List-decodable mean estimation

Sitan Chen

October 11, 2023

The following is an exposition of the basic guarantees of the SIFT (Subspace Isotropic FilTering) algorithm from [DKK+21] for list-decodable mean estimation for bounded-covariance distributions. This writeup has not been proofread carefully and is only meant to serve as supplemental notes for Lecture 10 of CS 224.

1 Setup and notation

There is a distribution q with mean $\mu^* \in \mathbb{R}^d$ and covariance $\Sigma \preceq \mathbb{1}$. Let $\alpha > 0$ be a small parameter corresponding to the fraction of inliers. We are given samples $\{X_i\}_{i \in T}$, such that an α fraction of them are i.i.d. draws from q, and the rest are arbitrary. Our goal is to output a list of estimates $\widehat{\mu}_1, \dots, \widehat{\mu}_m$ for $m = O(1/\alpha)$ such that one of them is guaranteed to satisfy $\|\widehat{\mu}_i - \mu^*\| \lesssim 1/\sqrt{\alpha}$.

Let Δ^T denote the solid probability simplex, i.e. the set of weights $\{w_i\}_{i \in T}$ for which $w_i \geq 0$ and $\sum_i w_i \leq 1$.

Given weights $w \in \Delta^T$ and a subset $T' \subseteq T$, define the weighted mean and covariance by

$$\mu_w(T') \triangleq \frac{1}{\|w_{T'}\|_1} \sum_{i \in T'} w_i X_i \quad \text{and} \quad \Sigma_w(T') \triangleq \frac{1}{\|w_{T'}\|_1} \sum_{i \in T'} w_i (X_i - \mu_w(T'))(X_i - \mu_w(T'))^\intercal \,.$$

The only property we need about the dataset is that there exists a subset of points G for which, for weights w^* given by

$$w_i^* \triangleq \frac{1}{|G|} \cdot \mathbb{1}[i \in G] \,,$$

we have

$$\sum_i w_i^* (X_i - \mu^*)(X_i - \mu^*)^\intercal \preceq \mathbb{1} \,. \tag{1}$$

To give intuition for this, note that if G were simply all uncorrupted points, then the left-hand side is just the empirical covariance of the uncorrupted points, and intuitively this should concentrates around the true covariance of q, which is bounded by $\mathbb{1}$. Unfortunately, this is not quite true because we are making a weak assumption on the tails of q. A classical result of Rudelson [Rud99] shows that the empirical covariance has operator norm bounded by that of the true covariance up to a $\log(n)$ factor, but recall that our goal is to obtain dimension-independent bounds.

Instead, it is proven in [CSV17] using tools from spectral sparsification that one can carefully sub-sample the set of uncorrupted points in order to produce G satisfying Eq. (1). Proving this will take us a bit further afield from the key ideas on list-decodable learning, so we refer the interested reader to Proposition 1.1 in [CSV17].

2 Safe scores and saturated weights

Definition 2.1. *A set of scores $\{\tau_i\}_{i\in T}$ is* safe *with respect to a set of weights $w \in \Delta^T$ if*

$$\frac{1}{\|w_G\|_1}\sum_{i\in G} w_i\tau_i \le \frac{1}{2\|w\|_1}\sum_{i\in T} w_i\tau_i\,. \tag{2}$$

Definition 2.2. *A set of weights $w \in \Delta^T$ is* saturated *if $w_i \le 1/n$ for all i and*

$$\|w_G\|_1 \ge \alpha\sqrt{\|w\|_1}\,.$$

Note that the set of uniform weights $w_i = \frac{1}{n}$ is saturated by assumption. We also record the following basic implication of saturation, namely that the total weight cannot be too small.

Fact 2.3. *If w is saturated, then $\|w\|_1 \ge \alpha^2$.*

Proof. This follows from rearranging the following:

$$\|w\|_1 \ge \|w_G\|_1 \ge \alpha\sqrt{\|w\|_1}\,. \qquad \square$$

The next lemma shows that the downweighting rule we used in the previous lecture maintains the invariant that the weights are saturated, provided that the scores used are safe.

Lemma 2.4. *If $\{\tau_i\}$ are safe with respect to a set of saturated weights w, then the weights given by downweighting, that is,*

$$w_i' \triangleq \Big(1 - \frac{\tau_i}{\tau_{\max}}\Big)w_i \;\text{ where }\tau_{\max} \triangleq \max_{i\in T: w_i\neq 0}\tau_i\,,$$

are also saturated.

