q_0\,,$$

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

$$\mathbf{x}_t = \exp(-t)\mathbf{x}_0 + \sqrt{1-\exp(-2t)}\mathbf{z}_t \quad \text{with} \;\; \mathbf{x}_0 \sim q_0, \;\; \mathbf{z}_t \sim \mathcal{N}(0, \mathbf{Id})\,. \tag{4}$$

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

$$\mathrm{d}\mathbf{x}_t^{\leftarrow} = \{\mathbf{x}_t^{\leftarrow} + 2\nabla_{\mathbf{x}} \ln q_{T-t}(\mathbf{x}_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{w}_t \quad \text{with} \quad \mathbf{x}_0^{\leftarrow} \sim q_T\,, \tag{5}$$

where now $\mathbf{w}_t$ is the reversed Brownian motion. In this reverse process, the iterate $\mathbf{w}_t^{\leftarrow}$ is distributed according to q_{T-t} for every $t \in [0, T]$, so that the final iterate $\mathbf{x}_T^{\leftarrow}$ is distributed according to the data distribution q_0. The function $\nabla_{\mathbf{x}} \ln q_t$ is called the *score function* and is required so that we are able to run the reverse SDE and generate samples from the unknown distribution. Ideally, we would like to have access to an approximate oracle $\widehat{\mathbf{s}}(\mathbf{x})$ such that for all $t \in [0, T]$ it is a good approximation to the score function $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$:

$$\underset{\mathbf{x}_t \sim q_t}{\mathbb{E}}[\|\nabla_{\mathbf{x}} \ln q_t(\mathbf{x}_t) - \widehat{\mathbf{s}}_t(\mathbf{x}_t)\|^2] \le \epsilon_{\text{score}} \,. \tag{6}$$

To obtain such a function $\widehat{\mathbf{s}}_t(\mathbf{x})$, one would an expressive enough set of candidate functions $\mathcal{G}$ and then try to optimize the score matching loss:

$$\min_{\mathbf{g}_t \in \mathcal{G}} \underset{\mathbf{x}_t \sim q_t}{\mathbb{E}}[\|\nabla_{\mathbf{x}} \ln q_t(\mathbf{x}_t) - \mathbf{g}_t(\mathbf{x})\|^2] \le \epsilon_{\text{score}} \,.$$

However, as the density function of q_t is unknown the above minimization problem cannot be solved directly. A standard calculation (see e.g. Appendix A of Chen et al. (2023c)) shows that this is equivalent to minimizing the *DDPM objective* in which one wants to predict the noise $\mathbf{z}_t$ from the noisy observation $\mathbf{x}_t$, i.e.

$$\min_{\mathbf{g} \in \mathcal{G}} \quad L_t(\mathbf{g}_t) = \underset{\mathbf{x}_0, \mathbf{z}_t}{\mathbb{E}}\left[\left\|\mathbf{g}_t(\mathbf{x}_t) + \frac{\mathbf{z}_t}{\sqrt{1 - \exp(-2t)}}\right\|^2\right] \,. \tag{7}$$

In this work we focus specifically on the optimization problem (7) and show that it can be solved efficiently when the underlying target density q_0 is a mixture of k Gaussian distributions.

Algorithm 1: GENERATESAMPLE

Input: Score estimation error ϵ_{score}, confidence δ, sequence of time steps $t_0, t_1, \ldots, t_N$
Output: A sample $\mathbf{y}_N \in \mathbb{R}^d$
for $\ell \in \{0, \ldots, N-1\}$ **do**
  $\widehat{\mathbf{s}}(\cdot, T - t_\ell) \leftarrow$ LEARNSCORE$(t_\ell, \epsilon_{\text{score}}, \delta)$ ▷ Learn the score function at all time steps
end
for $\ell \in \{0, \ldots, N-1\}$ **do**
  Set $\rho_\ell = e^{(t_{\ell+1} - t_\ell)/2}$
  Sample $\mathbf{z}_\ell \sim \mathcal{N}(\mathbf{0}, \mathbf{Id})$
  $\mathbf{y}_{\ell+1} \leftarrow \rho_\ell\, \mathbf{y}_\ell + 2(\rho_\ell - 1)\, \widehat{\mathbf{s}}(\mathbf{y}_\ell, T - t_\ell) + \sqrt{\rho_\ell^2 - 1}\, \mathbf{z}_\ell$ ▷ Run the (discretized) reverse SDE
end
return $\mathbf{y}_N$

We are now ready to present and prove our main result: an efficient algorithm for learning well-conditioned GMMs.

Theorem 11 (Efficient Sampler for GMMs) *Fix $\varepsilon, \delta_f \in (0, 1)$ and let $\mathcal{M}$ be an (α, β, R)-well-conditioned mixture of k Gaussians. Let $\tau = (\beta/\alpha) \log R$, $\varepsilon_{score} = \varepsilon / \log(R/(\alpha\varepsilon))$, $\delta_{stop} = \alpha\varepsilon/R$, and let time sequence $t_1, \ldots, t_N$ be as defined in Theorem 12. Then with probability at least $1 - \delta_f$, Algorithm 1 draws $M = d^{\mathrm{poly}(k\tau/(\lambda_{min}\varepsilon))} \log \frac{1}{\delta_f}$ samples from $\mathcal{M}$, runs in sample-polynomial time, and generates a sample $\mathbf{y}_N$ whose distribution is ε-close in total variation to $\mathcal{M}$.*

Proof

We are going to use the following result on the convergence of the discretized reverse SDE with the score approximation that we use in Algorithm 1.

Lemma 12 (Convergence given approximate scores, (Benton et al., 2023a)) *Fix some $\delta_{stop} \in (0,1)$, $T \geq 1$ and let N be some even integer larger than $\log(1/\delta_{stop})$ and let $\kappa > 0$ be larger than a sufficiently large constant multiple of $(T + \log(1/\delta_{stop}))/N$. Set $t_0 = 0$, $t_{N/2} = T - 1$, $t_N = T - \delta_{stop}$. Moreover, set $t_1, \dots, t_{N/2-1}$ equally spaced on $[0, T-1]$, i.e., $t_{\ell+1} - t_\ell = \kappa > 0$ for all $\ell \in \{0, \dots, N/2-1\}$ and $T - t_{N/2+1}, \dots T - t_{N-1}$ exponentially decaying, i.e., $t_{N/2+\ell+1} - t_{N/2+\ell} = \kappa/(1+\kappa)^\ell$ for all $\ell \in \{0, \dots, N/2-2\}$ and $\gamma_\ell \leq \kappa \min(1, T - t_\ell)$. Assume that the data distribution and the score function satisfy the following assumptions.*

1. $\sum_{\ell=0}^{N-1} \gamma_\ell \, \mathbb{E}_{\mathbf{x} \sim q_{t_\ell}}[\|\nabla \log q_{T-t_\ell}(\mathbf{x}) - \widehat{\mathbf{s}}(\mathbf{x}, T - t_\ell)\|_2^2] \leq \epsilon_{\text{score}}^2$.
2. *The target distribution q_0 on $\mathbb{R}^d$ has finite second moment.*

For any $t \in [0,T]$ denote by q_t the distribution of $\exp(-t)\mathbf{x}_0 + \sqrt{1 - \exp(-2t)}\mathbf{z}_t$, where $\mathbf{x}_0 \sim q_0$ and $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Id})$ and denote by p_{t_N} the distribution of the output $\mathbf{y}_N$ of Algorithm 1. It holds that

$$\mathrm{KL}(q_{\delta_{stop}} \| p_{t_N}) \lesssim \epsilon_{\text{score}}^2 + \kappa^2 dN + \kappa dT + \mathrm{KL}(q_T \| \mathcal{N}(\mathbf{0}, \mathbf{Id}))\,.$$

We first show that the guarantee of Theorem 12 yields a total variation bound between p_{t_N} and the target Gaussian mixture $\mathcal{M}$. By Pinsker's inequality, we obtain that $\mathrm{TV}(p_{t_N}, q_{\delta_{\text{stop}}}) \lesssim \sqrt{\mathrm{KL}(q_{\delta_{\text{stop}}} \| p_{t_N})}$. Moreover, by a triangle inequality, we obtain that $\mathrm{TV}(q_0, p_{t_N}) \leq \mathrm{TV}(p_{t_N}, q_{\delta_{\text{stop}}}) + \mathrm{TV}(q_{\delta_{\text{stop}}}, q_0)$. Therefore, we have to control $\mathrm{TV}(q_0, q_{\delta_{\text{stop}}})$. Using again Pinsker's inequality we obtain that $\mathrm{TV}(q_0, q_{\delta_{\text{stop}}}) \lesssim \sqrt{\mathrm{KL}(q_0 \| q_{\delta_{\text{stop}}})}$. To control the Kullback-Leibler divergence between the target q_0 that corresponds to the well-conditioned mixture and $q_{\delta_{\text{stop}}}$. We observe that $q_{\delta_{\text{stop}}}$ is also a Gaussian mixture with parameters $\widehat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i \exp(-\delta_{\text{stop}})$ and $\widehat{\mathbf{Q}} = \mathbf{Q}_i e^{-2\delta_{\text{stop}}} + (1 - e^{-2\delta_{\text{stop}}})\, \mathbf{Id}$. We denote this mixture by $\mathcal{M}_{\delta_{\text{stop}}}$ Since KL is convex we obtain that

$$\mathrm{KL}(q_{\delta_{\text{stop}}} \| q_0) \leq \sum_{i=1}^k \lambda_i \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i) \| \mathcal{N}(\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i)) \leq \max_{i=1}^k \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i) \| \mathcal{N}(\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i))\,.$$

We can now use the following standard bound for the Kullback-Leibler distance between two Normal distributions $\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{Q}_1) \| \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{Q}_2)) \lesssim \|\mathbf{Id} - \mathbf{Q}_2^{-1/2} \mathbf{Q}_1 \mathbf{Q}_2^{-1/2}\|_F^2 + \|\mathbf{Q}_2^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_2^2$. We have that

$$\|\mathbf{Q}_i^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i e^{-\delta_{\text{stop}}})\|_2^2 \leq \frac{R^2}{\alpha}(1 - e^{-\delta_{\text{stop}}})^2 \leq \frac{R^2}{\alpha}\, \delta_{\text{stop}}^2\,,$$

where the last inequality follows by the fact that $\alpha \mathbf{Id} \preceq \mathbf{Q}_i$ and the fact that $\|\boldsymbol{\mu}_i\|_2 \leq R$ and the inequality $e^x \geq x + 1$. Moreover, if $s_1, \dots s_d$ are the eigenvalues of $\mathbf{Q}_1$, we have that

$$\|\mathbf{Id} - \mathbf{Q}_2^{-1/2} \mathbf{Q}_1 \mathbf{Q}_2^{-1/2}\|_F^2 = \sum_{i=1}^d \Big(1 - \frac{s_i e^{-2\delta_{\text{stop}}} + (1 - e^{-2\delta_{\text{stop}}})}{s_i}\Big)^2 = (1 - e^{-2\delta_{\text{stop}}})^2 \sum_{i=1}^d \Big(\frac{1 - s_i}{s_i}\Big)^2 \lesssim \frac{\delta_{\text{stop}}^2}{\alpha^2} R^2\,,$$

where the last inequality follows by the assumption that $\|\mathbf{Q}_i - \mathbf{Id}\|_F^2 \lesssim R$ and the fact that $s_i \geq \alpha$ for all i. Putting the above together, we obtain that $\mathrm{TV}(p_{t_N}, \mathcal{M}) \lesssim \sqrt{\mathrm{KL}(q_{\delta_{\text{stop}}} \| p_{t_N})} + \delta_{\text{stop}} R/\alpha$.

Similarly, we have to control the convergence error of the forward OU process $\mathrm{KL}(q_T \| \mathcal{N}(\mathbf{0}, \mathbf{Id}))$. Similarly to the above argument, by the convexity of the Kullback-Leibler, we obtain that it suffices to control the KL divergence between any component of the mixture and the standard normal $\mathcal{N}(\mathbf{0}, \mathbf{Id})$. Using the same bound for the KL divergence as above, we have that

$$\mathrm{KL}(q_T \| \mathcal{N}(\mathbf{0}, \mathbf{Id}) \le \max_{i=1}^{k} (e^{-2T} \|\boldsymbol{\mu}_i\|_2^2 + e^{-4T} \|\mathbf{Q}_i - \mathbf{Id}\|_F^2) \lesssim e^{-2T} R^2 \,.$$

To make the forward process converge to an ε-approximate Gaussian, we take $T = \log(R/\varepsilon)$. We choose $\varepsilon_{\text{score}} = \varepsilon / \log(R/(\alpha\varepsilon))$ and $\delta_{\text{stop}} = \alpha\varepsilon/R$. Additionally, we have $\gamma_\ell \le \kappa$ for all ℓ. Therefore, we have

$$\sum_{\ell=0}^{N-1} \gamma_\ell \mathop{\mathbb{E}}_{\mathbf{x} \sim q_{t_\ell}} [\|\nabla \log q_{T-t_\ell}(\mathbf{x}) - \widehat{\mathbf{s}}(\mathbf{x}, T - t_\ell)\|_2^2] \le \varepsilon \,.$$

The above choice also yields $\kappa^2 dN \lesssim (\log^2(R/\alpha\varepsilon)d)/N$ and $\kappa dT \lesssim (\log^2(R/\alpha\varepsilon)d)/N$. Choosing $N = (\log^2(R/\alpha\varepsilon)d)/\varepsilon$ and combining all the terms in Theorem 12, we obtain that $\mathrm{TV}(p_{t_N}, \mathcal{M}) \le \varepsilon$. We obtain sample complexity and runtime of the algorithm by putting $\varepsilon_{\text{score}} = \varepsilon / \log(R/(\alpha\varepsilon))$ and failure probability $\delta = \delta_f / N$ in Theorem 49. ■

Appendix B. Obtaining crude estimates for the parameters

In this section, we prove the next lemma showing that we can construct a list of candidates for the unknown parameters of the mixture, containing "crude" approximation to the true target parameters.

Lemma 13 *There is an algorithm* CrudeEstimate*(q) which returns a list $\mathcal{W}$ such that for every $i \in [k]$, there exists $(\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i) \in \mathcal{W}$ for which $\|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i\|^2 \lesssim \beta/\lambda_{\min}$ and $\|\mathbf{Q}_i - \widehat{\mathbf{Q}}_i\|_F \lesssim k^{3/2}\beta/\lambda_{\min} + k^2\alpha \log R$. Furthermore, $|\mathcal{W}| \le (R/\sqrt{\beta})^{O(k^2)} \cdot d^{O(k)}$, and the algorithm runs in time $(R/\sqrt{\beta})^{O(k^2)} \cdot (\mathrm{poly}(d, 1/\beta) + d^{O(k)})$ and draws $\mathrm{poly}(dR/\beta)$ samples.*

The algorithm operates in two phases. First, we obtain a crude estimate for the subspace spanned by the means, after which we brute-force within this subspace to find points close to each of the means. Second, we use these mean estimates to form an estimator for the subspace spanned by the covariances, after which we can similarly brute-force to find points close to each of the covariances.

B.1. Estimating the means

This phase is straightforward: we simply take the top-k singular subspace of the empirical second moment matrix (see Algorithm 2 below).

Lemma 14 *There is an algorithm* CrudeEstimateMeans*(q) which returns a list $\mathcal{W}$ such that for each $i \in [k]$, there exists $\widehat{\boldsymbol{\mu}}_i \in \mathcal{W}$ for which $\|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i\|^2 \lesssim \beta/\lambda_{\min}$. Furthermore, $|\mathcal{W}| \le (R/\sqrt{\beta})^{O(k)}$, and the algorithm runs in time $\mathrm{poly}(dR/\beta) + (R/\sqrt{\beta})^{O(k)}$ and draws $\mathrm{poly}(dR/\beta)$ samples.*

The analysis (as well as subsequent parts of our proof) uses the following standard bound for k-SVD:

Lemma 15 *Let $\mathbf{A} = \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\intercal + \mathbf{E}$ for $\|\mathbf{E}\|_{\mathrm{op}} \le \epsilon$. The top-$k$ singular subspace of $\mathbf{A}$ contains vectors $\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_k$ for which $\|\mathbf{v}_i - \widehat{\mathbf{v}}_i\|^2 \le 2\epsilon$ for all $i \in [k]$.*

Algorithm 2: CRUDEESTIMATEMEANS(q)

Input: Sample access to q
Output: List $\mathcal{W}$ containing approximations to $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$
Draw samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from q for $N \leftarrow \mathrm{poly}(dR/\beta)$
$\widehat{\mathbf{M}} \leftarrow \frac{1}{N}\sum_i \mathbf{x}_i\mathbf{x}_i^\intercal$
$\widehat{V} \leftarrow$ top-k singular subspace of $\widehat{\mathbf{M}}$
$\mathcal{W} \leftarrow$ a $\beta^{1/2}$-net over vectors in $\widehat{V}$ with L_2 norm at most $2R$
return $\mathcal{W}$

Proof Define $\mathbf{A}^* \triangleq \sum_{i=1}^k \mathbf{v}_i\mathbf{v}_i^\intercal$. Let $\Pi^\perp$ denote the projector to the orthogonal complement of the top-k singular subspace of $\mathbf{A}$, and define $\mathbf{r}_i \triangleq \Pi^\perp \mathbf{v}_i$. Then

$$\mathbf{r}_i^\intercal \mathbf{A}^* \mathbf{r}_i = \sigma_{k+1}(\mathbf{A}) \cdot \|\mathbf{r}_i\|^2 - \mathbf{r}_i^\intercal \mathbf{E} \mathbf{r}_i \le 2\epsilon \|\mathbf{r}_i\|^2 \,,$$

where in the last step we used Weyl's inequality to bound $\sigma_{k+1}(\mathbf{A})$.

On the other hand,

$$\mathbf{r}_i^\intercal \mathbf{A}^* \mathbf{r}_i = \sum_j \langle \mathbf{r}_i, \mathbf{v}_j\rangle^2 \ge \langle \mathbf{r}_i, \mathbf{v}_i\rangle^2 = \|\mathbf{r}_i\|^4 \,,$$

so we conclude that $\|\mathbf{r}_i\|^2 \le 2\epsilon$. If we define $\widehat{\mathbf{v}}_i = \Pi \mathbf{v}_i$, where Π is the projector to the top-k singular subspace of $\mathbf{A}$, then $\|\widehat{\mathbf{v}}_i - \mathbf{v}_i\|^2 = \|\mathbf{r}_i\|^2 \le 2\epsilon$ as claimed. ■

We now show that the empirical second moment matrix can be used to extract a rough approximation to the span of the means:

Lemma 16 *For $\mathbf{x} \sim \mathcal{M}$, let $\mathbf{M} \triangleq \mathbb{E}[\mathbf{x}\mathbf{x}^\intercal]$. Given $\widehat{\mathbf{M}}$ for which $\|\mathbf{M} - \widehat{\mathbf{M}}\|_{\mathrm{op}} \lesssim \beta$, let $\widehat{V}$ denote the top-k singular subspace of $\widehat{\mathbf{M}}$. Then for every $i \in [k]$, there exists $\widehat{\boldsymbol{\mu}}_i \in \widehat{V}$ for which $\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 \lesssim \beta/\lambda_{\min}$.*

Proof Define $\mathcal{E} = \sum_i \lambda_i \mathbf{Q}_i$ and $\mathbf{M}^* \triangleq \sum_i \lambda_i \mu_i \mu_i^\intercal$. We have that

$$\mathbf{M} = \mathbf{M}^* + \mathcal{E} \,,$$

and $\|\mathcal{E}\|_{\mathrm{op}} \lesssim \beta$.

By Theorem 15, where we take $\mathbf{A}$ and $\mathbf{E}$ therein to be $\widehat{\mathbf{M}}$ and $\mathcal{E} + \widehat{\mathbf{M}} - \mathbf{M}$, we find that $\widehat{V}$ contains vectors $\boldsymbol{\mu}'_1, \ldots, \boldsymbol{\mu}'_k$ for which $\|\boldsymbol{\mu}'_i - \sqrt{\lambda_i}\boldsymbol{\mu}_i\|^2 \lesssim \beta$. So if we take $\widehat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}'_i/\sqrt{\lambda_i}$, the claimed bound follows. ■

Proof [Proof of Theorem 14] By standard matrix concentration (see, e.g., Vershynin (2018)) with $N = \mathrm{poly}(dR/\beta)$ samples (as set in Algorithm 2) we have that the matrix $\widehat{\mathbf{M}}$ constructed therein satisfies $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{op}} \le \beta$, where $\mathbf{M} \triangleq \mathbb{E}_{\mathcal{M}}[\mathbf{x}\mathbf{x}^\intercal] = \sum_i \lambda_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\intercal + \sum_i \lambda_i \mathbf{Q}_i$. We have $\|\widehat{\mathbf{M}} - \sum_i \lambda_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\intercal\|_{op} \le 2\beta$, so by Theorem 16, the β-net constructed in Algorithm 2 contains points which are $O(\beta/\lambda_{\min})$-close to each of the means $\boldsymbol{\mu}_i$ as claimed. ■

B.2. Estimating the covariances

Next, we show how to recover a rough approximation to the span of the covariance matrices and, as a consequence, produce a net containing rough approximations to each of the covariance matrices. The algorithm is summarized in Algorithm 3.

Lemma 17 *Suppose $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k \in \mathbb{R}^d$ satisfy $\|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i\|^2 \le \upsilon_{\mathrm{mean}}$ for all $i \in [k]$. Then there is an algorithm* CRUDEESTIMATECOVARIANCES*$(q, \{\widehat{\boldsymbol{\mu}}_i\})$ which returns a list $\mathcal{W}$ such that for each $i \in [k]$, there exists $\widehat{\mathbf{Q}}_i \in \mathcal{W}$ for which $\|\mathbf{Q}_i - \widehat{\mathbf{Q}}_i\|_F \lesssim \beta^{1/2}\upsilon_{\mathrm{mean}}^{1/2} + k^{3/2}\upsilon_{\mathrm{mean}} + k^{5/2}\beta + k^2\alpha\log R$. Furthermore $|\mathcal{W}| \le d^{O(k)}$, and the algorithm runs in time $\mathrm{poly}(dR/\beta) + d^{O(k)}$ and draws $\mathrm{poly}(dR/\beta)$ samples.*

The intuition behind our approach is that if the means of the mixture were all sufficiently close to zero, then the top-k singular subspace of the matrix $\mathbb{E}[\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)^\intercal]$ can be shown to contain points close to $\mathrm{vec}(\mathbf{Q}_1), \ldots, \mathrm{vec}(\mathbf{Q}_k)$. In general, if the means are arbitrary, then we can use the estimates $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$ derived in the previous section to approximately "recenter" the mixture components near zero. We now make this intuition precise.

Proof preliminaries. Define

$$\widehat{\boldsymbol{\Pi}} \triangleq \mathrm{span}(\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k) \quad \text{and} \quad \widehat{\boldsymbol{\Pi}}^\perp \triangleq \mathbf{Id} - \widehat{\boldsymbol{\Pi}}\,.$$

Let $\boldsymbol{\mu}_i^\parallel \triangleq \widehat{\boldsymbol{\Pi}}\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_i^\perp \triangleq \widehat{\boldsymbol{\Pi}}^\perp\boldsymbol{\mu}_i$. Note that

$$\|\boldsymbol{\mu}_i^\perp\|^2 = \|\widehat{\boldsymbol{\Pi}}^\perp(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i)\|^2 \le \upsilon_{mean}\,.$$

Also define

$$\boldsymbol{\zeta}_i \triangleq \boldsymbol{\mu}_i^\parallel - \widehat{\boldsymbol{\mu}}_i$$

and note that

$$\|\boldsymbol{\zeta}_i\|^2 = \|\widehat{\boldsymbol{\Pi}}(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i)\|^2 \lesssim \upsilon_{\mathrm{mean}}\,.$$

Define $\Delta \ge 1$ by

$$\Delta \triangleq C(\sqrt{\upsilon_{mean}} + \sqrt{k\beta} + k^{1/4}\sqrt{\alpha\log R}) \tag{8}$$

for sufficiently large absolute constant $C > 0$. Given $i \in [k]$, define

$$S_{\mathrm{far}}[i] \triangleq \{j \in [k] : \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \ge \Delta\} \quad \text{and} \quad S_{close}[i] \triangleq \{j \in [k] : \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \le \Delta\}\,.$$

The algorithm we give in this section (Algorithm 3) does not require knowledge of $S_{far}[i], S_{close}[i]$; these sets are only defined here for the purpose of analysis.

To approximately "recenter" the mixture components around zero, we will subtract from each sample the mean estimate which is closest to it in the subspace given by $\widehat{\boldsymbol{\Pi}}$. Formally, given $\mathbf{x} \sim \mathbb{R}^d$, define $\widehat{\boldsymbol{\mu}}(\mathbf{x})$ by

$$\widehat{\boldsymbol{\mu}}(\mathbf{x}) \triangleq \widehat{\boldsymbol{\mu}}_i \quad \text{for} \quad i = \operatorname*{argmin}_{j\in[k]} \|\widehat{\boldsymbol{\mu}}_j - \widehat{\boldsymbol{\Pi}}\mathbf{x}\|\,. \tag{9}$$

For every $i \in [k]$, define

$$\mathcal{K}_i \triangleq \{\mathbf{x} \in \mathbb{R}^d : \widehat{\boldsymbol{\mu}}(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_i\}\,,$$

i.e. the set of points which are closest to $\widehat{\boldsymbol{\mu}}_i$ in the subspace given by $\widehat{\boldsymbol{\Pi}}$.

Finally, given $\mathbf{z} \in \mathbb{R}^d$, define

$$\begin{aligned}
\mathbf{\Psi}_{00}(\mathbf{z}) &\triangleq \mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{z}\mathbf{z}^\intercal\widehat{\mathbf{\Pi}})\mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{z}\mathbf{z}^\intercal\widehat{\mathbf{\Pi}})^\intercal \\
\mathbf{\Psi}_{01}(\mathbf{z}) &\triangleq \mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{z}\mathbf{z}^\intercal\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{z}\mathbf{z}^\intercal\widehat{\mathbf{\Pi}}^\perp)^\intercal \qquad (10)\\
\mathbf{\Psi}_{11}(\mathbf{z}) &\triangleq \mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp\mathbf{z}\mathbf{z}^\intercal\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp\mathbf{z}\mathbf{z}^\intercal\widehat{\mathbf{\Pi}}^\perp)^\intercal\,.
\end{aligned}$$

We will assemble an estimate for the span of the covariances out of the top-k singular subspaces of empirical estimates of $\mathbb{E}_{\mathcal{M}}[\mathbf{\Psi}_{00}(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))], \mathbb{E}_{\mathcal{M}}[\mathbf{\Psi}_{01}(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))], \mathbb{E}_{\mathcal{M}}[\mathbf{\Psi}_{11}(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))]$.

For any $i \in [k]$ and $s \in \{00, 01, 11\}$, note that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))] &= \mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)\cdot \mathbb{1}[\mathbf{x}\in\mathcal{K}_i]] + \sum_{j\in S_{close}[i]\backslash i}\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_j)\cdot \mathbb{1}[\mathbf{x}\in\mathcal{K}_j]] \\
&\qquad + \sum_{j\in S_{\mathrm{far}}[i]}\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_j)\cdot \mathbb{1}[\mathbf{x}\in\mathcal{K}_j]] \\
&= \mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)] + \sum_{j\in S_{close}[i]\backslash i}\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_j)-\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i))\cdot \mathbb{1}[\mathbf{x}\in\mathcal{K}_j]] \\
&\qquad + \sum_{j\in S_{\mathrm{far}}[i]}\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_j)-\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i))\cdot \mathbb{1}[\mathbf{x}\in\mathcal{K}_j]]\,,
\end{aligned} \qquad (11)$$

where we used that $\mathcal{K}_1, \ldots, \mathcal{K}_m$ forms a partition of $\mathbb{R}^d$.

