A Quasi-Polynomial Time Mean Estimator Under Mean-Shift Contamination with Unknown Covariance

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin–Madison

Jingyi Gao JINGYIG@CS.WISC.EDU
University of Wisconsin–Madison

Giannis Iakovidis IAKOVIDIS@WISC.EDU
University of Wisconsin–Madison

Daniel M. Kane DAKANE@UCSD.EDU
University of California San Diego

Sihan Liu SIL046@UCSD.EDU
University of California San Diego

Thanasis Pittas PITTAS@WISC.EDU
University of Wisconsin–Madison

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study the algorithmic problem of robust Gaussian mean estimation in the mean-shift contamination model with unknown covariance. Specifically, we are allowed to draw samples from a statistical mixture of an unknown target Gaussian $\mathcal{N}(\mu, \Sigma)$ (with weight at least $1-\alpha$), and arbitrary (unknown) mean-shifts of it, i.e., $\{\mathcal{N}(\mu_i, \Sigma)\}_i$, and the goal is to estimate μ up to any desired accuracy ϵ in ℓ_2-norm. In the special case where Σ is known to be the identity, prior work gave an algorithm with a near-optimal sample complexity of $\mathrm{poly}(d, 2^{\epsilon^{-2}})$ and sample-polynomial time. In this work, we provide a quasi-polynomial time algorithm with sample complexity $2^{\mathrm{poly}(\log d/\epsilon)}$ in the more general unknown covariance case, markedly improving upon the only previously known estimator for this setting that incurs exponential runtime.

Keywords: mean estimation, robustness, mean-shift contamination

1. Introduction

Robust statistics studies statistical estimation in the presence of data contamination. Originating in the pioneering works of Huber (1992); Tukey (1960), the field developed minimax-optimal robust estimators for a wide range of structured distributions. More recently, a line of work starting from the theoretical computer science community (Diakonikolas et al., 2016; Lai et al., 2016) has focused on the computational complexity of robust estimation for high dimensional statistical tasks, including but not limited to mean/covariance estimation (Diakonikolas et al., 2017, 2020a; Hopkins and Li, 2019; Cheng et al., 2019; Diakonikolas et al., 2020b, 2022; Kane et al., 2024), learning mixture models (Charikar et al., 2017; Li and Schmidt, 2017; Steinhardt et al., 2018; Diakonikolas et al., 2018), and linear regression (Bhatia et al., 2017; Prasad et al., 2019; Klivans et al., 2018; Bakshi and Prasad, 2021; Das et al., 2023; Diakonikolas et al., 2025b). See Diakonikolas and Kane (2023) for a textbook regarding the recent developments.

The classical model in robust statistics, known as Huber's contamination model (Huber, 1964), however, suffers from inherent information-theoretic limitations. Specifically, as it allows for an arbitrary α-fraction of outlier data points, no estimator can achieve an error better than $\Omega(\alpha)$ regardless of the sample size for many natural structured distributions, e.g., Gaussians, uniform distributions, distributions supported on the boolean hypercube, etc. (see Section 1 of Diakonikolas and Kane (2023)). To achieve consistency, where the estimation error vanishes as the sample size increases, further assumptions are needed on the specific types of corruption. In this work, we focus on the mean-shift contamination model, where the outlier distribution consists of a statistical mixture of arbitrary mean-shifts of a clean base distribution.

The mean-shift contamination model has recently received attention from both the statistics (Cai and Jin, 2010; Collier and Dalalyan, 2019; Carpentier et al., 2021; Kotekal and Gao, 2025) and the computer science (Li, 2023; Diakonikolas et al., 2025a, 2026) communities, and is related to large-scale hypothesis testing (Efron, 2004, 2007, 2008), where the data is often assumed to follow a mean-shift version of the null distribution in the alternative hypothesis.

For univariate random variables including Gaussian and Laplace distributions, Li (2023); Kotekal and Gao (2025) provide minimax optimal estimators. While these estimators can be extended using standard covering techniques to handle high-dimensional distributions, the extension often comes with a cost in exponential blowup in the runtime of the estimator (see Section 2.1 of Diakonikolas et al. (2025a) for an example of the covering technique).

On the high-dimensional algorithmic side, Diakonikolas et al. (2025a) provide the first sample-polynomial time estimators for identity-covariance Gaussian distributions with a sample complexity of $\tilde{O}\left(d/\epsilon^{2+o(1)} + 2^{O(1/\epsilon^2)}\right)$. The main question left open by their work is how to algorithmically address the more general and realistic setting where the shared covariance Σ is unknown and arbitrary.

In this work, we take the first step towards answering this open question. In particular, we study the following Gaussian Mean-Shift Contamination Model with *unknown covariance*.

Definition 1 (Gaussian Mean-Shift Contamination Model with Unknown Covariance) *We say that a distribution D is an α-mean-shift corrupted distribution with clean mean μ and shared covariance Σ if it is a statistical mixture of the form $(1-\alpha)\mathcal{N}(\mu,\Sigma) + \frac{\alpha}{n}\sum_{i=1}^{n}\mathcal{N}(\mu_i,\Sigma)$,* [1] *where $\{\mu_i\}_{i=1}^{n}$ are arbitrary mean vectors.*

Our main result is the first algorithm in the *unknown* covariance setting with quasi-polynomial sample/runtime complexities in the dimension d.

Theorem 2 (Main Algorithmic Result) *Let D be an unknown α-mean-shift corrupted distribution with clean mean $\mu \in \mathbb{R}^d$ and shared positive definite covariance $\Sigma \in \mathbb{R}^{d\times d}$. Assume that $\alpha \in (0,1/2)$ is at most a sufficiently small universal constant. Then there exists an algorithm that takes as input an accuracy parameter $\epsilon \in (0,1/2)$, draws $m = 2^{O(\log d/\epsilon)^8}$ i.i.d. samples from $\mathcal{N}(\mu,\Sigma)$ under the mean-shift model (Definition 1), runs in $\mathrm{poly}(m,d)$ time, and outputs $\widehat{\mu}$ such that with probability at least $2/3$* [2] *it holds that $\|\mu-\widehat{\mu}\| \le \epsilon\sqrt{\|\Sigma\|_2}$.*

1. Neither the runtime nor the sample complexity of our algorithm depends on the number of mixture components n. Hence, n here can be taken to be arbitrarily large.
2. The success probability can be boosted to $1-\tau$ by running the algorithm for $O(\log(1/\tau))$ many times and returning the candidate that minimizes the sum of the ℓ_2 distances to the rest.

Since the information-theoretic lower bounds of Kotekal and Gao (2025) show that even univariate Gaussian mean estimation under the mean-shift contamination model requires $2^{\mathrm{poly}(1/\epsilon)}$ samples, our algorithm's ϵ-dependence is qualitatively tight. Moreover, when ϵ is treated as a fixed constant, our algorithm achieves a sample complexity and runtime of $2^{\mathrm{polylog}(d)}$, marking the first quasi-polynomial time guarantee in the high-dimensional regime for unknown covariance. The only previously known consistent estimators for this problem required runtime exponential in d.

1.1. Technical Overview

For simplicity, we will assume that $\|\mu\|_2 \le O(1)$ and $\|\Sigma - I\|_F \le O(1)$ throughout this subsection. It is easy to reduce to this case by first invoking an adversarially robust efficient Gaussian mean estimator from the robust statistics literature (see Lemma 6) and then normalizing the input distribution with the estimates. Besides, we will also focus on the regime $d \gg \mathrm{poly}(1/\epsilon)$ as otherwise one could just run a brute-force algorithm in the full space.

Prior Techniques with Identity Covariance We begin by recalling the analysis for the identity covariance case, i.e., $\Sigma = I$. At a high level, the framework of Diakonikolas et al. (2025a), which is also the starting point of this work, is to utilize *data-dependent dimensionality reduction* to isolate a low-dimensional subspace, where the mean μ retains a significant projection, before applying a brute-force generalization of the $1d$ estimator developed in Kotekal and Gao (2025). To achieve the goal, the algorithm proceeds to iteratively trim subspaces in which μ provably does not have a large projection. The idea behind the iterative trimming approach is the following observation.

Proposition 3 (Mean Size Certification (Informal)) *If all components of D have identity covariance and the (uncentered) second moment of D along a direction v is close to 1, then the mean μ of the clean component must be small along v.*

Thus, if one can identify a subspace V in which the second moment matrix is close to identity in spectral norm, V can then be safely removed. An immediate obstacle is that the outlier components within the original mixture D may inflate the second moment matrix along nearly all directions. To reduce their impact, Diakonikolas et al. (2025a) propose to reweight each sample $x \in \mathbb{R}^d$ by the exponential function $\exp(-\|x\|_2^2/(\sqrt{d}C_{\mathrm{w}}))$ for some carefully chosen parameter $C_{\mathrm{w}} = \mathrm{poly}(\epsilon)$ controlling the strength of the reweighting. This effectively suppresses extreme outlier components whose means are significantly larger than $d^{1/4}C_{\mathrm{w}}^{1/2}$, and consequently ensures that the reweighted second moment matrix will be equal to the identity matrix (contributed by the clean component) plus a PSD matrix of trace at most $\sqrt{d}C_{\mathrm{w}} \ll d$ (contributed by the outlier components that are relatively close to the origin). It is then not hard to show the existence of a non-trivial subspace in which the reweighted second moment matrix is spectrally close to the identity, suggesting that it can be safely removed.[3]

Covariance-Agnostic Mean Certification via Moment Matching While our algorithm for the unknown Σ setting retains the original framework, it encounters a significant hurdle: the foundational observation in Proposition 3 breaks down when individual mixture components possess non-identity covariances. To see this, recall that the second moment can be alternatively written

3. While a naive assessment might suggest that the approach only bounds the reweighted mean of the clean component within the removed subspace V, we can ensure the reweighting bias remains negligible by strategically selecting the parameter C.

as the square of the mean plus the variance; consequently, a substantial projection of mean vector μ along a direction v can be masked if the corresponding variance $v^\top \Sigma v$ is sufficiently small. To fix the issue, we build a more robust covariance-agnostic certification leveraging higher moment information:

Proposition 4 (Covariance-Agnostic Mean Size Certification (Informal)) *If the first 4 moments of D along a direction v are close to those of a standard Gaussian, then $v^\top \mu$ must be small.*

See Proposition 12 for the formal statement. Before sketching the proof of this structural result, we first show how one could complete the algorithm with it. Similar to the prior work, we will reweight the samples with the exponential function $\exp(-\|x\|_2^2/(\sqrt{d}C_{\mathrm{w}}))$ with $C_{\mathrm{w}} = \mathrm{poly}(\epsilon)$ so as to remove the outlier components with mean more than $d^{1/4}\mathrm{poly}(\epsilon)$. In this way, to introduce $\delta = \mathrm{poly}(\epsilon)$ amount of deviation along a direction v on the degree-4 moment, the adversary has to assign at least $w \gg 1/d$ mass on some outlier component with mean $d^{1/4}\mathrm{poly}(\epsilon)v$. Consequently, the adversary could only hope to introduce significant corruption to the moment tensor on a subspace of dimension at most $\ll d$. This therefore ensures the existence of a safe-to-remove subspace of dimension $\Omega(d)$, where the first degree-4 moments are all $\mathrm{poly}(\epsilon)$ close to the standard Gaussian moments. [4]

Moment Matching Subspace Identification It then remains for us to reliably identify such a subspace. We first consider the ideal case where we have perfect estimates of the reweighted low-degree moment tensor T of D. Note that unlike the case of second moments, there is strong evidence showing that it is computationally challenging to find even a single direction v in which the projection of an arbitrary degree-4 moment tensor is large[5]. We therefore adopt the following relaxation: instead of searching for a direction v in which $\langle v^{\otimes 4}, T\rangle$ is small, we transform T into T_H using the Hermite polynomial basis, flatten it into a matrix M of dimension $\mathbb{R}^{d^3} \times \mathbb{R}^d$, and simply search for a subspace V in which the singular values of M are smaller than $\delta = \mathrm{poly}(\epsilon)$. To see why the relaxation is sufficient, we note that for any vector $v \in V$, $\langle v^{\otimes 4}, T_H\rangle \le \sup_{u\in\mathbb{R}^{d^3}:\|u\|_2=1}\langle u, Mv\rangle \le \delta$. On the other hand, the hermite moments of a standard Gaussian are indeed precisely 0, ensuring that we have the desired moment matching condition up to a small slackness of δ. Fortunately, even after the relaxation, we manage to prove the existence of a non-trivial subspace in which the maximum singular value of M is small. In particular, we compute the Gaussian Hermite moment tensors explicitly, and demonstrate that its nuclear norm after flattening is still $d\,\mathrm{poly}(\epsilon)$, which implies the existence of a subspace of dimension $\Omega(d)$ such that the singular values are bounded from above by $\mathrm{poly}(\epsilon)$. See proof of item (5) in Lemma 8 for the detailed argument.

Next we turn to the case where we have to compute empirical estimates of the reweighted moment tensor. Similar to the prior work, we show that one can effectively estimate the low-degree moments up to error δ in Frobenius norm with $\mathrm{poly}(d/\delta)\exp(C_{\mathrm{w}}^2)$ many samples. We remark that the exponential term $\exp(C_{\mathrm{w}}^2)$ is qualitatively tight. Indeed, for x distributed like a standard Gaussian, $\|x\|_2^2$ is a χ^2 distribution with mean d and standard deviation about $\sqrt{d}$. This means that after reweighting, most of the weight will come from points with $\|x\|_2^2 < d - \Theta(\sqrt{d}C_{\mathrm{w}})$. By

4. The astute readers may note that this only ensures that the moments are close to those of the clean Gaussian component, which is not necessarily a standard Gaussian. However, since the covariance matrix is close to the identity in Frobenius norm, it is not hard to show that there must at least exist a subspace of dimension $\Omega(d)$ such that the clean component has variance $d^{-1/2}$, which is less than $\mathrm{poly}(\epsilon)$ by our regime assumption.
5. For second moments, this boils down to computing the eigen-decomposition of the second moment matrix.

Gaussian norm concentration, the mass of such points is at most $\exp(-\Theta(\sqrt{d}/(\sqrt{d}C_{\mathrm{w}}))^2)$ and so one would need to sample at least $\exp(\Omega(C_{\mathrm{w}}^2))$ many points to see one with reasonable weight.

We remark that this indicates that the use of higher moments beyond the 4th is unfortunately out of reach. Indeed, to ensure the existence of a non-trivial moment-matching subspace up to degree 6, one would need to choose C_{w} to be at least $d^{1/6}$, which would consequently require at least $\exp(\Theta(d^{1/3}))$ many samples for accurate estimation of the moment tensor, a far cry from the desired sample complexity budget.

Proof of Proposition 4 Fortunately, moment matching up to degree 4 is already sufficient for certifying the boundedness of μ according to our structural result. We now provide its proof sketch for completeness of the argument. Given an arbitrary direction v, D is distributed as $Y+Z$ where Y is a mean 0 Gaussian (potentially with non-unit variance) and Z is an independent random variable. Our goal is to show that if $Y+Z$ has its first degree-4 moments matched with a standard Gaussian up to slackness δ, then Z cannot put too much mass on a single point far from the origin. We first provide some intuition why this should be the case. In particular, standard results from the theory of moments state that if $Y+Z$ has its low-degree moments matched with a Gaussian distribution, then the two must be close in CDF distance (see, e.g., Klebanov and Rachev (1996); Rachev et al. (2013)). If Z were to put much mass far from the origin, it would make the CDF of $Y+Z$ deviate significantly from the standard Gaussian, constituting a contradiction. This effectively certifies that the point mass of Z corresponding to the clean component must sit near the origin. However, classic results typically require a distribution to match at least $O(1/\epsilon^2)$ many moments to certify ϵ-closeness in CDF distance. As discussed previously, it is crucial for us to leverage only the information from the first four moments to construct the certificate.

To provide a tight characterization, we appeal to a more delicate argument based on LP duality. As the first step, we show that the moment matching condition implies that the variance of Y, which we denote as σ^2, cannot be too small. Assume for the sake of contradiction that $\sigma^2 < 1 - C\delta^{1/2}$ for some sufficiently large constant C. Then, by deconvolution, the moment matching condition implies that $\bar{Z} := Z/\sqrt{1-\sigma^2}$ must have its first 4 moments approximately matched with the standard Gaussian up to a slackness of C^{-1}. We argue that this certifies sufficient anti-concentration of $\bar{Z}$. Specifically, we formulate a linear program searching for the maximum amount of mass any probability measure could put on a single point t subject to the degree-4 moment constraints. Via explicit construction of solutions to the dual of this linear program, we demonstrate that the point mass can be at most $2/3 + O(C^{-1})$ for $\bar{Z}$. [6] Yet, $\bar{Z}$ clearly should have a point mass of at least $2/3 + c$ corresponding to the mean of the clean component as assumed by the noise model. This then leads to a contradiction if $c > O(C^{-1})$. This therefore allows us to conclude that σ^2 must be at least $1 - O(\delta^{1/2})$. After that, we note that the second moment of $Y+Z$ is on the one hand at most $1+\delta$ as implied by the moment matching constraints, and on the other hand at least $1.5\mu^2 + \sigma^2$. Combining this with the bound $\sigma^2 > 1 - O(\delta^{1/2})$ then shows that $\mu < O(\delta^{1/4})$. See Section 5 for the detailed argument.

6. We remark that if one is allowed to leverage information from higher moments, we could tighten the bound to be arbitrarily close to $1/2$.

1.2. Open Problems

We conclude the section with two open problems. First, we remark that the information theoretic sample complexity of the problem is $\mathrm{poly}\left(d2^{\epsilon^{-2}}\right)$, where the lower bound follows from the univariate lower bound of $2^{\Omega(\epsilon^{-2})}$ from Kotekal and Gao (2025) and the standard high-dimensional mean estimation lower bound of $\Omega(d)$, and the upper bound follows from the univariate algorithm from Kotekal and Gao (2025) plus a standard covering technique. Hence, the most obvious direction is whether one can further improve the sample complexity of our algorithm to this information-theoretic limit (while making sure the runtime is still sample-polynomial). It would also be interesting if one can demonstrate that the problem exhibits a significant information-computation tradeoff.

