Lecture 10: Efficient algorithms for list-decodable mean estimation

October 27, 2025

1 Introduction and setup

In this lecture, we will sketch out the broad strokes of how to obtain an efficient algorithm for list-decodable mean estimation that achieves the optimal statistical guarantees (up to constant factors) in the bounded covariance setting. For the rest of the lecture, assume that S_{good} is a set of αn points in $\mathbb{R}^d$ with mean μ and which satisfy that

$$\Sigma := \frac{1}{|S_{\mathrm{good}}|} \sum_{i \in S_{\mathrm{good}}} (X_i - \mu)(X_i - \mu)^\top \preceq I \; . \tag{1}$$

We've already established (or more precisely, the homework asks you to show) that if S_{good} is a set of sufficiently many samples from a distribution with bounded covariance, then a large subset of it will satisfy these conditions (perhaps up to constant factors), and will also satisfy that the distance between μ and the true mean of the distribution is small. So, learning μ to small error is more or less equivalent to learning the true mean of the distribution.

The algorithm will be given a set of n samples S so that $S_{\mathrm{good}} \subset S$. Recall our goal here is to, given S, output a small list of $L = O(1/\alpha)$ candidate means $\mu_1, \ldots, \mu_L$ so that $\|\mu_i - \mu\|_2 \leq O(1/\sqrt{\alpha})$ for some $i \in [L]$. This lecture will primarily follow and be a simplified exposition of the algorithm in [1].

From list-decodable mean estimation to robust PCA Our first observation is that the problem is easy, if instead of asking for error $O(1/\sqrt{\alpha})$, we asked for a weaker notion of error:

Lemma 1.1. *There is a randomized polynomial time algorithm which outputs a list of $L' = O(1/\alpha)$ candidate means $\mu_1, \ldots, \mu_{L'}$ so that with high probability, $\|\mu_i - \mu\|_2 \leq O(\sqrt{d})$ for some $i \in 1, \ldots, L'$.*

Proof. The algorithm is very simple. We observe that by taking traces of both sides of (1), we obtain that

$$\frac{1}{|S_{\mathrm{good}}|} \sum_{i \in S_{\mathrm{good}}} \|X_i - \mu\|_2^2 \leq d \; .$$

Thus, by Markov's inequality, a random point in S_{good} will have $\|X_i - \mu\|_2^2 \leq 100d$ with probability at least 0.99. In other words, to obtain ℓ_2 error $O(\sqrt{d})$ from μ, it suffices to output a list which contains an random element of S_{good}. To do this, we can just output a list of $O(1/\alpha)$ random points from the dataset S. In particular, if we output a list of size, say $100/\alpha$, then the probability we do not output any points from S_{good} is at most 0.01. So by a union bound, this list will contain a point with $\|X_i - \mu\|_2^2 \leq 100d$ with probability at least 0.98. □

A somewhat unsatisfying aspect of this bound is that the method is both randomized, and moreover, needs to output a very large list if we want to ensure that the algorithm succeeds with very high probability. However, it turns out that by initially taking a large list, and combining this with some simple postprocessing steps, one can guarantee that the final list can simultaneously be small, say of size $\leq 2/\alpha$, and the list will have error at most $2d$ with probability $1 - \delta$. We leave this postprocessing as an exercise for the reader.

At first glance, this guarantee may not seem very interesting: it says that if d is small compared to $1/\alpha$, we can obtain good error guarantees. But why is this useful when d is large? The key insight is that this says that to solve list-decodable mean estimation, it suffices to find a subspace V of dimension $O(1/\alpha)$ so that V approximately contains μ. Once we've found such a subspace, we can then project the data into this subspace, and call this routine to solve the problem to the desired accuracy. In this way, we've reduced the problem to a problem of *robust PCA*, which turns out to be algorithmically more tractable to think about.

