Spectral Sample Reweighting, and Robust and “Heavy-Tailed” Mean Estimation

Logan Gnanapragasam

1 Introduction

Robust mean estimation is the problem of returning a close approximation of the mean of a distribution on $\mathbb{R}^d$ given n points that are drawn from the distribution and are then corrupted by an adversary. A recent paper on robust mean estimation ([3]) describes the “spectral sample reweighting[1]” problem. They then show that using an algorithm that solves the reweighting problem, it is possible to do robust mean estimation for covariance-bounded distributions and (non-robust) mean estimation in the “heavy-tailed” setting.

In Section 2, we will define many terms that are commonly used in the robust mean estimation literature. Then, in Section 3, we will describe the spectral sample reweighting problem and the algorithm that can be used to solve it. In Section 4, we show how to solve the robust mean estimation problem using an algorithm for the spectral sample reweighting problem. In Section 5, we do the same for the “heavy-tailed” mean estimation problem.

2 Background

2.1 Randomized Algorithms

Let $\mathcal{I}$ be the “set of inputs” of an algorithm $\texttt{alg}$ and $\mathcal{O}$ be the “set of outputs” of the algorithm. An **algorithm** takes an input $i \in \mathcal{I}$, does some computation, and returns an output in $\mathcal{O}$. When doing the computations, a standard algorithm is allowed to initialize some variables, often called “local variables” since they are only used by the algorithm. However, it must explicitly state the values that it will initialize the variables with. For example, a standard algorithm is allowed to say “set the local variable v to $(1,0,0)$”. Thus, the output of the algorithm is specified by its input, since all of its local variables have explicit values that they are initialized with. That is, for each input i, there is an output $o_i^{\texttt{alg}} \in \mathcal{O}$ such that the algorithm always returns $o_i^{\texttt{alg}}$ when given the input i.

Note that a standard algorithm cannot use the following instruction: “set the local variable v to a vector in $\mathbb{R}^3$ sampled from the standard Gaussian distribution,” since this does not specify the value of v uniquely. Random algorithms will be allowed to do this. That is, a **random algorithm** will take an input $i \in \mathcal{I}$, then initialize its local variables $v_1, \ldots, v_j$

[1] The original paper calls it the spectral sample reweighing problem. There is no difference in the content, but the problem involves putting weights on data points, so the name “reweighting” is more appropriate

by sampling from a distribution[2] D on the set of all possible initializations of local variables, and then it will return an output. Therefore, the output of the algorithm is not only a function of i, but also a function of the local variables chosen. We write $o_i^{\mathtt{alg}}(v_1, \ldots, v_j)$ for the output of the algorithm on input i when $v_1, \ldots, v_j$ are the variables chosen – that is, $o_i^{\mathtt{alg}}$ is a function dependent on the initialization of the local variables.

Now, if $E \subset \mathcal{O}$, then $(o_i^{\mathtt{alg}})^{-1}(E)$ is a set of possible initializations of local variables. If this is a measurable set, we define $p_i^{\mathtt{alg}}(E)$**, the probability that** $o_i^{\mathtt{alg}}$ **is in** E, as the probability that $(v_1, \ldots, v_j) \in (o_i^{\mathtt{alg}})^{-1}(E)$ – that is, as the measure of $(o_i^{\mathtt{alg}})^{-1}(E)$. Note that $p_i^{\mathtt{alg}}$ is a probability distribution on $\mathcal{O}$, whence the random algorithm can be used as a part of other random algorithms.

For problems where each input has a set of outputs that "solve" the problem, we can define a probability of success (or a success rate) of a randomized algorithm. For each input i, let E_i be the set of all outputs that solve the problem. The **success rate** is $\inf_i p_i^{\mathtt{alg}}(E_i)$. The **failure rate** is defined similarly (replacing the inf with a sup, and replacing $p_i^{\mathtt{alg}}(E_i)$ with $p_i^{\mathtt{alg}}(E_i^c)$). While we will use this definition later, it isn't enough to state the robust mean estimation problem. To see why, a sample $\{x_1, x_2, \ldots, x_n\}$ can be drawn from many probability distributions on $\mathbb{R}^d$, so when we define the problem we need to include a distribution. That is, a "solution" to the problem depends not only on the given input of the problem, but also on the unknown underlying distribution that it was sampled from, so we can't define an E_i as we did here.

2.2 Problem Statement

Now that we know what a randomized algorithm is, we can define the robust mean estimation problem more carefully. To begin, we clarify the statement involving the "adversary".

Definition 1 (ϵ-corruption of a set)**.** Let $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be an arbitrary finite set. We say that a set S' is *ϵ-corrupted from S* if $S' = (S \setminus S_r) \cup S_b$, where $|S_r| = |S_b| \le \epsilon|S|$ and $S_r \subset S$. That is, S' is obtained from S by removing an ϵ-fraction of the points (S_r) and replacing them with arbitrary "bad" points (S_b). Also, we write $\mathrm{corr}_\epsilon(S)$ for the set of all S' that are ϵ-corrupted from S.

Let D be an arbitrary probability distribution on $\mathbb{R}^d$ with a finite mean. Let μ_D be the mean of D. Let $\mathtt{alg}$ be an arbitrary randomized algorithm with $\mathcal{I} = \{U \subset \mathbb{R}^d : |U| < \infty\}$, the set of finite subsets of $\mathbb{R}^d$, and $\mathcal{O} = \mathbb{R}^d$ – so the randomized algorithm takes as input a finite subset of $\mathbb{R}^d$ and outputs a vector in $\mathbb{R}^d$. Now, for any $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and any $\rho, \epsilon > 0$, define (writing $p_{\mathtt{alg}}(i; E)$ instead of $p_i(E)$)

$$q(\mathtt{alg}, \rho, D, \epsilon, S) = \inf_{S' \in \mathrm{corr}_\epsilon(S)} p_{\mathtt{alg}}(S'; B(\mu_D, \rho)).$$

Note that $p_{\mathtt{alg}}(S'; B(\mu_D, \rho))$ is the probability that on input S', $\mathtt{alg}$ returns a vector $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mu_D\| < \rho$. If S is the set of clean samples from D that are "given to the

[2] Usually the algorithm specifies each distribution separately (e.g., let v_1 be drawn from a standard distribution on $\mathbb{R}^d$ and v_2 be drawn from the uniform distribution on $[0, 1]$), but sometimes they are specified jointly (e.g., let v_1 be drawn from a Poisson distribution and let v_2 be drawn from a Gaussian with mean 0 and variance v_1), in which case it's convenient to write D as a joint distribution.

adversary", then the input to the algorithm can be any $S' \in \mathrm{corr}_\epsilon(S)$. So $q(\mathtt{alg}, \rho, D, \epsilon, S)$ can be thought of as **the probability that $\mathtt{alg}$ returns an output that is within ρ of the true mean of D when the input is ϵ-corrupted from a set of good samples S.** Then we define

$$\mathrm{robustsuccess}(\mathtt{alg}, \rho, D, \epsilon, n) = \mathbb{E}_{X_j \sim D \text{ i.i.d.}, 1 \leq j \leq n}[q(\mathtt{alg}, \rho, D, \epsilon, \{X_1, \ldots, X_n\})].$$

Think of $\mathrm{robustsuccess}(\mathtt{alg}, \rho, D, \epsilon, n)$ as the **probability that the output of $\mathtt{alg}$ on an ϵ-corrupted set of n samples from D lies within ρ of the true mean**.

With this definition, an attempt at stating the robust mean estimation problem is:

Given a family $\mathcal{D}$ of distributions on $\mathbb{R}^d$, find a randomized algorithm $\mathtt{alg}$, a "breakdown point" $\epsilon_0 > 0$, and a "sample complexity bound" n_0 and an "error bound" $\rho > 0$ such that $\mathrm{robustsuccess}(\mathtt{alg}, \rho, D, \epsilon, n)$ is large for all $D \in \mathcal{D}$, $\epsilon < \epsilon_0$, and $n \geq n_0$.

However, in nearly every theorem about robust mean estimation, ρ and n_0 depend on ϵ. So, rather than attempt to state the robust mean estimation problem in a fully general way, we give an example of how the definition of "robustsuccess" is used in many theorems about robust mean estimation.

In this paper, we make the convention that $\|\cdot\|$ is the Euclidean norm for vectors and the corresponding operator norm for matrices. When there is ambiguity, we will be more explicit by putting a subscript on the norm (e.g., $\|\cdot\|_p$ for the p-norm on vectors and corresponding operator norm).

Theorem 1 (Restated Theorem 4.1 of [3]). *Let $\mathcal{D}$ be the family of distributions D over $\mathbb{R}^d$ with covariance Σ_D satisfying $\|\Sigma_D\| \leq 1$. There is an algorithm $\mathtt{alg}$ and constants $c, c' > 0$ such that for all $D \in \mathcal{D}$, $\epsilon < \frac{1}{10}$, and $n \geq c' d \log d/\epsilon$, the success probability $\mathrm{robustsuccess}(\mathtt{alg}, c\sqrt{\epsilon}, D, \epsilon, n)$ is large.*

In this theorem, the algorithm is $\mathtt{alg}$, the breakdown point is $\frac{1}{10}$, the sample complexity bound is $c' d \log d/\epsilon$, and the error bound is $c\sqrt{\epsilon}$.

While we should be careful about measurability, we'll ignore it nearly everywhere. Most of the time the algorithm's output depends continuously on the randomly chosen samples and the sets we need are Borel sets, so there's not much to be concerned about.

2.2.1 Relation to Regularity Conditions

To show that $\mathrm{robustsuccess}(\mathtt{alg}, c\sqrt{\epsilon}, D, \epsilon, n)$ is large, a common technique is to consider sample sets that satisfy some additional conditions which make the robust mean estimation problem more feasible. This technique is often referred to as introducing a "regularity condition". For example, here is a regularity condition.