Proof. Note that it suffices to show that

$$\frac{\|w_G'\|_1}{\|w_G\|_1} \ge \sqrt{\frac{\|w'\|_1}{\|w\|_1}}\,. \tag{3}$$

We can rewrite the left-hand side of Eq. (3) as

$$\begin{aligned}
\frac{\|w_G'\|_1}{\|w_G\|_1} &= 1 - \frac{1}{\|w_G\|_1}\sum_{i\in G}(w_i - w_i') \\
&= 1 - \frac{1}{\|w_G\|_1}\sum_{i\in G}\frac{w_i\tau_i}{\tau_{\max}} \\
&\ge 1 - \frac{1}{2\|w\|_1}\sum_{i\in T}\frac{w_i\tau_i}{\tau_{\max}} \\
&= 1 - \frac{1}{2}\frac{\|w'\|_1}{\|w\|_1}\,,
\end{aligned}$$

where the second step follows by the definition of the downweighting update, and the inequality follows by the assumption that τ's are safe. Finally, note that $1 - x/2 \ge \sqrt{1-x}$ for all $0 \le x \le 1$, so Eq. (3) follows. $\square$

3 Finding a $1/\alpha$-dimensional subspace suffices

First note that if we could somehow produce a subspace V of dimension

$$k = O(1/\alpha)$$

which is very close to the true mean μ^* of the good points, then the following simple algorithm works. Let Π denote the projection to V. Then select $\Omega(1/\alpha)$ random points from T and project them to V. With high probability one of the points x is from G, and with high probability Πx will be $O(\sqrt{\alpha})$-close to $\Pi\mu^*$ (and thus to μ^*) by the following basic fact:

Fact 3.1. *Let q be a distribution over $\mathbb{R}^k$ with mean μ and covariance $\Sigma \preceq \mathbb{1}$. Then for $x \sim q$, $\|x-\mu\| \lesssim \sqrt{k}$ with probability at least $4/5$.*

Proof. $\mathbb{E}\,\|x-\mu\|^2 = \mathrm{Tr}(\Sigma) \le k$, so by Markov's, $\mathbb{P}[\|x-\mu\|^2 \ge 5k] \le 1/5$, and the fact follows. □

Why is it reasonable to hope for the existence of such a subspace? We know that information-theoretically, list-decodable mean estimation is possible, so at the very least there is a computationally inefficient algorithm that outputs a list of $O(1/\alpha)$ candidate means, one of which is guaranteed to be $O(\sqrt{\alpha})$-close to μ^*. We can simply take V to be the span of the these candidate means, and then μ^* would certainly be close to V.

The whole challenge is thus to find V with an efficient algorithm. In fact, this is overkill: if we had an efficient algorithm for producing an $O(1/\alpha)$-dimensional subspace V and an accurate estimate for $\Pi^\perp\mu^*$, where $\Pi^\perp$ is the projector to the orthogonal complement of V, this would also suffice.

This suggests a natural termination condition for the filtering algorithm we will consider: check whether the current weighted covariance matrix has small k-th eigenvalue. Formally, we will terminate as soon as

$$\lambda_k(\Sigma_w(T)) \le \zeta\,,$$

where ζ will be a threshold to be tuned later (we will eventually set $\zeta \asymp 1/\sqrt{\|w\|_1}$).

Intuitively, the point at which this happens will correspond to the point at which we have successfully learned μ^* in all of the directions orthogonal to the top-k eigenspace V_w of the weighted covariance $\Sigma_w(T)$, at which point we can apply the trivial random sampling algorithm above to learn $\Pi\mu^*$.

Here we verify that if we hit the termination condition and w is saturated, then the current weighted mean is $O(1/\sqrt{\alpha})$-close to μ^* in the directions orthogonal to V, which is sufficient for our purposes:

Lemma 3.2. *If $w \in \Delta^T$ is saturated, then*

$$\|\mu_w(T) - \mu^*\| \lesssim \sqrt{\frac{\sqrt{\|w\|_1}}{\alpha}\,\|\Sigma_w(T)\|_{\mathsf{op}} + \frac{1}{\alpha}}\,. \tag{4}$$

To see why this implies what we want, we apply the above lemma to the dataset projected via $\Pi^\perp$ to the orthogonal complement of the top-k eigenspace of $\Sigma_w(T)$. Then the $\|\Sigma_w(T)\|_{\mathsf{op}}$ term is simply the $(k+1)$-st eigenvalue of $\Sigma_w(T)$, which is at most ζ by the termination condition. If we take $\zeta \asymp 1/\sqrt{\|w\|_1}$, then the right-hand side of Eq. (4) is $O(\sqrt{1/\alpha})$ as desired.

