CS 224 Fall 2023 Scribes: Kerem Dayi
October 11th, 2023

Lecture 10: Robust statistics II: list-decodable learning, subspace isotropic filtering

1 Motivation

In the previous lecture, we looked at the problem of recovering the true mean of a distribution when we had samples from an unknown distribution q, where an η fraction could be adversarially corrupted. In that setting, there is a fundamental limitation that imposes strict upper bounds on η. For instance, if $\eta \geq 1/2$, there is no hope that we can recover the true mean μ^* as there is no way of identifying which part of the distribution comes from the adversary vs. true dataset.

In this lecture, we will investigate a surprising result: If we relax our goal to outputting a set of 'means,' one of which is close to the true mean, then we can actually create an efficient algorithm that even works when $\eta \geq 1/2$. So, the setup is as follows.

2 Setup

Let $0 < \alpha < 1/2$ be a small constant, denoting the fraction of good points (non-corrupted). Let q be an arbitrary distribution with mean $\mu \in \mathbb{R}^d$ and covariance Σ with $\Sigma \preceq \mathrm{Id}_d$. Nature draws samples $x_1^*, \ldots, x_n^*$ from q, and an adversary corrupts an arbitrary $1-\alpha$ fraction of this dataset. We are given the corrupted dataset $\{x_i\}_{i=1}^n$.

Notice that the number of corrupted points is larger than the number of good points. For instance, the adversary can corrupt the dataset by creating clusters of size αn, and there would be many different explanations for the data. In this case, recovering the true mean can be impossible. To get around this, we change the problem formulation in the following way.

Goal: Output a list of mean estimates $\hat{\mu}_1, \ldots, \hat{\mu}_m$ for $m = O(1/\alpha)$ such that one of these is close to the true mean, i.e., $\exists j \in [m]$ such that $\left\|\mu - \hat{\mu}_j\right\|$ is small.

Here, you might wonder what a good baseline for this problem is. Let's consider the following scenario. If the corrupted dataset is a mixture of $k = O(1/\alpha)$ bounded-covariance distributions, then each cluster has radius $\approx \sqrt{d}$. Then, projecting the clusters onto the subspace spanned by the means, then we expect each cluster to have radius $\sqrt{k}$. So, we should expect that as long as the clusters are $\sqrt{k}$-separated,

we can hope to produce a list of estimates, at least one of which is $O(\sqrt{k}) = O(1/\sqrt{\alpha})$ close to μ.

3 Algorithm

In this lecture, we look at a simplified version of the algorithm given in [DKK+21]. Their main theorem, Theorem 1, is given below:

Theorem 1 (informal)**.** *Let $\alpha \in (0, 1/2)$. Let $\mathcal{D}$ be a distribution with unknown mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \preceq \sigma^2 \mathrm{Id}$. Let $T \subset \mathbb{R}^d$ have $|T| = n$, an α fraction of which is drawn independently from $\sim \mathcal{D}$. For $n = \Omega\left(\frac{d}{\alpha}\right)$, algorithm outputs a list of $m = O(1/\alpha)$ estimates $\{\hat{\mu}_j\}_{j\in[m]}$ so that $\min_{j\in[m]} \left\|\mu - \hat{\mu}_j\right\| = O(\sigma\sqrt{1/\alpha})$ with high probability. The runtime of the algorithm is*

$$\tilde{O}\left(\frac{nd}{\alpha} + \frac{1}{\alpha^6}\right)$$

In the paper, in theorem 2, they also talk about a way of getting rid of the $1/\alpha^6$ dependency, by adding a $\sqrt{\log \alpha^{-1}}$ term to the estimation error. Now, we will give our setup and simplified algorithm that runs in $\tilde{O}(n^2 d/\alpha)$ time. First, to not worry about concentration inequalities, we assume the following[1]

Assumption 1. *There is a $\Omega(\alpha)$ fraction of "good points" $G \subseteq [n]$ such that*

$$\left\| \frac{1}{|G|} \sum_{i\in G} (x_i - \mu)(x_i - \mu)^\top \right\|_{\mathsf{op}} \lesssim 1$$

3.1 Finding subspace V close to true mean suffices

Observation: Suppose we could find a subspace $V \subset \mathbb{R}^d$ s.t. μ is close to V, i.e.

$$\left\|\Pi_V^\perp \mu\right\| \lesssim O(1/\sqrt{\alpha})$$

where $\Pi_V^\perp$ is the projector to the orthogonal complement of V. Then, we could just do the following

1. Pick $O(1/\alpha)$ points at random from the dataset.

[1]This holds when the adversary is additive, ie. there are αn i.i.d. draws from q in the dataset. See [CSV17]. Also see problem set 3, problem 1 for justifications of assumptions like these.