Definition 2 (Regularity Condition of Lemma 4.2 of [3]). Let $\mu \in \mathbb{R}^d$ and $\Sigma \preceq I$, $\Sigma \in \mathbb{R}^{d \times d}$. We say that a set S of size n "satisfies the (μ, Σ)-regularity condition with c and c'" if, for all $S' \in \mathrm{corr}_\epsilon(S)$, there is $G \subset S'$ such that $|G| \geq (1-\epsilon)n$, and with $\frac{1}{|G|}\sum_{x_i \in G} x_i = \mu_G$, we have $\|\mu - \mu_G\| \leq c\sqrt{\epsilon}$, and $\left\|\frac{1}{|G|}\sum_{x_i \in G}(x_i - \mu_G)(x_i - \mu_G)^T\right\| \leq c'$.

Then, it is shown that (a) the probability $q(\texttt{alg}, \rho, D, \epsilon, S)$ is high whenever S satisfies the regularity conditions with universal constants c, c' and (b) with high probability, $\{X_1, \ldots, X_n\}$ satisfies the regularity conditions with c, c' when $X_j \sim D$ are i.i.d.

To see why regularity conditions are useful, let Ω be the set of all n-element sets in $\mathbb{R}^d$, and let λ_D be the probability distribution on this set corresponding to sampling independently from D. Then robustsuccess$(\texttt{alg}, \rho, D, \epsilon, n) = \int_\Omega q(\texttt{alg}, \rho, D, \epsilon, S)\, d\lambda_D$. Let $\Omega_{\mathrm{reg},D}$ be the set of n-element sets that satisfy the regularity condition with respect to D, and suppose that there is $q_0 > 0$ such that $q(\texttt{alg}, \rho, D, \epsilon, S) \geq q_0$ whenever $S \in \Omega_{\mathrm{reg},D}$ (this is usually shown in a theorem about the randomized algorithm). Then

$$\int_\Omega q(\texttt{alg}, \rho, D, \epsilon, S)\, d\lambda_D \geq \int_{\Omega_{\mathrm{reg},D}} q(\texttt{alg}, \rho, D, \epsilon, S)\, d\lambda_D \geq q_0 \lambda_D(\Omega_{\mathrm{reg},D}).$$

So if both of these factors are close to 1, then robustsuccess$(\texttt{alg}, \rho, D, \epsilon, n)$ is also close to 1.

2.3 Heavy-Tailed Distributions

The term heavy-tailed is used somewhat frequently in the robust mean estimation literature. However, it often used without a definition. In some contexts, the phrase "heavy-tailed distribution" is used to distinguish a more general type of distribution from a specific type of distribution. For example, [4] uses the term "heavy-tailed" to describe a distribution that may not be sub-Gaussian, but does have finite covariance.

However, it is often unclear what the distinction between heavy-tailed and non-heavy-tailed is. For example, in [3], the distributions in the robust setting (Section 4) have bounded covariance, and the distributions in the "heavy-tailed" setting (Section 6) have finite covariance. So both parts consider the same types of distributions, and it's unclear what the difference is (or if there is a difference at all). However it appears that important part of the "heavy-tailed" estimation is that the error guarantee is stronger than the guarantee in Section 4. In this case, the phrase "heavy-tailed" is used as a reminder that the only assumptions about the distribution are that it has finite mean and covariance. So, **from now on, we will avoid using the term heavy-tailed**.

3 Spectral Sample Reweighting

First, we define the **α-spectral sample reweighting** problem. The problem involves putting weights on a collection of samples in order to make the spectral norm of the reweighed covariance of the samples small. Before we state the problem, we'll define some notation.

Notation 1. For each $0 < \epsilon < \frac{1}{2}$, $\mathcal{W}_{n,\epsilon} = \{(w_1, \ldots, w_n) \in \mathbb{R}^n : \sum w_i = 1, 0 \leq w_i \leq \frac{1}{(1-\epsilon)n}\}$ is called the set of "good weights".

Notation 2. Let A and B be symmetric matrices. We write $A \preceq B$ if $B - A$ is positive semidefinite (PSD) and $A \prec B$ if $B - A$ is positive definite. The symbols $\succeq$ and $\succ$ are defined similarly.

Note that for a symmetric matrix A, we have $A \preceq I$ if and only if $\|A\| \leq 1$.

Problem 1 (α-Spectral Sample Reweighting, $\alpha > 0$)**.** Let $\{x_1, \ldots, x_n\}$ be a set of points in $\mathbb{R}^d$ and let λ be positive. Suppose that there are $\nu \in \mathbb{R}^d$ and $w \in \mathcal{W}_{n,\epsilon}$ such that

$$\sum_{i=1}^{n} w_i(x_i - \nu)(x_i - \nu)^T \preceq \lambda I.$$

Given $\{x_1, \ldots, x_n\}$ and λ (but not ν and w), find $\nu' \in \mathbb{R}^d$ and $w' \in \mathcal{W}_{n,\epsilon}$ such that

$$\sum_{i=1}^{n} w'_i(x_i - \nu')(x_i - \nu')^T \preceq \alpha\lambda I.$$

The assumption that there are $w \in \mathcal{W}_{n,\epsilon}$ and $\nu \in \mathbb{R}^d$ such that $\sum_{i=1}^n w_i(x_i - \nu)(x_i - \nu)^T \preceq \lambda I$ is called the **spectral centrality assumption**. To assist in the later exposition, we define the weighted mean and covariance.

Notation 3. Let $S = \{x_1, \ldots, x_n\}$ be a set of points. If $w \in \mathbb{R}^n$ is any weight vector (i.e., a vector of positive components that sum to 1), then we write $\nu_S(w) = \sum_{1\leq i\leq n} w_i x_i$ and $M_S(w) = \sum_{1\leq i\leq n} w_i(x_i - \nu_S(w))(x_i - \nu_S(w))^T$.

Note that if there are $\nu \in \mathbb{R}^d$ and $w \in \mathcal{W}_{n,\epsilon}$ such that $\sum_{i=1}^n w_i(x_i - \nu)(x_i - \nu)^T \preceq \lambda I$, then we can take $\nu = \nu_S(w)$ and $\sum_{i=1}^n w_i(x_i - \nu_S(w))(x_i - \nu_S(w))^T \preceq \lambda I$. So the α-spectral sample reweighting problem can be stated without ν, ν': if there is $w \in \mathcal{W}_{n,\epsilon}$ such that $M_S(w) \preceq \lambda I$, then given λ and S, without knowing w, find w' such that $M_S(w') \preceq \alpha\lambda I$. However, the reason for including ν in the statement of Problem 1 is that any $\nu \in \mathbb{R}^d$ that satisfies the requirement in the problem is sometimes called a "spectral center" of the dataset – we will state this more rigorously when we discuss definitions of the center of a dataset.

The exact value of the constant α doesn't matter for our purposes: any randomized algorithm that solves the α-spectral sample reweighting problem will be sufficient. For concreteness, we'll show that the randomized algorithm for spectral sample reweighting in [3] solves the 60-spectral sample reweighting problem (with high probability). This algorithm is framed as a regret minimization algorithm. In Section 3.1, we will introduce the regret minimization problem and the algorithm given in [1] that solves it. Then, in Section 3.2, we explain how it applies to the spectral sample reweighting problem.

3.1 Regret Minimization

Let $\Delta_n \subset \mathbb{R}^n$ be the set of probability distributions on $\{1, \ldots, n\}$, and let $\mathcal{P}$ be a convex, compact subset of Δ_n. We will build up to the regret minimization problem on $\mathcal{P}$ by starting with simple situations and adding complexity.

Consider a situation where an algorithm is given n possible decisions, each with a cost. Let $m \in \mathbb{R}^n$ be the "cost vector" for the decisions: that is, it is a vector such that $m_i \in [-1, 1]$ is the "cost" of decision i. For any $p \in \mathcal{P}$, if p_i is "the probability of choosing i", then the expected cost of the decision is $m \cdot p$.

Now consider a similar situation in which there are n possible decisions and the algorithm must choose a probability distribution; however, in this situation, the cost is allowed to vary with time. Specifically, at time $t = 1, 2, \ldots, T$, the cost vector is $m^{(t)}$, and the total expected cost of the probability distribution p is $\sum_{i=1}^{n} m^{(t)} \cdot p$.

Now, we come to regret minimization. Again, there are n decisions, and the cost is allowed to vary with time. However, in regret minimization, the algorithm doesn't know what the costs will be in the present time. It only knows the historical information. More specifically, at each time t, it is given the costs of the decisions *in the past* (i.e., $m^{(1)}$ through $m^{(t-1)}$), and it must choose a probability distribution $p^{(t)}$ that will be used at time t. *After* the algorithm makes the decision, a cost vector for time t is chosen. The cost vector may be chosen according to the probability distribution $p^{(t)}$, so in particular it might be chosen to make the expected cost large. The total expected cost of the algorithm's choices is $\sum_{t=1}^{T} m^{(t)} \cdot p^{(t)}$. If the algorithm knew all the cost vectors ahead of time, it would choose $p \in \mathcal{P}$ to minimize $\sum_{t=1}^{T} m^{(t)} \cdot p$. So, the **regret** of the algorithm on the costs is defined by $\sum_{t=1}^{T} m^{(t)} \cdot p^{(t)} - \min_{p \in \mathcal{P}} \sum_{t=1}^{T} m^{(t)} \cdot p$, and the algorithm's goal is to minimize the regret (so that its choices are not much worse[3] than the optimal choice for the sum).

The Multiplicative Weights algorithm (see Algorithm 1 below) is given in [1]. It uses the **relative entropy** from one distribution to another: if p and q are distributions on $\{1, \ldots, n\}$, then the relative entropy from q to p is $\mathrm{RE}(p \parallel q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$.

Algorithm 1 The Multiplicative Weights Algorithm, [1]

```
Initialize $0 < \eta \leq \frac{1}{2}$ and $p^{(1)}$ as the uniform distribution on $\{1, \ldots, n\}$
for $t = 1, 2, \ldots, T$ do
    "Observe" the costs of the decisions $m^{(t)}$.
    Define $\widehat{p}^{(t+1)} \in \mathbb{R}^n$ by $\widehat{p}_i^{(t+1)} \leftarrow p_i^{(t)}(1 - \eta m_i^{(t)}) / (\sum_{j=1}^{n} p_j^{(t)}(1 - \eta m_j^{(t)}))$.
    "Project $\widehat{p}^{(t+1)}$ onto $\mathcal{P}$" via $p^{(t+1)} \leftarrow \operatorname{argmin}_{p \in \mathcal{P}} \mathrm{RE}(p \parallel \widehat{p}^{(t+1)})$.
end for
```

Now, we can state the theorem. In the algorithm for the spectral sample reweighing problem, the components of the cost vectors will be nonnegative, so we can state the theorem in a less general way.