Constructing an approximation for $\sum_i \lambda_i \mathrm{vec}(\mathbf{Q}_i)\mathrm{vec}(\mathbf{Q}_i)^\intercal$. We will now argue that the two sums in Eq. (11) are negligible compared to the term $\mathbb{E}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)]$. This will allow us to construct a matrix that is close to $\sum_i \lambda_i \mathrm{vec}(\mathbf{Q}_i)\mathrm{vec}(\mathbf{Q}_i)^\intercal$.

In the expression $\mathbb{E}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))]$ above, we are recentering $\mathbf{x}$ around $\widehat{\boldsymbol{\mu}}(\mathbf{x})$. We first show that the probability that a sample from the i-th component lands in $\mathcal{K}_j$ for some $j \in S_{\mathrm{far}}[i]$ is small, meaning that with high probability we are correctly recentering $\mathbf{x}$ around $\widehat{\boldsymbol{\mu}}_j$ for some $j \in S_{close}[i]$.

Lemma 18 *For any $i \in [k]$, $\Pr_{\mathcal{N}_i}[\mathbf{x} \in \mathcal{K}_j$ for some $j \in S_{\mathrm{far}}[i]] \le 1/R^8$.*

Proof Note that $\mathrm{tr}(\mathbf{Q}_i\widehat{\mathbf{\Pi}}) \le k\beta$ and $\|\mathbf{Q}_i^{1/2}\widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\|_F^2 = \mathrm{tr}(\mathbf{Q}_i\widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}}) \ge k\alpha^2$. Therefore, for $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$, we may apply Hanson-Wright (Theorem 19 to control the tails of $\|\widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\mathbf{z}\|^2$).

Fact 19 (Hanson-Wright) *Suppose $\mathbf{A} \in \mathbb{R}^{d\times d}$ satisfies $\|\mathbf{A}\|_F^2/\|\mathbf{A}\|_{\mathrm{op}}^2 \ge r$. Then for any $s > 0$,*

$$\Pr_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{Id})}[\mathbf{x}^\intercal\mathbf{A}\mathbf{x} - \mathrm{tr}(\mathbf{A}) > s\|\mathbf{A}\|_F] \le \exp(-\Omega(\min(s\sqrt{r}, s^2))) \qquad (12)$$

$$\Pr_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{Id})}[\mathbf{x}^\intercal\mathbf{A}\mathbf{x} - \mathrm{tr}(\mathbf{A}) < -s\|\mathbf{A}\|_F] \le \exp(-\Omega(\min(s\sqrt{r}, s^2)))\,. \qquad (13)$$

By taking r in Theorem 19 to be 1, we find that there is an absolute constant $C' > 0$ such that

$$\Pr\big[\|\widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\mathbf{z}\|^2 > k\beta + C'\alpha\sqrt{k}\log R\big] \le 1/R^8\,.$$

Given $\mathbf{x} \sim \mathcal{N}_i$, note that $\widehat{\boldsymbol{\mu}}_i - \widehat{\mathbf{\Pi}}\mathbf{x} = \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + \widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$. Thus, conditioned on the above event,

$$\|\widehat{\boldsymbol{\mu}}_i - \widehat{\mathbf{\Pi}}\mathbf{x}\| \le \sqrt{\upsilon_{\mathrm{mean}}} + \sqrt{k\beta} + k^{1/4}\sqrt{C'\alpha \log R}\,.$$

For any $j \in [k]$ and $\mathbf{x} \sim \mathcal{N}_i$, note that $\widehat{\boldsymbol{\mu}}_j - \widehat{\mathbf{\Pi}}\mathbf{x} = \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_j - \widehat{\boldsymbol{\mu}}_i) + \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + \widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$. If $j \in S_{far}[i]$, we have

$$\|\widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_j - \widehat{\boldsymbol{\mu}}_i) + \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\| \ge \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\| - 3\sqrt{\upsilon_{\mathrm{mean}}} \ge \Delta - 3\sqrt{\upsilon_{\mathrm{mean}}}\,.$$

Thus, conditioned on the above event,

$$\|\widehat{\boldsymbol{\mu}}_j - \widehat{\mathbf{\Pi}}\mathbf{x}\| \ge \Delta - 3\sqrt{\upsilon_{\mathrm{mean}}} - \sqrt{k\beta} - k^{1/4}\sqrt{C'\alpha \log R}\,.$$

By our choice of Δ in Eq. (8), if C therein is a sufficiently large constant, the above is larger than $\|\widehat{\boldsymbol{\mu}}_i - \widehat{\mathbf{\Pi}}\mathbf{x}\|$ as desired. ■

Next, we argue that the "signal terms" $\mathbb{E}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)]$ in Eq. (11) are well-approximated by the rank-one matrices $\mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}})\mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}})^\intercal$, $\mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}}^\perp)^\intercal$, and $\mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp\mathbf{Q}_i\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp\mathbf{Q}_i\widehat{\mathbf{\Pi}}^\perp)^\intercal$.

Lemma 20

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)] - \mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\mathbf{\Pi}})\mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\mathbf{\Pi}})^\intercal\|_{op} \lesssim \beta^2 + \beta\upsilon_{\mathrm{mean}}$$

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_{01}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)] - \mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)^\intercal\|_{op} \lesssim \beta^2 + \beta\upsilon_{\mathrm{mean}}$$

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_{11}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)] - \mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp(\mathbf{Q}_i + \boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp(\mathbf{Q}_i + \boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)^\intercal\|_{op} \lesssim \beta^2 + \beta\upsilon_{\mathrm{mean}}\,.$$

Proof We will be bounding the operator norm of matrices of the form of $\mathbb{E}[\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)]$ where $\mathbf{x}$ is a Gaussian vector. To do so we take any test vector $\mathbf{A} \in \mathbb{R}^{d^2}$ for which $\|\mathbf{A}\|_F = 1$; we will regard it interchangeably as a vector or as a $d \times d$ matrix. We then bound $\mathrm{vec}(\mathbf{A})^\intercal\,\mathbb{E}_{\mathbf{x}}[\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)\mathrm{vec}(\mathbf{x}\mathbf{x}^\intercal)]\mathrm{vec}(\mathbf{A})$ using the following simple lemma (that follows from Wicks' identity for the fourth Gaussian moments).

Lemma 21 *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be any matrix and $\mathbf{Q}$ be a covariance matrix. Then for $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q})$, we have*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{Q})}[(\mathbf{x}^\intercal\mathbf{A}\mathbf{x})^2] = \langle\mathbf{A}, \mathbf{Q}\rangle^2 + 2\|\mathbf{Q}^{1/2}\mathbf{A}\mathbf{Q}^{1/2}\|_F^2 + \|\mathbf{Q}^{1/2}\mathbf{A}^\intercal\boldsymbol{\mu}\|^2 + \|\mathbf{Q}^{1/2}\mathbf{A}\boldsymbol{\mu}\|^2 + (\boldsymbol{\mu}^\intercal\mathbf{A}\boldsymbol{\mu})^2$$
$$+ 2\boldsymbol{\mu}^\intercal\mathbf{A}\boldsymbol{\mu}\cdot\langle\mathbf{Q}, \mathbf{A}\rangle + 2\mathrm{tr}(\mathbf{Q}^{1/2}\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^\intercal\mathbf{A}\mathbf{Q}^{1/2}).$$

Moreover, if $\|\mathbf{A}\|_F \le 1$ and $\|\mathbf{Q}\|_{op} \le \beta$, then

$$\left|\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{Q})}[(\mathbf{x}^\intercal\mathbf{A}\mathbf{x})^2] - \langle\mathbf{A}, \mathbf{Q}\rangle^2 - 2(\boldsymbol{\mu}^\intercal\mathbf{A}\boldsymbol{\mu})\langle\mathbf{A}, \mathbf{Q}\rangle - (\boldsymbol{\mu}^\intercal\mathbf{A}\boldsymbol{\mu})^2\right| \lesssim \max(\beta^2, \beta\|\mathbf{A}^\intercal\boldsymbol{\mu}\|^2, \beta\|\mathbf{A}\boldsymbol{\mu}\|^2)\,. \tag{14}$$

Proof Writing $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q})$ as $\mathbf{x} = \mathbf{Q}^{1/2}\mathbf{g} + \boldsymbol{\mu}$ for $\mathbf{g} \sim \mathcal{N}$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}}[((\mathbf{Q}^{1/2}\mathbf{g} + \boldsymbol{\mu})^\intercal \mathbf{A}(\mathbf{Q}^{1/2}\mathbf{g} + \boldsymbol{\mu}))^2] = \\
&\quad \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}}[(\mathbf{g}^\intercal \mathbf{Q}^{1/2}\mathbf{A}\mathbf{Q}^{1/2}\mathbf{g})^2] + \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}}[(\boldsymbol{\mu}^\intercal \mathbf{A}\mathbf{Q}^{1/2}\mathbf{g})^2] + \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}}[(\mathbf{g}^\intercal \mathbf{Q}^{1/2}\mathbf{A}\boldsymbol{\mu})^2] + \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}}[(\boldsymbol{\mu}^\intercal \mathbf{A}\boldsymbol{\mu})^2] \\
&\quad + 2\boldsymbol{\mu}^\intercal \mathbf{A}\boldsymbol{\mu} \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}}[(\mathbf{g}^\intercal \mathbf{Q}^{1/2}\mathbf{A}\mathbf{Q}^{1/2}\mathbf{g})] + 2 \mathop{\mathbb{E}}_{\mathbf{g}\sim\mathcal{N}}[(\mathbf{g}^\intercal \mathbf{Q}^{1/2}\mathbf{A}\boldsymbol{\mu})(\boldsymbol{\mu}^\intercal \mathbf{A}\mathbf{Q}^{1/2}\mathbf{g})].
\end{aligned}$$

Using the definition of $\mathbf{B} = \mathbf{Q}^{1/2}\mathbf{A}\mathbf{Q}^{1/2}$, we have

$$\mathbb{E}[(\mathbf{g}^\intercal \mathbf{B}\mathbf{g})^2] = \sum_{i,j=1}^{d} \mathbf{B}_{i,i}\mathbf{B}_{j,j}\, \mathbb{E}[\mathbf{g}_i^2 \mathbf{g}_j^2] + 2\sum_{i,j=1}^{d} \mathbf{B}_{i,j}^2\, \mathbb{E}[\mathbf{g}_i^2 \mathbf{g}_j^2] = \operatorname{tr}(\mathbf{B})^2 + 2\|\mathbf{B}\|_F^2.$$

Using the fact that $\mathbb{E}[\mathbf{g}^\intercal \mathbf{M}\mathbf{g}] = \operatorname{tr}(\mathbf{M})$ for any matrix $\mathbf{M}$, we obtain the result. ∎

For the first claimed inequality, we apply Eq. (14) from Theorem 21 to $\widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}$ and $\mathbf{x}' \sim \mathcal{N}(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i, \mathbf{Q}_i)$ to get

$$\left|\mathbb{E}[(\mathbf{x}'^\intercal \widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}\mathbf{x}')^2] - \langle \mathbf{A}, \widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\mathbf{\Pi}}\rangle^2\right| \lesssim \beta^2 + \beta\|\widehat{\mathbf{\Pi}}\mathbf{A}^\intercal\boldsymbol{\zeta}_i\|^2 + \beta\|\widehat{\mathbf{\Pi}}\mathbf{A}\boldsymbol{\zeta}_i\|^2 .$$

Note that $\|\widehat{\mathbf{\Pi}}\mathbf{A}^\intercal\boldsymbol{\zeta}_i\|^2 \le \upsilon_{\text{mean}}$ and $\|\widehat{\mathbf{\Pi}}\mathbf{A}\boldsymbol{\zeta}_i\|^2 \le \upsilon_{\text{mean}}$, so

$$\mathbb{E}[(\mathbf{x}'^\intercal \widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}\mathbf{x}')^2] = \langle \mathbf{A}, \widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\mathbf{\Pi}}\rangle^2 \pm O(\beta^2 + \beta\upsilon_{\text{mean}}) .$$

Furthermore, $\mathbf{A}^\intercal \mathbf{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)\mathbf{A} = (\mathbf{x}'^\intercal \widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}\mathbf{x}')^2$ for $\mathbf{x}' = \mathbf{x} - \widehat{\boldsymbol{\mu}}_i$, so because the above bound holds for all $\mathbf{A}$ for which $\|\mathbf{A}\|_F = 1$, the first claimed inequality follows.

The proof of the second inequality proceeds similarly. By Eq. (14) applied to $\widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}^\perp$ and $\mathbf{x}' \sim \mathcal{N}(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i, \mathbf{Q}_i)$, we get

$$\left|\mathbb{E}[(\mathbf{x}'^\intercal \widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}^\perp\mathbf{x}')^2] - \langle \mathbf{A}, \widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp\rangle^2\right| \lesssim \beta^2 + \beta\|\widehat{\mathbf{\Pi}}\mathbf{A}\boldsymbol{\mu}_i^\perp\|^2 + \beta\|\widehat{\mathbf{\Pi}}^\perp\mathbf{A}^\intercal\boldsymbol{\zeta}_i\|^2 .$$

Note that $\|\widehat{\mathbf{\Pi}}\mathbf{A}\boldsymbol{\mu}_i^\perp\|^2 \le \upsilon_{\text{mean}}$ and $\|\widehat{\mathbf{\Pi}}^\perp\mathbf{A}^\intercal\boldsymbol{\zeta}_i\|^2 \le \upsilon_{\text{mean}}$, so

$$\mathbb{E}[(\mathbf{x}'\widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}^\perp\mathbf{x}')^2] = \langle \mathbf{A}, \widehat{\mathbf{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp\rangle^2 \pm O(\beta^2 + \beta\upsilon_{\text{mean}}) .$$

Furthermore, $\mathbf{A}^\intercal \mathbf{\Psi}_{01}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)\mathbf{A} = (\mathbf{x}'^\intercal \widehat{\mathbf{\Pi}}\mathbf{A}\widehat{\mathbf{\Pi}}^\perp\mathbf{x}')^2$ for $\mathbf{x}' = \mathbf{x} - \widehat{\boldsymbol{\mu}}_i$, so because the above bound holds for all $\mathbf{A}$ for which $\|\mathbf{A}\|_F = 1$, the second claimed inequality follows.

For the third inequality, by Eq. (14) applied to $\widehat{\mathbf{\Pi}}^\perp\mathbf{A}\widehat{\mathbf{\Pi}}^\perp$ and $\mathbf{x}' \sim \mathcal{N}(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i, \mathbf{Q}_i)$, we get

$$\left|\mathbb{E}[(\mathbf{x}'^\intercal \widehat{\mathbf{\Pi}}^\perp\mathbf{A}\widehat{\mathbf{\Pi}}^\perp\mathbf{x}')^2] - \langle \mathbf{A}, \widehat{\mathbf{\Pi}}^\perp(\mathbf{Q}_i + \boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp\rangle^2\right| \lesssim \beta^2 + \beta\|\widehat{\mathbf{\Pi}}^\perp\mathbf{A}\boldsymbol{\mu}_i^\perp\|^2 + \beta\|\widehat{\mathbf{\Pi}}^\perp\mathbf{A}^\intercal\boldsymbol{\mu}_i^\perp\|^2 .$$

Note that $\|\widehat{\mathbf{\Pi}}^\perp\mathbf{A}\boldsymbol{\mu}_i^\perp\|^2 \le \upsilon_{\text{mean}}$ and $\|\widehat{\mathbf{\Pi}}^\perp\mathbf{A}^\intercal\boldsymbol{\mu}_i^\perp\|^2 \le \upsilon_{\text{mean}}$, so

$$\mathbb{E}[(\mathbf{x}'\widehat{\mathbf{\Pi}}^\perp\mathbf{A}\widehat{\mathbf{\Pi}}^\perp\mathbf{x}')^2] = \langle \mathbf{A}, \widehat{\mathbf{\Pi}}^\perp(\mathbf{Q}_i + \boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp\rangle^2 \pm O(\beta^2 + \beta\upsilon_{\text{mean}}) .$$

Furthermore, $\mathbf{A}^\top \mathbf{\Psi}_{11}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)\mathbf{A} = (\mathbf{x}'^\top \widehat{\mathbf{\Pi}}^\perp \mathbf{A} \widehat{\mathbf{\Pi}}^\perp \mathbf{x}')^2$ for $\mathbf{x}' = \mathbf{x} - \widehat{\boldsymbol{\mu}}_i$, so because the above bound holds for all $\mathbf{A}$ for which $\|\mathbf{A}\|_F = 1$, the third claimed inequality follows. ■

Now if we can show that the remaining terms in Eq. (11) have small norm, then we can argue that we can read off a rough approximation of $\sum_i \lambda_i \mathrm{vec}(\mathbf{Q}_i)\mathrm{vec}(\mathbf{Q}_i)^\top$ from $\mathbb{E}_{\mathcal{N}_i}[\mathbf{\Psi}_s(\mathbf{x} - \widehat{\boldsymbol{\mu}}(\mathbf{x}))]$. In the following Lemma, we show the remaining terms in Eq. (11) are indeed bounded:

Lemma 22 *Let $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\mu}}'$ be any vectors from among $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$. Suppose that either of the following holds:*

- *$j \in S_{\mathrm{far}}[i]$, or*
- *$j \in S_{\mathrm{close}}[i]$ and additionally $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\mu}}'$ are centers of components in $S_{\mathrm{close}}[i]$.*

Then

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\mathbf{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}) - \mathbf{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}')) \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]]\|_{op} \lesssim \beta^2 k^2 + \Delta^4$$

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\mathbf{\Psi}_{01}(\mathbf{x} - \widehat{\boldsymbol{\mu}}) - \mathbf{\Psi}_{01}(\mathbf{x} - \widehat{\boldsymbol{\mu}}')) \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]]\|_{op} \lesssim \beta^{3/2}\Delta + \beta\Delta^2$$

$$\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\mathbf{\Psi}_{11}(\mathbf{x} - \widehat{\boldsymbol{\mu}}) - \mathbf{\Psi}_{11}(\mathbf{x} - \widehat{\boldsymbol{\mu}}')) \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]] = 0\,.$$

Proof For $\mathbf{x} \sim \mathcal{N}_i$, define $\widetilde{\mathbf{x}} \triangleq \widehat{\mathbf{\Pi}}\mathbf{x} - \widehat{\boldsymbol{\mu}}$, $\widetilde{\mathbf{x}}' \triangleq \widehat{\mathbf{\Pi}}\mathbf{x} - \widehat{\boldsymbol{\mu}}'$, and $\mathbf{x}^\perp \triangleq \widehat{\mathbf{\Pi}}^\perp \mathbf{x}$ so that $\mathbf{x} - \widehat{\boldsymbol{\mu}} = \widetilde{\mathbf{x}} + \mathbf{x}^\perp$ and $\mathbf{x} - \widehat{\boldsymbol{\mu}}' = \widetilde{\mathbf{x}}' + \mathbf{x}^\perp$.

Let $\mathbf{A} \in \mathbb{R}^{d^2}$ be a test vector which we regard interchangeably as a vector and as a $d \times d$ matrix, and which satisfies $\|\mathbf{A}\|_F = 1$.

<u>*Proof for $\mathbf{\Psi}_{00}$*</u>: We have

$$\left|\mathbf{A}^\top(\mathbf{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}) - \mathbf{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}'))\mathbf{A} \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]\right| = \left|(\widetilde{\mathbf{x}}^\top \mathbf{A}\widetilde{\mathbf{x}})^2 - (\widetilde{\mathbf{x}}'^\top \mathbf{A}\widetilde{\mathbf{x}}')^2\right| \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]\,.$$

To bound the expectation of this over $\mathbf{x} \sim \mathcal{N}_i$, it suffices to bound $\mathbb{E}_{\mathcal{N}_i}[(\widetilde{\mathbf{x}}^\top \mathbf{A}\widetilde{\mathbf{x}})^2 \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]]$ and $\mathbb{E}_{\mathcal{N}_i}[(\widetilde{\mathbf{x}}'^\top \mathbf{A}\widetilde{\mathbf{x}}')^2 \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]]$. These can be handled in the same way, so here we consider the former.

First suppose that $j \in S_{\mathrm{far}}[i]$. By Cauchy-Schwarz,

$$\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\widetilde{\mathbf{x}}^\top \mathbf{A}\widetilde{\mathbf{x}})^2 \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]] \le \mathop{\mathbb{E}}_{\mathcal{N}_i}[(\widetilde{\mathbf{x}}^\top \mathbf{A}\widetilde{\mathbf{x}})^4]^{1/2} \cdot \Pr[\mathbf{x} \in \mathcal{K}_j]^{1/2}\,.$$

Note that

$$\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\widetilde{\mathbf{x}}^\top \mathbf{A}\widetilde{\mathbf{x}})^4]^{1/2} \le \mathop{\mathbb{E}}_{\mathcal{N}_i}[\|\widetilde{\mathbf{x}}\|^8]^{1/2} \lesssim \mathop{\mathbb{E}}_{\mathbf{h} \sim \mathcal{N}(0, \widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}})}[\|\mathbf{h}\|^8]^{1/2} + \|\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})\|^4 \lesssim \beta^2 k^2 + R^4\,.$$

The proof of the first part of the Lemma then follows by the fact that $\Pr[\mathbf{x} \in \mathcal{K}_j]^{1/2} \le 1/R^4$ by Theorem 18, so we get an overall bound of $\beta^2 k^2/R^4 + 1 \le \beta^2 k^2 + \Delta^4$ (as $\Delta, R \ge 1$ by assumption).

Next, suppose that $j \in S_{\mathrm{close}}[i]$ and additionally $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\mu}}'$ are centers of components in $S_{\mathrm{close}}[i]$. Then

$$\mathop{\mathbb{E}}_{\mathcal{N}_i}[(\widetilde{\mathbf{x}}^\top \mathbf{A}\widetilde{\mathbf{x}})^2 \cdot \mathbb{1}[\mathbf{x} \in \mathcal{K}_j]] \le \mathop{\mathbb{E}}_{\mathcal{N}_i}[(\widetilde{\mathbf{x}}^\top \mathbf{A}\widetilde{\mathbf{x}})^2] \le \mathop{\mathbb{E}}_{\mathcal{N}_i}[\|\widetilde{\mathbf{x}}\|^4] \lesssim \mathop{\mathbb{E}}_{\mathbf{h} \sim \mathcal{N}(0, \widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}})}[\|\mathbf{h}\|^4] + \|\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})\|^4 \lesssim \beta^2 k^2 + \Delta^4\,,$$

thus establishing the third part of the Lemma.

Proof for $\mathbf{\Psi}_{01}$: We have

$$\mathbf{A}^\intercal(\mathbf{\Psi}_{01}(\mathbf{x}-\widehat{\boldsymbol{\mu}})-\mathbf{\Psi}_{01}(\mathbf{x}-\widehat{\boldsymbol{\mu}}'))\mathbf{A}\cdot\mathbb{1}[\mathbf{x}\in\mathcal{K}_j]=\big((\widetilde{\mathbf{x}}^\intercal\mathbf{A}\mathbf{x}^\perp)^2-(\widetilde{\mathbf{x}}'^\intercal\mathbf{A}\mathbf{x}^\perp)^2\big)\cdot\mathbb{1}[\mathbf{x}\in\mathcal{K}_j]\,.$$

Note that the event that $\mathbf{x}\in\mathcal{K}_j$ only depends on $\mathbf{x}^\perp$, so the expectation of the above over $\mathbf{x}\sim\mathcal{N}_i$ is given by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{N}_i}\big[\big((\widetilde{\mathbf{x}}^\intercal\mathbf{A}\mathbf{x}^\perp)^2-(\widetilde{\mathbf{x}}'^\intercal\mathbf{A}\mathbf{x}^\perp)^2\big)\cdot\mathbb{1}[\mathbf{x}\in\mathcal{K}_j]\big]\\
&\quad=\mathop{\mathbb{E}}_{\widetilde{\mathbf{x}},\widetilde{\mathbf{x}}'}\big[\mathbb{1}[\mathbf{x}\in\mathcal{K}_j]\cdot\mathop{\mathbb{E}}_{\mathbf{x}^\perp}[(\widetilde{\mathbf{x}}^\intercal\mathbf{A}\mathbf{x}^\perp)^2-(\widetilde{\mathbf{x}}'^\intercal\mathbf{A}\mathbf{x}^\perp)^2]\big]\\
&\quad=\mathop{\mathbb{E}}_{\widetilde{\mathbf{x}},\widetilde{\mathbf{x}}'}\big[\mathbb{1}[\mathbf{x}\in\mathcal{K}_j]\cdot\langle\mathbf{A}\widehat{\mathbf{\Pi}}^\perp\mathbf{Q}_i\widehat{\mathbf{\Pi}}^\perp\mathbf{A}^\intercal,\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\intercal-\widetilde{\mathbf{x}}'\widetilde{\mathbf{x}}'^\intercal\rangle\big]\\
&\quad\le\Pr[\mathbf{x}\in\mathcal{K}_j]^{1/2}\cdot\mathop{\mathbb{E}}_{\widetilde{\mathbf{x}},\widetilde{\mathbf{x}}'}[\langle\mathbf{A}\widehat{\mathbf{\Pi}}^\perp\mathbf{Q}_i\widehat{\mathbf{\Pi}}^\perp\mathbf{A}^\intercal,\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\intercal-\widetilde{\mathbf{x}}'\widetilde{\mathbf{x}}'^\intercal\rangle^2]^{1/2}\\
&\quad\lesssim\Pr[\mathbf{x}\in\mathcal{K}_j]^{1/2}\cdot\beta\mathop{\mathbb{E}}_{\widetilde{\mathbf{x}},\widetilde{\mathbf{x}}'}[\|\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}-\widetilde{\mathbf{x}}'\widetilde{\mathbf{x}}'^\intercal\|_F^2]^{1/2}\\
&\quad=\Pr[\mathbf{x}\in\mathcal{K}_j]^{1/2}\cdot\beta\mathop{\mathbb{E}}_{\mathbf{h}\sim\mathcal{N}(0,\widehat{\mathbf{\Pi}}\mathbf{Q}_i\widehat{\mathbf{\Pi}})}[\|(\mathbf{h}+\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}))(\mathbf{h}+\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}))^\intercal\\
&\qquad\qquad\qquad\qquad-(\mathbf{h}+\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}'))(\mathbf{h}+\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}'))^\intercal\|_F^2]^{1/2}\\
&\quad=\Pr[\mathbf{x}\in\mathcal{K}_j]^{1/2}\cdot\beta\mathop{\mathbb{E}}_{\mathbf{h}}[\|\mathbf{h}\widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}'-\widehat{\boldsymbol{\mu}})^\intercal+(\widehat{\boldsymbol{\mu}}'-\widehat{\boldsymbol{\mu}})\widehat{\mathbf{\Pi}}\mathbf{h}^\intercal\\
&\qquad\qquad\qquad+\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}})^\intercal\widehat{\mathbf{\Pi}}-\widehat{\mathbf{\Pi}}(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}')(\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}')^\intercal\widehat{\mathbf{\Pi}}\|_F^2]^{1/2}\\
&\quad\lesssim\Pr[\mathbf{x}\in\mathcal{K}_j]^{1/2}\cdot\beta\big(\beta^{1/2}\|\widehat{\boldsymbol{\mu}}'-\widehat{\boldsymbol{\mu}}\|+\|\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}\|^2+\|\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}'\|^2\big)\,,
\end{aligned}\tag{15}$$

where in the second step we used that the covariance of $\mathbf{x}^\perp$ is $\widehat{\mathbf{\Pi}}^\perp\mathbf{Q}_i\widehat{\mathbf{\Pi}}^\perp$.

Suppose that $j\in S_{\mathrm{far}}[i]$. Then by Theorem 18, the above can be upper bounded by $\beta^{3/2}/R^3+\beta/R^2\le\beta^{3/2}\Delta+\beta\Delta^2$ (as $\Delta,R\ge1$ by assumption), completing the proof of the second part of the Lemma.

Next, suppose that $j\in S_{close}[i]$ and additionally $\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\mu}}'$ are centers of components in $S_{close}[i]$. Then Eq. (15) can be upper bounded by $\beta^{3/2}\Delta+\beta\Delta^2$, completing the proof of the fourth part of the Lemma.