2. Project to V and output.

To see why this works, we will bound the distance $\left\|\Pi_V(x-\mu)\right\|_2$ with high probability. Notice that the covariance matrix of $\Pi_V(x_i-\mu)$ satisfies

$$\begin{aligned}\frac{1}{|G|}\sum_{i\in G}\Pi_V(x_i-\mu)(x_i-\mu)^\top\Pi_V &= \Pi_V\frac{1}{|G|}\sum_{i\in G}(x_i-\mu)(x_i-\mu)^\top\Pi_V \\ &\preceq \Pi_V \mathrm{Id}\Pi_V = \Pi_V\end{aligned}$$

where G is the subset of 'good' points that satisfy Assumption 1 and the last step follows from Assumption 1. Then, taking traces, we have

$$\frac{1}{|G|}\sum_{i\in G}\left\|\Pi_V(x_i-\mu)\right\|_2^2 \leq \mathrm{Tr}\Pi_V = \dim(V) = O(1/\alpha)$$

Then, the expected square distance of the projection a point in G to the projection of μ is $O(1/\alpha)$, where the expectation is due to the randomness of randomly selecting a point from G. By markov we can conclude that 99% of points in G satisfy $\left\|\Pi_V(x_i-\mu)\right\| = cO(1/\alpha) = O(1/\alpha)$ by choosing c suitably. Then, because G contains αn points, choosing $\Theta(1/\alpha)$ (with a sufficiently large constant term) points at random ensures that we get a point from G with high probability, and since a high fraction of points in G satisfy the bound above, we have that with high probability we output a point that has projection close to the projection of the mean. Because we assumed the component outside of V is small, we are done.

Now, we will actually change our claim a bit. In fact it suffices to find a subspace V for which we can estimate $\Pi_V^\perp\mu$ up to small error $O(1/\sqrt{\alpha})$. We would just need to add our estimate (correction) to the projections $\Pi_V x_i$ we obtained in the previous part.

3.2 Finding subspace V with iterative filtering

Now, we get to the main algorithm which will enable us to find a subspace V for which $\Pi_V^\perp\mu$ can be estimated to small error. With a similar approach to last lecture, we want to iteratively elimiate outliers by looking at how they contribute to the top $k = O(1/\alpha)$ components of Σ_w. Recall, from last lecture that we defined the following weighted mean and covariances

$$\begin{aligned}\mu_w &= \frac{1}{\sum_i w_i}\sum_i w_i x_i \\ \Sigma_w &= \frac{1}{\sum_i w_i}\sum_i w_i(x_i-\mu_w)(x_i-\mu_w)^\top\end{aligned}$$

Algorithm 1: SUBSPACE ISOTROPIC FILTERING (SIFT)

Input: Corrupted dataset $\{x_i\}_i$, number of components $k = O(1/\alpha)$

Output: Estimates $\{\hat{\mu}_i\}_{i=1}^m$

$w_i \leftarrow 1/n$ `// initially assign equal weights`

while $\lambda_k(\Sigma_w) \gtrsim \frac{1}{\sqrt{\sum_i w_i}}$ **do**

  $V_w \leftarrow$ top k eigenvectors of Σ_w

  $\Sigma_w^{(k)} = V_w^\top \Sigma_w V$ `// top k 'components' of` Σ_w

  $\tau_i \leftarrow \left\| (\Sigma_w^{(k)})^{-1/2} V_w^\top (x_i - \mu) \right\|_2^2$

  $\tau_{\max} \leftarrow \max_{i, w_i > 0} \tau_i$

  $w_i = w_i(1 - \tau_i/\tau_{\max})$

end

$\tilde{w} \leftarrow \Pi_{V_w}^\perp \mu_w$ `// orthogonal component`

Pick $\Theta(1/\alpha)$ points at random, $\{x_i\}_{i=1}^T$

Return $\{\Pi_V x_i + \tilde{w}\}_{i=1}^T$

We define the SIFT algorithm in Algorithm 1.

