Fast Mean Estimation with Sub-Gaussian Rates

Yeshwanth Cherapanamjeri   YESHWANTH@BERKELEY.EDU
UC Berkeley

Nicolas Flammarion   FLAMMARION@BERKELEY.EDU
UC Berkeley

Peter L. Bartlett   PETER@BERKELEY.EDU
UC Berkeley

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We propose an estimator for the mean of a random vector in $\mathbb{R}^d$ that can be computed in time $O(n^{3.5} + n^2 d)$ for n i.i.d. samples and that has error bounds matching the sub-Gaussian case. The only assumptions we make about the data distribution are that it has finite mean and covariance; in particular, we make no assumptions about higher-order moments. Like the polynomial time estimator introduced by Hopkins (2018), which is based on the sum-of-squares hierarchy, our estimator achieves optimal statistical efficiency in this challenging setting, but it has a significantly faster runtime and a simpler analysis.

1. Introduction

Estimating the mean of a population given a finite sample is arguably the most fundamental statistical estimation problem. Despite the broad applicability and the fundamental nature of this problem, an estimator achieving the optimal statistical rate has only been discovered recently. However the optimal computational complexity of such an estimator is not well-understood.

In this paper, we are interested in obtaining high confidence estimates of the mean in the simple setting where only the existence of the covariance of the distribution is assumed. That is, we would like to find the smallest r_δ such that given samples $X_1, \ldots, X_n$ from a distribution $\mathcal{D}$ with mean μ our estimator $\hat{X}$ satisfies:

$$\mathbb{P}\left\{ \|\hat{X} - \mu\| \geq r_\delta \right\} \leq \delta.$$

To understand the inherent statistical limit of this problem, let us consider the simplified setting where the covariance is the identity. The most natural estimator for the mean of the population is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. From the Central Limit Theorem, the distribution of $\bar{X}$ satisfies $\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} \mathcal{N}(0, I)$, and assuming this conclusion holds for any n allows an r_δ satisfying

$$r_\delta = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}} \right).$$

Catoni (2012) shows that this r_δ is the optimal statistical performance achievable under such mild assumptions. However, the above confidence interval only holds true asymptotically when the number of samples goes to infinity or when the distribution is sub-Gaussian. For finite sample results

with a heavy-tailed distribution, applying Chebyshev's inequality to the empirical mean gives only

$$r_\delta = \Omega\left(\sqrt{\frac{d}{n\delta}}\right).$$

The above bound is weaker than the one obtained by the Central Limit Theorem in two ways, the dependence on the failure probability δ is polynomial in $1/\delta$ instead of logarithmic and the term depending on δ is multiplied by the dimensionality d as opposed to being part of a smaller additive term. Unfortunately, Catoni (2012) also shows the above result is tight. That is, for any n, δ, there exists a distribution $\mathcal{D}_{n,\delta}$ for which the bound guaranteed by Chebyshev's inequality is optimal.

The poor performance of the empirical mean is due to its sensitivity to large outliers that occur naturally as part of the sample. The median-of-means framework was devised as a means of circumventing such difficulties. It was independently developed by Nemirovsky and Yudin (1983); Jerrum et al. (1986); Alon et al. (1999) and was later analyzed by Hsu and Sabato (2016); Lerasle and Oliveira (2011); Minsker (2015). As part of this framework, the samples are first divided into k batches and the mean of the samples is computed within each batch to obtain k estimates $Z_1, \ldots, Z_k$. Each of these has mean μ and variance $\frac{k}{n}I$. The empirical mean is simply the mean of these k estimates, which is sensitive to outliers. The median-of-means estimator instead is the geometric median of the k estimates, which has greater tolerance to outliers. The success of the median-of-means estimator is due to the fact that it relies on only a fraction of estimates Z_i being close to the mean as opposed to all the estimates being close. Minsker (2015) shows this gives an improved value of r_δ as follows:

$$r_\delta = O\left(\sqrt{\frac{d\log 1/\delta}{n}}\right).$$

The confidence interval guaranteed by the median-of-means estimator is better than the one for the empirical mean by improving the dependence on $1/\delta$, but it is still poorer than we might expect from the Central Limit Theorem. Subsequent work attempting to bridge this gap achieves better rates than those guaranteed by the median-of-means but with stronger assumptions on the data generating distribution[1] (Joly et al., 2017). The question of whether it was statistically feasible to obtain confidence intervals of the form guaranteed by the Central Limit Theorem was finally resolved by Lugosi and Mendelson (2019). They devised an improved estimator, based on the median-of-means framework, called the median-of-means tournament, which achieves CLT-like confidence intervals. While the median-of-means estimator relies on the concentration of the number of Z_i close to the mean in Euclidean norm, the median-of-means tournament relies on the fact that along every direction v, the number of Z_i close to the projection of the mean concentrates. The freedom to choose a different set of Z_i for each direction allow one to obtain a much smaller confidence interval than the one for the median-of-means estimator. In subsequent work, following the PAC-Bayesian approach of Catoni (2012), Catoni and Giulini (2017a) proposed a soft-truncation based estimator which obtains CLT-like confidence intervals provided one has access to estimates of the trace and spectral norm of the covariance matrix.

However, it is not known whether the estimators from Lugosi and Mendelson (2019); Catoni and Giulini (2017a) are computationally feasible, as there are no known polynomial time algorithms to

---

1. A rate of $O\left(\sqrt{d/n} + \sqrt{\log(\frac{\log d}{\delta})/n}\right)$ is achieved under a fourth moment assumption on the distribution.

compute them. In contrast, the median-of-means and empirical mean can be computed in nearly-linear time (Cohen et al., 2016). To alleviate this computational intractability, Catoni and Giulini (2017b) proposed an efficient polynomial time estimator which achieves optimal statistical performance up to second order terms, assuming the existence of higher order moments. The question of computational tractability was subsequently resolved by Hopkins (2018), who showed that an algorithm based on a sum-of-squares relaxation of the median-of-means tournament estimator achieves the statistically optimal CLT-like confidence intervals. However, the runtime of this algorithm is exorbitantly large[2] ($O(nd + (\log 1/\delta)^{28} + d^{28})$). In the interesting regime where $\delta = e^{-d}$, this reduces to a runtime of $O(nd + d^{28})$ and in the worst case setting where $\delta = e^{-\Omega(n)}$, the runtime reduces to $O(nd + n^{28} + d^{28})$.

In this paper, we propose a novel algorithm which achieves a drastically reduced runtime—$\widetilde{O}((\log 1/\delta)^{3.5} + (\log 1/\delta)^2 d + nd)$—and a significantly simpler analysis. Our algorithm is a descent-based method that iteratively improves an estimate of the mean. The main challenge of such an approach is to estimate the descent direction. To this end, we crucially leverage the structure of the solutions to semidefinite programming relaxations of polynomial optimization problems designed to test whether an estimate is close to the mean. Our main contributions are twofold; we first show how exact solutions to the polynomial optimization problem furnish suitable descent directions and that such descent directions can also be efficiently extracted from relaxations of such problems and secondly, we show that these descent directions can be used in a descent style algorithm for mean estimation. Our paper is organized as follows: in Section 2, we present our main result, then in Section 3, as a warm-up, we devise a descent style algorithm for the case where we are given exact solutions to the polynomial optimization problems mentioned previously and prove that this algorithm achieves optimal statistical efficiency. This sets the stage for Section 4, where we present our main algorithm based on semidefinite relaxations of the previously defined polynomial optimization problems, leading to computationally efficient sub-Gaussian mean estimation.