Proof for $\mathbf{\Psi}_{11}$: We have

$$\mathbf{A}^\intercal(\mathbf{\Psi}_{11}(\mathbf{x}-\widehat{\boldsymbol{\mu}})-\mathbf{\Psi}_{11}(\mathbf{x}-\widehat{\boldsymbol{\mu}}'))\mathbf{A}\cdot\mathbb{1}[\mathbf{x}\in\mathcal{K}_j]=\big((\mathbf{x}^{\perp\intercal}\mathbf{A}\mathbf{x}^\perp)^2-(\mathbf{x}^{\perp\intercal}\mathbf{A}\mathbf{x}^\perp)^2\big)\cdot\mathbb{1}[\mathbf{x}\in\mathcal{K}_j]=0\,.$$

As this holds for all $\mathbf{A}$, the last part of the Lemma follows. ∎

By combining Eq. (11) with Theorem 20 and Theorem 22, we conclude the following:

Corollary 23

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_{00}(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))]-\mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i+\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\mathbf{\Pi}})\mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i+\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\mathbf{\Pi}})^\intercal\|_{\mathrm{op}}\lesssim\beta\upsilon_{\mathrm{mean}}+\beta^2k^3+k\Delta^4$$

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_{01}(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))]-\mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i+\boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}(\mathbf{Q}_i+\boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)^\intercal\|_{\mathrm{op}}\lesssim\beta^2+\beta\upsilon_{\mathrm{mean}}+k\beta^{3/2}\Delta+k\beta\Delta^2$$

$$\|\mathop{\mathbb{E}}_{\mathcal{N}_i}[\mathbf{\Psi}_{11}(\mathbf{x}-\widehat{\boldsymbol{\mu}}(\mathbf{x}))]-\mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp(\mathbf{Q}_i+\boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)\mathrm{vec}(\widehat{\mathbf{\Pi}}^\perp(\mathbf{Q}_i+\boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\mathbf{\Pi}}^\perp)^\intercal\|_{\mathrm{op}}\lesssim\beta^2+\beta\upsilon_{\mathrm{mean}}\,.$$

Using Theorem 23 and Theorem 15, we are now ready to state our algorithm and prove the main guarantee of this section.

Algorithm 3: CRUDEESTIMATECOVARIANCES($q, \{\widehat{\boldsymbol{\mu}}_i\}$)

Input: Sample access to q, estimates $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$
Output: List $\mathcal{W}$ containing approximations to $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$
$\widehat{\boldsymbol{\Pi}} \leftarrow$ span of $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$
Define the functions $\boldsymbol{\Psi}_s$ from Eq. (10) and $\widehat{\boldsymbol{\mu}}(\cdot)$ from Eq. (9) using $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$.
Initialize $\mathcal{W}$ to the empty set.
Draw samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from q for $N \leftarrow \mathrm{poly}(dR/\beta)$.
for $s \in \{00, 01, 11\}$ **do**
  $\widehat{\mathbf{C}}_s \leftarrow \frac{1}{N}\sum_{j=1}^N \boldsymbol{\Psi}_s(\mathbf{x}_j - \widehat{\boldsymbol{\mu}}(\mathbf{x}_j))$
  $\widehat{V}_s \leftarrow$ top-k singular subspace of $\widehat{\mathbf{C}}_s$
  $\mathcal{W}_s \leftarrow$ a β-net over vectors in $\widehat{V}_s$ with L_2 norm at most $\beta\sqrt{d}$
end
for $\widehat{\mathbf{Q}}^{00} \in \mathcal{W}_{00}$, $\widehat{\mathbf{Q}}^{01} \in \mathcal{W}_{01}$, $\widehat{\mathbf{Q}}^{11} \in \mathcal{W}_{11}$ **do**
  Add $\widehat{\mathbf{Q}}^{00} + \widehat{\mathbf{Q}}^{01} + (\widehat{\mathbf{Q}}^{01})^\intercal + \widehat{\mathbf{Q}}^{11}$ to $\mathcal{W}$.
end
return $\mathcal{W}$

Proof [Proof of Theorem 17] Consider the matrix $\mathbf{C}_{00} \triangleq \mathbb{E}_{\mathcal{M}}[\boldsymbol{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}(\mathbf{x}))] = \sum_i \lambda_i \mathbb{E}_{\mathcal{N}_i}[\boldsymbol{\Psi}_{00}(\mathbf{x} - \widehat{\boldsymbol{\mu}}(\mathbf{x}))]$. By standard matrix concentration, for $N = \mathrm{poly}(dR/\beta)$ given in Algorithm 3, we have that the matrix $\widehat{\mathbf{C}}_{00}$ constructed in Step Algorithm 3 of Algorithm 3 satisfies $\|\widehat{\mathbf{C}}_{00} - \mathbf{C}_{00}\|_{op} \le \beta$. Therefore, by triangle inequality and Theorem 23,

$$\|\widehat{\mathbf{C}}_{00} - \sum_i \lambda_i \mathrm{vec}(\widehat{\boldsymbol{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\boldsymbol{\Pi}})\mathrm{vec}(\widehat{\boldsymbol{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\boldsymbol{\Pi}})^\intercal\|_{op} \lesssim \beta\upsilon_{\mathrm{mean}} + \beta^2 k^3 + k\Delta^4\,.$$

By Theorem 15, this means that the top-k singular subspace of $\widehat{\mathbf{C}}_{00}$ contains d^2-dimensional vectors $\widehat{\mathbf{Q}}_1^{00}, \ldots, \widehat{\mathbf{Q}}_k^{00}$ which, regarded as $d \times d$ matrices, satisfy

$$\|\widehat{\mathbf{Q}}_i^{00} - \widehat{\boldsymbol{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal)\widehat{\boldsymbol{\Pi}}\|_F^2 \lesssim \beta\upsilon_{\mathrm{mean}} + \beta^2 k^3 + k\Delta^4$$

for all $i \in [k]$.

In an entirely analogous fashion, we can show that the top-k singular subspace of $\widehat{\mathbf{C}}_{01}$ contains d^2-dimensional vectors $\widehat{\mathbf{Q}}_1^{01}, \ldots, \widehat{\mathbf{Q}}_k^{01}$ satisfying

$$\|\widehat{\mathbf{Q}}_i^{01} - \widehat{\boldsymbol{\Pi}}(\mathbf{Q}_i + \boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\boldsymbol{\Pi}}^\perp\|_F^2 \lesssim \beta^2 + \beta\upsilon_{\mathrm{mean}} + k\beta^{3/2}\Delta + k\beta\Delta^2$$

Likewise, the top-k singular subspace of $\widehat{\mathbf{C}}_{11}$ contains d^2-dimensional vectors $\widehat{\mathbf{Q}}_1^{11}, \ldots, \widehat{\mathbf{Q}}_k^{11}$ satisfying

$$\|\widehat{\mathbf{Q}}_i^{11} - \widehat{\boldsymbol{\Pi}}^\perp(\mathbf{Q}_i + \boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal)\widehat{\boldsymbol{\Pi}}^\perp\|_F^2 \lesssim \beta^2 + \beta\upsilon_{\mathrm{mean}}\,.$$

Finally, note that

$$\|\widehat{\boldsymbol{\Pi}}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\intercal\widehat{\boldsymbol{\Pi}}\|_F, \|\widehat{\boldsymbol{\Pi}}\boldsymbol{\zeta}_i(\boldsymbol{\mu}_i^\perp)^\intercal\widehat{\boldsymbol{\Pi}}^\perp\|_F, \|\widehat{\boldsymbol{\Pi}}^\perp\boldsymbol{\mu}_i^\perp(\boldsymbol{\mu}_i^\perp)^\intercal\widehat{\boldsymbol{\Pi}}^\perp\|_F \le \upsilon_{\mathrm{mean}}\,.$$

Combining all of these bounds we find that

$$\|\widehat{\mathbf{Q}}_i^{00}+\widehat{\mathbf{Q}}_i^{01}+(\widehat{\mathbf{Q}}^{01})^\intercal+\widehat{\mathbf{Q}}_i^{11}-\mathbf{Q}_i\|_F \lesssim \beta^{1/2}\upsilon_{\text{mean}}^{1/2}+k^{3/2}(\beta+\Delta^2) \lesssim \beta^{1/2}\upsilon_{\text{mean}}^{1/2}+k^{3/2}\upsilon_{\text{mean}}+k^{5/2}\beta+k^2\alpha\log R\,.$$

The claim then follows from the fact that $\mathcal{W}_{00}, \mathcal{W}_{01}, \mathcal{W}_{11}$ in Step Algorithm 3 contain approximations to $\widehat{\mathbf{Q}}_i^{00}, \widehat{\mathbf{Q}}_i^{01}, \widehat{\mathbf{Q}}_i^{11}$ that are β-close in operator norm. Finally, note that the size of $\mathcal{W}$ is bounded by $d^{O(k)}$, by standard bounds on epsilon-nets. ∎

B.3. Putting everything together

It is straightforward to combine the results of the previous two sections to derive the proof of Theorem 13. First, for completeness, we provide the pseudocode for the algorithm:

Algorithm 4: CRUDEESTIMATE(q)

Input: Sample access to q
Output: List $\mathcal{W}$ containing approximations to $(\boldsymbol{\mu}_1, \mathbf{Q}_1), \ldots, (\boldsymbol{\mu}_k, \mathbf{Q}_k)$

```
W ← ∅
W^(μ) ← CrudeEstimateMeans(q)
for μ̂_1, ..., μ̂_k ∈ W^(μ) do
    W^(Q) ← CrudeEstimateCovariances(q, {μ̂_i})
    for i ∈ [k], Q̂ ∈ W^(Q) do
        Insert (μ̂_i, Q̂) into W
    end
end
return W
```

Proof [Proof of Theorem 13] By Theorem 14, in some iteration of Line Algorithm 4 of Algorithm 4, we get $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$ which satisfy $\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 \le \upsilon_{\text{mean}}$ for $\upsilon_{\text{mean}} = O(\beta/\lambda_{\min})$. Substituting this into Theorem 17, we conclude that for each $i \in [k]$, in some iteration of Line Algorithm 4 of Algorithm 4, we get $\widehat{\mathbf{Q}}$ satisfying $\|\widehat{\mathbf{Q}} - \mathbf{Q}_i\|_F \lesssim k^{3/2}\beta/\lambda_{\min} + k^2\alpha\log R$, where we used that $\lambda_{\min} \le 1/k$ to simplify the bound in Theorem 17.

For the bound on $|\mathcal{W}|$, note that there are $(R/\sqrt{\beta})^{O(k^2)}$ iterations of the outer loop, within each of which there are $d^{O(k)}$ iterations of the inner loop, so $|\mathcal{W}| = (R/\sqrt{\beta})^{O(k^2)} \cdot d^{O(k)}$ as claimed. For the runtime, CRUDEESTIMATEMEANS is called exactly once, and CRUDEESTIMATECOVARIANCES is called $(R/\sqrt{\beta})^{O(k^2)}$ times, so the overall runtime of the algorithm is $(R/\sqrt{\beta})^{O(k^2)} \cdot (\mathrm{poly}(d, 1/\beta) + d^{O(k)})$. ∎

Appendix C. Clustering via likelihood ratio estimates

In this section we present our main clustering guarantee, which leverages the estimates for the parameters we obtained from the previous section. As those estimates are only crude approximations to the true parameters, we will obtain a commensurately crude clustering. First, we formalize the notion of "clusters" and what it means to give an accurate clustering:

Definition 24 *Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ and $\mathcal{T} = \{T_1, \ldots, T_n\}$ be partitions of $[k]$.*

$(\mathcal{S}, \mathcal{T})$ *is a* $(\Delta_{\text{in}}^{(\boldsymbol{\mu})}, \Delta_{\text{in}}^{(\mathbf{Q})}, \Delta_{out}^{(\boldsymbol{\mu})}, \Delta_{out}^{(\mathbf{Q})})$-separated partition pair *if:*

- *For all $a \in [m]$ and $i, i' \in S_a$, we have that $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}\| \le \Delta_{\text{in}}^{(\boldsymbol{\mu})}$.*
- *For all distinct $a, a' \in [m]$ and $i \in S_a, i' \in S_{a'}$, we have that $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}\| \ge \Delta_{\text{out}}^{(\boldsymbol{\mu})}$.*
- *For all $b \in [n]$ and $i, i' \in T_b$, we have that $\|\mathbf{Q}_i - \mathbf{Q}_{i'}\|_F \le \Delta_{\text{in}}^{(\mathbf{Q})}$.*
- *For all distinct $b, b' \in [n]$ and $i \in T_b, i' \in T_{b'}$, we have that $\|\mathbf{Q}_i - \mathbf{Q}_{i'}\|_F \ge \Delta_{\text{out}}^{(\mathbf{Q})}$.*

Roughly speaking, $\mathcal{S}$ (resp. $\mathcal{T}$) partitions the mixture components into groups such that any two components in the same group have means (resp. covariances) that are not far, and any two components from two different groups have means (resp. covariances) that are not close. Their common refinement is a partition $\mathcal{U}$ such that any two components in the same group have both means and covariances not too far, and any two components from two different groups either have means not too close or covariances not too close.

By brute-forcing over pairs of partitions of $[k]$ (of which there are at most k^{2k}), we may assume we have access to $\mathcal{S}$ and $\mathcal{T}$, and thus to $\mathcal{U}$. Our goal is then to assign to every $\mathbf{x} \in \mathbb{R}^d$ an index into the partition $\mathcal{U}$. For $\mathbf{x}$ sampled from the i-th component of the mixture which belongs to the t-th group in $\mathcal{U}$, we would like our assignment to be t with high probability. The main result of this section is to show that this is indeed possible:

Proposition 25 *Suppose $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k \in \mathbb{R}^d$ and $\widehat{\mathbf{Q}}_1, \ldots, \widehat{\mathbf{Q}}_k \in \mathbb{R}^{d\times d}$ satisfy $\|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i\|^2 \le \upsilon_{\text{mean}}$ and $\|\mathbf{Q}_i - \widehat{\mathbf{Q}}_i\|_F \le \upsilon_{cov}$.*

Let $(\mathcal{S} = \{S_1, \ldots, S_m\}, \mathcal{T} = \{T_1, \ldots, T_n\})$ denote a $(\Delta_{\text{in}}^{(\boldsymbol{\mu})}, \Delta_{\text{in}}^{(\mathbf{Q})}, \Delta_{\text{out}}^{(\boldsymbol{\mu})}, \Delta_{out}^{(\mathbf{Q})})$-separated partition of $[k]$, where

$$\Delta_{out}^{(\mathbf{Q})} \ge \max(5(\beta/\alpha)^3 \upsilon_{\text{cov}}, c\alpha)\,, \qquad \Delta_{out}^{(\boldsymbol{\mu})} \ge \max(6\sqrt{\upsilon_{\text{mean}}}, 6\sqrt{\beta k})\,, \qquad \sqrt{\upsilon_{\text{mean}}} + \Delta_{\text{in}}^{(\boldsymbol{\mu})} \le c\Delta_{\text{out}}^{(\mathbf{Q})}\sqrt{\alpha}/\beta\,. \tag{16}$$

for sufficiently small constant $c > 0$. Let $\{U_1, \ldots, U_{n_c}\}$ denote the common refinement of $\mathcal{S}$ and $\mathcal{T}$.

Then there is an explicit deterministic function $\mathsf{c} : \mathbb{R}^d \to [n_{\text{c}}]$ using $\mathcal{S}$, $\mathcal{T}$, and $\{\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i\}$, such that for any $t \in [n_{\text{c}}]$ and $i \in U_t$,

$$\Pr_{\mathcal{N}_i}[\mathsf{c}(\mathbf{x}) \neq t] \le k^3 \exp\Big(-\Omega\Big(\frac{(\Delta_{out}^{(\boldsymbol{\mu})})^2}{\alpha\sqrt{k}} \wedge \frac{\alpha^6 (\Delta_{out}^{(\mathbf{Q})})^2}{\beta^6 \upsilon_{cov}^2} \wedge \frac{\alpha^2 \Delta_{out}^{(\mathbf{Q})}}{\beta^3}\Big)\Big)$$

At a high level, the idea is as follows. It is not too hard to determine which group in $\mathcal{S}$ a given point $\mathbf{x}$ should belong to, simply by checking which mean estimate $\widehat{\boldsymbol{\mu}}_i$ is closest to $\mathbf{x}$ after projecting to the subspace spanned by $\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_k$. For each group in $\mathcal{S}$, we can then effectively restrict our attention to components within that group and focus on clustering them according to their covariances. Roughly speaking, we accomplish this by comparing log-likelihoods of sampling $\mathbf{x}$ under $\mathcal{N}(\widehat{\boldsymbol{\mu}}_1, \widehat{\mathbf{Q}}_1), \ldots, \mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{Q}}_k)$ and choosing the group in $\mathcal{T}$ containing the component maximizing log-likelihood.

C.1. Proof preliminaries

First, we need the following basic lemma which implies that given estimates $\widehat{\mathbf{Q}}_1, \ldots, \widehat{\mathbf{Q}}_k$ for the covariances of the components, we can produce estimates $\widehat{\mathbf{K}}_1, \ldots, \widehat{\mathbf{K}}_k$ for the *inverse* covariances:

Lemma 26 *If $\widehat{\mathbf{Q}} \in \mathbb{R}^{d\times d}$ is a psd matrix satisfying $\|\mathbf{Q} - \widehat{\mathbf{Q}}\|_F \le \upsilon_{\rm cov}$, and $\alpha\,\mathbf{Id} \preceq \mathbf{Q} \preceq \beta\,\mathbf{Id}$, then $\|\mathbf{Q}'^{-1} - \mathbf{Q}^{-1}\|_F \le 4\upsilon_{\rm cov}/\alpha^2$ for $\mathbf{Q}' \in \mathbb{R}^{d\times d}$ defined as follows. Let $\widehat{\mathbf{Q}}$ have singular value decomposition $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$, and define $\mathbf{Q}' \triangleq \mathbf{U}\mathbf{\Lambda}'\mathbf{U}^\intercal$, where $\mathbf{\Lambda}'$ is given by replacing every diagonal entry of $\mathbf{\Lambda}$ less than $\alpha/2$ with $\alpha/2$.*

Proof Note that there are at most $4\upsilon_{\rm cov}^2/\alpha^2$ diagonal entries of Λ less than $\alpha/2$, or else we would violate the assumption that $\|\mathbf{Q} - \widehat{\mathbf{Q}}\|_F \le \upsilon_{\rm cov}$. So $\|\mathbf{Q}' - \widehat{\mathbf{Q}}\|_F \le \upsilon_{cov}$ and thus $\|\mathbf{Q}' - \mathbf{Q}\|_F \le 2\upsilon_{\rm cov}$. Finally, note that $\|\mathbf{Q}'^{-1}\|_{op} = \sigma_{\min}(\mathbf{Q}')^{-1} \le 2/\alpha$. We have

$$\|\mathbf{Q}'^{-1} - \mathbf{Q}^{-1}\|_F = \|\mathbf{Q}'^{-1}(\mathbf{Q}' - \mathbf{Q})\mathbf{Q}^{-1}\|_F \le 4\upsilon_{\rm cov}/\alpha^2\,. \tag{17}$$

■

Given $i, j \in [k]$ and $\mathbf{x}, \widehat{\boldsymbol{\mu}} \in \mathbb{R}^d$, define

$$\mathbf{\Lambda}_{ij}(\mathbf{x}; \widehat{\boldsymbol{\mu}}) = (\mathbf{x} - \widehat{\boldsymbol{\mu}})^\intercal \widehat{\mathbf{K}}_j (\mathbf{x} - \widehat{\boldsymbol{\mu}}) - \langle \mathbf{Q}_i, \widehat{\mathbf{K}}_j\rangle\,.$$

Note that for any $\boldsymbol{\mu}, \mathbf{Q}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(\boldsymbol{\mu},\mathbf{Q})}[\mathbf{\Lambda}_{ii}(\mathbf{x}; \widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(x; \widehat{\boldsymbol{\mu}})] = \langle (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}})^\intercal + \mathbf{Q} - \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle\,.$$

Provided $\boldsymbol{\mu}$ and $\widehat{\boldsymbol{\mu}}$ are close, if $\mathbf{Q} = \mathbf{Q}_i$ then this quantity is close to zero, but if $\mathbf{Q} = \mathbf{Q}_j$ then this quantity scales as

$$\langle \mathbf{Q}_j - \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle \approx \langle \mathbf{Q}_j - \mathbf{Q}_i, \mathbf{Q}_i^{-1} - \mathbf{Q}_j^{-1}\rangle = \mathrm{tr}(\mathbf{Q}_j\mathbf{Q}_i^{-1}) + \mathrm{tr}(\mathbf{Q}_i\mathbf{Q}_j^{-1}) - 2d\,,$$

which can be quite large in comparison. Motivated by this, we will use $\mathbf{\Lambda}_{ii}(\mathbf{x}; \widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(\mathbf{x}; \widehat{\boldsymbol{\mu}})$ to cluster the samples according to the covariances of the components generating them.

C.2. Properties of $\mathbf{\Lambda}_{ij}$

Lemma 27 *Suppose $\Delta^{(\mathbf{Q})}_{\rm out} \ge 5(\beta/\alpha)^3\upsilon_{\rm cov}$. Let $i, j \in [k]$. Suppose $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^d$ satisfies*

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_j\| \le c\Delta^{(\mathbf{Q})}_{\rm out}\sqrt{\alpha}/\beta \tag{18}$$

for some $c > 0$.

If $\|\mathbf{Q}_j - \mathbf{Q}_i\|_F \ge \Delta^{(\mathbf{Q})}_{out}$, then for any $c' > 0$, with probability at least $1 - \exp(-\Omega(c'^2(\alpha^4/\beta^6) \cdot \|\mathbf{Q}_j - \mathbf{Q}_i\|_F^2 \cdot \min(1, \alpha^2/\upsilon_{\rm cov}^2)))$ over $\mathbf{x} \sim \mathcal{N}_j$,

$$\mathbf{\Lambda}_{ii}(\mathbf{x}; \widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(\mathbf{x}; \widehat{\boldsymbol{\mu}}) \ge \langle \mathbf{Q}_j - \mathbf{Q}_i, \mathbf{Q}_i^{-1} - \mathbf{Q}_j^{-1}\rangle - E\,,$$

where

$$E \triangleq (c^2 + 2c')\|\mathbf{Q}_j - \mathbf{Q}_i\|_F^2/\beta^2 + (4\upsilon_{\rm cov}/\alpha^2) \cdot \|\mathbf{Q}_j - \mathbf{Q}_i\|_F\,.$$

Proof Define $\mathbf{B} \triangleq \mathbf{Q}_j^{1/2}(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)\mathbf{Q}_j^{1/2}$ and $\mathbf{w} \triangleq \mathbf{Q}_j^{1/2}(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)(\boldsymbol{\mu}_j - \widehat{\boldsymbol{\mu}})$. Then for $\mathbf{x} \sim \mathcal{N}_j$, writing this as $\mathbf{x} = \boldsymbol{\mu}_j + \mathbf{Q}_j^{1/2}\mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$, we see that the quantity $\boldsymbol{\Lambda}_{ii}(\mathbf{x}) - \boldsymbol{\Lambda}_{ij}(\mathbf{x})$ is distributed as

$$\mathbf{z}^\intercal \mathbf{B}\mathbf{z} - 2\langle \mathbf{z}, \mathbf{w}\rangle + \langle (\boldsymbol{\mu}_j - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \widehat{\boldsymbol{\mu}})^\intercal - \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle\,. \tag{19}$$

Controlling $\mathbf{z}^\intercal\mathbf{B}\mathbf{z}$: We would like to apply Theorem 19. Note that

$$\begin{aligned}
\|\mathbf{B}\|_F &\ge \|\mathbf{Q}_j^{1/2}(\mathbf{Q}_i^{-1} - \mathbf{Q}_j^{-1})\mathbf{Q}_j^{1/2}\|_F - 4\beta\upsilon_{cov}/\alpha^2 \\
&= \|\mathbf{Q}_j^{1/2}\mathbf{Q}_i^{-1}(\mathbf{Q}_j - \mathbf{Q}_i)\mathbf{Q}_j^{-1}\mathbf{Q}_j^{1/2}\|_F - 4\beta\upsilon_{cov}/\alpha^2 \\
&\ge (\alpha/\beta^2)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F - 4\beta\upsilon_{cov}/\alpha^2 \gtrsim (\alpha/\beta^2)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F\,,
\end{aligned}$$

where in the last step we used the fact that $\Delta_{out}^{(\mathbf{Q})}$ satisfies $\Delta_{out}^{(\mathbf{Q})} \ge 5(\beta/\alpha)^3\upsilon_{cov}$ by hypothesis. Furthermore, $\|\mathbf{B}\|_{op} \lesssim (\beta/\alpha)\cdot(\upsilon_{cov}/\alpha + 1)$, so $\|\mathbf{B}\|_F/\|\mathbf{B}\|_{op} \gtrsim (\alpha^2/\beta^3)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F \cdot \min(1, \alpha/\upsilon_{cov})$.

Additionally,

$$\begin{aligned}
\|\mathbf{B}\|_F &\le \|\mathbf{Q}_j^{1/2}\mathbf{Q}_i^{-1}(\mathbf{Q}_j - \mathbf{Q}_i)\mathbf{Q}_j^{-1}\mathbf{Q}_j^{1/2}\|_F + \beta\upsilon_{cov}/\alpha^2 \\
&\le (\beta/\alpha^2)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F + \beta\upsilon_{cov}/\alpha^2 \\
&\lesssim (\beta/\alpha^2)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F\,,
\end{aligned}$$

where in the last step we used the assumption that $\Delta_{out}^{(\mathbf{Q})} \ge \upsilon_{cov}$.

By Theorem 19, for any $s > 0$, we have

$$\begin{aligned}
\Pr_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})}&\Big[\mathbf{z}^\intercal\mathbf{B}\mathbf{z} - \mathrm{tr}(\mathbf{Q}_j(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)) \le -s(\beta/\alpha^2)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F\Big] \\
&\le \exp(-\Omega(\min(s(\alpha^2/\beta^3)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F\cdot\min(1,\alpha/\upsilon_{cov}), s^2)))\,.
\end{aligned}\tag{20}$$

We will take

$$s = c'(\alpha^2/\beta^3)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F\cdot\min(1,\alpha/\upsilon_{cov})$$

for arbitrarily small constant $c' > 0$. By this choice of s, we have $s(\beta/\alpha^2)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F \le c'\|\mathbf{Q}_j - \mathbf{Q}_i\|_d^2 F/\beta^2$. Additionally, s^2 is the dominant term in the exponent in Eq. (20). Summarizing,

$$\Pr_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})}\Big[\mathbf{z}^\intercal\mathbf{B}\mathbf{z} - \mathrm{tr}(\mathbf{Q}_j(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)) \le -c'\|\mathbf{Q}_j - \mathbf{Q}_i\|_F^2/\beta^2\Big] \le \exp(-\Omega(s^2))\,. \tag{21}$$

Controlling $\langle\mathbf{z},\mathbf{w}\rangle$: Note that $\|\widehat{\mathbf{K}}_i\|_{op}, \|\widehat{\mathbf{K}}_j\|_{op} \lesssim 1/\alpha$, so $\|\mathbf{w}\| \lesssim \Delta_{out}^{(\mathbf{Q})}/\sqrt{\alpha\beta}$ by Eq. (18). Note that because $\Delta_{out}^{(\mathbf{Q})} \gtrsim \beta \ge \alpha^{5/2}/\beta^{3/2}$, we have that $s\Delta_{out}^{(\mathbf{Q})}/\sqrt{\alpha\beta} \le c'(\Delta_{out}^{(\mathbf{Q})})^2/\beta^2 \le c'\|\mathbf{Q}_j - \mathbf{Q}_i\|_F^2/\beta^2$. By standard Gaussian tail bounds, we conclude that

$$\Pr[|\langle\mathbf{z},\mathbf{w}\rangle| \ge c'\|\mathbf{Q}_j - \mathbf{Q}_i\|^2/\beta^2] \le \exp(-\Omega(s^2))\,. \tag{22}$$

Controlling $\langle(\boldsymbol{\mu}_j - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \widehat{\boldsymbol{\mu}})^\intercal, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle$: As $\|\widehat{\mathbf{K}}_i\|_{op}, \|\widehat{\mathbf{K}}_j\|_{op} \lesssim 1/\alpha$, by Eq. (18) we have that

$$|\langle(\boldsymbol{\mu}_j - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \widehat{\boldsymbol{\mu}})^\intercal, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle| \le c^2(\Delta_{out}^{(\mathbf{Q})})^2/\beta^2\,. \tag{23}$$

Putting things together: Conditioned on the events of Eq. (21) and (22) not holding, and also using the bound on the constant term in Eq. (23), we see from the decomposition of $\mathbf{\Lambda}_{ii}(\mathbf{x};\widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(\mathbf{x};\widehat{\boldsymbol{\mu}})$ in Eq. (19) that

$$\Pr_{\mathbf{x}\sim\mathcal{N}_j}\Big[\mathbf{\Lambda}_{ii}(\mathbf{x};\widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(\mathbf{x};\widehat{\boldsymbol{\mu}}) - \langle \mathbf{Q}_j - \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle \le -(c^2+2c')\|\mathbf{Q}_j-\mathbf{Q}_i\|_F^2/\beta^2\Big] \lesssim \exp(-\Omega(s^2))\,. \tag{24}$$

It remains to bound $\langle \mathbf{Q}_j - \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle$. We have

$$\langle \mathbf{Q}_j - \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle \ge \langle \mathbf{Q}_j - \mathbf{Q}_i, \mathbf{Q}_i^{-1} - \mathbf{Q}_j^{-1}\rangle - (4\upsilon_{\mathrm{cov}}/\alpha^2)\cdot\|\mathbf{Q}_j - \mathbf{Q}_i\|_F\,. \tag{25}$$

Combining this with Eq. (24), we obtain the desired bound. ∎

Lemma 28 *Let $i \in [k]$. As in Theorem 27, suppose $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^d$ satisfies*

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_i\| \le c\Delta^{(\mathbf{Q})}_{\mathrm{out}}\sqrt{\alpha}/\beta \tag{26}$$

for sufficiently small absolute constant $c > 0$.

