A More Technical Preliminaries

For vectors $u, v$, we denote the entrywise product by $u \odot v$; that is, the vector such that $w_i = u_i \cdot v_i$ for each $i$. For any symmetric matrix $A \in \mathbb{R}^{d \times d}$, let $\exp(A)$ denote the matrix exponential of $A$.

Definition A.1 (total variation distance). For probability distributions $p, q$, the total variation distance is defined as $\text{TV}(p, q) = \sup_{E} |p(E) - q(E)| = \frac{1}{2} \|p - q\|_1$, where the supremum is over the set of measurable events.

We write $\mathcal{U}_n$ for the uniform distribution over $[n]$.

B Technical Details of Section 2

B.1 Full Proof of Theorem 2.1

To prove the main theorem (Theorem 2.1) of Section 2, we first claim the following set of guarantees provided by Algorithm 1.

Lemma B.1 (analysis of filter). Let $\epsilon \in (0, 1/10]$ and $\{x_i\}_{i=1}^n$ be $n$ points in $\mathbb{R}^d$. Suppose there exists $\nu \in \mathbb{R}^d$ and $w \in \mathcal{W}_{n, \epsilon}$ such that

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

for some $\lambda > 0$. Then, given $\{x_i\}_{i=1}^n$, a failure rate $\delta$ and $\rho$ such that $\rho \geq \tau_i^{(t)}$ for all $i$ and $t$, Algorithm 1 finds $w' \in \mathcal{W}_{n, \epsilon}$ and $\nu' \in \mathbb{R}^d$ such that

$$\sum_{i=1}^n w_i' (x_i - \nu') (x_i - \nu')^T \preceq 60\lambda I,$$  \hspace{1cm} (B.1)

with probability at least $1 - \delta$.

The algorithm terminates in $T = O(\rho \epsilon / \lambda)$ iterations. Further, if $T = O(\text{poly}(n, d))$, then each iteration takes $O(nd \log(1/\delta))$ time.

We first see how to prove Theorem 2.1 via Lemma B.1. Note that it requires to bound the width parameter $\rho$. To ensure the condition $\rho \geq \tau_i^{(t)}$ for all $i$ and $t$, observe that as $\|v^{(t)}\| = 1$, we have

$$\tau_i^{(t)} = \left< v^{(t)}, x_i - \nu^{(t)} \right>^2 \leq \|x_i - \nu^{(t)}\|^2.$$

Also, since $\nu^{(t)}$ is a convex combination of $\{x_i\}_{i=1}^n$, we can set $\rho$ to be the squared diameter of the input data $\{x_i\}_{i=1}^n$. As the first step, we show that a $(1 - 2\epsilon)$ fraction of the points lie within a ball of radius $\sqrt{d\lambda / \epsilon}$ under the spectral centrality condition. Then a (folklore) pruning procedure can be used to extract such set.

Lemma B.2 (diameter bound). Let $\{x_i\}_{i=1}^n$ be $n$ points in $\mathbb{R}^d$. Suppose there exists $\nu \in \mathbb{R}^d$ and $w \in \mathcal{W}_{n, \epsilon}$ such that $\sum_{i=1}^n w_i (x_i - \nu) (x_i - \nu)^T \preceq \lambda I$ for some $\lambda > 0$ and $\epsilon \in (0, 1/2)$. Then there exists a ball of radius $\sqrt{d\lambda / \epsilon}$ that contains at least $r = (1 - 2\epsilon)n$ points of $\{x_i\}_{i=1}^n$.

Proof. We show that there exists a ball of radius $\sqrt{d\lambda / \epsilon}$ that contains at least $(1 - 3\epsilon)n$ points. Note that the spectral centrality condition $\sum_{i=1}^n w_i (x_i - \nu) (x_i - \nu)^T \preceq \lambda I$ implies that

$$\sum_{i=1}^n \text{Tr} \left( w_i (x_i - \nu)(x_i - \nu)^T \right) \preceq d \lambda.$$

By the cyclic property of trace, we get

$$\sum_{i=1}^n w_i \|x_i - \nu\|^2 \leq d \lambda.$$
Therefore, by Markov’s inequality,
\[ \Pr_{i\sim w} \left( \| x_i - \nu \|_2^2 \geq d\lambda/\epsilon \right) \leq \epsilon, \]  
(\text{B.2})
where \( i \sim w \) denotes \( i \) drawn from the discrete distribution defined by \( w \). Observe that since \( \mathcal{W}_{n, \epsilon} \) is the convex hull of all uniform distributions over a subset of size \((1 - \epsilon)n\), we have \( \| w - U_n \|_1 \leq 2\epsilon \). Thus, \( \text{TV}(w, U_n) \leq \epsilon \). Hence, using the definition of total variation distance, (\text{B.2}) implies that
\[ \Pr_{i\sim U_n} \left( \| x_i - \nu \|_2^2 \geq d\lambda/\epsilon \right) \leq 2\epsilon, \]  
(\text{B.3})
as desired. \( \square \)

**Lemma B.3** (folklore; see [21]). Let \( \epsilon < 1/2 \) and \( \delta > 0 \). Let \( S \subset \mathbb{R}^d \) be a set of \( n \) points. Assume there exists a ball \( B \) of radius \( r \) and a subset \( S' \subseteq S \) such that \( |S'| \geq (1 - \epsilon)n \) and \( S' \subset B \). Then there is an algorithm \( \text{PRUNE}(S, r, \delta) \) that runs in time \( O(nd \log (1/\delta)) \) and with probability \( 1 - \delta \) outputs a set \( R \subseteq S \) so that \( S' \subseteq R \), and \( R \) is contained in a ball of radius \( 4r \).

Using the lemmas above, we can immediately prove the main theorem.

**Proof of Theorem 2.1.** Given \( S = \{ x_i \}_{i=1}^n \), \( \lambda \) and \( \epsilon \), we first run the \( \text{PRUNE}(S, r, \delta/2) \) algorithm, with \( r = \sqrt{d\lambda/\epsilon} \). By Lemma B.2, the spectral centrality condition (\text{\dagger}) implies there exists a ball of radius \( r \) containing at least \((1 - 2\epsilon)n\) points of \( S \). Therefore, Lemma B.3 guarantees that it will return a set \( R \subseteq S \) of at least \((1 - 2\epsilon)n\) points contained in a ball of radius \( 4r \). Hence by Lemma B.1, given \( R, \rho = 16d\lambda/\epsilon \) and failure rate \( \delta/2 \), Algorithm 1 finds \( w' \in \mathcal{W}_{R, \epsilon} \) and \( \nu' \in \mathbb{R}^d \) such that
\[ \sum_{i \in R} w'_i (x_i - \nu') (x_i - \nu')^T \leq 60\lambda I, \]
with probability at least \( 1 - \delta/2 \). Let \( w''_i = w'_i \) if \( i \in R \) and \( w''_i = 0 \) otherwise. since \( \frac{1}{1 - 2\epsilon} \leq \frac{1}{1 - \epsilon} \) for \( \epsilon < 1/3 \), we have \( w'' \in \mathcal{W}_{n, 3\epsilon} \). Moreover, \( \sum_{i=1}^n w''_i (x_i - \nu') (x_i - \nu')^T \leq 60\lambda I \), as desired.

The overall procedure succeeds with probability at least \( 1 - \delta \) by a union bound, since Algorithm 1 and \( \text{PRUNE} \) are both set up to have a failure rate at most \( \delta/2 \). Now for the run-time, \( \text{PRUNE}(S, r, \delta) \) takes \( O(nd \log (1/\delta)) \) by Lemma B.3. Moreover, by Lemma B.1, Algorithm 1 runs in time \( \widetilde{O}(nd \log (1/\delta) \cdot T) \) time, with \( T = O(p \rho \epsilon / \lambda) \) being the iteration count. Since \( \rho = 16d\lambda/\epsilon \), we have \( T = O(d) \), and this immediately yields the desired runtime. \( \square \)

**Analysis via regret minimization** Now it remains to analyze Algorithm 1, proving Lemma B.1.

We will cast the algorithm under the framework of regret minimization using multiplicative weights update (MWU). To see that, we consider \( \{ x_i \}_{i=1}^n \) as the set of actions, \( w^{(t)} \) as our probability distribution over the actions at time \( t \), and we receive a loss vector \( \tau^{(t)} \) each round. The weights are updated in a standard fashion. Further, the weights must lie in the constraint set \( \mathcal{W}_{n, \epsilon} \) and thus the projection step. (Note that the algorithm is implementing both the player and the adversary.) The following is a classic regret bound of MWU for the online linear optimization problem.

**Lemma B.4** (regret bound [4]). Suppose \( \rho \geq \gamma_i^{(t)} \) for every \( t \) and \( i \). Then for any weight \( w \in \mathcal{W}_{n, \epsilon} \), Algorithm 1 satisfies that
\[ \frac{1}{T} \sum_{t=1}^T \left< w^{(t)}, \tau^{(t)} \right> \leq \frac{1}{T} (1 + \eta) \sum_{t=1}^T \left< w, \tau^{(t)} \right> + \rho \cdot \frac{\text{KL}(w \| w^{(t)})}{T\eta}, \]  
\[ \text{(B.4)} \]
where any choice of step size \( \eta \leq 1/2 \).

In addition, we claim the following lemma and delay its proof to the appendix (Lemma B.7).

**Lemma B.5.** Under the centrality promise (\text{\dagger}), for any \( w' \in \mathcal{W}_{n, \epsilon} \),
\[ \| \nu - \nu(w') \| \leq \frac{1}{1 - 2\epsilon} \left( \sqrt{2\lambda} + \sqrt{\epsilon \| M(w') \|} \right), \]  
\[ \text{(B.5)} \]
where \( \nu(w') = \sum_i w'_i x_i \) and \( M(w') = \sum_i w'_i (x_i - \nu(w')) (x_i - \nu(w'))^T \).
This type of inequality is generally known as the spectral signature lemmas from the recent algorithmic robust statistics literature; see [35, 18].

With these technical ingredients, we are now ready to analyze the algorithm.