2. Main result

Formally, our main result[3] is as follows:

Theorem 1 *Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{n \times d}$ be n i.i.d. random vectors with mean μ and covariance Σ. Then Algorithm 1 instantiated with Algorithms 4 and 5 and run with inputs $\boldsymbol{X}$, target confidence δ, stepsize $\gamma = 1/20$ and number of iterations $T = 1000 \log \|\mu\| / \epsilon$ returns a vector x^* satisfying:*

$$\|x^* - \mu\| \leq \max\left(\epsilon, 480000 \left(\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{\|\Sigma\| \log 1/\delta}{n}}\right)\right),$$

with probability at least $1 - \delta$.

We can make the following comments:

- Our estimator is both statistically optimal and computationally efficient. It achieves sub-Gaussian performance under minimal conditions on the distribution, and its runtime is $O((\log 1/\delta)^{3.5} + (\log 1/\delta)^2 d + nd)$. See Section 4.2 for details.

---

2. Assuming standard runtimes of the Interior Point method for semidefinite programming (Alizadeh, 1995)
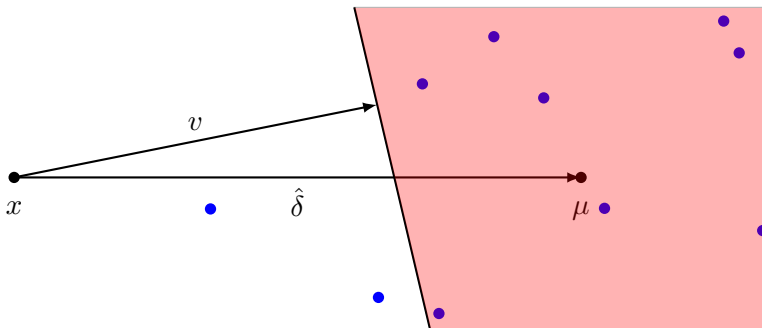3. The constants are explicit but we believe sub-optimal.

Figure 1: The direction v solution to MTE is well aligned with the vector joining the current estimate x to the true mean μ.

- In the regime where $\delta = e^{-d}$, we achieve a runtime of $\widetilde{O}(d^{3.5} + nd)$. In the worst case regime where $\delta = e^{-\Omega(n)}$, we obtain $\widetilde{O}(n^{3.5} + n^2 d)$. This is a significant improvement over the algorithm in Hopkins (2018) which obtains $O(nd + d^{28})$ and $O(nd + n^{28} + d^{28})$ respectively.

- The dependence of the number of iterations, T, on $\|\mu\|$ can be avoided by initializing the algorithm with the median-of-means estimate. In this case, we can instead use $T = 1000 \log d$ and obtain the same guarantees, avoiding any dependence on the knowledge of $\|\mu\|, \mathrm{Tr}(\Sigma), \|\Sigma\|$.

- The estimator depends on the confidence level δ. Devroye et al. (2016) propose an estimator which works for a whole range of δ but for a restricted class of distributions.

- Our result does not explicitly depend on the dimension d and our algorithm can be extended to a Hilbert space by working within the finite dimensional subspace containing the data points.

3. Warm-up

We present in this section a simple descent based algorithm. This algorithm is computationally inefficient but achieves the same guarantees of Theorem 1 with a much simpler analysis which nevertheless illustrates the main ideas behind the algorithm and proof of Theorem 1.

3.1. Intuition

We provide some intuition for our procedure, which iteratively improves an estimate of the mean. We first consider the simpler problem of testing whether a given point is close to the mean. We draw our inspiration from the main technical insight of Lugosi and Mendelson (2019), who show that along any direction, most of the bucket means, Z_i, are close to the mean, μ. Thus, to test whether a point, x, is far from the mean, it is sufficient to check whether there exists a direction where most of the Z_i are far away from x along that direction. This is formally expressed in the following polynomial optimization problem:

$$\max \sum_{i=1}^{k} b_i$$
$$b_i^2 = b_i$$
$$\|v\|^2 = 1$$
$$b_i\langle v, Z_i - x\rangle \geq b_i^2 r \quad \forall i \in [k] \qquad \textbf{(MTE)}$$

This polynomial problem over the set of variables $b_1, \ldots, b_k$ and $v_1, \ldots, v_d$ is parameterized by $r > 0$, the current estimate $x \in \mathbb{R}^d$ and the bucket means $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$. Its polynomial constraints are encoding the number of Z_i beyond a distance r from x when projected along a direction v. Intuitively, this program tries to find a direction v so as to maximize the number of Z_i beyond a distance r from x along that direction. Here, we know from (Lugosi and Mendelson, 2019) that for an appropriate choice of r, along all directions v, a large fraction of the Z_i are close to the mean. Formally, for all directions v, $|\{i : |\langle Z_i - \mu, v\rangle| \leq r\}| \geq 0.9k$ (see Corollary 5). Therefore this optimization problem has a large value when x is far from the mean and can be used to certify this.

Strikingly, the direction v returned by the solution of the above problem also contains information about the location of the mean when r is chosen appropriately, which enables improvement of the quality of the current estimate. As illustrated in Figure 1, the direction returned by this optimization problem is strongly correlated with the vector joining the current point x to the mean μ.

Algorithm 1 Mean Estimation

1: **Input**: Data Points $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, Target Confidence δ, Number of Iterations T, Stepsize γ
2: $k \leftarrow 3200 \log 1/\delta$
3: Split data points into k bins with bin $\mathcal{B}_i$ consisting of the points $X_{(i-1)\frac{n}{k}+1}, \ldots, X_{i\frac{n}{k}}$
4: $Z_i \leftarrow \text{Mean}(\mathcal{B}_i) \; \forall \, i \in [k]$ and $\boldsymbol{Z} \leftarrow (Z_1, \ldots, Z_k)$
5: $x^*, x_0 \leftarrow \boldsymbol{0}$ and $d^*, d_0 \leftarrow \infty$
6: **for** $t = 0 : T$ **do**
7: $\quad d_t \leftarrow \text{Distance Estimation}(\boldsymbol{Z}, x_t)$
8: $\quad g_t \leftarrow \text{Gradient Estimation}(\boldsymbol{Z}, x_t)$
9: $\quad$ **if** $d_t < d^*$ **then**
10: $\quad\quad x^* \leftarrow x_t$
11: $\quad\quad d^* \leftarrow d_t$
12: $\quad$ **end if**
13: $\quad x_{t+1} \leftarrow x_t + \gamma d_t g_t$
14: **end for**
15: **Return:** x^*

Algorithm 2 Distance Estimation

1: **Input**: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Current point x
2: $d^* = \max\{r > 0 : MTE(x, r, \boldsymbol{Z}) \geq 0.9k\}$
3: **Return:** d^*

Algorithm 3 Gradient Estimation

1: **Input**: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Current point x
2: $d^* = \text{Distance Estimation}(\boldsymbol{Z}, x)$
3: $(b, g) = MTE(x, d^*, \boldsymbol{Z})$
4: **Return:** g

Therefore, moving a small distance along the vector v should intuitively take us closer to the mean.

Given solutions to the polynomial optimization problem **MTE**, we may iteratively improve our estimate until no further change is necessary.

3.2. Algorithm

In this section we put the intuition provided previously into practice and propose a procedure that estimates the mean in the ideal situation where **MTE** can be exactly solved (the method is formally described in Algorithm 1):

1. First, following the median of means framework, the samples X_i are divided into k buckets and the mean of the samples within each bucket is computed as $Z_i = \frac{k}{n} \sum_{j=(i-1)n/k}^{in/k} X_j$.

