Lecture 9: Introduction to list-decodable learning

October 27, 2025

1 Introduction

So far in this class, we have heavily leveraged the assumption that the fraction of corrupted data is at most $1/2$, and usually, we assume that it is at most some relatively small constant (say 0.1 or so). This is, in many ways reasonable: when the fraction of corruptions exceeds $1/2$, then the problem of recovering the "uncorrupted" data is clearly ill-defined. Consider, for instance, an instance where, for some $\mu \in \mathbb{R}$, half of the data is drawn from $\mathcal{N}(\mu, I)$, and half of the data is drawn from $\mathcal{N}(-\mu, I)$. Then, clearly, this is a $1/2$-corrupted set of samples from $\mathcal{N}(\mu, I)$, but also a $1/2$-corrupted set of samples from $\mathcal{N}(-\mu, I)$, so as we take $\mu \to \infty$, no estimator can ever achieve any non-trivial error.

In the above example, the problem was that there were two equally valid, but very different, solutions, because of the corruption. But perhaps this is the only thing that an adversary can do? In this case, if we slightly relax the learning condition, and allow ourselves to output a list of 2 candidate solutions (namely, μ and $-\mu$), we can guarantee that this list contains all possible valid solutions. This motivates the following definition of list learning:

Definition 1.1 (List-decodable mean estimation, [1, 2])**.** Let $\mathcal{D}$ be a class of distributions over $\mathbb{R}^d$, and let $D \in \mathcal{D}$ with mean μ, and let $\alpha \in (0, 1)$. Given a set S of n samples, with the assumption that there exists a set $S_{\text{good}} \subset S$ of size $|S_{\text{good}}| = \alpha n$ so that S_{good} consists of αn independent draws from D, output a list $\{\mu_1, \ldots, \mu_L\}$ of candidate means minimizing

$$\min_{i \in [L]} \|\mu_i - \mu\|_2 \, .$$

We will refer to this quantity as the *error* of the list learning algorithm.

As before, we will largely focus on two canonical choices of $\mathcal{D}$, namely, the set of Gaussians with identity covariance, and the set of distributions with bounded covariance, although many other classes of distributions have been considered in the literature (and we will consider other assumptions in later classes as well).

Beyond the qualitative difference that we're typically assuming α (which is really just $1 - \varepsilon$ in the usual notions of contamination) is very small, one concrete difference between this setting and the ε-corruption setting we've been considering so far is that we're assuming an additive adversary. There are notions which also incorporate some notion of subtractive error, but it turns out that for any reasonable notion of subtractive error, there is no meaningful distinction between the types of guarantees we can achieve with and without subtractive error, so we will largely ignore this distinction.

A simple generalization of the toy argument given above also yields a simple lower bound on the list size L that is achievable:

Lemma 1.1. *Let $\mathcal{D} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$. Any algorithm that outputs a list of less than $\lfloor 1/\alpha \rfloor$ candidates cannot achieve any non-trivial error for this class of distributions.*

Proof. Let $\alpha = 1/k$, and for some parameters μ, consider the instance where a $1/k$-fraction of the samples are drawn from $\mathcal{N}(i \cdot \mu, 1)$, for $i = 1, \ldots, k$. Then, clearly, $\mu, 2\mu, \ldots, k\mu$ are all valid solutions, and so by letting $\mu \to \infty$, we see that no list of less than k elements can achieve any finite error. □

2 Information-theoretic upper bounds via resilience

We now show that something non-trivial is possible, as long as we output a sufficiently large list of candidate solutions. In fact, information-theoretically, we can do this by using the resilience criterion we established much earlier on. We recall:

Definition 2.1 (Resilience, see [3])**.** Let D be a distribution over $\mathbb{R}^d$ with mean μ. Let $\sigma, \varepsilon > 0$. We say that D is (σ, ε)-*resilient* if for all events E so that $\Pr_D[E] \geq 1 - \varepsilon$, then

$$\left\| \mathop{\mathbb{E}}_{D}[X|E] - \mu \right\|_2 \leq \sigma \ . \tag{1}$$

Given a dataset S of size n, we say that S is (σ, ε)-resilient if the uniform distribution over S is (σ, ε)-resilient.

