Lecture 7: Efficient filtering from spectral signatures for Gaussian data

October 20, 2025

In the last lecture, we established a stronger regularity condition that Gaussian data satisfies with high probability, as well as a stronger notion of spectral signatures that they satisfy. In this class, we'll show how to use these ideas to devise robust mean estimation algorithms for Gaussian data that are able to obtain much better error rates than the previous algorithm could achieve.

1 Recap: spectral signatures for Gaussian data

We briefly recall the definition of ε-goodness, and the resulting spectral signatures one obtains for Gaussian data:

Definition 1.1. A set of points S is ε-good with respect to μ if it satisfies the following conditions:

- we have that

$$\|\mu(S) - \mu\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon}\,, \qquad \text{and} \qquad \left\|\frac{1}{|S|}\sum_{i\in S}(X_i-\mu)(X_i-\mu)^\top\right\|_2 \lesssim \varepsilon \log 1/\varepsilon\,, \tag{1}$$

- for all $T \subset S$ with $|T| = \varepsilon n$ we have

$$\|\mu(T) - \mu\|_2 \lesssim \sqrt{\log 1/\varepsilon}\,, \qquad \text{and} \qquad \left\|\frac{1}{|T|}\sum_{i\in T}(X_i-\mu)(X_i-\mu)^\top\right\|_2 \lesssim \log 1/\varepsilon\,. \tag{2}$$

As we saw in the previous lecture, a set of samples from a Gaussian of sufficiently large size is ε-good with high probability. We then have the following spectral signature theorem:

Theorem 1.1. *Let $\mu \in \mathbb{R}^d$ and let $\varepsilon \in (0, 1/2)$. Let $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ be an ε-corrupted set of points where S_{good} is ε-good with respect to μ. Let $w \in \mathscr{W}_{S,\varepsilon}$. Then*

$$\|\mu(w) - \mu\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon} + \sqrt{\varepsilon\left(\|\Sigma(w) - I\|_2 + \varepsilon \log 1/\varepsilon\right)}\,.$$

2 Filtering for Gaussian data

Recall that each iteration of the filtering algorithm presented before worked in two steps: first, we find the top eigenvector and eigenvalue of $\Sigma(w)$, checked if the eigenvalue was below a threshold, and if not, used the top eigenvector to define scores, and applied the univariate filtering algorithm on these scores. The filtering algorithm we'll present here is extremely similar, but with one crucial difference in how we will apply the univariate filter.

First, let's see why second moment filter is insufficient here. Notice that it will achieve error $O(\sqrt{\varepsilon})$, since ε-good data also satisfies the bounded covariance property that the second moment filter needs to get error $O(\sqrt{\varepsilon})$. However, this analysis is tight: one can show that the second moment filter can only achieve error $\Omega(\sqrt{\varepsilon})$, as opposed to the $O(\varepsilon\sqrt{\log 1/\varepsilon})$ that we will obtain later on.

The main problem is that if univariate filter requires that the contribution to the weighted sum of the scores from the points in S_{bad} is greater than the contribution from the points in S_{good}. When $\|\Sigma(S)\|_2 > C$ for some sufficiently large constant C, this is definitely true: by regularity, we should expect that $\sum_{i \in S_{\text{good}}} \frac{1}{n} \langle X_i - \mu, v \rangle^2 \approx 1$, and so for $\|\Sigma(S)\|_2 > C$ it must be that $\sum_{i \in S_{\text{bad}}} \frac{1}{n} \langle X_i - \mu, v \rangle^2$ contributes an overwhelming fraction of the spectral norm. Indeed, this is roughly speaking how the proof for robust mean estimation with bounded second moments works.