Theorem 2 (Special case of Theorem 2.4 of [1])**.** *Algorithm 1 guarantees that for any $T \geq 1$, any $0 < \eta \leq \frac{1}{2}$, and any $p \in \mathcal{P}$, if $m_i^{(t)} \in [0, 1]$ for $1 \leq t \leq T$, then*

$$\sum_{t=1}^{T} m^{(t)} \cdot p^{(t)} \leq (1 + \eta) \sum_{t=1}^{T} m^{(t)} \cdot p + \frac{\mathrm{RE}(p \parallel p^{(1)})}{\eta}.$$

3.2 Application to reweighting

Note that each $\mathcal{W}_{n,\epsilon}$ is an intersection of closed convex sets, and it is a subset of the compact set Δ_n. Therefore, we can apply this algorithm with $\mathcal{P} = \mathcal{W}_{n,\epsilon}$.

[3] Interestingly, it is possible for the regret to be negative. As an example, if the cost functions are chosen so that they are minimized at $p^{(t)}$ and all the $p^{(t)}$ are distinct, then the regret could be negative. Concretely, take $n \geq 2$, $\mathcal{P} = \Delta_n$, $T = n$, $p^{(j)} = e_j$ is the jth standard basis vector in $\mathbb{R}^n$, and $m^{(j)} = e_{n+1-j}$. Then the total expected cost of the algorithm is 0, hence the regret is -1.

Before we do this, we introduce a randomized algorithm. For any PSD matrix M and any constant $c \in (0,1)$, a c-approximate largest eigenvector of M is a unit vector u such that $u^T M u \geq c \|M\|$. Let $\textsc{ApproximateTopEigenvector}(M, c; r)$ be a randomized algorithm that takes a PSD matrix M and $c \in (0,1)$ and outputs a c-approximate largest eigenvector of M with a failure rate smaller than r. We require that the algorithm always outputs a unit vector. The details of the algorithm are not necessary for the analysis, but according to [3], there is a power-method-based randomized algorithm that satisfies these requirements.

Algorithm 2 The Multiplicative Weights Algorithm for Spectral Sample reweighting

```
Input: $\{x_1, \dots, x_n\} = S$, $\lambda > 0$
Initialize $\eta = \frac{1}{2}$ and $w^{(1)}$ as the uniform distribution on $\{1, \dots, n\}$
Initialize $\rho$ to be the squared 2-norm diameter of $S$.
for $t = 1, 2, \dots, T$ do
    Compute the weighted mean and covariance by $\nu^{(t)} \leftarrow \nu_S(w^{(t)})$, $M^{(t)} \leftarrow M_S(w^{(t)})$.
    $u^{(t)} \leftarrow \textsc{ApproximateTopEigenvector}(M^{(t)}, 7/8; \delta/T)$.
    Compute the outlier score vector $\tau^{(t)}$ by $\tau_i^{(t)} \leftarrow \langle u^{(t)}, x_i - \nu^{(t)} \rangle^2$
    Compute the cost vector $m^{(t)} \leftarrow \tau^{(t)}/\rho$.
    Compute $\widehat{w}^{(t+1)} \in \mathbb{R}^n$ by $\widehat{w}_i^{(t+1)} \leftarrow w_i^{(t)}(1 - \eta m_i^{(t)}) / (\sum_{j=1}^n w_j^{(t)}(1 - \eta m_j^{(t)}))$.
    Compute $w^{(t+1)} \leftarrow \operatorname{argmin}_{w \in \mathcal{W}_{n,\epsilon}} \mathrm{RE}(w \parallel \widehat{w}^{(t+1)})$.
end for
Output $\nu^{(t^*)}$ and $w^{(t^*)}$ where $t^* = \operatorname{argmin}_t \left\|M^{(t)}\right\|$.
```

The algorithm is in Algorithm 2. Computing the cost vector corresponds to "observing the costs" in Line 3 of Algorithm 1. In robust mean estimation, a common technique is to assign each point an **outlier score**, which is a measure of how much the algorithm thinks the point is an outlier. So the cost of the data point i corresponds to how much the algorithm thinks it's an outlier. Moreover, as we set ρ to the diameter of S and $\nu^{(t)}$ is a convex combination of points in S, Cauchy's inequality shows that the components of the cost vectors in Algorithm 2 are in $[0,1]$. Hence Theorem 2 can be applied to this algorithm by using $\eta = \frac{1}{2}$. After multiplying each term by ρ to write the cost vectors in terms of outlier score vectors and then dividing by the iteration count, we get the following guarantee.

Lemma 1. *Algorithm 2 guarantees that for any $w \in \mathcal{W}_{n,\epsilon}$ and $T \geq 0$,*

$$\frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)}, \tau^{(t)} \rangle \leq \frac{3}{2T} \sum_{t=1}^{T} \langle w, \tau^{(t)} \rangle + \frac{2\rho \, \mathrm{RE}(w \parallel w^{(1)})}{T}$$

where $\tau^{(t)}$ are the score vectors computed in the algorithm.

We will use this to show that with a high probability, $\frac{1}{T} \sum_{t=1}^{T} \left\|M^{(t)}\right\|$ is small when T is sufficiently large. Here the probability is due to the randomized algorithm $\textsc{ApproximateTopEigenvector}$.

In the analysis of Algorithm 2, we will bound $\left\|\nu_S(w) - \nu^{(t)}\right\|$, where w is a good weight vector that satisfies the assumptions of Problem 1. For this, we'll first need a lemma.

Lemma 2. *The diameter of $\mathcal{W}_{n,\epsilon}$ in the* 1*-norm is at most* 4ϵ.

Proof. As $\{(w_1, \ldots, w_n) \in \mathbb{R}^n : \sum w_i = 1, 0 \leq w_i \leq r\} = S_r$ is compact, the sup in the definition of diameter is attained, so there are $u, v \in S_r$ such that $\|u - v\|_1 = \operatorname{diam}(S_r)$. As S_r is convex, u and v must be vertices of S_r (recall that a vertex of a convex set is a point that is not a midpoint of a nondegenerate segment in S_r)[4]. Since u is a vertex of S_r, we can only have $0 < u_i < r$ for at most one index i —otherwise, we would have indices u_i and u_j which both satisfy the strict inequality, and u is the midpoint of $(u_1, \ldots, u_i \pm \delta, \ldots, u_j \mp \delta, \ldots, u_n) \in S_r$ if δ is small enough. Therefore, the components of u and v consist of the following: $j = \lfloor \frac{1}{r} \rfloor$ coordinates that have value r, 1 coordinate that has value $1 - jr$, and $n - j - 1$ coordinates that have value 0. Now, we only need to find the maximum distance between any two points of this form.

By permuting the entries of u and v if necessary, we can assume $u_1 = 1 - jr$, $u_2 = u_3 = \cdots = u_{j+1} = r$, and $u_{j+2} = \cdots = u_n = 0$. There are three choices for the first component of v.

- $1 - jr$: in u and v, there are j components with value r and $n - j - 1$ components with value 0 that need to be matched. The largest distance occurs when $v = (1 - jr, 0, \ldots, 0, r, \ldots, r)$. As $j \geq \lfloor \frac{n}{2} \rfloor$, it follows that $n - j - 1 \leq j$ so $\|u - v\|_1 = r \cdot 2(n - j - 1)$.

- 0: the $1 - jr$ component in v is either paired with a 0 or with an r. If it's paired with an r component in u, then we swap the r component in u with the $1 - rj$ component in u. The distance doesn't change, since $r - (1 - rj) + (1 - rj) - 0 = r - 0 + (1 - rj) - (1 - rj)$. Thus this distance is handled by the first case already. So assume that the $1 - rj$ component of v is paired with a 0 in u. By permuting the entries of v if necessary, we can assume that $u = (1 - jr, r, r, \ldots, r, 0, 0, \ldots, 0)$ and $v_1 = 0$ and $v_n = 1 - jr$. Now, in both u and v, there are j components with value r and $n - j - 2$ components with value 0 that need to be matched. As $n - j - 2 < j$, the distance is $2(1 - jr) + 2r(n - j - 2)$.

- r: By the same argument, the only important case is when the $1 - jr$ term in v is paired with the r term of u. After that, there are $j - 1$ components with value 0 and $n - j - 1$ components with value r that need to be matched. So the distance is $2((j + 1)r - 1) + 2r(\min(j - 1, n - j - 1))$.

Each of these is less than $2rn - 2$: the first because it's equal to $2nr - 2r(j + 1)$; the second because $1 - jr < r$ and hence $2(1 - jr) + 2r(n - j - 2) < 2r(n - j - 1)$; and the third because $\min(j - 1, n - j - 1) \leq n - j - 1$. Thus, $\operatorname{diam}(\mathcal{W}_{n,\epsilon}) = \operatorname{diam}(S_{\frac{1}{(1-\epsilon)n}}) \leq 2\frac{1}{1-\epsilon} - 2$, and since $\frac{1}{1-\epsilon} \leq 2$, we have $2(\frac{1}{1-\epsilon} - 1) = \frac{2\epsilon}{1-\epsilon} \leq 4\epsilon$. □

As $\nu^{(t)} = \nu_S(w^{(t)})$, our goal is to bound $\left\|\nu_S(w) - \nu_S(w^{(t)})\right\|$. We will state the result in a slightly more general way than what's stated in [3], and it shows the results that they want to use.

[4] As the 1-norm isn't ***strictly*** convex (all points on the line segment from $(1, 0)$ to $(0, 1)$ have the same 1-norm), this isn't immediate. We don't know that the distance is strictly larger when we replace u or v with the endpoints of the segment. However, since S_r is convex and compact in $\mathbb{R}^n$, it is the convex hull of its vertices, so it's enough just to check the vertices.