For any $s \ge 1$, with probability at least $1 - O(k)\cdot\exp(-\Omega(s))$ over $\mathbf{x}\sim\mathcal{N}_i$, we have that for all $j \in [k]$,

$$\mathbf{\Lambda}_{ii}(\mathbf{x};\widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(\mathbf{x};\widehat{\boldsymbol{\mu}}) \le (s\beta/\alpha^2)\cdot\{\|\mathbf{Q}_j - \mathbf{Q}_i\|_F \vee \upsilon_{\mathrm{cov}}\} + c^2(\Delta^{(\mathbf{Q})}_{\mathrm{out}})^2/\beta^2 + c\Delta^{(\mathbf{Q})}_{\mathrm{out}}\sqrt{s/\alpha\beta}\,.$$

Proof Define $\mathbf{B} \triangleq \mathbf{Q}_i^{1/2}(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)\mathbf{Q}_i^{1/2}$ and $\mathbf{w} \triangleq \mathbf{Q}_i^{1/2}(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})$ (note these are slightly different from $\mathbf{B}$ defined in Theorem 27 as $\mathbf{x}$ is sampled from $\mathcal{N}_i$ instead of $\mathcal{N}_j$). Then for $\mathbf{x}\sim\mathcal{N}_i$, writing this as $\mathbf{x} = \boldsymbol{\mu}_i + \mathbf{Q}_i^{1/2}\mathbf{z}$ for $\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})$, we see that the quantity $\mathbf{\Lambda}_{ii}(\mathbf{x};\widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(\mathbf{x};\widehat{\boldsymbol{\mu}})$ is distributed as

$$\mathbf{z}^\intercal\mathbf{B}\mathbf{z} - 2\langle\mathbf{z},\mathbf{w}\rangle + \langle(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})^\intercal - \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle\,. \tag{27}$$

Controlling $\mathbf{z}^\intercal\mathbf{B}\mathbf{z}$: Note that

$$\begin{aligned}\|\mathbf{B}\|_F &\le \|\mathbf{Q}_i^{1/2}\mathbf{Q}_i^{-1}(\mathbf{Q}_j - \mathbf{Q}_i)\mathbf{Q}_j^{-1}\mathbf{Q}_i^{1/2}\|_F + 4\beta\upsilon_{\mathrm{cov}}/\alpha^2\\ &\lesssim (\beta/\alpha^2)\cdot\{\|\mathbf{Q}_j - \mathbf{Q}_i\|_F \vee \upsilon_{\mathrm{cov}}\}\end{aligned}$$

By Theorem 19, we have

$$\Pr_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})}\Big[|\mathbf{z}^\intercal\mathbf{B}\mathbf{z} - \mathrm{tr}(\mathbf{Q}_i(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j))| \le (s\beta/\alpha^2)\cdot\{\|\mathbf{Q}_j - \mathbf{Q}_i\|_F \vee \upsilon_{\mathrm{cov}}\}\Big] \ge 1 - 2\exp(-\Omega(s))\,. \tag{28}$$

Controlling $|\langle\mathbf{z},\mathbf{w}\rangle|$: Note that $\|\widehat{\mathbf{K}}_i\|_{\mathrm{op}}, \|\widehat{\mathbf{K}}_j\|_{\mathrm{op}} \lesssim 1/\alpha$, so $\|\mathbf{w}\| \le c\Delta^{(\mathbf{Q})}_{\mathrm{out}}/\sqrt{\alpha\beta}$ by Eq. (26). By standard Gaussian tail bounds, we conclude that with probability at least $1 - \exp(-\Omega(s))$,

$$|\langle\mathbf{z},\mathbf{w}\rangle| \le c\Delta^{(\mathbf{Q})}_{\mathrm{out}}\sqrt{s/\alpha\beta}\,. \tag{29}$$

Controlling $\langle(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})^\intercal, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle$: As $\|\widehat{\mathbf{K}}_i\|_{\mathrm{op}}, \|\widehat{\mathbf{K}}_j\|_{\mathrm{op}} \lesssim 1/\alpha$, by Eq. (26) we have that

$$|\langle(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}})^\intercal, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle| \le c^2(\Delta^{(\mathbf{Q})}_{\mathrm{out}})^2/\beta^2\,. \tag{30}$$

Putting things together: Conditioned on the events of Eq. (28) and (29) holding, and also using the bound on the constant term in Eq. (30), we see from the decomposition of $\mathbf{\Lambda}_{ii}(\mathbf{x};\widehat{\boldsymbol{\mu}}) - \mathbf{\Lambda}_{ij}(\mathbf{x};\widehat{\boldsymbol{\mu}})$ that

$$\Pr_{\mathbf{x}\sim\mathcal{N}_i}\Big[|\mathbf{\Lambda}_{ii}(\mathbf{x};\widehat{\boldsymbol{\mu}})-\mathbf{\Lambda}_{ij}(\mathbf{x};\widehat{\boldsymbol{\mu}})| > (s\beta/\alpha^2)\cdot\{\|\mathbf{Q}_j-\mathbf{Q}_i\|_F\vee\upsilon_{\mathrm{cov}}\}+c^2(\Delta^{(\mathbf{Q})}_{\mathrm{out}})^2/\beta^2+c\sqrt{s/\alpha\beta}\Big] \lesssim \exp(-\Omega(s))\,.$$

The claimed bound follows by a union bound. ■

C.3. Formally defining the clustering

We are now ready to define our clustering function.

Let $(\mathcal{S} = \{S_1,\ldots,S_m\}, \mathcal{T} = \{T_1,\ldots,T_n\})$ denote a $(\Delta^{(\boldsymbol{\mu})}_{\mathrm{in}}, \Delta^{(\mathbf{Q})}_{\mathrm{in}}, \Delta^{(\boldsymbol{\mu})}_{\mathrm{out}}, \Delta^{(\mathbf{Q})}_{\mathrm{out}})$-separated partition of $[k]$. First, define

$$\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) \triangleq a \in [m] \text{ for which } \operatorname*{argmin}_{i\in[k]} \|\widehat{\boldsymbol{\mu}}_i - \widehat{\mathbf{\Pi}}\mathbf{x}\| \in S_a\,,$$

where $\widehat{\mathbf{\Pi}}$ is the projector to the span of $\widehat{\boldsymbol{\mu}}_1,\ldots,\widehat{\boldsymbol{\mu}}_k$.

The following is a slight modification of Theorem 18:

Lemma 29 *Suppose that $\Delta^{(\boldsymbol{\mu})}_{\mathrm{out}} \ge \max(6\sqrt{\upsilon_{\mathrm{mean}}}, 6\sqrt{k\beta})$. Then for any $i \in S_a$ and $a' \neq a$,*

$$\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a'] \le \exp\Big(-\Omega\Big(\frac{1}{\alpha\sqrt{k}}\min_{i'\in S_{a'}}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}\|^2\Big)\Big)\,.$$

Equivalently,

$$\Pr_{\mathcal{N}_i}[\widehat{\boldsymbol{\mu}}(\mathbf{x}) \in \{\widehat{\boldsymbol{\mu}}_{i'} : i' \in S_{a'}\}] \le \exp\Big(-\Omega\Big(\frac{1}{\alpha\sqrt{k}}\min_{i'\in S_{a'}}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}\|^2\Big)\Big)\,.$$

Proof Note that $\mathrm{tr}(\mathbf{Q}_i\widehat{\mathbf{\Pi}}) \le k\beta$ and $\|\mathbf{Q}_i^{1/2}\widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\|_F^2 \ge k\alpha^2$, so for $\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})$, by Theorem 19 with r therein taken to be 1, for all $s > 0$ we have

$$\Pr[\|\widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\mathbf{z}\|^2 > k\beta + s\alpha\sqrt{k}] \le \exp(-\Omega(s))\,.$$

Given $\mathbf{x}\sim\mathcal{N}_i$, note that $\widehat{\boldsymbol{\mu}}_i - \widehat{\mathbf{\Pi}}\mathbf{x} = \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + \widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\mathbf{z}$ for $\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})$. Thus, conditioned on the above event,

$$\|\widehat{\boldsymbol{\mu}}_i - \widehat{\mathbf{\Pi}}\mathbf{x}\| \le \sqrt{\upsilon_{\mathrm{mean}}} + \sqrt{k\beta} + k^{1/4}\sqrt{\alpha s}\,.$$

Next, for any $i' \notin S_a$ and $\mathbf{x}\sim\mathcal{N}_i$, note that $\widehat{\boldsymbol{\mu}}_{i'} - \widehat{\mathbf{\Pi}}\mathbf{x} = \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_{i'} - \widehat{\boldsymbol{\mu}}_i) + \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + \widehat{\mathbf{\Pi}}\mathbf{Q}_i^{1/2}\mathbf{z}$ for $\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})$. We have

$$\|\widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_{i'} - \widehat{\boldsymbol{\mu}}_i) + \widehat{\mathbf{\Pi}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\| \ge \|\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_i\| - 3\sqrt{\upsilon_{\mathrm{mean}}} \ge \frac{1}{2}\|\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_i\|\,,$$

where in the last step we used that $\Delta^{(\boldsymbol{\mu})}_{\mathrm{out}} \ge 6\sqrt{\upsilon_{\mathrm{mean}}}$. Thus, conditioned on the above event,

$$\|\widehat{\boldsymbol{\mu}}_{i'} - \widehat{\mathbf{\Pi}}\mathbf{x}\| \ge \frac{1}{2}\|\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_i\| - \sqrt{k\beta} - k^{1/4}\sqrt{\alpha s}\,.$$

Provided that $s > (\frac{1}{2}\|\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_i\| - \sqrt{\upsilon_{\mathrm{mean}}} - \sqrt{k\beta})^2/\alpha\sqrt{k}$, we have that $\|\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\Pi}}\mathbf{x}\| < \|\widehat{\boldsymbol{\mu}}_{i'} - \widehat{\boldsymbol{\Pi}}\mathbf{x}\|$. As $\sqrt{\upsilon_{\mathrm{mean}}} \le \frac{1}{6}\Delta^{(\boldsymbol{\mu})}_{out}$ and $\sqrt{k\beta} \le \frac{1}{6}\Delta^{(\boldsymbol{\mu})}_{out}$, it suffices to take $s = \frac{\|\boldsymbol{\mu}_{i'} - \boldsymbol{\mu}_i\|^2}{36\alpha\sqrt{k}}$.

The second part of the Lemma follows by definition of $\widehat{\boldsymbol{\mu}}(\mathbf{x})$. ■

Define $\mathsf{c}^{(\mathbf{Q})}(\mathbf{x})$ as follows. First note that we can't directly use $\boldsymbol{\Lambda}_{ii} - \boldsymbol{\Lambda}_{ij}$ as it has a term $\langle \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle$ which depends on the true covariance $\mathbf{Q}_i$. Likewise, the lower and upper bounds on $\boldsymbol{\Lambda}_{ii} - \boldsymbol{\Lambda}_{ij}$ in Theorem 27 and Theorem 28 depend on the true covariances $\mathbf{Q}_i, \mathbf{Q}_j$.

Instead, we will brute force over guesses for these quantities. Henceforth, suppose we have access to numbers $\{t_{ij}\}$ satisfying

$$\left|t_{ij} - (\langle \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle + \langle \mathbf{Q}_j - \mathbf{Q}_i, \mathbf{Q}_i^{-1} - \mathbf{Q}_j^{-1}\rangle - E)\right| \le \eta$$

for sufficiently small parameter η, where E is the error term from Theorem 27. Because

$$|\langle \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j\rangle + \langle \mathbf{Q}_j - \mathbf{Q}_i, \mathbf{Q}_i^{-1} - \mathbf{Q}_j^{-1}\rangle - E| \lesssim \beta d/\alpha + \upsilon_{cov}\beta\sqrt{d}/\alpha^2 \lesssim \beta d/\alpha\,,$$

we can produce these numbers by brute-forcing over a grid of size $(\beta d/\alpha\eta)^{O(k^2)}$. We will eventually take

$$\eta = \frac{\Delta^{(\mathbf{Q})}_{out}}{100\beta^2}\,. \tag{31}$$

With these $\{t_{ij}\}$ in hand, given an index $\ell \in [n]$ into the partition $\{T_1, \ldots, T_n\}$, we define $\mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = b$ if there exists some $i \in T_b$ such that

$$(\mathbf{x} - \widehat{\boldsymbol{\mu}}(\mathbf{x}))^\intercal(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)(\mathbf{x} - \widehat{\boldsymbol{\mu}}(\mathbf{x})) < t_{ij} - \eta$$

for all $j \notin T_b$. If there exist multiple such b for which this is the case, then choose one arbitrarily. If no such b exists, then set $\mathsf{c}^{(\mathbf{Q})}(\mathbf{x})$ to be 0.

Corollary 30 *For any $i \in S_a \cap T_b$ and nonzero $b' \neq b$, we have that*

$$\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = b' \mid \mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a] \le 2k^2\exp(-\Omega(c'^2(\alpha^4/\beta^6)\cdot \min_{j\in T_{b'}}\|\mathbf{Q}_j - \mathbf{Q}_i\|_F^2 \cdot \min(1, \alpha^2/\upsilon_{\mathrm{cov}}^2)))\,.$$

Proof We can rewrite the conditional probability as

$$\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a]^{-1}\cdot \Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = b'] \le 2\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = b']\,,$$

where we used Theorem 29 and the fact $k\cdot\exp(-\Omega((\Delta^{(\boldsymbol{\mu})}_{out})^2/\alpha\sqrt{k})) \le 1/2$. Note that

$$\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = b'] = \sum_{i'\in S_a}\Pr_{\mathcal{N}_i}[\widehat{\boldsymbol{\mu}}(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_{i'} \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = b'] \tag{32}$$

$$\le \sum_{i'\in S_a}\sum_{j\in T_{b'}}\Pr_{\mathcal{N}_i}[(\mathbf{x} - \widehat{\boldsymbol{\mu}}_{i'})^\intercal(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_j)(\mathbf{x} - \widehat{\boldsymbol{\mu}}_{i'}) < t_{jj'} - \eta\;\forall\, j' \in [k]] \tag{33}$$

$$\le \sum_{i'\in S_a}\sum_{j\in T_{b'}}\Pr_{\mathcal{N}_i}[(\mathbf{x} - \widehat{\boldsymbol{\mu}}_{i'})^\intercal(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_i)(\mathbf{x} - \widehat{\boldsymbol{\mu}}_{i'}) < t_{ji} - \eta] \tag{34}$$

$$\le k^2\exp(-\Omega(c'^2(\alpha^4/\beta^6)\cdot \min_{j\in T_{b'}}\|\mathbf{Q}_j - \mathbf{Q}_i\|_F^2 \cdot \min(1, \alpha^2/\upsilon_{cov}^2)))\,, \tag{35}$$

where in the last step we used Theorem 27. ■

Corollary 31 *Suppose that*

$$\Delta_{out}^{(\mathbf{Q})} \geq C \max(\upsilon_{cov}\beta^2/\alpha^2, c^{2/3}(\Delta_{out}^{(\mathbf{Q})})^{2/3}\alpha^{1/3}, (\beta/\alpha)^3\upsilon_{cov}) \tag{36}$$

for sufficiently large absolute constant $C > 0$. Then for any $i \in S_a \cap T_b$, we have that

$$\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = 0 \mid \mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a] \leq 2k^3 \exp(-\Omega(\alpha^2 \Delta_{\mathrm{out}}^{(\mathbf{Q})}/\beta^3))\,.$$

Proof We can rewrite the conditional probability as

$$\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a]^{-1} \cdot \Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = 0] \leq 2 \Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = 0]\,,$$

where we used Theorem 29 and the fact $k \cdot \exp(-\Omega((\Delta_{out}^{(\boldsymbol{\mu})})^2/\alpha\sqrt{k})) \leq 1/2$. Note that

$$\begin{aligned}\Pr_{\mathcal{N}_i}[\mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x}) = a \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = 0] &= \sum_{i' \in S_a} \Pr_{\mathcal{N}_i}[\widehat{\boldsymbol{\mu}}(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_{i'} \text{ and } \mathsf{c}^{(\mathbf{Q})}(\mathbf{x}) = 0] && (37)\\ &\leq \sum_{i' \in S_a} \sum_{j \notin T_b} \Pr_{\mathcal{N}_i}[(\mathbf{x} - \widehat{\boldsymbol{\mu}}_{i'})^\intercal (\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)(\mathbf{x} - \widehat{\boldsymbol{\mu}}_{i'}) \geq t_{ij} - \eta] && (38)\end{aligned}$$

We wish to apply Theorem 28 here. Consider any $j \notin T_b$. Note that

$$\begin{aligned} t_{ij} - \eta - \langle \mathbf{Q}_i, \widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j \rangle &\geq \langle \mathbf{Q}_j - \mathbf{Q}_i, \mathbf{Q}_i^{-1} - \mathbf{Q}_j^{-1} \rangle - 2\eta - E \\ &\geq \operatorname{tr}((\mathbf{Q}_j - \mathbf{Q}_i)\mathbf{Q}_i^{-1}(\mathbf{Q}_j - \mathbf{Q}_i)\mathbf{Q}_j^{-1}) - 2\eta - E \\ &\geq (1/\beta^2) \cdot \|\mathbf{Q}_j - \mathbf{Q}_i\|_F^2 - 2\eta - E\,. \end{aligned}$$

In Theorem 28, take $s = (\alpha^2/\beta^3) \cdot \|\mathbf{Q}_j - \mathbf{Q}_i\|_F$. Then we can bound the above by

$$\geq (s\beta/\alpha^2) \cdot \{\|\mathbf{Q}_j - \mathbf{Q}_i\|_F \vee \upsilon_{cov}\} + c^2(\Delta_{out}^{(\mathbf{Q})})^2/\beta^2 + c\Delta_{out}^{(\mathbf{Q})}\sqrt{s/\alpha\beta}\,.$$

By Theorem 28, this happens with probability at most $O(k) \cdot \exp(-\Omega(s))$. There are at most k^2 terms in the sum in Eq. (38), so the claimed bound follows by a union bound. ■

We can now immediately conclude the proof of the main result of this section:

Proof [Proof of Theorem 25] Define $\mathsf{c}(\mathbf{x})$ as follows. Let $a = \mathsf{c}^{(\boldsymbol{\mu})}(\mathbf{x})$ and $b = \mathsf{c}^{(\mathbf{Q})}(\mathbf{x})$). If $b = 0$, or S_a and T_b do not intersect, then define $\mathsf{c}(\mathbf{x})$ arbitrarily. Otherwise, if they do intersect, let U_t denote the element of the common refinement of $\mathcal{S}$ and $\mathcal{T}$ corresponding to $S_a \cap T_b$, and define $\mathsf{c}(\mathbf{x}) = t$.

The bound on the misclassification error then follows from Theorem 29, Theorem 30, and Theorem 31, noting that the condition of Eq. (16) ensures that the hypotheses of these components are met. ■

For convenience, we summarize $\mathsf{c}(\mathbf{x})$ in Algorithm 5 below.

Algorithm 5: CLUSTERING

Input: Partitions $\mathcal{S} = \{S_1, \dots, S_m\}, \mathcal{T} = \{T_1, \dots, T_n\}$ of $[k]$, estimates $\{(\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i)\}$, thresholds $\{t_{ij}\}$

Output: Clustering function $\mathsf{c} : \mathbb{R}^d \to [n_c]$

$\eta \leftarrow \Delta_{out}^{(\mathbf{Q})}/100\beta^2$.

Let $U_1, \dots, U_{n_c}$ denote the common refinement of the partitions $\mathcal{S}, \mathcal{T}$.

Let $\widehat{\mathbf{\Pi}}$ denote the projector to the span of $\widehat{\boldsymbol{\mu}}_1, \dots, \widehat{\boldsymbol{\mu}}_k$.

Define $c^{(\boldsymbol{\mu})}(\mathbf{x})$ to be the index a of the piece S_a of $\mathcal{S}$ containing $\mathrm{argmin}_{i\in[k]} \|\widehat{\boldsymbol{\mu}}_i - \widehat{\mathbf{\Pi}}\mathbf{x}\|$.

Define $\widehat{\boldsymbol{\mu}}(\mathbf{x})$ to be $\widehat{\boldsymbol{\mu}}_i$ for $i = \mathrm{argmin}_{j\in[k]} \|\widehat{\boldsymbol{\mu}}_j - \widehat{\mathbf{\Pi}}\mathbf{x}\|$.

Define $c^{(\mathbf{Q})}(\mathbf{x})$ to be the index b if there exists $i \in T_b$ such that $(\mathbf{x} - \widehat{\boldsymbol{\mu}}(\mathbf{x}))^\intercal(\widehat{\mathbf{K}}_i - \widehat{\mathbf{K}}_j)(\mathbf{x} - \widehat{\boldsymbol{\mu}}(\mathbf{x})) < t_{ij} - \eta$ for all $j \notin T_b$.

if $b = 0$ *or* $S_a \cap T_b = \emptyset$ **then**

  Define $\mathsf{c}(\mathbf{x})$ arbitrarily.

else

  Let U_t denote an element of the common refinement corresponding to $S_a \cap T_b$.

  return $\mathsf{c}(\mathbf{x}) = t$.

end

Appendix D. Score simplification

The main difficulty in providing a polynomial approximation of the score function arises when it involves multiple Gaussians that are far apart. Without further structural assumptions about the function and/or the underlying measure, the degree of the polynomial approximation depends on (1) the smoothness properties of the target function (e.g., Lipschitz constant or higher-order derivative bounds) and (2) the radius of the support over which the polynomial is guaranteed to be close to the target function.

Recall that the score function of a mixture $\mathcal{M}$ of k Gaussian distributions with means $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$ and covariances $\mathbf{Q}_1, \dots, \mathbf{Q}_k$ is given by

$$\mathbf{s}(\mathbf{x}; \mathcal{M}) = -\sum_{i=1}^{k} w_i(\mathbf{x})\mathbf{Q}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \quad \text{where} \quad w_i(\mathbf{x}) = \frac{\lambda_i \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i; \mathbf{x})}{\sum_{j=1}^{k} \lambda_j \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i; \mathbf{x})}\,.$$

For simplicity, in what follows we will denote by $\mathcal{N}_i$ the i-th component of the above mixture, $\mathcal{N}_i = \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{Q}_i)$. For Gaussian mixtures, the effective support of the score function is roughly proportional to the radius of the parameter space which scales with the dimension and the parameter distance $\mathrm{poly}(d, R)$. This is the case as we consider a mixture over d-dimensional Gaussians with mean and covariances bounded (in parameter distance) by R. Moreover, the Lipschitz constant of the score function can also scale as $\mathrm{poly}(d, R)$. Therefore, applying black-box polynomial approximation results (such as Jackson's theorem – see Theorem 41) would yield a polynomial of degree at least polynomial in the dimension d and the parameter radius R yielding a trivial (exponential) runtime. Instead of using the polynomial approximation results in a black-box manner, we will be constructing a piecewise polynomial approximation of the score function where the partition is given by the clustering algorithm we designed in Appendix C.

In this section, we show that given the "rough" clustering function of Appendix C we can simplify the score function inside each cell of the partition given by the clustering so that it is possible to prove the existence of a low-degree approximation inside each cell. More precisely, we require that the clustering function $\mathsf{c}(\mathbf{x})$ assigns each $\mathbf{x} \in \mathbb{R}^d$ to one of n_c subsets $U_1, \dots, U_{n_c}$ of $[k]$ that form a partition of the original k components such that if $\mathcal{N}_i, \mathcal{N}_j$ belong in different subsets U_t and $U_{t'}$ have to be at least $\mathrm{poly}(\beta/\alpha) \cdot \log(k/\varepsilon)$ far in parameter distance. In other words, we require that components in different subsets of the partition have to be sufficiently separated. Moreover, for every $i \notin U_t$, we require that the clustering function c incorrectly classifies a sample $\mathbf{x} \sim \mathcal{N}_i$ as belonging to U_t with probability at most ε. Under those assumptions, we show that for any given $\mathsf{c}(\mathbf{x}) = t$, we can "simplify" the score function by removing the contribution of all components $\mathcal{N}_j$ that do not belong in U_t.

In what follows, given a subset U_t of indices of $[k]$ we denote by $\mathcal{M}(U_t)$ the submixture containing the components $\mathcal{N}_i$ for $i \in U_t$ and by $\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))$ the score function containing only the contribution of components from U_t, i.e.,

$$\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) = \sum_{i \in U_t} \lambda_i \mathbf{g}_i(\mathbf{x}) \frac{\mathcal{N}_i(\mathbf{x})}{\sum_{j \in U_t} \lambda_j \mathcal{N}_j(\mathbf{x})}$$

The main result of this section is the following proposition showing that, inside each cell t of the partition given by $\mathsf{c}(\cdot)$, we can replace the original score function $\mathbf{s}(\mathbf{x}; \mathcal{M})$ by the score function of the sub-mixture $\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))$.