**Proof of Lemma B.1.** Notice first that since \( v^{(t)} \) is a 7/8-approximate largest eigenvector of \( M^{(t)} = \sum w^{(t)}_i (x_i - \nu^{(t)}) (x_i - \nu^{(t)})^T \), then for all \( t \),

\[
\sum_i w^{(t)}_i \tau^{(t)}_i = \sum_i w_i \left( v^{(t)}^T x_i - \nu^{(t)} \right)^2 = v^{(t)}^T M^{(t)} v^{(t)} \geq \frac{7}{8} \| M^{(t)} \|_2^2. \tag{B.6}
\]

Let \( w \) be the good weights that satisfies our centrality promise (\( \dagger \)). Summing over the \( T \) rounds and applying the the regret bound (Lemma B.4), we obtain that

\[
\frac{7}{8T} \sum_{t=1}^T \| M^{(t)} \|_2^2 \leq \frac{1}{T} \sum_{t=1}^T \left( w^{(t)}, \tau^{(t)} \right) \leq (1 + \eta) \frac{1}{T} \sum_{t=1}^T \left( w, \tau^{(t)} \right) + \frac{\eta \cdot \text{KL}(w \| w^{(1)})}{T}. \tag{B.7}
\]

The KL term can be bounded because \( w \) and \( w^{(1)} \) are both close to uniform. Indeed, it is a simple calculation to verify that \( \text{KL}(w \| w^{(1)}) \leq 5 \epsilon \), using the fact \( w_i \leq 1/(1 + \epsilon) n \) (Lemma B.8). Plugging in \( \eta = 1/2 \), we get

\[
\frac{7}{8T} \sum_{t=1}^T \| M^{(t)} \|_2^2 \leq \frac{3}{2T} \sum_{t=1}^T \left( w, \tau^{(t)} \right) + \frac{10 \kappa \rho}{T}. \tag{B.8}
\]

Our eventual goal is to bound this by \( O(\lambda) \). Note that the second term is easy to control—just set \( T = \Omega(\rho \epsilon / \lambda) \), and this will determine the iteration count and thus the runtime.

The remaining is mostly tedious calculations to bound the first term. The reader can simply skip forward to (B.13). For those interested: we proceed by expanding the first term on the right-hand side,

\[
\frac{3}{2T} \sum_{t=1}^T \left( w, \tau^{(t)} \right) = \frac{3}{2T} \sum_{t=1}^T \sum_{i=1}^n w_i \left( x_i - \nu^{(t)} \right)^2 \tag{B.9}
\]

where (B.8) is by the definition that \( \tau^{(t)}_t = \left( \nu^{(t)}, x_i - \nu^{(t)} \right)^2 \), (B.9) uses the definition of \( \nu^{(t)} \), (B.10) follows from the spectral centrality assumption (\( \dagger \)), and (B.11) is by the fact that \( \| \nu^{(t)} \| = 1 \). Since

\[
\nu^{(t)} = \sum_{i=1}^n w_i^{(t)} x_i,
\]

we can apply Lemma B.5 to bound \( \| \nu - \nu^{(t)} \| \) and it follows that

\[
\frac{3}{2T} \sum_{t=1}^T \left( w, \tau^{(t)} \right) \leq \frac{3}{2 \lambda} + \frac{3}{2T} \left( \sum_{i=1}^T \frac{25}{2} \lambda + \frac{1}{3} \left( M^{(t)} \right) \right), \tag{B.12}
\]

for \( \epsilon \leq 1/10 \). Plugging the bound into (B.11), we obtain

\[
\frac{3}{2T} \sum_{t=1}^T \left( w, \tau^{(t)} \right) \leq \frac{3}{2 \lambda} + \frac{3}{2T} \left( \sum_{i=1}^T \left( M^{(t)} \right) \right) = \frac{81}{4} \lambda + \frac{1}{2T} \sum_{t=1}^T \left( M^{(t)} \right) \tag{B.13}
\]

Finally, substituting this back into (B.7), we see that

\[
\frac{7}{8T} \sum_{t=1}^T \left( M^{(t)} \right) \leq \frac{81}{4} \lambda + \frac{1}{2T} \sum_{t=1}^T \left( M^{(t)} \right) + \frac{10 \kappa \rho}{T}. \tag{B.13}
\]
Now if we set $T = 10\rho \e / \lambda$, then the last term is $\lambda$. Rearranging yields that $\frac{1}{T} \sum_{t=1}^{T} \| M^{(t)} \|_2 \leq 60\lambda$. This shows that within $T = O(\rho \e / \lambda)$ iterations we have achieved our goal (B.1).

Now it remains to argue the cost of each iteration. For approximating the largest eigenvector, the well-known power method computes a constant-approximation in $O(n d \log(1/\alpha))$ time with a failure probability at most $\alpha$ [29]. We set $\alpha = \delta / T$, and an application of union bound implies that all the $T$ calls to the power method jointly succeed with probability at least $1 - \delta$. This gives a total run-time of $O(n d \log(1/\delta))$, since $T = O(\text{poly}(n, d))$, and bounds the overall failure probability of the algorithm by $\delta$. Finally, we remark that the KL projection onto $\mathcal{W}_{n, \epsilon}$ can be computed exactly in $O(n)$ time, by the deterministic procedures in [24, 45]. This completes the run-time analysis.  

### B.2 Proof of a Spectral Signature

Towards a proof of Lemma B.5, we need the following technical lemma.

**Lemma B.6.** Let $\{x_i\}_{i=1}^{n}$ be $n$ points in $\mathbb{R}^d$. Suppose there exists $\nu \in \mathbb{R}^d$ and a set of good weights $w_0 \in \mathcal{W}_{n, \epsilon}$ such that

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

(B.14)

Then there exists a set $G \subseteq [n]$ of size $(1 - \epsilon)n$ such that

$$
\frac{1}{(1 - \epsilon)n} \sum_{i \in G} (x_i - \nu) (x_i - \nu)^T \leq \lambda I.
$$

(B.15)

**Proof.** Define $w(T)$ to be the uniform distribution over a subset $T \subseteq [n]$ of data. Let $\mathcal{T}$ denote the collection of all subsets of size $(1 - \epsilon)n$. Observe that the set of good weights $\mathcal{W}_{n, \epsilon}$ is simply the convex hull of $\{w(T) : T \in \mathcal{T}\}$. Thus, for each $i$, we can rewrite $w_i = \sum_{T \in \mathcal{T}} \alpha_T w(T)_i$ for some distribution $\alpha$ over $\mathcal{T}$. Then we get that

$$
\sum_{i=1}^{n} w_i (x_i - \nu) (x_i - \nu)^T = \sum_{i=1}^{n} \sum_{T \in \mathcal{T}} \alpha_T w(T)_i (x_i - \nu) (x_i - \nu)^T
$$

$$
= \sum_{T \in \mathcal{T}} \alpha_T \sum_{i=1}^{n} w(T)_i (x_i - \nu) (x_i - \nu)^T
$$

$$
= \sum_{T \in \mathcal{T}} \alpha_T A_T,
$$

where $A_T = \frac{1}{(1 - \epsilon)n} \sum_{i \in T} (x_i - \nu) (x_i - \nu)^T$. It follows that the spectral centrality condition (B.14) is equivalent to

$$
\sum_{T \in \mathcal{T}} \alpha_T A_T \leq \lambda I.
$$

Therefore, there must exist a $G \in \mathcal{T}$ such that $A_G \leq \lambda I$, as we desired in (B.15).

**Lemma B.7.** Let $\{x_i\}_{i=1}^{n}$ be $n$ points in $\mathbb{R}^d$. Suppose there exists $\nu \in \mathbb{R}^d$ and a set of good weights $w_0 \in \mathcal{W}_{n, \epsilon}$ such that

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

(B.16)

for some $\lambda > 0$. Then for any $w' \in \mathcal{W}_{n, \epsilon}$,

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

(B.17)

where $\nu(w') = \sum_i w_i x_i$ and $M(w') = \sum_i w_i (x_i - \nu(w')) (x_i - \nu(w'))^T$.

The lemma and its proof strategy is similar to the spectral signature lemma in robust statistics and is now somewhat standard in the literature; see, e.g., [21, 35].
Proof. First, by Lemma B.6, there exists a set $G$ of data of size $(1 - \epsilon)n$ such that
\[
\frac{1}{(1 - \epsilon)n} \sum_{i \in G} (x_i - \nu) (x_i - \nu) \succeq \lambda I. \tag{B.18}
\]
Let $\mu_G = \frac{1}{(1 - \epsilon)n} \sum_{i \in G} x_i$ and $B = [n] \setminus G$.

Next, to bound $\|\nu - \nu(w')\|$, we note
\[
\|\nu(w') - \nu\|^2 = \sum_i w_i' (\nu(w') - \nu, x_i - \nu)
= \sum_{i \in G} \frac{1}{(1 - \epsilon)n} (\nu(w') - \nu, x_i - \nu) + \sum_{i \in G} \left( w_i' - \frac{1}{(1 - \epsilon)n} \right) (\nu(w') - \nu, x_i - \nu) + \sum_{i \in B} w_i' (\nu(w') - \nu, x_i - \nu) \tag{B.19}
\]
We bound the three terms respectively as follows.

(i) For the first term, by Cauchy-Schwarz,
\[
\sum_{i \in G} \frac{1}{(1 - \epsilon)n} (\nu(w') - \nu, x_i - \nu) = \langle \nu(w') - \nu, \mu_G - \nu \rangle \leq \|\nu(w') - \nu\| \cdot \|\mu_G - \nu\|.
\]
By Jensen’s inequality and since $A_G \succeq \lambda I$, we have for all unit $u$,
\[
\langle \mu_G - \nu, u \rangle^2 = \left( \frac{1}{(1 - \epsilon)n} \sum_{i \in G} x_i - \nu, u \right)^2 \leq \frac{1}{(1 - \epsilon)n} \sum_{i \in G} (x_i - \nu, u)^2 \leq \lambda.
\]
Thus, $\|\mu_G - \nu\| \leq \sqrt{\lambda}.$

(ii) For the second term, let $\alpha_i = w_i' - 1/(1 - \epsilon)n$. Then
\[
\left( \sum_{i \in G} \alpha_i (\nu(w') - \nu, x_i - \nu) \right)^2 \leq \left( \sum_{i \in G} (1 - \epsilon)n \alpha_i^2 \right) \cdot \sum_{i \in G} \frac{1}{(1 - \epsilon)n} (\nu(w') - \nu, x_i - \nu)^2
\leq \left( \sum_{i \in G} (1 - \epsilon)n \alpha_i^2 \right) \cdot \lambda \cdot \|\nu(w') - \nu\|^2 \tag{B.20}
\leq \left( \sum_{i \in G} |\alpha_i| \right) \cdot \lambda \cdot \|\nu(w') - \nu\|^2 \tag{B.21}
\leq 2 \epsilon \lambda \cdot \|\nu(w') - \nu\|^2, \tag{B.22}
\]
where (B.20) is by the covariance bound that $A_G \succeq \lambda I$, (B.21) follows since $|(1 - \epsilon)n \alpha_i| \leq 1$, and (B.22) since $\sum_{i=1}^n |\alpha_i| \leq \epsilon/(1 - \epsilon) \leq 2 \epsilon$.