2. Second, the estimate of the mean is iteratively updated using a descent approach, based on the solution of **MTE**. As mentioned in Section 3.1, we need to run **MTE** with an appropriate choice of r for the solution v to be correlated with the direction $x - \mu$. In the Distance Estimation step of our algorithm, we estimate a suitable choice of r (see Algorithm 2). This value of r is subsequently used in the Gradient Estimation step, to obtain an appropriate descent direction g (see Algorithm 3).

From this point on, we refer to the solution of polynomial equations **MTE** as $(b, v) = MTE(x, r, \boldsymbol{Z})$.

3.3. Analysis warm-up

In this simplified setting, we provide an analysis of our method and show that it obtains the same guarantees as those presented in Theorem 1. This is formally expressed in the following theorem for Algorithm 1 instantiated with Algorithms 2 and 3.

Theorem 2 *Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{n \times d}$ be n i.i.d. random vectors with mean μ and covariance Σ. Then Algorithm 1 instantiated with Algorithms 2 and 3 and run with inputs $\boldsymbol{X}$, target confidence δ, stepsize $\gamma = 1/4$ and number of iterations $T = 50 \log \|\mu\|/\epsilon$ returns a vector x^* satisfying:*

$$\|x^* - \mu\| \leq \max\left(\epsilon, 108000\left(\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{\|\Sigma\| \log 1/\delta}{n}}\right)\right),$$

with probability at least $1 - \delta$.

The main steps involved in the proof are the following:

1. **Distance Estimation:** We show that the Distance Estimation step in Algorithm 2 provides an accurate estimate of the distance of the current point from the mean. See Lemma 3.

2. **Gradient Estimation:** Next, we show that when x is far away from the mean μ, the vector g obtained by solving **MTE** in Algorithm 3 is well aligned with the vector joining the current point x to the mean μ. See Lemma 4.

3. **Gradient Descent:** Combining the previous two steps, we prove that we eventually converge to a good approximation to the mean.

In the proofs of our lemmas relating to the correctness of the Distance Estimation and the Gradient Estimation steps, we make use of the following assumption:

Assumption 1 *For the bucket means,* $\mathbf{Z} = (Z_1, \ldots, Z_k)$, *we have:*

$$\forall v \in \mathbb{R}^d, \|v\| = 1 \;\Rightarrow\; \left| \{ i : \langle Z_i - \mu, v \rangle \geq 300 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right) \} \right| \leq 0.05k$$

The assumption is a formalization of the insight of (Lugosi and Mendelson, 2019), which shows that along all directions, v, most of the bucket means are within a small radius of the true mean, μ, with high probability[4].

First, we prove that the **Distance Estimation** step defined in Algorithm 2 is correct.

Lemma 3 *Under Assumption 1, for all* $t \in \{0, \ldots, T\}$ *in the running of Algorithm 1, d_t satisfies:*

$$\left| d_t - \|x_t - \mu\| \right| \leq 300 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{\|\Sigma\|k/n} \right).$$

Proof Let $r^* = 300 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{\|\Sigma\|k/n} \right)$. We first prove the lower bound $\|x_t - \mu\| - r^* \leq d_t$. We may assume that $\|x_t - \mu\| > r^*$, as the alternate case is trivially true. For $r = \|x_t - \mu\| - r^*$, we can simply pick the vector $v = \Delta$ where Δ is the unit vector in the direction of $\mu - x_t$. Under Assumption 1, we have that for at least $0.95k$ points:

$$\langle Z_i - x_t, v \rangle = \langle Z_i - \mu, v \rangle + \langle \mu - x_t, v \rangle \geq \|x_t - \mu\| - r^* = r.$$

This implies the lower bound holds in the case where $\|x_t - \mu\| > r^*$.

For the upper bound $d_t \leq \|x_t - \mu\| + r^*$, suppose, for the sake of contradiction, there is a value of $r > \|x_t - \mu\| + r^*$ for which the optimal value of $MTE(x_t, r, \mathbf{Z})$ is greater than $0.9k$. Let v be the solution of $MTE(x_t, r, \mathbf{Z})$. This means that for $0.9k$ of the Z_i, we have:

$$\langle Z_i - \mu, v \rangle = \langle Z_i - x_t, v \rangle + \langle x_t - \mu, v \rangle \geq r - \|x_t - \mu\| > r^*.$$

This contradicts Assumption 1 and proves the upper bound. ■

Next, we prove the correctness of the **Gradient Estimation** step from Algorithm 3.

Lemma 4 *In the running of Algorithm 1, let us assume x_t satisfies:*

$$\|\mu - x_t\| \geq 1200 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{\|\Sigma\|k/n} \right), \tag{1}$$

and let Δ denote the unit vector in the direction of $\mu - x_t$. Then, under Assumption 1, we have that:

$$\langle g_t, \Delta \rangle \geq \frac{1}{2}.$$

Proof Let $r^* = 300 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{\|\Sigma\|k/n} \right)$. We have, from the definition of d_t, that for $0.9k$ of the Z_i, $\langle Z_i - x_t, g_t \rangle \geq d_t$. We also have, under Assumption 1, that $\langle Z_i - \mu, g_t \rangle \leq r^*$ for $0.95k$

---

4. This will be made precise in Corollary 5.

of the Z_i. From the pigeonhole principle, there exists a Z_j which satisfies both those inequalities. Therefore, for that Z_j, the lower bound from Lemma 3 implies

$$\|\mu - x_t\| - r^* \le d_t \le \langle Z_j - x_t, g_t \rangle = \langle Z_j - \mu, g_t \rangle + \langle \mu - x_t, g_t \rangle \le r^* + \|\mu - x_t\| \langle \Delta, g_t \rangle.$$

By rearranging the above inequality and using the assumption on $\|\mu - x_t\|$ in Eq. (1), we get the required conclusion. ∎

To control the probability that Assumption 1 holds, we assume the correctness of the following corollary of Lemma 11, formalizing the insight of (Lugosi and Mendelson, 2019):

Corollary 5 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k) \in \mathbb{R}^{k \times d}$ be k i.i.d. random vectors with mean μ and covariance Λ. Furthermore, assume $k \ge 3200 \log 1/\delta$. Then we have for all $v \in \mathbb{R}$ such that $\|v\| = 1$:*

$$\left| \{ i : \langle Y_i - \mu, v \rangle \ge 300 \left(\sqrt{\operatorname{Tr} \Lambda / k} + \sqrt{\|\Lambda\|} \right) \} \right| \le 0.05k$$

with probability at least $1 - \delta$.

By instantiating Corollary 5 with the $Y_i = Z_i$, we see that Assumption 1 holds with high probability.

Finally, we put the results of Lemma 3, Lemma 4 and Corollary 5 together to prove Theorem 2.

Proof [of Theorem 2] Assume first that Assumption 1 holds. Let $r^* = 1200 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{\|\Sigma\| k / n} \right)$.
To start with, let us define the set $\mathcal{G} = \{ x : \|x - \mu\| \le r^* \}$. We prove the theorem in two cases:

Case 1: None of the iterates x_t lie in $\mathcal{G}$. In this case, note that by Lemma 3 and the definition of r^*, we have:

$$\frac{3}{4} \|x_t - \mu\| \le d_t \le \frac{5}{4} \|x_t - \mu\|. \tag{2}$$

Moreover, we have by the definition of the update rule of x_t in Algorithm 1:

$$\|x_{t+1} - \mu\|^2 = \|x_t - \mu\|^2 + \frac{1}{2} d_t \langle x_t - \mu, g_t \rangle + \frac{d_t^2}{16} \le \|x_t - \mu\|^2 - \frac{d_t \|x_t - \mu\|}{4} + \frac{d_t^2}{16}$$

$$\le \|x_t - \mu\|^2 - \frac{3}{16} \|x_t - \mu\|^2 + \frac{25}{256} \|x_t - \mu\|^2 \le \frac{23}{25} \|x_t - \mu\|^2,$$

where we have used Lemma 4 for the first inequality and the inequalities in Eq. (2) for the second inequality. By iteratively applying the above inequality, we get the conclusion of the theorem in this case.

