[Lecture 14]

0.1 Efficient Algorithm for Robust Regression

We now turn to the question of efficient algorithms, focusing on linear regression (we will address finite-sample issues later). Recall that information-theoretically, we found that two conditions are sufficient to imply resilience:

- *Hypercontractivity:* For all v, $\mathbb{E}_{x\sim p}[\langle x, v\rangle^4] \le \kappa \mathbb{E}_{x\sim p}[\langle x, v\rangle^2]^2$.
- *Bounded noise:* $\mathbb{E}_{x\sim p}[xz^2x^\top] \preceq \sigma^2 \mathbb{E}_{x\sim p}[xx^\top]$.

As for mean estimation under bounded covariance, our strategy will be to check whether these two properties hold for the empirical distribution, and if they don't we will filter out points such that we guarantee removing more bad points than good points.

Unfortunately, the hypercontractivity condition is difficult to verify because it involves fourth moments. We will thus need to assume a stronger condition, called *certifiable hypercontractivity*:

$$\mathbb{E}_{x\sim p}[\langle x, v\rangle^4] \preceq_{\text{sos}} \kappa \mathbb{E}_{x\sim p}[\langle x, v\rangle^2]^2, \tag{1}$$

where the LHS and RHS are considered as polynomials in v.

We will also need to introduce one additional piece of sum-of-squares machinery, called *pseudoexpectations*:

Definition 0.1. A *degree-$2k$ pseudoexpectation* is a linear map E from the space of degree-$2k$ polynomials to $\mathbb{R}$ satisfying the following two properties:

- $E[1] = 1$ (where 1 on the LHS is the constant polynomial).
- $E[p^2] \ge 0$ for all polynomials p of degree at most k.

We let $\mathcal{E}$ or $\mathcal{E}_{2k}$ denote the set of degree-$2k$ pseudoexpectations.

The space $\mathcal{E}$ can be optimized over efficiently, because it has a separation oracle expressible as a sum-of-squares program. Indeed, checking that $E \in \mathcal{E}$ amounts to solving the problem $\min\{E[p] \mid p \succeq_{\text{sos}} 0\}$, which is a sum-of-squares program because $E[p]$ is a linear function of p.

We are now ready to define our efficient algorithm for linear regression, Algorithm 1. It is closely analogous to the filter for mean estimation (Algorithm **??**).

Algorithm 1 `FilterLinReg`

1: Input: $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$.
2: Initialize weights $c_1, \dots, c_n = 1$.
3: Compute the empirical least squares regressor: $\hat{\theta}_c \stackrel{\text{def}}{=} (\sum_{i=1}^n c_i x_i x_i)^{-1}(\sum_{i=1}^n c_i x_i y_i)$.
4: Find, if possible, a pseudoexpectation $E \in \mathcal{E}_4$ such that $E[\frac{1}{n}\sum_{i=1}^n c_i \langle x_i, v\rangle^4] \ge 3\kappa E[(\frac{1}{n}\sum_{i=1}^n c_i \langle x_i, v\rangle^2)^2]$.
5: If E exists, let $\tau_i = E[\langle x_i, v\rangle^4]$ and update $c_i \leftarrow c_i \cdot (1 - \tau_i / \max_j \tau_j)$, and return to line 3.
6: Otherwise, find, if possible, a vector $v \in \mathbb{R}^d$ such that $\sum_{i=1}^n c_i \langle x_i, v\rangle^2 (y_i - \langle \hat{\theta}_c, x_i\rangle)^2 \ge 24\sigma^2 \sum_{i=1}^n c_i \langle x_i, v\rangle^2$.
7: If v exists, let $\tau_i = \langle x_i, v\rangle^2 (y_i - \langle \hat{\theta}_c, x_i\rangle)^2$ and update $c_i \leftarrow c_i \cdot (1 - \tau_i / \max_j \tau_j)$, and return to line 3.
8: Otherwise, output $\hat{\theta}_c$.

The algorithm first optimizes over $E \in \mathcal{E}_4$ to try to refute hypercontractivity; if it does so successfully, it filters according to $E[\langle x_i, v\rangle^4]$. Otherwise, it tries to refute the bounded noise condition, using $\hat{\theta}_c$ as a proxy for θ^* to approximate $z = y - \langle \theta^*, x\rangle$. Again, if it successfully refutes bounded noise it filters based on this. If it fails to refute either condition, we can safely output $\hat{\theta}_c$, which will be close to θ^* by resilience.

Analyzing Algorithm 1. We will show that Algorithm 1 enjoys the following loss bound:

Proposition 0.2. *Suppose that a good set S of $(1-\epsilon)n$ of the x_i satisfy:*

$$\frac{1}{n}\sum_{i\in S}\langle x_i, v\rangle^4 \preceq_{\text{sos}} \kappa(\frac{1}{n}\sum_{i\in S}\langle x_i, v\rangle^2)^2 \text{ and } \frac{1}{n}\sum_{i\in S} z_i^2 x_i x_i^\top \preceq \sigma^2 \frac{1}{n}\sum_{i\in S} x_i x_i^\top. \tag{2}$$

Then assuming $\epsilon \le \frac{1}{100}$ and $\kappa\epsilon \le \frac{1}{50}$, the output of Algorithm 1 has excess loss at most $250\sigma^2\epsilon$.