We will primarily be using the empirical notion of resilience here. An important, but straightforward, consequence of resilience for us will be that if we have a set S of n points with mean μ that is (σ, ε)-resilient, then this implies that for all $T \subset S$ with $|T| = \varepsilon n$, then it must be that

$$\left\| \frac{1}{|T|} \sum_{i \in T} X_i - \mu \right\|_2 \leq \frac{(1-\varepsilon)\sigma}{\varepsilon} \ , \tag{2}$$

and moreover, the fact that this occurs for all $T \subset S$ of size $|T| = \varepsilon n$ is equivalent to resilience. One can also easily show the following fact:

Fact 2.1. *Let S be a (σ, ε)-resilient set of points for $\varepsilon \leq 1/2$. Then for all $\varepsilon' \geq \varepsilon$, S is also a (σ, ε')-resilient set of points.*

With this, we can now show:

Theorem 2.2. *Let S_{good} be a subset of αn points that is $(\sigma, \alpha/4)$-resilient with mean μ. Then, there is an (inefficient) estimator which takes in a set S of n points so that $S_{\text{good}} \subset S$, and outputs a list of $L \leq 2/\alpha$ means $\mu_1, \ldots, \mu_L$ so that $\|\mu_i - \mu\|_2 \leq O(\sigma/\alpha)$ for some $i = 1, \ldots, L$.*

Proof. The algorithm will be as follows. Let $S_1, \ldots, S_L$ be a maximal collection of disjoint subsets of S satisfying:

- $|S_i| = \alpha n/2$, for all $i = 1, \ldots, L$, and
- S_i is $(\sigma, \alpha/2)$-resilient, for all $i = 1, \ldots, L$.

We claim that if we let μ_i be the mean of S_i, then the list $\mu_1, \ldots, \mu_L$ must satisfy the desired condition. Clearly $L \leq 2/\alpha$, which together with this claim, yields the desired conclusion.

To do this, we first observe that S_{good} must have large intersection with at least one of the S_i. Indeed, by maximality, it follows that $S' = S_{\text{good}} \setminus (S_1 \cup \ldots \cup S_L)$ cannot be added to the collection of sets. Now, if S' had size larger than $\alpha n/2$, then by (2) this implies that S' itself would be a (σ, α)-resilient subset of points of size at least $\alpha n/2$, which would contradict the maximality of $S_1, \ldots, S_L$.[1]

Therefore $|S'| \leq \alpha n/2$. By the pigeonhole principle, this implies that $|S_{\text{good}} \cap S_i| \geq \frac{\alpha^2}{4} n = \frac{\alpha}{2}|S_i|$ for some i. Let $T = S_{\text{good}} \cap S_i$. By the resilience of S_{good}, we know that $\|\mu(T) - \mu\| \leq \sigma/\alpha$, and similarly, by the resilience of S_i, we also have that $\|\mu(T) - \mu(S_i)\|_2 \leq \sigma/\alpha$. Combining this with a triangle inequality completes the proof. □

By using the machinery we've already developed in this class, it is not hard to show the following instantiations of this resilience bound:

[1] There is a small subtlety to this argument—do you see it?

Lemma 2.3. *Let S_{good} be a sufficiently large set of points from a distribution D with mean μ and covariance $\Sigma \preceq I$. Then, with high probability, S_{good} is a $(\sqrt{\varepsilon}, \varepsilon)$-resilient set of points satisfying $\|\mu(S_{\text{good}}) - \mu\|_2 \leq O(\sqrt{\varepsilon})$.*

Lemma 2.4. *Let S_{good} be a sufficiently large set of points from $\mathcal{N}(\mu, I)$. Then, with high probability, S_{good} is a $(\varepsilon\sqrt{\log 1/\varepsilon}, \varepsilon)$-resilient set of points satisfying $\|\mu(S_{\text{good}}) - \mu\|_2 \leq O(\varepsilon\sqrt{\log 1/\varepsilon})$.*

Combining these lemmas with our theorem, we obtain the following corollaries:

Corollary 2.5. *Let S_{good} be a set of αn points from a distribution D with mean μ and covariance $\Sigma \preceq I$ for n sufficiently large, and let $S \supset S_{\text{good}}$ be of size n. Then, with high probability, there is a list learning algorithm that, given S, outputs a list of $O(1/\alpha)$ means which achieves error $O(1/\sqrt{\alpha})$.*

Corollary 2.6. *Let S_{good} be a set of αn points from $\mathcal{N}(\mu, I)$ for n sufficiently large, and let $S \supset S_{\text{good}}$ be of size n. Then, with high probability, there is a list learning algorithm that, given S, outputs a list of $O(1/\alpha)$ means which achieves error $O(\sqrt{\log 1/\alpha})$.*

Note that these error bounds are, in some sense, quite bad! Namely, while they are dimension-free, their error goes to ∞ as $\alpha \to 0$, as opposed to in the previous lectures, where we achieved vanishing error (i.e. $o(1)$) error. However, it turns out that in this setting, this is the best type of error one can hope to achieve, as we will now see.