Case 2: At least one of the iterates x_t lies in $\mathcal{G}$. Therefore, we have from Lemma 3:

$$d_t \le 1500 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{\|\Sigma\| k / n} \right).$$

We also have at the completion of the algorithm, from another application of Lemma 3:

$$\|x^* - \mu\| - 300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{\|\Sigma\| k / n} \right) \le d^* \le d_t \le 1500 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{\|\Sigma\| k / n} \right).$$

By re-arranging the above inequality, we get the desired result.

By Corollary 5, Assumption 1 holds with probability at least $1 - \delta$ and therefore, the conclusions from Case 1 and Case 2 hold with probability $1 - \delta$. ∎

Bearing in mind that the polynomial optimization problem **MTE** is non-convex, we consider a convex relaxation in the following section.

4. Efficient Algorithm for Mean Estimation

In this section, we define a semi-definite programming relaxation of the polynomial optimization problem **MTE**. We then design new Distance Estimation and Gradient Estimation algorithms that use the tractable solutions to the relaxation instead of the original polynomial optimization problem. We then use these solutions to update our mean estimate along the same lines as those from Section 3, albeit with some added technical difficulty. Finally, we provide the analysis of the method and prove Theorem 1.

4.1. The Semi-Definite Relaxation of **MTE**

Here, we propose a semidefinite programming relaxation of **MTE**, a variant of the Threshold-SDP from (Hopkins, 2018). We first define a semidefinite matrix $X \in \mathbb{R}^{(k+d+1)\times(k+d+1)}$ symbolically indexed by 1, the variables b_i and v_j and denote by the vector $v_{b_i} := (X_{b_i,v_1}, \ldots, X_{b_i,v_d})$:

$$\max \sum_{i=1}^{k} X_{1,b_i}$$
$$X_{1,b_i} = X_{b_i,b_i}$$
$$X_{1,1} = 1$$
$$\sum_{j=1}^{d} X_{v_j,v_j} = 1$$
$$\langle v_{b_i}, Z_i - x \rangle \geq X_{b_i,b_i} r \ \forall i \in [k]$$
$$X \succcurlyeq 0 \tag{MT}$$

Similar to the polynomial optimization **MTE**, this optimization problem is also parameterized by a vector $x \in \mathbb{R}^d$, $r > 0$ and a matrix $\boldsymbol{Z} \in \mathbb{R}^{k\times d}$. We refer to solutions of this program as $(X, m) = MT(x, r, \boldsymbol{Z})$ with m denoting the optimal value and X denoting the optimal solution.

Algorithm 4 Distance Estimation

1: **Input**: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k\times d}$, Current point x
2: $d^* = \max\{r > 0 : MT(x, r, \boldsymbol{Z}) \geq 0.9k\}$
3: **Return:** d^*

Algorithm 5 Gradient Estimation

1: **Input**: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k\times d}$, Current point x
2: $d^* = $ Distance Estimation$(\boldsymbol{Z}, x)$
3: $(X, m) = MT(x, d^*, \boldsymbol{Z})$
4: $X_v = $ Submatrix of X corresponding to the indices v_i
5: $g = $ Top singular vector of X_v
6: $\mathcal{H} = \{i : \langle Z_i - x, g \rangle \geq 0\}$
7: **if** $|\mathcal{H}| \geq 0.9k$ **then**
8:     **Return:** g
9: **else**
10:     **Return:** $-g$
11: **end if**

The main contribution of our paper is in showing that the solutions to the relaxed optimization problem **MT** can be used to improve the mean estimate similarly to those of **MTE**. We redefine the Distance and Gradient Estimation steps in Algorithm 1 using **MT** in Algorithms 4 and 5.

4.2. Algorithm

To efficiently estimate the mean, we instantiate Algorithm 1 to use solutions of **MT** instead of **MTE**. The new Distance Estimation and Gradient Estimation procedures are stated in Algorithms 4 and 5.

As opposed to the polynomial optimization problem, solutions to the relaxation may not necessarily return a single vector v but rather a semidefinite matrix which corresponds to the relaxation of v. This matrix may not uniquely determine a direction of improvement. We, therefore, parse the solution to isolate a provably good direction of improvement and use this to iteratively improve our estimate. It is noteworthy that the singular value decomposition does not provide a sign direction. Thankfully the correct orientation is easily ascertained using the data points.

To analyze the runtime of Algorithm 1 with Algorithms 4 and 5, we first note that the semidefinite relaxation has $O(k^2 + d^2)$ variables. However, by projecting all the data down to a subspace containing the k bucket means, we may effectively reduce the number of variables to $O(k^2)$ with an $O(k^2 d)$ time pre-processing step. Therefore, we are now left with $O(k^2)$ variables. The runtime of interior point methods for solving semidefinite programs with $O(k^2)$ variables and $O(k)$ constraints is $O(k^{3.5})$ (Alizadeh, 1995). Furthermore, a single call of the Distance Estimation procedure can be efficiently implemented using $\widetilde{O}(1)$ rounds of binary search on the parameter r. Therefore, the total cost of a single call to Algorithm 4 is $\widetilde{O}(k^{3.5})$. Similarly, the total cost of a call to Algorithm 5 is $\widetilde{O}(k^{3.5})$. Since the cost of each iteration is dominated by a single call of Algorithm 4 and 5, the total cost per iteration is $\widetilde{O}(k^{3.5})$. Since, we only run $\widetilde{O}(1)$ iterations, the total cost of the Algorithm 1 instantiated with Algorithms 4 and 5 is $\widetilde{O}(k^{3.5} + k^2 d + nd)$.

4.3. Analysis

We now prove Theorem 1. We follow the same lines as the proof of Theorem 2, but with the added technical difficulties arising from the use of the semi-definite relaxation.

1. **Distance Estimation:** We show that the Distance Estimation step in Algorithm 4 provides an accurate estimate of the distance of the current point from the mean. See Section 4.3.1.

2. **Gradient Estimation:** Next, we show that when x is far away from the mean μ, the vector g output by Algorithm 5 is well aligned with the vector joining the current point x to the mean μ. See Section 4.3.2.

3. **Gradient Descent:** Combining the previous two steps, we prove that we eventually converge to a good approximation to the mean. See Section 4.3.3.

The following assumption is required to prove the correctness of the Distance Estimation and Gradient Estimation steps:

Assumption 2 *For the bucket means, $\mathbf{Z} = (Z_1, \ldots, Z_k)$, let $\mathcal{S}_r$ denote the set of feasible solutions for $MT(\mu, r, \mathbf{Z})$. Then, we have for all $r \geq 300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right)$,*

$$\max_{X \in \mathcal{S}_r} \sum_{i=1}^{k} X_{b_i, b_i} \leq \frac{k}{20}.$$

The above assumption is a strengthening of Assumption 1 for the case where we use **MT** instead of **MTE**. We use the following fact at several points in the subsequent analysis:

Remark 6 *Note that Assumption 2 implies Assumption 1.*

4.3.1. DISTANCE ESTIMATION STEP

In this subsection, we analyze the Distance Estimation step from Algorithm 4. We show that an accurate estimate of the distance of the current point from the mean can be found. We begin by stating a lemma that shows that a feasible solution for $MT(x, r, \boldsymbol{Z})$ can be converted to a feasible solution for $MT(\mu, 300 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right), \boldsymbol{Z})$ with a reduction in optimal value. The proof of the following lemma may be found in Appendix C.