Proof. We analyze Algorithm 1 similarly to Algorithm **??**. Specifically, we will establish the invariant that we always remove more bad points than good points. This requires showing that $\sum_{i\in S} c_i\tau_i \le \frac{1}{2}\sum_{i=1}^n c_i\tau_i$ for both choices of τ_i in the algorithm. Concretely, we need to show:

$$\sum_{i\in S} c_i E[\langle x_i, v\rangle^4] \le_? \frac{1}{2}\sum_{i=1}^n c_i E[\langle x_i, v\rangle^4] \text{ and } \sum_{i\in S} c_i\langle x_i, v\rangle^2(y_i - \langle\hat{\theta}_c, x_i\rangle)^2 \le_? \frac{1}{2}\sum_{i=1}^n c_i\langle x_i, v\rangle^2(y_i - \hat{\theta}_c, x_i))^2. \tag{3}$$

For both of these we will want the following intermediate lemma, which states that deletions of hypercontractive distributions are hypercontractive:

Lemma 0.3. *Suppose that the set S of good points is hypercontractive in the sense that $\frac{1}{n}\sum_{i\in S}\langle x_i, v\rangle^4 \preceq_{\text{sos}} \kappa(\frac{1}{n}\sum_{i\in S}\langle x_i, v\rangle^2)^2$. Then, for any c_i such that $\frac{1}{n}\sum_{i\in S}(1-c_i) \le \epsilon$, we have*

$$\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^4 \preceq_{\text{sos}} \frac{\kappa}{1-\kappa\epsilon}(\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2)^2. \tag{4}$$

Proof. We expand directly; let

$$A = \frac{1}{n}\sum_{i\in S}\langle x_i, v\rangle^4, \quad B = \frac{1}{n}\sum_{i\in S}\langle x_i, v\rangle^2, \tag{5}$$

$$C = \frac{1}{n}\sum_{i\in S}(1-c_i)\langle x_i, v\rangle^4, \quad D = \frac{1}{n}\sum_{i\in S}(1-c_i)\langle x_i, v\rangle^2. \tag{6}$$

Then our goal is to show that $\frac{\kappa}{1-\kappa\epsilon}(B-D)^2 - (A-C) \succeq_{\text{sos}} 0$. We are also given that (i) $\kappa B^2 \succeq_{\text{sos}} A$ and we observe that (ii) $C \succeq_{\text{sos}} D^2/(\frac{1}{n}\sum_{i=1}^n(1-c_i) \succeq_{\text{sos}} D^2/\epsilon$ by Cauchy-Schwarz. We thus have

$$\frac{\kappa}{1-\kappa\epsilon}(B-D)^2 - (A-C) = \frac{\kappa}{1-\kappa\epsilon}B^2 - \frac{2\kappa}{1-\kappa\epsilon}BD + \frac{\kappa}{1-\kappa\epsilon}D^2 - A + C \tag{7}$$

$$\succeq^{(i)}_{\text{sos}} (\frac{\kappa}{1-\kappa\epsilon} - \kappa)B^2 - \frac{2\kappa}{1-\kappa\epsilon}BD + (\frac{\kappa}{1-\kappa\epsilon}D^2 + C) \tag{8}$$

$$\succeq^{(ii)}_{\text{sos}} (\frac{\kappa}{1-\kappa\epsilon} - \kappa)B^2 - \frac{2\kappa}{1-\kappa\epsilon}BD + (\frac{\kappa}{1-\kappa\epsilon} + \frac{1}{\epsilon})D^2 \tag{9}$$

$$= \frac{\kappa^2\epsilon}{1-\kappa\epsilon}B^2 - \frac{2\kappa}{1-\kappa\epsilon}BD + \frac{1/\epsilon}{1-\kappa\epsilon}D^2 \tag{10}$$

$$= \frac{\epsilon}{1-\kappa\epsilon}(\kappa B - D/\epsilon)^2 \succeq_{\text{sos}} 0, \tag{11}$$

as was to be shown. □

With Lemma 0.3 in hand, we proceed to analyze the filtering steps by establishing the inequalities in (3). For the first, observe that

$$\frac{1}{n}\sum_{i\in S} c_i E[\langle x_i, v\rangle^4] \overset{(i)}{\le} \frac{\kappa}{1-\kappa\epsilon}E[(\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2)^2] \tag{12}$$

$$\overset{(ii)}{\le} \frac{\kappa}{1-\kappa\epsilon}E[(\frac{1}{n}\sum_{i=1}^n c_i\langle x_i, v\rangle^2)^2] \tag{13}$$

$$\overset{(iii)}{\le} \frac{1}{3(1-\kappa\epsilon)}\frac{1}{n}\sum_{i=1}^n c_i E[\langle x_i, v\rangle^4]. \tag{14}$$

Here (i) is by Lemma 0.3 (and the fact that $E[p] \le E[q]$ if $p \preceq_{\text{sos}} q$), (ii) is by the fact that adding the $c_i\langle x_i, v\rangle^2$ terms for $i \notin S$ is adding a sum of squares, and (iii) is by the assumption that E refutes hypercontractivity. Thus as long as $\kappa\epsilon \le \frac{1}{3}$ we have the desired property for the first filtering step.