Proposition 32 (Score Simplification) *Fix $\varepsilon > 0$ and let $\mathcal{M}$ be a mixture of k Gaussian distributions $\mathcal{N}_1, \dots, \mathcal{N}_k$ with mean and covariances $\boldsymbol{\mu}_i, \mathbf{Q}_i$ such that for every pair i, j $D_p(\mathcal{N}_i, \mathcal{N}_j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 + \|\mathbf{Q}_i - \mathbf{Q}_j\|_F^2 \leq R$ for some $R > 1$ Moreover, assume that for some $\alpha \leq 1 \leq \beta$ it holds that $\alpha \mathbf{Id} \preceq \mathbf{Q}_i \preceq \beta \mathbf{Id}$ for all $i \in [k]$ for $\alpha \leq 1 \leq \beta$.*

1. *Let $n_c \in [k]$ and let $U_1, \dots, U_{n_c}$ be a partition of $[k]$ such that for every $i \in U_t$, and $j \notin U_t$ it holds that $D_p(\mathcal{N}_i, \mathcal{N}_j)$ is larger than a sufficiently large absolute constant multiple of $\beta^4/\alpha^2 \log(k\beta/(\alpha\varepsilon))$.*
2. *Assume that $\mathsf{c} : \mathbb{R}^d \mapsto [n_c]$ is a ε-approximate clustering function, i.e., $\mathbb{P}_{\mathbf{x} \sim \mathcal{N}_i}[\mathsf{c}(\mathbf{x}) = t] \leq \varepsilon$ for all $t \in [n_c]$ and $i \notin U_t$.*

Define the following piecewise approximation to the score function

$$s(\mathbf{x}; \mathsf{c}(\cdot)) = \sum_{t=1}^{n_c} s(\mathbf{x}; \mathcal{M}(U_t)) \, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\} \,.$$

It holds that

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}}[\|s(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}; \mathsf{c}(\cdot))\|_2^2] \lesssim k^{5/4} R \frac{\beta^5}{\alpha^6} \sqrt{\varepsilon}.$$

Proof We first observe that since $\sum_{t=1}^{n_c} \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\} = 1$ for all $\mathbf{x}$ (i.e., each point x is only assigned to a single set U_t), we can write $s(\mathbf{x}) = \sum_{t=1}^{n_c} s(\mathbf{x}) \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}$ and therefore, we have that

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}}[\|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}; \mathsf{c}(\cdot))\|_2^2] = \sum_{t=1}^{n_c} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}}[\|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))\|_2^2 \, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}] \,.$$

We break down the total L_2^2 error into the case where $\mathbf{x}$ was actually generated by a mixture component that belongs to the set U_t (as predicted by the clustering function $\mathsf{c}(\mathbf{x})$) and the case where $\mathbf{x}$ was generated by some mixture component that is not in U_t. Recall that we denote by $\mathcal{M}^J$ the joint density of the indexed pair $(i, \mathbf{x})$ where i corresponds to the index of the mixture component that generates $\mathbf{x}$. We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}[\|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))\|_2^2 \, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}] && (39)\\
&\quad = \mathop{\mathbb{E}}_{(i,\mathbf{x})\sim\mathcal{M}^J}[\|s(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}; \mathcal{M}(U_t))\|^2 \, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t, i \in U_t\}] && (40)\\
&\quad + \mathop{\mathbb{E}}_{(i,\mathbf{x})\sim\mathcal{M}^J}[\|s(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}; \mathcal{M}(U_t))\|^2 \, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t, i \notin U_t\}] \,. && (41)
\end{aligned}$$

We first focus on the first part of the error, i.e., when the example $\mathbf{x}$ is generated by some component $\mathcal{N}_i$ that belongs to the set U_t. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{(i,\mathbf{x})\sim\mathcal{M}^J}[\|s(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}; \mathcal{M}(U_t))\|^2 \, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t, i \in U_t\}] &\le \sum_{i\in U_t} \lambda_i \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|s(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}; \mathcal{M}(U_t))\|^2] \\
&\le \sum_{i\in U_t} \lambda_i \sqrt{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|s(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}; \mathcal{M}(U_t))\|^4]},
\end{aligned}$$

where the last inequality follows by Jensen's.

We show that as long as a component $\mathcal{N}_j$ that we remove is far from the component $i \in U_t$ in parameter distance, their removal induces an exponentially small error in the score function.

Lemma 33 *Let $\mathcal{N}_1, \ldots, \mathcal{N}_k$ be Normal distributions with means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and covariances $\mathbf{Q}_1, \ldots, \mathbf{Q}_k$ such that for all i, $\alpha\mathbf{Id} \le \mathbf{Q}_i \le \beta\mathbf{Id}$. For any $i \in U_t$, it holds that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x}; \mathcal{M}) - \mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))\|_2^4] \lesssim \frac{k\beta^{10}}{\sqrt{\lambda_i}\alpha^{12}} \sum_{j\notin U_t} \exp\left(-c\frac{\alpha^2}{\beta^4} D_p(\mathcal{N}_i, \mathcal{N}_j)\right) .$$

for some universal constant $c > 0$. Moreover if $i \notin U_t$ it holds that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x}; \mathcal{M}) - \mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))\|_2^4] \\
&\quad\lesssim \frac{\beta^2}{\alpha^8} \sum_{\ell=1,\ell\ne i}^{k} (D_p(\mathcal{N}_i, \mathcal{N}_\ell)^2 + D_p(\mathcal{N}_i, \mathcal{N}_\ell)) + \sum_{j\notin U_t, j\ne i} \frac{k\beta^{10}}{\sqrt{\lambda_i}\alpha^{12}} \exp\left(-c\frac{\alpha^2}{\beta^4} D_p(\mathcal{N}_i, \mathcal{N}_j)\right) .
\end{aligned}$$

Using Theorem 33 we obtain that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(i,\mathbf{x})\sim\mathcal{M}^J}[\|\mathbf{s}(\mathbf{x}; \mathcal{M}) - \mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\} \mid i \in U_t] \\
&\quad\le \frac{1}{\sqrt{\sum_{i\in U_t} \lambda_i}} \sum_{i\in U_t} \lambda_i \sqrt{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x}; \mathcal{M}) - \mathbf{s}(\mathbf{x}; \mathcal{M}(U_t))\|^4]} \\
&\quad\lesssim \frac{\sqrt{k}\beta^5}{\alpha^6} \frac{\sum_{i\in U_t} \lambda_i^{3/4}}{\sqrt{\sum_{i\in U_t} \lambda_i}} \max_{j\notin U_t} e^{-c\frac{\alpha^2}{\beta^4} D_p(\mathcal{N}_i, \mathcal{N}_j)} \lesssim \frac{k^{3/4}\beta^5}{\alpha^6} \max_{j\notin U_t} e^{-c\frac{\alpha^2}{\beta^4} D_p(\mathcal{N}_i, \mathcal{N}_j)},
\end{aligned}$$

where the last inequality follows from the fact that $\sum_{i\in U_t}\lambda_i^{3/4} \le |U_t|^{1/4}(\sum_{i\in U_t}\lambda_i)^{3/4} \le k^{1/4}(\sum_{i\in U_t}\lambda_i)^{3/4}$. Therefore, using this estimate we obtain that in the case where the sample is generated by some component in U_t, the error is

$$\sum_{t=1}^{n_c} \frac{k^{3/4}\beta^5}{\alpha^6} \max_{j\notin U_t} e^{-c\frac{\alpha^2}{\beta^4}D_p(\mathcal{N}_i,\mathcal{N}_j)} \le \frac{k^{7/4}\beta^5}{\alpha^6} e^{-c\frac{\alpha^2}{\beta^4}\Delta_{out}} .$$

We next bound the error in the difference of the score functions when the clustering function makes a mistake, i.e., $\mathsf{c}(\mathbf{x}) = t$ but $\mathbf{x}$ is generated by $\mathcal{N}_i$ for $i \notin U_t$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(i,\mathbf{x})\sim\mathcal{M}^J}[\|\mathbf{s}(\mathbf{x};\mathcal{M}) - \mathbf{s}(\mathbf{x};\mathcal{M}(U_t))\|^2\, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t, i\notin U_t\}] \\
&\qquad = \sum_{i\notin U_t} \lambda_i \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x};\mathcal{M}) - \mathbf{s}(\mathbf{x};\mathcal{M}(U_t))\|^2\, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}] \\
&\qquad \le \sum_{i\notin U_t} \lambda_i \sqrt{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x};\mathcal{M}) - \mathbf{s}(\mathbf{x};\mathcal{M}(U_t))\|^4]}\sqrt{\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{N}_i}[\mathsf{c}(\mathbf{x}) = t]} \\
&\qquad \le \sqrt{2\varepsilon} \sum_{i\notin U_t} \lambda_i \left(\frac{\beta}{\alpha^4}\sqrt{\sum_{\ell=1,\ell\neq i}^{k} D_p(\mathcal{N}_i,\mathcal{N}_\ell)^2 + D_p(\mathcal{N}_i,\mathcal{N}_\ell)} + \frac{\sqrt{k}\beta^5}{\lambda_i^{1/4}\alpha^6}\sqrt{\sum_{j\notin U_t, j\neq i} e^{-c\frac{\alpha^2}{\beta^4}D_p(\mathcal{N}_i,\mathcal{N}_j)}} \right) \\
&\qquad \lesssim \sqrt{\varepsilon} \left(\frac{\beta}{\alpha^4} \max_{i\notin U_t} \sum_{\ell=1,\ell\neq i}^{k} (D_p(\mathcal{N}_i,\mathcal{N}_\ell) + \sqrt{D_p(\mathcal{N}_i,\mathcal{N}_\ell)}) + \frac{k^{5/4}\beta^5}{\alpha^6} \right),
\end{aligned}$$

where for the third step we used the fact that by our assumption it holds that $\mathbb{P}_{\mathbf{x}\sim\mathcal{N}_i}[\mathsf{c}(\mathbf{x}) = t] \le \varepsilon$ when $i \notin U_t$ and for the last inequality we used the fact that there are at most k elements that do not belong in U_t and, similarly to the previous derivation, the fact that $\sum_{i\in U_t}\lambda_i^{3/4} \le |U_t|^{1/4}(\sum_{i\in U_t}\lambda_i)^{3/4} \le k^{1/4}(\sum_{i\in U_t}\lambda_i)^{3/4}$. ■

D.1. Proof of Lemma 33

We first show the following lemma capturing the effect of removing a single component from the score function. We show that the induced error is exponentially small in the distance of the removed component j and the component i.

Lemma 34 *Let $\mathcal{N}_1,\dots,\mathcal{N}_k$ be Normal distributions with means $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_k$ and covariances $\mathbf{Q}_1,\dots,\mathbf{Q}_k$ such that for all i $\alpha\mathbf{Id} \preceq \mathbf{Q}_i \preceq \beta\mathbf{Id}$ for some $\alpha \le 1 \le \beta$. Let $\mathcal{M}$ be the mixture of $\mathcal{N}_1,\dots,\mathcal{N}_k$ with weights $\lambda_1,\dots,\lambda_k$. Let $c > 0$ be some universal constant. For all $i \neq j$, it holds that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x};\mathcal{M}) - \mathbf{s}^{-j}(\mathbf{x})\|_2^4] \lesssim \frac{k\beta^{10}}{\sqrt{\lambda_i}\alpha^{12}} \exp\left(-c\frac{\alpha^2}{\beta^4} D_p(\mathcal{N}_i,\mathcal{N}_j)\right),$$

where $\mathbf{s}^{-j}(\mathbf{x}) = \mathbf{s}(\mathbf{x};\mathcal{M}([k]\setminus j))$ is the score function of the mixture after we drop the contribution of component j. Moreover, it holds $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}_j}[\|\mathbf{s}(\mathbf{x}) - \mathbf{s}^{-j}(\mathbf{x})\|_2^4] \lesssim \frac{\beta^2}{\alpha^8}\sum_{\ell=1,\ell\neq j}^{k}(D_p(\mathcal{N}_j,\mathcal{N}_\ell)^2 + D_p(\mathcal{N}_j,\mathcal{N}_\ell))$.

By iteratively applying Theorem 34, and the (almost) triangle inequality $\|\mathbf{a}+\mathbf{b}\|_2^4 \le 8\|\mathbf{a}\|_2^4 + 8\|\mathbf{b}\|_2^4$ we can remove all the components that do not belong in the set U_t and obtain the error guarantee of Theorem 33.

Proof [Proof Theorem 34] We first show the following claim bounding the gap between the original score function and the version where we drop the contribution of a component. We remark that the following claim is a pointwise fact about the score function and holds for every $\mathbf{x} \in \mathbb{R}^d$.

Claim 35 (Softmax Simplification) *Moreover let $D_1, \ldots, D_k$ be non-negative weight functions on $\mathbb{R}^d$ and $\mathbf{g}_1, \ldots, \mathbf{g}_k$ be functions $\mathbf{g}_i : \mathbb{R}^d \mapsto \mathbb{R}^d$. Define $\mathbf{s}(\mathbf{x}) = \sum_{i=1}^k \mathbf{g}_i(\mathbf{x}) D_i(\mathbf{x}) / (\sum_{i=1}^k D_i(\mathbf{x}))$ and*

$$\mathbf{s}^{-j}(\mathbf{x}) = \sum_{i=1, i\neq j}^{k} \mathbf{g}_i(\mathbf{x}) D_i(\mathbf{x}) / \Big(\sum_{i=1, i\neq j}^{k} D_i(\mathbf{x}) \Big).$$

For every $i = 1, \ldots, k$, it holds that

$$\|\mathbf{s}(\mathbf{x}) - \mathbf{s}^{-j}(\mathbf{x})\|_2^4 \le 8 \sum_{\ell=1, \ell\neq j}^{k} \Big(\frac{D_j(\mathbf{x})}{A(\mathbf{x})}\Big)\Big(\frac{D_\ell(\mathbf{x})}{B(\mathbf{x})}\Big) \|\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2^4 + 8\Big(\frac{D_j(\mathbf{x})}{A(\mathbf{x})}\Big)\|\mathbf{g}_j(\mathbf{x}) - \mathbf{g}_i(\mathbf{x})\|_2^4,$$

where we denote by $A(\mathbf{x}) = \sum_{i=1}^k D_i(\mathbf{x})$ and $B(\mathbf{x}) = \sum_{i=1, i\neq j}^k D_i(\mathbf{x})$.

Proof By a direct computation, we observe that

$$\mathbf{s}(\mathbf{x}) - \mathbf{s}^{-j}(\mathbf{x}) = \frac{D_j(\mathbf{x})}{A(\mathbf{x})} \left(\mathbf{g}_j(\mathbf{x}) - \sum_{\ell=1, \ell\neq j}^{k} \mathbf{g}_\ell(\mathbf{x}) \frac{D_\ell(\mathbf{x})}{B(\mathbf{x})} \right).$$

Adding and subtracting $\mathbf{g}_i$, we obtain that the above expression is equal to

$$\frac{D_j(\mathbf{x})}{A(\mathbf{x})} \left(\mathbf{g}_j(\mathbf{x}) - \mathbf{g}_i(\mathbf{x}) + \sum_{\ell=1, \ell\neq j}^{k} \frac{D_\ell(\mathbf{x})}{B(\mathbf{x})} (\mathbf{g}_\ell(\mathbf{x}) - \mathbf{g}_i(\mathbf{x})) \right).$$

We observe that the normalized weights $D_\ell(\mathbf{x})/B(\mathbf{x})$ form a distribution over $\ell \in [k] \setminus j$ and therefore, using Jensen's inequality, we obtain that

$$\Big\| \sum_{\ell=1, \ell\neq j}^{k} \frac{D_\ell(\mathbf{x})}{B(\mathbf{x})} (\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})) \Big\|_2^4 \le \sum_{\ell=1, \ell\neq j}^{k} \frac{D_\ell(\mathbf{x})}{B(\mathbf{x})} \|\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x}))\|_2^4.$$

Combining the above we obtain the following upper bound for the ℓ_2 error induced in the score function when we remove the contribution of the j-th component. We use the fact that $\|\mathbf{a}+\mathbf{b}\|_2^4 \le 8\|\mathbf{a}\|_2^4 + 8\|\mathbf{b}\|_2^4$ to obtain:

$$\begin{aligned}
\|\mathbf{s}(\mathbf{x}) - \mathbf{s}^{-j}(\mathbf{x})\|_2^4 &\le \Big(\frac{D_j(\mathbf{x})}{A(\mathbf{x})}\Big)^4 \Big(8 \Big\| \sum_{\ell=1, \ell\neq j}^{k} \frac{D_\ell(\mathbf{x})}{B(\mathbf{x})} (\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})) \Big\|_2^4 + 8\|\mathbf{g}_j(\mathbf{x}) - \mathbf{g}_i(\mathbf{x})\|_2^4 \Big) \\
&\le 8 \sum_{\ell=1, \ell\neq j}^{k} \Big(\frac{D_j(\mathbf{x})}{A(\mathbf{x})}\Big)\Big(\frac{D_\ell(\mathbf{x})}{B(\mathbf{x})}\Big) \|\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2^4 + 8\Big(\frac{D_j(\mathbf{x})}{A(\mathbf{x})}\Big)\|\mathbf{g}_j(\mathbf{x}) - \mathbf{g}_i(\mathbf{x})\|_2^4,
\end{aligned}$$

where for the last inequality we used the fact that $D_j(\mathbf{x})/A(\mathbf{x}) \leq 1$ for all $\mathbf{x}$ and Jensen's inequality, since $D_\ell(\mathbf{x})/B(\mathbf{x})$ is a distribution over $\ell \neq j$ and $\|\cdot\|_2^4$ is convex. ■

Using Theorem 35, with D corresponding to the component $\mathcal{N}_i$ in the statement of Theorem 34, we obtain that we have to control the terms

$$A^{(i,j,\ell)} = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\Big(\frac{\lambda_j\mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\Big)\Big(\frac{\lambda_\ell\mathcal{N}_\ell(\mathbf{x})}{S^{-j}(\mathbf{x})}\Big)\|\mathbf{g}_i(\mathbf{x})-\mathbf{g}_\ell(\mathbf{x})\|_2^4\Big]\,, \tag{42}$$

where $S(\mathbf{x}) = \sum_{s=1}^k \lambda_s\mathcal{N}_s(\mathbf{x})$ and $S^{-j}(\mathbf{x}) = S(\mathbf{x}) - \lambda_j\mathcal{N}_j(\mathbf{x})$. Moreover, we have to control the term

$$B^{(i,j)} = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\Big(\frac{\lambda_j\mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\Big)\|\mathbf{g}_j(\mathbf{x})-\mathbf{g}_i(\mathbf{x})\|_2^4\Big]\,. \tag{43}$$

Using the above notation, and Theorem 35, we obtain that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x})-\mathbf{s}^{-j}(\mathbf{x})\|_2^2] \leq 8B^{(i,j)} + 8\sum_{\ell=1,\ell\neq j}^{k} A^{(i,j,\ell)}\,. \tag{44}$$

We first bound the term $B^{(i,j)}$. By Cauchy-Schwarz we have

$$\begin{aligned}
B^{(i,j)} &\leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\Big(\frac{\lambda_j\mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\Big)\|\mathbf{g}_j(\mathbf{x})-\mathbf{g}_i(\mathbf{x})\|_2^4\Big] \\
&\leq \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\Big(\frac{\lambda_j\mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\Big)^2\Big]\right)^{1/2}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\|\mathbf{g}_j(\mathbf{x})-\mathbf{g}_i(\mathbf{x})\|_2^8\Big]\right)^{1/2} \\
&\leq \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\frac{\lambda_j\mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\Big]\right)^{1/2}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\|\mathbf{g}_j(\mathbf{x})-\mathbf{g}_i(\mathbf{x})\|_2^8\Big]\right)^{1/2} \\
&\leq \frac{1}{\sqrt{\lambda_i}}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\frac{\mathcal{N}_j(\mathbf{x})}{\mathcal{N}_j(\mathbf{x})+\mathcal{N}_i(\mathbf{x})}\Big]\right)^{1/2}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\Big[\|\mathbf{g}_j(\mathbf{x})-\mathbf{g}_i(\mathbf{x})\|_2^8\Big]\right)^{1/2}\,,
\end{aligned} \tag{45}$$

where the third inequality follows because the ratio of weighted densities is pointwise smaller than 1, and the last inequality follows by the fact that $\lambda_j\mathcal{N}_j(\mathbf{x})/(\lambda_i\mathcal{N}_i(\mathbf{x})+\lambda_j\mathcal{N}_j(\mathbf{x})) \leq \frac{1}{\lambda_i}\mathcal{N}_j(\mathbf{x})/(\mathcal{N}_i(\mathbf{x})+\mathcal{N}_j(\mathbf{x}))$ for all $\mathbf{x}$.

We now need to control the following correlation between $\mathcal{N}_j$ and $\mathcal{N}_i$, $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}_j}\Big[\frac{\mathcal{N}_i(\mathbf{x})}{\mathcal{N}_i(\mathbf{x})+\mathcal{N}_j(\mathbf{x})}\Big]$. We show that as long as the parameters of $\mathcal{N}_\ell$ are far in ℓ_2 from those of $\mathcal{N}_j$ this correlation is exponentially small. We prove the following claim.

Claim 36 *Let $\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{Q}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\mathbf{Q}_2)$ be normal distributions with $\alpha I \leq \mathbf{Q}_1 \leq \beta I$, $\alpha I \leq \mathbf{Q}_2 \leq \beta I$. For $c = 16(1+\beta/\alpha)^2\beta^2$, it holds that*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{Q}_1)}\Big[\frac{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\mathbf{Q}_2)}{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\mathbf{Q}_1)+\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_2,\mathbf{Q}_2)}\Big] \leq \exp\Big(-\frac{1}{\beta}\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_2^2 - \frac{1}{c}\,\|\mathbf{Q}_1-\mathbf{Q}_2\|_F^2\Big)\,.$$

Proof We first observe that we can bound by above the correlation between the two normals by their Hellinger distance. For brevity, we will denote $\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{Q}_1)$ as $\mathcal{N}_1$ and $\mathcal{N}(\boldsymbol{\mu}_2,\mathbf{Q}_2)$ as $\mathcal{N}_2$. Using the inequality $2tz/(t+z) \leq \sqrt{tz}$ we obtain that $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}_1}[\mathcal{N}_2(\mathbf{x})/(\mathcal{N}_1(\mathbf{x})+\mathcal{N}_2(\mathbf{x})] \leq \frac{1}{2}(1-\mathbf{H}^2(\mathcal{N}_1,\mathcal{N}_2))$, where $\mathbf{H}^2$ is the squared Hellinger distance between $\mathcal{N}_1$ and $\mathcal{N}_2$. For two normal distributions, we have that

$$1 - \mathbf{H}^2(\mathcal{N}_1, \mathcal{N}_2) = \frac{|\mathbf{Q}_1|^{1/4}|\mathbf{Q}_2|^{1/4}}{|\mathbf{Q}_1/2 + \mathbf{Q}_2/2|^{1/2}} \exp(-(1/8)\mathbf{u}^T(\mathbf{Q}_1/2 + \mathbf{Q}_2/2)^{-1}\mathbf{u}),$$

where $\mathbf{u} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Assuming that λ_i^1 and λ_i^2 are the eigenvalues of $\mathcal{N}_1, \mathcal{N}_2$, we observe that we can write

$$\frac{|\mathbf{Q}_1|^{1/4}|\mathbf{Q}_2|^{1/4}}{|\mathbf{Q}_1/2 + \mathbf{Q}_2/2|^{1/2}} = \exp\Big(\sum_{i=1}^{d} \frac{1}{4}\log\Big(\frac{\lambda_i^1}{\lambda_i^2}\Big) - \frac{1}{2}\log\Big(\frac{1}{2} + \frac{\lambda_i^1}{2\lambda_i^2}\Big)\Big).$$

We can now use the following inequality showing that as long as the ratio λ_i^1/λ_i^2 is not very large the above difference of logarithms behaves roughly as $(1 - \lambda_i^1/\lambda_i^2)^2$.

Fact 37 *Let $x > 0$. It holds $\frac{1}{4}\log x - \log(1/2 + x/2) \le -\frac{1}{16}\frac{(1-x)^2}{(1+x)^2}$.*

Proof We first use the following integral representation of the logarithm difference

$$-\frac{1}{4}\log x + \frac{1}{2}\log(1/2 + x/2) = \frac{1}{2}\int_1^x \frac{1}{1+t} - \frac{1}{2t}dt = \frac{1}{4}\int_1^x \frac{t-1}{(1+t)t}dt.$$

We observe that if $0 < x \le 1$ we have that $(1+t)t \le 2$ when $t \in [1, x]$. In that case, by using the integral identity above, we obtain that $-\frac{1}{4}\log x + \frac{1}{2}\log(1/2 + x/2) \le -(1/16)(1-x)^2$. When $x \ge 1$ we similarly obtain the upper bound $-(1/8)(1-x)^2/((1+x)x)$. Combining the two cases, we obtain the inequality. ∎

Using Theorem 37 we obtain that $\frac{|\mathbf{Q}_1|^{1/4}|\mathbf{Q}_2|^{1/4}}{|\mathbf{Q}_1/2+\mathbf{Q}_2/2|^{1/2}} \le \exp\Big(-\frac{1}{16C^2}\|\mathbf{Id} - \mathbf{Q}_2^{-1/2}\mathbf{Q}_1\mathbf{Q}_2^{-1/2}\|_F^2\Big)$, where $C = 1 + \max_{i=1}^d \lambda_i^1/\lambda_i^2 \le 1 + \beta/\alpha$. Moreover, since $\mathbf{Q}_2^{-1} \ge (1/\beta)\mathbf{Id}$ we obtain that

$$\frac{|\mathbf{Q}_1|^{1/4}|\mathbf{Q}_2|^{1/4}}{|\mathbf{Q}_1/2 + \mathbf{Q}_2/2|^{1/2}} \le \exp\Big(-\frac{1}{16C^2\beta^2}\|\mathbf{Q}_1 - \mathbf{Q}_1\|_F^2\Big).$$

∎

In the following claim, we give a bound for the $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}_1}\Big[\|\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})\|_2^8\Big]$ term that appears in the bound of term $B^{(i,j)}$ of Equation (45).

Claim 38 *Let $\mathcal{N}_1 = \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{Q}_1)$, $\mathcal{N}_2 = \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{Q}_2)$ and define $\mathbf{g}_1(\mathbf{x}) = \mathbf{Q}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)$, $\mathbf{g}_2(\mathbf{x}) = \mathbf{Q}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)$. Assuming that $\alpha\mathbf{Id} \le \mathbf{Q}_1, \mathbf{Q}_2 \le \beta\mathbf{Id}$, it holds*

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_1}\Big[\|\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})\|_2^4\Big] &\lesssim \frac{\beta^2}{\alpha^8}(\|\mathbf{Q}_1 - \mathbf{Q}_2\|_F^2 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2)^2 + \frac{1}{\alpha^2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 \\ &\lesssim \frac{\beta^2}{\alpha^8}(D_p(\mathcal{N}_1, \mathcal{N}_2)^2 + D_p(\mathcal{N}_1, \mathcal{N}_2)).\end{aligned}$$

Moreover, for $t \ge 2$ we have

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_1}\Big[\|\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})\|_2^{2t}\Big] \lesssim t^t\Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_1}\Big[\|\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})\|_2^4\Big]\Big)^{t/2}.$$

Proof We first observe that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_1}\left[\|\mathbf{g}_1(\mathbf{x})-\mathbf{g}_2(\mathbf{x})\|_2^4\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\|(\mathbf{Q}_1^{-1/2}-\mathbf{Q}_2^{-1}\mathbf{Q}_1^{1/2})\,\mathbf{x}+\mathbf{Q}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\|_2^4\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{b},\mathbf{A})}\left[\|\mathbf{x}\|_2^4\right],$$

where $\mathbf{b} = \mathbf{Q}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ and $\mathbf{A} = \mathbf{S}\mathbf{S}^T$ with $\mathbf{S} = \mathbf{Q}_1^{-1/2} - \mathbf{Q}_2^{-1}\mathbf{Q}_1^{1/2}$. By Theorem 21 we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{b},\mathbf{A})}\left[\|\mathbf{x}\|_2^4\right] &= \mathrm{tr}(\mathbf{A})^2 + 2\|\mathbf{A}\|_F^2 + 2\|\mathbf{A}^{1/2}\mathbf{b}\|^2 + \|\mathbf{b}\|_2^2(1+2\mathrm{tr}(\mathbf{A})) + 2\mathbf{b}^T\mathbf{A}\mathbf{b} + \|\mathbf{b}\|_2^4 \\ &\lesssim \|\mathbf{S}\|_F^4 + \|\mathbf{b}\|_2^2(1+\|\mathbf{S}\|_F^2) + \|\mathbf{b}\|_2^4 \lesssim (\|\mathbf{S}\|_F^2 + \|\mathbf{b}\|_2^2)^2 + \|\mathbf{b}\|_2^2\,.\end{aligned}$$

We observe that $\|\mathbf{S}\|_F = \|\mathbf{Q}_1^{-1}(\mathbf{Q}_2 - \mathbf{Q}_1)\mathbf{Q}_2^{-1}\mathbf{Q}_1^{1/2}\|_F \le \frac{\sqrt{\beta}}{\alpha^2}\,\|\mathbf{Q}_1 - \mathbf{Q}_2\|_F$, where the inequality follows by the fact that $\|\mathbf{A}\mathbf{B}\|_F \le \|\mathbf{A}\|_2\|\mathbf{B}\|_F$ and the spectral bounds on $\mathbf{Q}_1, \mathbf{Q}_2$. Moreover, $\|\mathbf{b}\|_2 \le (1/\alpha)\,\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2$, since $\|\mathbf{Q}_2^{-1}\|_2 \le 1/\alpha$. Therefore, we obtain that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_1}\left[\|\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})\|_2^4\right] \lesssim \frac{\beta^2}{\alpha^8}(\|\mathbf{Q}_1 - \mathbf{Q}_2\|_F^2 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2)^2 + \frac{1}{\alpha^2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2\,.$$

To obtain the second bound of the claim, we will use the standard hypercontractivity inequality for polynomials (Theorem 39).