Lemma 7 *Let us assume Assumption 2. Let $X \in \mathbb{R}^{(k+d+1)\times(k+d+1)}$ be a positive semi-definite matrix, symbolically indexed by 1 and the variables b_i and v_j. Moreover, suppose that X satisfies:*

$$X_{1,1} = 1, \quad X_{b_i, b_i} = X_{1, b_i}, \quad \sum_{j=1}^{d} X_{v_j, v_j} = 1, \quad \sum_{i=1}^{k} X_{b_i, b_i} \geq 0.9k.$$

Then, there is a set of at least $0.85k$ indices $\mathcal{T}$ such that for all $i \in \mathcal{T}$:

$$\langle Z_i - \mu, v_{b_i} \rangle < X_{b_i, b_i} 300 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right),$$

and a set of at least $k/3$ indices $\mathcal{R}$ such that for all $j \in \mathcal{R}$, we have $X_{b_j, b_j} \geq 0.85$.

The following lemma, proved in Appedix D, shows that if the distance between the mean μ and a point x is small then the estimate returned by Algorithm 4 is also small.

Lemma 8 *Suppose a point $x \in \mathbb{R}^d$ satisfies $\|x - \mu\| \leq 6000 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right)$. Then, under Assumption 2, Algorithm 4 returns a value d' satisfying*

$$d' \leq 7500 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right).$$

The next lemma, whose proof may be found in Appendix E, shows that the distance between the mean μ and a point x can be accurately estimated as long as x is sufficiently far from μ.

Lemma 9 *Suppose a point x satisfies $\tilde{d} = \|x - \mu\| \geq 6000 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right)$. Then, under Assumption 2, Algorithm 4 returns a value d' satisfying:*

$$0.95\tilde{d} \leq d' \leq 1.25\tilde{d}.$$

4.3.2. GRADIENT ESTIMATION STEP

In this section, we analyze the Gradient Estimation step of the algorithm. We show that an approximate gradient can be found as long as the current point x is not too close to the mean μ. The following lemma shows that we obtain a non-trivial estimate of the gradient in Algorithm 5. The proof of the following result may be found in Appendix F.

Lemma 10 *Suppose a point x satisfies $\|x - \mu\| \geq 6000 \left(\sqrt{\operatorname{Tr} \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right)$ and let Δ be the unit vector along $\mu - x$. Then under Assumption 2, Algorithm 5 returns a vector g satisfying:*

$$\langle g, \Delta \rangle \geq \frac{1}{15}.$$

4.3.3. GRADIENT DESCENT STEP

The following lemma guarantees that Assumption 2 holds with high probability and is used analogously to Corollary 5 in the proof of Theorem 2:

Lemma 11 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k) \in \mathbb{R}^{k \times d}$ be k i.i.d. random vectors with mean μ and covariance Λ and let $\mathcal{S}$ denote the set of feasible solutions of $MT(\mu, r, \boldsymbol{Y})$. Then, we have for $r \geq 300 \left(\sqrt{\operatorname{Tr} \Lambda / k} + \sqrt{\|\Lambda\|} \right)$ and $k \geq 3200 \log 1/\delta$:*

$$\max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{b_i, b_i} \leq \frac{k}{20},$$

with probability at least $1 - \delta$.

The proof of the lemma is an application of standard empirical process theory and concentration inequalities (Lugosi and Mendelson, 2019; Hopkins, 2018) and is proven in Appendix B.

The rest of the proof of Theorem 1 follows the same lines as that of Theorem 2 and is postponed to Appendix G.

5. Conclusion

In this paper, we proposed a computationally efficient estimator for the mean of a random vector which obtains the statistically optimal performance. This estimator has a significantly faster runtime together with a simpler analysis than previous works. Our algorithm is based on a descent method, where a current estimate of the mean is iteratively improved.

Considering the extension to M-estimation procedures (Brownlees et al., 2015; Hsu and Sabato, 2016; Lugosi and Mendelson, 2017) is a promising direction for further research, with as first step, the particular example of linear regression with heavy tailed noise and covariates (Audibert and Catoni, 2011).

References

F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM journal on Optimization*, 5(1):13–51, 1995.

N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 10 2011.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6):2507–2536, 12 2015.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017a.

O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *NIPS 2017 workshop; (Almost) 50 shades of Bayesian learning: PAC-Bayesian trends and insights*, 2017b.

M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford. Geometric median in nearly linear time. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, 2016.

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016.

S. B Hopkins. Sub-Gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17, 2016.

M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

E. Joly, G. Lugosi, and R. Oliveira. On the estimation of the mean of a random vector. *Electron. J. Statist.*, 11(1):440–451, 2017.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

M. Lerasle and R. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, to appear, 2017.

G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794, 04 2019.

S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, 1983.

Y. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization methods and software*, 9(1-3):141–160, 1998.

Appendix A. Auxiliary lemma

Lemma 12 *For any $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$ and $x \in \mathbb{R}^d$, the optimal value of $MT(x, r, \boldsymbol{Z})$ is monotonically non-increasing in r.*

Proof The lemma follows trivially from the fact that a feasible solution X of $MT(x, r, \boldsymbol{Z})$ is also a feasible solution for $MT(x, r', \boldsymbol{Z})$ for $r' \leq r$. ■

Appendix B. Proof of Lemma 11

We first show that the optimal value of the semi-definite program **MT** satisfies a bounded-difference condition with respect to the Z_i's.

Lemma 13 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ be any set of k vectors in $\mathbb{R}^d$. Now, let $\boldsymbol{Y}' = (Y_1, \ldots, Y'_i, \ldots, Y_k)$ be the same set of k vectors with the ith vector replaced by $Y'_i \in \mathbb{R}^d$. If m and m' are the optimal values of $MT(x, r, \boldsymbol{Y})$ and $MT(x, r, \boldsymbol{Y}')$, we have:*

$$|m - m'| \leq 1$$

Proof Firstly, assume that X is a feasible solution to $MT(x, r, \boldsymbol{Y})$. Now, let us define X' as:

$$X'_{i,j} = \begin{cases} X_{i,j} & \text{if } i, j \neq b_i \\ 0 & \text{otherwise} \end{cases}$$

That is X' is equal to X except with the row and column corresponding to b_i being set to 0. We see that X' forms a feasible solution to $MT(x, r, \boldsymbol{Y}')$. Therefore, we have that:

$$\sum_{j=1}^{k} X_{b_j, b_j} = \sum_{j=1, j \neq i}^{k} X'_{b_j, b_j} + X_{b_i, b_i} \leq \sum_{j=1, j \neq i}^{k} X'_{b_j, b_j} + 1 \leq m' + 1$$

where the bound $X_{b_i, b_i} \leq 1$ follows from the fact that the 2×2 sub-matrix of X formed by the rows and columns indexed by 1 and b_i is positive semidefinite and the constraint that $X_{b_i, b_i} = X_{1, b_i}$. Since the above series of equalities holds for all feasible solutions X of $MT(x, r, \boldsymbol{Y})$, we get:

$$m \leq m' + 1.$$

Through a similar argument, we also conclude that $m' \leq m + 1$. Putting the above two inequalities together, we get the required conclusion. ■

For the next few lemmas, we are concerned with the case where $x = \mu$. Since we already know that the optimal SDP value satisfies the bounded differences condition, we need to verify that the expectation is small. As a first step towards this, we define the 2-to-1 norm of a matrix M.