In Algorithm 1, we initialize the weights w_i equally and then at each step, downweigh the points that contribute the most to the top k components of Σ_w. We do this until the k'th eigenvalue of Σ_w is still large, which allows us to bound the error in $V_w^\perp \mu$ when we terminate. Notice that in τ_i we 'whiten' the projected distance to mean $V_w^\top (x_i - \mu)$ with $(\Sigma_w^{(k)})^{-1/2}$ to make sure components contribute equally to the score. In Table 1, we compare this algorithm to the algorithm from the previous lecture where we were only considering the top eigenvector while eliminating outliers.

	$\eta < 1/2$ (previous lecture)	list decodable learning
Invariant on w_i	$\sum_{i \in \mathsf{clean}} \left(\frac{1}{n} - w_i\right) < \sum_{i \in \mathsf{bad}} \left(\frac{1}{n} - w_i\right)$	$\sum_{i \in \mathsf{clean}} w_i \geq \alpha \sqrt{\sum_{i \in [n]} w_i}$
Termination cond.	$\|\Sigma_w\|_{\mathsf{op}} \lesssim 1$	$\lambda_k(\Sigma_w) \lesssim \frac{1}{\sqrt{\sum_i w_i}}$, $k = \Theta(1/\alpha)$
Scores (τ_i)	$\tau_i = \langle u, x_i - \mu_w \rangle^2$, u top eig.vec.	$\tau_i = \left\| (\Sigma_w^{(k)})^{-1/2} V_w^\top (x_i - \mu_w) \right\|_2^2$
Condition on τ_i	$\sum_{i \in \mathsf{clean}} w_i \tau_i < \frac{1}{2} \sum_{i \in \mathsf{all}} w_i \tau_i$	$\frac{\sum_{i \in \mathsf{clean}} w_i \tau_i}{\sum_{i \in \mathsf{clean}} w_i} \leq \frac{1}{2} \frac{\sum_{i \in \mathsf{all}} w_i \tau_i}{\sum_{i \in \mathsf{all}} w_i}$
Spectral Signature	$\|\mu_w - \mu\| \lesssim \sqrt{\eta}(1 + \|\Sigma_w\|_{\mathsf{op}}^{1/2})$	$\|\mu_w - \mu\| \lesssim \sqrt{\frac{1 + \sqrt{\sum_i w_i} \|\Sigma_w\|_{\mathsf{op}}}{\alpha}}$

Table 1: Comparison of SIFT to Simple iterative filtering from previous lecture

Here, the condition on τ_i is the condition that has to hold for the invariant to

be maintained. Notice that the termination condition now looks at the magnitude of the k'th eigenvector instead of the operator norm. Similarly, we have that the condition on τ_i is now normalized, by dividing by the sum of the weights.

4 Analysis of SIFT

Here, we analyse SIFT. The goal is to show that

1. When the algorithm terminates, the output is correct.
2. We maintain the invariant on w_i given in Table 1 when the condition on τ_i holds.
3. The condition on the τ_i holds when the algorithm is still running .
4. Prove the spectral signature lemma.

4.1 Termination condition implies output is correct

Recall that V_w has columns which are top-k eigenvectors of Σ_w. We initially prove that if we hit the termination condition, we estimate $\Pi_{V_w}^{\perp}\mu$ to small error. Notice that when we apply the spectral signature lemma to data projected to $V_w^{\perp}$, we get

$$\begin{aligned}\left\|\Pi_{V_w}^{\perp}\mu_w - \Pi_{V_w}^{\perp}\mu\right\| &\lesssim \frac{1}{\sqrt{\alpha}}\sqrt{1+\sqrt{\sum_i w_i}\left\|\Pi_{V_w}^{\perp}\Sigma_w\Pi_{V_w}^{\perp}\right\|_{\mathsf{op}}} \\ &\leq \frac{1}{\sqrt{\alpha}}\sqrt{1+\sqrt{\sum w_i}\lambda_k(\Sigma_w)} \\ &\lesssim \frac{1}{\sqrt{\alpha}}\sqrt{1+\sqrt{\sum w_i}/\sqrt{\sum w_i}} \lesssim \frac{1}{\sqrt{\alpha}}\end{aligned}$$

So, we have a $O(1/\alpha)$ dimensional subspace V_w and an $O(1/\sqrt{\alpha})$-accurate estimate of $\Pi_{V_w}^{\perp}\mu$. This is exactly the condition required to run our random selection algorithm as described in Section 3.1.

