STATS 314A: Advanced Statistical Theory
The Sum-of-Squares Algorithmic Paradigm in Statistics
Instructor: Tselil Schramm

Lecture 8
April 25, 2022

Lecture 8: Robust Linear Regression

In this lecture we'll apply the sum-of-squares paradigm to the problem of linear regression with adversarial corruptions. We'll depart slightly from previous lectures and work in a setting where the data is not necessarily well-modeled as a linear regression, so there is no parameter that we are trying to identify. Rather than giving an SoS proof of identifiability, we'll show (using an SoS proof) that the if D_1, D_2 are two distributions which are close in total variation distance, then the fit of a linear function for D_1 can be bounded by the fit of the linear function for D_2 (so long as D_1 is hypercontractive).

These notes have not been reviewed with the same scrutiny applied to formal publications. There may be errors.

1 Robust Linear Regression

In *linear regression*, we have sample access to a distribution $\mathcal{D}$ over pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, where x are the *covariates* and y are the *labels*, and our goal is to find the best *linear* function which relates the covariates and labels. The way we measure fit can vary, but in this lecture we'll consider the *squared loss.* For a linear function defined by $\theta \in \mathbb{R}^d$, we define the squared loss error,

$$\mathrm{err}_{\mathcal{D}}(\theta) = \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}} (\langle \theta, x\rangle - y)^2,$$

and we define the minimum error achievable on $\mathcal{D}$,

$$\mathrm{err}(\mathcal{D}) = \arg\min_{\theta\in\mathbb{R}^d} \mathrm{err}_{\mathcal{D}}(\theta).$$

Here, we won't make the assumption that the data is generated by a linear model (which is referred to as the *realizable* setting). We are just looking for the best-fit linear function.

The non-robust setting. In the non-robust setting, we are given a dataset of pairs $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ sampled from $\mathcal{D}$. Let $\hat{\mathcal{D}}$ be the uniform distribution over these samples. There is a simple linear-algebraic closed form for the best-fit line for this set of samples. Letting X be the $d \times n$ matrix whose ith row is given by x_i, and $y \in \mathbb{R}^n$ be the vector whose ith entry is y_i,

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d} \mathop{\mathbf{E}}_{(x,y)\sim\hat{\mathcal{D}}} (\langle x, \theta\rangle - y)^2 = \arg\min_{\theta} \frac{1}{n}\|X\theta - y\|^2 = (X^{-1}X)X^{-1}y.$$

The idea is then to ensure that we have enough samples n so that $\hat{\theta}$ generalizes; that is, that it achieves an error on the "population" data from $\mathcal{D}$, $\mathrm{err}_{\mathcal{D}}(\theta)$, which is not too far from the error on the sample, $\mathrm{err}(\hat{\mathcal{D}}) = \mathrm{err}_{\hat{\mathcal{D}}}(\hat{\theta})$. For the purposes of this lecture, we will completely ignore this aspect of generalization/concentration; we'll assume that any linear function with bounded error on $\hat{\mathcal{D}}$ has a similar bound on the error in $\mathcal{D}$, and instead, we will be concerned with handling adversarial corruptions to $\hat{\mathcal{D}}$.

The strong contamination model. Here, we'll be interested in the following setting: the points (x_1, y_1), $\dots, (x_n, y_n)$ are first sampled from $\mathcal{D}$. Then, an adversary makes arbitrary corruptions to an ε fraction of points, so that we observe $(x_1', y_1'), \dots, (x_n', y_n')$, and our only promise is that for a subset $I \subset [n]$ of size $|I| = (1-\varepsilon)n$, each $i \in I$ has $x_i' = x_i$ and $y_i' = y_i$. Our goal is, given access only to this corrupted sample set, to find some $\hat{\theta} \in \mathbb{R}^d$ which makes $\mathrm{err}_{\hat{D}}(\theta)$ as small as possible. Ideally, we would like that the fit not be too much worse than in the uncorrupted setting, matching the uncorrupted setting closely when the fraction of corruptions ε is small:

$$\mathrm{err}_{\hat{D}}(\theta) \leqslant (1 + f(\varepsilon))\mathrm{err}(\hat{D}),$$

where $f : \mathbb{R} \to \mathbb{R}$ is such that $\lim_{\varepsilon \to 0} f(\varepsilon) = 0$.