Proof of Lemma 3.2. Let $w^* \in \Delta^T$ be given by $w_i^* \triangleq 1/|G| \cdot \mathbb{1}[i \in G]$ so that $\sum_{i\in T} w_i^* X_i = \mu^*$.

We will bound the distance from both $\mu_w(T)$ and μ^* to the following point:

$$\widehat{\mu} \triangleq \frac{1}{\langle w, w^*\rangle} \sum_{i\in T} w_i w_i^* X_i$$

The intuition for $\widehat{\mu}$ is that it is the weighted mean of the dataset where the weights are given by taking w and "tilting" them in the direction of the true weights w^*.

Writing $\|\widehat{\mu} - \mu^*\|^2 = \sup_{u\in\mathbb{S}^{d-1}} \langle \widehat{\mu} - \mu^*\rangle^2$, we have

$$\begin{aligned}
\sup_u \langle \widehat{\mu} - \mu^*\rangle^2 &= \sup_u \Big\langle \sum_{i\in T} \frac{w_i w_i^*}{\langle w, w^*\rangle}(X_i - \mu^*), u\Big\rangle^2 \\
&\le \sup_u \sum_{i\in T} \frac{w_i w_i^*}{\langle w, w^*\rangle}\langle X_i - \mu^*, u\rangle^2 \\
&\le \frac{1}{n\langle w, w^*\rangle} \sup_u \sum_{i\in T} w_i^* \langle X_i - \mu^*, u\rangle^2 \\
&\le \frac{1}{n\langle w, w^*\rangle} = \frac{\alpha}{\|w_G\|_1} \le \frac{1}{\sqrt{\|w\|_1}} \le \frac{1}{\alpha}\,.
\end{aligned}$$

where in the second step we used Jensen's inequality, in the third step we used that $w_i \le 1/n$, in the fourth step we used Eq. (1), in the fifth step we used that $\langle w, w^*\rangle = \frac{1}{|G|}\sum_{i\in G} w_i = \frac{\|w_g\|_1}{\alpha n}$, and in the last two steps we used saturation of w.

Similarly, writing $\|\widehat{\mu} - \mu_w(T)\|^2 = \sup_u \langle \widehat{\mu} - \mu_w(T)\rangle^2$, we have

$$\begin{aligned}
\sup_u \langle \widehat{\mu} - \mu_w(T)\rangle^2 &= \sup_u \Big\langle \sum_{i\in T} \frac{w_i w_i^*}{\langle w, w^*\rangle}(X_i - \mu_w(T)), u\Big\rangle^2 \\
&\le \sup_u \sum_{i\in T} \frac{w_i w_i^*}{\langle w, w^*\rangle}\langle X_i - \mu_w(T), u\rangle^2 \\
&= \frac{\|w\|_1}{|G|\cdot\langle w, w^*\rangle} \sup_u \sum_{i\in T} \frac{w_i}{\|w\|_1}\langle X_i - \mu_w(T), u\rangle^2 \\
&= \frac{\|w\|_1}{\|w_G\|}\|\Sigma_w(T)\|_{\mathsf{op}} \le \frac{\sqrt{\|w\|_1}}{\alpha}\|\Sigma_w(T)\|_{\mathsf{op}}\,,
\end{aligned}$$

where in the second step we used Jensen's inequality, in the third step we used the definition of w^*, in the fourth step we used the definition of $\Sigma_w(T)$, and in the last step we used saturation. □

4 Safety maintained before termination

Finally, it remains to prove that prior to termination, there is an appropriate scoring rule which is safe and thus maintains the saturation invariant by Lemma 2.4.