2 Filtering for robust PCA

So now our goal is to efficiently find a low-dimensional subspace which approximately contains μ. A natural candidate subspace would simply be the subspace spanned by the k largest eigenvectors of the empirical covariance $\widehat{\Sigma}$ of the data, for some appropriate parameter k. Let us assume without loss of generality that the empirical mean of the dataset is 0, which we can always achieve by shifting the data, so that

$$\widehat{\Sigma} = \frac{1}{|S|} \sum_{i \in S} X_i X_i^\top \; .$$

It is not too hard to see that if μ is far from 0, then the good points must induce a reasonably large eigenvector. This is because

$$\begin{aligned} \widehat{\Sigma} &\succeq \frac{1}{|S|} \sum_{i \in S_{\text{good}}} X_i X_i^\top \\ &= \alpha \frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}}} X_i X_i^\top \\ &= \alpha \left(\sum_{i \in S_{\text{good}}} (X_i - \mu)(X_i - \mu)^\top + \mu\mu^\top \right) \; , \end{aligned}$$

and hence if we let v be the unit vector in the direction of μ, we have that

$$v^\top \widehat{\Sigma} v \geq \alpha \|\mu\|^2 - O(\alpha) \; ,$$

by (1). In particular, if $\|\mu\|_2 \gg O(1/\sqrt{\alpha})$, then this induces a direction with energy at least $\Omega(1)$. The contrapositive of this statement is as follows: if we can find a subspace V so that for all u orthogonal to V, we have that $u^\top \Sigma u \leq 1$, then this implies that μ must be largely contained in V. Formally, if we let $\Pi_V : \mathbb{R}^d \to \mathbb{R}^d$ denote the projection onto V, this implies that $\|\Pi_{V^\perp} \mu\|_2 \lesssim 1/\sqrt{\alpha}$. Combining this with Lemma 1.1 suggests the following general recipe for list-decodable mean estimation:

- Remove outliers until we can find a subspace V of dimension $k = O(1/\alpha)$ so that $\Pi_{V^\perp} \widehat{\Sigma} \Pi_{V^\perp}$ has maximum eigenvalue $O(1)$.
- Use Lemma 1.1 to learn μ to error $O(1/\sqrt{\alpha})$ in V, and output this list.

If we can do this, then Lemma 1.1 implies that $\|\Pi_V \mu - \mu_i\|_2 \leq O(1/\sqrt{\alpha})$ for some $i = 1, \ldots, L$, and combining this with the bound that $\|\Pi_{V^\perp} \mu\|_2 \leq O(1/\sqrt{\alpha})$, we obtain that $\|\mu - \mu_i\|_2 \leq O(1/\sqrt{\alpha})$, which is what we wanted.

The key algorithmic routine is the first bullet point. Now, such a subspace exists if and only if $\widehat{\Sigma}$ has at most k eigenvalues above $O(1)$. So, the main work is to somehow remove the influence of "bad" points if $\widehat{\Sigma}$ has more than k large eigenvalues. Intuitively, why should this be possible? Recall that the hard instances we've seen so far all have the form that there are $1/\alpha$ disjoint collections of possible solutions, each with mean μ_i, for $i = 1, \ldots, 1/\alpha$. But note that even for such instances, there can only be at most $O(1/\alpha)$ large eigenvalues, since if we let $V = \text{span}(\mu_1, \ldots, \mu_{1/\alpha})$, then outside of this subspace, all of the eigenvalues should be at most constant. So if there are more than $O(1/\alpha)$ large eigenvalues, this should intuitively be because of a collection of bad points, which we can hope to filter out. But we have to be careful: after all, the guarantees for the filter we've previously demonstrated only guarantee that we remove more bad points than good points—but this is clearly insufficient for our setting here, where the majority of the points will be bad.

A filter in the isotropic setting Let us first see how we can do this in a slightly simplified setting. Assume that $\widehat{\Sigma}$ has k eigenvalues which exceed C for some constant C sufficiently large, but let's also assume (and this is less justified) that these k eigenvalues of $\widehat{\Sigma}$ are all *exactly equal* to C. Let us also continue to assume that the empirical mean of S is 0. Let V be the span of the first k eigenvectors (if there's a tie, choose arbitrarily), and let Π be the projection onto the subspace spanned by these eigenvectors. The key idea will be to define the following scores:

$$\tau_i = \|\Pi X_i\|_2^2 \; , \tag{2}$$

for all $i \in S$. These scores have two key properties. First, we have that

$$\begin{aligned}\frac{1}{|S|}\sum_{i\in S}\tau_i &= \frac{1}{|S|}\sum_{i\in S}\|\Pi X_i\|_2^2 \\ &= \frac{1}{|S|}\sum_{i\in S}\operatorname{tr}\left(\Pi X_i X_i^\top \Pi\right) \\ &= \operatorname{tr}\left(\Pi\left(\frac{1}{|S|}\sum_{i\in S}X_iX_i^\top\right)\Pi\right) \\ &= \operatorname{tr}\left(\Pi\widehat{\Sigma}\Pi\right) = Ck \; .\end{aligned}$$