2 Relating error for distributions close in total variation

Here, we will show that if we have two distributions which are close in total variation distance, and if at least one of the two distribution satisfies a hyperccontractivity condition, then any linear function achieves similar error on both. The condition we will need is the following:

Definition 2.1. Let k be an even integer. We say a distribution $\mathcal{D}$ over $\mathbb{R}^d$ is (k, C_k)-hypercontractive if for all $v \in \mathbb{R}^d$,

$$\mathop{\mathbf{E}}_{X\sim\mathcal{D}_X}[\langle v, X\rangle^k] \leqslant C_k^{k/2} \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_X}[\langle v, X\rangle^2]^{k/2}.$$

This is similar to the subgaussianity condition we saw in Lecture 4.

Lemma 2.2. *Suppose $\mathcal{D}_1, \mathcal{D}_2$ are distributions over $\mathbb{R}^d \times \mathbb{R}$ which satisfy $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \varepsilon$, and suppose $\mathcal{D}_1$ is (k, C_k)-hypercontractive for k an even positive integer.*[1] *Then for any $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$\mathrm{err}_{\mathcal{D}_1}(\theta_2) \leqslant (1 + O(C_k {\varepsilon}^{1-2/k})) \cdot \mathrm{err}_{\mathcal{D}_2}(\theta_2) + O(C_k {\varepsilon}^{1-2/k}) \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_1).$$

Before we prove this lemma, a word or two about its intended use. Suppose we are in a setting where $\hat{\mathcal{D}}$ is k-hypercontractive. We have access only to the corrupted distribution $\hat{\mathcal{D}}'$, but we can try to find a distribution $\mathcal{D}_2$ which minimizes $\mathrm{err}(\mathcal{D}_2)$ subject to $\mathrm{d}_{\mathrm{TV}}(\hat{\mathcal{D}}', \mathcal{D}_2) \leqslant \varepsilon$, which implies by the triangle inequality $\mathrm{d}_{\mathrm{TV}}(\hat{\mathcal{D}}, \mathcal{D}_2) \leqslant 2\varepsilon$. Because $\mathcal{D}_2$ minimizes $\mathrm{err}(\mathcal{D}_2)$ subject to $\mathrm{d}_{\mathrm{TV}}(\hat{\mathcal{D}}', \mathcal{D}_2) \leqslant \varepsilon$, in particular $\mathrm{err}(\mathcal{D}_2) \leqslant \mathrm{err}(\hat{\mathcal{D}})$. We could then take $\theta_2 = \arg\min_\theta \mathrm{err}_{\mathcal{D}_2}(\theta)$, and then, choosing $\theta_1 = \arg\min_\theta \mathrm{err}_{\hat{D}}(\theta)$, we have from Lemma 2.2 that

$$\mathrm{err}_{\hat{D}}(\theta_2) \leqslant (1 + O(C_k {\varepsilon}^{1-2/k})) \cdot \mathrm{err}(\mathcal{D}_2) + O(C_k {\varepsilon}^{1-2/k}) \cdot \mathrm{err}_{\hat{D}}(\theta_1) \leqslant (1 + O(C_k {\varepsilon}^{1-2/k})) \cdot \mathrm{err}(\hat{\mathcal{D}})$$

so that θ_2 matches the optimal error within a factor of $(1 + f(\varepsilon))$ for f going to zero with epsilon; the stronger the hypercontractive property of $\hat{\mathcal{D}}$, the faster f goes to zero with epsilon, and the better our approximation. Later, our strategy will be to prove a sum-of-squares version of this lemma and then use a degree-$O(k)$ sum-of-squares to find a pseudodistribution with the desired properties of $\mathcal{D}_2$.

Notice that, perhaps surprisingly, we did not require $\mathcal{D}_2$ to be hypercontractive.

[1] If we make the weaker requirement that only the marginal of $\mathcal{D}_1$ on the covariates x is hypercontractive, a weaker version of this lemma, with a second error term, is true. In some contexts it is important to let the marginal on y *not* satisfy hypercontractivity. Here we'll make this more stringent assumption for clarity of exposition; we'll comment in the proof where it could be relaxed.