(iii) For the third term, we have
\[
\sum_{i \in B} w_i' (\nu(w') - \nu, x_i - \nu) = \sum_{i \in B} w_i' (\nu(w') - \nu, x_i - \nu(w')) + \left( \sum_{i \in B} w_i \right) \|\nu(w') - \nu\|^2
\leq \sum_{i \in B} w_i' (\nu(w') - \nu, x_i - \nu(w')) + \epsilon \cdot \|\nu(w') - \nu\|^2
\]
By Cauchy-Schwarz,
\[
\left( \sum_{i \in B} w_i' (\nu(w') - \nu, x_i - \nu(w')) \right)^2 \leq \left( \sum_{i \in B} w_i' \right) \left( \sum_{i \in B} w_i' (\nu(w') - \nu, x_i - \nu(w'))^2 \right)
\leq \epsilon \cdot \sum_{i=1}^n w_i' (\nu(w') - \nu, x_i - \nu(w'))^2
= \epsilon \cdot (\nu(w') - \nu)^\top M(w') (\nu(w') - \nu)
\leq \epsilon \cdot \|M(w')\|_2 \cdot \|\nu(w') - \nu\|^2
\]

Substituting the three bounds back into (B.19) immediately yields the result (B.17).

### B.3 Proof of the KL Divergence Bound

**Lemma B.8.** Let \( p \in \mathcal{W}_{n, \varepsilon} \) and \( q \) be the uniform distribution over \( n \) points. Then \( KL(p||q) \leq 5\varepsilon \).

**Proof.** The lemma follows from direct calculations. By definition of KL divergence,

\[
KL(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \\
\leq \sum_i \frac{1}{(1-\varepsilon)n} \log \frac{1}{(1-\varepsilon)} \\
= \frac{1}{1-\varepsilon} \log \frac{1}{1-\varepsilon} \\
\leq 5\varepsilon.
\]

where the last inequality holds when \( 0 < \varepsilon \leq 1/2 \).

### B.4 Further Technical Comments

#### Faster algorithm

Under the same assumptions, the spectral sample reweighing problem can be solved in \( \tilde{O}(nd \log(1/\delta)) \) time, by adapting a matrix multiplicative weight scheme, due to Dong, Hopkins and Li [21]. The algorithm and its analysis generally follow from the proofs therein. The details can be found in Appendix G.

As we can show in the corresponding sections, applying this procedure directly match the fastest known algorithms for both robust and heavy-tailed settings.

#### Gradient descent analysis

As we argued, Algorithm 1 is essentially an online linear optimization scheme, with the objective of minimizing \( \sum_{t=1}^T \langle w(t), \tau^t \rangle \). It is known that the multiplicative weights rule employed here can be seen an entropic mirror descent update [42]. Therefore, it is natural to ask whether an additive update/gradient descent procedure would solve the problem as well. In Appendix H, we provide such an analysis (Theorem H.3). More importantly, the resulting scheme is equivalent to the gradient descent algorithm analyzed by [9], and our analysis improves upon the iteration complexity from their work (in the concrete settings of robust mean estimation, under bounded second moment and sub-gaussian distributions).

### C Solving Robust Mean Estimation

We now apply Algorithm 1 for the robust mean estimation problem. We focus on the bounded second moment distributions, where Algorithm 1 can be invoked in a black-box fashion. A slight variant of it can be used for the sub-gaussian setting, where we achieve a more refined analysis; see Appendix F.

The problem is formally defined below.

**Definition C.1** (robust mean estimation). Given a distribution \( D \) over \( \mathbb{R}^d \) with bounded covariance and a parameter \( 0 \leq \varepsilon < 1/2 \), the adversary draws \( n \) i.i.d. samples \( D \), inspects the samples, then removes at most \( \varepsilon n \) points and replaces them with arbitrary points. We call the resulting dataset \( \varepsilon \)-corrupted (by an adaptive adversary).

The goal is to estimate the mean of \( D \) only given the \( \varepsilon \)-corrupted set of samples.

Using a meta-algorithm for approximating the spectral sample reweighing problem, we will show the following. In particular, using Algorithm 1 matches the run-time and statistical guarantee of the original FILTER algorithm.
Theorem C.1 (robust mean estimation via sample reweighing). Let $D$ be a distribution over $\mathbb{R}^d$ with mean $\mu$ and covariance $\Sigma \preceq \sigma^2 I$ and $\epsilon \leq 1/10$. Given an $\epsilon$-corrupted set of $n = \Omega(d \log d/\epsilon)$ samples, there is an algorithm that runs in time $O(n d^2)$ that with constant probability outputs an estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \leq O(\sigma \sqrt{\epsilon})$.

Further, the algorithm is via a black-box application of Algorithm 1, which can be replaced by any algorithm solving the spectral sample reweighing problem (Definition 2.1).

Information-theoretically, Theorem C.1 is near optimal. It is known that the sample complexity of $d \log d/\epsilon$ is tight, only up to the log factor. The estimation error $O(\sqrt{\epsilon})$ is tight up to constant factor.

Our analysis requires a set of deterministic conditions to hold for the input, which follow from Lemma A.18 of [16]. This is meant to obtain the desired spectral centrality condition and to bound the final estimation error.

Lemma C.2 (deterministic conditions [16]). Let $S$ be an $\epsilon$-corrupted set of $\Omega(d \log d/\epsilon)$ samples from $D$ with mean $\mu$ and covariance $\Sigma \preceq I$. With high constant probability, $S$ contains a subset $G$ of size at least $(1-\epsilon)n$ such that

$$\|\mu - \mu_G\| \leq O(\sqrt{\epsilon})$$

(C.1)

$$\left\| \frac{1}{|G|} \sum_{i \in G} (x_i - \mu_G) (x_i - \mu_G)^T \right\|_2 \leq O(1),$$

(C.2)

where $\mu_G = \frac{1}{|G|} \sum_{i \in G} x_i$.

We now prove the main result of this section—using the meta-algorithm to solve the robust mean estimation problem. Observe that it suffices to prove the theorem with $\sigma^2 = 1$. Without loss of generality, we can first divide every input sample by $\sigma$, execute the algorithm and then multiply the output by $\sigma$.

Proof of Theorem C.1. First, we check that the centrality promise (i) is satisfied. This would ensure that we are in the setting of the spectral sample reweighing problem so that the meta-algorithm applies. Assume the conditions from Lemma C.2. Then suppose we let $w_i = 1/|G|$ if $x_i \in G$ and $w_i = 0$ otherwise, so we have that $w \in W_{n,\epsilon}$, and let $\nu = \mu_G$ and $\lambda = O(1)$. Observe that (C.2) is exactly the spectral centrality condition (i). Then we can apply Theorem 2.1 and obtain that the algorithm will find $\nu' \in \mathbb{R}^d$ and $w' \in W_{n,3\epsilon}$ such that

$$M(w') := \sum_{i=1}^n w_i' (x_i - \nu') (x_i - \nu')^T \preceq O(1) \cdot I$$

Furthermore, by definition of the algorithm, $\nu'$ is a weighted average of the points $\{x_i\}_{i=1}^n$; that is, $\nu' = \nu(w') = \sum_{i=1}^n w_i' x_i$. This allows us again to apply the spectral signature lemma. In particular, Lemma C.5 implies

$$\|\mu_G - \nu'\| \leq \frac{1}{1 - 6\epsilon} \left( \sqrt{6\epsilon \lambda} + \sqrt{3\epsilon \|M(w')\|} \right) = O(\sqrt{\epsilon})$$

since $\lambda = O(1)$ and $\|M(w')\| = O(1)$. Finally, by triangle inequality and (C.1),

$$\|\mu - \nu'\| \leq \|\mu_G - \nu'\| + \|\mu - \mu_G\| \leq O(\sqrt{\epsilon}).$$

Therefore, the output $\nu'$ estimates the true mean up to an error of $O(\sqrt{\epsilon})$, as desired.

Finally, the run-time guarantee follows directly from the statement of Theorem 2.1, since we apply the meta-algorithm in a black-box fashion. This completes the proof.

Other algorithms. To improve the computational efficiency, applying the same argument and using the matrix multiplicative weight algorithm (Theorem G.1), we can obtain a near linear time algorithm, which matches the fastest known algorithm for robust mean estimation [21, 10].

Corollary C.3 (faster robust mean estimation [21]). Let $D$ be a distribution over $\mathbb{R}^d$ with mean $\mu$ and covariance $\Sigma \preceq \sigma^2 I$ and $\epsilon$ be a sufficiently small constant. Given an $\epsilon$-corrupted set of $n = \Omega(d \log d/\epsilon)$ samples, there is a matrix multiplicative update algorithm that runs in time $O(n d)$ and with constant probability computes an estimate of error $O(\sigma \sqrt{\epsilon})$. 

18
Since $\lambda = O(1)$ in the robust mean estimation problem under bounded covariance (Lemma C.2), our analysis of the gradient descent algorithm (Theorem H.3) implies the following.

**Corollary C.4** (robust mean estimation via gradient descent). Let $D$ be a distribution over $\mathbb{R}^d$ with mean $\mu$ and covariance $\Sigma \preceq \sigma^2 I$ and $c$ be a sufficiently small constant. Given an $c$-corrupted set of $n = \Omega(d \log d/c)$ samples, there is a gradient-descent based algorithm that computes an estimate of error $O(\sigma \sqrt{c})$ with constant probability in $\tilde{O}(n^2 d/c)$ iterations.$^{2}$

A variant of the gradient descent-based algorithm can be used for robust mean estimation in the sub-gaussian setting as well; see Appendix H.2.