On the other hand, by a similar calculation, we have that

$$\begin{aligned}\frac{1}{|S_{\text{good}}|}\sum_{i\in S_{\text{good}}}\tau_i &= \operatorname{tr}\left(\Pi\left(\frac{1}{|S_{\text{good}}|}\sum_{i\in S_{\text{good}}}X_iX_i^\top\right)\Pi\right) \\ &= \operatorname{tr}\left(\Pi\left(\Sigma + \mu\mu^\top\right)\Pi\right) \\ &\le k + \|\Pi\mu\|_2^2 \\ &\le k + \|\mu\|_2^2 \; .\end{aligned}$$

Now, if $\|\mu\|_2^2 > C/\alpha$, by the same calculation as before, this would imply that the top eigenvalue of $\widehat{\Sigma}$ would exceed C, which is impossible by asssumption. Hence if we take k to be a sufficiently large multiple of $1/\alpha$, we have that $\|\mu\|_2^2 \le C/\alpha \le \frac{Ck}{10}$. So what we've shown is that

$$\frac{1}{|S_{\text{good}}|}\sum_{i\in S_{\text{good}}}\tau_i \ll \frac{1}{2}\cdot\frac{1}{|S|}\sum_{i\in S}\tau_i \; , \tag{3}$$

that is, the average value of the scores on the good points is much smaller than the average score of the overall set of points. Note that this is a significantly stronger condition than what we had in the small ε regime, and it is this stronger guarantee that allows us to ensure that we don't ever remove too many good points.

Let's see how this plays out with an unweighted filter, i.e. the randomized filter where we keep each point X_i from the dataset with probability $1 - \tau_i/\tau_{\max}$, where $\tau_{\max} = \max_{i\in S}\tau_i$. In this case, the expected total number of points we removed is exactly

$$\Delta S := \sum_{i\in S}\Pr[X_i \text{ was removed}] = \frac{1}{\tau_{\max}}\sum_{i\in S}\tau_i \; ,$$

and by the above, the expected number of points in S_{good} we removed is at most

$$\frac{1}{\tau_{\max}} \sum_{i \in S_{\text{good}}} \tau_i \leq \frac{\alpha}{2} \cdot \Delta S \; .$$

If we pretend that the actual number of points removed was exactly the expectation, then now we have a set of $n - \Delta S$ points, of which an $\alpha n - \frac{\alpha}{2}\Delta S$ belong to S_{good}. The key point is that

$$\frac{\alpha n - \alpha \Delta S/2}{n - \Delta S} > \alpha \; ,$$

so in particular, the fraction of good points remaining in the dataset has actually increased, and so the filtering step has actually strictly improved (in some sense) the quality of the data!

But this is not quite the end of the story, unfortunately. The problem is that because we've removed some good points, we can no longer guarantee that the covariance of the remaining good points has spectral norm at most 1. So not only do we need that the fraction of points which are good remains the same, we actually need that it *strictly increases* from this process. A clean way of expressing this will be through the notion of a *saturated* set. Formally, we say that a subset S' is *saturated* if

$$\frac{|S' \cap S_{\text{good}}|}{|S_{\text{good}}|} \geq \left(\frac{|S'|}{|S|} \right)^{1/2} \; .$$

Note that after one iteration, once again assuming that the number of points removed is exactly the expected number, if S' is the set of points remaining, and using the identity that $1 - \delta/2 \geq \sqrt{1 - \delta}$ for all $\delta \in (0, 1)$, we have that

$$\begin{aligned} \frac{|S_{\text{good}} \cap S'|}{|S_{\text{good}}|} &\geq \frac{\alpha n - \alpha \Delta S/2}{\alpha n} \\ &= 1 - \frac{1}{2} \frac{\Delta S}{n} \geq \sqrt{1 - \frac{\Delta S}{n}} = \left(\frac{|S'|}{|S|} \right)^{1/2} \; , \end{aligned}$$

so after one iteration, the set of points remains saturated. By inductively applying this argument, one can show that as long as the scores we filter with always satisfy the safety condition (3), then the final set of points will be saturated.[1]

Let us see why saturation intuitively should suffice for solving the problem. First, observe that saturated sets cannot be too small: indeed, if $|S'| < \alpha^2 n$, then

$$\frac{|S' \cap S_{\text{good}}|}{|S_{\text{good}}|} \leq \frac{|S'|}{|S_{\text{good}}|} < \left(\frac{|S'|}{|S|} \right)^{1/2} \; .$$

Now, suppose we have a S' which is a saturated set of points of size βn, where $\beta \geq \alpha^2$ by the above. Then, the number of good points remaining must be at least $\beta^{1/2} \alpha n$. There are several balancing considerations.