Second, we note that there may be room for improvement for the breakdown point of Theorem 2, which currently requires α to be smaller than some universal constant instead of $1/2$. There are mainly two bottlenecks. First, our algorithm begins by invoking the adversarially robust covariance estimator from Diakonikolas and Kane (2023), which requires the corruption rate to be at most some universal constant. To the best of our knowledge, we are not aware of any such estimators with a breakdown point close to $1/2$. Second, our mean size certification (Proposition 12) also requires an upper bound of $1/3 - c$ on the corruption rate. Although the rate can be improved to approach $1/2$ by leveraging higher moments, as discussed earlier, the use of higher moments beyond 4 is unfortunately out of reach within the current dimensionality reduction framework.

2. Preliminaries

For a vector x, we denote by $\|x\|_2$ the ℓ_2 norm of x. For an integer $k \in \mathbb{Z}_+$ we denote by $\mathbb{R}_k[x]$ the space of polynomials of degree $\leq k$ with coefficients in $\mathbb{R}$. For a matrix A, we denote by $\|A\|_*$: the nuclear norm of A. For a square matrix $\Sigma \in \mathbb{R}^{d\times d}$, we denote by $\sigma_i(\Sigma)$ the i^{th} eigenvalue of Σ, $\|\Sigma\|_2$ its operator norm and $\|\Sigma\|_F$ its Frobenius norm. For V a subspace of $\mathbb{R}^d$ with $\dim(V) = k$, we define the projection matrix $\Pi_V \in \mathbb{R}^{k\times d}$ to be any matrix whose rows form an orthonormal basis of V. For a measurable function $f : \mathbb{R} \mapsto \mathbb{R}$ and $i \in \mathbb{Z}_{\geq 1}$, we denote by $\|f\|_i$ the ℓ_i norm of f under standard Gaussian, i.e. $\|f\|_i := \left(\int_{\mathbb{R}} \left(|f(x)|^i p_{\mathcal{N}(0,1)}(x)\right) dx\right)^{1/i} = \left(\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[|f(x)|^i\right]\right)^{1/i}$. For a distribution D, we denote by p_D its probability density function. Given two vector spaces $V_1 \subseteq V_2$, we denote by $V_2 \ominus V_1$ the orthogonal complement of V_1 within V_2.

A tensor T of order k and dimensions $(d_i)_{i\in[t]}$ is a multilinear map defined by a $k-$dimensional array with real entries $T_{l_1,\cdots,l_k}$, where $l_i \in [d_i]$. We denote by $\|T\|_F := \sqrt{\sum_{l_1,\cdots,l_k} T^2_{l_1,\cdots,l_k}}$ the Frobenius norm of T. For two tensors T_1, T_2 of same order k and dimensions $(d_i)_{i\in[t]}$, we denote by $\langle T_1, T_2\rangle := \sum_{l_1,\cdots,l_k} (T_1)_{l_1,\cdots,l_k} (T_2)_{l_1,\cdots,l_k}$ their inner product. For positive integers x, y satisfying $x + y = k$, we denote by $\mathrm{flat}_{(x,y)}(T)$ the flattening of T into a matrix of dimension $d^x \times d^y$, where the vectorization operation is conducted in lexicographic order. We will also be extensively working with Hermite tensors to perform some basic computations about them.

Definition 5 (Normalized k-th Hermite Tensor) *For $k \in \mathbb{N}$ and $x \in \mathbb{R}^d$, denote by $H_k : \mathbb{R}^d \mapsto \mathbb{R}^{d^{\otimes k}}$ the k^{th} Hermite tensor of order k, where for $(i_1, i_2, \cdots, i_k) \in [d]^k$, $(H_k(x))_{i_1,i_2,\cdots,i_k} := \frac{1}{\sqrt{k!}} \sum_{\text{Partitions } P \text{ of } [k] \text{ into sets of size 1 and 2}} \prod_{\{a,b\}\in P} (-(I_d)_{i_a,i_b}) \prod_{\{c\}\in P} x_{i_c}$. For a distribution D, denote by $\mathbb{E}_{x\sim D}[H_k(x)]$ its degree-k Hermite moment tensor.*

We require the following robust estimation algorithm that can give a coarse constant-error estimate of the unknown mean and covariance for an adversarially corrupted Gaussian distribution. We defer the proof to Section A.

Lemma 6 (Proposition 4.15 in Diakonikolas and Kane (2023)) *Let $d \in \mathbb{Z}_{>0}$ and let $\alpha \in (0,1)$ denote the contamination rate. Let $\mu \in \mathbb{R}^d$ and let $\Sigma \in \mathbb{R}^{d\times d}$ be positive definite. Assume that α is at most a sufficiently small universal constant. There exists an algorithm that, given $n = \mathrm{poly}(d/\alpha)$ α-adversarially corrupted samples from $\mathcal{N}(\mu,\Sigma)$, runs in $\mathrm{poly}(n)$ time and outputs $\widehat{\mu} \in \mathbb{R}^d$ and a positive definite matrix $\widehat{\Sigma} \in \mathbb{R}^{d\times d}$ such that: (1) $\left\|\widehat{\Sigma}^{-1/2}\,\Sigma\,\widehat{\Sigma}^{-1/2} - I\right\|_2 \le 0.1$, (2) $\left\|\widehat{\Sigma}^{-1/2}\,\Sigma\,\widehat{\Sigma}^{-1/2} - I\right\|_F \le 0.1$, and (3) $\left\|\widehat{\Sigma}^{-1/2}(\widehat{\mu}-\mu)\right\|_2 \le 0.1$.*

We will also need the following estimator whose runtime scales exponentially in the sample dimension. The proof is deferred to Section A.

Lemma 7 (Inefficient Estimator) *Let $d \in \mathbb{Z}_{>0}$, $0 < \epsilon < \alpha \le 1/4$, $\delta \in (0,1)$ be parameters, and denote by $\mu \in \mathbb{R}^d$ the unknown target mean and positive definite $\Sigma \in \mathbb{R}^{d\times d}$ be the unknown covariance matrix where $\|\Sigma - I_d\|_2 = O(1)$. Then, there exists an algorithm that on input ϵ and any set of $n \ge 2^{O(1/\epsilon^2)}(d+\log(1/\delta))$ α-corrupted set of points from $\mathcal{N}(\mu,\Sigma)$ under the mean-shift contamination model (Definition 1), outputs a $\hat{\mu}$ such that $\|\hat{\mu}-\mu\| \le \epsilon$ with probability at least $1-\delta$. Moreover, it runs in time $2^{O(d)}\mathrm{poly}(n,d)$.*

3. Gaussian Mixture Reweighting

The key ingredient of our algorithm is the weight function $w : \mathbb{R}^d \mapsto \mathbb{R}_+$ defined as $\exp\left(-C_{\mathrm{w}}\frac{\|\mu_i\|_2^2}{d^{1/2}}\right)$, where C_{w} is some weight parameter that will be chosen later. In this subsection, we analyze the effect of this weight function on an α-mean-shift corrupted distribution D. As we have mentioned in Section 1.1, since the algorithm starts by normalizing the input distribution with the mean and covariance learned from the standard adversarial robust estimator (Lemma 6), we will readily assume that the clean mean of D is of size $O(1)$ and its covariance matrix is close to the identity.

In particular, our main result of this subsection shows that the reweighted distribution will be a new mixture where the mixing weight of each component with mean μ_i will be adjusted by an extra multiplicative factor proportional to roughly $\exp\left(-C_{\mathrm{w}}\frac{\|\mu_i\|_2^2}{d^{1/2}}\right)$. As a result, the noisy components with mean $> d^{1/4}/\sqrt{C_{\mathrm{w}}}$ weigh significantly less compared to before, and the weight of the target component remains roughly the same. As an important consequence of the mixing weight adjustment, the size (in a technical sense) of the Hermite moments of the new mixture will become much smaller. The formal statement is as follows, and we defer the proof to Section B.

Lemma 8 (Properties of Reweighted Mixture) *Let $C_w > 1$, $c > 0$ be a small constant, $\alpha \in (0, 1/3 - c)$, $\mu, z_1, \cdots z_n, \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Assume that $\|\mu\| < 0.1$, $\|\Sigma - I(1 - C_w/\sqrt{d})^{-1}\|_F = 0.1$, and $C_w \ll d^{1/2}$. Furthermore, define*

$$\tilde{\Sigma} := \left(\Sigma^{-1} + \frac{C_w}{d^{1/2}} I\right)^{-1}, \tilde{\mu} = \tilde{\Sigma}\Sigma^{-1}\mu\,, \tilde{z}_i = \tilde{\Sigma}\Sigma^{-1} z_i. \tag{1}$$

Let D be a Gaussian mixture defined as $D := (1-\alpha)\mathcal{N}(\mu,\Sigma) + \frac{\alpha}{n}\sum_{i=1}^n \mathcal{N}(z_i,\Sigma)$, $w(x)$ be a weight function of the form $w(x) = \exp\left(-\frac{C_w\|x\|^2}{2d^{1/2}}\right)$. Then there exist another Gaussian mixture

$D_w = \beta^* \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}) + \sum_{i=1}^n \beta_i \mathcal{N}(\tilde{z}_i, \tilde{\Sigma})$ for some mixing coefficients $\beta^*, \beta_1, \cdots, \beta_n$ such that the following holds: (1) $\tilde{\Sigma}$ is positive definite and $\|\tilde{\Sigma}\|_2 < 2$, (2) $p_D(x)w(x) \propto p_{D_w}(x)$, (3) $\beta^* > 2/3 + c/2$, (4) $\forall v \in \mathbb{S}^{d-1}$, $|v^\top \tilde{\mu} - v^\top \mu| = O(C_w/d^{1/2})$, and (5) for all positive integers $k \le 4$, the nuclear norm of the $(k-1,1)$-flattening of degree-k Hermite moment tensor of D_w satisfies $\left\|\mathrm{flat}_{(k-1,1)}(\mathbb{E}_{x\sim D_w}[H_k(x)])\right\|_* \le O(d/C_w)$.

4. Subspace Identification

Recall that in Lemma 8 we have shown that the low-degree Hermite moments of the reweighted mixture will be small in a technical sense. In particular, we show that its nuclear norm will be of the order of $O(d/C_{\mathrm{w}})$, where C_{w} is the weight parameter, after flattening it into a linear transformation taking vectors from $\mathbb{R}^d$ as input. Thus, by a simple counting argument, there must exist a subspace of dimension at least $\Omega(d)$ such that the projected Hermite moments of the reweighted mixture is at most $O(1/C_{\mathrm{w}})$. In this subsection, we present an algorithm that can reliably identify the subspace with samples from the unweighted mixture D.

Lemma 9 (Subspace Identification) *Let $C_w > 1, \alpha, \epsilon, \tau, \in (0, 1/2)$, $\mu, z_1, \cdots z_n \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Assume that $\|\mu\|^2 < 0.1$, $\|\Sigma - (1 - C_w/\sqrt{d})^{-1} I\|_F \le 0.1$, $\|\Sigma\| = O(1)$, and $C_w \ll d^{1/2}$. Let D be a Gaussian mixture defined as $D := (1-\alpha)\mathcal{N}(\mu, \Sigma) + \frac{\alpha}{n}\sum_{i=1}^n \mathcal{N}(z_i, \Sigma)$, and $w(x)$ be a weight function of the form $w(x) = \exp\left(-\frac{C_w\|x\|^2}{d^{1/2}}\right)$. Define $C_N = \mathbb{E}_{x\sim D}[w(x)]$. Then Algorithm 1 given $m = d^4\epsilon^{-2}\exp(\Theta\left(C_w^2\right))\tau^{-1}$ i.i.d. samples from D, runs in time $\mathrm{poly}(m, d)$, and with probability at least $1 - \tau$ outputs a subspace V of dimension at least $d/2$ satisfying $\frac{1}{C_N}\left|\mathbb{E}_{x\sim D}\left[h_j(\langle v, x\rangle)w(x)\right]\right| = O(C_w^{-1})$. for all $j \in [4]$ and unit vector $v \in V$.*

Towards showing Lemma 9, we will first show that the reweighted Hermite moments have good concentration properties. The formal statement is as follows, and we defer the proof to Section C.

Lemma 10 (Concentration) *Let $C_w > 1, \alpha \in (0, 1/2)$, $\mu, z_1, \cdots z_n \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Assume that $\|\mu\| = O(1)$, $\|\Sigma\|_2 = O(1)$, and $C_w \ll d^{1/2}$. Let D be a Gaussian mixture defined as $D := (1-\alpha)\mathcal{N}(\mu, \Sigma) + \frac{\alpha}{n}\sum_{i=1}^n \mathcal{N}(z_i, \Sigma)$, and $w(x)$ be a weight function of the form $w(x) = \exp\left(-\frac{C_w\|x\|^2}{d^{1/2}}\right)$. Define $C_N = \mathbb{E}_{x\sim D}[w(x)]$. Let $\{x_i\}_{i=1}^m$ be $m = \exp(CC_w^2)\cdot d^4/(\delta\epsilon^2)$ i.i.d. samples drawn from D, where $C > 0$ is a sufficiently large constant. Then we have that with probability at least $1 - \delta$, for $j = 0, 1, \cdots, 4$, it holds $\left\|\frac{1}{m}\sum_{i=1}^m \frac{H_j(x_i)w(x_i)}{C_N} - \mathbb{E}_{x\sim D}\left[\frac{H_j(x)w(x)}{C_N}\right]\right\|_F \le \epsilon$.*

Note that the actual empirical moment tensor $\widehat{T}_i$ computed in line 7 of Algorithm 1 uses the normalization factor $\bar{C}_N$, which is an empirical estimation of the true normalization factor $C_N = \mathbb{E}[w(x)]$. However, since Lemma 10 (applied with $j = 0$) guarantees that $\bar{C}_N$ will be a good approximation to C_N up to small multiplicative factor, as a simple corollary, we obtain the following concentration bounds for the actual empirical moment tensor $\widehat{T}_i$, and we defer the proof to Section C.

Corollary 11 *Let $\alpha, \epsilon, \tau, \in (0, 1/2)$, $\mu, z_1, \cdots z_n, \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Assume that $\|\mu\| = O(1)$, $\|\Sigma\|_2 = O(1)$, and $\epsilon^{-12} \ll d^{1/2}$. Let D be a Gaussian mixture defined as $D := (1-\alpha)\mathcal{N}(\mu, \Sigma) + \frac{\alpha}{n}\sum_{i=1}^n \mathcal{N}(z_i, \Sigma)$, $w : \mathbb{R}^d \mapsto \mathbb{R}_+$ be the weight function defined as in line 4 and $C_N := \mathbb{E}_{x\sim D}[w(x)]$. Then the empirical moment tensor $\widehat{T}_i$ computed in line 7 satisfies $\left\|\widehat{T}_i - C_N^{-1}\mathbb{E}_{x\sim D}\left[H_i(x)w(x)\right]\right\|_F \le \epsilon$ with probability at least $1 - \tau$.*

The desired algorithm is Algorithm 1, and the full analysis (proof of Lemma 9) is in Section C.

Algorithm 1 Subspace Identification

Input: dimension d, constant C_{w}, set S of m i.i.d. samples from a distribution over $\mathbb{R}^d$.
Output: A subspace V.
Let $C > 0$ be a sufficiently large universal constant.
Define the weight function $w(z) = \exp(-C_{\mathrm{w}}\|z\|^2/d^{1/2})$.
Compute the empirical normalization factor $\bar{C}_N := \frac{1}{m}\sum_{x\in S} w(x)$.
For $i \in [4]$, compute the degree-i empirical weighted Hermite tensors using S: $\widehat{T}_i = \frac{1}{m}\sum_{x\in S} w(x)\, H_i(x)\bar{C}_N^{-1}$.
Flatten every $\widehat{T}_i$ into a matrix $\widehat{M}_i \in \mathbb{R}^{d^{i-1}\times d}$.
Compute the right singular vectors $v_1^{(i)}, \ldots, v_d^{(i)}$ and let $\sigma_1^{(i)}, \ldots, \sigma_d^{(i)}$ be their singular values.
Let $\mathcal{I} = \mathrm{Span}\left\{v_j^{(i)} : \exists i \in [4], j \in [d] \text{ s.t. } \sigma_j^{(i)} > CC_{\mathrm{w}}^{-1}\right\}$.
return the orthogonal complement of $\mathcal{I}$.

5. Mean Deviation and Hermite Moment Matching

In this section, we prove the main one-dimensional structural result described informally in Proposition 4. We view a one-dimensional mixture of shifted $\mathcal{N}(0,\sigma^2)$ Gaussians as a convolution $X := Z + \sigma Y$, where $Y \sim \mathcal{N}(0,1)$ and Z encodes the random mean displacement. Since the target component carries more weight than the sum of all noisy components, Z has substantial probability mass at the target mean μ (i.e., $\Pr[Z = \mu]$ is large). We show that if X approximately matches the first four Hermite moments of $\mathcal{N}(0,1)$, then this heavy point of Z must be near the origin—in particular, μ is small.

Proposition 12 (Moment Matching implies no heavy point) *Let $\sigma > 0$, $c > 0$ be a small constant, Y be standard Gaussian random variable and Z be another random variable over $\mathbb{R}$ independent of Y. Define the random variable X as the convolution $X := Z + \sigma Y$. If $\Pr[Z = \mu] > 2/3 + c/2$, and for $i = 1, 2, 3, 4$, $|\mathbb{E}[h_i(X)]| \le \epsilon$, then $\mu = O(\epsilon^{1/4})$.*

Intuitively, classic results from the theory of moments imply that X and the standard Gaussian must be close in CDF distance. If Z were to have a point mass far away from the origin where the standard Gaussian has relatively low density, that would result in X having significantly more mass within the nearby interval while compared to the standard Gaussian. We can then conclude that the heavy point mass of Z, which corresponds exactly to μ in Proposition 12, must be small.

