[Lecture 24]

0.1 Clustering Under Resilience

The mixture of Gaussians case is unsatisfying because data are unlikely to actually be Gaussian mixtures in practice, yet common algorithms like k-means still do a good job at clustering data. We therefore move to the agnostic setting, and show that we only need the distributions to be *resilient* in order to cluster successfully.

We will start by proving an even stronger result—that if a set of points contains a (ρ, α)-resilient subset S of size αn, then it is possible to output an estimate $\hat{\mu}$ that is close to the true mean μ of S, regardless of the other $(1-\alpha)n$ points. As stated, this is impossible, since there could be $\mathcal{O}(1/\alpha)$ identical clusters in the data. So what we will actually show is a *list-decoding* result—that it is possible to output $\mathcal{O}(1/\alpha)$ "candidates" $\hat{\mu}_l$ such that one of them is close to the mean of S:

Proposition 0.1. *Suppose that a set of points $\tilde{S} = \{x_1, \dots, x_n\}$ contains a $(\rho, \alpha/4)$-resilient set S with mean μ. Then if $|S| \geq \alpha n$ (even if $\alpha < \frac{1}{2}$), it is possible to output $m \leq \frac{2}{\alpha}$ candidates $\hat{\mu}_1, \dots, \hat{\mu}_m$ such that $\|\hat{\mu}_j - \mu\| \leq \frac{8\rho}{\alpha}$ for some j.*

Proof. The basic intuition is that we can cover the points in $\tilde{S}$ by resilient sets $S'_1, \dots, S'_{2/\alpha}$ of size $\frac{\alpha}{2}n$. Then by the pigeonhole principle, the resilient set S must have large overlap with at least one of the S', and hence have similar mean. This is captured in Figure 1 below.

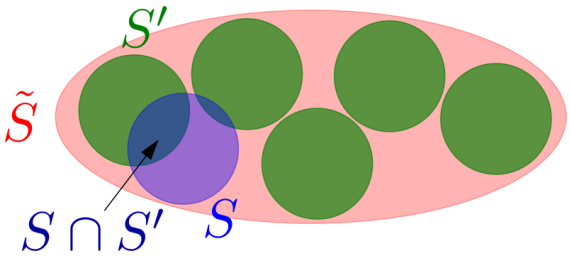

Figure 1: If we cover $\tilde{S}$ by resilient sets, at least one of the sets S' has large intersection with S.

The main difference is that S and S' may have relatively small overlap (in a roughly α-fraction of elements). We thus need to care about resilience when the subset T is small compared to S. The following lemma relates resilience on large sets to resilience on small sets:

Lemma 0.2. *For any $0 < \epsilon < 1$, a distribution/set is (ρ, ϵ)-resilient if and only if it is $(\frac{1-\epsilon}{\epsilon}\rho, 1-\epsilon)$-resilient.*

This was already proved in Appendix **??** as part of Lemma **??**. Given Lemma 0.2, we can prove Proposition 0.1 with a similar triangle inequality argument to how we showed that resilient sets have small modulus of continuity. However, we now need to consider multiple resilient sets S_i rather than a single S'.

Suppose S is $(\rho, \frac{\alpha}{4})$-resilient around μ–and thus also $(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{4})$-resilient by Lemma 0.2–and let $S_1, \dots, S_m$ be a maximal collection of subsets of $[n]$ such that:

1. $|S_j| \geq \frac{\alpha}{2}n$ for all j.
2. S_j is $(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{2})$-resilient (with mean μ_j).
3. $S_j \cap S_{j'} = \emptyset$ for all $j \neq j'$.

Clearly $m \leq \frac{2}{\alpha}$. We claim that S has large intersection with at least one of the S_j and hence μ_j is close to μ. By maximality of the collection $\{S_j\}_{j=1}^m$, it must be that $S_0 = S \backslash (S_1 \cup \dots \cup S_m)$ cannot be added to the collection. First suppose that $|S_0| \geq \frac{\alpha}{2}n$. Then S_0 is $(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{2})$-resilient (because any subset of $\frac{\alpha}{2}|S_0|$ points in S_0 is a subset of at least $\frac{\alpha}{4}|S|$ points in S). This contradicts the maximality of $\{S_j\}_{j=1}^m$, so we must have $|S_0| < \frac{\alpha}{2}n$.