Lemma 3 (Lemma A.1 of [3]). *Let $S = \{x_1, \dots, x_n\}$ be a set of points in $\mathbb{R}^d$ and let λ be positive. Suppose that there are $w \in \mathcal{W}_{n,\epsilon}$ and $\nu \in \mathbb{R}^d$ such that*

$$M = \sum_{i=1}^{n} w_i (x_i - \nu)(x_i - \nu)^T \preceq \lambda I.$$

Then for any $w' \in \mathcal{W}_{n,\epsilon}$,

$$\begin{aligned} \|\nu - \nu_S(w')\| &\leq \frac{1}{1-\sqrt{2\epsilon}} \left(\|\nu_S(w) - \nu\| + \sqrt{2\epsilon\lambda} + \sqrt{2\epsilon \, \|M_S(w')\|} \right) \\ &\leq \frac{1}{1-\sqrt{2\epsilon}} \left(\sqrt{\lambda} + \sqrt{2\epsilon\lambda} + \sqrt{2\epsilon \, \|M_S(w')\|} \right). \end{aligned}$$

In the special case $\nu = \nu_S(w)$ (which we are interested in), the first inequality can be thought of as "removing the $\sqrt{\lambda}$ term" because it shows

$$\|\nu_S(w) - \nu_S(w')\| \leq \frac{1}{1-\sqrt{2\epsilon}} \left(\sqrt{2\epsilon\lambda} + \sqrt{2\epsilon \, \|M_S(w')\|} \right).$$

Proof. Let $w = (w_1, \dots, w_n)$ and $w' = (w'_1, w'_2, \dots, w'_n) \in \mathcal{W}_{n,\epsilon}$. Then we have

$$\begin{aligned} \|\nu_S(w') - \nu\|^2 = \langle \nu_S(w') - \nu, \sum_{i=1}^{n} w'_i (x_i - \nu) \rangle &= \sum_{i=1}^{n} w'_i \langle \nu_S(w') - \nu, x_i - \nu \rangle \\ &= \sum_{i=1}^{n} w_i \langle \nu_S(w') - \nu, x_i - \nu \rangle \\ &\quad + \sum_{i: w_i > w'_i} (w'_i - w_i) \langle \nu_S(w') - \nu, x_i - \nu \rangle \\ &\quad + \sum_{i: w'_i > w_i} (w'_i - w_i) \langle \nu_S(w') - \nu, x_i - \nu \rangle \end{aligned}$$

We will bound each term and then rearrange.

By Cauchy's inequality,

$$\begin{aligned} \sum_{i=1}^{n} w_i \langle \nu_S(w') - \nu, x_i - \nu \rangle &= \langle \nu_S(w') - \nu, \sum_{i=1}^{n} w_i (x_i - \nu) \rangle \\ &= \langle \nu_S(w') - \nu, \nu_S(w) - \nu \rangle \leq \|\nu_S(w') - \nu\| \, \|\nu_S(w) - \nu\|. \end{aligned}$$

So the first term is bounded by $\|\nu_S(w') - \nu\| \, \|\nu_S(w) - \nu\|$.

For the next two terms, we define $\alpha_i = w'_i - w_i$. Then the 1-norm of $w' - w$ is $\sum_{i: w'_i > w_i} \alpha_i - \sum_{i: w_i > w'_i} \alpha_i$, and as $w, w' \in \mathcal{W}_{n,\epsilon}$, this difference is at most 4ϵ by Lemma 2. On the other hand, w and w' are weight vectors, so the components of w and w' sum to 1. So we have $\sum_{i: w'_i > w_i} \alpha_i + \sum_{i: w_i > w'_i} \alpha_i = \sum_{i=1}^{n} \alpha_i = 0$. This yields $\sum_{i: w'_i > w_i} |\alpha_i| = \sum_{i: w'_i > w_i} \alpha_i = -\sum_{i: w_i > w'_i} \alpha_i = \sum_{i: w_i > w'_i} |\alpha_i|$. Therefore,

$$\sum_{i: w_i > w'_i} |\alpha_i| = \sum_{i: w_i > w'_i} |\alpha_i| \leq 2\epsilon.$$

We bound $\sum_{i:w_i>w_i'}(w_i'-w_i)\langle \nu_S(w')-\nu, x_i-\nu\rangle = \sum_{i:w_i>w_i'}\alpha_i\langle \nu_S(w')-\nu, x_i-\nu\rangle$. If $w_i > w_i'$, then $w_i > 0$ and $\left|\frac{\alpha_i}{w_i}\right| = \left|\frac{w_i'}{w_i}-1\right| \le 1$, so $\frac{\alpha_i^2}{w_i} \le |\alpha_i|$. By Cauchy's inequality,

$$\begin{aligned}
\left(\sum_{i:w_i>w_i'}\frac{\alpha_i}{\sqrt{w_i}}\sqrt{w_i}\langle \nu_S(w')-\nu, x_i-\nu\rangle\right)^2 &\le \left(\sum_{i:w_i>w_i'}\frac{\alpha_i^2}{w_i}\right)\left(\sum_{i:w_i>w_i'} w_i\langle \nu_S(w')-\nu, x_i-\nu\rangle^2\right)\\
&\le \left(\sum_{i:w_i>w_i'}|\alpha_i|\right)\left(\sum_{i=1}^n w_i\langle x_i-\nu, \nu_S(w')-\nu\rangle^2\right)\\
&= \left(\sum_{i:w_i>w_i'}|\alpha_i|\right)(\nu_S(w')-\nu)^T M_S(w)(\nu_S(w')-\nu)\\
&\le 2\epsilon\lambda \left\|\nu_S(w')-\nu\right\|^2.
\end{aligned}$$

Thus, the second term is bounded by $\sqrt{2\epsilon\lambda}\,\|\nu_S(w')-\nu\|$.

Finally, we bound $\sum_{i:w_i'>w_i}(w_i'-w_i)\langle \nu_S(w')-\nu, x_i-\nu\rangle = \sum_{i:w_i'>w_i}\alpha_i\langle \nu_S(w')-\nu, x_i-\nu\rangle$. We begin in the same way as the previous part. If $w_i' > w_i$, then $w_i' > 0$ and $\left|\frac{\alpha_i}{w_i'}\right| = \left|\frac{w_i}{w_i'}-1\right| \le 1$, so $\frac{\alpha_i^2}{w_i'} \le |\alpha_i|$. As before,

$$\begin{aligned}
\left(\sum_{i:w_i'>w_i}\frac{\alpha_i}{\sqrt{w_i'}}\sqrt{w_i'}\langle \nu_S(w')-\nu, x_i-\nu\rangle\right)^2 &\le \left(\sum_{i:w_i'>w_i}\frac{\alpha_i^2}{w_i'}\right)\left(\sum_{i:w_i'>w_i} w_i'\langle \nu_S(w')-\nu, x_i-\nu\rangle^2\right)\\
&\le 2\epsilon\sum_{i=1}^n w_i'\langle \nu_S(w')-\nu, x_i-\nu\rangle^2.
\end{aligned}$$

Now we simplify:

$$\begin{aligned}
\sum_{i=1}^n w_i'\langle \nu_S(w')-\nu, x_i-\nu\rangle^2 &= \sum_{i=1}^n w_i'\langle \nu_S(w')-\nu, x_i-\nu_S(w')+\nu_S(w')-\nu\rangle^2\\
&= \sum_{i=1}^n w_i'\left(\langle \nu_S(w')-\nu, x_i-\nu_S(w')\rangle^2 + \langle \nu_S(w')-\nu, \nu_S(w')-\nu\rangle^2\right)\\
&\quad + \sum_{i=1}^n 2w_i'\langle \nu_S(w')-\nu, x_i-\nu_S(w')\rangle\langle \nu_S(w')-\nu, \nu_S(w')-\nu\rangle.
\end{aligned}$$

But as $\sum_{i=1}^n w_i'\langle \nu_S(w')-\nu, x_i-\nu_S(w')\rangle = 0$, the sum on the last line is 0. Moreover, $\sum_{i=1}^n w_i'\langle \nu_S(w')-\nu, x_i-\nu_S(w')\rangle^2 = (\nu_S(w')-\nu)^T M_S(w')(\nu_S(w')-\nu)$, which is at most $\|M_S(w')\|\,\|\nu_S(w')-\nu\|^2$. As $\langle \nu_S(w')-\nu, \nu_S(w')-\nu\rangle^2$ is a constant independent of i and the coefficients of w' sum to 1, we get

$$\left(\sum_{i:w_i'>w_i}\alpha_i\langle \nu_S(w')-\nu, x_i-\nu\rangle\right)^2 \le 2\epsilon\left(\|M_S(w')\|\,\|\nu_S(w')-\nu\|^2 + \|\nu_S(w')-\nu\|^4\right)$$

Thus, the third term is bounded by $\sqrt{2\epsilon\left(\|M_S(w')\| + \|\nu_S(w') - \nu\|^2\right)}\,\|\nu_S(w') - \nu\|$.

So, we have

$$\|\nu_S(w') - \nu\|^2 \le \left(\|\nu_S(w) - \nu\| + \sqrt{2\epsilon\lambda} + \sqrt{2\epsilon\left(\|M_S(w')\| + \|\nu_S(w') - \nu\|^2\right)}\right)\|\nu_S(w') - \nu\|\,.$$

Divide by $\|\nu_S(w') - \nu\|$ to get

$$\|\nu_S(w') - \nu\| \le \|\nu_S(w) - \nu\| + \sqrt{2\epsilon\lambda} + \sqrt{2\epsilon\,\|M_S(w')\|} + \sqrt{2\epsilon}\,\|\nu_S(w') - \nu\|\,.$$

Note that this also holds immediately if $\|\nu_S(w') - \nu\| = 0$ (in which case division is not legal). Isolating $\|\nu_S(w') - \nu\|$ yields the first inequality.

For the second inequality we can apply Jensen's inequality to the convex function $\phi(x) = x^2$ to bound $\|\nu_S(w) - \nu\|$:

$$\begin{aligned}\|\nu_S(w) - \nu\|^2 = \sup_{\|u\|=1}\langle\nu(w) - \nu, u\rangle^2 &= \sup_{\|u\|=1}\left(\sum_{i=1}^n w_i\langle x_i - \nu, u\rangle\right)^2\\ &\le \sup_{\|u\|=1}\sum_{i=1}^n w_i\langle x_i - \nu, u\rangle^2\\ &= \sup_{\|u\|=1}\sum_{i=1}^n w_i u^T(x_i - \nu)(x_i - \nu)^T u\\ &= \sup_{\|u\|=1} u^T M u \le \lambda,\end{aligned}$$

because $M \preceq \lambda I$. So $\|\nu_S(w) - \nu\| \le \sqrt{\lambda}$. □

We are now ready to give an analysis of Algorithm 2.