### C.1 Proof of the Spectral Signature

**Lemma C.5.** Let $\{x_i\}_{i=1}^n$ be $n$ points. Suppose there exists a subset $G \subset [n]$ of size $(1 - \epsilon)$ such that $\frac{1}{|G|} \sum_{i \in G} (x_i - \mu_G) (x_i - \mu_G)^\top \preceq \lambda I$ for some $\lambda > 0$, where $\mu_G = \frac{1}{|G|} \sum_{i \in G} x_i$. Then for any $w \in \mathcal{W}_{n,c}$,

$$
\|\mu_G - \mu(w)\| \leq \frac{1}{1 - 2\epsilon} \left( \sqrt{2\epsilon \lambda} + \sqrt{\epsilon \|M(w)\|} \right). 
$$

(C.3)

**Proof.** The proof follows from the same argument of Lemma B.7, with $\nu = \mu_G$. Observe that the first term in (B.19) becomes $\frac{1}{|G|} \sum_{i \in G} \langle \mu(w) - \nu, x_i - \nu \rangle$, which equals 0 when $\nu = \mu_G$, shaving the $\sqrt{\lambda}$ term in the final bound. $\Box$

### D Technical Details of Section 3

#### D.1 Proof of the Duality Theorems

To pave way for the proofs, a key observation, first made by [8], is that the left-side of (spectral center) is an SDP objective. (This is because it is simply minimizing the maximum eigenvalue of $\sum_i w_i (x_i - \nu)(x_i - \nu)^\top$.) And strong duality allows us to swap the min and max, so

$$
\min_{w \in \mathcal{W}_{k,c}} \max_{M} \sum_{i=1}^k w_i \langle (x_i - \nu)(x_i - \nu)^\top, M \rangle = \max_{M} \min_{w \in \mathcal{W}_{k,c}} \sum_{i=1}^k w_i \langle (x_i - \nu)(x_i - \nu)^\top, M \rangle,
$$

(D.1)

where the maximization is over the set of density matrices. Using this, we prove the following two propositions, showing (by contrapositives) that the two notions of centrality are equivalent. The constants in the statements are chosen only to serve the purpose of heavy-tailed mean estimation, and they can be tweaked easily by the same arguments.

We are now ready for proving the duality theorems claimed in Section 3.

**Proof of Proposition 3.1.** The assumption (3.2) immediately implies that

$$
\sum_{i=1}^k 1 \{ (x_i - \nu, \nu)^2 \geq 100 \lambda \} \geq 0.4 k
$$

This means that there are (at least) 0.4k points in $\{x_i\}_{i=1}^k$ such that $t_i := \langle (x_i - \nu)(x_i - \nu)^\top, M \rangle \geq 100 \lambda$, where $M = \nu \nu^\top$. We call them outliers.

Now by the SDP duality (D.1), we only need to show that for any feasible $w$ the objective is at least $\lambda$. Observe first that for a fixed $M$, the optimal $w^*$ for the max-min objective is to put weight $1/(1 - \epsilon)k$ on the $(1 - \epsilon)k$ points with the smallest $t_i$. Recall we set $\epsilon = 0.3$. Hence, by pigeonhole principle, the support of $w^*$ must have an overlap of size $0.1k$ with the outliers. It follows that

$$
\sum_{i=1}^k w_i^* \langle (x_i - \nu)(x_i - \nu)^\top, M \rangle \geq 0.1 k \cdot \frac{1}{(1 - 0.3)k} \cdot 100 \lambda \geq 10 \lambda.
$$

$^{2}$The $1/\epsilon$ dependence in the run-time can be removed by a simple bucketing trick due to [12]; also see Lemma B.1 of [21].
Since \( w^* \) is the optimal choice, this completes the proof. \( \square \)

The other direction is a bit more involved. The key idea is to round the maximizing PSD matrix \( M \) into a single vector \( v \), via gaussian sampling, and this part of the argument is due to [12].

**Proof of Proposition 3.2.** Strong duality (D.1) implies that there exists PSD \( M \) of unit trace such that

\[
\sum_{i=1}^{k} w_i \langle (x_i - \nu)(x_i - \nu)^\top, M \rangle \geq \lambda
\]

for all \( w \in \mathcal{W}_{k, \epsilon} \). As we observed, the optimal \( w^* \) for a fixed \( M \) would put weights on the points with smallest value of \( t_i = \langle (x_i - \nu)(x_i - \nu)^\top, M \rangle \). The fact that the objective is large implies that there must be more than \( \epsilon k = 0.1k \) points with \( t_i \geq \lambda \). Let \( B \) be this set of points such that \( t_i \geq \lambda \).

It remains to demonstrate a vector \( v \) such that

\[
\sum_{i=1}^{k} I \left\{ |\langle x_i - \nu, v \rangle| \geq 0.1\sqrt{\lambda} \right\} \geq 0.01k.
\]  

(D.2)

The idea is to round the PSD matrix \( M \) to a single vector \( v \) that achieves this inequality. The right rounding method is simply gaussian sampling. Namely, if we draw \( v_M \sim \mathcal{N}(0, M) \), then it can be shown that with constant probability \( v = v_M/\|v_M\| \) satisfies the property above.

For that, we apply the argument from [12]. First let \( g_i = \langle x_i - \nu, v_M \rangle \) for each \( i \in [k] \). Note that \( g_i \) is a mean-zero Gaussian random variable with variance \( \sigma_i^2 = t_i \). A standard anti-concentration calculation shows that for any \( i \in B \), \( \Pr(|g_i| \geq 0.5\sqrt{\lambda}) \geq 1/2 \). Therefore, if we define

\[
Y = \sum_{i=1}^{k} I \left\{ |\langle x_i - \nu, v \rangle| \geq 0.5\sqrt{\lambda} \right\},
\]

then by linearity of expectations we have \( \mathbb{E} Y \geq 0.05k \). It follows from the Payley-Zigmund inequality that \( \Pr(Y \geq 0.01k) \geq 0.0018 \). Moreover, by Borell-TIS inequality (Theorem 7.1 of [31]), we can bound that with probability at least 0.999,

\[
\|v_M\| \leq \mathbb{E} \|v_M\| + 4\sqrt{\|M\|} \leq \sqrt{\text{Tr}(M)} + 4\sqrt{\text{Tr}(M)} \leq 5,
\]

since \( \text{Tr}(M) = 1 \). Combing these facts immediately prove (D.2). \( \square \)

### E Technical Details of Section 4

#### E.1 Proof of Lemma 4.2

The proof is via a simple “pigeonhole + triangle inequality” argument.

**Proof of Lemma 4.2.** Let \( v \) be the unit vector in the direction of \( \mu - \hat{\mu} \). Then since \( \hat{\mu} \) is an \((\epsilon, O(r_3^2))\)-combinatorial center with \( \epsilon < 1/2 \), we have \( |\langle Z_i - \hat{\mu}, v \rangle| \leq r_3 \) for most \( Z_i \). Also, \( |\langle Z_i - \mu, v \rangle| \leq O(r_3) \) for most \( \{Z_i\}_{i=1}^{k} \) by our assumption from Lugosi-Mendelson lemma. By the pigeonhole principle, there must be a \( Z_j \) such that \( |\langle Z_j - \mu, v \rangle| \leq O(r_3) \) and \( |\langle Z_j - \hat{\mu}, v \rangle| \leq O(r_3) \). By triangle inequality,

\[
\|\hat{\mu} - \mu\| = |\langle Z_i - \mu, v \rangle| \leq |\langle Z_i - \hat{\mu}, v \rangle| + |\langle Z_i - \hat{\mu}, v \rangle| \leq O(r_3).
\]

as desired, and this completes the proof. \( \square \)

#### E.2 Proof of Theorem 4.3

**Proof of Theorem 4.3.** Let \( \{Z_i\}_{i=1}^{k} \) be the bucket means with \( k = \lceil 800\log(1/\delta) \rceil \) and let \( \lambda = 3000r_3 \). We assume that the true mean \( \mu \) is a \((0.01, \lambda^2)\)-combinatorial center of \( \{Z_i\}_{i=1}^{k} \). Suppose that we can obtain an \( \alpha \)-factor approximation the spectral sample reweighing, with the input being \( \{Z_i\}_{i=1}^{k} \).
• **Promise:** First let’s check the spectral centrality condition holds. Since, by assumption, \( \mu \) is a 
\((0.01, \lambda^2)\)-combinatorial center of \( \{Z_i\}_{i=1}^k \), we have that for all unit \( v \)
\[
\sum_{i=1}^k \mathbb{1}\{|\langle x_i - \mu, v \rangle| \geq \lambda\} \leq 0.01k.
\]

Thus, Proposition 3.2 (with \( \nu = \mu \)) implies that
\[
\min_{w \in \mathcal{W}_{k, \epsilon}} \max_{M \geq 0, \text{Tr}(M) = 1} \sum_{i=1}^k w_i \langle (x_i - \mu)(x_i - \mu)^T, M \rangle \leq 100\lambda^2,
\]
where \( \epsilon = 0.1 \). This means that there exists \( w \in \mathcal{W}_{k, \epsilon} \) such that
\[
\left\| \sum_{i=1}^n w_i (x_i - \mu)(x_i - \mu)^T \right\| \leq 100\lambda^2.
\]

• **Output:** Now the guarantee of an \( \alpha \)-factor approximation for spectral sample reweighing (Definition 2.1) is that we have \( \mu \in \mathbb{R}^d \) and \( w' \in \mathcal{W}_{k, 3\epsilon} \) such that
\[
\left\| \sum_{i=1}^n w'_i (x_i - \mu)(x_i - \mu)^T \right\| \leq 100\alpha\lambda^2.
\]

It immediately follows that
\[
\min_{w \in \mathcal{W}_{k, \epsilon}} \max_{M \geq 0, \text{Tr}(M) = 1} \sum_{i=1}^k w_i \langle (x_i - \mu)(x_i - \mu)^T, M \rangle \leq 100\alpha\lambda^2.
\]

Now we can apply Proposition 3.1. Since \( \alpha \) is a constant by assumption, we obtain that for all 
unit \( v \),
\[
\sum_{i=1}^k \mathbb{1}\{|\langle x_i - \hat{\mu}, v \rangle| \geq C(\alpha) \cdot \lambda\} \leq 0.4k,
\]
for some constant \( C(\alpha) = O(1) \) that depends on \( \alpha \). Therefore, we get that a majority of the 
points cluster around \( \hat{\mu} \), along any direction \( v \), so it is a \((0.4, O(\lambda))\)-combinatorial center. It 
follows from Lemma 4.2 that \( \|\hat{\mu} - \mu\| \leq O(r_\delta) \), as \( \lambda = O(r_\delta^2) \).

Finally, note that the only condition of the argument is that the true mean is a combinatorial center, 
which occurs with probability at least \(1 - \delta\), by Lemma 4.1.

We remark that the exact constants we choose in the proof are immaterial, and no efforts have been 
given in optimizing them.

**E.3 Proof of Corollary 4.4**

**Proof of Corollary 4.4.** Given the input, we first compute the bucket means \( \{Z_i\}_{i=1}^{2k} \), which takes 
\( O(nd) \) time. Assume that the condition of the Lugosi-Mendelson structural lemma (Lemma 4.1) 
holds; that is, \( \mu \) is a \((0.01, \lambda^2)\)-combinatorial center of \( \{Z_i\}_{i=1}^{k} \), where \( \lambda = 3000r_\delta \). We use the filter 
algorithm (Algorithm 1) with the input being a pruned subset of \( \{Z_i\}_{i=1}^{k} \) and apply its guarantees.