For the second, observe that

$$\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2 (y_i - \langle\hat{\theta}_c, x_i\rangle)^2 \le \frac{2}{n}\sum_{i\in S} c_i(\underbrace{\langle x_i, v\rangle^2 (y_i - \langle\theta^*, x_i\rangle)^2}_{(a)} + \underbrace{\langle x_i, v\rangle^2\langle\hat{\theta}_c - \theta^*, x_i\rangle^2}_{(b)}). \tag{15}$$

We will bound (a) and (b) in turn. To bound (a) note that

$$\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2 (y_i - \langle\theta^*, x_i\rangle)^2 = \frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2 z_i^2 \tag{16}$$

$$\le \frac{1}{n}\sum_{i\in S} \langle x_i, v\rangle^2 z_i^2 \tag{17}$$

$$\le \frac{\sigma^2}{n}\sum_{i\in S} \langle x_i, v\rangle^2 \tag{18}$$

$$\le \frac{\sigma^2}{1-\kappa\epsilon}\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2, \tag{19}$$

where the last line invokes Lemma 0.3 and the middle inequality is by the bounded noise assumption for S.

To bound (b), let $R = \frac{1}{(1-\epsilon)n}\sum_{i\in S}\langle\hat{\theta}_c - \theta^*, x_i\rangle^2$, which is the excess loss of $\hat{\theta}_c$ and what we eventually hope to bound when the algorithm terminates. We use Cauchy-Schwarz and hypercontractivity:

$$\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2\langle\hat{\theta}_c - \theta^*, x_i\rangle^2 \le \Big(\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^4\Big)^{1/2}\Big(\frac{1}{n}\sum_{i\in S} c_i\langle\hat{\theta}_c - \theta^*, x_i\rangle^4\Big)^{1/2} \tag{20}$$

$$\le \frac{\kappa}{1-\kappa\epsilon}\Big(\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2\Big)\Big(\frac{1}{n}\sum_{i\in S} c_i\langle\hat{\theta}_c - \theta^*, x_i\rangle^2\Big) \tag{21}$$

$$\le \frac{\kappa R}{1-\kappa\epsilon}\Big(\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2\Big). \tag{22}$$

Combining these, we obtain

$$\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2 (y_i - \langle\hat{\theta}_c, x_i\rangle)^2 \le \frac{2\sigma^2 + 2\kappa R}{1-\kappa\epsilon}\Big(\frac{1}{n}\sum_{i\in S} c_i\langle x_i, v\rangle^2\Big). \tag{23}$$

But we are assuming that overall

$$\frac{1}{n}\sum_{i=1}^{n} c_i\langle x_i, v\rangle^2 (y_i - \langle\hat{\theta}_c, x_i\rangle)^2 = S\cdot\Big(\frac{1}{n}\sum_{i=1}^{n} c_i\langle x_i, v\rangle^2\Big), \tag{24}$$

with $S \ge 10\sigma^2$. Thus we are safe as long as $\frac{2\sigma^2+2\kappa R}{1-\kappa\epsilon} \le S/2$, and the main remaining issue is to bound R in terms of S. To do so, note that the distribution weighted by c_i satisfies the hypercontractive and bounded noise conditions with parameters $\frac{3\kappa}{1-\epsilon}$ and $\frac{S}{1-\epsilon}$. It follows from Proposition **??** (the resilience bound for linear regression) that $R \le 10S\epsilon/(1-\epsilon)$ as long as $\epsilon(\kappa/(1-\epsilon) - 1) \le \frac{1}{6}$ and $\epsilon \le \frac{1}{8}$. We thus need to verify that

$$\frac{2\sigma^2 + 20\kappa S\epsilon/(1-\epsilon)}{1-\kappa\epsilon} \le_? S/2, \tag{25}$$

which if $\epsilon\kappa \le 0.02$ and $\epsilon \le 0.01$ reduces to $2\sigma^2 + 0.4S/0.99 \le 0.49S$, which holds if $S \ge 24\sigma^2$, which is the cutoff in the algorithm.

Since the algorithm terminates with $S \le 24\sigma^2$, we incidentally also have that $R \le 250\sigma^2\epsilon$, as claimed. □