Definition 14 *The 2-to-1 norm of $M \in \mathbb{R}^{n \times d}$ is defined as*

$$\|M\|_{2 \to 1} = \max_{\substack{\|v\|=1 \\ \sigma_i \in \{\pm 1\}}} \sigma^\top M v = \max_{\|v\|=1} \|Mv\|_1$$

We consider the classical semidefinite programming relaxation of the 2-to-1 norm. To start with, we will define a matrix $X \in \mathbb{R}^{(n+d+1) \times (n+d+1)}$ with the rows and columns indexed by 1 and the elements σ_i and v_j. The semidefinite programming relaxation is defined as follows:

$$
\begin{aligned}
\max \sum_{i,j} & M_{i,j} X_{\sigma_i, v_j} \\
X_{1,1} &= 1 \\
\sum_{j=1}^{d} X_{v_j, v_j} &= 1 \\
X_{\sigma_i, \sigma_i} &= 1 \\
X &\succeq 0
\end{aligned}
\tag{TOR}
$$

We now state a theorem of Nesterov as stated in (Hopkins, 2018):

Theorem 15 *(Nesterov (1998)) There is a constant $K_{2 \to 1} = \sqrt{\pi/2} \leq 2$ such that the optimal value, m, of the semidefinite programming relaxation TOR satisfies:*

$$
m \leq K_{2 \to 1} \|M\|_{2 \to 1}.
$$

In the next step, we will bound the expected 2-to-1 norm of the random matrix Z. To do this, we begin by stating the famous Ledoux-Talagrand Contraction Theorem (Ledoux and Talagrand, 1991).

Theorem 16 *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors, $\mathcal{F}$ be a class of real-valued functions on $\mathbb{R}^d$ and $\sigma_i, \ldots, \sigma_n$ be independent Rademacher random variables. If $\phi : \mathbb{R} \to \mathbb{R}$ is an L-Lipschitz function with $\phi(0) = 0$, then:*

$$
\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \phi(f(X_i)) \leq L \cdot \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(X_i).
$$

We are now ready to bound the expected 2-to-1 norm of the random matrix Z.

Lemma 17 *Let $\mathbf{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^{n \times d}$ be a set of n i.i.d. random vectors such that $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i Y_i^\top] = \Lambda$. Then, we have:*

$$
\mathbb{E}\|\mathbf{Y}\|_{2 \to 1} \leq 2\sqrt{n \operatorname{Tr} \Lambda} + n \|\Lambda\|^{1/2}.
$$

Proof Denoting by Y and Y_i' random vectors that are independently and identically distributed as Y_i and by σ_i independent Rademacher random variables, we have:

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{Y}\|_{2 \to 1}] &= \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} |\langle Y_i, v \rangle| \right] = \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} |\langle Y_i, v \rangle| + \mathbb{E}|\langle v, Y_i \rangle| - \mathbb{E}|\langle v, Y_i \rangle| \right] \\
&\leq \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} |\langle Y_i, v \rangle| - \mathbb{E}|\langle Y_i', v \rangle| \right] + n \max_{\|v\|=1} \mathbb{E}[|\langle v, Y \rangle|] \\
&\leq \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} \sigma_i(|\langle Y_i, v \rangle| - |\langle Y_i', v \rangle|) \right] + n \max_{\|v\|=1} \mathbb{E}\left[|\langle v, Y \rangle| \right].
\end{aligned}
$$

15

Now, we have for the second term:

$$\max_{\|v\|=1} \mathbb{E}[|\langle v, Y \rangle|] \leq \max_{\|v\|=1} \sqrt{\mathbb{E}\langle v, Y \rangle^2} \leq \|\Lambda\|^{1/2}.$$

For the first term, we get via a standard symmetrization argument:

$$\mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} \sigma_i(|\langle Y_i, v \rangle| - |\langle Y_i', v \rangle|)\right] \leq \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} \sigma_i|\langle Y_i, v \rangle|\right] + \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} -\sigma_i|\langle Y_i', v \rangle|\right]$$

$$= 2\mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} \sigma_i|\langle v, Y_i \rangle|\right] \leq 2\mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} \sigma_i\langle v, Y_i \rangle\right]$$

$$= 2\mathbb{E}\left[\left\|\sum_{i=1}^{n} \sigma_i Y_i\right\|\right] \leq 2\left(\mathbb{E}\left[\left\|\sum_{i=1}^{n} \sigma_i Y_i\right\|^2\right]\right)^{1/2}$$

$$= 2\left(\mathbb{E}\sum_{1 \leq i,j \leq n} \sigma_i\sigma_j\langle Y_i, Y_j \rangle\right)^{1/2} = 2\sqrt{n \operatorname{Tr} \Lambda},$$

where the second inequality follows from the Ledoux-Talagrand Contraction Principle (Theorem 16) By putting the above two bounds together, we get the lemma. ∎

We now bound the expected value of $MT(\mu, r, \boldsymbol{Y})$ by relating it to $\|\boldsymbol{Y}\|_{2 \to 1}$.

Lemma 18 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k) \in \mathbb{R}^{k \times d}$ be a collection of k i.i.d. random vectors with mean μ and covariance Λ. Now, denoting by $\mathcal{S}$ the set of feasible solutions for $MT(\mu, r, \boldsymbol{Y})$, we have:*

$$\mathbb{E}\max_{x \in \mathcal{S}} \sum_{i=1}^{k} X_{1,b_i} \leq \frac{1}{2r}\left(5\sqrt{k \operatorname{Tr} \Lambda} + 2k\|\Lambda\|^{1/2}\right).$$

Proof Firstly, let X be a feasible solution for $MT(\mu, r, \boldsymbol{Y})$. We construct a new matrix W which is indexed by σ_i and v_j as opposed to b_i and v_j for X:

$$W_{\sigma_i,\sigma_j} = 4X_{b_i,b_j} - 2X_{1,b_i} - 2X_{1,b_j} + 1, \quad W_{v_i,v_j} = X_{v_i,v_j}, \quad W_{1,1} = 1,$$
$$W_{1,v_i} = X_{1,v_i}, \quad W_{1,b_i} = 2X_{1,b_i} - 1, \quad W_{v_i,b_j} = 2X_{v_i,b_j} - X_{1,v_i}.$$

We prove that Y is a feasible solution to the SDP relaxation TOR of $\boldsymbol{Y} - \mu$. We see that:

$$W_{\sigma_i,\sigma_i} = 1 \text{ and } \sum_{i=1}^{d} W_{v_i,v_i} = 1.$$

Then, we simply need to verify that Y is PSD. Let $w \in \mathbb{R}^{k+d+1}$ indexed by 1, σ_i and v_j. We construct from w a new vector w', indexed by 1, b_i and v_j and defined as follows:

$$w_1' = w_1 - \sum_{i=1}^{k} w_{\sigma_i}, \quad w_{b_i}' = 2w_{\sigma_i}, \quad w_{v_j}' = w_{v_j}.$$

With w' defined as above, we have the following equality:

$$w^\top W w = (w')^\top X w' \geq 0.$$

Since the above condition holds for all $w \in \mathbb{R}^{k+d+1}$, we get that $Y \succcurlyeq 0$. Therefore, we conclude that Y is a feasible solution to the SDP relaxation TOR of $\boldsymbol{Y} - \mu$.