Alas, these classic results are only applicable when we match $\Omega(1)$ many moments. A more delicate argument is hence needed to establish Proposition 12, which crucially only assumes moment matching up to degree 4. In what follows, we provide a more detailed proof sketch, deferring the formal arguments to Section D.

We first use some basic variance related equalities to show that σ^2 cannot significantly exceed 1. Next, we proceed to show that with the approximate Hermite-moment-matching constraints, σ^2 cannot be much smaller than 1 either. Assume for the sake of contradiction that σ^2 is much less than 1. We can then explicitly write $X = \sigma Y + \sqrt{1-\sigma^2}\frac{Z}{\sqrt{1-\sigma^2}}$, which views X as the result of applying

the Gaussian noise operator U_σ on $\frac{Z}{\sqrt{1-\sigma^2}}$. Then it is not hard to show via a deconvolution argument that the Hermite moments of $\frac{Z}{\sqrt{1-\sigma^2}}$ is as small as the Hermite moments of X up to polynomial inflation in $\left(1-\sigma^2\right)^{-1}$. The formal statement is as follows, and the proof is deferred to Section D.

Lemma 13 *Let $\sigma \in (0,1)$, Z and X be real-valued random variables and Y be standard Gaussian random variable, where X is defined as the convolution $X = \sigma Y + Z$, and let $\bar{Z} = \frac{Z}{\sqrt{1-\sigma^2}}$. If for $i = 1, \cdots, k$, $|\mathbb{E}[h_i(X)]| = \Theta(\epsilon)$, then for $i = 1, \cdots, k$, $\left|\mathbb{E}\left[h_i\left(\bar{Z}\right)\right]\right| = \Theta\left(\frac{\epsilon}{(1-\sigma^2)^{i/2}}\right)$.*

From Lemma 13 and the assumption that σ^2 is much less than 1, we know that $\bar{Z} = Z/\sqrt{1-\sigma^2}$ must also have its moments approximately matched with $\mathcal{N}(0,1)$. This ends up putting significant constraints on how much mass $\bar{Z}$ can put on any single point. In particular, we show the following anti-concentration property.

Lemma 14 *For $\delta \in \mathbb{R}_{\geq 0}$, let $\bar{Z}$ be a real-valued random variable, and $k \geq 2$ be an even integer. If for $i = 1, \cdots, k$, $\left|\mathbb{E}[h_i(\bar{Z})]\right| \leq \delta$, where h_i is the degree-i Hermite polynomial, then $\forall t \in \mathbb{R}$, $\Pr[\bar{Z} = t] < \frac{1+\delta\cdot\sqrt{k}\cdot 3^{k/2}}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)}$.*

To show Lemma 14, we formulate the problem of maximizing the point mass at a single point with approximate Hermite-moment-matching constraints as a linear program in the measure space, and construct an explicit feasible solution for its dual using the Hermite polynomials. The actual calculations are rather technical and thus deferred to Section D.

By invoking Lemma 14 with appropriate parameters on $\bar{Z}$, we demonstrate that $\bar{Z}$ can put at most $2/3 + c/2$ mass at any single point for a small constant c. On the other hand, we know that $\Pr\left[\bar{Z} = \mu/\sqrt{1-\sigma^2}\right] = \Pr[Z = \mu] > 2/3 + c/2$ by the assumption of Proposition 12, whence giving the desired contradiction.

At this point, we have shown that the variance σ^2 must live within a small interval around 1. This essentially reduces the problem to the unit-variance case and so we can employ an almost identical argument as the one used in Diakonikolas et al. (2025a) to conclude the proof of Proposition 12.

6. Proof of Main Theorem

We first argue that the property is satisfied by the whitened distribution $\bar{D}$.

Lemma 15 (Coarse Normalization) *Let $d \in \mathbb{Z}_+$ and C_w the parameter defined in Algorithm 2. Assume that $C_w/\sqrt{d} < 0.1$. With probability at least 0.99, it holds that the whitened distribution $\bar{D}$ (line 7 of Algorithm 2) is an α-mean-shift corrupted distribution with clean mean $\bar{\mu}$ and positive definite shared covariance $\bar{\Sigma}$ satisfying that $\|\bar{\mu}\|_2 = O(1)$, $\|\bar{\Sigma} - (1 - C_w/\sqrt{d})^{-1} I\|_F = O(1)$.*

We defer the proof to Section E. We next argue that after each iteration of the main loop of Algorithm 2, the estimation difference $\widehat{\Sigma}_0^{-1/2}(\mu - \widehat{\mu}_0)$ projected onto the subspace removed, i.e, $V_t \ominus V_{t+1}$, must have small ℓ_2 norm.

Lemma 16 (Effective Dimensionality Reduction) *Let $d \in \mathbb{Z}_+$. Let C_w be the parameter defined at line 4 of Algorithm 2, and V_t, $k := \mathsf{dim}(V_t)$, be the subspace of $\mathbb{R}^d$ identified at iteration t (line d). Assume that $C_w \ll \sqrt{d}$ and the whitened distribution $\bar{D}$ (line 7) is an α-mean-shift corrupted distribution with clean mean $\bar{\mu}$ and covariance $\bar{\Sigma}$ satisfying that $\|\bar{\mu}\|_2 =$*

Algorithm 2 Robust Mean Estimation via Reweighted Moments

1: **Input**: dimension d, contamination rate $\alpha < 1/4$, accuracy $\epsilon \in (0,1)$, sample access to an α-mean-shift corrupted distribution D (see Definition 1).
2: **Output**: estimator $\widehat{\mu} \in \mathbb{R}^d$ with $\|\widehat{\mu} - \mu\| \leq \epsilon\sqrt{\|\Sigma\|}$ with probability at least $2/3$.
3: Let C be a sufficiently large absolute constant.
4: Define the weight parameter $C_{\mathrm{w}} := C\left(\frac{\log d}{\epsilon}\right)^4$.
5: **/*Coarse normalization*/**
6: Draw $n_0 = (d/\alpha)^C$ samples from D and apply the Robust Gaussian Estimator of Lemma 6 to obtain $\widehat{\mu}_0 \in \mathbb{R}^d$ and $\widehat{\Sigma}_0 \in \mathbb{R}^{d\times d}$ satisfying $\|\widehat{\Sigma}_0^{-1/2}(\mu - \widehat{\mu}_0)\| = O(1)$ and $\|\widehat{\Sigma}_0^{-1/2}\Sigma\widehat{\Sigma}_0^{-1/2} - I\|_F = O(1)$.
7: Define the whitened distribution $\bar{D}$ whose sample is given by $\left(1 - C_{\mathrm{w}}/\sqrt{d}\right)^{-1/2}\widehat{\Sigma}_0^{-1/2}(x - \widehat{\mu_0})$, where $x \sim D$.
8: **/*Dimension–reduction loop*/**
9: Initialize $t \leftarrow 1$, $k \leftarrow d$, and $V_t \leftarrow \mathbb{R}^d$. **while** $k > C\,C_w^2$ **do**
   a Define D_t to be the distribution of $\Pi_{V_t}x \in \mathbb{R}^k$ where $x \sim \bar{D}$.
   b Draw a sample set S of $n_1 = 2^{CC_{\mathrm{w}}^2}/\epsilon^2$ i.i.d. points from D_t.
   c $W \leftarrow$ Algorithm 1(k, C_{w}, S), where $W \subseteq \mathbb{R}^k$.
   d Update $V_{t+1} \leftarrow \Pi_{V_t}^\top W^\perp$, $k \leftarrow \mathsf{dim}(V_{t+1})$, $t \leftarrow t+1$.
   end
10: **/*Low–dimensional exhaustive search*/**
11: Run exhaustive search of Lemma 7 using $n_2 = \log^4(d)2^{C/\epsilon^2}$ samples to obtain $\widehat{\mu}_1 \in \mathbb{R}^k$.
12: **return** $\left(1 - C_{\mathrm{w}}/\sqrt{d}\right)^{1/2}\widehat{\Sigma}_0^{1/2}\Pi_{V_t}^\top\widehat{\mu}_1 + \widehat{\mu}_0$.

$O(1)$, $\|\bar{\Sigma} - (1 - C_w/\sqrt{d})^{-1}I\|_F = O(1)$. *Then with probability at least* $1 - 1/(10\log(d))$ *it holds that* $\|\,\mathrm{Proj}_{V_t\ominus V_{t+1}}(\bar{\mu})\,\| = O(C_w^{-1/4})$.

Proof Let $w : \mathbb{R}^k \mapsto \mathbb{R}_+$ be the weight function defined as in line 4 of Algorithm 1. Let $\Pi_{V_t} \in \mathbb{R}^{k\times d}$ denote the (orthogonal) projection matrix onto the subspace V_t. Let $W \subseteq \mathbb{R}^k$ be the subspace returned by Algorithm 1. By line d, $V_{t+1} = \Pi_{V_t}^\top W^\perp$ and hence $V_t \ominus V_{t+1} = \Pi_{V_t}^\top W$. Denote by D_t the projection of $\bar{D}$ onto V_t (line a), i.e. the distribution of $\Pi_{V_t}x$ where $x \sim \bar{D}$. Since $\bar{D}$ is an α-mean-shift corrupted distribution with clean component $\mathcal{N}(\bar{\mu}, \bar{\Sigma})$, D_t must also be an α-mean-shift corrupted distribution with clean component $\mathcal{N}(\mu_t, \Sigma_t)$, where $\mu_t = \Pi_{V_t}\bar{\mu}$ and $\Sigma_t = \Pi_{V_t}\bar{\Sigma}\Pi_{V_t}^\top$. Moreover, since Π_{V_t} is a projection matrix, it holds that $\|\mu_t\|_2 = \|\Pi_{V_t}\bar{\mu}\|_2 \leq \|\bar{\mu}\|_2 = O(1)$. Denote by $c := (1 - C_{\mathrm{w}}/\sqrt{d})^{-1}$. We then have that $\|\Sigma_t - cI_k\|_F = \left\|\Pi_{V_t}\left(\bar{\Sigma} - cI_d\right)\Pi_{V_t}^\top\right\|_F \leq \|\Pi_{V_t}\|_2\,\|\bar{\Sigma} - cI_d\|_F\,\|\Pi_{V_t}^\top\|_2$
$= \|\bar{\Sigma} - cI_d\|_F = O(1)$. Therefore, since D_t satisfies the assumptions of Lemma 9 and the sample size given is at least $d^4C_{\mathrm{w}}^2\exp(CC_{\mathrm{w}}^2)$ (by line b), with probability at least $1 - 1/(10\log(d))$ for all $j \in [4]$ and every unit vector $u \in W$ we have

$$\left|C_N^{-1}\mathbb{E}_{x\sim D_t}\left[h_j(\langle v, x\rangle)w(x)\right]\right| = O(C_{\mathrm{w}}^{-1})\,, \tag{2}$$

for $C_N = \mathbb{E}_{x\sim D_t}[w(x)]$. Denote by D_t^w the distribution proportional to D_t weighted by w. By Lemma 8, D_t^w is exactly proportional to another $\tilde{\alpha}$-mean-shift corrupted distribution with clean mean $\tilde{\mu}$ and covariance $\tilde{\Sigma}$ satisfying that

$$\tilde{\alpha} \geq 3/4\,, \tag{3}$$

$$\|\tilde{\mu} - \mu_t\|_2 = O(C_{\mathrm{w}}/\sqrt{d}). \tag{4}$$

In particular, Equation (3) implies that D_t^w projected onto an arbitrary unit direction $u \in W$ must be a univariate distribution X that can be written as a convolution of the form $X = Y + Z$, where Y is a univariate Gaussian random variable and Z is some random variable satisfying $\Pr[Z = \langle u, \tilde{\mu}\rangle] \geq 3/4$. Moreover, Equation (2) implies that $\mathbb{E}[h_i(X)] = O(C_{\mathrm{w}}^{-1})$ for $i \in [4]$. Hence, Proposition 12 is applicable and yields $|\langle u, \tilde{\mu}\rangle| = O(C_{\mathrm{w}}^{-1/4})$. Combining this with Equation (4) gives $|\langle u, \mu_t\rangle| = O(C_{\mathrm{w}}^{-1/4})$ for every unit vector $u \in W$. Therefore,

$$\|\operatorname{Proj}_{V_t \ominus V_{t+1}}(\bar{\mu})\|_2 = \|\operatorname{Proj}_W(\mu_t)\|_2 = O(C_{\mathrm{w}}^{-1/4}).$$

This concludes the proof of Lemma 16. ■

We are now ready to conclude the proof of Theorem 2.

Proof By Lemma 15, with probability at least 0.99, the whitened distribution $\bar{D}$ (line 7) is an α-mean-shift corrupted distribution with clean mean $\bar{\mu}$ and shared covariance $\bar{\Sigma}$ satisfying that $\|\bar{\mu}\|_2 = O(1)$, $\|\bar{\Sigma} - (1 - C_{\mathrm{w}}/\sqrt{d})^{-1} I\|_F = O(1)$. Condition on this event, we thus have that $\bar{D}$ satisfies the assumption of Lemma 16. Therefore, for any iteration t, it holds that the Subspace Identification algorithm of Lemma 9 succeeds in identifying a subspace V_{t+1} satisfying that $\operatorname{Proj}_{V_t \ominus V_{t+1}}(\bar{\mu}) \leq O(C_{\mathrm{w}}^{-1/4})$. with probability at least $1 - 1/(10 \log d)$. Since $\mathsf{dim}(W) \geq k/2$, line d gives $\mathsf{dim}(V_{t+1}) = \mathsf{dim}(W^{\perp}) \leq k/2$. Hence, with high constant probability, the while loop will run for at most $\log(d)$ iterations. Denote by T the last iteration when the algorithm exits the main while loop. With high constant probability, we must have that

$$\operatorname{Proj}_{V_t \ominus V_{t+1}}(\bar{\mu}) = O(C_{\mathrm{w}}^{-1/4}). \tag{5}$$

for all $t \in [T]$. We will condition on the above events in the rest of the analysis. By Lemma 7, the final mean $\widehat{\mu}_1$ obtained with probability 0.99 must satisfy

$$\|\widehat{\mu}_1 - \Pi_{V_{T+1}}\bar{\mu}\|_2 \leq \epsilon/2. \tag{6}$$

In particular, Equations (5) and (6) together implies that

$$\begin{aligned}
\|\Pi_{V_{T+1}}^{\top}\widehat{\mu}_1 - \bar{\mu}\|_2^2 &= \sum_{t=1}^{T} \|\operatorname{Proj}_{V_t \ominus V_{t+1}} \bar{\mu}\|_2^2 + \|\widehat{\mu}_1 - \Pi_{V_{T+1}}\bar{\mu}\|_2^2 \\
&\leq O(\log d\, C_{\mathrm{w}}^{-1/2}) + \epsilon^2/2 \\
&\leq \epsilon^2.
\end{aligned}$$

Recall that Lemma 15 shows that: $\bar{\mu} = \left(1 - C_{\mathrm{w}}/\sqrt{d}\right)^{-1/2} \widehat{\Sigma}_0^{-1/2}(\mu - \widehat{\mu}_0)$, which implies

$$\left\| \left(1 - \frac{C_{\mathrm{w}}}{\sqrt{d}}\right)^{1/2} \widehat{\Sigma}_0^{1/2} \Pi_{V_{T+1}}^{\top}\widehat{\mu}_1 + \widehat{\mu}_0 - \mu \right\|_2 \leq \left(1 - \frac{C_{\mathrm{w}}}{\sqrt{d}}\right)^{1/2} \|\widehat{\Sigma}_0^{1/2}\|_2 \|\Pi_{V_{T+1}}^{\top}\widehat{\mu}_1 - \bar{\mu}\|_2 \leq O(\|\widehat{\Sigma}_0^{1/2}\|\epsilon).$$

where in the last inequality we have again used our regime assumption that $C_{\mathrm{w}} \ll \sqrt{d}$. Note that $\|\widehat{\Sigma}_0\|_2 = O(\|\Sigma\|_2)$, which is implied by the guarantee of robust covariance estimation (Lemma 6) stating that $\|\widehat{\Sigma}_0^{-1/2}\widehat{\Sigma}\widehat{\Sigma}_0^{-1/2} - I\|_F \leq 0.1$. Moreover using the union bound we have that the probability of success is at least $2/3$.

Lastly, we analyze the sample complexity and the runtime of the algorithm. The Robust Gaussian Estimator of Lemma 6 consumes at most $n_0 = \mathrm{poly}(d/\alpha)$ many samples and runs in $\mathrm{poly}(dn_0)$ time. The Subspace Identification algorithm consumes at most $n_1 = 2^{O(C_{\mathrm{w}}^2)} = 2$, for $C_{\mathrm{w}} = O\left(\log(d)/\epsilon\right)^4$, and runs in $\mathrm{poly}(dn_1)$ time. Moreover, we invoke it for at most $\log d$ many times. The exhaustive search algorithm is executed on a distribution of dimension $O(C_{\mathrm{w}}^2)$. Therefore, it consumes at most $n_2 = O(C_{\mathrm{w}}^2)2^{O(\epsilon^{-2})}$ samples and runs in time $2^{O(\log d/\epsilon)^8}$. Hence, both the sample complexity and the run-time complexity of the algorithm is given by $2^{C_{\mathrm{w}}^2+1/\epsilon^2}$. This concludes the proof of Theorem 2. ■

Acknowledgments

Ilias Diakonikolas acknowledges support under NSF Medium Award CCF-2107079, ONR award number N00014-25-1-2268, and an H.I. Romnes Faculty Fellowship. Giannis Iakovidis acknowledges support under ONR award number N00014-25-1-2268. Daniel M. Kane acknowledges support under NSF Medium Award CCF-2107547. Thanasis Pittas acknowledges support under NSF Medium Award CCF-2107079.