Here, we will not use the pruning step (Lemma B.3), since it requires the knowledge of \( \lambda \). Instead, 
we first compute the coordinate-wise median-of-means \( \bar{\mu}_0 \) of \( \{Z_i\}_{i=k+1}^{2k} \) and the distances \( d_i = \|Z_i - \bar{\mu}_0\| \) for each \( i \in [k] \). We then sort the points by \( d_i \) (in descending order) and remove the top 
0.01\( k \) points in \( \{Z_i\}_{i=1}^{k} \) with large \( d_i \). It can be shown that the remaining points has diameter at 
most \( O(\sqrt{d_\delta}) \); see Lemma E.1 of [32]. Let \( S \) the remaining points in \( \{Z_i\}_{i=1}^{k} \). 

For the run-time, we can apply the guarantee of the filter algorithm (Lemma B.1), given the input \( S \) 
and a failure probability \( \delta/\delta' \). Since the squared diameter is \( \rho = O(d_\delta^2) \) and \( \lambda = O(r_\delta^2) \), this gives a 
run-time of \( O(k^2d^2) \), since \( k = O(\log(1/\delta)) \).
We now have a constant-factor approximation for the spectral sample reweighing problem. By Theorem 4.3, this gives an estimate with the sub-gaussian error (4.1). Finally, the procedure’s success depends on the condition of Lugosi-Mendelson (Theorem 4.3), success of the pruning procedure, and the guarantees of constant-approximation of spectral sample reweighing (Theorem 2.1). The failure probability of each event can be bounded by $\delta/3$. Applying union bound completes the proof.

E.4 Further Technical Comments

Other algorithms for heavy-tailed mean estimation This argument also enables us to solve the heavy-tailed mean estimation problem using other approximation algorithms for the spectral sample reweighing problem. Let $\lambda = 3000r_\delta$. Recall that the argument for Theorem 4.3 shows that there is a $(0.1, O(\lambda^2))$-spectral center (which is the true mean $\mu$). Moreover, the pruning step in the proof of Corollary 4.4 allows us to bound the squared diameter of a large subset of $\{Z_i\}_{i=1}^n$ by $\rho = O(d\lambda^2)$.

This implies that the gradient descent-based algorithm that we analyze in Appendix H solves the heavy-tailed setting in $O(kd^2)$ iterations.

**Corollary E.1** (heavy-tailed mean estimation via gradient descent). Assume the setting of Corollary 4.4. A black-box application of the gradient descent-based algorithm (Algorithm 4, Appendix H) solves the heavy-tailed mean estimation problem with optimal error rate within $O(nd^3)$ iterations and $O(nd^3)$ time.

The quantum entropy scoring scheme (Appendix G), however, runs in $\tilde{O}(\log(p/\lambda))$ number of iterations. Setting its failure probability to be $\delta/3$, we obtain the following, which matches the fastest-known algorithm for the problem [12, 32].

**Corollary E.2** (heavy-tailed mean estimation via quantum entropy scoring). Assume the setting of Corollary 4.4. A black-box application of the matrix multiplicative update algorithm (Algorithm 3, Appendix G) solves the heavy-tailed mean estimation problem with optimal error rate, in $O(1)$ iterations and $O(k^2d)$ total run-time.

F Extension to sub-gaussian distributions

We now consider a variant of the filter algorithm (Algorithm 1) analyzed in Section 2. The difference is that instead of fixing the step size to be $\eta = 1/2$, we set it as $\epsilon$. That is, we will perform the multiplicative update less aggressively when there are few bad points. In addition, we require a stronger approximation for the largest eigenvector computation. This increases the the run-time by an $O(\text{poly}(1/\epsilon))$ factor. For technical reasons, we also ask the algorithm to stop early if the weighted covariance has been reduced to a desired value. Formally, the algorithm is described by the pseudo-code below (Algorithm 2).

**Algorithm 2:** Multiplicative weights for sub-gaussian robust mean estimation

**Input:** A set of points $\{x_i\}_{i=1}^n$, an iteration count $T$, and parameter $\rho, \delta$

**Output:** A set of weights $w \in \mathcal{W}_{n,\epsilon}$.

1. Let $w^{(1)} = \frac{1}{n} \mathbb{I}_n$.
2. For $t$ from 1 to $T$
   3. Let $w^{(t)} = \sum_i w_i^{(t)} x_i$, $M^{(t)} = \sum_i w_i^{(t)} (x_i - \mu^{(t)})(x_i - \mu^{(t)})^T$.
   4. Let $\nu^{(t)}$ be a $(1 - \epsilon^2)$-approximate largest eigenvector of $M^{(t)}$ (with $\|\nu^{(t)}\| = 1$). Fail with probability at most $\delta/T$.
   5. If $\lambda^{(t)} = \nu^{(t)^T} M^{(t)} \nu^{(t)} \leq 1$, return $w^{(t)}$.
   6. Compute $v_i^{(t)} = \nu^{(t)^T} x_i - \nu^{(t)}$.
   7. Set $w_i^{(t+1)} \leftarrow w_i^{(t)} \left(1 - \epsilon v_i^{(t)}/\rho\right)$ for each $i$.
   8. Project $w^{(t+1)}$ onto the set of good weights $\mathcal{W}_{n,\epsilon}$ (under KL divergence).
9. Return $w^{(t^*)}$, where $t^* = \arg\min_t \|M^{(t)}\|$. 


First, we need a stronger spectral signature lemma.

**Lemma F.1** ([21]). Let \( S = \{ x_i \}_{i=1}^{n} \) be an \( \epsilon \)-corrupted set of \( n \) samples from a sub-gaussian distribution over \( \mathbb{R}^d \), with mean \( \mu \) and identity covariance. Suppose \( n \geq \tilde{\Omega}(d/\epsilon^2) \). If \( \|M(w)\| \leq 1 + \lambda \), for some \( \lambda \geq 0 \), then for any \( w \in \mathcal{W}_{n, 2\epsilon} \),

\[
\| \mu - \mu(w) \| \leq \frac{1}{1 - \epsilon} \left( \sqrt{\epsilon \lambda + C \epsilon \sqrt{\log(1/\epsilon)}} \right),
\]

for some universal constant \( C > 0 \).

Moreover, we assume that for all \( w \in \mathcal{W}_{n, \epsilon} \) we have

\[
\left\| \sum_{i \in G} w_i (x_i - \mu)(x_i - \mu)^T - I \right\| \leq \lambda = O(\epsilon \log(1/\epsilon)). \tag{F.1}
\]

This condition holds with high probability over the draws of samples [14].

**Lemma F.2** (analysis of sub-gaussian filter). Let \( \epsilon \) be a sufficiently small constant and \( \{ x_i \}_{i=1}^{n} \) be \( n \) points in \( \mathbb{R}^d \). Assume the following (deterministic) conditions hold.

(i) There exists \( \nu \in \mathbb{R}^d \) and \( w \in \mathcal{W}_{n, \epsilon} \) such that

\[
\left\| \sum_{i=1}^{n} w_i (x_i - \nu)(x_i - \nu)^T \right\| \leq 1 + O(\epsilon \log(1/\epsilon)) \tag{F.2}
\]

(ii) If \( \|M(w)\| \leq 1 + \lambda \), for some \( \lambda \geq 0 \), then for any \( w \in \mathcal{W}_{n, \epsilon} \),

\[
\| \nu - \mu(w) \| \leq \frac{1}{1 - \epsilon} \left( \sqrt{\epsilon \lambda + C \epsilon \sqrt{\log(1/\epsilon)}} \right) \tag{F.3}
\]

Then, given \( \{ x_i \}_{i=1}^{n} \), a failure rate \( \delta \) and \( \rho \) such that \( \rho \geq \tau_i^{(t)} \) for all \( i \) and \( t \), Algorithm 2 finds \( w' \in \mathcal{W}_{n, \epsilon} \) such that

\[
\|M(w')\| \leq 1 + O(\epsilon \log(1/\epsilon)) \tag{F.4}
\]

with probability at least \( 1 - \delta \).

The algorithm terminates in \( T = O(\rho/\epsilon) \) iterations. Further, if \( T = O(\text{poly}(n, d)) \), then each iteration takes \( \tilde{O}(nd \log (1/\delta)/\epsilon^2) \) time.

**Proof of Lemma F.2.** If the algorithm gets stopped early (at Line 5), then it means that

\[
\|M^{(t)}\| \leq \lambda^{(t)}/(1 - \epsilon^2) \leq 1/(1 - \epsilon^2) \leq 1 + O(\epsilon^2),
\]

since \( v^{(t)} \) is a \( (1 - \epsilon^2) \) approximate largest eigenvector of \( M^{(t)} \). Hence, in this case, we immediately achieves the goal (F.4).

Now assume the algorithm did not stop early and so \( \|M^{(t)}\| > 1 \) for all \( t \). Then we have

\[
\sum_{i} w_i^{(t)} \tau_i^{(t)} = \sum_{i} w_i^{(t)} \left( v^{(t)}, x_i - \nu^{(t)} \right)^2 = v^{(t)\top} M^{(t)} v^{(t)} \geq (1 - \epsilon^2) \left\| M^{(t)} \right\|_2,
\]

for all \( t \). Since the step size \( \epsilon < 1/2 \) and \( \rho \geq \tau_i^{(t)} \) for all \( i, t \) by assumption, we can apply the regret bound of MWU (Lemma B.4) and conclude that for \( w \) that satisfies assumption (F.2),

\[
\frac{1 - \epsilon^2}{T} \sum_{t=1}^{T} \left\| M^{(t)} \right\|_2 \leq \frac{1}{T} \sum_{t=1}^{T} \left( w^{(t)}, \tau^{(t)} \right) \leq (1 + \epsilon) \frac{1}{T} \sum_{t=1}^{T} \left( w, \tau^{(t)} \right) + \frac{\rho \cdot \text{KL}(w||w^{(1)})}{T \epsilon}. \tag{F.6}
\]

We now focus on bounding \( \frac{1}{T} \sum_{t=1}^{T} \left( w, \tau^{(t)} \right) \).