Proof of Lemma 2.2. We'll sample $(x, y) \sim \mathcal{D}_1$ and $(a, b) \sim \mathcal{D}_2$ according to the total variation coupling between the distributions, so that $(x, y) = (a, b)$ with probability $1 - \varepsilon$. Then by definition,

$$\begin{aligned}
\mathrm{err}_{\mathcal{D}_1}(\theta_2) &= \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_1} (\langle\theta_2, x\rangle - y)^2 \\
&= \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)=(x,y)}(\langle\theta_2, x\rangle - y)^2\right] + \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)\neq(x,y)}(\langle\theta_2, x\rangle - y)^2\right] \\
&= \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)=(x,y)}(\langle\theta_2, a\rangle - b)^2\right] + \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)\neq(x,y)}(\langle\theta_2, x\rangle - y)^2\right] \\
&\leqslant \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[(\langle\theta_2, a\rangle - b)^2\right] + \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)\neq(x,y)}(\langle\theta_2, x\rangle - y)^2\right] \\
&= \mathrm{err}_{\mathcal{D}_2}(\theta_2) + \mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)\neq(x,y)}(\langle\theta_2, x\rangle - y)^2\right],
\end{aligned}$$

where in the first step we used that the sum of the indicators is 1, in the second step we used that $(a, b) = (x, y)$ in the first term, in the third step we were able to drop the indicator because $(\langle\theta_2, a\rangle - b)^2$ is a square, and finally we applied the definition of the error.

Now we'll bound this second term in which we have the $(a, b) \neq (x, y)$. We'll use Hölder's inequality to separate the event over a, b from the term involving x, y. Hölder's inequality states that for any $p, q \geqslant 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, $\langle u, v\rangle \leqslant \|u\|_p \|v\|_q$. We'll apply it with $p = \frac{k}{k-2}$ and $q = \frac{k}{2}$.

$$\begin{aligned}
\mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)\neq(x,y)}(\langle\theta_2, x\rangle - y)^2\right] &\leqslant \left(\mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)\neq(x,y)}^{k/(k-2)}\right]\right)^{1-2/k} \left(\mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[(\langle\theta_2, x\rangle - y)^k\right]\right)^{2/k} \\
&= \left(\mathop{\mathbf{E}}_{\substack{(x,y)\sim\mathcal{D}_1\\(a,b)\sim\mathcal{D}_2}} \left[\mathbf{1}_{(a,b)\neq(x,y)}\right]\right)^{1-2/k} \left(\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_1} \left[(\langle\theta_2, x\rangle - y)^k\right]\right)^{2/k} \\
&\leqslant \varepsilon^{1-2/k} \left(\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_1} \left[(\langle\theta_2, x\rangle - y)^k\right]\right)^{2/k}, \qquad (1)
\end{aligned}$$

where we have used that $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \varepsilon$. We now bound this last term. We add and subtract zero to get that

$$(\langle\theta_2, x\rangle - y)^k = (\langle\theta_2 - \theta_1, x\rangle + (\langle\theta_1, x\rangle - y)^k \leqslant 2^{k-1}\left(\langle\theta_2 - \theta_1, x\rangle^k + (\langle\theta_1, x\rangle - y)^k\right),$$

where we've used the SoS inequality $(a + b)^k \leqslant 2^{k-1}(a^k + b^k)$ from a previous lecture.

Now it is time to use the k-hypercontractivity of $\mathcal{D}_1$. Note that the first term, $\langle\theta_2 - \theta_1, x\rangle^k$ only involves x, while the second term involves y too. If we only were to assume that the marginal of $\mathcal{D}_1$ on x is hypercontractive, then we could just leave this term, and get an additive error of $(2\varepsilon)^{1-2/k}\,\mathbf{E}_{\mathcal{D}_1}[(\langle\theta_1, x\rangle - y)^k]^{2/k}$. In some cases, we would want to allow for only this marginal distribution to be hypercontractive. Here, for a simpler exposition, we have assumed $\mathcal{D}_1$ is hypercontractive on all coordinates, so

$$\mathop{\mathbf{E}}_{\mathcal{D}_1}[(\langle\theta_1, x\rangle - y)^k] \leqslant C_k^{k/2} \mathop{\mathbf{E}}_{\mathcal{D}_1}[(\langle\theta_1, x\rangle - y)^2]^{k/2} = \left(C_k \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_1)\right)^{k/2}.$$