References

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA, 1997.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Aline Bonami. Étude des coefficients de fourier des fonctions de $l^p(g)$. In *Annales de l'institut Fourier*, volume 20, pages 335–402, 1970.

T Tony Cai and Jiashun Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145, 2010.

Alexandra Carpentier, Sylvain Delattre, Etienne Roquain, and Nicolas Verzelen. Estimating minimum effect with outlier selection. *The Annals of Statistics*, 49(1):272–294, 2021.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 47–60, 2017. doi: 10.1145/3055399.3055491.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2755–2771, 2019. doi: 10.1137/1.9781611975482.171.

Olivier Collier and Arnak Dalalyan. Multidimensional linear functional estimation in sparse gaussian models and robust estimation of the mean. *Electronic Journal of Statistics*, 13:2830–2864, 2019.

Abhimanyu Das, Ayush Jain, Weihao Kong, and Rajat Sen. Efficient list-decodable regression using batches. In *International Conference on Machine Learning*, pages 7025–7065. PMLR, 2023.

Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664. IEEE, 2016.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 999–1008, 2017. URL https://arxiv.org/abs/1703.00893.

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. In *IEEE Symposium on Foundations of Computer Science (FOCS) 2020*, 2020a.

Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1830–1841, 2020b.

Ilias Diakonikolas, Daniel M Kane, Ankit Pensia, and Thanasis Pittas. Streaming algorithms for high-dimensional robust statistics. In *International Conference on Machine Learning*, pages 5061–5117. PMLR, 2022.

Ilias Diakonikolas, Giannis Iakovidis, Daniel Kane, and Thanasis Pittas. Efficient multivariate robust mean estimation under mean-shift contamination. In *Forty-second International Conference on Machine Learning*, 2025a.

Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Sihan Liu, and Thanasis Pittas. Batch list-decodable linear regression via higher moments. In *Forty-second International Conference on Machine Learning*, 2025b.

Ilias Diakonikolas, Giannis Iakovidis, Daniel M Kane, and Sihan Liu. Sample complexity bounds for robust mean estimation with mean-shift contamination. *To appear in the Forty-third International Conference on Machine Learning*, 2026.

Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

Bradley Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, 2007.

Bradley Efron. Microarrays, empirical bayes and the two-groups model. 2008.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Samuel B. Hopkins and Jerry Li. How hard is robust mean estimation? In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, pages 1649–1682, 2019.

Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

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

Daniel M Kane, Ilias Diakonikolas, Hanshen Xiao, and Sihan Liu. Online robust mean estimation. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3197–3235. SIAM, 2024.

L.B. Klebanov and S.T. Rachev. Proximity of probability measures with common marginals in a finite number of directions. *Lecture Notes-Monograph Series*, pages 162–174, 1996.

Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 1420–1430, 2018.

Subhodh Kotekal and Chao Gao. Optimal estimation of the null distribution in large-scale inference. *IEEE Transactions on Information Theory*, 2025.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 1302–1382, 2017. URL https://proceedings.mlr.press/v65/li17a.html.

Shuchen Li. Robust mean estimation against oblivious adversaries. Master's thesis, Carnegie Mellon University, Pittsburgh, PA, 2023.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook. *Technical University of Denmark*, 2012.

Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

Svetlozar T Rachev, Lev B Klebanov, Stoyan V Stoyanov, and Frank Fabozzi. *The methods of distances in the theory of probability and statistics*, volume 10. Springer, 2013.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS)*, 2018. doi: 10.4230/LIPIcs.ITCS.2018.45.

John Wilder Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

R Tyrrell Rockafellar. Convex analysis. *Princeton mathematical series*, 28, 1970.

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

Appendix

The appendix is structured as follows: Section A includes additional preliminaries required in subsequent technical sections. Section B contains the proof of Lemma 8, establishing key properties of the reweighted samples. Section C proves the guarantee for the subspace identification algorithm (Algorithm 1). Section D contains the full proof of our moment matching result for a gaussian mixture with a very heavy component (Proposition 12). Finally, Section E contains the proof omitted from Section 6.

Appendix A. Omitted Facts and Preliminaries

Algebra While Definition 5 provides an entry-wise definition, $H_k(x)$ has a coordinate-free equivalent representation that sums over permutations to capture the underlying partitions as follows

$$H_k(x) = \frac{1}{\sqrt{k!}} \sum_{i=0}^{\lfloor k/2 \rfloor} c_{k,i} \operatorname{Sym}\left(x^{\otimes(k-2i)} \otimes I_d^{\otimes i}\right),$$

where the operator $\operatorname{Sym}(T)_{i_1,\cdots i_k} := \frac{1}{k!}\sum_{\text{Permutations } \sigma \text{ of } [k]} T_{i_{\sigma(1)},\cdots,i_{\sigma(k)}}$ maps any order k tensor T to a symmetric tensor with same order, and $c_{k,i} := \frac{k!}{(k-2i)!i!2^i}$ is the constant scalar that ensures each unique partitions is only counted once.

We will also use the following fact on entries of Hermite moment tensor under the Gaussian distribution.

Fact 17 (Lemma 2.7 in Kane (2021)) *For $x \sim \mathcal{N}(\mu, I+\Sigma)$, $k \in \mathbb{N}$, and $i_1, \cdots, i_k \in [d]^k$*

$$(\mathbb{E}[H_k(x)])_{i_1,\cdots,i_k} = \frac{1}{\sqrt{k!}} \sum_{\substack{\text{Partitions } P \text{ of } [k] \text{ into} \\ \text{sets of size 1 and 2}}} \prod_{\{a,b\}\in P} \Sigma_{i_a,i_b} \prod_{\{c\}\in P} \mu_{i_c}.$$

Equivalently, the coordinate-free representation is as follows

$$\mathbb{E}[H_k(x)] = \frac{1}{\sqrt{k!}} \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{k!}{(k-2i)!i!} Sym\left(\mu^{\otimes(k-2i)} \otimes \Sigma^{\otimes i}\right).$$

Probability Theory and Robust Statistics We require the following facts in our proof.

Definition 18 (Gaussian noise (Ornstein-Uhlenbeck) operator) *For $\rho \in [0,1]$, the Gaussian noise operator U_ρ maps a univariate distribution $P : \mathbb{R} \to [0,1]$ to the distribution of random variable $\rho X + \sqrt{1-\rho^2} \cdot Y$ where $X \sim P$ and $Y \sim \mathcal{N}(0,1)$ are independent.*

Fact 19 (Gaussian Noise Operator Preserves Hermite Moments) *For $\rho \in (0,1)$ and univariate distribution $P : \mathbb{R} \to [0,1]$, it holds that $E_{x\sim U_\rho(P)}[h_i(x)] = \rho^i \cdot \mathbb{E}_{x\sim P}[h_i(x)]$.*

Fact 20 (Gaussian Hypercontractivity inequality Bonami (1970); Gross (1975)) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a degree $\le k$ polynomial and $q > 2$, then $\|p\|_q \le (q-1)^{k/2}\|p\|_2$. Specifically, when $q = 4$ we have that $\|p\|_4 \le 3^{k/2}\|p\|_2$.*

Lemma 6 (Proposition 4.15 in Diakonikolas and Kane (2023)) *Let $d \in \mathbb{Z}_{>0}$ and let $\alpha \in (0,1)$ denote the contamination rate. Let $\mu \in \mathbb{R}^d$ and let $\Sigma \in \mathbb{R}^{d\times d}$ be positive definite. Assume that α is at most a sufficiently small universal constant. There exists an algorithm that, given $n = \mathrm{poly}(d/\alpha)$ α-adversarially corrupted samples from $\mathcal{N}(\mu,\Sigma)$, runs in $\mathrm{poly}(n)$ time and outputs $\widehat{\mu} \in \mathbb{R}^d$ and a positive definite matrix $\widehat{\Sigma} \in \mathbb{R}^{d\times d}$ such that: (1) $\left\|\widehat{\Sigma}^{-1/2}\,\Sigma\,\widehat{\Sigma}^{-1/2} - I\right\|_2 \le 0.1$, (2) $\left\|\widehat{\Sigma}^{-1/2}\,\Sigma\,\widehat{\Sigma}^{-1/2} - I\right\|_F \le 0.1$, and (3) $\left\|\widehat{\Sigma}^{-1/2}(\widehat{\mu}-\mu)\right\|_2 \le 0.1$.*

Proof We first show property 1. $\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I\|_2 \le \|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I\|_F = O(\alpha\log(1/\alpha))$. Equivalently, we have that $\forall x \in \mathbb{S}^{d-1}, x(\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I)x^T = O(\alpha\log(1/\alpha))$, then since $\widehat{\Sigma}^{-1/2}$ and $\Sigma^{1/2}$ nonsingular whence injective, suffices to substitute x with $x'\widehat{\Sigma}^{-1/2}\Sigma^{1/2}$. Indeed, $x'(\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I)(x')^T = O(\alpha\log(1/\alpha)) \le 0.1, \forall x' \in \mathbb{S}^{d-1}$, for carefully chosen constant in n as desired.

We next show property 2.

$$\begin{aligned}\|\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I\|_F &= \|(\widehat{\Sigma}^{-1/2}\Sigma^{1/2})(\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I)(\widehat{\Sigma}^{-1/2}\Sigma^{1/2})^T\|_F \\ &\le \|(\widehat{\Sigma}^{-1/2}\Sigma^{1/2})\|_2\|(\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I)\|_F\|(\widehat{\Sigma}^{-1/2}\Sigma^{1/2})^T\|_2 \\ &\le O(1)\|(\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I)\|_F = O(\alpha\log(1/\alpha)) \le 0.1,\end{aligned}$$

where the second inequality follows from singular value decomposition, Cauchy-Shwarz inequality, and the fact that Frobenius norm is unitarily invariant, and the last inequality follows from Lemma 6 and the property of operator norm that $\|A^TA\|_2 = \|AA^T\|_2 = \|A\|_2^2$, as desired.

Property 3 was shown in Diakonikolas and Kane (2023), which concludes the proof. ■

We need the following facts to prove Lemma 7.

Fact 21 (see, e.g., Corollary 4.2.13 in Vershynin (2018)) *Let $\xi > 0$. There exists a set $\mathcal{C}$ of unit vectors of $\mathbb{R}^d$ such that $|\mathcal{C}| < (1+2/\xi)^d$ and for every $u \in \mathbb{R}^d$ with $\|u\| = 1$ it holds $\min_{y\in\mathcal{C}}\|y - u\| \le \xi$.*

Fact 22 (One-dimensional estimator, see, e.g., Kotekal and Gao (2025)) *There exists a sufficiently large constant $C > 0$ such that the following hold. Let $\mu \in \mathbb{R}, \sigma > 0$ be the (unknown) mean and variance and $\alpha \le 0.49$. There is an algorithm that given, $\epsilon > 0$, $\delta \in (0,1)$ and a set of $n = 2^{C(\alpha\sigma/\epsilon)^2}\log(1/\delta)/\alpha^2$ α-corrupted samples from $\mathcal{N}(\mu,\sigma)$ according to the mean-shift model (Definition 1), finds $\widehat{\mu} \in \mathbb{R}$ such that, with probability at least $1-\delta$, it holds $|\widehat{\mu} - \mu| \le \epsilon$. The runtime of the algorithm is $\mathrm{poly}(n)$.*

Lemma 7 (Inefficient Estimator) *Let $d \in \mathbb{Z}_{>0}$, $0 < \epsilon < \alpha \le 1/4$, $\delta \in (0,1)$ be parameters, and denote by $\mu \in \mathbb{R}^d$ the unknown target mean and positive definite $\Sigma \in \mathbb{R}^{d\times d}$ be the unknown covariance matrix where $\|\Sigma - I_d\|_2 = O(1)$. Then, there exists an algorithm that on input ϵ and any set of $n \ge 2^{O(1/\epsilon^2)}(d+\log(1/\delta))$ α-corrupted set of points from $\mathcal{N}(\mu,\Sigma)$ under the mean-shift contamination model (Definition 1), outputs a $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \le \epsilon$ with probability at least $1-\delta$. Moreover, it runs in time $2^{O(d)}\mathrm{poly}(n,d)$.*

Proof

Denote by $T = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$ the points from the α-corrupted version of $\mathcal{N}(\mu,\Sigma)$ and denote by $\mathcal{C}$ the cover set of Fact 21. The algorithm is the following: First, using the algorithm

from Fact 22, calculate a m_v for each $v \in \mathcal{C}$ such that $\left|m_v - v^\top \mu\right| \leq \epsilon/8$ (see next paragraph for more details on this step). Then, output the solution of the following linear program (note that the program always has a solution, as it is satisfied by $\widehat{\mu} = \mu$):

$$\begin{aligned} &\text{Find } \widehat{\mu} \in \mathbb{R}^d \text{ s.t.} \\ &\left|v^\top \widehat{\mu} - m_v\right| \leq \epsilon/4, \forall v \in \mathcal{C} \ . \end{aligned}$$

The claim is that this solution $\widehat{\mu}$ is indeed close to the target μ, since

$$\begin{aligned} \|\mu - \widehat{\mu}\|_2 &\leq 2 \max_{v \in \mathcal{C}} \left|v^\top (\mu - \widehat{\mu})\right| && \text{(using Fact 21)} \\ &\leq 2 \max_{v \in \mathcal{C}} (\left|v^\top \mu - m_v\right| + \left|m_v - v^\top \widehat{\mu}\right|) \\ &\leq 2(\epsilon/8 + \epsilon/4) < \epsilon \ . && (7) \end{aligned}$$

We now explain how to obtain the approximations m_v with the guarantee $\left|m_v - v^\top \mu\right| \leq \epsilon/8$. Fixing a direction $v \in \mathcal{C}$, we note that when $x \sim \mathcal{N}(\mu, \Sigma)$ then $v^\top x \sim \mathcal{N}(v^\top \mu, v^\top \Sigma v)$, thus $\{v^\top x_i\}_{i=1}^m$ is a set of samples from an α-corrupted version of $\mathcal{N}(v^\top \mu, v^\top \Sigma v)$. Note that since $\|\Sigma - I_d\|_F = O(1)$ we have that for any $y \in \mathbb{R}^d : \|y\| = 1$ it holds $y^\top \Sigma y = y^\top (\Sigma - I_d) y + y^\top I_d y = O(1)$. Hence the variance in the direction v is $v^\top \Sigma v = O(1)$. Thus, if we apply algorithm from Fact 22 with probability of failure $\delta' = \delta/|\mathcal{C}|$, the event $\left|m_v - v^\top \mu\right| \leq \epsilon/8$ will hold with probability at least $1 - \delta/|\mathcal{C}|$. By union bound, the probability all the events for $v \in \mathcal{C}$ hold simultaneously is at least $1 - \delta$. The number of samples for this application of Fact 22 is $2^{O((\alpha/\epsilon)^2)} \frac{\log(1/\delta')}{\alpha^2} = 2^{O((\alpha/\epsilon)^2)} \frac{d + \log(1/\delta)}{\alpha^2}$.

We conclude with the runtime analysis. The runtime to find the m_v's is $O(|\mathcal{C}| \mathrm{poly}(nd)) = 2^{O(d)} \mathrm{poly}(nd)$ since for each fixed $v \in \mathcal{C}$ we need $\mathrm{poly}(nd)$ time to calculate the projection $\{x_i^\top v\}$ of our dataset onto v and $\mathrm{poly}(n)$ time to run the one-dimensional estimator. The linear program can be solved using the ellipsoid algorithm. Consider the separation oracle that exhaustively checks all $2^{O(d)}$ constraints. We need $\mathrm{poly}(d) \log(\frac{R}{r})$ calls to that separation oracle, where R, r are the radii of the bounding spheres of the feasible region. First, $R \leq \epsilon$, because we have already shown in (7) that the feasible set belongs in a ball of radius ϵ around μ. Regarding the upper bound r, note that all $\widehat{\mu}$ inside a ball of radius $\epsilon/8$ around μ are feasible since $\left|v^\top \widehat{\mu} - m_v\right| \leq \left|v^\top \widehat{\mu} - v^\top \mu\right| + \left|v^\top \mu - m_v\right| \leq \|\widehat{\mu} - \mu\| + \epsilon/8 \leq \epsilon/4$. This means that $r = \epsilon/4$. Hence the total runtime for solving the LP is $2^{O(d)} \mathrm{poly}(d)$ or simply $2^{O(d)}$. Noting that for any $C > 0$, $2^{C(\alpha/\epsilon)^2}/\alpha^2 \leq 2^{3C/\epsilon^2}$ concludes the proof. ■

Linear Programming We derive duality of Linear Programs to view the problem through a different lens. Below are the two different dualities and the corresponding property that we use in our proofs. We start by introducing the widely used weak duality of Lagrange duality that works for any optimization problem and can be easily derived by hand via the standard method of linear combinations (see e.g. Bertsimas and Tsitsiklis (1997)).

Lemma 23 (Weak Duality of Lagrange duality (Chapter 8 of Luenberger (1997))) *Any constrained optimization problem, including those defined over infinite dimensional measurable spaces, has a Lagrange dual. Furthermore, if the primal is the maximization problem with optimal objective value p^*, and denote by d^* the optimal objective value of its Lagrange dual, then $p^* \leq d^*$.*

While weak duality of Lagrange duality captures the relation of the optimal objective values between primal and dual, Gauge duality establishes a fundamental correspondence between the set of optimal rays for the primal and dual problems (see, e.g. Tyrrell Rockafellar (1970)). We will use the homogeneous duality stated as below which is a special case of Gauge duality suited to a more constrained family of problems under consideration.