**Claim F.3.** In the setting above, we have

\[
\frac{1}{T} \sum_{t=1}^{T} \left( w, \tau^{(t)} \right) \leq 1 + O(\epsilon \log(1/\epsilon)) + \frac{2\epsilon}{(1 - \epsilon)^2} \frac{1}{T} \sum_{t=1}^{T} \left\| M^{(t)} \right\| - \frac{2\epsilon}{(1 - \epsilon)^2}
\]
Proof. Note that
\[
\frac{1}{T} \sum_{t=1}^{T} \langle w, \tau(t) \rangle = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} w_i \langle x_i - \nu(t), v(t) \rangle
\]
\[
= \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} w_i \left( \langle x_i - \nu, v(t) \rangle^2 + \langle \nu - \nu(t), v(t) \rangle^2 \right)
\]
\[
\leq 1 + O(\epsilon \log(1/\epsilon)) + \frac{1}{T} \sum_{t=1}^{T} \langle \nu - \nu(t), v(t) \rangle^2 \quad \text{(F.7)}
\]
\[
\leq 1 + O(\epsilon \log(1/\epsilon)) + \frac{1}{T} \sum_{t=1}^{T} \| \nu - \nu(t) \|_2^2, \quad \text{(F.8)}
\]
where (F.7) follows from the assumption (F.4). Now we apply assumption (F.3) to bound $\| \nu - \nu(t) \|_2^2$.
Since we may assume $\| M(t) \| \geq 1$ by the early stopping of Line 5, we have
\[
\| \nu - \nu(t) \|_2^2 \leq \frac{2(1 + \epsilon)\epsilon}{(1 - \epsilon)^2} \left( \epsilon \left( \| M(t) \| - 1 \right) + C^2 \epsilon^2 \log(1/\epsilon) \right)
\]
\[
= \frac{2\epsilon}{(1 - \epsilon)^2} \| M(t) \| - \frac{2\epsilon}{(1 - \epsilon)^2} + O(\epsilon^2 \log(1/\epsilon)).
\]
Substituting the bound back into (F.8) completes the proof.

Using Claim F.3, the KL bound (Lemma B.8) and (F.6), we have
\[
\frac{1 - \epsilon^2}{T} \sum_{t=1}^{T} \| M(t) \|_2^2 \leq \frac{2(1 + \epsilon)\epsilon}{(1 - \epsilon)^2} \frac{1}{T} \sum_{t=1}^{T} \| M(t) \| + 1 - \frac{2(1 + \epsilon)\epsilon}{(1 - \epsilon)^2} + O(\epsilon \log(1/\epsilon)) + \frac{5\rho}{T}.
\]
For sufficiently small $\epsilon$, we rearrange and divide through to obtain
\[
\frac{1}{T} \sum_{t=1}^{T} \| M(t) \|_2 \leq 1 + O(\epsilon \log(1/\epsilon)) + O(\epsilon) + \frac{O(\rho)}{T}.
\]
Setting $T = O(\rho/\epsilon)$ completes the correctness proof. Finally, the per-iteration cost follows from the run-time of using power method to approximate the largest eigenvector.

Using the lemma we can prove our main theorem.

Theorem F.4 (sub-gaussian robust mean estimation, [14]). Let $S = \{x_i\}_{i=1}^{n}$ be an $\epsilon$-corrupted set of $n$ samples from a sub-gaussian distribution over $\mathbb{R}^d$, with mean $\mu$ and identity covariance. Suppose $n \geq \Omega(d/\epsilon^2)$. Given $S$, there is an algorithm that outputs $\hat{\mu}$ such that $\| \hat{\mu} - \mu \| \leq O(\epsilon \log(1/\epsilon))$ with high constant probability. The algorithm runs in time $O(nd^2/\epsilon^3)$.

Proof. Let $\delta = 0.01$. We apply Algorithm 2 with a simple pruning procedure as a preprocessing. By standard concentration of sub-gaussian random vectors, with high constant probability, $\| x_i - \mu \| \leq r = O(\sqrt{d \log n})$ for all $i \in G$. Hence, we apply PRUNE($S, r, \delta$), and by Lemma B.3 it guarantees to terminate in $O(nd)$ time and removes at most $\epsilon n$ (bad) points.
We feed the remaining (at least) $\epsilon n$ points $R \supseteq G$ into Algorithm 2 with $\rho = r^2$. Notice that $\frac{1}{|1 - \epsilon^2|} \leq \frac{1}{1 - 2\epsilon}$ for $\epsilon \leq 1/2$, so assumptions (i)-(ii) of Lemma F.2 are satisfied by the claim of (F.1) and Lemma F.1, respectively.

It then follows from Lemma F.2 that Algorithm 2 outputs $w' \in \mathcal{W}_{|R|, \epsilon}$ such that
\[
\left\| \sum_{i \in R} (x_i - \mu(w'))(x_i - \mu(w'))^T \right\| \leq 1 + O(\epsilon \log(1/\epsilon)),
\]
where $\mu(w') = \sum_{i \in R} w'_i x_i$. Let $w''_i = w'_i$ if $i \in R$ and $w''_i = 0$ otherwise. We obtain $w'' \in \mathcal{W}_{n, 2\epsilon}$ such that $\| M(w'') \| \leq 1 + O(\epsilon \log(1/\epsilon))$. Applying the spectral signature (Lemma F.1) proves that $\mu(w'')$ attains the desired estimation error. Moreover, the run-time simply follows from Lemma F.2.

24
G Sampling reweighing via Matrix Multiplicative Update

We now show that the spectral sample reweighing problem (Definition 2.1) can be solved in near linear time via a matrix multiplicative update scheme from the recent work of [21]. Our analysis will closely resemble the arguments therein.

**Theorem G.1.** Let \( \{x_i\}_{i=1}^n \) be \( n \) points in \( \mathbb{R}^d \). Suppose there exists \( \nu \in \mathbb{R}^d \) and \( w \in \mathcal{W}_{n,2} \) such that \( \sum_{i=1}^n w_i (x_i - \nu) (x_i - \nu)^\top \leq \lambda I \) for some \( \lambda > 0 \) and a sufficiently small \( \epsilon \). Then, given \( \{x_i\}_{i=1}^n \), \( \lambda \), the squared diameter \( \rho \) of the points and a failure rate \( \delta \), there is a matrix multiplicative weights-based algorithm (Algorithm 3) that, with probability at least \( 1 - \delta \), finds \( w' \in \mathcal{W}_{n,\epsilon} \) and \( \nu' \in \mathbb{R}^d \) such that

\[
\sum_{i=1}^n w_i' (x_i - \nu') (x_i - \nu')^\top \leq O(\lambda) I.
\]

Further, the algorithm terminates in \( O(\log(\lambda/\rho)) \) iterations, where \( \rho \) is the squared diameter of the input points \( \{x_i\}_{i=1}^n \), and each iteration can be implemented in \( O(nd\log(1/\delta)) \) time.

**Remark G.1.** In the following, we will consider an idealized version of the algorithm and omit the detail of implementing the numerical linear algebra primitives in \( O(nd\log(1/\delta)) \) time each iteration. The exact details can be found in [21].

The algorithm is based on the matrix multiplicative weights update. For a sequence of PSD matrices \( M_1 \geq M_2 \geq \cdots \geq M_{l-1} \), we will apply the matrix multiplicative weight (MMW) update, given by

\[
\text{MMW}(M_0, M_1, \cdots, M_{l-1}) = \exp \left( \frac{1}{\|M_0\|_2} \sum_{k=1}^{l-1} M_k \right) / \text{tr} \exp \left( \frac{1}{\|M_0\|_2} \sum_{k=1}^{l-1} M_k \right) .
\]

(G.1)

For technical reasons, we will not maintain a set of weights that is a probability distribution throughout. Instead, recall by Lemma B.6 that there exists a subset \( G \) of size \( (1 - \epsilon)n \) such that \( A_G - \lambda I \). We will maintain the mostly-good property of the input weights.

**Definition G.1** (mostly-good weight vector). Given \( \{x_i\}_{i=1}^n \) that satisfy the spectral centrality condition (i), let \( G \) be a subset of size \( (1 - \epsilon)n \) such that \( A_G - \lambda I \). The set of mostly-good weight vectors is

\[
C_{n,\epsilon} = \left\{ w \in \mathbb{R}^n : 0 \leq w_i \leq \frac{1}{n} \quad \text{and} \quad \frac{1}{n} \|x_G\|_2 - w_G \leq \frac{1}{n} \|x_B\|_2 - w_B \right\}
\]

A crucial subroutine we use is a deterministic down-weighting scheme, directly from [21], that maintains the mostly-good property of the input weights.

**Lemma G.2** (1D Filter [21]). Let \( \eta \in (0, 1/2) \), let \( b \geq 2\eta \), and let \( w_1, \ldots, w_m \) and \( \tau_1, \ldots, \tau_m \) be non-negative numbers so that \( \sum_{i=1}^m w_i \leq 1 \). Let \( \tau_{\max} = \max_{i \in [m]} \tau_i \). Suppose there exist two disjoint sets \( G, B \) so that \( G \cup B = [m] \), and moreover,

\[
\sum_{i \in G} w_i \tau_i \leq \eta \sigma, \text{ where } \sigma = \sum_{i=1}^n w_i \tau_i .
\]

Then \( \text{1DFilter}(w, \tau, b) \) runs in time \( O(1 + \log \frac{\tau_{\max}}{\eta b}) ) \) and outputs \( 0 \leq w' \leq w \) so that:

- more weight is removed from \( B \) than from \( G \): \( \sum_{i \in G} w_i - w_i' \leq \sum_{i \in B} w_i - w_i' \), and
- the weighted sum of \( \tau_i \) is decreased: \( \sum_{i=1}^m w_i' \tau_i \leq b \sigma \).

The algorithm is formally described in Algorithm 3. Throughout let \( M^{(s)} = M(w^{(s)}) \) and \( M^{(s)} = M(w^{(s)}) \), where \( M(w) = \sum_{i=1}^n w_i (x_i - \mu(w)) (x_i - \mu(w))^\top \). The procedure runs by epochs, where each epoch \( s \) reduces the largest eigenvalue of \( M^{(s)} \) by a constant factor. We will show that the inner loop achieves the reduction within \( O(\log d) \) iterations while maintaining the invariant that the weights are mostly-good (Definition G.1).
Algorithm 3: Matrix multiplicative update for spectral sample reweighing (Definition 2.1)

**Input:** A set of points \( x_1, \ldots, x_n \), \( \lambda, \rho \) and a failure rate \( \delta \)

**Output:** A point \( \nu' \in \mathbb{R}^d \) and weights \( w' \in \mathcal{W}_{n, \epsilon} \) that satisfy (2.1) up to a constant factor.