Fact 39 (Gaussian hypercontractivity) *Let $p : \mathbb{R}^d \mapsto \mathbb{R}$ be a polynomial of degree at most ℓ and let $t \ge 2$. It holds $\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{N}}[p^t(\mathbf{x})]\right)^{1/t} \le (t-1)^{\ell/2}\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{N}}[p^2(\mathbf{x})]\right)^{1/2}$.*

We have that $p(\mathbf{x}) = \|\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})\|_2^2$ is a degree 2 polynomial and therefore the claimed bound follows from the previous bound on $\|\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})\|_2^4 = |p(\mathbf{x})|^2$ and the hypercontractivity inequality of Theorem 39. ∎

We can now apply Theorem 36 and Theorem 38 to the bound of Equation (45) and obtain the following bound for some universal constant $c > 0$:

$$\begin{aligned} B^{(i,j)} &\lesssim \frac{\beta^2}{\sqrt{\lambda_i}\alpha^8}\left((\|\mathbf{Q}_i - \mathbf{Q}_j\|_F^2 + \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2)^2 + \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2\right) e^{-c\frac{\alpha^2}{\beta^4}\,(\|\mathbf{Q}_i-\mathbf{Q}_j\|_F^2+\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2)} \\ &\lesssim \frac{\beta^2}{\sqrt{\lambda_i}\alpha^8}(D_p(\mathcal{N}_i,\mathcal{N}_j)^2 + D_p(\mathcal{N}_i,\mathcal{N}_j))\, e^{-c\frac{\alpha^2}{\beta^4}D_p(\mathcal{N}_i,\mathcal{N}_j)} \\ &\lesssim \frac{\beta^{10}}{\sqrt{\lambda_i}\alpha^{12}}\, e^{-(c/4)\frac{\alpha^2}{\beta^4}D_p(\mathcal{N}_i,\mathcal{N}_j)}\,, \end{aligned}$$

where for the last inequality, we used the fact that for all $t \ge 0$, it holds that $t^2e^{-t} \le e^{-t/4}$ and $te^{-t} \le e^{-t/2}$.

We now bound the cross-error term $A^{(i,j,\ell)}$ of Equation Equation (42). We first observe that $A^{(i,j,\ell)}$ (in contrast with term $B^{(i,j)}$ that we bounded previously) does not vanish when $i = j$. We first focus on the case where $i \ne j$. Using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned} A^{(i,j,\ell)} &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\left[\left(\frac{\lambda_j\mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\right)\left(\frac{\lambda_\ell\mathcal{N}_\ell(\mathbf{x})}{S^{-j}(\mathbf{x})}\right)\|\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2^4\right] \\ &\le \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\left[\left(\frac{\lambda_j\mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\right)^4\right]\right)^{1/4}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\left[\left(\frac{\lambda_\ell\mathcal{N}_\ell(\mathbf{x})}{S^{-j}(\mathbf{x})}\right)^4\right]\right)^{1/4}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\left[\|\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2^8\right]\right)^{1/2} \\ &\le \frac{1}{\sqrt{\lambda_i}}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\left[\frac{\mathcal{N}_j(\mathbf{x})}{\mathcal{N}_i(\mathbf{x}) + \mathcal{N}_j(\mathbf{x})}\right]\right)^{1/4}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\left[\frac{\mathcal{N}_\ell(\mathbf{x})}{\mathcal{N}_i(\mathbf{x}) + \mathcal{N}_\ell(\mathbf{x})}\right]\right)^{1/4}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}\left[\|\mathbf{g}_i(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2^8\right]\right)^{1/2}, \end{aligned}$$

where the third inequality follows because the ratio of weighted densities is pointwise smaller than 1. We remark that the last inequality holds true because in the case where $i \neq j$ it holds that $S^{-j}(\mathbf{x}) \geq \lambda_i \mathcal{N}_i(\mathbf{x}) + \lambda_\ell \mathcal{N}_\ell(\mathbf{x})$. We can now use Theorem 36 and Theorem 38 to bound each of the three terms of the above expression for $A^{(i,j,\ell)}$ separately:

$$A^{(i,j,\ell)} \lesssim \frac{\beta^2}{\alpha^8 \sqrt{\lambda_i}} e^{-c' \frac{\alpha^2}{\beta^4}(D_p(\mathcal{N}_i,\mathcal{N}_j)+D_p(\mathcal{N}_i,\mathcal{N}_\ell))} (D_p(\mathcal{N}_i,\mathcal{N}_\ell)^2 + D_p(\mathcal{N}_i,\mathcal{N}_\ell)) \lesssim \frac{\beta^{10}}{\alpha^{12}\sqrt{\lambda_i}} e^{-c' \frac{\alpha^2}{\beta^4} D_p(\mathcal{N}_i,\mathcal{N}_j)} ,$$

where c' is some universal constant and for the last inequality we used the fact that for all t where for the last inequality, we used the fact that for all $t \geq 0$, it holds that $t^2 e^{-t} \leq e^{-t/4}$ and $te^{-t} \leq e^{-t/2}$.

Putting together the bounds for $A^{(i,j,\ell)}$ and $B^{(i,j)}$ we obtain that

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_i}[\|\mathbf{s}(\mathbf{x}) - \mathbf{s}^{-j}(\mathbf{x})\|_2^4] \lesssim \sum_{\ell=1,\ell\neq j}^{k} A^{(i,j,\ell)} + B^{(i,j)} \lesssim \frac{k}{\sqrt{\lambda_i}} \frac{\beta^{10}}{\alpha^{12}} \exp\left(-c\frac{\alpha^2}{\beta^4}\, D_p(\mathcal{N}_i,\mathcal{N}_j)\right) .$$

We now work out the case where $i = j$ (see the second estimate in Theorem 34). Using Theorem 35, for $i = j$, we obtain the following estimate

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_j}[\|\mathbf{s}(\mathbf{x}) - \mathbf{s}^{-j}(\mathbf{x})\|_2^4] \leq 8 \sum_{\ell=1,\ell\neq j}^{k} A^{(j,j,\ell)} .$$

In this case, we cannot guarantee that the weight terms $\lambda_j \mathcal{N}_j(\mathbf{x})/S(\mathbf{x})$ and $\lambda_\ell \mathcal{N}_\ell(\mathbf{x})/S^{-j}(\mathbf{x})$ will be exponentially small and therefore we simply use the fact that they are at most 1:

$$\begin{aligned} A^{(j,j,\ell)} &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_j}\left[\left(\frac{\lambda_j \mathcal{N}_j(\mathbf{x})}{S(\mathbf{x})}\right)\left(\frac{\lambda_\ell \mathcal{N}_\ell(\mathbf{x})}{S^{-j}(\mathbf{x})}\right)\|\mathbf{g}_j(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2^4\right] \leq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}_j}\left[\|\mathbf{g}_j(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2^4\right] \\ &\lesssim \frac{\beta^2}{\alpha^8}(D_p(\mathcal{N}_j,\mathcal{N}_\ell)^2 + D_p(\mathcal{N}_j,\mathcal{N}_\ell)) , \end{aligned}$$

where for the last inequality we used Theorem 38. Substituting the estimate for $A^{(i,j,\ell)}$ yields the claimed bound. ∎

Appendix E. Existence and learning of a piecewise polynomial

E.1. Existence of a piecewise polynomial

In this section, we will show the existence of a piecewise polynomial approximation for the score function. To show the desired polynomial existence result, we start by showing the polynomial existence result for the score function of each subset U_i and combine the results with the clustering guarantee (Theorem 25) and the score simplification guarantee (Theorem 32) to obtain the result for the complete mixture.

E.1.1. Polynomial approximation of a sub-mixture with small parameter distance

We will first obtain the result for a mixture $\mathcal{M}(U)$ where the mixture has $|U| = m \leq k$ components and the parameter distance between any two components $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| + \|\mathbf{Q}_i - \mathbf{Q}_j\| \leq \Delta_{\text{in}}$ for all $i, j \in [m]$. Our main result of this section is the following proposition.

Proposition 40 *Let $\mathcal{M}(U)$ be a mixture of m well-conditioned Gaussians with $\alpha \mathbf{Id} \preceq \mathbf{Q}_i \preceq \beta \mathbf{Id}$ and parameters satisfying $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| + \|\mathbf{Q}_i - \mathbf{Q}_j\|_F \le \Delta_{\text{in}}$ for all $i, j \in [m]$. Let $\{\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i, \widehat{\mathbf{K}}_i\}_{i=1}^m$ be the estimates of the parameters $\{\boldsymbol{\mu}_i, \mathbf{Q}_i, \mathbf{Q}_i^{-1}\}_{i=1}^m$ within parameter distance $\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\| + \|\widehat{\mathbf{Q}}_i - \mathbf{Q}_i\|_F + \|\widehat{\mathbf{K}}_i - \mathbf{Q}_i^{-1}\|_F \le \upsilon$ and with the operator norm satisfying $\|\widehat{\mathbf{K}}_i\|_{\mathsf{op}} \lesssim \frac{1}{\alpha}$ for all $i \in U$. Then, there exists a polynomial $p(\mathbf{x}; \mathcal{M}(U))$ of degree $\widetilde{O}(\frac{\beta^2 m^2 \upsilon^5 \Delta_{\text{in}}^6}{\alpha^6 \varepsilon})$ and coefficients bounded in magnitude by $dR \exp(\widetilde{O}(\frac{\beta^2 m^2 \upsilon^5 \Delta_{\text{in}}^6}{\alpha^6 \varepsilon}))$ such that for all $\mathbf{x}$, the following holds*

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}(U)}[\|s(\mathbf{x}; \mathcal{M}(U)) - \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U))\|^2] \le \varepsilon \,,$$

where the approximating function is $\widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U)) \triangleq p(\mathbf{x}; \mathcal{M}(U)) \mathbb{1}\{\widehat{B}(\mathbf{x}; U)\} + \widehat{\mathbf{K}}_i(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i) \mathbb{1}\{\widehat{B}^c(\mathbf{x}; U)\}$ for some $i \in U$ where $\widehat{B}(\mathbf{x}; U)$ denotes the region $\widehat{B}(x; \widehat{\theta}_1, \widehat{\theta}_2)$ of the polynomial approximation for cluster U. where $\widehat{B}(\mathbf{x}) : \mathbb{R}^d \to \{0, 1\}$ function that only depends on the estimates $\{\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i, \widehat{\mathbf{K}}_i\}_{i=1}^m$.

Observe that the score function for the mixture can be written as a product between linear functions (i.e., $\mathbf{Q}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$) and the softmax function. We define the softmax function $w : \mathbb{R}^m \mapsto [0, 1]^m$ as follows:

$$w_i(\mathbf{y}; \boldsymbol{\theta}) = \frac{e^{\mathbf{y}_i + \boldsymbol{\theta}_i}}{\sum_{j=1}^m e^{\mathbf{y}_j + \boldsymbol{\theta}_j}} \tag{46}$$

for some fixed parameters $\{\boldsymbol{\theta}_i\}_{i=1}^m$. We start by showing that in this special case, the score can be pointwise approximated by a low-degree polynomial over a bounded domain (Theorem 43 below).

For this, we will need the following classical polynomial approximation result for functions with bounded gradients:

Lemma 41 (Multivariate Jackson's Approximation, (Newman and Shapiro, 1964; Diakonikolas et al., 2010)) *For $F : \mathbb{R}^n \to \mathbb{R}$, define the modulus of continuity*

$$\omega(F, \delta) = \sup_{\substack{\|\mathbf{x}\|_2, \|\mathbf{y}\|_2 \le 1 \\ \|\mathbf{x} - \mathbf{y}\| \le \delta}} |F(\mathbf{x}) - F(\mathbf{y})|.$$

For any $\ell \ge 1$, there exists a polynomial p_ℓ of degree ℓ such that

$$\sup_{\|\mathbf{x}\|_2 \le 1} |F(\mathbf{x}) - p_\ell(\mathbf{x})| \lesssim \omega(F, n/\ell) \,.$$

To prove an upper bound on the coefficients of the polynomial, we will use the following result.

Lemma 42 (Coefficients of bounded polynomials, (Ben-David et al., 2018)) *Let p be a polynomial with real coefficients on d variables with degree ℓ such that for all $\mathbf{x} \in [0, L]^d$, $|p(\mathbf{x})| \le R$. Then, the sum of the magnitude of all coefficients of p is at most $R(2L(d + \ell))^{3\ell}$ for any $L \ge 1$.*

We now show the polynomial approximation result for the softmax function and, as a consequence, for the product of a linear function with the softmax function:

Lemma 43 (Polynomial Approximation) *Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ and $w_i(\mathbf{y}; \boldsymbol{\theta})$ be the softmax function defined in (46). Let $\mathbf{G}(\mathbf{x}) = (\mathbf{g}_1(\mathbf{x}), \ldots, \mathbf{g}_m(\mathbf{x})) : \mathbb{R}^d \mapsto \mathbb{R}^{d \times m}$ be such that $\|\mathbf{g}_i(\mathbf{x})\|_2 \le M$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{g}_i(\mathbf{x})$ is linear in $\mathbf{x}$. Let $\mathbf{r} : \mathbb{R}^d \mapsto \mathbb{R}^m$ with $\mathbf{r} = (\mathbf{r}_1(\mathbf{x}), \ldots, \mathbf{r}_m(\mathbf{x}))$ be such that $|\mathbf{r}_i(\mathbf{x})| \le L$ for all $\mathbf{x} \in \mathcal{X}$. There exists a polynomial transformation $\mathbf{q} : \mathbb{R}^m \mapsto \mathbb{R}^m$ of degree at most $O(LMm^2/\varepsilon)$ such that for all $\mathbf{x} \in \mathcal{X}$ it holds that $\|\mathbf{G}(\mathbf{x}) w(r(\mathbf{x}); \boldsymbol{\theta}) - \mathbf{G}(\mathbf{x}) \mathbf{q}(r(\mathbf{x}))\|_2 \le \varepsilon$. The sum of the magnitudes of the coefficients of $\mathbf{q}$ is at most $m \exp(\widetilde{O}(LMm^2/\varepsilon))$.*

Proof The gradient of the softmax function is given by

$$\frac{\partial w_i(\mathbf{y};\boldsymbol{\theta})}{\partial \mathbf{y}_j} = \begin{cases} w_i(\mathbf{y};\boldsymbol{\theta})(1-w_i(\mathbf{y};\boldsymbol{\theta})) & \text{if} \quad i=j \\ -w_i(\mathbf{y};\boldsymbol{\theta})w_j(\mathbf{y};\boldsymbol{\theta}) & \text{otherwise.} \end{cases}$$

We conclude that $\|\nabla w_i(\mathbf{y};\boldsymbol{\theta})\| \le \sqrt{m}$ for all $i \in [m]$ and any $\mathbf{y} \in \mathbb{R}^m$. Using multivariate Jackson's theorem (Theorem 41) for $w_i(\mathbf{y};\boldsymbol{\theta})$, we obtain that there exists a polynomial $q(\mathbf{y})$ of degree ℓ such that

$$\sup_{\|\mathbf{y}\| \le Lm} |w_i(\mathbf{y};\boldsymbol{\theta}) - q(\mathbf{y})| \lesssim \frac{Lm^{\frac{3}{2}}}{\ell}\,.$$

This implies that we have a set of polynomials $\{q_i(\mathbf{y})\}_{i=1}^m$ of degree $O(\frac{Lm^{3/2}}{\varepsilon})$ such that for all $\mathbf{y}$ in L_2-ball of radius $\|\mathbf{y}\| \le Lm$, we have $\|w(\mathbf{y};\boldsymbol{\theta}) - \mathbf{q}(\mathbf{y})\| \le \varepsilon$. Additionally, $\|\mathbf{g}_i(\mathbf{x})\|_2 \le M$ implies that $\|\mathbf{G}(\mathbf{x})\| \le M\sqrt{m}$. Therefore, we have

$$\|\mathbf{G}(\mathbf{x})\mathbf{w}(r(\mathbf{x})) - \mathbf{G}(\mathbf{x})\mathbf{q}(r(\mathbf{x}))\|_2 \le \|\mathbf{G}(\mathbf{x})\|\|\mathbf{w}(r(\mathbf{x})) - \mathbf{q}(r(\mathbf{x}))\|_2 \le M\sqrt{m}\varepsilon.$$

We obtain the result by rescaling ε. To obtain the bounds on the sum of the magnitude of coefficients, we use the fact that $|\mathbf{q}_i(\mathbf{y})| \le 2$ for all $\|\mathbf{y}\| \le Lm$. Therefore, using Theorem 42, we obtain that the bounds on the sum of the magnitude of coefficients is at most $O((2Lm(m+\frac{LMm^2}{\varepsilon})^{\frac{LMm^2}{\varepsilon}})) = \exp(\widetilde{O}(LMm^2/\varepsilon))$. ■

Lemma 44 *Let $\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{Q}_1)$ be a Gaussian distribution with $\alpha\mathbf{Id} \preceq \mathbf{Q}_1 \preceq \beta\mathbf{Id}$. Let $(\widehat{\boldsymbol{\mu}}_2,\widehat{\mathbf{Q}}_2,\widehat{\mathbf{K}}_2)$ and $(\widehat{\boldsymbol{\mu}}_3,\widehat{\mathbf{Q}}_3,\widehat{\mathbf{K}}_3)$ be any triplets of the same shape as $(\boldsymbol{\mu}_1,\mathbf{Q}_1,\mathbf{Q}_1^{-1})$ with condition that $\|\widehat{\mathbf{K}}_2\|_{\mathsf{op}}, \|\widehat{\mathbf{K}}_3\|_{\mathsf{op}} \lesssim \frac{1}{\alpha}$. Then, with probability at least $1-\delta$ over $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1,\mathbf{Q}_1)$, we have*

$$\begin{aligned}\big|\|\mathbf{x}-\widehat{\boldsymbol{\mu}}_2\|^2_{\widehat{\mathbf{K}}_2} - \|\mathbf{x}-\widehat{\boldsymbol{\mu}}_3\|^2_{\widehat{\mathbf{K}}_3} - \langle \mathbf{Q}_1, (\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3)\rangle\big| &\lesssim \beta\|\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3\|_F \log\frac{1}{\delta} \\ &+ \frac{1}{\alpha}\big(\|\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2\|^2 + \|\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_3\|^2\big) + \sqrt{\beta}\log\frac{1}{\delta}(\|\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3\|_{\mathsf{op}}\|\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2\| + \frac{1}{\alpha}\|\widehat{\boldsymbol{\mu}}_3-\widehat{\boldsymbol{\mu}}_2\|)\end{aligned}$$

Proof For $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1,\mathbf{Q}_1)$, we rewrite $\|\mathbf{x}-\widehat{\boldsymbol{\mu}}_2\|^2_{\widehat{\mathbf{K}}_2} - \|\mathbf{x}-\widehat{\boldsymbol{\mu}}_3\|^2_{\widehat{\mathbf{K}}_3}$ by writing $\mathbf{x} = \mathbf{Q}_1^{1/2}\mathbf{z} + \boldsymbol{\mu}_1$ for $\mathbf{z} \sim \mathcal{N}(0,\mathbf{Id})$, obtaining:

$$\begin{aligned}\|\mathbf{x}-\widehat{\boldsymbol{\mu}}_2\|^2_{\widehat{\mathbf{K}}_2} - \|\mathbf{x}-\widehat{\boldsymbol{\mu}}_3\|^2_{\widehat{\mathbf{K}}_3} &= \|\mathbf{Q}_1^{1/2}\mathbf{z}\|^2_{\widehat{\mathbf{K}}_2} - \|\mathbf{Q}_1^{1/2}\mathbf{z}\|^2_{\widehat{\mathbf{K}}_3} + \|\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2\|^2_{\widehat{\mathbf{K}}_2} - \|\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_3\|^2_{\widehat{\mathbf{K}}_3} \\ &+ 2(\mathbf{Q}_1^{1/2}\mathbf{z})^{\mathsf{T}}\widehat{\mathbf{K}}_2(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2) - 2(\mathbf{Q}_1^{1/2}\mathbf{z})^{\mathsf{T}}\widehat{\mathbf{K}}_3(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_3)\end{aligned} \tag{47}$$

We would like to bound the first two terms in the above equation using Hanson-Wright (Theorem 19). Using $\|\mathbf{Q}_1\| \le \beta$, we have $\|\mathbf{Q}_1^{1/2}(\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3)\mathbf{Q}_1^{1/2}\| \le \beta\|\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3\|_F$. Using Hanson-Wright on the quadratic form $\mathbf{z}^{\mathsf{T}}\mathbf{Q}_1^{1/2}(\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3)\mathbf{Q}_1^{1/2}\mathbf{z}$, we have for any $\delta > 0$ that

$$\Pr_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})}\Big[|\|\mathbf{Q}_1^{1/2}\mathbf{z}\|^2_{\widehat{\mathbf{K}}_2} - \|\mathbf{Q}_1^{1/2}\mathbf{z}\|^2_{\widehat{\mathbf{K}}_3} - \langle \mathbf{Q}_1, (\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3)\rangle| \gtrsim \beta\|\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3\|_F \log\frac{1}{\delta}\Big] \le \delta.$$

We simplify the sum of the last two terms in (47) to obtain

$$(\mathbf{Q}_1^{1/2}\mathbf{z})^\intercal\widehat{\mathbf{K}}_2(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2)-(\mathbf{Q}_1^{1/2}\mathbf{z})^\intercal\widehat{\mathbf{K}}_3(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_3)=(\mathbf{Q}_1^{1/2}\mathbf{z})^\intercal(\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3)(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2)+(\mathbf{Q}_1^{1/2}\mathbf{z})^\intercal\widehat{\mathbf{K}}_3(\widehat{\boldsymbol{\mu}}_3-\widehat{\boldsymbol{\mu}}_2). \tag{48}$$

Using the bounds $\|\mathbf{Q}_1\|_{op}\le\beta$ and $\|\widehat{\mathbf{K}}_3\|_{op}\lesssim 1/\alpha$, we can upper bound $\|\mathbf{Q}_1^{1/2}(\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3)(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2)\|\lesssim\sqrt{\beta}\|\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3\|_{\mathsf{op}}\|\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2\|$ and $\|\mathbf{Q}_1^{1/2}\widehat{\mathbf{K}}_3(\widehat{\boldsymbol{\mu}}_3-\widehat{\boldsymbol{\mu}}_2)\|\lesssim\sqrt{\beta}\|\widehat{\boldsymbol{\mu}}_3-\widehat{\boldsymbol{\mu}}_2\|/\alpha$. So with probability at least $1-\delta$, we have

$$\|(\mathbf{Q}_1^{1/2}\mathbf{z})^\intercal\widehat{\mathbf{K}}_2(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2)-(\mathbf{Q}_1^{1/2}\mathbf{z})^\intercal\widehat{\mathbf{K}}_3(\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_3)\|\le\sqrt{\beta}\log\frac{1}{\delta}(\|\widehat{\mathbf{K}}_2-\widehat{\mathbf{K}}_3\|_{\mathsf{op}}\|\boldsymbol{\mu}_1-\widehat{\boldsymbol{\mu}}_2\|+\frac{\|\widehat{\boldsymbol{\mu}}_3-\widehat{\boldsymbol{\mu}}_2\|}{\alpha}).$$

Putting everything together in (47) and assuming $\alpha\le 1$ and $\beta\ge 1$ to simplify, we obtain the result. ■

Lemma 45 *Let $\mathcal{M}(U)$ be a mixture of m Gaussians with well-conditioned covariances $\alpha\mathbf{Id}\preceq\mathbf{Q}_i\preceq\beta\mathbf{Id}$ for all $i\in[m]$. Let Δ_{in} be an upper bound on the parameter distance between any two components, i.e., $\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|+\|\mathbf{Q}_i-\mathbf{Q}_j\|_F\le\Delta_{\text{in}}$ for all $i,j\in[m]$. Then, for $\mathbf{x}\sim\mathcal{M}(U)$ and for any $j\in[m]$, with probability at least $1-\delta$, we have*

$$|\|\mathbf{x}-\boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}}-\|\mathbf{x}-\boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}}-\langle\mathbf{Q}_1,(\mathbf{Q}_j^{-1}-\mathbf{Q}_1^{-1})\rangle|\lesssim\zeta_1\quad\textit{where}\quad\zeta_1\triangleq\frac{\beta\Delta_{\text{in}}^2}{\alpha^2}\log\frac{m}{\delta}$$
$$\textit{and}\quad\|(\mathbf{Q}_i^{-1}-\mathbf{Q}_1^{-1})(\mathbf{x}-\boldsymbol{\mu}_i)\|\lesssim\zeta_2\quad\textit{where}\quad\zeta_2\triangleq\frac{\sqrt{\beta}\Delta_{\text{in}}^2}{\alpha^2}\log\frac{m}{\delta}\ .$$

Combining it with Theorem 43, we obtain that there exists a polynomial $p(\mathbf{x};\mathcal{M}(U))$ of degree $O(\frac{\zeta_1\zeta_2m^2}{\varepsilon})$ and coefficients bounded in magnitude by $dR\exp(\widetilde{O}(\frac{\zeta_1\zeta_2m^2}{\varepsilon}))$ such that

$$\Pr_{\mathbf{x}\sim\mathcal{M}(U)}\left[\|s(\mathbf{x};\mathcal{M}(U))-p(\mathbf{x};\mathcal{M}(U))\|\le\varepsilon\right]\ge1-\delta\ .$$

Proof Recall that the score function for the mixture is

$$s(\mathbf{x};\mathcal{M}(U))=\sum_{i\in U}w_i(\mathbf{x})\mathbf{Q}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\quad\text{ where }\quad w_i(\mathbf{x})=\frac{\lambda_i\det(\mathbf{Q}_i)^{-1/2}\,e^{-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}_i\|^2_{\mathbf{Q}_i^{-1}}}}{\sum_{j\in U}\lambda_i\det(\mathbf{Q}_j)^{-1/2}\,e^{-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}}}}.$$

We can rewrite the score function as $s(\mathbf{x};\mathcal{M}(U))=s_1(\mathbf{x};\mathcal{M}(U))+s_2(\mathbf{x};\mathcal{M}(U))+\mathbf{Q}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)$ where $s_1(\mathbf{x};\mathcal{M}(U))$ and $s_2(\mathbf{x};\mathcal{M}(U))$ are defined as

$$s_1(\mathbf{x};\mathcal{M}(U))=\sum_{i\in U}w_i(\mathbf{x})(\mathbf{Q}_i^{-1}-\mathbf{Q}_1^{-1})(\mathbf{x}-\boldsymbol{\mu}_i)\text{ and }s_2(\mathbf{x};\mathcal{M}(U))=-\sum_{i\in U}w_i(\mathbf{x})\mathbf{Q}_1^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_1)$$
$$\text{and}\quad w_i(\mathbf{x})=\frac{e^{-\frac{1}{2}(\|\mathbf{x}-\boldsymbol{\mu}_i\|^2_{\mathbf{Q}_i^{-1}}-\|\mathbf{x}-\boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}}+\log(\frac{\det(\mathbf{Q}_i)}{\det(\mathbf{Q}_1)}))+\log\frac{\lambda_i}{\lambda_1}}}{1+\sum_{j=2}^m e^{-\frac{1}{2}(\|\mathbf{x}-\boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}}-\|\mathbf{x}-\boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}}+\log(\frac{\det(\mathbf{Q}_j)}{\det(\mathbf{Q}_1)}))+\log\frac{\lambda_j}{\lambda_1}}}\ .$$