However, to get error $\varepsilon\sqrt{\log 1/\varepsilon}$, we must now consider regimes where $\|\Sigma(S)\|_2 \approx 1 + o(1)$, and now it is possible to construct instances where $\sum_{i \in S_{\text{good}}} \frac{1}{n} \langle X_i - \mu, v \rangle^2 \approx 1$, and $\sum_{i \in S_{\text{good}}} \frac{1}{n} \langle X_i - \mu, v \rangle^2 = o(1)$, and yet, the mean shift is $\omega(\varepsilon\sqrt{\log 1/\varepsilon})$. In this case, if we naively applied the univariate filter, it would remove substantially more mass from the good points from the bad points, which would violate the safety condition and cause bad results.

To get around this, we will instead only apply the univariate filter on the top 2ε-quantile of scores. Recall that given a set of weights $w^{(t)}$ at time t, the scores are given by

$$\tau_i^{(t)} = \left\langle v^{(t)}, X_i - \mu\left(w^{(t)}\right) \right\rangle^2 , \tag{3}$$

where $v^{(t)}$ is the top eigenvector of $\Sigma^{(t)}$. At a high level, ε-goodness implies that that largest ε-quantile of scores from the good points are at most $O(\log 1/\varepsilon)$ on average, but as we'll see, Theorem 1.1 will imply that if the mean shift is $\omega(\varepsilon\sqrt{\log 1/\varepsilon})$, then the scores of the largest bad points must be $\omega(\log 1/\varepsilon)$, and hence the univariate filter will remove more mass from the bad points than from the good points. The formal pseudocode for this algorithm is presented in Algorithm 1. Our guarantee for this algorithm will be:

Algorithm 1 GaussianFilter

procedure GaussianFilter(S, ε)
  Let $w^{(0)} = w(S)$.
  Let $C \geq 11$ be a universal constant.
  Let $\Sigma^{(0)} = \Sigma(w^{(0)})$.
  Let $t = 0$.
  while $\left\|\Sigma^{(t)} - I\right\|_2 > C\varepsilon \log 1/\varepsilon$ **do**
    Let $\tau^{(t)}$ be as defined in (3).
    Sort the $\tau_i^{(t)}$ in decreasing order. WLOG assume that $\tau_1^{(t)} \geq \tau_2^{(t)} \geq \ldots \geq \tau_n^{(t)}$.
    Let N be the first index so that $\sum_{i=1}^{N} w_i^{(t)} > 2\varepsilon$.
    Let τ' and w' be the restriction of $\tau^{(t)}$ and $w^{(t)}$ to the first N indices.
    Let $w'' = 1\text{DFilter}(\tau', w')$.
    Let $w_i^{(t+1)} = w_i''$ for $i = 1, \ldots, N$, and $w_i^{(t+1)} = w_i^{(t)}$ otherwise.
    Let $\Sigma^{(t+1)} = \Sigma(w^{(t)})$.
    Let $t \leftarrow t + 1$.
  end while
  return $\mu(w^{(t)})$
end procedure

Theorem 2.1. *Let $\varepsilon < c$ for some universal constant $c > 0$ sufficiently small, and let $\mu \in \mathbb{R}^d$. Let $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ be ε-corrupted of size n, so that S_{good} is 3ε-good with respect to μ. Suppose that $n = \Omega(1/\varepsilon)$. Then,* Gaussian Filter$(S, \varepsilon)$ *runs in time* $\text{poly}(n, d)$ *and outputs $\widehat{\mu}$ so that $\|\widehat{\mu} - \mu\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon}$.*

As before, the main work will be to establish the following invariant, which is the same as in the second moment case:

$$\sum_{i \in S_{\text{good}} \cap S} \frac{1}{n} - w_i^{(t)} < \sum_{i \in S_{\text{bad}}} \frac{1}{n} - w_i^{(t)} . \tag{4}$$

Specifically, we will show:

Lemma 2.2. *Let $s \geq 0$, and suppose that in iteration $t = s$, we have that (4) is satisfied and furthermore $\left\|\Sigma(w^{(t)}) - I\right\|_2 \gtrsim \varepsilon \log 1/\varepsilon$. Then (4) is satisfied at iteration $t = s + 1$.*

As before, the proof of the Theorem given the Lemma is not so hard:

Proof of Theorem 2.1 given Lemma 2.2. It is not hard to see that each iteration can be implemented in polynomial time, and again we cannot run for more than n iterations as the one dimensional filter removes at least one point from the support of the $w^{(t)}$ in each iteration. Thus the overall runtime is polynomial in n and d. We now turn our attention to correctness. The invariant (4) guarantees that at the final iterate T, we have that $w^{(T)} \in \mathscr{W}_{S,2\varepsilon}$. Moreover, the termination condition guarantees that $\left\|\Sigma(w^{(T)}) - I\right\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon}$. These two facts, along with Theorem 1.1, immediately yield correctness. □

Thus the remaining work will be to show Lemma 2.2. As in the analysis of the bounded second moment filter, Lemma 2.2 follows immediately from the following per-iterate guarantee, which is what we will actually show:

Lemma 2.3. *Let $s \geq 0$, and suppose that in iteration $t = s$, we have that (4) is satisfied and furthermore $\left\|\Sigma(w^{(t)}) - I\right\|_2 \gtrsim \varepsilon \log 1/\varepsilon$. Then*

$$\sum_{i \in S_{\text{good}} \cap S} w_i^{(t)} - w_i^{(t+1)} < \sum_{i \in S_{\text{bad}}} w_i^{(t)} - w_i^{(t+1)} \,. \tag{5}$$

In the rest of this section we will prove Lemma 2.3. For simplicity of notation, since we consider a fixed iteration t, let us drop the superscripts, and denote $w^{(t+1)}$ by w'. As in the algorithm, assume without loss of generality that $\tau_1 \geq \tau_2 \geq \ldots \tau_n$, and let T denote the smallest index so that $\sum_{i=1}^{T} w_i \geq 2\varepsilon$. Note that since $n = \Omega(1/\varepsilon)$, we may assume that $\sum_{i=1}^{T} w_i \leq 3\varepsilon$. Then, since $w_i = w_i'$ for $i > T$, (5) is equivalent to the condition that

$$\sum_{i \in S_{\text{good}} \cap [T]} w_i - w_i' < \sum_{i \in S_{\text{bad}} \cap [T]} w_i - w_i' \,. \tag{6}$$

Since we apply the one-dimensional filter only to the first T weights, the safety condition of the one-dimensional filter guarantees that (7) holds so long as

$$\sum_{i \in S_{\text{good}} \cap [T]} w_i \tau_i < \sum_{i \in S_{\text{bad}} \cap [T]} w_i \tau_i \,. \tag{7}$$

Proving this will be a bit more delicate than the corresponding proof presented previously. There are two places in which this analysis must deviate from the previous analysis. First, we are working with the matrix $\Sigma(w) - I$, which is not necessarily PSD. This causes a number of obnoxious technical issues: for instance, the largest eigenvalue of $\Sigma(w) - I$ in magnitude could a priori be negative, which would be bad for the arguments we wish to make. However, Corollary 1.6 from Lecture 6 essentially argues that this matrix can never be too negative. As a result, by being sufficiently careful, we can essentially pretend that the matrix is PSD.

Secondly, our deviations are much smaller, which is why we needed to only filter a 2ε-quantile of the scores. Therefore ultimately we care about "tail" properties of the scores, i.e. what is going on with relatively small mass subsets of scores. However, our per-iteration guarantee, namely, that $\|\Sigma(w) - I\|$ is large, is a global guarantee, over the whole distribution of scores. Thus, our analysis will need to use global properties of the scores as well to argue about the relative mass of the tails of the scores.