One natural attempt at scoring points would be their distance to the weighted mean in the top-k eigenspace V_w of the current weighted covariance $\Sigma_w(T)$. We will abuse notation and let V_w also denote a $d \times k$ matrix whose columns are a basis for V_w. The motivation for this choice of score is its compatibility with the quantity $\lambda_k(\mathbf{\Sigma}_w(T))$ in the termination condition: the sum of squares of scores is

$$\sum_i \|V_w^\intercal (X_i - \mu_w(T))^2\| = V_w^\intercal \mathbf{\Sigma}_w(T) V_w\,,$$

i.e. the sum of squares of the top-k eigenvalues of the weighted covariance (the squared "Ky-fan k-norm"). Points for which this is large get progressively downweighted, so that eventually the top-k eigenvalues of the weighted covariance are sufficiently small that we hit the termination condition.

This is however not quite the right scoring rule, because the k large eigendirections need not be of equal magnitude, so points that deviate from $\mu_w(T)$ in certain directions within V_w get more severely penalized than points that deviate in other directions. To place all k directions on level playing field, we tweak the above score by "whitening" V_w. Let $\Sigma_w^{(k)} \triangleq V_w^\intercal \Sigma_w(T) V$ and define the scores for a particular set of weights w to be

$$\tau_i \triangleq \left\|(\Sigma_w^{(k)})^{-1/2} V_w^\intercal (X_i - \mu_w(T))\right\|^2 .$$

The effect of the $(\Sigma_w^{(k)})^{-1/2}$ term is simply to transform the projected data to be isotropic, i.e. so that the weighted k-dimensional covariance becomes $\mathbb{1}_k$

Our analysis is complete upon proving the following:

Lemma 4.1. *If $\lambda_k(\Sigma_w(T)) \geq 8/\sqrt{\|w\|_1}$ and $k \geq 8/\alpha$, then the scores $\{\tau_i\}$ are safe with respect to w.*

Proof. Because the projected data is isotropic after being transformed by $(\Sigma_w^{(k)})^{-1/2}$, the right-hand side of Eq. (2) in the definition of safety turns out to be exactly $k/2$:

$$\begin{aligned}\frac{1}{2\|w\|_1}\sum_{i\in T} w_i \tau_i &= \frac{1}{2}\Big\langle (\Sigma_w^{(k)})^{-1}, V_w^\intercal \Big(\frac{1}{\|w\|_1}\sum_{i\in T}(X_i - \mu_w(T))(X_i - \mu_w(T))^\intercal\Big) V_w \Big\rangle \\ &= \frac{1}{2}\langle (\Sigma_w^{(k)})^{-1}, \Sigma_w^{(k)}\rangle = k/2 .\end{aligned}$$

For the left-hand side of Eq. (2), we can first split up $X_i - \mu_w(T)$ into $X_i - \mu_w(G)$ and $\mu_w(G) - \mu_w(T)$ and thus decompose

$$\tau_i \leq 2\left\|(\Sigma_w^{(k)})^{-1/2} V_w^\intercal (X_i - \mu_w(G))\right\|^2 + 2\left\|(\Sigma_w^{(k)})^{-1/2} V_w^\intercal (\mu_w(G) - \mu_w(T))\right\|^2 . \tag{5}$$

The contribution of the former term for each i to $\frac{1}{\|w_G\|_1}\sum_{i\in G} w_i\tau_i$ is

$$\begin{aligned}&\frac{1}{\|w_G\|_1}\sum_{i\in G} w_i \cdot 2\left\|(\Sigma_w^{(k)})^{-1/2} V_w^\intercal (X_i - \mu_w(G))\right\|^2 \\ &\quad= \frac{2}{\|w_G\|_1}\sum_{i\in G} w_i \left\|(\Sigma_w^{(k)})^{-1/2} V_w^\intercal (X_i - \mu_w(G))\right\|^2 \\ &\quad= 2\langle (\Sigma_w^{(k)})^{-1}, V_w^\intercal \Sigma_w(G) V_w\rangle \\ &\quad\leq \frac{2\alpha}{\|w_G\|_1}\cdot \mathrm{Tr}((\Sigma_w^{(k)})^{-1}) \\ &\quad\leq \frac{2k\alpha\sqrt{\|w\|_1}}{8\|w_G\|_1} \leq k/4 ,\end{aligned}$$

where in the first inequality we used Lemma 4.2 below and in the last step we used the fact that $\lambda_k(\Sigma_w(T)) \geq 8/\sqrt{\|w\|_1}$.