We show the polynomial approximation result for $s_1(\mathbf{x}; \mathcal{M}(U))$ and $s_2(\mathbf{x}; \mathcal{M}(U))$ using Theorem 43. To prove an upper bound on $\|\mathbf{g}_i(\mathbf{x})\|$ in Theorem 43, we apply Theorem 44 for all $j, \ell \in [m]$ and have that with probability at least $1-\delta$ over $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \mathbf{Q}_\ell)$ (and hence over $\mathbf{x} \sim \mathcal{M}(U)$), we have

$$\left|\|\mathbf{x}-\boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}} - \|\mathbf{x}-\boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}} - \langle \mathbf{Q}_\ell, \mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\rangle\right| \lesssim \beta\|\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\|_F \log\frac{m}{\delta}$$
$$+ \frac{1}{\alpha}(\|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_j\|^2 + \|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_1\|^2) + \sqrt{\beta}\log\frac{m}{\delta}(\|\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\|_{\mathsf{op}}\|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_j\| + \frac{1}{\alpha}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_1\|).$$

Using $\|\mathbf{Q}_i^{-1}\|_{op} \le 1/\alpha$ for all $i \in [k]$, we have $\|\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\|_F = \|\mathbf{Q}_j^{-1}(\mathbf{Q}_j - \mathbf{Q}_1)\mathbf{Q}_1^{-1}\|_F \le \Delta_{in}/\alpha^2$, we have

$$\beta\|\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\|_F \log\frac{m}{\delta} + \frac{1}{\alpha}(\|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_j\|^2 + \|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_1\|^2)$$
$$+ \sqrt{\beta}\log\frac{m}{\delta}(\|\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\|_{\mathsf{op}}\|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_j\| + \frac{1}{\alpha}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_1\|) \le \frac{\beta\Delta_{\text{in}}^2}{\alpha^2}\log\frac{m}{\delta}$$

We add and subtract $\langle \mathbf{Q}_1, (\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\rangle$ on the left side and rearranging the terms and

$$\left|\|\mathbf{x}-\boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}} - \|\mathbf{x}-\boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}} - \langle \mathbf{Q}_1, (\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\rangle\right| \lesssim \frac{\beta\Delta_{\text{in}}^2}{\alpha^2}\log\frac{m}{\delta} + \|\mathbf{Q}_\ell - \mathbf{Q}_1\|_F\|\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\|_F$$
$$\lesssim \frac{\beta\Delta_{\text{in}}^2}{\alpha^2}\log\frac{m}{\delta}\,.$$

We have $\|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})\mathbf{Q}_\ell^{1/2}\|_F^2 \le \frac{\beta\Delta_{in}^2}{\alpha^4}$. For a fixed $\ell \in [m]$, when $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \mathbf{Q}_\ell)$, we can rewrite $\|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x}-\boldsymbol{\mu}_i)\|$ by expressing $\mathbf{x} = \mathbf{Q}_\ell^{1/2}\mathbf{z} + \boldsymbol{\mu}_\ell$ for $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$ to get:

$$\|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x}-\boldsymbol{\mu}_i)\| \le \|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})\mathbf{Q}_\ell^{1/2}\mathbf{z}\| + \|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_i)\| \tag{49}$$

Using Hanson-Wright (Theorem 19), with at least $1-\delta$ probability over $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$, $\|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})\mathbf{Q}_\ell^{1/2}\mathbf{z}\| \lesssim \|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})\mathbf{Q}_\ell^{1/2}\|_F(1 + \log\frac{1}{\delta}) \lesssim \frac{\sqrt{\beta}\Delta_{\text{in}}}{\alpha^2}\log\frac{1}{\delta}$. Using this bound in (49), with probability at least $1-\delta$ over $\mathbf{x} \sim \mathcal{M}(U)$, we have

$$\|(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x}-\boldsymbol{\mu}_i)\| \lesssim \frac{\Delta_{in}^2}{\alpha} + \frac{\sqrt{\beta}\Delta_{\text{in}}}{\alpha^2}\log\frac{m}{\delta} \lesssim \frac{\sqrt{\beta}\Delta_{\text{in}}^2}{\alpha^2}\log\frac{m}{\delta}.$$

We apply Theorem 43 to $s_1(\mathbf{x}; \mathcal{M}(U))$ with the softmax function taking input $r_j(\mathbf{x}) = -\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}} + \frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}} + \frac{1}{2}\langle \mathbf{Q}_1, (\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\rangle$ and $\boldsymbol{\theta}_j = \log\frac{\lambda_j}{\lambda_1} - \frac{1}{2}\langle \mathbf{Q}_1, (\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\rangle + \frac{1}{2}\log\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_j)}$. We take L and M therein to be of order $\frac{\beta\Delta_{\text{in}}^2}{\alpha^2}\log\frac{m}{\delta}$ and $\frac{\sqrt{\beta}\Delta_{in}^2}{\alpha^2}\log\frac{m}{\delta}$ respectively. We conclude that there exists a polynomial transformation $p_1(\mathbf{x}; \mathcal{M}(U))$ with degree $O(LMm^2/\varepsilon) = O(\zeta_1\zeta_2 m^2/\varepsilon)$ such that with probability at least $1-\delta$ over $\mathbf{x} \sim \mathcal{M}(U)$, we have

$$\|s_1(\mathbf{x}; \mathcal{M}(U)) - p_1(\mathbf{x}; \mathcal{M}(U))\| \le \varepsilon\,.$$

Note that the multiplication of $(\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x}-\boldsymbol{\mu}_i)$ to the polynomial approximation of the softmax can increase the sum of absolute values of coefficients at most by a factor of $\frac{dRm}{\alpha}$. The sum of absolute values of coefficients of the polynomial transformation $p_1(\mathbf{x}; \mathcal{M}(U))$ is $\frac{dRm}{\alpha}\exp(\widetilde{O}(\frac{\zeta_1\zeta_2 m^2}{\varepsilon}))$.

We also have $\|\mathbf{Q}_1^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_1)\| \leq \Delta_{\text{in}}/\alpha$. We apply Theorem 43 for $s_2(\mathbf{x}; \mathcal{M}(U))$ with the same choice of $r_j(\mathbf{x})$ and L but we take $\mathbf{g}_j(\mathbf{x})$ and M as $\mathbf{Q}_1^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_1)$ and $\Delta_{\text{in}}/\alpha$. Therefore, we obtain that there exists a polynomial $p_2(\mathbf{x}; \mathcal{M}(U))$ with degree $\frac{\beta m^2 \Delta_{\text{in}}^3}{\varepsilon \alpha^3} \log \frac{m}{\delta}$ such that with at least $1-\delta$ probability, we have

$$\|s_2(\mathbf{x}; \mathcal{M}(U)) - p_2(\mathbf{x}; \mathcal{M}(U))\| \leq \varepsilon\,.$$

Combining the polynomials $p_1(\mathbf{x}; \mathcal{M}(U))$ and $p_2(\mathbf{x}; \mathcal{M}(U))$, we obtain the result. ■

We define $V_1^{(j)}(\mathbf{x})$ to measure relative distance of j^{th} input of the softmax to its mean and $V_2^{(j)}(\mathbf{x})$ to measure norm of $(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x} - \boldsymbol{\mu}_j)$ as follows:

$$V_1^{(j)}(\mathbf{x}) \triangleq \|\mathbf{x} - \boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}} - \|\mathbf{x} - \boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}} - \langle \mathbf{Q}_1, (\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\rangle$$
$$V_2^{(j)}(\mathbf{x}) \triangleq \|(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x} - \boldsymbol{\mu}_j)\|^2$$

We similarly define $\widehat{V}_1^{(j)}$ and $\widehat{V}_2^{(j)}$ using estimates $\{\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i, \widehat{\mathbf{K}}_i\}_{i=1}^k$ instead of $\{\boldsymbol{\mu}_i, \mathbf{Q}_i, \mathbf{Q}_i^{-1}\}_{i=1}^k$. Define $B(\cdot)$ to be the indicator function for whether the input to the softmax is close to its mean and $(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x} - \boldsymbol{\mu}_j)$ is sufficiently small in norm:

$$B(\mathbf{x}, \theta_1, \theta_2) \triangleq \bigwedge_{j=1}^{k} B^{(j)}(\mathbf{x}, \theta_1, \theta_2) \quad \text{where} \quad B^{(j)}(\mathbf{x}, \theta_1, \theta_2) \triangleq \mathbb{1}\Big\{\Big(\big|V_1^{(j)}(\mathbf{x})\big| \leq \theta_1\Big) \wedge \Big(V_2^{(j)}(\mathbf{x}) \leq \theta_2\Big)\Big\}$$

Observe that the polynomial approximation result of Theorem 45 holds when $B(\mathbf{x}, \theta_1, \theta_2) = 1$. We also define $\widehat{B}$ and $\widehat{B}^{(j)}$ by replacing $V_1^{(j)}$ and $V_2^{(j)}$ with $\widehat{V}_1^{(j)}$ and $\widehat{V}_1^{(j)}$ in the definition of B and $B^{(j)}$.

Following the parameters used in the proof of Theorem 45, we will take

$$\theta_1 \triangleq \Theta\Big(\frac{\beta \Delta_{\text{in}}^2}{\alpha^2} \log \frac{m}{\delta}\Big) \qquad \theta_2 \triangleq \Theta\Big(\frac{\sqrt{\beta} \Delta_{\text{in}}^2}{\alpha^2} \log \frac{m}{\delta}\Big)\,. \tag{50}$$

Lemma 46 *Let $\mathcal{M}(U)$ be a mixture of m Gaussians with $\alpha \mathbf{Id} \preceq \mathbf{Q}_i \preceq \beta \mathbf{Id}$ and parameters satisfying $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| + \|\mathbf{Q}_i - \mathbf{Q}_j\|_F \leq \Delta_{\text{in}}$ for all $i, j \in [m]$. Let $\{\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i, \widehat{\mathbf{K}}_i\}_{i=1}^m$ be the estimates of the parameters $\{\boldsymbol{\mu}_i, \mathbf{Q}_i, \mathbf{Q}_i^{-1}\}_{i=1}^m$ within parameter distance $\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\| + \|\widehat{\mathbf{Q}}_i - \mathbf{Q}_i^{-1}\|_F + \|\widehat{\mathbf{K}}_i - \mathbf{Q}_i^{-1}\|_F \leq \upsilon$ and with the operator norm satisfying $\|\widehat{\mathbf{K}}_i\|_{\mathsf{op}} \lesssim \frac{1}{\alpha}$ for all $i \in U$. Then, for any $\mathbf{x} \sim \mathcal{M}(U)$, with probability at least $1-\delta$, the error in estimating $V_1^{(j)}(\mathbf{x})$ by $\widehat{V}_1^{(j)}(\mathbf{x})$ (similarly $V_2^{(j)}(\mathbf{x})$ by $\widehat{V}_2^{(j)}(\mathbf{x})$) is upper bounded by*

$$\big|V_1^{(j)}(\mathbf{x}) - \widehat{V}_1^{(j)}(\mathbf{x})\big| \lesssim \omega_1 \qquad \textit{where} \qquad \omega_1 \triangleq \frac{\beta \Delta_{\text{in}}^2 \upsilon^2}{\alpha} \log \frac{m}{\delta}\,,$$
$$\big|V_2^{(j)}(\mathbf{x}) - \widehat{V}_2^{(j)}(\mathbf{x})\big| \lesssim \omega_2 \qquad \textit{where} \qquad \omega_2 \triangleq \frac{\beta \Delta_{\text{in}}^4 \upsilon^3}{\alpha^4} \log \frac{m}{\delta}\,.$$

Proof The expression of $V_1^{(j)}(\mathbf{x}) - \widehat{V}_1^{(j)}(\mathbf{x})$ can be rewritten as

$$\begin{aligned} V_1^{(j)}(\mathbf{x}) - \widehat{V}_1^{(j)}(\mathbf{x}) = \big(&\|\mathbf{x} - \boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}} - \|\mathbf{x} - \widehat{\boldsymbol{\mu}}_j\|_{\widehat{\mathbf{K}}_j} - \langle \mathbf{Q}_\ell, \mathbf{Q}_j^{-1} - \widehat{\mathbf{K}}_j\rangle\big) \\ &- \big(\|\mathbf{x} - \boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}} - \|\mathbf{x} - \widehat{\boldsymbol{\mu}}_1\|_{\widehat{\mathbf{K}}_1} - \langle \mathbf{Q}_\ell, \mathbf{Q}_1^{-1} - \widehat{\mathbf{K}}_1\rangle\big) \\ &+ \langle \mathbf{Q}_\ell - \mathbf{Q}_1, \mathbf{Q}_j^{-1} - \widehat{\mathbf{K}}_j + \widehat{\mathbf{K}}_1 - \mathbf{Q}_1^{-1}\rangle + \langle \mathbf{Q}_1 - \widehat{\mathbf{Q}}_1, \widehat{\mathbf{K}}_1 - \widehat{\mathbf{K}}_j\rangle. \end{aligned} \tag{51}$$

Using Theorem 44 by choosing $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{Q}_1)$ as $\mathcal{N}(\boldsymbol{\mu}_\ell, \mathbf{Q}_\ell)$ and $(\widehat{\boldsymbol{\mu}}_2, \widehat{\mathbf{Q}}_2, \widehat{\mathbf{K}}_2), (\widehat{\boldsymbol{\mu}}_3, \widehat{\mathbf{Q}}_3, \widehat{\mathbf{K}}_3)$ as $(\boldsymbol{\mu}_j, \mathbf{Q}_j, \mathbf{Q}_j^{-1})$ and $(\widehat{\boldsymbol{\mu}}_j, \widehat{\mathbf{Q}}_j, \widehat{\mathbf{K}}_j)$ and applying the union bound over $j, \ell \in U$, for $(\ell, \mathbf{x}) \sim \mathcal{M}^J(U)$, with at least $1-\delta$ probability, we have

$$\begin{aligned}\left|\|\mathbf{x}-\boldsymbol{\mu}_j\|^2_{\mathbf{Q}_j^{-1}} - \|\mathbf{x}-\widehat{\boldsymbol{\mu}}_j\|_{\widehat{\mathbf{K}}_j} - \langle \mathbf{Q}_\ell, \mathbf{Q}_j^{-1} - \widehat{\mathbf{K}}_j\rangle\right| &\lesssim \beta\upsilon \log\frac{m}{\delta} + \frac{1}{\alpha}(\Delta_{\text{in}}^2 + \upsilon^2) + \sqrt{\beta}\log\frac{m}{\delta}(\upsilon\Delta_{\text{in}} + \frac{\upsilon}{\alpha}) \\ &\lesssim \frac{\beta\Delta_{\text{in}}^2\upsilon^2}{\alpha}\log\frac{m}{\delta}.\end{aligned}$$

For $j = 1$ in the above equation, we also have

$$\left|\|\mathbf{x}-\boldsymbol{\mu}_1\|^2_{\mathbf{Q}_1^{-1}} - \|\mathbf{x}-\widehat{\boldsymbol{\mu}}_1\|_{\widehat{\mathbf{K}}_1} - \langle \mathbf{Q}_\ell, \mathbf{Q}_1^{-1} - \widehat{\mathbf{K}}_1\rangle\right| \lesssim \frac{\beta\Delta_{\text{in}}^2\upsilon^2}{\alpha}\log\frac{m}{\delta}$$

Note that $\langle \mathbf{Q}_1 - \widehat{\mathbf{Q}}_1, \widehat{\mathbf{K}}_1 - \widehat{\mathbf{K}}_j\rangle \lesssim \upsilon(\upsilon + \frac{\Delta_{\text{in}}}{\alpha^2})$ therefore, the last term in (51) can be upper bounded as

$$\left|\langle \mathbf{Q}_\ell - \mathbf{Q}_1, \mathbf{Q}_j^{-1} - \widehat{\mathbf{K}}_j + \widehat{\mathbf{K}}_1 - \mathbf{Q}_1^{-1}\rangle + \langle \mathbf{Q}_1 - \widehat{\mathbf{Q}}_1, \widehat{\mathbf{K}}_1 - \widehat{\mathbf{K}}_j\rangle\right| \lesssim \frac{\upsilon^2\Delta_{\text{in}}}{\alpha^2}.$$

When $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$, using Hanson-Wright (Theorem 19), we have $\|(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\mathbf{Q}_\ell^{1/2}\mathbf{z}\|^2 \lesssim \frac{\beta\Delta_{\text{in}}^2}{\alpha^4}\log\frac{1}{\delta}$ with probability at least $1-\delta$. Additionally, we have $\|(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_j)(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})^\intercal(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\mathbf{Q}_\ell^{1/2}\| \lesssim \sqrt{\beta}\Delta_{\text{in}}^3/\alpha^4$. Therefore, with at least $1-\delta$ probability, we obtain $|(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_j)(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})^\intercal(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})\mathbf{Q}_\ell^{1/2}\mathbf{z}| \le \frac{\sqrt{\beta}\Delta_{\text{in}}^3}{\alpha^4}\log\frac{m}{\delta}$ Moreover, $\|(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_j)\|^2 \le \beta\Delta_{\text{in}}^4/\alpha^4$. Therefore, for $\mathbf{x} \sim \mathcal{M}(U)$, with probability at least $1-\delta$, we have

$$\|(\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1})(\mathbf{x} - \boldsymbol{\mu}_j)\|^2 \lesssim \frac{\beta\Delta_{\text{in}}^4}{\alpha^4}\log\frac{m}{\delta}. \tag{52}$$

Similarly, for any $\ell \in [m]$, we have $\|(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)\mathbf{Q}_\ell^{1/2}\|_F^2 \lesssim \beta(\|\widehat{\mathbf{K}}_j - \mathbf{Q}_j^{-1}\|_F^2 + \|\widehat{\mathbf{K}}_1 - \mathbf{Q}_1^{-1}\|_F^2 + \|\mathbf{Q}_j^{-1} - \mathbf{Q}_1^{-1}\|_F^2) \lesssim \beta(\upsilon^2 + \Delta_{\text{in}}^2/\alpha^4)$. Using Hanson-Wright inequality (Theorem 19) for $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$, with probability at least $1-\delta$, we have $\|(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)\mathbf{Q}_\ell^{1/2}\mathbf{z}\|^2 \lesssim (\beta\upsilon^2\Delta_{\text{in}}^2\log(m/\delta))/\alpha^4$. We also have $\|(\boldsymbol{\mu}_\ell - \widehat{\boldsymbol{\mu}}_j)(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)^\intercal(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)\mathbf{Q}_\ell^{1/2}\| \lesssim \sqrt{\beta}(\upsilon + \Delta_{\text{in}})(\upsilon + \Delta_{\text{in}}/\alpha^2)^2 \lesssim \sqrt{\beta}\upsilon^3\Delta_{\text{in}}^3/\alpha^4$. This implies that with probability at least $1-\delta$, we have $\left|(\boldsymbol{\mu}_\ell - \widehat{\boldsymbol{\mu}}_j)(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)^\intercal(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)\mathbf{Q}_\ell^{1/2}\right| \lesssim \frac{\sqrt{\beta}\upsilon^3\Delta_{\text{in}}^3}{\alpha^4}\log\frac{m}{\delta}$. We also have $\|(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)(\boldsymbol{\mu}_\ell - \widehat{\boldsymbol{\mu}}_j)\|^2 \lesssim \Delta_{\text{in}}^2(\upsilon^2 + \frac{\Delta_{\text{in}}^2}{\alpha^4}) \lesssim \frac{\upsilon^2\Delta_{\text{in}}^4}{\alpha^4}$. Combining all the bounds, for $\mathbf{x} \sim \mathcal{M}(U)$, with probability at least $1-\delta$, we have

$$\|(\widehat{\mathbf{K}}_j - \widehat{\mathbf{K}}_1)(\mathbf{x} - \widehat{\boldsymbol{\mu}}_j)\| \lesssim \frac{\beta\upsilon^3\Delta_{\text{in}}^4}{\alpha^4}\log\frac{m}{\delta}.$$

Combining this bound with (52), we obtain the result. ■

We now prove our main proposition of this section.

Proof [Proof of Theorem 40] We set $\widehat{\theta}_1 = c_1\frac{\beta\Delta_{\text{in}}^2\upsilon^2}{\alpha^2}\log\frac{m}{\delta}$ and $\widehat{\theta}_2 = c_2\frac{\beta\upsilon^3\Delta_{\text{in}}^4}{\alpha^4}\log\frac{m}{\delta}$ for some large constant c_1 and c_2.

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U)}[\|s(\mathbf{x};\mathcal{M}(U)) - p(\mathbf{x};\mathcal{M}(U))\mathbb{1}\{\widehat{B}(\mathbf{x},\widehat{\theta}_1,\widehat{\theta}_2) = 1\} - \widehat{\mathbf{K}}_1(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1)\mathbb{1}\{\widehat{B}(\mathbf{x},\widehat{\theta}_1,\widehat{\theta}_2) = 0\}\|^2] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U)}[\|s(\mathbf{x};\mathcal{M}(U)) - p(\mathbf{x};\mathcal{M}(U))\|^2\mathbb{1}\{\widehat{B}(\mathbf{x},\widehat{\theta}_1,\widehat{\theta}_2) = 1\}] \\ &\quad + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U)}[\|s(\mathbf{x};\mathcal{M}(U)) - \widehat{\mathbf{K}}_1(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1)\|^2\mathbb{1}\{\widehat{B}(\mathbf{x},\widehat{\theta}_1,\widehat{\theta}_2) = 0\}]\end{aligned}$$

Theorem 46 gives us that $|\widehat{V}_1^{(j)}(\mathbf{x})| \le \widehat{\theta}_1$ implies that $V^{(j)}(\mathbf{x}) \le \widehat{\theta}_1 + \omega_1$ for all $\mathbf{x}$ and for all $j \in U$ and hence, $B(\mathbf{x}, \widehat{\theta}_1 + \omega_1, \widehat{\theta}_2 + \omega_2)$. We apply Theorem 45 with ζ_1 as $c_1 \frac{\beta \Delta_{\text{in}}^2 \upsilon^2}{\alpha^2} \log \frac{m}{\delta}$ and ζ_2 as $c_2 \frac{\beta \upsilon^3 \Delta_{\text{in}}^4}{\alpha^4} \log \frac{m}{\delta}$ and obtain that there exist a polynomial $p(\mathbf{x}; \mathcal{M}(U))$ of degree $O(\frac{\beta^2 m^2 \upsilon^5 \Delta_{\text{in}}^6}{\alpha^6 \varepsilon} \log^2 \frac{m}{\delta})$ and coefficients bounded in magnitude by $dR \exp(\widetilde{O}(\frac{\beta^2 m^2 \upsilon^5 \Delta_{\text{in}}^6}{\alpha^6 \varepsilon} \log^2 \frac{m}{\delta}))$ such that the following holds:

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}(U)} [\|s(\mathbf{x}; \mathcal{M}(U)) - p(\mathbf{x}; \mathcal{M}(U))\|^2 \mathbb{1}\{\widehat{B}(\mathbf{x}, \widehat{\theta}_1, \widehat{\theta}_2) = 1\}] \lesssim \varepsilon.$$

We can upper bound the error when $\widehat{B}(\mathbf{x}, \widehat{\theta}_1, \widehat{\theta}_2) = 0$ using Cauchy-schwarz inequality as follows:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}(U)} [\|s(\mathbf{x}; \mathcal{M}(U)) - \widehat{\mathbf{K}}_1(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)\|^2 \mathbb{1}\{\widehat{B}(\mathbf{x}, \widehat{\theta}_1, \widehat{\theta}_2)\}] \\ =& \Big(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}(U)} [\|s(\mathbf{x}; \mathcal{M}(U)) - \widehat{\mathbf{K}}_1(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)\|^4] \Big)^{1/2} \big(\Pr[\widehat{B}(\mathbf{x}, \widehat{\theta}_1, \widehat{\theta}_2) = 0] \big)^{1/2} \end{aligned}$$

We know that $\Pr_{\mathbf{x} \sim \mathcal{M}(U)}[\widehat{B}(\mathbf{x}, \widehat{\theta}_1, \widehat{\theta}_2) = 0] \le \delta$. We upper bound the other term as follows:

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}(U)} [\|s(\mathbf{x}; \mathcal{M}(U)) - \widehat{\mathbf{K}}_1(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)\|^4] \le m^4 \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}} [\|\mathbf{Q}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \widehat{\mathbf{K}}_1(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)\|^4]. \tag{53}$$

Writing $\mathbf{x}$ in terms of standard Gaussian $\mathbf{z} \sim \mathcal{N}(0, \mathbf{Id})$ for any $i, \ell \in [m]$, we have

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{z} \sim \mathcal{N}} [\|\mathbf{Q}_i^{-1}(\mathbf{Q}_\ell^{1/2}\mathbf{z} + \boldsymbol{\mu}_\ell - \boldsymbol{\mu}_i) - \ \widehat{\mathbf{K}}_1(\mathbf{Q}_\ell^{1/2}\mathbf{z} + \boldsymbol{\mu}_\ell - \widehat{\boldsymbol{\mu}}_1)\|^4] \\ &\quad \lesssim \mathop{\mathbb{E}}_{\mathbf{z} \sim \mathcal{N}} [\|(\mathbf{Q}_i^{-1} - \widehat{\mathbf{K}}_1)\mathbf{Q}_\ell^{1/2}\mathbf{z}\|^4] + \|\mathbf{Q}_i^{-1}(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_i)\|^4 + \|\widehat{\mathbf{K}}_1(\boldsymbol{\mu}_\ell - \widehat{\boldsymbol{\mu}}_1)\|^4 \\ &\quad \lesssim \frac{\beta^2 \Delta_{\text{in}}^4}{\alpha^8} + \beta^2 \upsilon^4 + \frac{\Delta_{\text{in}}^4}{\alpha^4} + \frac{\Delta_{\text{in}}^2}{\alpha^4} \lesssim \frac{\beta^2 \upsilon^4 \Delta_{\text{in}}^4}{\alpha^8}, \end{aligned}$$

where the last inequality follows from Theorem 21 and $\|(\mathbf{Q}_i^{-1} - \widehat{\mathbf{K}}_1)\mathbf{Q}_\ell^{1/2}\|^4 \lesssim \beta^2(\|\mathbf{Q}_i^{-1} - \mathbf{Q}_1^{-1}\|_F^4 + \|\mathbf{Q}_1^{-1} - \widehat{\mathbf{K}}_1\|_F^4) \lesssim \frac{\beta^2 \Delta_{\text{in}}^4}{\alpha^8} + \beta^2 \upsilon^4$. Putting together the above bounds, we obtain that there exists a polynomial $p(\mathbf{x})$ such that

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}(U)} [\|s(\mathbf{x}; \mathcal{M}(U)) - p(\mathbf{x}; \mathcal{M}(U))\mathbb{1}\{B(\mathbf{x}; \mathcal{M}(U)) = 1\} - \widehat{\mathbf{K}}_1(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)\mathbb{1}\{B(\mathbf{x}\mathcal{M}(U)) = 0\}\|^2] \tag{54}$$

$$\lesssim \varepsilon + \sqrt{\delta} \frac{\beta^2 \upsilon^4 \Delta_{\text{in}}^4}{\alpha^8} \tag{55}$$

Choosing $\delta = \frac{\varepsilon^2 \alpha^{16}}{\beta^4 \widehat{\beta}^8 \Delta^8 \Delta_{\text{in}}^8}$, we obtain the result. ■

E.1.2. Piecewise polynomial approximation of the complete mixture

The goal of this section is to prove that there exists a piecewise polynomial that can approximate the $\mathsf{s}(\mathbf{x}; \mathcal{M})$. More precisely, there exists $\widehat{\mathsf{s}}(\mathbf{x}; \mathcal{M}(U_t))$ when used with the $\mathsf{c}(\cdot)$, $\widehat{\mathsf{s}}$ is ε-approximate to the true score function s, i.e.,

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{M}} \left[\|\mathsf{s}(\mathbf{x}; \mathcal{M}) - \widehat{\mathsf{s}}(\mathbf{x}, \mathsf{c}(\cdot))\|^2\right] \le \varepsilon,$$

where $\widehat{\mathbf{s}}(\mathbf{x}, \mathsf{c}(\cdot))$ is defined as

$$\widehat{\mathbf{s}}(\mathbf{x}, \mathsf{c}(\cdot)) = \sum_{t=1}^{n_c} \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))\, \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}$$

We will bound the error for every subset U_t. The error for the subset corresponding to U_t can be decomposed into an error due to the score simplification of $\mathcal{M}$ to $\mathcal{M}(U_t)$ and an error due to the approximation $\mathcal{M}(U_t)$ to the piecewise polynomial score function.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}) - \widehat{\mathbf{s}}(\mathbf{x}, \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right] \\
&\quad= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}) - s(\mathbf{x}, \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right] && (56)\\
&\quad\quad+ \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}, \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right]. && (57)
\end{aligned}$$

Recall that the score simplification (Theorem 32) bounds the term in (56). We rewrite (57) in two parts, when samples are coming from $\mathcal{M}(U_t)$ and $\mathcal{M}(U_t^c)$ as follows

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}, \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right] \\
&\quad= \Pr[j \in U_t] \cdot \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t)}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right] && (58)\\
&\quad\quad+ \Pr[j \in U_t^c] \cdot \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right] && (59)
\end{aligned}$$

The term in (58) is upper bounded by ε using Theorem 40. In the following Lemma, we upper bound the term in (59).