4.2 Maintaining invariant given condition on τ_i

We show that if the condition on τ_i holds, ie

$$\frac{\sum_{i\in\mathsf{clean}} w_i\tau_i}{\sum_{i\in\mathsf{clean}} w_i} \leq \frac{1}{2}\frac{\sum_{i\in\mathsf{all}} w_i\tau_i}{\sum_{i\in\mathsf{all}} w_i}$$

then, downweighting maintains the invariant on the w_i, which is $\sum_{i\in\mathsf{clean}} w_i \geq \alpha\sqrt{\sum_{i\in\mathsf{all}} w_i}$. First, note that $w_i' \leftarrow w_i\left(1-\frac{\tau_i}{\tau_{\max}}\right)$. Then, it will suffice to show that

$$\frac{\sum_{i\in\mathsf{clean}} w_i'}{\sum_{i\in\mathsf{clean}} w_i} \geq \sqrt{\frac{\sum_{i\in\mathsf{all}} w_i'}{\sum_{i\in\mathsf{all}} w_i}}$$

This is because multiplying the cross terms tells us that the relative weight of the clean terms has increased. Now, we have

$$\begin{aligned}
\frac{\sum_{i\in\mathsf{clean}} w_i'}{\sum_{i\in\mathsf{clean}} w_i} = \frac{\sum_{i\in\mathsf{clean}} w_i\left(1-\frac{\tau_i}{\tau_{\max}}\right)}{\sum_{i\in\mathsf{clean}} w_i} &= 1 - \frac{1}{\tau_{\max}}\cdot\frac{\sum_{i\in\mathsf{clean}} w_i\tau_i}{\sum_{i\in\mathsf{clean}} w_i} \\
&\overset{(a)}{\geq} 1 - \frac{1}{\tau_{\max}}\cdot\frac{1}{2}\frac{\sum_{i\in\mathsf{all}} w_i\tau_i}{\sum_{i\in\mathsf{all}} w_i} \\
&\overset{(b)}{=} 1 - \frac{1}{2}\left(1-\frac{\sum_{i\in\mathsf{all}} w_i'}{\sum_{i\in\mathsf{all}} w_i}\right) \\
&\overset{(c)}{\geq} \sqrt{\frac{\sum_{i\in\mathsf{all}} w_i'}{\sum_{i\in\mathsf{all}} w_i}}
\end{aligned}$$

where (a) follows from the condition on τ_i, (b) follows from the definition of w_i', and (c) follows from the inequality $1-\frac{1}{2}x \geq \sqrt{1-x}$ when $x<1$.[2]

4.3 Maintaining Condition on τ_i given we haven't terminated

Now, we show that if we haven't hit the termination condition, the condition on τ_i holds. We want to show

$$\frac{1}{\sum_{i\in\mathsf{clean}} w_i}\sum_{i\in\mathsf{clean}} w_i\tau_i \leq \frac{1}{2}\frac{1}{\sum_{i\in\mathsf{all}} w_i}\sum_{i\in\mathsf{all}} w_i\tau_i \tag{1}$$

We initially note that

$$\begin{aligned}
\tau_i &= \left\|(\Sigma_w^{(k)})^{-1/2}V_w^\top(x_i-\mu_w)\right\|^2 \\
&= \mathrm{Tr}\left((\Sigma_w^{(k)})^{-1/2}V_w^T(x_i-\mu_w)(x_i-\mu)^\top V_w(\Sigma_w^{(k)})^{-1/2}\right) \\
&\overset{(a)}{=} \mathrm{Tr}\left((\Sigma_w^{(k)})^{-1/2}(\Sigma_w^{(k)})^{-1/2}V_w^T(x_i-\mu_w)(x_i-\mu)^\top V_w\right) \\
&= \langle(\Sigma_w^{(k)})^{-1}, V_w^\top(x_i-\mu_w)(x_i-\mu_w)^\top V)\rangle
\end{aligned}$$

[2]We have $(1-x/2)^2 = 1-x+x^2/4 \geq 1-x$. Taking square roots of both sides gives desired result.