Applying the k-hypercontractivity of $\mathcal{D}_1$ to the first term as well,

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_1}[\langle\theta_2 - \theta_1, x\rangle^k] \leqslant C_k^{k/2} \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_1}[\langle\theta_2 - \theta_1, x\rangle^2]^{k/2}$$

and again adding and subtracting y, then applying $(a+b)^2 \leqslant 2a^2 + 2b^2$,

$$\begin{aligned}
&= C_k^{k/2} \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_1} [(((\langle\theta_2, x\rangle - y) - (\langle\theta_1, x\rangle - y))^2]^{k/2} \\
&\leqslant C_k^{k/2} \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_1} [2\left((\langle\theta_2, x\rangle - y)^2 + (\langle\theta_1, x\rangle - y)^2\right)]^{k/2} \\
&= \left(2C_k \cdot \left(\mathrm{err}_{\mathcal{D}_1}(\theta_2) + \mathrm{err}_{\mathcal{D}_1}(\theta_1)\right)\right)^{k/2}.
\end{aligned}$$

Plugging this back in to (1),

$$\begin{aligned}
(1) &\leqslant (2\varepsilon)^{1-2/k}\left(\left(C_k \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_1)\right)^{k/2} + \left(2C_k \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_1) + 2C_k \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_2)\right)^{k/2}\right)^{2/k} \\
&\leqslant (2\varepsilon)^{1-2/k}\left(\left(C_k \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_1)\right)^{k/2} + \left(2C_k \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_1) + 2C_k \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_2)\right)^{k/2}\right)^{2/k} \\
&\leqslant (2\varepsilon)^{1-2/k} C_k \left(3\mathrm{err}_{\mathcal{D}_1}(\theta_1) + 2\mathrm{err}_{\mathcal{D}_1}(\theta_2)\right).
\end{aligned}$$

Putting everything together,

$$\begin{aligned}
\mathrm{err}_{\mathcal{D}_1}(\theta_2) &\leqslant \mathrm{err}_{\mathcal{D}_2}(\theta_2) + 3(2\varepsilon)^{1-2/k} C_k \mathrm{err}_{\mathcal{D}_1}(\theta_1) + 2C_k(2\varepsilon)^{1-1/2k}\mathrm{err}_{\mathcal{D}_1}(\theta_2) \\
(1 - 2C_k(2\varepsilon)^{1-2/k}) \cdot \mathrm{err}_{\mathcal{D}_1}(\theta_2) &\leqslant \mathrm{err}_{\mathcal{D}_2}(\theta_2) + 3(2\varepsilon)^{1-2/k} C_k \mathrm{err}_{\mathcal{D}_1}(\theta_1)
\end{aligned}$$

and dividing through by $1 - 2C_k(2\varepsilon)^{1-2/k}$ finishes the proof, since $1/(1-\delta) = 1 + O(\delta)$ when δ is small. □

3 Sum-of-squares algorithm for robust regression

As discussed above, if we could minimize $\mathrm{err}(\mathcal{D}_2)$ over distributions $\mathcal{D}_2$ which satisfy $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_2, \hat{\mathcal{D}}') \leqslant \varepsilon$, Lemma 2.2 guarantees that the minimizing $\theta_2 = \arg\min_\theta \mathrm{err}_{\mathcal{D}_2}(\theta)$ would have bounded error for $\hat{\mathcal{D}}$. We turn to the sum-of-squares paradigm to replace this optimization over distributions to optimization over pseudodistributions. We take the following polynomial optimization probelm over variables $w_1, \ldots, w_n \in \mathbb{R}$ where w_i represents the indicator that (x_i', y_i') is in the support of $\mathcal{D}_2$, $a_1, \ldots, a_n \in \mathbb{R}^d$ and $b_1, \ldots, b_n \in R$ representing the points in the support of $\mathcal{D}_2$, and $\theta \in \mathbb{R}^d$ the linear regression coefficients:

$$\begin{aligned}
\min \quad & \left(\frac{1}{n}\sum_{i=1}^n (\langle a_i, \theta\rangle - b_i)^2\right)^{k/2} \\
\text{subject to} \quad & \\
w_i^2 &= w_i \ \forall i \in [n] \\
w_i(a_i - x_i') &= 0 \ \forall i \in [n] \\
w_i(b_i - y_i') &= 0 \ \forall i \in [n] \\
\sum_{i=1}^n w_i &= (1-\varepsilon)n.
\end{aligned}$$

We can then think of $\mathcal{D}_2$ be the uniform distribution over the a_i, b_i. Our objective is $\min \mathrm{err}_{\mathcal{D}_2}(\theta)^{k/2}$ rather than $\min \mathrm{err}_{\mathcal{D}_2}(\theta)$l; this is because we will ultimately need to work with the $k/2$th power in our sum-of-squares proof of Lemma 2.2.