Lemma 24 (Homogeneous Duality) *Let $\mathcal{K}$ be a cone, and $F, G : \mathcal{K} \to \mathbb{R}$ be positively homogeneous functionals of degree 1, i.e. $\forall a \in \mathbb{R}_{\geq 0}$ and $f \in \mathcal{K}$, $F(af) = aF(f)$ and $G(af) = aG(f)$. Then the optimal value $opt(LP : inf)$ of LP:inf and the optimal value $opt(LP : sup)$ of LP:sup are reciprocals, i.e. $opt(LP : sup) = (opt(LP : inf))^{-1}$. Moreover, LP:inf and LP:sup are radially equivalent, i.e. the optimal solutions of LP:inf and these of LP:sup are positive constant scalar multiple of each other.*

$$\inf_{f \in C} F(f) \quad \text{s.t.} \quad G(f) = 1. \qquad \text{(LP:inf)}$$

$$\sup_{f \in C} G(f) \quad \text{s.t.} \quad F(f) = 1. \qquad \text{(LP:sup)}$$

Proof Denote by v_{inf} and v_{sup} the optimal solution of LP:inf and LP:sup, respectively. We will show that $v_{inf} = (v_{sup})^{-1}$ by showing $v_{inf} \leq (v_{sup})^{-1}$ and $v_{inf} \geq (v_{sup})^{-1}$. And we can see along the proof that each optimizer of LP:inf can induce a corresponding optimizer of LP:sup in the same direction, and vice versa.

We first show that $v_{sup} \geq (v_{inf})^{-1}$. Let f^* be a minimizer of LP:inf, then $F(f^*) = v_{inf}$ and $G(f^*) = 1$. Denote by $g := \frac{f^*}{v_{inf}}$, then by degree-1 homogeneity of F we have that $F(g) = \frac{1}{v_{inf}} F(f^*) = 1$ whence g is in the feasible region of LP:sup. Since $G(g) = \frac{1}{v_{inf}}$, we know that $v_{sup} \geq \frac{1}{v_{inf}}$, as desired.

We next show that $v_{sup} \leq (v_{inf})^{-1}$. Let g^* be an maximizer of LP:sup. Then $G(g^*) = v_{sup}$ and $F(g^*) = 1$. Denote by $f := \frac{g^*}{v_{sup}}$, then by degree-1 homogeneity of G we have that $G(f) = \frac{1}{v_{sup}} G(g^*) = 1$ whence f is in the feasible region of LP:inf. Since $F(f) = \frac{1}{v_{sup}}$, we know that $v_{inf} \leq \frac{1}{v_{sup}}$, which concludes the proof.

The radial equivalence of LP:inf and LP:sup follows from noting that the above construction of g via minimizer f^* of LP:inf gives an optimizer of LP:sup in the same direction as f^*, and similarly the above construction of f via maximizer g^* of LP:sup gives an optimizer of LP:inf in the same direction as g^*. ■

Appendix B. Omitted Proofs for Section 3

Lemma 8 (Properties of Reweighted Mixture) *Let $C_w > 1$, $c > 0$ be a small constant, $\alpha \in (0, 1/3 - c)$, $\mu, z_1, \cdots z_n, \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d \times d}$ be a positive definite matrix. Assume that $\|\mu\| < 0.1$, $\|\Sigma - I(1 - C_w/\sqrt{d})^{-1}\|_F = 0.1$, and $C_w \ll d^{1/2}$. Furthermore, define*

$$\tilde{\Sigma} := \left(\Sigma^{-1} + \frac{C_w}{d^{1/2}} I\right)^{-1}, \tilde{\mu} = \tilde{\Sigma}\Sigma^{-1}\mu\,, \tilde{z}_i = \tilde{\Sigma}\Sigma^{-1} z_i. \tag{1}$$

Let D be a Gaussian mixture defined as $D := (1 - \alpha)\mathcal{N}(\mu, \Sigma) + \frac{\alpha}{n}\sum_{i=1}^{n} \mathcal{N}(z_i, \Sigma)$, $w(x)$ be a weight function of the form $w(x) = \exp\left(-\frac{C_w \|x\|^2}{2d^{1/2}}\right)$. Then there exist another Gaussian mixture

$D_w = \beta^* \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}) + \sum_{i=1}^n \beta_i \mathcal{N}(\tilde{z}_i, \tilde{\Sigma})$ for some mixing coefficients $\beta^*, \beta_1, \cdots, \beta_n$ such that the following holds: (1) $\tilde{\Sigma}$ is positive definite and $\|\tilde{\Sigma}\|_2 < 2$, (2) $p_D(x)w(x) \propto p_{D_w}(x)$, (3) $\beta^* > 2/3 + c/2$, (4) $\forall v \in \mathbb{S}^{d-1}$, $|v^\top \tilde{\mu} - v^\top \mu| = O(C_w/d^{1/2})$, and (5) for all positive integers $k \leq 4$, the nuclear norm of the $(k-1,1)$-flattening of degree-k Hermite moment tensor of D_w satisfies $\left\|\mathrm{flat}_{(k-1,1)}(\mathbb{E}_{x\sim D_w}[H_k(x)])\right\|_* \leq O(d/C_w)$.

Proof We give the proof of each of the items in the lemma statement in order.

Covariance Spectral Analysis First, we show that $\tilde{\Sigma} = \left(\Sigma^{-1} + \frac{C_{\mathrm{w}}}{d^{1/2}} I\right)^{-1}$ is positive definite and $\|\tilde{\Sigma}\|_2 < 2$. Indeed, the eigenvalues of $\tilde{\Sigma}$ are all positive follows from the fact that Σ is positive definite and $\frac{C_{\mathrm{w}}}{d^{1/2}} > 0$. Furthermore,

$$\|\tilde{\Sigma}\|_2 \leq \left(\frac{1}{\|\Sigma\|_2} + \frac{C_{\mathrm{w}}}{d^{1/2}}\right)^{-1} \leq \|\Sigma\|_2 < 2,$$

where the last inequality follows from Lemma 6, which concludes the proof of (1).

New Gaussian Mixture Expression Secondly, we analyze the effect of the weight function $w(x)$ on a single Gaussian component.

Claim 25 (Reweighting a single component) *For any vector $u \in \mathbb{R}^d$ and any positive definite matrix $S \in \mathbb{R}^{d\times d}$, it holds that*

$$p_{\mathcal{N}(u,S)}(x) \cdot w(x) = \det\left(I_d + \frac{C_w}{d^{1/2}} S\right)^{-1/2} \exp\left(-\frac{1}{2} u^\top \left(S + \frac{d^{1/2}}{C_w} I_d\right)^{-1} u\right) p_{\mathcal{N}(\tilde{u},\tilde{S})}(x)\,,$$

where $\tilde{S} = \left(S^{-1} + (C_w/d^{1/2})I\right)^{-1}$, $\tilde{u} = \tilde{S}S^{-1}u$.

Proof For convenience, we define $\lambda = C_{\mathrm{w}}/d^{1/2}$. Note that we have

$$p_{\mathcal{N}(u,S)}(x) \cdot w(x) = (2\pi)^{-d/2} \det(S)^{-1/2} \exp\left(-\frac{1}{2}(x-u)^\top S^{-1}(x-u) - \frac{\lambda}{2} x^\top I_d x\right)\,.$$

By treating $w(x)$ as $\left(\frac{2\pi}{\lambda}\right)^{d/2} \cdot p_{\mathcal{N}\left(0, \frac{1}{\lambda} I_d\right)}(x)$ then applying Section 8.1.8 of Petersen and Pedersen (2012) we have that,

$$p_{\mathcal{N}(u,S)}(x) \cdot w(x) = \det\left(I_d + \lambda S\right)^{-1/2} \exp\left(-\frac{1}{2} u^\top \left(S + \frac{1}{\lambda} I_d\right)^{-1} u\right) p_{\mathcal{N}(\tilde{u},\tilde{S})}(x)\,.$$

This concludes the proof of Claim 25. ■

For convenience, we define

$$C_{\mathrm{det}} = \det\left(I_d + \frac{C_{\mathrm{w}}}{d^{1/2}}\Sigma\right)^{-1/2}, \tag{8}$$

$$\bar{\beta}^* = \exp\left(-\frac{1}{2}\mu^\top \left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}} I_d\right)^{-1} \mu\right)\,,$$

$$\bar{\beta}_i = \exp\left(-\frac{1}{2} z_i^\top \left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}} I_d\right)^{-1} z_i\right) \forall i \in [n]$$

Applying Claim 25 to each of the Gaussian component within D gives that $p_D(x)w(x)$ is equal to

$$p_D(x)w(x) = (1-\alpha)\ C_{\mathrm{det}}\bar{\beta}^* p_{\mathcal{N}(\tilde{\mu},\tilde{\Sigma})}(x) + \frac{\alpha}{n} C_{\mathrm{det}} \sum_{i=1}^{n} \bar{\beta}_i p_{\mathcal{N}(\tilde{z}_i,\tilde{\Sigma})}(x)\,,$$

where $\tilde{\mu}, \tilde{z}_i, \tilde{\Sigma}$ are defined as in Equation (1). Furthermore, if we define C_N as the sum of the coefficients in front of the Gaussian density functions on the right hand side, i.e.,

$$C_N := C_{\mathrm{det}} \left((1-\alpha)\ \bar{\beta}^* + \frac{\alpha}{n} \sum_{i=1}^{n} \bar{\beta}_i\right)\,, \tag{9}$$

then we immediately have that

$$\frac{(1-\alpha)\ C_{\mathrm{det}}\ \bar{\beta}^*}{C_N} \mathcal{N}\left(\tilde{\mu}, \tilde{\Sigma}\right) + \sum_{i=1}^{n} \frac{\alpha n^{-1}\ C_{\mathrm{det}}\ \bar{\beta}_i}{C_N} \mathcal{N}\left(\tilde{z}_i, \tilde{\Sigma}\right) \tag{10}$$

is a proper Gaussian mixture. We can therefore set

$$\beta^* := \frac{(1-\alpha)\ C_{\mathrm{det}}\ \bar{\beta}^*}{C_N}\,, \beta_i := \frac{\alpha n^{-1}\ C_{\mathrm{det}}\ \bar{\beta}_i}{C_N}\,, \tag{11}$$

and conclude that $p_D(x)w(x) = C_N\ p_{D_w}(x)$ for $D_w := \beta^* \mathcal{N}\left(\tilde{\mu}, \tilde{\Sigma}\right) + \sum_{i=1}^{n} \beta_i \mathcal{N}\left(\tilde{z}_i, \tilde{\Sigma}\right)$. This concludes the proof of (2).

Lower Bound on Clean Component Weight We next proceed to bound from below β^*. Since $\beta^* := \frac{(1-\alpha)\ C_{\mathrm{det}}\ \bar{\beta}^*}{C_N}$, it suffices if we can bound from above C_N and bound from below $\bar{\beta}^*$. For $\bar{\beta}^*$, we have that

$$\begin{aligned}
\bar{\beta}^* &= \exp\left(-\frac{1}{2}\mu^\top \left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}} I_d\right)^{-1} \mu\right) \\
&\geq \exp\left(-\frac{1}{2}\|\mu\|_2^2 \frac{C_{\mathrm{w}}}{d^{1/2}}\right) \\
&\geq \frac{4+3c}{4+6c}\,,
\end{aligned} \tag{12}$$

where in the first inequality we use that the minimum singular value of $\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}} I$ is at least $d^{1/2}/C_{\mathrm{w}}$, and the last inequality follows from the assumption that $\|\mu\|_2 < 0.1$ and $C_w < 200 d^{1/2} \log\left(1 + \frac{3c}{4+3c}\right)$.

For C_N, since Σ is positive definite, then $\left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}} I_d\right)^{-1} \succ 0$, which implies that $\mu^T \left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}} I_d\right)^{-1} \mu > 0$. Hence, $\bar{\beta}^*, \bar{\beta}_i < 1$ for all $i = 1, \cdots n$. Thus,

$$\Omega\left(C_{\mathrm{det}}\right) \leq C_N \leq C_{\mathrm{det}}\,, \tag{13}$$

where the first inequality follows from Equations (9) and (12). Combining Equations (11) to (13) gives that $\beta^* \geq (1-\alpha)\ \frac{4+3c}{4+6c} > \frac{2+3c}{3}\ \frac{4+3c}{4+6c} = \frac{2}{3} + \frac{c}{2}$. This concludes the proof of (3).

Bound Reweighting Bias Next, we show that at any direction the mean of the clean component $\mathcal{N}(\mu, \Sigma)$ won't deviate too much after reweighting. Indeed, for all $v \in \mathbb{S}^{d-1}$

$$\left| v^\top \left(\Sigma^{-1} + \frac{C_{\mathrm{w}}}{d^{1/2}} I \right)^{-1} \Sigma^{-1}\mu - v^\top \mu \right| = \left| v^\top \left(\left(I + \frac{C_{\mathrm{w}}}{d^{1/2}} \Sigma \right)^{-1} - I_d \right) \mu \right|$$
$$\leq \|\mu\| \left\| \left(I + \frac{C_{\mathrm{w}}}{d^{1/2}} \Sigma \right)^{-1} - I \right\|_2 = O(C_{\mathrm{w}}/d^{1/2}),$$

where the second inequality follows from Cauchy Schwartz inequality and the definition of operator norm, and the last inequality follows from the fact that $\|\mu\| = O(1)$ and $O(1) I_d \succeq \Sigma \succ 0$ hence $\left\| \left(I_d + \frac{C_{\mathrm{w}}}{d^{1/2}} \Sigma \right)^{-1} - I_d \right\|_2 = 1 - \frac{1}{1 + O\left(C_{\mathrm{w}}/d^{1/2}\right)} = O\left(C_{\mathrm{w}}/d^{1/2}\right)$ as desired, which concludes the proof of (4).

Bound Low-degree Hermite Moments Lastly, we show that the $(k-1, 1)$-flattened degree-k Hermite moment tensor of the new mixture distribution D_w has its nuclear norm bounded from above by $O(d/C_{\mathrm{w}})$. For this purpose, we will first bound from above the Frobenius norm between the reweighted covariance matrix $\tilde{\Sigma}$ and the identity.

Claim 26 *It holds that* $\left\| \tilde{\Sigma} - I \right\|_F = O(1)$.

Proof For convenience, we define the scalars $\gamma = (1 - C_{\mathrm{w}}/\sqrt{d})^{-1}$ and $\zeta = C_{\mathrm{w}}/\sqrt{d}$, noting that $1 - \zeta = \gamma^{-1}$. We can then rewrite the deviation of the updated matrix as

$$\tilde{\Sigma} - I = (\Sigma^{-1} + \zeta I)^{-1} - I = (\Sigma^{-1}(I + \zeta\Sigma))^{-1} - I = (I + \zeta\Sigma)^{-1}\Sigma - I.$$

By factoring $(I + \zeta\Sigma)^{-1}$ to the left, we get that

$$\begin{aligned} \tilde{\Sigma} - I &= (I + \zeta\Sigma)^{-1}(\Sigma - (I + \zeta\Sigma)) \\ &= (I + \zeta\Sigma)^{-1}(\Sigma(1 - \zeta) - I) \\ &= (I + \zeta\Sigma)^{-1}(1 - \zeta)(\Sigma - (1 - \zeta)^{-1} I) \\ &= (I + \zeta\Sigma)^{-1}(1 - \zeta)(\Sigma - \gamma I). \end{aligned}$$

Given the assumption that $\zeta = C_{\mathrm{w}}/\sqrt{d} \leq 1/2$ hence $|1 - \zeta| = O(1)$. Since $\Sigma \succeq 0$ and $\zeta \geq 0$, all eigenvalues of $I + \zeta\Sigma$ are at least 1, hence $\|(I + \zeta\Sigma)^{-1}\|_2 \leq 1$. Therefore $\|(I + \zeta\Sigma)^{-1}(1 - \zeta)\|_2 \leq |1 - \zeta| = O(1)$.

Recall by assumption $\|\Sigma - I\gamma\|_F^2 = O(1)$. Hence, we have that the i-th singular value of a matrix product can be bounded from above by the product of the spectral norm of one factor and the i-th singular value of the other. Thus, we have that

$$\sigma_i(\tilde{\Sigma} - I) \leq \|(I + \zeta\Sigma)^{-1}(1 - \zeta)\|_2 \sigma_i(\Sigma - \gamma I) \lesssim \sigma_i(\Sigma - \gamma I).$$

Consequently, we can conclude that $\|\tilde{\Sigma} - I\|_F^2 = \sum_i \sigma_i(\tilde{\Sigma} - I)^2 \lesssim \sum_i \sigma_i(\Sigma - \gamma I)^2 = O(1)$. This concludes the proof of Claim 26. ∎

We then proceed to bound the low-degree Hermite moment tensor of each of the Gaussian distribution $\mathcal{N}(u, \tilde{\Sigma})$ within the reweighted mixture D_w.