We bound the expected value of $MT(\mu, r, \boldsymbol{Y})$ as follows, denoting by v_{b_i} the vector $(X_{b_i, v_1}, \ldots, X_{b_i, v_d})$ and by v the vector $(X_{1, v_1}, \ldots, X_{1, v_d})$:

$$\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{1, b_i} = \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{b_i, b_i} \leq \frac{1}{r} \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v_{b_i}, Y_i - \mu \rangle$$

$$= \frac{1}{2r} \mathbb{E} \max_{X \in S} \left[\sum_{i=1}^{k} \langle 2 v_{b_i} - v, Y_i - \mu \rangle + \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \right]$$

$$\leq \frac{1}{2r} \left(\mathbb{E} \max_{X \in S} \sum_{i=1}^{k} \langle 2 v_{b_i} - v, Y_i - \mu \rangle + \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \right).$$

We note that from the fact that X is PSD, we have that (from the fact that the 2×2 submatrix indexed by v_i and b_j is PSD):

$$X_{v_i, b_j}^2 \leq X_{v_i, v_i} X_{b_j, b_j} \leq X_{v_i, v_i} \implies \|v_{b_j}\|^2 = \sum_{i=1}^{d} X_{v_i, b_j}^2 \leq \sum_{i=1}^{d} X_{v_i, v_i} = 1.$$

Therefore, we get for the second term in the above equation:

$$\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \leq \mathbb{E} \left\| \sum_{i=1}^{k} Y_i - \mu \right\| \leq \left(\mathbb{E} \left\| \sum_{i=1}^{k} Y_i - \mu \right\|^2 \right)^{1/2} = (k \operatorname{Tr} \Lambda)^{1/2}.$$

We bound the first term using the following series of inequalities where Y is constructed from X as described above:

$$\mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^{k} \langle 2 v_{b_i} - v, Y_i - \mu \rangle = \mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^{k} \sum_{j=1}^{d} (Y_i - \mu)_j W_{\sigma_i, v_j} = \mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^{k} \sum_{j=1}^{d} (\boldsymbol{Y}_{i,j} - \mu_j) W_{\sigma_i, v_j}$$

$$\leq 2 \mathbb{E} \| \boldsymbol{Y} - \mathbf{1} \mu^\top \|_{2 \to 1} \leq 4 \sqrt{k \operatorname{Tr} \Lambda} + 2k \|\Lambda\|^{1/2},$$

where the first inequality follows from Theorem 15 and the second inequality follows from Lemma 17. By combining the above three inequalities, we finally get:

$$\mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^{k} X_{1, b_i} \leq \frac{1}{2r} \left(5 \sqrt{k \operatorname{Tr} \Lambda} + 2k \|\Lambda\|^{1/2} \right).$$

∎

We are now able to prove Lemma 11.

Proof [Lemma 11] From Lemma 18, we see that:

$$\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{b_i, b_i} \leq \frac{k}{40}.$$

Now from Lemma 13 and an application of the bounded difference inequality (see, for example, Theorem 6.2 in (Boucheron et al., 2013)), with probability at least $1 - \delta$:

$$\max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{b_i, b_i} \leq \frac{k}{20}.$$

∎

Appendix C. Proof of Lemma 7

Let $r = 300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right)$. We prove the lemma by contradiction. Firstly, note that X is infeasible for $MT(\mu, r, \boldsymbol{Z})$ as the optimal value for $MT(\mu, r, \boldsymbol{Z})$ is less than $k/20$ (Assumption 2). Note that the only constraints of $MT(\mu, r, \boldsymbol{Z})$ that are violated by X are constraints of the form:

$$\langle Z_i - \mu, v_{b_i} \rangle < X_{b_i, b_i} r.$$

Now, let $\mathcal{T}$ denote the set of indices for which the above inequality is violated. We can convert X to a feasible solution for $MT(\mu, r, \boldsymbol{Z})$ by setting to 0 the rows and columns corresponding to the indices in $\mathcal{T}$. Let X' be the matrix obtained by the above operation. We have from Assumption 2:

$$0.05k \geq \sum_{i=1}^{k} X'_{b_i, b_i} = \sum_{i=1}^{k} X_{b_i, b_i} - \sum_{i \in \mathcal{T}} X_{b_i, b_i} \geq 0.9k - |\mathcal{T}|,$$

where the last inequality follows from the fact that $X_{b_i, b_i} \leq 1$. By rearranging the above inequality, we get the first claim of the lemma.

For the second claim, let $\mathcal{R}$ denote the set of indices j satisfying $X_{b_j, b_j} \geq 0.85$. We have:

$$0.9k \leq \sum_{j=1}^{k} X_{b_j, b_j} = \sum_{j \in \mathcal{R}} X_{b_j, b_j} + \sum_{j \notin \mathcal{R}} X_{b_j, b_j} \leq |\mathcal{R}| + 0.85k - 0.85|\mathcal{R}| \implies \frac{k}{3} \leq |\mathcal{R}|.$$

This establishes the second claim of the lemma. ∎

Appendix D. Proof of Lemma 8

Let $r' = 7500 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right)$ and $r = 300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right)$. Suppose that the optimal value of $MT(x, r', \boldsymbol{Z})$ is greater than $0.9k$ and let its optimal solution be X. Let $\mathcal{R}$ and $\mathcal{T}$ denote the two sets whose existence is guaranteed by Lemma 7. From, the cardinalities of $\mathcal{R}$ and $\mathcal{T}$, we see that their intersection is not empty. For $j \in \mathcal{R} \cap \mathcal{T}$, we have:

$$0.85r' \leq \langle Z_j - x, v_{b_j} \rangle = \langle Z_j - \mu, v_{b_j} \rangle + \langle \mu - x, v_{b_j} \rangle < r + \|x - \mu\|,$$

where the first inequality follows from the fact that $j \in \mathcal{R}$ and the fact that X is feasible for $MT(x, r', \mathbf{Z})$ and the last inequality follows from the inclusion of j in $\mathcal{T}$ and Cauchy-Schwarz.

By plugging in the bounds on r' and r, we get:

$$\|x - \mu\| > 6075 \left(\sqrt{\mathrm{Tr}\, \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right).$$

This contradicts the assumption on $\|x - \mu\|$ and concludes the proof of the lemma. ∎

Appendix E. Proof of Lemma 9

Let us define the direction Δ to be the unit vector in the direction of $x - \mu$. From Assumption 1 (which is implied by Assumption 2), the number of Z_i satisfying $\langle Z_i - \mu, \Delta \rangle \geq 300 \left(\sqrt{\mathrm{Tr}\, \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right)$ is less than $k/20$. Therefore, we have that for at least $0.95k$ points:

$$\langle Z_i - x, -\Delta \rangle = \langle x - \mu + \mu - Z_i, \Delta \rangle = \|x - \mu\| - 300 \left(\sqrt{\mathrm{Tr}\, \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right) \geq 0.95\tilde{d}.$$

Along with the monotonicity[5] of $MT(x, r, \mathbf{Z})$ in r, this implies the lower bound.