where we used the cyclic property of trace in (a), and we refer to the trace inner product at the end. So,

$$\begin{aligned}\frac{1}{2}\frac{1}{\sum_{i\in\mathsf{all}} w_i}\sum_{i\in\mathsf{all}} w_i\tau_i &= \frac{1}{2}\frac{1}{\sum_{i\in\mathsf{all}} w_i}\sum_{i\in\mathsf{all}}\langle(\Sigma_w^{(k)})^{-1}, V_w^\top(x_i-\mu_w)(x_i-\mu_w)^\top V)\rangle \\ &= \frac{1}{2}\langle(\Sigma_w^{(k)})^{-1}, V_w^\top\Sigma_w V_w\rangle \\ &= \frac{1}{2}\langle(\Sigma_w^{(k)})^{-1}, \Sigma_w^{(k)}\rangle = \frac{1}{2}\mathrm{Tr}(\mathrm{Id}_k) = k/2\end{aligned}$$

which is the RHS of Eq. (1). Then, for the LHS, we initially do the following for $x_i - \mu_w$

$$\left(x_i - \frac{1}{|G|}\sum_{i\in\mathsf{clean}} w_i x_i\right) + \left(\frac{1}{|G|}\sum_{i\in\mathsf{clean}} w_i x_i - \mu_w\right)$$

With this motivation, let $\mu_{w,g} = \frac{1}{\sum_{i\in\mathsf{clean}}}\sum_{i\in\mathsf{clean}} w_i x_i$ and we define the 'good covariance'

$$\Sigma_{w,g} \triangleq \frac{1}{\sum_{i\in\mathsf{clean}} w_i}\sum_{i\in\mathsf{clean}} w_i(x_i-\mu_{w,g})(x_i-\mu_{w,g})^\top$$

So that

$$\begin{aligned}\tau_i &\le 2\left\|(\Sigma_w^{(k)})^{-1/2}V_w^\top(x_i-\mu_{w,g})\right\|^2 + 2\left\|(\Sigma_w^{(k)})^{-1/2}V_w^\top(\mu_{w,g}-\mu)\right\|^2 \\ &= 2\gamma_i + 2\delta_i \end{aligned} \tag{2}$$

where we used the inequality $\|a+b\|^2 \le 2\|a\|^2 + \|b\|^2$ and $\mu_{w,g}$ is the mean of the good points. Then,

$$\frac{2}{\sum_{i\in\mathsf{clean}} w_i}\sum_{i\in\mathsf{clean}} w_i\gamma_i = 2\langle(\Sigma_w^{(k)})^{-1}, V_w^\top\Sigma_{w,g}V_w\rangle$$

similar to before. Now, we note that $\Sigma_{w,g} \preceq \frac{1}{\sum_{i\in\mathsf{clean}} w_i}\sum_{i\in\mathsf{clean}} w_i(x_i-\mu)(x_i-\mu)^\top$ because $\Sigma_{w,g}$ is the empirical covariance of the good points. Then, using $w_i \le 1/n \le \alpha/|G|$ we get

$$\begin{aligned}\Sigma_{w,g} &\preceq \frac{1}{\sum_{i\in\mathsf{clean}} w_i}\sum_{i\in\mathsf{clean}} w_i(x_i-\mu)(x_i-\mu)^\top \\ &\preceq \frac{\alpha}{\sum_{i\in\mathsf{clean}} w_i}\cdot\frac{1}{|G|}\sum_{i\in\mathsf{clean}}(x_i-\mu)(x_i-\mu)^\top \\ &\overset{(a)}{\preceq} \frac{\alpha}{\sum_{i\in\mathsf{clean}} w_i}\cdot\mathrm{Id}\end{aligned}$$

where we used Assumption 1 in (a). Then,

$$2\langle(\Sigma_w^{(k)})^{-1}, V_w^\top \Sigma_{w,g} V_w\rangle \leq \frac{2\alpha}{\sum_{i\in\mathsf{clean}} w_i} \cdot \mathrm{Tr}(\Sigma_w^{(k)})^{-1}$$