The contribution of the latter term in Eq. (5) to $\frac{1}{\|w_G\|_1}\sum_{i\in G} w_i\tau_i$ is

$$\begin{aligned}
&\frac{1}{\|w_G\|_1}\sum_{i\in G} w_i \cdot 2\left\|(\Sigma_w^{(k)})^{-1/2}V_w^\intercal(\mu_w(G)-\mu_w(T))\right\|^2 \\
&\quad= \frac{2}{\|w_G\|_1}\sum_{i\in G} w_i\left\|(\Sigma_w^{(k)})^{-1/2}V_w^\intercal(X_i-\mu_w(G))\right\|^2 \\
&\quad= 2\langle(\Sigma_w^{(k)})^{-1}, V_w^\intercal(\mu_w(G)-\mu_w(T))(\mu_w(G)-\mu_w(T))^\intercal V_w\rangle \\
&\quad\le \frac{2\|w\|_1}{\|w_G\|_1} \\
&\quad\le \frac{2\sqrt{\|w\|_1}}{\alpha} \le \frac{2}{\alpha},
\end{aligned}$$

where in the first inequality we used Lemma 4.3 below.

We conclude that

$$\frac{1}{\|w_G\|_1}\sum_{i\in G} w_i\tau_i \le k/4 + \frac{2}{\alpha} \le k/2$$

as desired, where in the last step we used the assumption that $k \ge 8/\alpha$. □

Lemma 4.2. *If $w \in \Delta^T$ satisfies $w_i \le 1/n$ for all i, then $\mathbf{\Sigma}_w(G) \preceq \frac{\alpha}{\|w_G\|_1}\mathbb{1}$.*

Proof. Let $w^* \in \Delta^T$ be given by $w_i^* \triangleq 1/|G| \cdot \mathbb{1}[i \in G]$. Note that by assumption, $w_i \le \alpha w_i^*$. Furthermore,

$$\sum_{i\in G} w_i(X_i-\mu_w(G))(X_i-\mu_w(G))^\intercal \preceq \sum_{i\in G} w_i(X_i-\mu')(X_i-\mu')^\intercal$$

for any vector $\mu' \in \mathbb{R}^d$ (i.e. to minimize the variance of a distribution, one should center it around its mean and not around any other point μ').

We thus have

$$\mathbf{\Sigma}_w(G) \preceq \frac{1}{\|w_G\|_1}\sum_{i\in G} w_i(X_i-\mu^*)(X_i-\mu^*) \preceq \frac{\alpha}{\|w_G\|_1}\sum_{i\in G} w_i^*(X_i-\mu^*)(X_i-\mu^*) \preceq \frac{\alpha}{\|w_G\|_1}\mathbb{1}\,,$$

where in the last step we used Eq. (1). □

Lemma 4.3. *For any $w \in \Delta^T$, we have*

$$(\mu_w(G)-\mu_w(T))(\mu_w(G)-\mu_w(T))^\intercal \preceq \frac{\|w\|_1}{\|w_G\|_1}\Sigma_w(T)\,.$$

Proof. For any distribution over vector x, we have $\mathbb{E}[x]\mathbb{E}[x]^\intercal \preceq \mathbb{E}[xx^\intercal]$. By applying this to the distribution which samples x by selecting $i \in G$ with probability $w_i/\|w\|_1$ and outputting $x = X_i - \mu_w(T)$, we conclude that

$$(\mu_w(G)-\mu_w(T))(\mu_w(G)-\mu_w(T))^\intercal \preceq \frac{1}{\|w_G\|_1}\sum_{i\in G} w_i(X_i-\mu_w(T))(X_i-\mu_w(T))^\intercal\,.$$

We can bound the right-hand side by a sum over T and rewrite $1/\|w_G\|_1$ by $\frac{\|w\|_1}{\|w_G\|_1} \cdot \frac{1}{\|w\|_1}$ to get

$$\begin{aligned}
&\preceq \frac{\|w\|_1}{\|w_G\|_1} \frac{1}{\|w\|_1} \sum_{i \in T} w_i (X_i - \mu_w(T))(X_i - \mu_w(T))^\intercal \\
&= \frac{\|w\|_1}{\|w_G\|_1} \Sigma_w(T)
\end{aligned}$$

as claimed. □

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.

[Rud99] Mark Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.