Claim 27 *Let $u \in \mathbb{R}^d$ and $M_k(u)$ be the $(k-1,1)$-flattening of the degree-k Hermite moment tensor of $\mathcal{N}(u, \tilde{\Sigma})$. Assume that $\|\tilde{\Sigma} - I\|_F \leq O(1)$. Then it holds that*

$$\|M_4(u)\|_* \leq O\left(\|u\|_2^4 + \sqrt{d}\,\|u\|^2 + \sqrt{d}\right),$$
$$\|M_3(u)\|_* \leq O\left(\|u\|_2^3 + \sqrt{d}\|u\|\right),$$
$$\|M_2(u)\|_* \leq O\left(\|u\|_2^2 + \sqrt{d}\right),$$
$$\|M_1(u)\|_* \leq O\left(\|u\|_2\right).$$

Proof We focus on bounding from above the nuclear norm of the $(3,1)$-flattening of the degree-4 Hermite moment tensor as the arguments for the lower degree Hermite moment tensor are similar and simpler. To analyze the nuclear norm of the flattened matrix $M_4(u)$ derived from the fourth-order Hermite tensor $\mathbb{E}_{x\sim\mathcal{D}}[H_4(x)]$, we first apply Fact 17 to expand the components of M into sums over partitions of $\{1,2,3,4\}$ into sets of size one and two. Specifically, the entries M_{i_1,i_2,i_3,i_4} are given by a constant multiplied by a sum of terms involving tensor products of the mean components u and covariance perturbations $\Delta := \left(\tilde{\Sigma} - I\right)$. For all the terms whose last tensor component is u, we note that the resulting tensor product will be a rank-1 matrix after flattening. In such cases, the nuclear norm is the same as the Frobenius norm of the tensor, which is at most $O\left(\|u\|^4\right)$ for $u^{\otimes 4}$ or $O\left(\|u\|^2\right)$ for $u \otimes \Delta \otimes u$ and $\Delta \otimes u \otimes u$. For all the terms whose last tensor component is Δ, the nuclear norm after flattening is at most the nuclear norm of Δ multiplied with the Frobenius norm of either $uu^\top$ or Δ. In other words, the nuclear norm can be bounded from above by $\sqrt{d}\|\Delta\|_F \ \left(\|\Delta\|_F + \|u\|_2^2\right) \leq O\left(\sqrt{d} + \sqrt{d}\|u\|_2^2\right)$. The bound for $M_3(u)$, $M_2(u)$, and $M_1(u)$ can be established via a similar case analysis on the tensors $u^{\otimes 3}$, $u \otimes \Delta \otimes u$, $\Delta \otimes u^{\otimes 2}$, $u^{\otimes 2}$, Δ, u. This concludes the proof of Claim 27. ■

We note that the $(k-1,1)$-flattening of degree-k Hermite moment tensor of D_w can be bounded from above by a convex combination of

$$\exp\left(-\frac{1}{2}u^\top\left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}}I_d\right)^{-1}u\right)\|M_k(u)\|_*$$

for $u \in \{\mu, z_1, \cdots z_n\}$. Combining Claims 26 and 27 gives that

$$\exp\left(-\frac{1}{2}u^\top\left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}}I_d\right)^{-1}u\right)\|M_k(u)\|_*$$
$$\leq O(1)\ \exp\left(-\frac{1}{2}u^\top\left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}}I_d\right)^{-1}u\right)\ \left(\|u\|_2^4 + \|u\|_2^3 + \sqrt{d}\|u\|_2^2 + \sqrt{d}\|u\|_2 + \sqrt{d}\right)$$
$$\leq O(1)\ \exp\left(-\frac{C_{\mathrm{w}}}{2\sqrt{d}}\|u\|_2^2\right)\ \left(\|u\|_2^4 + \|u\|_2^3 + \sqrt{d}\|u\|_2^2 + \sqrt{d}\|u\|_2 + \sqrt{d}\right),$$

where in the last inequality we have again used the fact that the minimal singular value of $\left(\Sigma + \frac{d^{1/2}}{C_{\mathrm{w}}} I_d\right)^{-1}$ is at least $C_{\mathrm{w}}/\sqrt{d}$.

Note that the each of the term is of the form $f_k(x) = x^k \exp(-\gamma x^2)$, where $\gamma = C_{\mathrm{w}}/\sqrt{d} = o(1)$ and $k \in [4]$, multiplied by powers of d. Hence, we will first prove the following generic claim.

Claim 28 *For any $x > 0$ and $k \in [4]$, it holds that $f_k(x) = x^k \exp(-\gamma x^2) \le (k/(2e\gamma))^{k/2}$.*

Proof This follows by solving the maximum value of f_k using calculus. ■

It thereby follows that

$$\begin{aligned}
&\exp\left(-\frac{C_{\mathrm{w}}}{2\sqrt{d}}\|u\|_2^2\right)\left(\|u\|_2^4 + \|u\|_2^3 + \sqrt{d}\|u\|_2^2 + \sqrt{d}\|u\|_2 + \sqrt{d}\right)\\
&\le O(1)\left(\left(\sqrt{d}/C_{\mathrm{w}}\right)^2 + \left(\sqrt{d}/C_{\mathrm{w}}\right)^3 + \sqrt{d}\left(\sqrt{d}/C_{\mathrm{w}}\right) + \sqrt{d}\right)\\
&\le O(d/C_{\mathrm{w}})\,,
\end{aligned}$$

whenever $1 < C_{\mathrm{w}} \lesssim \sqrt{d}$. This concludes the proof of (5) as well as Lemma 8. ■

Appendix C. Omitted Proofs for Section 4

Lemma 10 (Concentration) *Let $C_w > 1, \alpha \in (0, 1/2)$, $\mu, z_1, \cdots z_n \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Assume that $\|\mu\| = O(1)$, $\|\Sigma\|_2 = O(1)$, and $C_w \ll d^{1/2}$. Let D be a Gaussian mixture defined as $D := (1-\alpha)\mathcal{N}(\mu, \Sigma) + \frac{\alpha}{n}\sum_{i=1}^n \mathcal{N}(z_i, \Sigma)$, and $w(x)$ be a weight function of the form $w(x) = \exp\left(-\frac{C_w\|x\|^2}{d^{1/2}}\right)$. Define $C_N = \mathbb{E}_{x\sim D}[w(x)]$. Let $\{x_i\}_{i=1}^m$ be $m = \exp(CC_w^2)\cdot d^4/(\delta\epsilon^2)$ i.i.d. samples drawn from D, where $C > 0$ is a sufficiently large constant. Then we have that with probability at least $1-\delta$, for $j = 0, 1, \cdots, 4$, it holds $\left\|\frac{1}{m}\sum_{i=1}^m \frac{H_j(x_i)w(x_i)}{C_N} - \mathbb{E}_{x\sim D}\left[\frac{H_j(x)w(x)}{C_N}\right]\right\|_F \le \epsilon$.*

Proof Fix $j \in \{0, \ldots, 4\}$. For simplicity, denote by $T := \frac{1}{m}\sum_{i=1}^m \frac{H_j(x_i)w(x_i)}{C_N} - \mathbb{E}_{x\sim D}\left[\frac{H_j(x)w(x)}{C_N}\right]$ the difference tensor between the empirical and the population weighted Hermite moment tensors. Our goal is to show that $\|T\|_F^2 \le \epsilon^2$ with high probability. To do so, it suffices for us to show an upper bound on the expected value $\mathbb{E}\left[\|T\|_F^2\right] \ll \epsilon^2$, and then the result will follow from Markov's inequality. Note that $x_1, \cdots, x_m$ are i.i.d. samples drawn from some Gaussian mixture where all the Gaussian components share the same covariance matrix Σ. We will condition on an arbitrary sequence of Gaussian components from which each sample is drawn and $x_1, \cdots, x_m$ can be viewed as independent samples drawn from the Gaussian distributions $\mathcal{N}(u_1, \Sigma), \cdots, \mathcal{N}(u_m, \Sigma)$ respectively. We therefore have that

$$\sum_{l_1,\cdots,l_j} \mathbb{E}\left[T_{l_1,\cdots,l_j}^2\right] = \sum_{l_1,\cdots,l_j} \mathrm{Var}_{x_1,\cdots,x_m\sim D}\left[\frac{1}{m}\sum_{i=1}^m \frac{(H_j(x_i))_{l_1,\cdots,l_j}\, w(x_i)}{C_N}\right]$$

$$
\begin{aligned}
&= \frac{1}{m} \sum_{l_1,\cdots,l_j} \mathrm{Var}_{x\sim D}\left[\frac{(H_j(x))_{l_1,\cdots,l_j}\, w(x_i)}{C_N}\right] \\
&\le \frac{1}{m} \sum_{l_1,\cdots,l_j} \mathbb{E}_{x\sim D}\left[\frac{(H_j(x))^2_{l_1,\cdots,l_j}\, w^2(x_i)}{C_N^2}\right] \\
&= \frac{1}{m}\mathbb{E}_{x\sim D}\left[\frac{\|H_j(x)\|_F^2\, w^2(x_i)}{C_N^2}\right] \\
&\le \frac{1}{m}\max_u \mathbb{E}_{x\sim N(u,\Sigma)}\left[\|H_j(x)\|_F^2\, \frac{w^2(x_i)}{C_N^2}\right] \qquad (14)
\end{aligned}
$$

where the first inequality follows from the definition of variance, and the second inequality follows from the fact that D is a mixture of Gaussian with covariance Σ. We therefore focus on bounding from above the term $\mathbb{E}_{x\sim\mathcal{N}(u,\Sigma)}\left[\|H_j(x)\|_F^2\left(\frac{w(x)}{C_N}\right)^2\right]$ for an arbitrary $u\in\mathbb{R}^d$. Note that

$$
w(x)^2 = \exp\left(-\frac{C_{\mathrm{w}}\|x\|^2}{d^{1/2}}\right) = \exp\left(-\frac{(2C_{\mathrm{w}})\|x\|^2}{2d^{1/2}}\right).
$$

Therefore, by Claim 25 applied with the parameter $2C_w$,

$$
p_{\mathcal{N}(u,\Sigma)}(x)\cdot w(x)^2 = \det\left(I_d + \frac{2C_{\mathrm{w}}}{d^{1/2}}\Sigma\right)^{-1/2} \exp\left(-\frac{1}{2}u^T\left(\Sigma + \frac{d^{1/2}}{2C_{\mathrm{w}}}I_d\right)^{-1}u\right)\; p_{\mathcal{N}(\tilde{u},\tilde{\Sigma})}(x),
$$

where $\tilde{\Sigma} = \left(\Sigma^{-1} + \frac{2C_{\mathrm{w}}}{d^{1/2}}I_d\right)^{-1}$ and $\tilde{u} = \tilde{\Sigma}\Sigma^{-1}u$. Thus, we have that

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{N}(u,\Sigma)}\left[\|H_j(x)\|_F^2\left(\frac{w(x)}{C_N}\right)^2\right] \\
&= \frac{1}{C_N^2}\det\left(I_d + \frac{2C_{\mathrm{w}}}{d^{1/2}}\Sigma\right)^{-1/2} \exp\left(-\frac{1}{2}u^T\left(\Sigma + \frac{d^{1/2}}{2C_{\mathrm{w}}}I_d\right)^{-1}u\right)\mathbb{E}_{x\sim\mathcal{N}(\tilde{u},\tilde{\Sigma})}\left[\|H_j(x)\|_F^2\right]. \qquad (15)
\end{aligned}
$$

We analyze the right hand side of the equation term by term. The following claim helps us control the square Frobenius norm of the Hermite tensor under $\mathcal{N}(\tilde{u},\tilde{\Sigma})$.

Claim 29 *Let $x\sim\mathcal{N}(u,S)$ be a random vector in $\mathbb{R}^d$, where the covariance matrix satisfies $\|S\|_2\le O(1)$. Then, for any $j\in\{1,2,3,4\}$, the expected squared Frobenius norm satisfies:*

$$
\mathbb{E}\left[\|H_j(x)\|_F^2\right] \lesssim \|u\|_2^{2j} + d^j.
$$

Proof Recall that the degree-j Hermite tensor $H_j(x)$ can be expanded as a sum of terms involving tensor products of x and the identity matrix I_d. Specifically, we have that

$$
H_j(x) = \frac{1}{\sqrt{j!}}\sum_{l=0}^{\lfloor j/2\rfloor} c_{j,l}\,\mathrm{Sym}\left(x^{\otimes(j-2l)}\otimes I_d^{\otimes l}\right).
$$

By the triangle inequality and the fact that Sym is an orthogonal projection (hence non-expansive in Frobenius norm), we have

$$\begin{aligned}\|H_j(x)\|_F &\le \frac{1}{\sqrt{j!}} \sum_{l=0}^{\lfloor j/2 \rfloor} |c_{j,l}| \left\|\mathrm{Sym}\big(x^{\otimes(j-2l)} \otimes I_d^{\otimes l}\big)\right\|_F \\ &\le \frac{1}{\sqrt{j!}} \sum_{l=0}^{\lfloor j/2 \rfloor} |c_{j,l}| \left\|x^{\otimes(j-2l)} \otimes I_d^{\otimes l}\right\|_F.\end{aligned}$$

Moreover, using $\|A \otimes B\|_F = \|A\|_F \|B\|_F$, $\|x^{\otimes k}\|_F = \|x\|_2^k$, and $\|I_d\|_F = \sqrt{d}$, we get

$$\left\|x^{\otimes(j-2l)} \otimes I_d^{\otimes l}\right\|_F = \|x\|_2^{j-2l} \|I_d\|_F^l = \|x\|_2^{j-2l}\, d^{l/2}.$$

Defining $c'_{j,l} := |c_{j,l}|/\sqrt{j!}$, we obtain the pointwise bound

$$\|H_j(x)\|_F \le \sum_{l=0}^{\lfloor j/2 \rfloor} c'_{j,l} \|x\|_2^{j-2l}\, d^{l/2} \lesssim (\|x\|_2 + \sqrt{d})^j,$$

where the last inequality follows from the binomial theorem and the fact that $\max_l c'_{j,l} = O(1)$ for fixed j (in particular, for $j \le 4$). Squaring this expression yields:

$$\|H_j(x)\|_F^2 \lesssim \left(\|x\|_2^{2j} + d^j\right).$$

Next, we decompose the random vector as $x = u + S^{1/2}z$, where $z \sim \mathcal{N}(0, I_d)$. Using the inequality $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ and the operator norm bound $\|S^{1/2}\|_2 = O(1)$, we have $\|x\|_2^2 \le 2\|u\|_2^2 + O(1)\|z\|_2^2$. Consequently, the $2j$-th moment satisfies:

$$\mathbb{E}\left[\|x\|_2^{2j}\right] \lesssim \|u\|_2^{2j} + \mathbb{E}[\|z\|_2^{2j}] + d^j\ .$$

Since z is a standard Gaussian vector, $\mathbb{E}[\|z\|_2^{2j}] \lesssim d^j$. Substituting this back into the bound for the Hermite tensor, we conclude that:

$$\mathbb{E}\left[\|H_j(x)\|_F^2\right] \lesssim \|u\|_2^{2j} + d^j\ .$$

This concludes the proof of Claim 29. ■

For the distribution $\mathcal{N}(\tilde{u}, \tilde{\Sigma})$, recall that $\tilde{\Sigma} = \left(\Sigma^{-1} + \frac{2C_{\mathrm{w}}}{\sqrt{d}} I_d\right)^{-1}$. Hence, writing $a := \frac{2C_{\mathrm{w}}}{\sqrt{d}}$,

$$\|\tilde{\Sigma}\|_2 = \left\|\left(\Sigma^{-1} + aI_d\right)^{-1}\right\|_2 = \frac{1}{\lambda_{\min}(\Sigma^{-1} + aI_d)} = \frac{1}{\lambda_{\min}(\Sigma^{-1}) + a} = \frac{1}{\frac{1}{\lambda_{\max}(\Sigma)} + a} \le \lambda_{\max}(\Sigma) = \|\Sigma\|_2.$$

In particular, as long as $\|\Sigma\|_2 = O(1)$, we get $\|\tilde{\Sigma}\|_2 \le O(1)$. Moreover, the ℓ_2 norm of $\tilde{u}$ is at most $\|\tilde{\Sigma}\Sigma^{-1}u\|_2 = O(\|u\|_2)$. Applying Claim 29 then gives that

$$\exp\left(-\frac{1}{2} u^T \left(\Sigma + \frac{d^{1/2}}{2C_{\mathrm{w}}} I_d\right)^{-1} u\right) \mathbb{E}_{x \sim \mathcal{N}(\tilde{u}, \tilde{\Sigma})}\left[\|H_j(x)\|_F^2\right]$$

$$\lesssim \left(d^j + \|u\|^{2j}\right) \exp\left(-\frac{1}{2}u^T\left(\Sigma + \frac{d^{1/2}}{2C_{\mathrm{w}}}I_d\right)^{-1}u\right)$$

$$= \|u\|^{2j} \cdot \exp\left(-c\frac{C_{\mathrm{w}}}{\sqrt{d}}\|u\|^2\right) + d^j \qquad \text{(for a sufficiently small constant } c > 0\text{)}$$

$$\lesssim \left(\left(\sqrt{d}/C_{\mathrm{w}}\right)^j + d^j\right) = O(d^j) \tag{16}$$

where the last line uses Claim 28 and the assumption that $C > 1$.