Then, because the eigenvalues of $\mathrm{Tr}(\Sigma_w^{(k)})^{-1}$ are $\lesssim \sqrt{\sum_{i\in\mathsf{all}} w_i}$ we have

$$\frac{2\alpha}{\sum_{i\in\mathsf{clean}} w_i} \cdot \mathrm{Tr}(\Sigma_w^{(k)})^{-1} \lesssim \frac{2\alpha\sqrt{\sum_{i\in\mathsf{all}} w_i}}{\sum_{i\in\mathsf{clean}} w_i} \cdot k \leq^* k/4$$

where the last inequality can hold by choosing the constant in the termination condition large enough. So, we have bounded the first term by $k/4$ and it suffices to bound the second one by $k/4$ too. So,

$$\frac{2}{\sum_{i\in\mathsf{clean}} w_i} \sum_{i\in\mathsf{clean}} w_i\delta_i = 2\left\|(\Sigma_w^{(k)})^{-1/2} V_w^\top(\mu_w - \mu_{w,g})\right\|_2^2$$

Now, we note that $\mathbb{E}[x]\mathbb{E}[x]^\top \preceq \mathbb{E}[xx^\top]$ by Jensen. So,

$$\begin{aligned}
(\mu_w - \mu_{w,g})(\mu_w - \mu_{w,g})^\top &\preceq \frac{1}{\sum_{i\in\mathsf{clean}} w_i} \sum_{i\in\mathsf{clean}} w_i(x_i - \mu_w)(x_i - \mu_w)^\top \\
&\overset{(a)}{\preceq} \frac{1}{\sum_{i\in\mathsf{clean}} w_i} \sum_{i\in\mathsf{all}} w_i(x_i - \mu_w)(x_i - \mu_w)^\top \\
&\preceq \frac{\sum_{i\in\mathsf{all}} w_i}{\sum_{i\in\mathsf{clean}} w_i}\Sigma_w
\end{aligned}$$

So,

$$2\left\|(\Sigma_w^{(k)})^{-1/2} V_w^\top(\mu_w - \mu_{w,g})\right\|_2^2 \leq \frac{2\sum_{i\in\mathsf{all}} w_i}{\sum_{i\in\mathsf{clean}} w_i} \leq 2\frac{\sqrt{\sum_{i\in\mathsf{all}} w_i}}{\alpha} \leq \frac{2}{\alpha} \leq \frac{k}{4}$$

by taking $k \geq 8/\alpha$ since $k = \Theta(1/\alpha)$. Therefore, in Eq. (2) we have bounded both terms by $k/4$, bounding the LHS by $k/2$. Since in Eq. (1) the RHS was $k/2$, we have the desired result.

4.4 Spectral Signature Lemma

Now, what remains to show to conclude the analysis of SIFT is the spectral signature lemma. I.e.,

$$\|\mu_w - \mu\| \lesssim \frac{1}{\sqrt{\alpha}}\sqrt{1 + \frac{1}{\sqrt{\sum_i w_i}}\|\Sigma_w\|_{\mathsf{op}}}$$

To do this, we will show that both μ_w and μ are close to $\hat{\mu}$ where

$$\hat{\mu} \triangleq \frac{1}{\langle w, w^* \rangle} \sum_{i \in \mathsf{all}} w_i w_i^* x_i$$

where $w_i^* \triangleq \frac{1}{|G|} \cdot 1[\mathsf{i} \in \mathsf{G}]$ is the indicator of the good points. Here, $\hat{\mu}$ is the averaging of the good points under the weighting given by w. We want to do this because we can relate the distance $\|\mu - \mu_w\|$ to $\|\mu - \hat{\mu}\|$ and $\|\mu_w - \hat{\mu}\|$.

Then, we start with $\|\hat{\mu} - \mu\|^2$. We have

$$\begin{aligned}
\|\hat{\mu} - \mu\|^2 &= \sup_{u \in \mathbb{S}^{d-1}} \langle \hat{\mu} - \mu, u \rangle^2 \\
&= \sup_{u \in \mathbb{S}^{d-1}} \langle \frac{1}{\langle w, w^* \rangle} \sum_{i \in \mathsf{all}} w_i w_i^* (x_i - \mu), u \rangle^2 \\
&\overset{(a)}{\leq} \sup_{u \in \mathbb{S}^{d-1}} \frac{1}{\langle w^*, w \rangle} \sum_i w_i w_i^* \langle x_i - \mu, u \rangle^2 \\
&\leq \frac{1}{n} \frac{1}{\langle w^*, w \rangle} \sup_{u \in \mathbb{S}^{d-1}} \sum_i w_i^* \langle x_i - \mu, u \rangle^2 \\
&\overset{(b)}{\lesssim} \frac{1}{n \langle w, w^* \rangle}
\end{aligned}$$