1. Let \( w^{(0)} = \frac{1}{n}(1, 1, \ldots, 1) \).
2. For \( s \) from 0 to \( O(\log \rho) \)
   3. Compute \( \lambda^{(s)} = \|M^{(s)}\| \).
   4. If \( \lambda^{(s)} \leq 300\lambda \) return \( w^{(s)}/\|w^{(s)}\|_1, \mu(w^{(s)}) \).
5. For \( t \) from 0 to \( O(\log d) \)
   6. Compute \( \lambda_t^{(s)} = \|M_t^{(s)}\| \) and terminate epoch if \( \lambda_t^{(s)} \leq \frac{3}{4}\lambda_0^{(s)} \).
   7. Compute \( U_t^{(s)} = \text{MMW}(M_1^{(s)}, M_2^{(s)}, \ldots, M_{t-1}^{(s)}) \).
   8. Compute \( \tau_t^{(s)} = (x_i - \mu(w_t^{(s)}))^T U_t^{(s)}(x_i - \mu(w_t^{(s)})) \) (G.2)
   9. Let \( w_{t+1}^{(s)} = w_t^{(s)} \) if \( \sum_i w_{t,i} \tau_{t,i}^{(s)} \leq \frac{1}{4}\lambda_t^{(s)} \); otherwise \( w_{t+1}^{(s)} = \text{1DFILTER}(w_t^{(s)}, \tau_t^{(s)}) \).
10. Let \( w^{(s+1)} = w_t^{(s)} \).

To start with the analysis, we first establish certain invariants of the algorithm. This requires the following lemma. The proof follows from exactly the same argument for Lemma B.7, which we omit for the sake of brevity.

**Lemma G.3.** Let \( \{x_i\}_{i=1}^n \) be \( n \) points in \( \mathbb{R}^d \). Suppose there exists \( \nu \in \mathbb{R}^d \) and a mostly-good weight vector \( \nu \in \mathcal{C}_{n, \epsilon} \) such that \( \sum_{i=1}^n w_i (x_i - \nu) (x_i - \nu)^T \leq \lambda I \), for some \( \lambda > 0 \). Then for any \( w' \in \mathcal{C}_{n, \epsilon} \),

\[
\|\nu - \nu(w')\| \leq \frac{1}{1 - 2\epsilon} \left( 2\sqrt{\lambda} + \sqrt{\epsilon\|M(w')\|} \right),
\]

(G.3)

where \( \nu(w') = \sum_i w_i x_i \) and \( M(w') = \sum_i w_i(x_i - \nu(w'))(x_i - \nu(w'))^T \).

Using this, we establish a key lemma of the inner loop of the algorithm.

**Lemma G.4.** Let \( w \in \mathcal{C}_{n, \epsilon} \) be such that \( \beta = \|M(w)\|_2 \geq 300\lambda \) and \( U \) be a density matrix. Let \( \tau_i = (x_i - \mu(w))^T U (x_i - \mu(w)) \). If \( \beta \geq \frac{3}{4}\beta \) and \( w = \text{1DFILTER}(w, \tau, 1/4) \), then we have \( w' \in \mathcal{C}_{n, \epsilon} \) and \( (M(w'), U) \leq \frac{1}{4}(M(w), \bar{U}) \).

**Proof.** Let \( \bar{w}_i = 1/n \) if \( i \in G \) and \( \bar{w}_i = 0 \) otherwise. Let \( \mu(\bar{w}) = \sum_i \bar{w}_i x_i \). Then for any unit vector \( u \), we have

\[
\langle \mu(\bar{w}) - \nu, u \rangle^2 \leq \frac{1}{(1 - \epsilon)n} \sum_{i \in G} x_i - \nu, u \rangle^2 \leq \frac{1}{(1 - \epsilon)n} \sum_{i \in G} (x_i - \nu, u)^2 \leq \lambda.
\]
We are now ready to prove the main theorem of this section. Consider a fixed epoch and drop the super script for simplicity of notation. The analysis will be through the classic regret guarantee of online gradient descent ([Definition 2.1]). The analysis will be through the classic regret guarantee of online gradient descent for convex optimization [46]. Though the resulting run-time is higher than the MWU scheme
we analyzed in Section 2, it nonetheless betters the recent work of [9], where essentially the same
gradient descent-based algorithm is studied.

We will leverage the following regret guarantee of online gradient descent; the definition of the
algorithm in the general setting can be found in standard text [23].

**Lemma H.1 (Theorem 3.1 [23], originally due to [46]).** Let \( f_t : \mathcal{X} \rightarrow \mathbb{R} \) be the convex cost function
revealed at iteration \( t \), where \( \mathcal{X} \) is a convex feasible set. Suppose \( f_t \) is \( L \)-Lipschitz (in \( \ell_2 \) norm) and
\[
\| x_0 - x^* \|_2 \leq R \text{ for some } x^* \in \arg\min_{x \in \mathcal{X}} \sum_t f_t(x).
\]
The online gradient descent algorithm with step sizes \( \eta_t = \frac{R}{L\sqrt{T}} \) achieves
\[
\sum_{t=1}^T f_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^T f_t(x) \leq \frac{3LRT}{2}.
\]

Our algorithm implicitly defines the cost functions \( f_t(w) = \langle w, \tau(t) \rangle \), where the feasible set is
\( \mathcal{W}_{n,\epsilon} \), and implements the online gradient descent algorithm for the linear objective. Note that
\( \nabla f_t(w) = \tau(t) \), and the main difference of this algorithm from the MWU scheme (Algorithm 1) is
that we use an additive/gradient-descent update, in lieu of the multiplicative update.

**Algorithm 4:** Gradient descent for spectral sample reweighing (Definition 2.1)

**Input:** A set of points \( \{ x_i \}_{i=1}^n \), an iteration count \( T \), and step sizes \( \eta_t \)

**Output:** A point \( \nu \in \mathbb{R}^d \) and weights \( w \in \mathcal{W}_{n,\epsilon} \).

1. Let \( w^{(1)} = \frac{1}{n}(1, 1, \ldots, 1) \).
2. For \( t \) from 1 to \( T \)
   3. Let \( \nu^{(t)} = \sum_i w_i^{(t)} x_i, M^{(t)} = \sum_i w_i^{(t)} (x_i - \nu^{(t)})(x_i - \nu^{(t)})^T \).
   4. Let \( \nu^{(t)} \) be the top eigenvector of \( M^{(t)} \) (with \( \| \nu^{(t)} \| = 1 \)).
   5. Compute \( \tau_i^{(t)} = \langle \nu^{(t)}, x_i - \nu^{(t)} \rangle^2 \).
   6. Set \( w_i \leftarrow w_i - \eta_t \tau_i^{(t)}. \)
   7. Project \( w^{(t+1)} \) onto the set of good weights \( \mathcal{W}_{n,\epsilon} \) (under \( \ell_2 \) distance).
8. Return \( \nu^{(t^*)}, w^{(t^*)} \), where \( t^* = \arg\min_t \| M^{(t)} \| \).

**Lemma H.2.** Let \( \rho \) be the squared diameter of the inputs points \( \{ x_i \}_{i=1}^n \). The cost function \( f_t(\cdot) \) is.
\( \sqrt{n} \rho \)-Lipschitz (in \( \ell_2 \) norm), for all \( t \).

**Proof.** Since \( f_t \) is differentiable, we only need the bound \( \| \nabla f_t \| \). We have that for all \( t \) and \( i \),
\[
\tau_i^{(t)} = \langle \nu^{(t)}, x_i - \nu^{(t)} \rangle^2 \leq \| x_i - \nu^{(t)} \|_2^2 \leq \rho.
\]
Therefore, \( \| \nabla f_t \| = \| \tau_i^{(t)} \| \leq \sqrt{n} \rho. \)

**Theorem H.3.** Given \( \{ x_i \}_{i=1}^n \) and \( \eta_t = R/L\sqrt{T} \) with \( L = \sqrt{n} \rho, R = \sqrt{2} \), the online gradient
descent algorithm (based on Algorithm 4) yields a constant-factor approximation for the spectral
sample reweighing problem (Definition 2.1) in \( O(n^2d^2/\epsilon^2) \) iterations and \( \Omega(n^2d^3/\epsilon^2) \) total run-time.

**Proof.** We first apply the PRUNE procedure of Lemma B.3 to bound the diameter. By Lemma B.2
and the guarantee of PRUNE, we can have \( \rho = 16d\lambda/\epsilon \). Then we apply Algorithm 4.

We will use Lemma H.1 to analyze Algorithm 4. First, by Lemma H.2, we have \( L = \sqrt{n} \rho, \) and
further, since the \( \ell_2 \) diameter of the probability simplex can be (trivially) bounded by \( \sqrt{2}, R = \sqrt{2} \).
Moreover, observe for any \( t \),
\[
f_t(\nu^{(t)}) = \langle \nu^{(t)}, \tau^{(t)} \rangle = \sum_i w_i^{(t)} \langle \nu^{(t)}, x_i - \nu^{(t)} \rangle^2 = \nu^{(t)}^T M^{(t)} \nu^{(t)} = \| M^{(t)} \|_2.
\]
Let \( w \in \mathcal{W}_{n, \varepsilon} \) be a weight that satisfies the spectral centrality condition. Then, from the regret guarantee (H.1),

\[
\frac{1}{T} \sum_{t=1}^{T} \left\| M^{(t)} \right\|_2 \leq \frac{1}{T} \sum_{t=1}^{T} \langle w, r^{(t)} \rangle + \frac{3Ld}{2T}.
\] (H.2)

We bound the two terms on the right side individually.

(i) A bound on the first term follows exactly from the calculations we did in the analysis of MWU algorithm (Algorithm 1). In particular, from (B.12) we have

\[
\frac{1}{T} \sum_{t=1}^{T} \langle w, r^{(t)} \rangle \leq 15\lambda + \frac{1}{3T} \sum_{t=1}^{T} \left\| M^{(t)} \right\|_2.
\]

(ii) Observe that it suffices to set \( T = 3L^2 R^2 / \lambda^2 \) to bound the second term by \( \lambda \).

Substituting the two bounds back into (H.2),

\[
\frac{1}{T} \sum_{t=1}^{T} \left\| M^{(t)} \right\|_2 \leq 16\lambda + \frac{1}{3T} \sum_{t=1}^{T} \left\| M^{(t)} \right\|_2.
\] (H.3)

Rearranging and dividing through immediately yields the desired guarantee.

Given that \( L = \sqrt{n} \rho, R = \sqrt{2} \), we have that the iteration count \( T = 6n \rho^2 / \lambda^2 \). Since \( \rho = 16d \lambda / \varepsilon \), \( T = O(n^d \varepsilon^2 / \varepsilon^2) \). For the run-time, note that instead of computing the exact largest eigenvector, we can use power method to find an \( 7/8 \)-approximate one. Observe that this suffices for our analysis of the method above. Finally, the Euclidean projection onto \( \mathcal{W}_{n, \varepsilon} \) can be computed in \( O(n \log n) \) time [44]. This yields the desired run-time.