3 Lower bounds for small lists

Recall that ε-additive corruption is the natural "adaptive" analog of the "oblivious" setting, where we assume we are given samples from some distribution D' so that $D' = (1-\varepsilon)D + \varepsilon P$ for some arbitrary distribution P, which in our setting is equivalent to saying that $D' = \alpha D + (1-\alpha)P$ for some distribution P. Formally, this means that if we generate a set of n samples from D' for n sufficiently large, then with high probability, there exists a subset of $(1-o(1))\alpha n$ of these samples so that these samples are independent draws from D. A helpful, equivalent way of phrasing this is that $\alpha D(x) \leq D'(x)$ for all x; that is, D could be the good distribution so long as its pdf is pointwise upper bounded by $1/\alpha$ times the pdf of $D'(x)$. With this, we can also prove the following lower bound on the error achievable by any algorithm that outputs a $\text{poly}(1/\alpha)$-sized list of candidate solutions:

Lemma 3.1. *Let $\mathcal{D} = \{\mathcal{N}(\mu, I) : \mu \in \mathbb{R}^d\}$, for d sufficiently large. Then, any algorithm which achieves error $o(\sqrt{\log 1/\alpha})$ must output a list of size $\min(2^{cd}, (1/\alpha)^{\omega(1)})$.*

Proof sketch. By the logic above, it suffices the demonstrate a distribution D', and a collection of N means $\mu_1, \ldots, \mu_N$, where $N = \min(2^{cd}, (1/\alpha)^{\omega(1)})$ so that:

- $\|\mu_i - \mu_j\|_2 \geq c\sqrt{\log 1/\alpha}$ for all $i \neq j$, and
- $D'(x) \geq \alpha \mathcal{N}(\mu_i, I)(x)$, for all $i = 1, \ldots, N$.

To construct our set of μ_i, we use the following fact, which we will not prove:

Fact 3.2. *For $c > 0$ sufficiently small, there exists a collection of $L \geq 2^{cd}$ vectors $\mu_1, \ldots, \mu_L$ so that $\|\mu_i\| = c\sqrt{\log 1/\alpha}$ and $\|\mu_i - \mu_j\|_2 \geq \frac{c}{100}\sqrt{\log 1/\alpha}$ for all $i \neq j$.*

We will construct our distribution in steps. First, let $D_1(x) = \frac{1}{2}\mathcal{N}(0, I)(x)$. Note that this is not yet a distribution, but the key observation is that this already is already quite an effective way of "covering" the pdf of $\mathcal{N}(\mu_i, I)$. Indeed, we observe that $D_1(x) \leq \alpha \mathcal{N}(\mu_i, I)(x)$ if and only if

$$\frac{1}{2}\exp\left(-\frac{1}{2}\|x\|^2\right) \leq \alpha \cdot \exp\left(-\frac{1}{2}\|x - \mu_i\|^2\right),$$

which simplifies to the condition that

$$\langle x, \mu_i \rangle \geq \left(1 - \frac{c^2}{2}\right) \log 1/\alpha - \log 2 \,,$$

and call S_i the set of points which satisfies this property. Now, let $P_i(x) = (\alpha \mathcal{N}(\mu_i, I)(x) - D_1(x)) \cdot \mathbf{1}_{x \in S_i}$, so that $P_i(x) \geq 0$, and moreover,

$$\begin{aligned}
p_i := \int_{x \in \mathbb{R}^d} P_i(x) dx &= \int_{x \in S_i} \alpha \mathcal{N}(\mu_i, I)(x) - D_1(x) dx \\
&\leq \int_{x \in S_i} \alpha \mathcal{N}(\mu_i, I)(x) dx \\
&= \Pr_{X \sim \mathcal{N}(\mu_i, I)} \left[\langle X, \mu_i \rangle \geq \left(1 - \frac{c^2}{2}\right) \log 1/\alpha - \log 2 \right] \\
&= \Pr_{X \sim \mathcal{N}(0, I)} \left[\langle X, \mu_i \rangle \geq \left(1 - \frac{3c^2}{2}\right) \log 1/\alpha - \log 2 \right] \,.
\end{aligned}$$

But since $\langle X, \mu_i \rangle \sim \mathcal{N}(0, c^2 \log 1/\alpha)$, we have that $p_i \leq \alpha^{\Omega(c^{-2})}$. Note that by symmetry, $p_i = p_j$ for all $i \neq j$, so let this shared value be denoted p.

Now, let $N = \frac{1}{2} \min(2^{cd}, 1/p)$, and consider the distribution given by $D = D_1 + \sum_{i=1}^{L} P_i(x)$. By assumption (ignoring issues of rounding), we know that D is a valid distribution, and by construction, $D(x) \geq \alpha \mathcal{N}(\mu_i, I)(x)$ for all $i = 1, \ldots, L$, which yields the desired conclusion. □

This establishes that in the Gaussian setting, we cannot really hope to obtain error better than $O(\sqrt{\log 1/\alpha})$ with any polynomial-sized list. Similarly, one can show that one cannot achieve error better than $O(1/\sqrt{\alpha})$ in the bounded second moment setting with a polynomial-sized list.

References

[1] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):1–27, 2017.

[2] 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. ACM, 2017.

[3] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.