Theorem 3 (Lemma 3.2 of [3]). *Let $\epsilon \in (0, 1/24]$. Then Algorithm 2 solves the 60-spectral sample reweighting problem in $T = 6\rho\epsilon/\lambda$ (round up if necessary) iterations with a failure rate of (at most) δ. Here ρ is the diameter of the set of $\{x_i\}$.*

Proof. We will show that if every call to ApproximateTopEigenvector returns a valid solution, then the algorithm solves the spectral sample reweighting problem. Note that ApproximateTopEigenvector is called T times with a failure rate of $\frac{\delta}{T}$ at each time. So by the union bound, this would show that Algorithm 2 has a failure rate of (at most) δ.

Let $w \in \mathcal{W}_{n,\epsilon}$ be such that $M_S(w) \preceq \lambda I$ (by the assumptions of Problem 1, such a w exists). Also, let $\nu = \nu_S(w)$. The discussion after the statement of the problem shows that this is a valid choice of ν for the spectral centrality assumption. Note that $\langle w^{(t)}, \tau^{(t)}\rangle = \sum_{i=1}^n w_i^{(t)}\langle u^{(t)}, x_i - \nu_S(w^{(t)})\rangle^2 = (u^{(t)})^T M_S(w^{(t)}) u^{(t)}$, as we saw in the proof of Lemma 1. By the guarantees of ApproximateTopEigenvector, we have $\langle w^{(t)}, \tau^{(t)}\rangle \ge \left\|M_S(w^{(t)})\right\|$. Then Lemma 1 shows that

$$\frac{7}{8T}\sum_{t=1}^T\left\|M_S(w^{(t)})\right\| \le \frac{3}{2T}\sum_{t=1}^T\langle w, \tau^{(t)}\rangle + \frac{2\rho\,\mathrm{RE}(w \parallel w^{(1)})}{T}$$

We first handle the relative entropy term. Since $w \in \mathcal{W}_{n,\epsilon}$ and each component of $w^{(1)}$ is $\frac{1}{n}$:

$$\mathrm{RE}(w \parallel w^{(1)}) = \sum_{i=1}^{n} w_i \log(nw_i) \leq \sum_{i=1}^{n} \frac{1}{n(1-\epsilon)} \log \frac{1}{1-\epsilon} = \frac{1}{1-\epsilon} \log\left(\frac{1}{1-\epsilon}\right),$$

and the last term is a convex function for $\epsilon \in \left[0, \frac{1}{2}\right]$ by the second derivative test. So $\mathrm{RE}(w \parallel w^{(1)}) \leq (4 \ln 2)\epsilon < 3\epsilon$. Thus, if $T \geq \frac{6\rho\epsilon}{\lambda}$ then $\frac{2\rho\,\mathrm{RE}(w \| w^{(1)})}{T} < \lambda$.

Now we handle the sum.

$$\begin{aligned}
\frac{3}{2T}\sum_{t=1}^{T}\langle w, \tau^{(t)}\rangle &= \frac{3}{2T}\sum_{t=1}^{T}\sum_{i=1}^{n} w_i \left(\langle x_i - \nu, u^{(t)}\rangle + \langle \nu - \nu_S(w^{(t)}), u^{(t)}\rangle\right)^2 \\
&= \frac{3}{2T}\sum_{t=1}^{T}\sum_{i=1}^{n} w_i\langle x_i - \nu, u^{(t)}\rangle^2 + w_i\langle \nu - \nu_S(w^{(t)}), u^{(t)}\rangle^2 \\
&\quad + \frac{3}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} w_i\langle x_i - \nu, u^{(t)}\rangle\langle \nu - \nu_S(w^{(t)}), u^{(t)}\rangle
\end{aligned}$$

As $\nu = \nu_S(w)$, we have $\sum_{i=1}^{n} w_i\langle x_i - \nu, u^{(t)}\rangle\langle \nu - \nu_S(w^{(t)}), u^{(t)}\rangle = 0$. Also, Cauchy's inequality shows $\langle \nu - \nu_S(w^{(t)}), u^{(t)}\rangle^2 \leq \left\|\nu_S(w) - \nu_S(w^{(t)})\right\|^2$, and finally $\sum_{i=1}^{n} w_i\langle x_i - \nu_S(w), u^{(t)}\rangle^2 = (u^{(t)})^T M_S(w) u^{(t)} \leq \lambda$ as we have seen before. So

$$\frac{3}{2T}\sum_{t=1}^{T}\langle w, \tau^{(t)}\rangle \leq \frac{3}{2T}\sum_{t=1}^{T}\lambda + \left\|\nu_S(w) - \nu_S(w^{(t)})\right\|^2 = \frac{3}{2}\lambda + \frac{3}{2T}\sum_{t=1}^{T}\left\|\nu_S(w) - \nu_S(w^{(t)})\right\|^2.$$

We use Lemma 3 to bound the sum. As $\epsilon \leq \frac{1}{2k}$ with $k = 12$, we have

$$\begin{aligned}
\left\|\nu_S(w) - \nu_S(w^{(t)})\right\|^2 &\leq \frac{1}{1 - 2\sqrt{\frac{1}{k}} + \frac{1}{k}}\left(\sqrt{\frac{\lambda}{k}} + \sqrt{\frac{1}{k}\left\|M_S(w^{(t)})\right\|}\right)^2 \\
&\leq \frac{\lambda + \left\|M_S(w^{(t)})\right\| + 2\sqrt{\lambda\left\|M_S(w^{(t)})\right\|}}{(k+1) - 2\sqrt{k}} \\
&\leq \frac{3}{(k+1) - 2\sqrt{k}}\lambda + \frac{3}{(k+1) - 2\sqrt{k}}\left\|M_S(w^{(t)})\right\|.
\end{aligned}$$

Sum from $t = 1$ to T:

$$\frac{3}{2T}\sum_{t=1}^{T}\left\|\nu_S(w) - \nu_S(w^{(t)})\right\|^2 \leq \frac{3}{2}\cdot\frac{3}{(k+1) - 2\sqrt{k}}\lambda + \frac{3}{2}\cdot\frac{3}{(k+1) - 2\sqrt{k}}\cdot\frac{1}{T}\sum_{t=1}^{T}\left\|M_S(w^{(t)})\right\|,$$

so

$$\frac{7}{8T}\sum_{t=1}^{T}\left\|M_S(w^{(t)})\right\| \leq \frac{3}{2}\cdot\left(1 + \frac{3}{(k+1) - 2\sqrt{k}}\right)\lambda + \frac{3}{2}\cdot\frac{3}{(k+1) - 2\sqrt{k}}\cdot\frac{1}{T}\sum_{t=1}^{T}\left\|M_S(w^{(t)})\right\|$$

Isolating $\frac{1}{T}\sum_{t=1}^{T}\left\|M_S(w^{(t)})\right\|$ and then doing the arithmetic with $k = 12$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\left\|M_S(w^{(t)})\right\| \leq 17\lambda.$$

So the output $M_S(w^{(t^*)})$ has spectral norm at most 17λ, hence $M^{(t^*)} \preceq 60\lambda I$ as needed. □

A few comments about the proof: first, the computation is tedious, but some care is required. When we isolate $\frac{1}{T}\sum_{t=1}^{T}\left\|M_S(w^{(t)})\right\|$, we need to ensure that it's multiplied by a positive number so that division preserves the direction of the inequality. Second, this proof shows that we can do 17-spectral sample reweighting, which is better than 60-spectral sample reweighting. When $k = 11$ (i.e., $\epsilon \leq \frac{1}{22}$) we can do the same analysis and show that the algorithm solves the 65-spectral sample reweighting problem. However, putting $k = 10.5$ results in a negative factor on the average of the $\left\|M_S(w^{(t)})\right\|$, so we can't use the same analysis there. As we mentioned earlier, the exact constant α will not matter for us in the future.

In [3], a few more facts are shown. Specifically, they show that if a set satisfies the assumptions of Problem 1, then most points lie in a ball of radius $\sqrt{\frac{d\lambda}{\epsilon}}$, and there is a randomized algorithm that finds a ball with radius $4\sqrt{\frac{d\lambda}{\epsilon}}$ so that most points lie in this ball. This allows them to bound the number of iterations in a way that's independent of the points in the input set. As the inputs to the algorithm may be adversarial, without this step, the runtime of the algorithm could be arbitrarily large (the adversary can put points as far from the origin as is necessary to increase the iteration count). The interested reader should see Lemma 3.3 and Lemma 3.4 of [3], but we won't be concerned too much with runtime in our exposition, so we can ignore it. The authors of [3] also give a number of other reweighting algorithms which improve efficiency and breakdown point. Specifically, in Appendix C, they show a more efficient algorithm that uses a "matrix-multiplicative update approach"; in Appendix D, they show an approach using gradient descent; and in Appendix E of they show that there is a variant of the algorithm that works for all $0 < \epsilon < \frac{1}{2}$ (i.e., its "breakdown point" is optimal). We will not prove any of these, but we'll use the algorithm with optimal breakdown point if necessary.

4 Robust Mean Estimation

In this section, we see how the spectral sample reweighting problem can be used to do robust mean estimation for distributions with a covariance bounded by I. We will show that if a set satisfies a regularity condition, then it satisfies the spectral centrality assumption, and a solution to the reweighting problem yields a solution to the mean estimation problem.