Now, this implies that $|S \cap (S_1 \cup \dots \cup S_m)| \geq \frac{\alpha}{2}n$, so by pigeonhole we must have $|S \cap S_j| \geq \frac{\alpha}{2}|S_j|$ for some j. Letting $T = S \cap S_j$ as before, we find that $|T| \geq \frac{\alpha}{2}|S_j| \geq \frac{\alpha}{4}|S|$ and hence by resilience of S_j and S we have $\|\mu - \mu_j\| \leq 2 \cdot (\frac{4}{\alpha}\rho) = \frac{8}{\alpha}\rho$ by the same triangle inequality argument as before. □

Better bounds for well-separated clusters. Proposition 0.1 is powerful because it holds under very minimal conditions (we do not need to assume anything about separation of clusters or even about any of the clusters other than the one we are estimating). However, its guarantees are also minimal—we only know that we get approximate parameter recovery in the list-decoding model, and cannot say anything about cluster recovery. We next obtain a stronger bound assuming that the data can actually be separated into clusters (with a small fraction of outliers) and that the means are well-separated. This stronger result both gives cluster recovery, and gives better bounds for parameter recovery:

Proposition 0.3. *Suppose that a set of points $\{x_1, \dots, x_n\}$ can be partitioned into k sets $C_1, \dots, C_k$ of size $\alpha_1 n, \dots, \alpha_k n$, together with a fraction ϵn of outliers ($\epsilon = 1 - (\alpha_1 + \dots + \alpha_k)$), where $2\epsilon \le \alpha = \min_{k=1}^{k} \alpha_j$. Further suppose that*

- *Each cluster is (ρ_1, ϵ)-resilient and $(\rho_2, 2\epsilon/\alpha)$-resilient.*
- *The means are well-separated: $\Delta > \frac{4\rho}{\epsilon}$ where $\Delta = \min_{j \neq j'} \|\mu_j - \mu_{j'}\|_2$.*

Then we can output clusters $\hat{C}_1, \dots, \hat{C}_k$ such that:

- *$|C_j \triangle \hat{C}_j| \le \mathcal{O}(\epsilon/\alpha)|C_j|$ (cluster recovery)*
- *The mean $\hat{\mu}_j$ of $\hat{C}_j$ satisfies $\|\hat{\mu}_j - \mu_j\|_2 \le 2\rho_2$ (parameter recovery).*

Proof. We will construct a covering by resilient sets as before, but this time make use of the fact that we know the data can be approximately partitioned into clusters. Specifically, let $S_1, \dots, S_k$ be a collection of k sets such that:

- $|S_l| \ge \alpha n$
- The S_l are disjoint and contain all but ϵn points.
- Each S_l is (ρ_1, ϵ)-resilient.

We know that such a collection exists because we can take the C_j themselves. Now call a set S "j-like" if it contains at least $\alpha_j(\epsilon/\alpha)|S|$ points from C_j. We claim that each S_l is j-like for exactly one j. Indeed, by pigeonhole it must be j-like for at least one j since $\epsilon/\alpha \le 1/2 < 1$.

In the other direction, note that if S if j-like then $S \cap C_j$ contains at least $(\alpha_j/\alpha)\epsilon$ of the points in S, and at least $(|S|/n)(\epsilon/\alpha) \ge \epsilon$ of the points in C_j. Thus by resilience of both sets, the means of both S and C_j are within $\frac{\rho_1}{\epsilon}$ of the mean of $S \cap C_j$ and hence within $\frac{2\rho_1}{\epsilon}$ of each other. In summary, $\|\mu_j - \mu_S\|_2 \le \frac{2\rho_1}{\epsilon}$. Now if S were j-like and also j'-like, then we would have $\|\mu_j - \mu_{j'}\|_2 \le \frac{4\rho_1}{\epsilon}$, which contradicts the separation assumption.

Since S_l is j-like for a unique j, it contains at most $(\epsilon/\alpha)|S_l|$ points from any of the other $C_{j'}$, together with at most ϵn outliers. Moreover, since the other $S_{l'}$ are not j-like, S_l is missing at most $\alpha_j(\epsilon/\alpha)n$ points from C_j. Thus $S_l \cap C_j$ is missing at most $2\epsilon/\alpha|S_l|$ points from S_l and at most $\epsilon/\alpha|C_j|$ points from C_j. By resilience their means are thus within $2\rho_2$ of each other, as claimed. □