For the upper bound, we show that the optimal value of $MT(x, 1.25\tilde{d}, \mathbf{Z})$ is less than $0.9k$. For the sake of contradiction, suppose that this optimal value is greater than $0.9k$. Let X be a feasible solution of $MT(x, 1.25\tilde{d}, \mathbf{Z})$ that achieves $0.9k$. Let $\mathcal{R}$ and $\mathcal{T}$ be the two sets whose existence is guaranteed by Lemma 7 and j be an element in their intersection. We have for j:

$$0.85(1.25\tilde{d}) \leq X_{b_j,b_j} 1.25\tilde{d} \leq \langle Z_j - x, v_{b_j} \rangle = \langle Z_j - \mu, v_{b_j} \rangle + \langle \mu - x, v_{b_j} \rangle$$
$$< X_{b_j,b_j} 300 \left(\sqrt{\mathrm{Tr}\, \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right) + \|\mu - x\| = X_{b_j,b_j} 300 \left(\sqrt{\mathrm{Tr}\, \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right) + \tilde{d},$$

where the first inequality follows from the inclusion of j in $\mathcal{R}$ and the last inequality follows from the inclusion of j in $\mathcal{T}$ and Cauchy-Schwarz. By re-arranging the above inequality, we get:

$$X_{b_j,b_j} > (1.0625\tilde{d} - \tilde{d}) \left(300 \left(\sqrt{\mathrm{Tr}\, \Sigma/n} + \sqrt{k\|\Sigma\|/n} \right) \right)^{-1} > 1,$$

which is a contradiction. Therefore, we get from the monotonicity of $MT(x, r, \mathbf{Z})$ (see Lemma 12), that $d' \leq 1.25\tilde{d}$ and this concludes the proof of the lemma. ∎

Appendix F. Proof of Lemma 10

In the running of Algorithm 5, let X denote the solution of $MT(x, d^*, \mathbf{Z})$. We begin by factorizing the solution X into $UU^\top$ with the rows of U denoted by $u_1, u_{b_1}, \ldots, u_{b_k}$ and $u_{v_1}, \ldots, u_{v_d}$. We also define the matrix $U_v = (u_{v_1}, \ldots, u_{v_d})$ in $\mathbb{R}^{(k+d+1)\times d}$. From the constraints in **MT**, we have:

$$X_{b_i,b_i} = \|u_{b_i}\|^2 \leq 1 \implies \|u_{b_i}\| \leq 1, \quad \sum_{j=1}^{d} X_{v_j,v_j} = \sum_{j=1}^{d} \|u_{v_j}\|^2 = \|U_v\|_F^2 = 1 \implies \|U_v\|_F = 1.$$

---

5. See Lemma 12 in Appendix A.

Let $\mathcal{R}$ and $\mathcal{T}$ denote the sets defined in Lemma 7. Let $j \in \mathcal{T} \cap \mathcal{R}$. By noting that $v_{b_j} = u_{b_j}^\top U_v$, we have for j:

$$0.85d^* \leq \langle Z_j - \mu, v_{b_j} \rangle + \langle \mu - x, v_{b_j} \rangle \leq X_{b_j, b_j} 300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right) + u_{b_j}^\top U_v (\mu - x),$$

where the first inequality follows from the inclusion of j in $\mathcal{R}$ and the second from its inclusion in $\mathcal{T}$. We get by rearranging the above equation and using our bound on d^* from Lemma 9:

$$0.80 \|\mu - x\| \leq 0.85 d^* \leq X_{b_j, b_j} 300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right) + u_{b_j}^\top U_v (\mu - x). \tag{3}$$

By rearranging Eq. (3), using Cauchy-Schwarz, $\|u_{b_i}\| \leq 1$ and the assumption on $\|x - \mu\|$:

$$\|U_v(\mu - x)\| \geq u_{b_j}^\top U_v(\mu - x) \geq 0.75 \|\mu - x\|.$$

We finally get that:

$$\|U_v \Delta\| \geq 0.75.$$

Now, we have:

$$1 = \|U_v\|_F^2 = \|U_v \mathcal{P}_\Delta\|_F^2 + \|U_v \mathcal{P}_\Delta^\perp\|_F^2 \geq \|U_v \mathcal{P}_\Delta^\perp\|_F^2 + (0.75)^2 \implies \|U_v \mathcal{P}_\Delta^\perp\|_F \leq 0.67.$$

Let y be the top singular vector of X_v. Note that $X_v = U_v^\top U_v$ and y is also the top right singular vector of U_v. We have that:

$$0.75 \leq \|U_v y\| \leq \|U_v \mathcal{P}_\Delta y\| + \|U_v \mathcal{P}_\Delta^\perp y\| \leq \|\mathcal{P}_\Delta y\| + \|U_v \mathcal{P}_\Delta^\perp\|_F \leq \|\mathcal{P}_\Delta y\| + 0.67.$$

This means that we have:

$$|\langle y, \Delta \rangle| \geq \frac{1}{15}.$$

Note that the algorithm returns either y or $-y$. Firstly, consider the case where $\langle y, \Delta \rangle > 0$. From Assumption 1 (implied by Assumption 2), we have for at least $0.95k$ points:

$$\langle Z_i - \mu, y \rangle \leq 300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right).$$

Therefore, we have for $0.95k$ points:

$$\langle Z_i - x, y \rangle = \langle Z_i - \mu, y \rangle + \langle \mu - x, y \rangle$$

$$\geq -300 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right) + \frac{6000 \left(\sqrt{\operatorname{Tr} \Sigma / n} + \sqrt{k \|\Sigma\| / n} \right)}{15} > 0.$$

This means that in the case where $\langle y, \Delta \rangle > 0$, we return y which satisfies $\langle \mu - x, y \rangle > 0$. This implies the lemma in this case. The case where $\langle y, \Delta \rangle < 0$ is similar with $-y$ used instead of y. This concludes the proof of the lemma. ∎

Appendix G. Proof of Theorem 1

Let $\mathcal{G} = \{x : \|x - \mu\| \leq 6000 \left(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{k\|\Sigma\|/n} \right)\}$. Also, we assume that Assumption 2 holds. We prove the theorem differentiating between two cases:

Case 1: None of the iterates x_t fall into the set $\mathcal{G}$. In this case, we have from Lemma 9 that:

$$0.95\|x_t - \mu\| \leq d_t \leq 1.25\|x_t - \mu\| \tag{4}$$

Now, we get:

$$\|x_{t+1} - \mu\|^2 = \|x_t - \mu\|^2 - 2\frac{d_t}{20}\langle g_t, \mu - x_t \rangle + \frac{d_t^2}{400} \leq \|x_t - \mu\|^2 - \frac{d_t\|\mu - x_t\|}{150} + \frac{d_t^2}{400}$$

$$\leq \|x_t - \mu\|^2 - d_t \left(\frac{\|\mu - x_t\|}{150} - \frac{d_t}{400} \right) \leq \left(1 - \frac{1}{500} \right) \|x_t - \mu\|^2.$$

where the first inequality follows from Lemma 10 and the last inequality follows by substituting the lower bound on d_t in the first term and the upper bound on d_t in the second term (Equation (4)). By an iterated application of the above inequality, we get the required result.

Case 2: One of the iterates x_t falls into the set $\mathcal{G}$. If the algorithm returns an element from $\mathcal{G}$, the theorem is trivially true. From Lemma 8, we have for this iterate $x_t \in \mathcal{G}$ that:

$$d_t \leq 7500 \left(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{k\|\Sigma\|/n} \right).$$

Therefore, we have at the completion of the algorithm a value $d^* \leq 7500 \left(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{k\|\Sigma\|/n} \right)$ together with x^* lying outside $\mathcal{G}$. Thus, we finally have from Lemma 9:

$$0.95\|x^* - \mu\| \leq 7500 \left(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{k\|\Sigma\|/n} \right) \implies \|x^* - \mu\| \leq 8000 \left(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{k\|\Sigma\|/n} \right).$$

By Lemma 11, Assumption 2 holds with probability at least $1 - \delta$ and therefore, the conclusions from Case 1 and Case 2 hold with probability $1 - \delta$.

Substituting the value of k, we obtain

$$\|x^* - \mu\| \leq \max \left(\epsilon, 8000 \left(\sqrt{\operatorname{Tr}\Sigma/n} + \sqrt{k\|\Sigma\|/n} \right) \right)$$

$$\leq \max \left(\epsilon, 480000 \left(\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{\|\Sigma\|\log 1/\delta}{n}} \right) \right),$$

with probability at least $1 - \delta$.

This concludes the proof of the theorem. ∎