We've already introduced the regularity condition in Definition 2. The authors claim:

Lemma 4 (Lemma 4.2 of [3]). *Let D be a distribution with mean μ and covariance $\Sigma \preceq I$. There are constants $c_{sample}, C, C' > 0$ such that if $n \geq c_{sample}(d \log d)/\epsilon$ and $X_j \sim D$ are i.i.d. samples for $1 \leq j \leq n$, then with a high probability, $\{X_1, \ldots, X_n\}$ satisfy the (μ, Σ)-regularity condition of Definition 2 with constants C, C'.*

Theorem 4 (Theorem 4.1 of [3])**.** *Let D be a distribution with mean μ and covariance $\Sigma \preceq I$. If $0 < \epsilon \leq \frac{1}{24}$ and if S is an n-element set that is ϵ-corrupted from a set of points that satisfy the (μ, Σ)-regularity condition of Definition 2 with constants c_{mean}, c_{cov}, then there is a constant $c_{err} > 0$ such that with high probability, Algorithm 2 returns $\widehat{\mu} = \nu^{(t^*)}$ such that $\|\widehat{\mu} - \mu\| \leq c_{err}\sqrt{\epsilon}$.*

Here the "probability" refers to the probability that the algorithm's output is in the correct ball.

Proof. First, we show that the points satisfy the spectral centrality assumption. Then, we apply the guarantees of Theorem 3. By the regularity condition there is $G \subset S$ such that $|G| \geq (1-\epsilon)n$ and $\left\|\frac{1}{|G|}\sum_{x_i \in G} x_i - \mu\right\| \leq c_{\text{mean}}\sqrt{\epsilon}$. Define $w \in \mathcal{W}_{n,\epsilon}$ by $w_i = \frac{1}{|G|}$ if $x_i \in G$, $w_i = 0$ otherwise (note that although we do not know G, we know that it exists, so we know that w exists). Then $M_S(w)$ is the empirical covariance of G, so $M_S(w) \preceq c_{\text{cov}} I$ by the regularity condition. By Theorem 3, with high probability, Algorithm 2 outputs $w^{(t^*)}$ and $\widehat{\mu} = \nu_S(w^{(t^*)})$ such that $M_S(w^{(t^*)}) \preceq 60 c_{\text{cov}} I$. So applying Lemma 3 with $\nu = \nu_S(w)$ gives $\|\nu_S(w) - \widehat{\mu}\| \leq \frac{1}{1-\sqrt{2\epsilon}}\left(\sqrt{2\epsilon c_{\text{cov}}} + \sqrt{2\epsilon 60 c_{\text{cov}}}\right)$. By the triangle inequality,

$$\|\mu - \widehat{\mu}\| \leq \|\mu - \nu_S(w)\| + \|\nu_S(w) - \widehat{\mu}\| \leq c_{\text{mean}}\sqrt{\epsilon} + \frac{1}{1-\sqrt{2\epsilon}}\left(\sqrt{2\epsilon c_{\text{cov}}} + \sqrt{2\epsilon 60 c_{\text{cov}}}\right).$$

As $\epsilon \leq \frac{1}{24}$, this is at most $\left(c_{\text{mean}} + 15\sqrt{c_{\text{cov}}}\right)\sqrt{\epsilon}$. Take $c_{\text{err}} = c_{\text{mean}} + 15\sqrt{c_{\text{cov}}}$.

The assertion about probability holds by Theorem 3. □

So, robust mean estimation for distributions with covariance bounded by identity can be done using an algorithm that solves the spectral sample reweighting problem. In Appendix B of [3], the authors show that a similar algorithm works for sub-Gaussian distributions with a stronger error guarantee of $O(\epsilon \log \frac{1}{\epsilon})$. To do this, they show that Algorithm 2 satisfies a stronger guarantee, as long as the inputs also do. But as before, we will not go through the proof here.

5 Mean Estimation with Sub-Gaussian Error Rate

In this section, we describe how spectral sample reweighting can be used to do mean-estimation with sub-Gaussian error rate. We first state the problem. Then, we describe the Lugosi-Mendelson estimator which gives the required error bound. This motivates a new definition of center. We show that the new definition of center coincides with the spectral centrality assumption, which will allow us to use the spectral sample reweighting algorithm.

Problem 2. Let D be a distribution with mean μ and covariance Σ. Given δ and n such that $2^{-n} \leq \delta < 1$, define

$$r_{\delta,n}(\Sigma) = \sqrt{\frac{\text{tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\| \log(\frac{1}{\delta})}{n}}.$$

Find a constant $c > 0$ and a randomized algorithm that takes n samples from D and outputs $\widehat{\mu}$ such that with probability at least $1-\delta$, $\|\widehat{\mu} - \mu\| \leq c r_{\delta,n}(\Sigma)$. The error $c r_{\delta,n}(\Sigma)$ is called **sub-Gaussian error**.

Here the probability is taken both with respect to the randomness of the algorithm (i.e., the distribution that the algorithm draws its parameters from) and with respect to the distribution of the samples. Note that we have dropped the requirement that the algorithm should be robust.

We now state the Lugosi-Mendelson lemma. As it isn't the focus of this paper, we won't prove it. To get the sub-Gaussian error rate in one dimension, the "median-of-means" approach can be used as described in [5]: first, partition the n data points into $k = \lceil 8\log(\frac{1}{\delta})\rceil$ disjoint buckets, each of size at least $\lfloor \frac{n}{k} \rfloor$, and compute the empirical mean of each bucket. Then, take the estimator $\widehat{\mu}_\delta$ to be the median of these means. If σ is the variance of the distribution (recall we are in one dimension for now), then with probability at least $1-\delta$ we have

$$|\widehat{\mu}_\delta - \mu| \leq 8\sigma\sqrt{\frac{\log\frac{2}{\delta}}{N}}.$$

However, it's unclear how to extend this to $d > 1$ dimensions, because the definition of "median" needs to be made. The lemma that motivates the definition of "median" is one of the contributions of [5]. We state it in the way that's stated in [3]:

Lemma 5 (Lemma 6.1 of [3])**.** *Let D be a distribution with mean μ and covariance Σ, and let δ and n be as in the statement of Problem 1. Let $\{X_1, \ldots, X_n\}$ be i.i.d. samples from D. Let $k = \lceil 800\log(\frac{1}{\delta})\rceil$. Group the indices $\{1, 2, \ldots, n\}$ into k buckets $B_1, \ldots, B_k$, each of size at least $\lfloor \frac{n}{k} \rfloor$. Let $\{Z_j : 1 \leq j \leq n\}$ be the random variables defined by $Z_j = \frac{1}{\operatorname{card} B_j}\sum_{i\in B_j} X_i$. Then, with probability at least $1-\delta$,*

$$\operatorname{card}\{1 \leq i \leq k : |\langle Z_i - \mu, v\rangle| \geq 3000 r_{\delta,N}(\Sigma)\} \leq 0.01k \textit{ for all } v \in \mathbb{R}^d, \|v\| = 1.$$

That is, with probability at least $1-\delta$, for every unit vector v, most of the bucket means Z_j achieve sub-Gaussian error rates from μ when projected along v.

5.1 Spectral and Combinatorial Centrality

We can now introduce the notions of center in [3] and show that they are the same. We first define a notion of center which is inspired by the guarantee of Lemma 5.

Definition 3 ((ϵ, λ)-combinatorial center)**.** The point $\nu \in \mathbb{R}^d$ is a (ϵ, λ)-combinatorial center of the set $\{z_1, \ldots, z_k\} \subset \mathbb{R}^d$ if for all unit vectors v in $\mathbb{R}^d$,

$$|\{1 \leq i \leq k : |\langle z_i - \nu, v\rangle| \geq \sqrt{\lambda}\}| \leq \epsilon k.$$

Thus, Lemma 5 shows that with high probability, μ is a $(0.01, (3000r_{\delta,N}(\Sigma))^2)$-combinatorial center of the bucket means.

We want to connect this to the spectral centrality assumption. So, we'll define "spectral center" in a way that's consistent with the spectral centrality assumption, then show that a combinatorial center is a spectral center and vice versa. Although the ideas are from [3], we'll provide some more exposition than what's there.

Definition 4 (Density Matrix)**.** A matrix $M \in \mathbb{R}^{d\times d}$ is called a "density matrix" if it is PSD and its trace is 1.

The following fact will be quite useful as it will allow us to apply the minimax theorem.

Lemma 6. *The set of density matrices is a convex, compact subset of* $\mathbb{R}^{d\times d}$.

Proof. All norms on $\mathbb{R}^{d\times d}$ are equivalent. By taking the Frobenius norm, which corresponds with the 2-norm on $\mathbb{R}^{d^2}$, it's enough to show that the set of density matrices is closed and bounded. Moreover, the set of PSD matrices is both closed and convex, so it suffices to show that the set of matrices with trace 1 is closed and convex, and that the set of density matrices is bounded. In both of these parts, we can use any norms on $\mathbb{R}^{d\times d}$.

To see that the set of matrices with trace 1 is closed, we show that $\operatorname{tr} : \mathbb{R}^{d\times d} \to \mathbb{R}$ is continuous: if $\{M_n\}$ is a sequence of matrices and $M_n \to M$ in the Frobenius norm, then $M_n \to U$ entrywise, so $\operatorname{tr} M_n \to \operatorname{tr} M$. So the preimage of 1 under tr is a closed subset of $\mathbb{R}^{d\times d}$. To see that it's convex, note that if $M, M' \in \mathbb{R}^{d\times d}$ and $0 \leq t \leq 1$ and $\operatorname{tr} M = \operatorname{tr} M' = 1$, then $\operatorname{tr}(tM + (1-t)M') = t\operatorname{tr}(M) + (1-t)\operatorname{tr}(M') = 1$.

To see that the set of density matrices is bounded in the 2-norm, let M be a density matrix. As U is PSD, $\|M\|_2$ is the maximum eigenvalue of M. Therefore, $\|M\|_2$ is at most the sum of the eigenvalues of M, which is $\operatorname{tr}(M) = 1$. So the set is also bounded (in the 2-norm). □

Density matrices are also useful for computing the norm of PSD matrices using the Frobenius inner product, defined by $\langle A, B\rangle = \operatorname{tr}(A^T B)$.

Lemma 7. *Let* $A \in \mathbb{R}^{d\times d}$ *be a PSD matrix. Then* $\max_{M\succeq 0, \operatorname{tr}(M)=1}\langle A, M\rangle = \|A\|$.