where (a) follows from the convexity of $f(x) = x^2$.[3], and (b) follows from the covariance bound on the good points, i.e. Assumption 1. Now, note that

$$\langle w, w^* \rangle = \frac{1}{|G|} \sum_{i \in \mathsf{clean}} w_i \geq \frac{1}{|G|} \alpha \sqrt{\sum_{i \in \mathsf{all}} w_i} \tag{3}$$

from the invariant on the w_i. We want to lower bound the sum of all w_i. Therefore, we use the invariant (lower bound on good w_i) to get

$$\sum_{i \in \mathsf{all}} w_i \geq \sum_{i \in \mathsf{clean}} w_i \geq \alpha \sqrt{\sum_{i \in \mathsf{all}} w_i}$$

So, $\sum_{i \in \mathsf{all}} w_i \geq \alpha^2$. Then, combining with Eq. (3), we have

$$\frac{1}{n \langle w, w^* \rangle} \leq \frac{|G|}{\alpha n} \frac{1}{\sqrt{\sum_{i \in \mathsf{all}} w_i}} \leq 1 \cdot \frac{1}{\sqrt{\alpha^2}} = 1/\alpha$$

Hence, we conclude that $\|\mu - \hat{\mu}\| \lesssim \frac{1}{\alpha}$.

[3]If f is convex and $\sum_i a_i = 1$ with $a_i \geq 0$, then $f(\sum_i a_i x_i) \leq \sum_i a_i f(x_i)$ This is by definition of convexity.

Now, we look at the distnace of weighted mean μ_w to the true mean μ, $\|\mu_w - \hat{\mu}\|$. Using a similar convexity inequality to the previous part, we have

$$\begin{aligned}
\|\hat{\mu} - \mu_w\|^2 &\leq \sup_{u \in \mathbb{S}^{d-1}} \frac{1}{\langle w, w^* \rangle} \sum_{i \in \mathsf{all}} w_i w_i^* \langle x_i - \mu_w, u \rangle^2 \\
&\overset{(a)}{\leq} \sup_{u \in \mathbb{S}^{d-1}} \frac{1}{\langle w, w^* \rangle} \frac{1}{|G|} \left(\sum_{i \in \mathsf{all}} w_i \right) \frac{1}{\sum_{i \in \mathsf{all}} w_i} \sum_{i \in \mathsf{all}} w_i \langle x_i - \mu_w, u \rangle^2 \\
&= \frac{\sum_{i \in \mathsf{all}} w_i}{\langle w, w^* \rangle |G|} \sup_{u \in \mathbb{S}^{d-1}} u^\top \Sigma_w u \\
&= \frac{\sum_{i \in \mathsf{all}} w_i}{\langle w, w^* \rangle |G|} \|\Sigma_w\|_{\mathsf{op}}
\end{aligned}$$

where (a) follows from $w_i^* \leq \frac{1}{|G|}$. Then, using the bound for $\langle w, w^* \rangle$ from Eq. (3), we have

$$\frac{\sum_{i \in \mathsf{all}} w_i}{\langle w, w^* \rangle |G|} \|\Sigma_w\|_{\mathsf{op}} \leq \frac{|G| \sum_{i \in \mathsf{all}} w_i}{|G| \alpha \sqrt{\sum_{i \in \mathsf{all}} w_i}} \|\Sigma_w\|_{\mathsf{op}} = \frac{1}{\alpha} \sqrt{\sum_{i \in \mathsf{all}} w_i} \|\Sigma_w\|_{\mathsf{op}}$$

Combining this with the bound of $\|\mu - \hat{\mu}\|$, we get

$$\|\mu - \mu_w\|^2 \leq 2\|\hat{\mu} - \mu\|^2 + 2\left\|\hat{\mu} - \mu_w^2\right\| \lesssim \frac{1}{\alpha} \left(1 + \sqrt{\sum_{i \in \mathsf{all}} w_i} \|\Sigma_w\|_{\mathsf{op}} \right)$$

as desired.

References

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

[DKK+21] Ilias Diakonikolas, Daniel Kane, Daniel Kongsgaard, Jerry Li, and Kevin Tian. List-decodable mean estimation in nearly-pca time. *Advances in Neural Information Processing Systems*, 34:10195–10208, 2021.