- First of all, as we argued before, any set of points with bounded covariance is automatically $(\sqrt{\alpha}, \alpha)$-resilient. Since the number of remaining good points is at least $\alpha^2 n$, this implies that the empirical mean of these good points is $O(1/\sqrt{\alpha})$ close to the original mean. So up to constant factors, it suffices to recover the mean of this remaining set of good points.
- Second, the empirical covariance of the good points has spectral norm at most $\beta^{-1/2}$, since

$$\frac{1}{|S' \cap S_{\text{good}}|} \sum_{i \in S' \cap S_{\text{good}}} (X_i - \mu)(X_i - \mu)^\top \preceq \beta^{-1/2} \frac{1}{|S_{\text{good}}|} \sum_{i \in S_{\text{good}}} (X_i - \mu)(X_i - \mu)^\top \; .$$

[1] Again, this technically only holds if the number of points we remove is exactly the expected number of points we expect to remove. Arguably, the cleanest way of getting around this is to once again use downweighting rather than point removal—see [1] for a full treatment with downweighting.

- Finally, the remaining good points form an $\alpha/\sqrt{\beta}$-fraction of the overall dataset.

But by scale invariance of this problem, in general, if you have an α'-fraction of good points, where the good points have covariance with spectral norm C, one should expect to be able to recover the mean of the good points to error $\sqrt{C}/\sqrt{\alpha'}$, so in our case, we can recover the mean of the good points to error

$$\beta^{-1/2} \cdot \frac{\beta^{1/2}}{\alpha^{1/2}} = \frac{1}{\alpha^{1/2}} \; ,$$

which is exactly what we wanted!

In other words, in the next iteration, what we should do is treat the remaining set of points as a set of points which contains a $\alpha/\beta^{1/2}$-fraction subset of points whose covariance has spectral norm at most $\beta^{1/2}$, and we should repeatedly run the filter, but with these updating parameters. This will always guarantee that, before termination, the set of points is saturated. The filter always removes at least one point at a time, and we argued that any saturated set has size at least $\alpha^2 n$, so this filter must terminate, and when it does, it must output a subspace V with the desired properties.

Addressing the isotropic assumption There is one final thing we need to deal with, which is that we assumed that all of the k largest eigenvalues of the empirical covariance were exactly C. But in general, we're allowed to assume that all of these eigenvalues are at least C. However, this is an easy enough fix—we simply scale down these k directions, so that the empirical covariance looks like it has eigenvalue C. Formally, we can define the transformation

$$X_i \mapsto C\Sigma^{-1/2}\Pi_V X_i + \Pi_{V^\perp} X_i \; ,$$

which has exactly this effect. Note that the resulting scores of the transformed points have a nice form: they are exactly

$$\tau_i = C\|\Sigma^{-1/2}\Pi_V X_i\|_2^2 \; .$$

Since this is only scaling down the points, one can check that this can only decrease the scores on the good points, and so this will still guarantee the safety condition.

Putting it all together With this, we can now finally state the pseudocode for the full algorithm. We note that this is still a simplification of the overall algorithm, since we need to deal with the fact that the filtering either needs to be randomized or incorporate weights. See [1] for the full description of the algorithm.

References

[1] 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.

Algorithm 1 List-Filter

```
procedure Filter($S$)
    Let $S_0 = S$.
    Let $C \geq 11$ be a universal constant.
    Let $\Sigma^{(0)} = \Sigma(S)$.
    Let $k = 10/\alpha$.
    Let $t = 0$.
    while the $(k+1)$-st eigenvalue of $\Sigma^{(t)}$ exceeds $C$: do
        Let $V$ be the span of the top $k$ eigenvalues of $\Sigma^{(t)}$
        Let
            $\tau_i^{(t)} = C\|(\Sigma^{(t)})^{-1/2} V X_i\|_2^2$
        Let $S_{t+1}$ be the result of filtering using these scores, and let $\Sigma^{(t+1)}$ be the empirical covariance
        Let $t \leftarrow t + 1$.
    end while
    return a set of $O(k)$ random points from $S^{(t)}$.
end procedure
```