Proof. As A is PSD, there is an orthonormal basis of $\mathbb{R}^d$ consisting only of eigenvectors of A, and all eigenvalues are nonnegative. Let $\{v_1, v_2, \ldots, v_d\}$ be a basis of eigenvectors with $Av_j = \lambda_j v_j$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Also let $Q = [v_1 v_2 \cdots v_d]^T$ be the matrix whose jth row is v_j. Then

$$\lambda_1 = \|A\| \text{ and } A = Q^T \operatorname{diag}(\lambda_1, \ldots, \lambda_d)Q.$$

We first show that $\|A\|$ is attained (so the maximum is at least $\|A\|$). Let $M = Q^T \operatorname{diag}(1, 0, 0, \cdots, 0)Q$. Then M is symmetric, has only nonnegative eigenvalues, and its trace is 1. So it is a density matrix. We also have $AM = Q^T \operatorname{diag}(\lambda_1, \ldots, 0)Q$, so the trace of AM is λ_1. This proves

$$\max_{M\succeq 0, \operatorname{tr}(M)=1}\langle A, M\rangle \geq \|A\|\,.$$

For the other direction of the inequality, any density matrix M can be written as $U^T\Lambda U$ where Λ is a diagonal matrix with nonnegative entries and trace 1, and U is an orthogonal matrix. Then $\langle A, M\rangle = \operatorname{tr}(AU^T\Lambda U) = \operatorname{tr}(UAU^T\Lambda) = \langle UAU^T, \Lambda\rangle$. Now $B = UAU^T$ is a symmetric matrix with the same norm as A, since U and U^T are isometries. Moreover, it is PSD because $x^T Bx = (U^T x)^T A(U^T x) \geq 0$ for all $x \in \mathbb{R}^d$. We need to show that $\langle B, \Lambda\rangle$ is at most $\|B\|$. As Λ is a diagonal matrix whose trace is 1, $\langle B, \Lambda\rangle$ is a weighted sum of the diagonal entries of B. No diagonal entry of B can be larger than $\|B\|$, since Be_j is a vector whose jth component is B_{jj}, so $\|B\| \geq \|Be_j\| \geq B_{jj}$. Thus,

$$\max_{M\succeq 0, \operatorname{tr}(M)=1}\langle A, M\rangle = \langle B, \Lambda\rangle \leq \|B\| = \|A\|\,,$$

as needed. □

Note that a weighted sum of PSD matrices is PSD (weights are nonnegative). As the spectral centrality condition is a statement about the norm of a weighted sum of PSD matrices, we can apply Lemma 7 and the linearity of the inner product to rewrite it.

Definition 5 ((ϵ, λ)-spectral center)**.** The point $\nu \in \mathbb{R}^d$ is a (ϵ, λ)-spectral center of the set $\{z_1, \ldots, z_k\} \subset \mathbb{R}^d$ if,

$$\min_{w \in \mathcal{W}_{n,\epsilon}} \max_{M \succeq 0, \mathrm{tr}(M)=1} \sum_{i=1}^{k} w_i \langle (z_i - \nu)(z_i - \nu)^T, M \rangle \leq \lambda.$$

So ν is a (ϵ, λ)-spectral center of the data if and only if there is $w \in \mathcal{W}_{n,\epsilon}$ such that $\{z_1, \ldots, z_k\}$ satisfy the spectral centrality assumption with respect to w and ν.

Note that $\mathcal{W}_{n,\epsilon}$ is an intersection of closed convex sets, one of which is bounded. Hence $\mathcal{W}_{n,\epsilon}$ is compact. Moreover, the set of density matrices is compact and convex. The minimax theorem allows us to swap the min and the max in the definition of spectral center:

Theorem 5 (Corollary of the minimax theorem, [2])**.** *Let $\{z_1, \ldots, z_k\} \subset \mathbb{R}^d$ and $\nu \in \mathbb{R}^d$. Then*

$$\max_{M \succeq 0, \mathrm{tr}(M)=1} \min_{w \in \mathcal{W}_{n,\epsilon}} \sum_{i=1}^{k} w_i \langle (z_i - \nu)(z_i - \nu)^T, M \rangle = \min_{w \in \mathcal{W}_{n,\epsilon}} \max_{M \succeq 0, \mathrm{tr}(M)=1} \sum_{i=1}^{k} w_i \langle (z_i - \nu)(z_i - \nu)^T, M \rangle.$$

5.2 Equivalence of Centrality

We're now ready to show that the two notions of center are "equivalent". We do this with particular choices of constants.

First, we show that if a point is a $(0.3, \lambda)$-spectral center, then it is a $(0.4, 100\lambda)$-combinatorial center. We show the contrapositive:

Lemma 8 (Proposition 5.1 of [3])**.** *Let $\nu \in \mathbb{R}^d$ and $\{z_1, \ldots, z_k\}$ be a set. If there is a unit vector $v \in \mathbb{R}^d$ such that*

$$|\{1 \leq i \leq k : |\langle z_i - \nu, v \rangle| \geq 10\sqrt{\lambda}\}| > 0.4k,$$

then

$$\min_{w \in \mathcal{W}_{n,0.3}} \max_{M \succeq 0, \mathrm{tr}(M)=1} \sum_{i=1}^{k} w_i \langle (z_i - \nu)(z_i - \nu)^T, M \rangle > \lambda.$$

Proof. There are more than $0.4k$ points z_i such that $\langle z_i - \nu, v \rangle^2 \geq 100\lambda$. We write the inner product as a matrix product and then express it at a trace, to write it as an inner product of matrices: $v^T(z_i - \nu)(z_i - \nu)^T v = \mathrm{tr}(v^T(z_i - \nu)(z_i - \nu)^T v) = \mathrm{tr}((z_i - \nu)(z_i - \nu)^T vv^T) = \langle (z_i - \nu)(z_i - \nu)^T, vv^T \rangle$. Let $M = vv^T$; then M is PSD and its trace is $\mathrm{tr}(vv^T) = \mathrm{tr}(v^T v) = 1$. So M is a density matrix and

$$\langle (z_i - \nu)(z_i - \nu)^T, M \rangle \geq 100\lambda$$

for more than $0.4k$ points.

By Theorem 5, it is enough to show that $\min_{w \in \mathcal{W}_{n,0.3}} \sum_{i=1}^{k} w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \lambda$. As each $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle$ is independent of w, the minimizer w puts the largest weights on the points that have the smallest $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle$. So, for the the $0.7k$ indices with the lowest values of $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle$, the weight vector sets $w_i = \frac{1}{0.7k}$. As there are more than $0.4k$ indices with $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle \geq 100\lambda$ and $0.7k$ indices with $w_i = \frac{1}{0.7k}$, **strictly** more than $0.1k$ indices have both $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle \geq 100\lambda$ and $w_i = \frac{1}{0.7k}$. Therefore, $\sum_{i=1}^{k} w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle > 0.1k \frac{1}{0.7k} 100\lambda = \frac{100}{7}\lambda > 10\lambda$. □

In case $0.7k$ is not an integer, we don't get that there are strictly over $0.1k$ points which have both $w_i = \frac{1}{0.7k}$ and $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle \geq 100\lambda$. However, if k is sufficiently large (e.g., larger than 40) then there are strictly more than $0.07k$ points with both large weight and large inner product, and the result follows.

Next, we prove that any $(0.01, 0.01\lambda)$-combinatorial center is also a $(0.1, \lambda)$-spectral center. In this direction, we will need the following fact.

Lemma 9. *Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be compact and $f : X \times Y \to \mathbb{R}$ be continuous. The function $\phi(x) = \min_{y \in Y} f(x, y)$ is a continuous function of x.*

Proof. Note that as Y is compact and f is continuous, ϕ is defined. Moreover, it is real valued. Let $\epsilon > 0$ be arbitrary. We show that there is $\delta > 0$ such that if $x, x' \in X$ and $\|x - x'\| < \delta$ then $\phi(x) - \phi(x') < \epsilon$.

As $X \times Y$ is compact, f is uniformly continuous. Take $\delta > 0$ such that if (x, y) and (x', y') are in $X \times Y$ and $\|x - x'\|, \|y - y'\| < \delta$, then $|f(x, y) - f(x', y')| < \epsilon$. Let $x, x' \in X$ be arbitrary and assume $\|x - x'\| < \delta$. There is $y' \in Y$ such that $f(x', y') = \phi(x')$. By uniform continuity, $|f(x', y') - f(x, y')| < \epsilon$, so that

$$\phi(x) \leq f(x, y') < f(x', y') + \epsilon = \phi(x') + \epsilon.$$

Hence $\phi(x) - \phi(x') < \epsilon$.

Now, as $\|x' - x\| = \|x - x'\| < \delta$, we have $\phi(x') - \phi(x) < \epsilon$ and $\phi(x) - \phi(x') < \epsilon$, so that $|\phi(x) - \phi(x')| < \epsilon$. Hence ϕ is continuous on X. □

Lemma 10 (Proposition 5.2 of[3])**.** *Let $\nu \in \mathbb{R}^d$ and $\{z_1, \ldots, z_k\}$ be a set. If*

$$\min_{w \in \mathcal{W}_{n,0.1}} \max_{M \succeq 0, \mathrm{tr}(M)=1} \sum_{i=1}^{k} w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \lambda,$$

then there is a unit vector $v \in \mathbb{R}^d$ such that

$$|\{1 \leq i \leq k : |\langle z_i - \nu, v\rangle| \geq 0.1\sqrt{\lambda}\}| > 0.01k,$$

Proof. First, we show that any density matrix M is a limit of positive definite density matrices. Although this argument is not given in [3], it is necessary because we will sample from a Gaussian whose covariance is a density matrix, but the definition of a Gaussian distribution on $\mathbb{R}^d$ requires the covariance matrix to be positive definite. For a sketch of the proof, we first diagonalize the density matrix $M = Q^T \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)Q$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. If the last j eigenvalues are 0, then

$$M_n = Q^T \operatorname{diag}\left(\lambda_1 - \frac{\lambda_1}{n}, \lambda_2, \ldots, \lambda_{d-j}, \lambda_{d-j+1} + \frac{\lambda_1}{nj}, \ldots, \lambda_d + \frac{\lambda_1}{nj}\right) Q$$

is a positive definite density matrix for $n \geq 2$. The continuity of matrix multiplication shows that $M_n \to M$.