Let us see how to do this. Because our proofs will be a bit more complicated, we will not try to optimize constants (i.e. breakdown point) in this proof. First, we will show:

Lemma 2.4. *In the setting above, we have*

$$\sum_{i \in S_{\text{good}} \cap [T]} w_i \tau_i \lesssim \varepsilon \log 1/\varepsilon + \varepsilon^2 \left\|\Sigma(w) - I\right\|_2 \,.$$

Proof. Let v be the top eigenvector of $\Sigma(w)$. Let w'' be the set of weights defined so that $w''_i = w_i$ for $i \in S_{\text{good}} \cap [T]$ and $w'' = 0$ otherwise. Observe that $w'' \in \mathscr{W}_{S,3\varepsilon}$. We have

$$\begin{aligned}
\sum_{i\in S} w''_i \tau_i &= \sum_{i\in S} w''_i \left\langle X_i - \mu(w), v\right\rangle^2 \\
&= \sum_{i\in S} w'' \left\langle X_i - \mu(w''), v\right\rangle^2 + \left\|w''\right\|_1 \cdot \left\langle \mu(w) - \mu(w''), v\right\rangle^2 \\
&\le \sum_{i\in S} w'' \left\langle X_i - \mu(w''), v\right\rangle^2 + 3\varepsilon \left\|\mu(w) - \mu(w'')\right\|_2^2 \\
&\le \sum_{i\in S} w'' \left\langle X_i - \mu, v\right\rangle^2 + 3\varepsilon \left\|\mu(w) - \mu(w'')\right\|_2^2 \qquad (8) \\
&\lesssim \varepsilon \log 1/\varepsilon + 3\varepsilon \left\|\mu(w) - \mu(w'')\right\|_2^2 \qquad (9)
\end{aligned}$$

where (8) follows since the summation is minimized by subtracting out $\mu(w'')$, and the last inequality follows from 3ε-goodness of S_{good}. We also have that

$$\begin{aligned}
\left\|\mu(w) - \mu(w'')\right\|_2 &\le \left\|\mu(w) - \mu\right\|_2 + \left\|\mu - \mu(w'')\right\|_2 \\
&\lesssim \left\|\mu(w) - \mu\right\|_2 + \sqrt{\log 1/\varepsilon} \\
&\lesssim \sqrt{\varepsilon(\left\|\Sigma(w) - I\right\|_2 + \log 1/\varepsilon)} + \sqrt{\log 1/\varepsilon}\,, \qquad (10)
\end{aligned}$$

where the last line follows from Theorem 1.1. Since $(a+b)^2 \lesssim a^2 + b^2$, we have that

$$\left\|\mu(w) - \mu(w'')\right\|_2^2 \lesssim \varepsilon \left\|\Sigma(w) - I\right\|_2 + \log 1/\varepsilon\,.$$

Hence

$$\sum_{i\in S} w''_i \tau_i \lesssim \varepsilon \log 1/\varepsilon + \varepsilon^2 \left\|\Sigma(w) - I\right\|_2\,,$$

as claimed. □

As a consequence of this, we have:

Corollary 2.5. *$\tau_i \lesssim \log 1/\varepsilon + \varepsilon \left\|\Sigma(w) - I\right\|_2$ for all $i > T$.*

Proof. This follows directly from Lemma 2.4 and since $\sum_{i\in S_{\text{good}}\cap[T]} w_i \ge \varepsilon$, as

$$\sum_{i\in S_{\text{bad}}\cap[T]} w_i \le \sum_{i\in S_{\text{bad}}} w_i \le \varepsilon\,,$$

and we know that $\sum_{i\in[T]} w_i \ge 2\varepsilon$. □

We also have:

Lemma 2.6. *We have*

$$\sum_{i\in S_{\text{good}}\cap S} w_i\tau_i - 1 \lesssim C\varepsilon \log 1/\varepsilon + C\varepsilon \left\|\Sigma(w) - I\right\|_2\,.$$