Next, we proceed to bound from above $\det\left(I_d + \frac{2C_{\mathrm{w}}}{d^{1/2}}\right)/C_N^2$. Using Equation (13), we have that $C_N \geq \Omega(C_{\mathrm{det}})$ for $C_{\mathrm{det}} = \det\left(I_d + \frac{C_{\mathrm{w}}}{d^{1/2}}\Sigma\right)^{-1/2}$. It follows that

$$\det\left(I_d + \frac{2C_{\mathrm{w}}}{d^{1/2}}\right)/C_N^2 \leq O(1)\,\frac{\det\left(I_d + \frac{C_{\mathrm{w}}}{d^{1/2}}\Sigma\right)}{\det\left(I_d + \frac{2C_{\mathrm{w}}}{d^{1/2}}\Sigma\right)^{1/2}}$$

$$= O(1)\,\frac{\prod_{i=1}^d\left(1 + \frac{C_{\mathrm{w}}}{d^{1/2}}\sigma_i(\Sigma)\right)}{\prod_{i=1}^d\left(1 + \frac{2C_{\mathrm{w}}}{d^{1/2}}\sigma_i(\Sigma)\right)^{1/2}}$$

$$= O(1)\,\exp\left(\sum_{i=1}^d\left(\ln\left(1 + \frac{C_{\mathrm{w}}}{d^{1/2}}\sigma_i(\Sigma)\right) - \frac{1}{2}\ln\left(1 + \frac{2C_{\mathrm{w}}}{d^{1/2}}\sigma_i(\Sigma)\right)\right)\right)$$

$$= O(1)\,\exp\left(\sum_{i=1}^d\left(\left(\frac{C_{\mathrm{w}}}{d^{1/2}}\sigma_i(\Sigma) - \frac{C_{\mathrm{w}}^2}{2d}\sigma_i(\Sigma)^2 + o\left(1/d\right)\right) - \left(\frac{C_{\mathrm{w}}}{d^{1/2}}\sigma_i(\Sigma) - \frac{C_{\mathrm{w}}^2}{d}\sigma_i(\Sigma)^2 + o(1/d)\right)\right)\right)$$

(Taylor Expansion of $\ln(\cdot)$.)

$$= O(1)\,\exp\left(\sum_{i=1}^d\left(\frac{C_{\mathrm{w}}^2}{2d}\sigma_i(\Sigma)^2 + o\left(1/d\right)\right)\right)$$

$$= \exp(o(1))\exp\left(\frac{C_{\mathrm{w}}^2}{2d}\sum_{i=1}^d\sigma_i(\Sigma)^2\right) \leq O(1)\,\exp\left(\frac{C_{\mathrm{w}}^2}{2d}\|\Sigma\|_F^2\right) \leq \exp(O(C_{\mathrm{w}}^2)). \tag{17}$$

where the last inequality uses the assumption that $\|\Sigma\|_F^2 = O(d)$. Combining Equations (15) to (17) then gives that

$$\mathbb{E}_{x\sim\mathcal{N}(u,\Sigma)}\left[\|H_j(x)\|_F^2\left(\frac{w(x)}{C_N}\right)^2\right] \lesssim d^j\exp(O(C_{\mathrm{w}}^2))$$

Substituting this back into Equation (14) yields that

$$\sum_{l_1,\cdots,l_j}\mathbb{E}\left[T_{l_1,\cdots,l_j}^2\right] \lesssim \frac{1}{m}d^j\exp(O(C_{\mathrm{w}}^2))\,.$$

Applying Markov's inequality to the nonnegative random variable $\|T\|_F^2$:

$$\Pr[\|T\|_F > \epsilon] = \Pr[\|T\|_F^2 > \epsilon^2] \leq \frac{\mathbb{E}\|T\|_F^2}{\epsilon^2} \leq \frac{O_j(1)d^j\exp(O(C_{\mathrm{w}}^2))}{m\epsilon^2}\,.$$

Hence, if $m = d^j \exp(O(C_{\mathrm{w}}^2))\epsilon^{-2}\delta^{-1}$ then $\Pr[\|T\|_F \le \epsilon] \ge 1-\delta$. For j at most a constant (4 in our case) we have that $m = d^4 \exp(O(C_{\mathrm{w}}^2))\epsilon^{-2}\delta^{-1}$ samples suffice. This concludes the proof of Lemma 10. ■

Corollary 11 *Let $\alpha, \epsilon, \tau, \in (0, 1/2)$, $\mu, z_1, \cdots z_n, \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Assume that $\|\mu\| = O(1)$, $\|\Sigma\|_2 = O(1)$, and $\epsilon^{-12} \ll d^{1/2}$. Let D be a Gaussian mixture defined as $D := (1-\alpha)\mathcal{N}(\mu, \Sigma) + \frac{\alpha}{n}\sum_{i=1}^n \mathcal{N}(z_i, \Sigma)$, $w : \mathbb{R}^d \mapsto \mathbb{R}_+$ be the weight function defined as in line 4 and $C_N := \mathbb{E}_{x\sim D}[w(x)]$. Then the empirical moment tensor $\widehat{T}_i$ computed in line 7 satisfies $\left\|\widehat{T}_i - C_N^{-1}\mathbb{E}_{x\sim D}\left[H_i(x)w(x)\right]\right\|_F \le \epsilon$ with probability at least $1-\tau$.*

Proof Fix $i \in \{1,2,3,4\}$ and let $x_1, \ldots, x_m \overset{i.i.d.}{\sim} D$. Write

$$\bar{C}_N := \frac{1}{m}\sum_{t=1}^m w(x_t), \qquad \widehat{T}_i := \frac{1}{m}\sum_{t=1}^m \frac{H_i(x_t)w(x_t)}{\bar{C}_N}, \qquad T_i := \frac{1}{C_N}\mathbb{E}_{x\sim D}[H_i(x)w(x)].$$

Also define the C_N-normalized empirical tensor

$$\bar{T}_i := \frac{1}{m}\sum_{t=1}^m \frac{H_i(x_t)w(x_t)}{C_N}.$$

Then

$$\begin{aligned}\widehat{T}_i - T_i &= \left(\frac{1}{m}\sum_{t=1}^m \frac{H_i(x_t)w(x_t)}{\bar{C}_N}\right) - \frac{1}{C_N}\mathbb{E}[H_i(x)w(x)] \\ &= \left(\frac{1}{m}\sum_{t=1}^m H_i(x_t)w(x_t)\right)\left(\frac{1}{\bar{C}_N} - \frac{1}{C_N}\right) \;+\; \left(\bar{T}_i - T_i\right).\end{aligned}$$

First we control the second term. Apply Lemma 10 with index $j = i$ and failure probability $\tau/3$ (and accuracy parameter $\epsilon/2$). With probability at least $1-\tau/3$,

$$\|\bar{T}_i - T_i\|_F \le \epsilon/2. \tag{18}$$

Now we control the first term. Note that the first term in frobenius norm is at most $\left|\frac{C_N}{\bar{C}_N} - 1\right| \cdot \|\bar{T}_i\|_F$. Apply Lemma 10 with index $j = 0$ and failure probability $\tau/3$ and error $d^{-4}\exp(-CC_{\mathrm{w}}^2)\epsilon$ for a sufficiently large constant $C > 0$. (Recall $H_0(x) \equiv 1$, so this is concentration of $\bar{C}_N/C_N$.) With probability at least $1-\tau/3$, $\left|\frac{\bar{C}_N}{C_N} - 1\right| \le d^{-4}\exp\big(-O(C_{\mathrm{w}}^2)\big)$. On this event, $\bar{C}_N \in [\frac{3}{4}C_N, \frac{5}{4}C_N]$, hence

$$\left|\frac{C_N}{\bar{C}_N} - 1\right| = \frac{\left|\frac{\bar{C}_N}{C_N} - 1\right|}{\frac{\bar{C}_N}{C_N}} \le \frac{\epsilon d^{-4}\exp\big(-CC_{\mathrm{w}}^2\big)}{1/2} = \epsilon d^{-4}\exp\big(-CC_{\mathrm{w}}^2\big). \tag{19}$$

Now it remains to bound $\|\bar{T}_i\|_F$. By the triangle inequality and (18),

$$\|\bar{T}_i\|_F \le \|T_i\|_F + \|\bar{T}_i - T_i\|_F \le \|T_i\|_F + \epsilon/2. \tag{20}$$

Moreover, by Jensen,

$$\|T_i\|_F^2 = \left\|\mathbb{E}_{x\sim D}\left[\frac{H_i(x)w(x)}{C_N}\right]\right\|_F^2 \le \mathbb{E}_{x\sim D}\left[\left\|\frac{H_i(x)w(x)}{C_N}\right\|_F^2\right].$$

The second-moment bound established in the proof of Lemma 10 gives

$$\mathbb{E}_{x\sim D}\left[\left\|\frac{H_i(x)w(x)}{C_N}\right\|_F^2\right] \lesssim d^4\exp\big(O(C_{\mathrm{w}}^2)\big), \tag{21}$$

and hence

$$\|T_i\|_F \lesssim d^2\exp\big(O(C_{\mathrm{w}}^2)\big). \tag{22}$$

Hence combining the above, since we set the constant C sufficiently large, we have that

$$\left|\frac{C_N}{\bar{C}_N} - 1\right| \cdot \|\bar{T}_i\|_F \le \epsilon/2\ .$$

Applying the union bound concludes the statement. ■

Lemma 9 (Subspace Identification) *Let $C_w > 1, \alpha, \epsilon, \tau, \in (0, 1/2)$, $\mu, z_1, \cdots z_n \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Assume that $\|\mu\|^2 < 0.1$, $\|\Sigma - (1 - C_w/\sqrt{d})^{-1}I\|_F \le 0.1$, $\|\Sigma\| = O(1)$, and $C_w \ll d^{1/2}$. Let D be a Gaussian mixture defined as $D := (1-\alpha)\mathcal{N}(\mu,\Sigma) + \frac{\alpha}{n}\sum_{i=1}^n \mathcal{N}(z_i,\Sigma)$, and $w(x)$ be a weight function of the form $w(x) = \exp\left(-\frac{C_w\|x\|^2}{d^{1/2}}\right)$. Define $C_N = \mathbb{E}_{x\sim D}[w(x)]$. Then Algorithm 1 given $m = d^4\epsilon^{-2}\exp(\Theta\left(C_w^2\right))\tau^{-1}$ i.i.d. samples from D, runs in time $\mathrm{poly}(m,d)$, and with probability at least $1-\tau$ outputs a subspace V of dimension at least $d/2$ satisfying $\frac{1}{C_N}\left|\mathbb{E}_{x\sim D}\left[h_j(\langle v, x\rangle)w(x)\right]\right| = O(C_w^{-1})$. for all $j \in [4]$ and unit vector $v \in V$.*

Proof We will show that Algorithm 1 is the desired algorithm. Fix $i \in [4]$. For simplicity, denote by $T_i := C_N^{-1}\mathbb{E}_{x\sim D}\left[H_i(x)w(x)\right]$ the normalized reweighted degree-i Hermite moment tensor, and denote by M_i its $(i-1,1)$-flattening. By (5) of Lemma 8, the $(i-1,1)$-flattening of T_i satisfies that

$$\|M_i\|_* = O(d/C_{\mathrm{w}}). \tag{23}$$

By Corollary 11, with probability $1-\tau$ the empirical moment tensor satisfies that

$$\left\|\widehat{T}_i - T_i\right\|_F \le \epsilon/\sqrt{d}, \tag{24}$$

with probability at least $1-\tau$. We will condition on Equation (23)in the rest of the analysis. In particular, by Cauchy's inequality, this implies that

$$\left\|\widehat{M}_i - M_i\right\|_* \le \epsilon. \tag{25}$$

Combining Equations (23) and (25) and the triangle inequality for the nuclear norm then gives that

$$\|\widehat{M}_i\|_* = \|\widehat{M}_i - M_i + M_i\|_* \le \|\widehat{M}_i - M_i\|_* + \|M_i\|_* = O(d/C_{\mathrm{w}}). \tag{26}$$

Now let $\sigma_1^{(i)}, \ldots, \sigma_d^{(i)}$ be the right singular values of $\widehat{M}_i$ and let $t > 0$. Define $r := |j : \sigma_j \geq t|$. Since all $\sigma_j \geq 0$, we have $\|\widehat{M}_i\|_* = \sum_j \sigma_j \geq rt$, hence $r \leq \|\widehat{M}_i\|_*/t = O(d/(C_{\mathrm{w}}t))$. If we set to be the cutoff threshold in Line 9,i.e., $t = CC_{\mathrm{w}}^{-1}$ for a sufficiently large constant $C > 0$ we have that $r \leq d/8$. Hence, the dimension of $\mathcal{I}$ is at most $d/2$, showing that V is a subspace of dimension $d/2$.

It then remains to show that the Hermite moments are small within the subspace identified. Denote by $\mathcal{I}$ the subspace constructed at line 9. By the definition of $\mathcal{I}$, we have that

$$\|\widehat{M}_i v\|_2 = O(C_{\mathrm{w}}^{-1}) \,,$$

for all unit vectors v over $\mathcal{I}^{\perp}$. Next note the flattening identity

$$\langle v^{\otimes i}, \widehat{T}_i \rangle = \langle v^{\otimes(i-1)}, \mathrm{flat}_{(i-1,1)}(\widehat{T}_i)\, v \rangle = \langle v^{\otimes(i-1)}, \widehat{M}_i v \rangle.$$

Hence, by Cauchy–Schwarz and $\|v\|_2 = 1$,

$$\left|\langle v^{\otimes i}, \widehat{T}_i \rangle\right| = \left|\langle v^{\otimes(i-1)}, \widehat{M}_i v \rangle\right| \leq \|v^{\otimes(i-1)}\|_2 \, \|\widehat{M}_i v\|_2 = \|\widehat{M}_i v\|_2 = O(C_{\mathrm{w}}^{-1}).$$

Finally,

$$\left|\langle v^{\otimes i}, T_i \rangle\right| \leq \left|\langle v^{\otimes i}, \widehat{T}_i \rangle\right| + \left|\langle v^{\otimes i}, T_i - \widehat{T}_i \rangle\right| \leq O(C_{\mathrm{w}}^{-1}) + \|T_i - \widehat{T}_i\|_F \cdot \|v^{\otimes i}\|_F \leq O(C_{\mathrm{w}}^{-1}) + \|T_i - \widehat{T}_i\|_F,$$

and under the concentration event, this is $O(C_{\mathrm{w}}^{-1})$. This concludes the proof of Lemma 9. ■

Appendix D. Omitted Proofs for Section 5

Lemma 13 *Let $\sigma \in (0,1)$, Z and X be real-valued random variables and Y be standard Gaussian random variable, where X is defined as the convolution $X = \sigma Y + Z$, and let $\bar{Z} = \frac{Z}{\sqrt{1-\sigma^2}}$. If for $i = 1, \cdots, k$, $|\mathbb{E}[h_i(X)]| = \Theta(\epsilon)$, then for $i = 1, \cdots, k$, $\left|\mathbb{E}\left[h_i\left(\bar{Z}\right)\right]\right| = \Theta\left(\frac{\epsilon}{(1-\sigma^2)^{i/2}}\right)$.*

Proof Indeed, $X = \sqrt{1-\sigma^2} \cdot \bar{Z} + \sigma \cdot Y = U_{\sqrt{1-\sigma^2}}\left(\bar{Z}\right)$. Thus, from Fact 19 we know that $\left|\mathbb{E}\left[h_i\left(\bar{Z}\right)\right]\right| = \left|\frac{1}{(1-\sigma^2)^{i/2}} \cdot \mathbb{E}[h_i(X)]\right| = \Theta\left(\frac{\epsilon}{(1-\sigma^2)^{i/2}}\right)$ as desired. ■

Lemma 14 *For $\delta \in \mathbb{R}_{\geq 0}$, let $\bar{Z}$ be a real-valued random variable, and $k \geq 2$ be an even integer. If for $i = 1, \cdots, k$, $\left|\mathbb{E}[h_i(\bar{Z})]\right| \leq \delta$, where h_i is the degree-i Hermite polynomial, then $\forall t \in \mathbb{R}$, $\Pr[\bar{Z} = t] < \frac{1+\delta\cdot\sqrt{k}\cdot 3^{k/2}}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)}$.*

Proof For an arbitrary point $t \in \mathbb{R}$, consider the following linear optimization problem over the cone of non-negative measures on $\mathbb{R}$ that maximizes the point mass at t satisfying normalization constraint and first k Hermite-moment-matching constraints up to slack δ:

$$\begin{aligned} \sup_{P \text{ is a non-negative measure on } \mathbb{R}} \quad & \Pr_{\bar{Z}\sim P}[\bar{Z} = t] \\ \text{subject to} \quad & \int_{\mathbb{R}} p_P(x)dx = 1 \\ & \left|\int_{\mathbb{R}} h_i(x) p_P(x) dx\right| \leq \delta \quad \text{for } i = 1, \cdots, k. \end{aligned} \tag{Primal}$$

Since there is no general closed form solution for measurable LP Equation (Primal), we derive its Lagrange dual by hand to provide a formulation that is more amenable to analyze. Formally, we have that

$$\begin{aligned} \inf_{c_i\in\mathbb{R}, i=0,\cdots,k} \quad & c_0 + \delta\cdot\sum_{i=1}^{k}|c_i| \\ \text{subject to} \quad & \sum_{i=0}^{k} c_i h_i(t) \geq 1, \\ & \sum_{i=0}^{k} c_i h_i(x) \geq 0, \quad \forall x\in\mathbb{R}\backslash\{t\}. \end{aligned} \tag{Dual}$$

By Lemma 23, the objective of Dual forms an upper bound on the objective of Primal. While in Primal the goal is to maximize the expected value of the indicator function $\mathbf{1}_{\{t\}}(x)$ over μ satisfying the Hermite moment constraints, the Dual seek to a degree k polynomial $f(x) := \sum_{i=0}^{k} c_i h_i(x)$ as an upper bound on $\mathbf{1}_{\{t\}}(x)$ and minimizes the worst case expected value $\int_{\mathbb{R}} f(x)d\mu(x)$ under all μ satisfying the Hermite moment constraints. This insight suggests that we upper bound the objective of Dual by an equation in $f(x)$ as a unity as oppose to dealing with c_i's separately. On one hand, since all but the 0^{th} Hermite moment of standard Gaussian vanishes, and the fact that f is nonnegative, we can replace c_0 in the objective by $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[f(X)] = \mathbb{E}_{X\sim\mathcal{N}(0,1)}[|f(X)|] = \|f\|_1$. On the other hand, to bound from above $\sum_{i=1}^{k}|c_i|$, we have that

$$\sum_{i=1}^{k}|c_i| \leq \sqrt{k}\cdot\sqrt{\sum_{i=0}^{k}c_i^2} = \sqrt{k}\cdot\sqrt{\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\left(\sum_{i=0}^{k}c_i h_i(X)\right)^2\right]} = \sqrt{k}\cdot\|f\|_2,$$

where the first inequality follows from Cauchy Schwartz, the first equality follows from the fact that the normalized probabilists' Hermite polynomials are orthonormal with respect to standard Gaussian measure. Thus, we can formulate a new LP that has objective value of any feasible solution upper bounding the optimal objective value of Dual as follows:

$$\begin{aligned} \inf_{f\in\mathbb{R}_k[x]} \quad & \|f\|_1 + \delta\sqrt{k}\cdot\|f\|_2 \\ \text{subject to} \quad & f(t)\geq 1, \\ & f(x)\geq 0, \qquad \forall x\in\mathbb{R}\backslash\{t\}. \end{aligned} \tag{UB}$$