**H.2 Extension to sub-gaussian setting**

Theorem H.3 implies that a gradient descent-based algorithm (Algorithm 4) can be used for robust mean estimation under bounded covariance. We now extend the result to the sub-gaussian setting, showing that the same iteration and run-time complexity holds. The optimal estimation error we will aim for is \( O(\varepsilon \sqrt{\log(1/\varepsilon)}) \). We assume the spectral signature Lemma F.1 and the deterministic condition (F.1).

**Algorithm 5:** Gradient descent for sub-gaussian robust mean estimation

**Input:** A set of points \( \{x_i\}_{i=1}^{n} \), step sizes \( \eta_t \), an iteration count \( T \), and parameter \( \rho \)

**Output:** A set of weights \( w \in \mathcal{W}_{n, \varepsilon} \).

1. Let \( w^{(1)} = \frac{1}{n} I_n \).
2. For \( t \) from 1 to \( T \)
3. \hspace{1em} Let \( w^{(t)} = \sum_i w_i^{(t)} x_i \), \( M^{(t)} = \sum_i w_i^{(t)} (x_i - \mu^{(t)})(x_i - \mu^{(t)})^T \).
4. Let \( \mu^{(t)} \) be a \( (1 - \varepsilon^2) \)-approximate largest eigenvector of \( M^{(t)} \) (with \( \| \mu^{(t)} \| = 1 \)). Fail with probability at most \( \delta / T \).
5. If \( \lambda^{(t)} = \mu^{(t)^T} M^{(t)} \mu^{(t)} \leq 1 \), return \( w^{(t)} \).
6. Compute \( v_{i}^{(t)} = \langle \mu^{(t)}, x_i - \mu^{(t)} \rangle \).
7. Set \( w_i \leftarrow w_i - \eta_t v_{i}^{(t)} \).
8. Project \( w^{(t+1)} \) onto the set of good weights \( \mathcal{W}_{n, \varepsilon} \) (under \( \ell_2 \) distance).
9. Return \( w^{(t_f)} \), where \( t_f = \arg \min_t \| M^{(t)} \| \).

In particular, we will analyze Algorithm 5 and prove the following set of guarantees.

**Lemma H.4.** Let \( \varepsilon \) be a sufficiently small constant and \( \{x_i\}_{i=1}^{n} \) be \( n \) points in \( \mathbb{R}^d \). Assume the following (deterministic) conditions hold.
(i) There exists \( \nu \in \mathbb{R}^d \) and \( w \in \mathcal{W}_{n,e} \) such that
\[
\left\| \sum_{i=1}^{n} w_i (x_i - \nu) (x_i - \nu)^T \right\| \leq 1 + O (\varepsilon \log (1/\varepsilon)). \tag{H.4}
\]

(ii) If \( \|M(w)\| \leq 1 + \lambda \), for some \( \lambda \geq 0 \), then for any \( w \in \mathcal{W}_{n,e} \),
\[
\|\nu - \mu(w)\| \leq \frac{1}{1 - \varepsilon} \left( \sqrt{\varepsilon \lambda} + C \varepsilon \sqrt{\log(1/\varepsilon)} \right), \tag{H.5}
\]
Then, given \( \{x_i\}_{i=1}^{n} \), a failure rate \( \delta \) and \( \rho \) such that \( \rho \geq \tau_i^{(t)} \) for all \( i \) and \( t \), Algorithm 5 finds \( w' \in \mathcal{W}_{n,e} \) such that
\[
\|M(w')\| \leq 1 + O (\varepsilon \log (1/\varepsilon)), \tag{H.6}
\]
with probability at least \( 1 - \delta \).

The algorithm terminates in \( T = \tilde{O}(n\rho^2/\varepsilon^2) \) iterations. Further, if \( T = O(\text{poly}(n, d)) \), then each iteration takes \( \tilde{O}(nd \log (1/\delta)/\varepsilon^2) \) time.

Proof. If the algorithm gets early stopped, then \( \|M(t)\| \leq 1 + O(\varepsilon^2) \), so assumption (H.4) guarantees that \( \mu(w^{(t)}) \) achieves the desired bound (H.6). We now assume that \( \|M(t)\| > 1 \) for any \( t \).

By the regret bound (Lemma H.1) and the inequality \( \langle w^{(t)}, \tau^{(t)} \rangle \geq (1 - \varepsilon^2) \|M^{(t)}\| \), for a \( w \) that satisfies assumption (H.4)
\[
\frac{1 - \varepsilon^2}{T} \sum_{t=1}^{T} \|M^{(t)}\| \leq \frac{1}{T} \sum_{t=1}^{T} \langle w, \tau^{(t)} \rangle + \frac{3LR}{2\sqrt{T}}, \tag{H.7}
\]
where \( L = \sqrt{\rho} \) and \( R = \sqrt{2} \). For the first term, note that we may apply Claim F.3 and obtain
\[
\frac{1}{T} \sum_{t=1}^{T} \langle w, \tau^{(t)} \rangle \leq 1 + O (\varepsilon \log(1/\varepsilon)) + \frac{2\varepsilon}{(1 - \varepsilon)^2} \frac{1}{T} \sum_{t=1}^{T} \|M^{(t)}\| - \frac{2\varepsilon}{(1 - \varepsilon)^2}
\]
By setting \( T = 3L^2R^2/\varepsilon^2 = O(n\rho^2/\varepsilon^2) \), we can bound the second term by \( O(\varepsilon) \)
Substituting the bounds back into (H.7), we obtain
\[
\frac{1 - \varepsilon^2}{T} \sum_{t=1}^{T} \|M^{(t)}\| \leq 1 - \frac{2\varepsilon}{(1 - \varepsilon)^2} + O(\varepsilon \log(1/\varepsilon)) + \frac{1}{T} \sum_{t=1}^{T} \frac{2\varepsilon}{(1 - \varepsilon)^2} \|M^{(t)}\|
\]
For sufficiently small \( \varepsilon \), we can move the last term to the left side and divide through. This immediately yields that
\[
\frac{1}{T} \sum_{t=1}^{T} \|M^{(t)}\| \leq 1 + O(\varepsilon \log(1/\varepsilon)).
\]
The run-time follows from the cost of computing \((1 - \varepsilon^2)\)-approximate largest eigenvector via power iteration.

Using the same argument for Theorem F.4, Lemma H.4 implies the following theorem.

**Theorem H.5.** Let \( S = \{x_i\}_{i=1}^{n} \) be an \( \epsilon \)-corrupted set of \( n \) samples from a sub-gaussian distribution over \( \mathbb{R}^d \), with mean \( \mu \) and identity covariance. Suppose \( n \geq \tilde{\Omega}(d/\varepsilon^2) \). Then given \( S \), there is an algorithm (based on Algorithm 5) that finds \( \tilde{\mu} \) such that with high constant probability \( \|\tilde{\mu} - \mu\| \leq O \left( \varepsilon \sqrt{\log(1/\varepsilon)} \right) \).

The algorithm runs in \( \tilde{O}(nd^2/\varepsilon^2) \) iterations and \( \tilde{O}(n^2 d^3/\varepsilon^2) \) total time.
H.3 Equivalence with \cite{9}

The recent work of Cheng, Diakonikolas, Ge and Soltanolkotabi \cite{9} studies a gradient-descent-based algorithm for solving the following non-convex formulation of robust mean estimation.

\[
\min \| \Sigma_w \| \quad \text{such that} \quad w \in \mathcal{W}_{n,\epsilon}.
\]

where \( \Sigma_w = \sum_{i=1}^n w_i (x_i - \mu(w)) (x - \mu(w))^\top \). This is equivalent to

\[
\min_w \max_{u \in S^{d-1}} F(w, u) = u^\top \Sigma_w u \quad \text{such that} \quad w \in \mathcal{W}_{n,\epsilon}.
\]

The sub-gradient of \( F(w, u) \) with respect to \( w \) (for a fixed \( u \)) is given by

\[
\nabla_u F(w, u) = X u \odot X u - 2 (w^\top X u) X u, \tag{H.8}
\]

where \( X \in \mathbb{R}^{n \times d} \) is the data matrix whose the \( i \)th row is \( x_i \).

Based on the observation, they consider and analyze an algorithm that computes a (approximately) maximizing \( u \) and performs a projected gradient descent on \( w \) each iteration.

Since Algorithm 4 can be directly applied to the same robust setting (Corollary C.4), it is natural to consider the relationships between the two algorithms. Indeed, one can argue that they are essentially the same. First, we unpack our gradient update (i.e., the spectral scores) of iteration \( t \). Note that

\[
\nabla_i f_i(w^{(t)}) = r_i^{(t)} = \left( v^{(t)}, x_i - \nu^{(t)} \right)^2 = \left( v^{(t)}, x_i \right)^2 + \left( v^{(t)}, \nu^{(t)} \right)^2 - 2 \left( v^{(t)}, x_i \right) \left( v^{(t)}, \nu^{(t)} \right) = \left( X v^{(t)} \odot X v^{(t)} \right)_i + \left( w^{(t)} X v^{(t)} \right)^2 - 2 \left( w^{(t)} X v^{(t)} \right) \left( X v^{(t)} \right)_i
\]

since \( \nu^{(t)} = \sum_i w_i^{(t)} x_i = X^T w^{(t)} \), where \( \odot \) denotes entrywise product of vectors. Let \( C_t = w^{(t)} X v^{(t)} \). Therefore, we can rewrite the gradient as

\[
\nabla f_i(w^{(t)}) = C_t^2 \cdot \mathbb{1}_n + X w^{(t)} X v^{(t)} - 2C_t \cdot X v^{(t)}
\]

Note that the gradient (H.8) used in \cite{9} is exactly the same as above, except without the term of all-one vector \( C_t^2 \cdot \mathbb{1}_n \). In the gradient update step, the additional term reduces the weight of every point uniformly by the same quantity \( C_t^2 \). However, observe that by Pythagorean theorem, the (Euclidean) projection onto \( \mathcal{W}_{n,\epsilon} \) can be decomposed into two (sequential) steps: (1) first an orthogonal projection onto the affine subspace containing \( \mathcal{W}_{n,\epsilon} \), and then (2) a projection onto \( \mathcal{W}_{n,\epsilon} \) itself. Note that reducing each coordinate by the same quantity or not results in the same vector by the first step. Therefore, the two algorithms yield the same sequence of iterates \( (w^{(t)})_t \).