Proof. Recall v is the top eigenvector of $\Sigma(w)$. Let w_g denote the restriction of w to $S_{\text{good}} \cap S$. Expanding, we have:

$$\begin{aligned}
\sum_{i \in S_{\text{good}} \cap S} w_i \tau_i &= \sum_{i \in S_{\text{good}} \cap S} w_i \left\langle v, X_i - \mu(w) \right\rangle^2 \\
&= \sum_{i \in S_{\text{good}} \cap S} w_i \left\langle v, X_i - \mu \right\rangle^2 + \left\| w \right\|_1 \cdot \left(\left\langle v, \mu(w_g) - \mu(w) \right\rangle^2 - \left\langle v, \mu(w) - \mu(w_g) \right\rangle^2 \right) \\
&\leq \sum_{i \in S_{\text{good}} \cap S} w_i \left\langle v, X_i - \mu(w) \right\rangle^2 + \left\| \mu(w_g) - \mu \right\|_2^2 \qquad (11) \\
&\leq 1 + c\varepsilon \log 1/\varepsilon + \left\| \mu(w_g) - \mu(w) \right\|_2^2 \ , \qquad (12)
\end{aligned}$$

where (11) from Cauchy-Schwartz and by removing a negative term, and (12) follows from Corollary 1.6 from Lecture 6.

To conclude, we observe that

$$\begin{aligned}
\left\| \mu(w_g) - \mu(w) \right\|_2 &\leq \left\| \mu(w_g) - \mu \right\|_2 + \left\| \mu - \mu(w) \right\|_2 \qquad (13) \\
&\lesssim \varepsilon \sqrt{\log 1/\varepsilon} + \sqrt{\varepsilon \left\| \Sigma(w) - I \right\|_2} \ , \qquad (14)
\end{aligned}$$

where the last inequality follows from Corollary 1.5 from Lecture 6 and Theorem 1.1. Squaring and combining terms yields the desired result. □

We are now (finally) ready to prove Lemma 2.3.

Proof of Lemma 2.3. As previously mentioned, it suffices to demonstrate (7). We are assuming that $\left\| \Sigma(w) - I \right\|_2 > C\varepsilon \log 1/\varepsilon$ for some constant $C > 0$ sufficiently large. By Lemma 2.4, and our assumption, it suffices to show that

$$\sum_{i \in S_{\text{bad}} \cap [T]} w_i \tau_i \gtrsim \left\| \Sigma(w) - I \right\|_2 \ . \qquad (15)$$

By definition, we know that

$$\sum_{i \in S} w_i \tau_i = v^\top \left(\Sigma(w) - I \right) v + 1 \ .$$

Notice in particular that Corollary 1.6 from Lecture 6 implies that the largest eigenvalue of $\Sigma(w) - I$ in magnitude must be positive, as all negative eigenvalues of $\Sigma(w) - I$ have magnitude at most $c\varepsilon \log 1/\varepsilon < C\varepsilon \log 1/\varepsilon$. Therefore we have that

$$\sum_{i \in S} w_i \tau_i - 1 \gtrsim \varepsilon \log 1/\varepsilon + \left\| \Sigma(w) - I \right\|_2 \ .$$

Combining this with Lemma 2.6, we have that for ε sufficiently small,

$$\sum_{i \in S_{\text{bad}}} w_i \tau_i \gtrsim \varepsilon \log 1/\varepsilon + \left\| \Sigma(w) - I \right\|_2 \ .$$

We now have

$$\begin{aligned}
\sum_{i \in S_{\text{bad}} \cap [T]} w_i \tau_i &\gtrsim \varepsilon \log 1/\varepsilon + \left\| \Sigma(w) - I \right\|_2 - \sum_{i \in S_{\text{bad}} \setminus [T]} w_i \tau_i \\
&\gtrsim \varepsilon \log 1/\varepsilon + \left\| \Sigma(w) - I \right\|_2 - \left(\sum_{i \in S_{\text{bad}}} w_i \right) \left(\log 1/\varepsilon + \varepsilon \left\| \Sigma(w) - I \right\|_2 \right) \qquad (16) \\
&\gtrsim \varepsilon \log 1/\varepsilon + \left\| \Sigma(w) - I \right\|_2 \gtrsim \left\| \Sigma(w) - I \right\|_2 \ ,
\end{aligned}$$

where (16) follows from Corollary 2.5. This proves (15), which completes the proof of the Lemma. □