We focus on finding the optimal solution s that minimizes $\|f\|_1$ subject to the above constraints, and using a similar argument we can show that s also has minimal l_2 norm over the feasible region, whence is the optimal solution of UB. Indeed, consider the following LPs:

$$\begin{aligned} \inf_{f\in\mathbb{R}_k[x]} & \|f\|_1 \\ \text{s.t.} \quad & f(t)\geq 1, \qquad \text{(LP 1)} \\ & f(x)\geq 0, \forall x\in\mathbb{R}\backslash\{t\}. \end{aligned} \qquad \begin{aligned} \inf_{f\in\mathbb{R}_k[x]} & \|f\|_1 \\ \text{s.t.} \quad & f(t)=1, \qquad \text{(LP 2)} \\ & f(x)\geq 0, \forall x\in\mathbb{R}\backslash\{t\}. \end{aligned} \qquad \begin{aligned} \sup_{f\in\mathbb{R}_k[x]} & \frac{f(t)}{\|f\|_1} \\ \text{s.t.} \quad & f(t)=1, \qquad \text{(LP 3)} \\ & f(x)\geq 0, \forall x\in\mathbb{R}\backslash\{t\}. \end{aligned}$$

$$\sup_{f\in\mathbb{R}_k[x]} f(t) \quad \text{s.t.} \quad \|f\|_1 = 1, \quad f(x)\geq 0, \forall x\in\mathbb{R}\backslash\{t\}. \qquad \text{(LP 4)}$$

$$\sup_{g\in\mathbb{R}_{k/2}[x]} (g(t))^2 \quad \text{s.t.} \quad \|g\|_2 = 1. \qquad \text{(LP 5)}$$

$$\sup_{c_i\in\mathbb{R}, i=0,\cdots,k/2} \left|\sum_{i=0}^{k/2} c_i h_i(t)\right| \quad \text{s.t.} \quad \sum_{i=0}^{k/2} c_i^2 = 1. \qquad \text{(LP 6)}$$

Our goal is to show that the optimal solution of LP 6 is a scalar multiple of the optimal solution of LP 1 by showing the pairwise radial equivalence of the above linear programs. And since LP 6 is the problem of finding the direction $(c_0,\cdots,c_{k/2})$ that maximizes the projection of the vector $(h_0(t),\cdots,h_{k/2}(t))$ onto it, we immediately know from basic linear algebra that the optimal solution of LP 6 is the unit vector in the same direction as $(h_0(t),\cdots,h_{k/2}(t))$, i.e. $\frac{1}{\left(\sum_{i=0}^{k/2}(h_i(t))^2\right)^{1/2}}(h_0(t),\cdots,h_{k/2}(t))$ whence the optimal solution of LP 1 is $s(x) := \frac{\left(\sum_{i=0}^{k/2} h_i(x)h_i(t)\right)^2}{\left(\sum_{i=0}^{k/2}(h_i(t))^2\right)^2}$ after rescaling according to the constraints in LP 1. Formally,

Claim 30 *LP 6 is radially equivalent to LP 1.*

Proof We start by showing that LP 1 and LP 2 are equivalent, and it suffices to show that for LP 1, $f(t)\geq 1$ is an active constraint. Indeed, for any g in the feasible region with $g(t)>1$, we can obtain a better solution $g' = \frac{g}{g(t)}$ in the feasible region with smaller l_1 norm $\|g'\|_1 < \|g\|_1$. Thus, the optimal solution f^* of LP 1 must have that $f^*(t)=1$, as desired.

The radial equivalence of LP 2 and LP 4 follows immediately from applying Lemma 24 to positively homogeneous degree-1 functionals $F: f\mapsto \|f\|_1$ and $G: f\mapsto f(t)$ over the cone of nonnegative degree $\leq k$ real polynomials.

Next, we show that LP 2 is equivalent to LP 3, and then we show that we may assume that the optimal solution of LP 3 is of the form $f(x)=(g(x))^2$, whence so does LP 2. Combining these facts, it immediately follows from the argument in the last paragraph that optimal solution of LP 4 is of the form $f(x)=(g(x))^2$, whence LP 4 is equivalent to LP 5. Indeed, we first show that LP 2 is equivalent to LP 3. This immediately follows from the fact that $f\neq 0$ because $f(t)=1$ whence $\|f\|_1>0$. Next, we proceed to show that we may assume that the optimal solution of LP 3 is of the form $f(x)=(g(x))^2$. Indeed, since any univariate nonnegative polynomial can be decomposed into sum of squares, let f^* be an optimal solution of LP 3, then we can write $f^* := \sum_{i=0}^{k/2} f_i^2$ where $f_i\in\mathbb{R}_i[x]$. Then, we have that

$$\frac{f^*(t)}{\|f^*\|_1} = \frac{\sum_{i=0}^{k/2} f_i(t)^2}{\sum_{i=0}^{k/2}\|f_i^2\|_1} \leq \max_{i=1,\cdots,k/2}\frac{f_i(t)^2}{\|f_i^2\|_1},$$

where the first equality follows from linearity of expectation and the inequality follows from the fact that $\frac{\sum_{i=1}^k a_i}{\sum_{i=1}^k b_i}\leq \max\frac{a_i}{b_i}$ for nonnegative values $a_1,\cdots,a_k,b_1,\cdots,b_k$. Thus, it must be the case that $f^2_{\text{argmax}_{i=1,\cdots,k/2}\frac{f_i(t)^2}{\|f_i^2\|_1}}$ is a solution that is no worse than f^*. Hence, we may safely assume that the optimal solution of LP 3 is of the form $f(x)=(g(x))^2$ for $g\in\mathbb{R}_{k/2}[x]$, as desired.

The equivalence of LP 5 and LP 6 simply follows from using Hermite analysis to decompose g, which concludes the proof. ■

We next proceed to compute the objective value of $s(x) := \frac{\left(\sum_{i=0}^{k/2} h_i(x)h_i(t)\right)^2}{\left(\sum_{i=0}^{k/2}(h_i(t))^2\right)^2}$ in UB. Indeed, for the first term $\|s\|_1 = \mathbb{E}_{X\sim\mathcal{N}(0,1)}[s(X)]$, we have that

$$\begin{aligned}
\mathbb{E}_{X\sim\mathcal{N}(0,1)}[s(X)] &= \frac{1}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)^2}\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\left(\sum_{i=0}^{k/2} h_i(t)h_i(X)\right)^2\right] \\
&= \frac{1}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)^2}\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\sum_{i=0}^{k/2}\sum_{j=0}^{k/2} h_i(t)h_j(t)h_i(X)h_j(X)\right] \qquad (27)\\
&= \frac{\sum_{i=0}^{k/2} h_i(t)^2}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)^2} = \frac{1}{\sum_{i=0}^{k/2} h_i(t)^2}, \qquad (28)
\end{aligned}$$

where the third equality follows from orthonormality of $h_i(x)$. On the other hand,

$$\begin{aligned}
\|s\|_2 &= \sqrt{\frac{1}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)^4}\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\left(\sum_{i=0}^{k/2} h_i(t)h_i(X)\right)^4\right]} = \frac{1}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)^2}\left\|\left(\sum_{i=0}^{k/2} h_i(t)\right)h_i\right\|_4^2 \\
&\le \frac{3^{k/2}}{\left(\sum_{i=0}^{k/2} h_i(t)^2\right)^2}\left\|\left(\sum_{i=0}^{k/2} h_i(t)\right)h_i\right\|_2^2 = \frac{3^{k/2}}{\sum_{i=0}^{k/2} h_i(t)^2}, \qquad (29)
\end{aligned}$$

where the first inequality follows from Fact 20, and the last equality follows from the same computation as in Equation (27). Denote by $J(\cdot)$ the objective function of UB. Combining Equation (28) and Equation (29) we have that $J(s(x)) \le \frac{1+\delta\cdot\sqrt{k}\cdot 3^{k/2}}{\sum_{i=0}^{k/2} h_i(t)^2} \le \frac{1+\delta\cdot\sqrt{k}\cdot 3^{k/2}}{\min_{t\in\mathbb{R}}\left(\sum_{i=0}^{k/2} h_i(t)^2\right)}$, which concludes the proof.

■

Equipped with Lemma 14, we are now ready to show the main result in this section.

Proposition 12 (Moment Matching implies no heavy point) *Let $\sigma > 0$, $c > 0$ be a small constant, Y be standard Gaussian random variable and Z be another random variable over $\mathbb{R}$ independent of Y. Define the random variable X as the convolution $X := Z + \sigma Y$. If $\Pr[Z = \mu] > 2/3 + c/2$, and for $i = 1, 2, 3, 4$, $|\mathbb{E}[h_i(X)]| \le \epsilon$, then $\mu = O(\epsilon^{1/4})$.*

Proof We first show that we have an upper bound on σ. Denote by $\tau_i > 0$ the moment-matching slack, i.e. for $i = 1, 2, 3, 4$ we have that $|\mathbb{E}[h_i(X)]| = \tau_i$ for $\tau_i \le \epsilon$. By law of total variance, $\mathrm{Var}(X) \ge \sigma^2$. Moreover, $\mathbb{E}[h_2(X)] = (\mathbb{E}[X^2] - 1)/\sqrt{2} = (\mathrm{Var}(X) + \mathbb{E}[X]^2 - 1)/\sqrt{2}$, which implies that $\mathrm{Var}(X) = 1 \pm \sqrt{2}\tau_2 - \tau_1^2$. Hence, we obtain the upper bound $\sigma^2 \le 1 + \sqrt{2}\tau_2$, and it suffices for us to consider σ^2 within $(0, 1 + \sqrt{2}\tau_2)$.

Towards showing the desired upper bound on μ, at a high level, we first consider the case (Claim 31) when $\sigma^2 \in \left(1 - 2\sqrt{6C/c}\,\epsilon^{1/2}, 1 + \sqrt{2}\tau_2\right)$ for a universal constant $C > 0$ to be determined later in the proof, we obtain the desired upper bound on μ. Then, we show Claim 32 that states that it must be the case that σ^2 falls within this interval $\left(1 - 2\sqrt{6C/c}\,\epsilon^{1/2}, 1 + \sqrt{2}\tau_2\right)$ that is close to 1 by the moment matching constraints. Combining the two claims concludes the proof. It remains to formalize the proof of the claims.

Claim 31 *If $\sigma^2 \in \left(1 - 2\sqrt{6C/c}\,\epsilon^{1/2}, 1 + \sqrt{2}\tau_2\right)$, we have that $\mu = O(\epsilon^{1/4})$.*

Proof Recall that by matching moments $\mathbb{E}[h_2(X)] = \pm\tau_2$ which implies that $\mathbb{E}[X^2] = 1 \pm \sqrt{2}\tau_2$. We proceed to show that $\mathbb{E}[X^2] \geq \sigma^2 + \frac{2}{3}\mu^2$. Indeed, write $X = \sigma Y + Z$ where $Y \sim \mathcal{N}(0,1)$ is independent of Z. Then

$$\mathbb{E}[X^2] = \mathbb{E}[(\sigma Y + Z)^2] = \sigma^2\mathbb{E}[Y^2] + 2\sigma\,\mathbb{E}[YZ] + \mathbb{E}[Z^2].$$

By independence and $\mathbb{E}[Y] = 0$, we have $\mathbb{E}[YZ] = \mathbb{E}[Y]\mathbb{E}[Z] = 0$, and $\mathbb{E}[Y^2] = 1$, whence $\mathbb{E}[X^2] = \sigma^2 + \mathbb{E}[Z^2] \geq \sigma^2 + \frac{2}{3}\mu^2$, which implies that $\mu^2 \leq \frac{3}{2}(1 + \sqrt{2}\tau_2 - \sigma^2) = O\left(\epsilon + \epsilon^{1/2}\right)$, and thus $\mu = O(\epsilon^{1/4})$ which concludes the proof of this Claim. ■

Claim 32 *It holds that $\sigma^2 \in \left(1 - 2\sqrt{6C/c}\,\epsilon^{1/2}, 1 + \sqrt{2}\tau_2\right)$.*

Proof Assume for contradiction that $\sigma^2 \in \left(0, 1 - 2\sqrt{6C/c}\,\epsilon^{1/2}\right]$. By applying Lemma 13 with $k = 4$, we obtain that for each $i \in \{1,2,3,4\}$,

$$\left|\mathbb{E}[h_i(\bar{Z})]\right| \leq C\,\frac{\epsilon}{(1-\sigma^2)^{i/2}}. \tag{30}$$

For simplicity, denote by

$$\bar{\delta} \;\coloneqq\; \max_{1\leq i\leq 4} C\,\frac{\epsilon}{(1-\sigma^2)^{i/2}} \;=\; C\,\frac{\epsilon}{(1-\sigma^2)^2}.$$

Since $\sigma^2 < 1 - 2\sqrt{6C/c}\,\epsilon^{1/2}$ implies $1 - \sigma^2 > 2\sqrt{6C/c}\,\epsilon^{1/2}$, we have that

$$\bar{\delta} \;\leq\; \frac{c}{24}\,.$$

On the other hand, since Z has point mass $> 2/3 + c/2$ at μ, $\bar{Z} = Z/\sqrt{1-\sigma^2}$ has point mass $> 2/3 + c/2$ at $\mu/\sqrt{1-\sigma^2}$, i.e.

$$\Pr\left[\bar{Z} = \frac{\mu}{\sqrt{1-\sigma^2}}\right] > \frac{2}{3} + c/2.$$

Applying Lemma 14 with $k = 4$ and slack $\bar{\delta}$ gives that for every $t \in \mathbb{R}$,

$$\Pr[\bar{Z} = t] \;\leq\; \frac{1 + \bar{\delta}\,2\,3^2}{\frac{3}{2} + \frac{t^4}{2}} \;\leq\; \frac{1 + \bar{\delta}\,2\,3^2}{3/2} \;=\; \frac{2}{3} + 12\,\bar{\delta} \leq \frac{2}{3} + \frac{c}{2},$$

which contradicts the assumption that $\Pr[\bar{Z} = \mu/\sqrt{1-\sigma^2}] > 2/3 + c/2$. Thus, we must have that

$$\sigma^2 \geq 1 - 2\sqrt{6C/c}\,\epsilon^{1/2},$$

which concludes the proof. ■

This concludes the proof of Proposition 12. ■

Appendix E. Omitted Proofs for Section 6

Lemma 15 (Coarse Normalization) *Let $d \in \mathbb{Z}_+$ and C_w the parameter defined in Algorithm 2. Assume that $C_w/\sqrt{d} < 0.1$. With probability at least 0.99, it holds that the whitened distribution $\bar{D}$ (line 7 of Algorithm 2) is an α-mean-shift corrupted distribution with clean mean $\bar{\mu}$ and positive definite shared covariance $\bar{\Sigma}$ satisfying that $\|\bar{\mu}\|_2 = O(1)$, $\|\bar{\Sigma} - (1 - C_w/\sqrt{d})^{-1} I\|_F = O(1)$.*

Proof By the guarantee of Lemma 6, the robust estimators $\widehat{\mu}_0$ and $\widehat{\Sigma}_0$ returned in line 6 satisfy

$$\left\|\widehat{\Sigma}_0^{-1/2}(\mu - \widehat{\mu}_0)\right\|_2 = O\left(\alpha \log \frac{1}{\alpha}\right) = O(1), \tag{31}$$

$$\left\|\widehat{\Sigma}_0^{-1/2}\,\Sigma\,\widehat{\Sigma}_0^{-1/2} - I\right\|_F = O(1). \tag{32}$$

Then the whitened distribution $\bar{D}$ defined as in line 7 must be an α-mean-shift corrupted distribution with clean mean $\bar{\mu}$ and shared covariance $\bar{\Sigma}$ defined as follows:

$$\bar{\mu} = \left(1 - C_{\mathrm{w}}/\sqrt{d}\right)^{-1/2} \widehat{\Sigma}_0^{-1/2}(\mu - \widehat{\mu}_0) \tag{33}$$

$$\bar{\Sigma} = \left(1 - C_{\mathrm{w}}/\sqrt{d}\right)^{-1} \widehat{\Sigma}_0^{-1/2}\Sigma\widehat{\Sigma}_0^{-1/2} \tag{34}$$

From the assumption it follows that $1 - C_{\mathrm{w}}/\sqrt{d} = \Theta(1)$. Combining Equations (31) and (33) then gives that $\|\bar{\mu}\| = O(1)$. Combining Equations (32) and (34) then gives that $\|\bar{\Sigma} - (1 - C_{\mathrm{w}}/\sqrt{d})^{-1} I\|_F = (1 - C_{\mathrm{w}}/\sqrt{d})^{-1}\|\widehat{\Sigma}_0^{-1/2}\Sigma\widehat{\Sigma}_0^{-1/2} - I\|_F = O(1)$. Finally, note that $\bar{\Sigma}$ is positive definite. This is because $(1 - C_{\mathrm{w}}/\sqrt{d}) > 0$ and $\Sigma_0^{-1/2}\Sigma\widehat{\Sigma}_0^{-1/2}$ is a conjugation of Σ. Since $\Sigma \succ 0$ and $\widehat{\Sigma}_0$ is invertible by Lemma 6, it follows that $\Sigma_0^{-1/2}\Sigma\widehat{\Sigma}_0^{-1/2} \succ 0$. Indeed, for any nonzero x,

$$x^\top \Sigma_0^{-1/2}\Sigma\widehat{\Sigma}_0^{-1/2} x = \left(\Sigma_0^{-1/2}x\right)^\top \Sigma \left(\widehat{\Sigma}_0^{-1/2}x\right) > 0,$$

and therefore $\bar{\Sigma} \succ 0$. ■