Lemma 47 *Let $\mathcal{M}$ be a (α, β, R)-well-conditioned mixture and let $U_t \subset [k]$ be a subset of components. Assume that the clustering function $\mathsf{c} : \mathbb{R}^d \to [n_c]$ satisfies $\Pr_{\mathbf{x}\sim\mathcal{N}_i}[\mathsf{c}(\mathbf{x}) = t] \le \delta$ for all $i \notin U_t$ and $t \in [n_c]$. Then, we have*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\|s(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right] \lesssim \frac{\beta^2}{\alpha^8} k^3 (\upsilon \Delta_{\text{in}} R)^4 \sqrt{\delta}\,.$$

Proof The term in (59) can be upper bounded by Cauchy-Schwarz inequality as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\|s(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right] \\
&\quad\le \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\|s(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))\|^4 \mathbb{1}\{\mathsf{c}(\mathbf{x}) = t\}\right]\right)^{1/2} \mathop{\Pr}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\mathsf{c}(\mathbf{x}) = t\right]^{1/2}.
\end{aligned}$$

Using the definition of $\widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))$, we can simplify the first term as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{s}}(\mathbf{x}; \mathcal{M}(U_t))\|^4\right] \\
&\quad= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) - p(\mathbf{x}; \mathcal{M}(U_t))\|^4 \mathbb{1}\{\widehat{B}(\mathbf{x}; U_t)\}\right] && (60)\\
&\quad\quad+ \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\|\mathbf{s}(\mathbf{x}; \mathcal{M}(U_t)) - \widehat{\mathbf{K}}_i(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)\|^4 \mathbb{1}\{\widehat{B}^c(\mathbf{x}; U_t)\}\right]
\end{aligned}$$

The first term in (60) is upper bounded by ε^4. The second term in (60) can be upper bounded by

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}[\|\mathbf{s}(\mathbf{x};\mathcal{M}(U_t)) - \widehat{\mathbf{K}}_i(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)\|^4] \le k^3 \sum_{j\in U_t} \mathbb{E}_{\mathbf{x}\sim\mathcal{M}(U_t^c)}[\|\mathbf{Q}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - \widehat{\mathbf{K}}_i(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)\|^4].$$

We can upper bound $\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}_\ell,\mathbf{Q}_\ell)$ by writing it in terms of the standard normal $\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mu_\ell,\mathbf{Q}_\ell)}[\|\mathbf{Q}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - \widehat{\mathbf{K}}_i(\mathbf{x}-\widehat{\boldsymbol{\mu}}_i)\|^4] \\
&\lesssim \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{Id})}\left[\|(\mathbf{Q}_j^{-1} - \widehat{\mathbf{K}}_i)\mathbf{Q}_\ell^{1/2}\mathbf{z}\|^4\right] + \|\mathbf{Q}_j^{-1}(\boldsymbol{\mu}_\ell-\boldsymbol{\mu}_j)\|^4 + \|\widehat{\mathbf{K}}_i(\boldsymbol{\mu}_\ell-\widehat{\boldsymbol{\mu}}_i)\|^4 \\
&\lesssim \|(\mathbf{Q}_j^{-1} - \widehat{\mathbf{K}}_i)\mathbf{Q}_\ell^{1/2}\|^4 + \frac{R^4}{\alpha^4} + \widehat{\beta}^4(R^4+\Delta^4) \\
&\lesssim \beta^2\left(\frac{\Delta_{\text{in}}^4}{\alpha^8} + \upsilon^4\right) + \frac{R^4}{\alpha^4} + \frac{(R^4+\upsilon^4)}{\alpha^4} \\
&\lesssim \frac{\beta^2}{\alpha^8}(\upsilon\Delta_{\text{in}}R)^4.
\end{aligned}$$

Additionally, we have

$$\Pr_{\mathbf{x}\sim\mathcal{M}(U_t^c)}\left[\mathsf{c}(\mathbf{x})=t\right] \le \max_{j\in[k]:j\notin U_t} \Pr_{\mathbf{x}\sim\mathcal{N}_j}[\mathsf{c}(\mathbf{x})=t] \le \delta.$$

Combining Equation (60) with the above bound, we obtain the result. ∎

Proposition 48 *Let $\mathcal{M}$ be (α,β,R)-well-conditioned mixture and then, there exists a piecewise polynomial $\widehat{\mathsf{s}}(\mathbf{x};\mathsf{c}(\cdot))$ such that it satisfies*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x};\mathcal{M}) - \widehat{\mathsf{s}}(\mathbf{x},\mathsf{c}(\cdot))\|^2\right] \le \varepsilon,$$

where $\widehat{\mathsf{s}}(\mathbf{x},\mathsf{c}(\cdot))$ is defined as

$$\widehat{\mathsf{s}}(\mathbf{x},\mathsf{c}(\cdot)) = \sum_{t=1}^{n_{\mathrm{c}}} \widehat{\mathsf{s}}(\mathbf{x};\mathcal{M}(U_t))\,\mathbb{1}\{\mathsf{c}(\mathbf{x})=t\}$$

and $$\begin{aligned}\widehat{\mathsf{s}}(\mathbf{x};\mathcal{M}(U_t)) &= p(\mathbf{x};\mathcal{M}(U_t))\mathbb{1}\{\widehat{B}(\mathbf{x};\mathcal{M}(U_t))=1\} \\ &\quad + \widehat{\mathbf{K}}_j(\mathbf{x}-\widehat{\boldsymbol{\mu}}_j)\mathbb{1}\{\widehat{B}(\mathbf{x};\mathcal{M}(U_t))=0\} \quad \text{for some } j\in U_t \text{ and } \widehat{B} \text{ defined in Equation (50)}\end{aligned}$$

Moreover, every polynomial $p(\mathbf{x};\mathcal{M}(U_t))$ has the degree at most $\mathrm{poly}(\frac{\beta k}{\alpha\lambda_{\min}\varepsilon}\log R)$ and coefficients of the polynomials are bounded in magnitude by $\mathrm{poly}(d)\exp(\mathrm{poly}(\frac{\beta k}{\alpha\lambda_{\min}\varepsilon}\log R))$.

Proof Combining Equation (58), Equation (59) and Theorem 47, for a fixed $t\in[n_{\mathrm{c}}]$, we have

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x};\mathcal{M}(U_t)) - \widehat{\mathsf{s}}(\mathbf{x},\mathcal{M}(U_t))\|^2\mathbb{1}\{\mathsf{c}(\mathbf{x})=t\}\right] \lesssim \varepsilon + \frac{\beta^2}{\alpha^8}k^3(\upsilon\Delta_{\text{in}}R)^4\sqrt{\delta}. \tag{61}$$

We now combine the bound of the above equation with the score simplification guarantee. The score simplification guarantee (Theorem 32) assumes that the clustering function $\mathsf{c}:\mathbb{R}^d\to[n_{\mathrm{c}}]$ satisfies $\Pr_{\mathbf{x}\sim\mathcal{N}_i}[\mathsf{c}(\mathbf{x})=t]\le\delta$ for all $i\notin U_t$ and $t\in[n_{\mathrm{c}}]$. and obtains that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{M}}[\|s(\mathbf{x};\mathcal{M}) - s(\mathbf{x};\mathsf{c}(\cdot))\|_2^2] \le O(k^{5/4}(\beta^3/\alpha^5)R)\sqrt{\delta},$$

Combining the above bound with Equation (61), we have

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x};\mathcal{M})-\widehat{\mathbf{s}}(\mathbf{x},\mathsf{c}(\cdot))\|^2\right]\lesssim k\varepsilon+\frac{\beta^3}{\alpha^8}k^3(\upsilon\Delta_{\text{in}}R)^4\sqrt{\delta}. \tag{62}$$

Using clustering guarantee from Theorem 25 for any $t\in[n_{\mathrm{c}}], i\in U_t$ and $t'\in[n_{\mathrm{c}}]$ and $t'\neq t$, we have

$$\Pr_{\mathbf{x}\sim\mathcal{N}_i}[\mathsf{c}(\mathbf{x})=t']\le\Pr_{\mathbf{x}\sim\mathcal{N}_i}[\mathsf{c}(\mathbf{x})\neq t]\le k^3\exp\Big(-\Omega\Big(\frac{(\Delta_{\text{out}}^{(\boldsymbol{\mu})})^2}{\alpha\sqrt{k}}\wedge\frac{\alpha^6(\Delta_{\text{out}}^{(\mathbf{Q})})^2}{\beta^6\upsilon_{cov}^2}\wedge\frac{\alpha^2\Delta_{\text{out}}^{(\mathbf{Q})}}{\beta^3}\Big)\Big).$$

Recall that $\upsilon_{mean}\lesssim\beta/\lambda_{\min}$ and $\upsilon_{cov}\lesssim k^{3/2}\beta/\lambda_{\min}+k^2\alpha\log R$. Therefore, we choose $\Delta_{\text{out}}^{(\boldsymbol{\mu})}$ and $\Delta_{\text{out}}^{(\mathbf{Q})}$ for some large constants c_1 and c_2 as follows which satisfies the conditions in Theorem 25.

$$\Delta_{\text{out}}^{(\boldsymbol{\mu})}=c_1\frac{\beta\sqrt{k}}{\lambda_{\min}}\log\frac{kR\beta}{\lambda_{min}\alpha\varepsilon}\quad\text{ and }\quad\Delta_{\text{out}}^{(\mathbf{Q})}=c_2\frac{\beta^4k^2\log R}{\alpha^3\lambda_{\min}}\log\frac{kR\beta}{\lambda_{min}\alpha\varepsilon}.$$

We also choose $\Delta_{\text{in}}^{(\boldsymbol{\mu})}\asymp k\Delta_{\text{out}}^{(\boldsymbol{\mu})}$ and $\Delta_{\text{in}}^{(\mathbf{Q})}\asymp k\Delta_{\text{out}}^{(\mathbf{Q})}$. Using the chosen values of $\Delta_{\text{out}}^{(\boldsymbol{\mu})}$ and $\Delta_{\text{out}}^{(\mathbf{Q})}$, we have

$$\Pr[\mathsf{c}(\mathbf{x})=t\mid j\notin U_t]\le\varepsilon^2\,\mathrm{poly}\Big(\frac{\alpha\lambda_{min}}{\beta kR}\Big).$$

Using this bound in Equation (60), we have

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{M}}\left[\|\mathbf{s}(\mathbf{x};\mathcal{M})-\widehat{\mathbf{s}}(\mathbf{x},\mathsf{c}(\cdot))\|^2\right]\lesssim k\varepsilon.$$

Rescaling ε as ε/k and using $\Delta_{\text{in}}=\Delta_{\text{in}}^{(\boldsymbol{\mu})}+\Delta_{\text{in}}^{(\mathbf{Q})}$ and $\upsilon=\upsilon_{mean}+\upsilon_{cov}$ in Theorem 40, we obtain the result. ∎

E.2. Learning polynomials using denoising objective

The goal of this section is to provide details about our learning algorithm using denoising objective. Recall that to sample from the data distribution, the diffusion reverse process uses an approximation to the score function $\nabla_{\mathbf{x}}\log q_t(\mathbf{x})$. To learn the score function, we minimize the following DDPM objective in which one wants to predict the noise $\mathbf{z}_t$ from the noisy observation $\mathbf{x}_t$, i.e.

$$\min_{\mathbf{g}\in\mathcal{G}}\ L_t(\mathbf{g}_t)=\mathop{\mathbb{E}}_{\mathbf{x}_0,\mathbf{z}_t}\left[\left\|\mathbf{g}_t(\mathbf{x}_t)+\frac{\mathbf{z}_t}{\sqrt{1-\exp(-2t)}}\right\|^2\right]. \tag{63}$$

Given parameter candidates $\{(\widehat{\boldsymbol{\mu}}_i,\widehat{\mathbf{Q}}_i)\}_{i=1}^k$ and a clustering function $\mathsf{c}(\cdot)$, our learning algorithm minimizes the following empirical loss

$$\begin{aligned}&\min_{\substack{p(\mathbf{x};\mathcal{M}(U_i))\\ \forall i\in[k]}}\ \frac{1}{n}\sum_{i=1}^n L_t^{(clip)}(\widehat{\mathbf{s}}_t,\mathbf{x}_t^{(i)},\mathbf{z}_t^{(i)})\\ &\text{where }=\left\|\widehat{\mathbf{s}}_t(\mathbf{x}_t,\mathsf{c}(\cdot))+\frac{\mathbf{z}_t}{\sqrt{1-\exp(-2t)}}\right\|^2\mathbb{1}\{\|\mathbf{x}_t\|\le R_{\mathbf{x}},\|\mathbf{z}_t\|\le R_{\mathbf{z}}\},\end{aligned} \tag{64}$$

for some large choices of $R_{\mathbf{x}},R_{\mathbf{z}}=\mathrm{poly}(dR\tau/\varepsilon)^{\ell}$. Clipping the loss for large values of $\|\mathbf{x}_t\|$ and $\|\mathbf{z}_t\|$ is for analysis purposes and in fact, we show that the choice of the value of $R_{\mathbf{x}}$ and $R_{\mathbf{z}}$ are sufficiently large such that the unclipped loss will be at most $O(\varepsilon)$ in expectation.

Proposition 49 *Let $\mathcal{M}$ be a (α, β, R)-well-conditioned mixture. Then, for any $\varepsilon, \delta > 0$ and noise scale $t \geq \varepsilon$, there exists an algorithm that runs in $O(d^{\mathrm{poly}(\frac{\beta k \log R}{\alpha \varepsilon \lambda_{\min}})}\mathrm{poly}(\log \frac{1}{\delta}))$ and returns a score function $\widehat{\mathbf{s}}_t$ such that with probability $1-\delta$ over samples generated from the mixture $\mathcal{M}$, we have*

$$\mathop{\mathbb{E}}_{\mathbf{x}_t \sim \mathcal{M}_t}\left[\|\widehat{\mathbf{s}}_t(\mathbf{x}_t) - \nabla_{\mathbf{x}} \log q_t(\mathbf{x})\|^2\right] \leq \varepsilon. \tag{65}$$

The algorithm to learn the score function takes input as noise scale t, target error ε and confidence δ and it is given by

- *Obtain a candidate list of parameters $\mathcal{W} \leftarrow$ CRUDEESTIMATE*
- *Brute force over the parameter candidate list $(\widehat{\boldsymbol{\mu}}_1, \widehat{\mathbf{Q}}_1) \ldots (\widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{Q}}_k) \in \mathcal{W}$*
  - *Brute force over number of mean-based partition (m), number of covariance-based partition (n), mean-based partition $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ and covariance-based partition $\mathcal{T} = \{T_1, \ldots, T_n\}$*
    - *Brute force over possible thresholds $\{t_{ij}\}_{i,j=1}^k$ in range $[-c\frac{\beta d}{\alpha}, c\frac{\beta d}{\alpha}]$ for some large constant c.*
      - *Clustering function $\mathsf{c} \leftarrow$ CLUSTERING$(S, T, \{(\widehat{\boldsymbol{\mu}}_i, \widehat{\mathbf{Q}}_i)\}_{i=1}^k, \{t_{i,j}\}_{i,j=1}^k)$*
      - *$\widehat{\mathbf{s}}_t \leftarrow$ minimizer of empirical loss Equation (64).*
      - *Compute the validation loss on the fresh samples for $\widehat{\mathbf{s}}_t$.*

In the end, the algorithm returns the $\widehat{\mathbf{s}}_t$ which has minimum validation loss across all brute force candidates.

E.3. Generalization error analysis

As we can decompose the learning problem into learning a polynomial in the piece given by the clustering function $\mathsf{c}(\cdot)$, we can start the generalization error argument by considering the loss function restricted to a fixed piece of the polynomial.

Observe that the DDPM objective can be unbounded in general however, the loss becomes bounded assuming that $\|\mathbf{x}_t\| \leq R_{\mathbf{x}}$ and $\|\mathbf{z}_t\| \leq R_{\mathbf{z}}$. Therefore, we first derive the generalization error bound when we restrict the loss function to points $\|\mathbf{x}_t\| \leq R_{\mathbf{x}}$ and $\|\mathbf{z}_t\| \leq R_{\mathbf{z}}$ and then argue that the points outside of this region follow with a small probability because of the sub-Gaussian tail of the mixture model outside an appropriate radius.

To simplify the notation, we define the clipped loss and clipped loss restricted to a particular piece as

$$L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t) = \left\|\widehat{\mathbf{s}}(\mathbf{x}_t, \mathsf{c}(\cdot)) + \frac{\mathbf{z}_t}{\sqrt{1-\exp(-2t)}}\right\|^2 \mathbb{1}\{\|\mathbf{x}_t\| \leq R_{\mathbf{x}}, \|\mathbf{z}_t\| \leq R_{\mathbf{z}}\}$$

$$L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t, U_i, \widehat{B}) = \left\|\widehat{\mathbf{s}}(\mathbf{x}_t, \mathsf{c}(\cdot)) + \frac{\mathbf{z}_t}{\sqrt{1-\exp(-2t)}}\right\|^2 \mathbb{1}\{\mathsf{c}(\mathbf{x}_t) = i, \widehat{B}(\mathbf{x}_t, U_i), \|\mathbf{x}_t\| \leq R_{\mathbf{x}}, \|\mathbf{z}_t\| \leq R_{\mathbf{z}}\}.$$

Similarly define $L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t, U_i, \widehat{B}^c)$ by replacing $\widehat{B}$ with $\widehat{B}^c$. Recall that for the region where $\mathbb{1}\{\mathsf{c}(\mathbf{x}_t) = i, \widehat{B}(\mathbf{x}_t, U_i)\} = 1$, $\widehat{\mathbf{s}}(\mathbf{x}_t, \mathsf{c}(\cdot))$ is simplified to $p_j(\mathbf{x}_t)$.

Lemma 50 (Sample complexity) *Assume that the sum of absolute values of the coefficient of the polynomial is M. Then, choosing $R_{\mathbf{x}}, R_{\mathbf{z}} = \Theta((\beta R d/\alpha)\log(1/\delta'))$ for some $\delta' > 0$ and taking number of samples $n \geq \mathrm{poly}(\frac{dMR\beta}{\alpha\varepsilon t_{\min}}\log\frac{1}{\delta})\mathrm{poly}(\frac{d\beta R}{\alpha}\log\frac{1}{\delta'})^{\ell}$, with probability at least $1-\delta$ over samples, we have*

$$\mathop{\mathbb{E}}_{\mathbf{x}_t,\mathbf{z}_t}[L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t)] \leq \frac{1}{n}\sum_{i=1}^{n} L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t^{(i)}, \mathbf{z}_t^{(i)}) + \varepsilon.$$

Proof Denote $\boldsymbol{\theta}$ as coefficients of the polynomials and $\phi(\mathbf{x})$ denote the monomials up to degree ℓ. Then, we know that $\|\boldsymbol{\theta}\|_2 \leq \|\boldsymbol{\theta}\|_1 \leq M$. Additionally, the bound on $\|\mathbf{x}\|$ implies that $\|\phi(\mathbf{x})\|_\infty \lesssim R_{\mathbf{x}}^{\ell}$. This implies that $\|\phi(\mathbf{x})\|_2 \lesssim (dR_{\mathbf{x}})^{\ell}$. The Lipschitz constant $L_t^{(clip)}$ for each coordinate can be upper bounded by

$$\|\nabla L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t, U_i, \widehat{B})\| \lesssim \frac{dMR_{\mathbf{z}}(dR_{\mathbf{x}})^{\ell}}{\sqrt{1-\exp(-2t)}} \leq \frac{dMR_{\mathbf{z}}(dR_{\mathbf{x}})^{\ell}}{\sqrt{t_{\min}}}.$$

Additionally, we have $\|L_t^{(clip)}\| \leq \frac{(dMR_{\mathbf{z}})^2(dR_{\mathbf{x}})^{2\ell}}{t_{\min}}$ for any $\|\mathbf{x}_t\| \leq R_{\mathbf{x}}$ and $\|\mathbf{z}_t\| \leq R_{\mathbf{z}}$. We choose $R_{\mathbf{x}}, R_{\mathbf{z}} \asymp \frac{\beta R d}{\alpha}\log(1/\delta')$ for some $\delta' > 0$ and apply standard generalization error analysis result using Rademacher complexity for linear function class (e.g., see Shalev-Shwartz and Ben-David (2014)). If we choose the total number of samples n to satisfy $n \geq \frac{(dMR_{\mathbf{z}})^4(dR_{\mathbf{x}})^{4\ell}}{t_{\min}^2\varepsilon^2}\log\frac{1}{\delta}$,then with at least $1-\delta$ probability, we have

$$\mathop{\mathbb{E}}_{\mathbf{x}_t,\mathbf{z}_t}[L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t, S_j, \widehat{B})] \leq \frac{1}{n}\sum_{i=1}^{n} L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t^{(i)}, \mathbf{z}_t^{(i)}, S_j, \widehat{B}) + \varepsilon$$

for all $j \in [n_c]$. Using a similar argument to prove the boundedness of $L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t, S_j, \widehat{B}^c)$, we also obtain

$$\mathop{\mathbb{E}}_{\mathbf{x}_t,\mathbf{z}_t}[L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t, S_j, \widehat{B}^c)] \leq \frac{1}{n}\sum_{i=1}^{n} L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t, S_j, \widehat{B}^c) + \varepsilon.$$

Because $\mathbb{1}\{\mathsf{c}(\mathbf{x}_t) = j\}$ for any single j for all $\mathbf{x}_t$, combining these bounds for all $j \in [n_c]$ for $n \geq$, we have

$$\mathop{\mathbb{E}}_{\mathbf{x}_t,\mathbf{z}_t}[L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t)] \leq \frac{1}{n}\sum_{i=1}^{n} L_t^{(clip)}(\widehat{\mathbf{s}}, \mathbf{x}_t^{(i)}, \mathbf{z}_t^{(i)}) + \varepsilon\ .$$

■

Proposition 51 *Let $\mathcal{M}$ be an (α,β,R)-well-conditioned mixture. Then, for any $\varepsilon > 0$ and any noise scale $t \geq t_{\min} \geq \alpha\varepsilon/R$, there exist an algorithm that takes number of samples $n \geq (\log\frac{1}{\delta})d^{\mathrm{poly}(\frac{\beta k \log R}{\alpha\varepsilon\lambda_{\min}})}$ and runs in sample-polynomial time and returns a score function $\widehat{\mathbf{s}}_t$ such that*

$$\mathop{\mathbb{E}}_{\mathbf{x}_t}[\|\nabla_{\mathbf{x}}\log q_t(\mathbf{x}_t) - \widehat{\mathbf{s}}_t(\mathbf{x}_t)\|^2] \leq \varepsilon\ .$$

Proof We define the loss function outside the radius $\|\mathbf{x}_t\| \geq \frac{\beta R d}{\alpha} \log \frac{1}{\delta'}$ or $\|\mathbf{z}_t\| \geq \frac{\beta R d}{\alpha} \log \frac{1}{\delta'}$ as

$$L_t^{(out)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t) = \left\| \widehat{\mathbf{s}}(\mathbf{x}_t, \mathsf{c}(\cdot)) + \frac{\mathbf{z}_t}{\sqrt{1-\exp(-2t)}} \right\|^2 \mathbb{1}\{\|\mathbf{x}_t\| \geq \frac{\beta R d}{\alpha} \log \frac{1}{\delta'} \vee \|\mathbf{z}_t\| \geq \frac{\beta R d}{\alpha} \log \frac{1}{\delta'}\}$$

The $L^{(out)}$ can be simplified as

$$\left| \mathop{\mathbb{E}}_{\mathbf{x}_t, \mathbf{z}_t} [L^{(out)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t)] \right| \lesssim \mathop{\mathbb{E}}_{\mathbf{x}_t} \Big[\|\widehat{\mathbf{s}}(\mathbf{x}_t, \mathsf{c}(\cdot))\|^2 \cdot \mathbb{1}\Big\{ \|\mathbf{x}_t\| \geq \frac{\beta R d}{\alpha} \log \frac{1}{\delta'} \Big\} \Big] \tag{66}$$

$$+ \frac{1}{t_{\min}} \mathop{\mathbb{E}}_{\mathbf{z}_t} \Big[\|\mathbf{z}_t\|^2 \cdot \mathbb{1}\Big\{ \|\mathbf{z}_t\| \geq \frac{\beta R d}{\alpha} \log \frac{1}{\delta'} \Big\} \Big]. \tag{67}$$

The second term in the above equation can be upper bounded by $(\Pr\{\|\mathbf{z}_t\| \geq R_{\mathbf{z}}\})^{1/2} (\mathbb{E}\left[\|\mathbf{z}_t\|^4\right])^{1/2} \lesssim \sqrt{\delta} d$. To upper-bound the first term, we first upper-bound $\mathbb{E}_{\mathbf{x}_t}[\|p(\mathbf{x}_t, \mathcal{M}(S_j))\|^4]$:

$$\mathbb{E}\left[\|p(\mathbf{x}_t, \mathcal{M}(S_j))\|^4\right] \leq M^4 \mathop{\mathbb{E}}_{\mathbf{x}_t} \left[\|\phi(\mathbf{x}_t)\|_1^4\right] \leq M^4 d^\ell \left(\max_{\mathbf{v}:\|\mathbf{v}\|_1 \leq 4\ell} \mathbb{E}\left[\prod_{i=1}^d |\mathbf{x}_t^{(i)}|^{\mathbf{v}_i} \right] \right)$$

$$\leq M^4 d^\ell \max_{\mathbf{v}:\|\mathbf{v}\|_1 \leq 4\ell} \prod_{i=1}^d \left(\mathbb{E}\left[|\mathbf{x}_t^{(i)}|^{\mathbf{v}_i d} \right] \right)^{1/d}$$

Using Gaussian hypercontractivity (Theorem 39), we can simplify $\mathbb{E}\left[|\mathbf{x}_t^{(i)}|^{\mathbf{v}_i d}\right] \lesssim \sum_{i=1}^k \lambda_i (\mathbf{v}_i d)^{\mathbf{v}_i d} (\beta R)^{\mathbf{v}_i d} \leq (4\ell d \beta R)^{4\ell d}$. Using this bound in (66), we have

$$\left| \mathop{\mathbb{E}}_{\mathbf{x}_t, \mathbf{z}_t} [L^{(out)}(\widehat{\mathbf{s}}, \mathbf{x}_t, \mathbf{z}_t)] \right| \lesssim \sqrt{\delta'/t_{\min}} d + M^4 d^\ell (4\ell d \beta R)^{4\ell} \sqrt{\delta'}.$$

Choosing $\delta' = \mathrm{poly}(\varepsilon t_{min}/(dM(4\ell d\beta R)^\ell)$, we obtain the result. ■