Now we can proceed. By Theorem 5, $\max_{M \succeq 0, \operatorname{tr}(M)=1} \min_{w \in \mathcal{W}_{n,0.1}} \sum_{i=1}^k w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \lambda$. As $\sum_{i=1}^k w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle$ is continuous in M and w, Lemma 9 shows that $\min_{w \in \mathcal{W}_{n,0.1}} \sum_{i=1}^k w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle$ is continuous in M. As any density matrix is a limit of positive definite density matrices, this shows that

$$\sup_{M \succ 0, \operatorname{tr}(M)=1} \min_{w \in \mathcal{W}_{n,0.1}} \sum_{i=1}^k w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \lambda.$$

So take a positive definite density matrix M such that $\min_{w \in \mathcal{W}_{n,0.1}} \sum_{i=1}^k w_i \langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \lambda$. As before, the minimizer puts a weight of $\frac{1}{0.9k}$ on the $0.9k$ indices with the smallest values of $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle$. So over $0.1k$ indices must have $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \lambda$; otherwise, all of the indices with nonzero weight have $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \leq \lambda$.

Let B be the set of indices such that $\langle (z_i - \nu)(z_i - \nu)^T, M\rangle > \lambda$. We show that if $v_M \sim \mathcal{N}(0, M)$, then with nonzero probability $v = \frac{v_M}{\|v_M\|}$ satisfies the given requirements. This will show that there is a unit vector with the requirements.

Let $g_i = \langle x_i - \nu, v_M\rangle$; then g_i is a Gaussian with mean 0 and variance $\sigma_i^2 = \operatorname{tr}(\mathbb{E}[v_M^T(x_i - \nu)(x_i - \nu)^T v_M]) = \operatorname{tr}((x_i - \nu)(x_i - \nu)^T \mathbb{E}[v_M v_M^T]) = \langle (x_i - \nu)(x_i - \nu)^T, M\rangle$. So if $i \in B$ then $\sigma_i^2 \geq \lambda$. We have $\Pr(|g_i| \geq 0.5\sqrt{\lambda}) \geq \Pr(|g_i| \geq 0.5\sigma_i) > 0.5$; to see this, we can write the probability as an integral, do a u-substitution to assume $\sigma_i = 1$, and then bound $\Pr(g_i \geq 0.5)$ below with the lower Riemann sum $\sum_{n=1}^{10} 0.3 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.5+0.3n)^2}{2}} > 0.25$.

Now, the expected value of $Y = |\{1 \leq i \leq k : |\langle z_i - \nu, v_M\rangle| \geq 0.5\sqrt{\lambda}\}| = |\{1 \leq i \leq k : |g_i| \geq 0.5\sqrt{\lambda}\}|$ is at least $0.5 \cdot |B| > 0.05k$, by linearity of expectation applied to the random variable $\chi_{\mathbb{R}\setminus(-0.5\sqrt{\lambda}, 0.5\sqrt{\lambda})}(g_i)$. By the Paley-Zygmund Inequality,

$$\Pr(Y \geq 0.01) \geq \Pr(Y \geq 0.2\mathbb{E}[Y]) \geq 0.64 \frac{0.0025k^2}{\mathbb{E}[Y^2]} \geq 0.64 \frac{0.0025k^2}{k^2} = 0.0016.$$

Finally, the authors cite the Borell-TIS inequality to show that that with probability at least 0.999, $\|v_M\| < 5$. By the pigeonhole principle there exists a $v_M \in \mathbb{R}^d$ such that $\|v_M\| < 5$ and $|\{1 \leq i \leq k : |\langle z_i - \nu, v_M\rangle| \geq 0.5\sqrt{\lambda}\}| \geq 0.01k$. For this v_M we have $|\{1 \leq i \leq k : |\langle z_i - \nu, v\rangle| \geq 0.1\sqrt{\lambda}\}| > 0.01k$ as needed. □

Thus, the two notions of center are equivalent.

5.3 Application to Sub-Gaussian Error Rate

Now, we can finish our exposition of mean estimation with sub-Gaussian error rate. We first show that every combinatorial center gives sub-Gaussian error rate, then show that we can find a combinatorial center by finding a spectral center.

Lemma 11. *Let D be a distribution with mean μ and covariance Σ, and let δ and n be as in the statement of Problem 1. Let $\{X_1, \dots, X_n\}$ be i.i.d. samples from D. Let $k = \lceil 800 \log(\frac{1}{\delta}) \rceil$. Group the indices $\{1, 2, \dots, n\}$ into k buckets $B_1, \dots, B_k$, each of size at least $\lfloor \frac{n}{k} \rfloor$. Let $\{Z_j : 1 \leq j \leq n\}$ be the random variables defined by $Z_j = \frac{1}{|B_j|} \sum_{i \in B_j} X_i$. If*

$$|\{1 \leq i \leq k : |\langle Z_i - \mu, v\rangle| \geq 3000 r_{\delta,N}(\Sigma)\}| \leq 0.01k \text{ for all unit vectors } v \in \mathbb{R}^d,$$

then any $(\epsilon, 600000 \cdot 3000^2 r_{\delta,n}(\Sigma)^2)$-combinatorial center $\widehat{\mu}$ achieves the sub-Gaussian error rate, $\epsilon < \frac{1}{2}$.

Proof. If $\widehat{\mu} - \mu = 0$ we are done. Otherwise, $v = \frac{\widehat{\mu} - \mu}{\|\widehat{\mu} - \mu\|}$ is a unit vector, and $|\{1 \leq i \leq k : |\langle Z_i - \mu, v\rangle| \geq 3000 r_{\delta,N}(\Sigma)\}| \leq 0.01k$ and $|\{1 \leq i \leq k : |\langle Z_i - \widehat{\mu}, v\rangle| \geq 600000 \cdot 3000 r_{\delta,N}(\Sigma)\}| \leq \epsilon k$. Therefore, there is $1 \leq j \leq k$ such that $|\langle Z_j - \mu, v\rangle| \leq 3000 r_{\delta,N}(\Sigma)$ and $|\langle Z_j - \widehat{\mu}, v\rangle| \leq \sqrt{600000} \cdot 3000 r_{\delta,N}(\Sigma)$. As v is a unit vector in the direction of $\widehat{\mu} - \mu$, the triangle inequality shows that

$$\|\widehat{\mu} - \mu\| = |\langle \widehat{\mu} - \mu, v\rangle| \leq |\langle Z_j - \widehat{\mu}, v\rangle| + |\langle Z_j - \mu, v\rangle| \leq (\sqrt{600000} + 1) 3000 r_{\delta,N}(\Sigma).$$

□

We remark that although this does satisfy the requirements of Problem 2, the constants are quite weak. The authors state that "no efforts have been given in optimizing them".

Finally, we show that the spectral sample reweighting problem can be used to solve this problem.

Theorem 6. *Let D be a distribution with mean μ and covariance Σ. If δ and n are as before, then an algorithm that solves the 60-spectral sample reweighting problem with $\epsilon = 0.1$ can be used to solve the mean estimation problem with sub-Gaussian error rate.*

Note that as $0.1 > \frac{1}{24}$ we cannot use Algorithm 2 (or at least, we can't use the analysis). Instead, we could use the estimator with optimal breakdown point.

Proof. To begin, construct $\{Z_1, \dots, Z_k\}$ as described in Lemma 11. By Lemma 5, with probability at least $1 - \delta$, the true mean is a $(0.01, 3000^2 r_{\delta,N}(\Sigma)^2)$-combinatorial center of $\{Z_1, \dots, Z_k\}$. Lemma 10 shows that μ is a $(0.1, 100 \cdot 3000^2 r_{\delta,N}(\Sigma)^2)$-spectral center of $\{Z_1, \dots, Z_k\}$. As the spectral centrality assumption is equivalent to Definition 5, there is $w \in \mathcal{W}_{n,0.1}$ such that $\left\|\sum_{i=1}^n w_i (x_i - \mu)(x_i - \mu)^T\right\| \leq 100 \cdot 3000^2 r_{\delta,N}(\Sigma)^2$. So a solution to the spectral sample reweighting problem (Problem 1) outputs $w' \in \mathcal{W}_{n,0.1} \subset \mathcal{W}_{n,0.3}$ and $\widehat{\mu}$ such that $\left\|\sum_{i=1}^n w'_i (x_i - \widehat{\mu})(x_i - \widehat{\mu})^T\right\| \leq 6000 \cdot 3000^2 r_{\delta,N}(\Sigma)^2$, and hence

$$\min_{w \in \mathcal{W}_{n,0.3}} \max_{M \succeq 0, \operatorname{tr}(M)=1} \sum_{i=1}^k w'_i \langle (Z_i - \widehat{\nu})(z_i - \widehat{\nu})^T, M\rangle \leq 6000 \cdot 3000^2 r_{\delta,N}(\Sigma)^2.$$

By Lemma 8, $\widehat{\mu}$ is a $(0.4, 600000 \cdot 3000^2 r_{\delta,N}(\Sigma)^2)$-combinatorial center. By 11, we are done.

□

A few remarks: in this paper, we said that an algorithm that solves the spectral sample reweighting problem outputs $w' \in \mathcal{W}_{n,0.1} \subset \mathcal{W}_{n,0.3}$. This is not true if the problem is phrased as it is in [3]; the requirement there is that the output lies in $\mathcal{W}_{n,3\epsilon}$. However, this proof (replacing $w' \in \mathcal{W}_{n,0.1} \subset \mathcal{W}_{n,0.3}$ with $w' \in \mathcal{W}_{n,0.3}$) shows that a solution to the problem as stated in [3] solves the mean estimation with sub-Gaussian error rates. $\in \mathcal{W}_{n,\epsilon}$

6 Conclusion

We have seen how to solve the spectral sample reweighting problem, and how it connects to the two mean estimation problems. In addition to the details of the particular problems, we've seen how randomized algorithms and regularity conditions can be used in real-world robust estimation. In future, some areas to consider are whether the constants in Section 5 can be optimized, and whether a similar analysis can be done for robust covariance estimation (to begin, whether it can be done for robust covariance estimation of Gaussian distributions).

References

[1] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.

[2] D. P. Bertsekas. Nonlinear programming, 1999.

[3] S. B. Hopkins, J. Li, and F. Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization, 2021.

[4] X. Liu, W. Kong, S. Kakade, and S. Oh. Robust and differentially private mean estimation, 2021.

[5] G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector, 2017.