Getting a SoS proof of Lemma 2.2 will require two ingredients: the first is an SoS version of Hölder's inequality. The second is an SoS version of hypercontractivity. We will say a distribution $\mathcal{D}$ is degree-t

SoS-certifiably (C_k, k)-hypercontractive if there is a degree-t proof that $\mathbf{E}_{X\sim\mathcal{D}}[\langle X, v\rangle^k] \leqslant C_k^{k/2}\,\mathbf{E}[\langle X, v\rangle^2]^{k/2}$. Recall for example that $\mathcal{N}(0, \mathbb{1})$ is degree-k certifiably (k, k)-hypercontractive, as is the uniform distribution over $x_1, \dots, x_n \sim \mathcal{N}(0, \mathbb{1})$ when $n = d^{\Omega(k)}$; a more thorough account of the distributions which are known to be certifiably hypercontractive can be found in [KS17].

Theorem 3.1. *Let $k > 2$ be a power of 2. If $\hat{\mathcal{D}}$ is a degree-t SoS-certifiably (C_k, k)-hypercontractive, then for $\widetilde{\mathbf{E}}$ a degree-t pseudoexpectation optimizing the program above,*

$$\mathrm{err}_{\hat{\mathcal{D}}}(\widetilde{\mathbf{E}}[\theta]) \leqslant (1 + O(C_k \varepsilon^{1-2/k})) \cdot \mathrm{err}(\hat{\mathcal{D}}).$$

Proof. The error achieved by $\widetilde{\mathbf{E}}[\theta]$ can be bounded by

$$\mathrm{err}_{\hat{\mathcal{D}}}(\widetilde{\mathbf{E}}[\theta]) = \frac{1}{n}\sum_{i=1}^{n}\left(\langle \widetilde{\mathbf{E}}[\theta], x_i\rangle - y_i\right)^2 \leqslant \widetilde{\mathbf{E}}\left[\frac{1}{n}\sum_{i=1}^{n}(\langle \theta, x_i\rangle - y_i)^2\right] = \widetilde{\mathbf{E}}[\mathrm{err}_{\hat{\mathcal{D}}}(\theta)],$$

since $\widetilde{\mathbf{E}}[f]^2 \leqslant \widetilde{\mathbf{E}}[f^2]$ by Cauchy-Schwarz. Hence it is enough to give a sum-of-squares proof that $\mathrm{err}_{\hat{\mathcal{D}}}(\theta)$ is bounded.

We'll prove an SoS version of Lemma 2.2. Let $w_i' = w_i \cdot \mathbf{1}[(x_i, y_i) = (x_i', y_i')]$, we can verify that the w_i' also satisfy the axioms $w_i' = (w_i')^2$. Now, we use these w_i' to simulate the "coupling" from the argument of Lemma 2.2:

$$\begin{aligned}
\mathrm{err}_{\hat{\mathcal{D}}}(\theta) &= \mathop{\mathbf{E}}_{i\sim[n]}[(\langle \theta, x_i\rangle - y_i)^2] \\
&= \mathop{\mathbf{E}}_{i\sim[n]}[w_i'(\langle \theta, x_i\rangle - y_i)^2] + \mathop{\mathbf{E}}_{i\sim[n]}[(1 - w_i')(\langle \theta, x_i\rangle - y_i)^2] \\
&= \mathop{\mathbf{E}}_{i\sim[n]}[w_i \mathbf{1}_{(x_i,y_i)=(x_i',y_i')}(\langle \theta, x_i'\rangle - y_i')^2] + \mathop{\mathbf{E}}_{i\sim[n]}[(1 - w_i')(\langle \theta, x_i\rangle - y_i)^2]
\end{aligned}$$

and since we have the constraint $w_i x_i' = w_i a_i$ and $w_i y_i' = w_i y_i$,

$$\begin{aligned}
&= \mathop{\mathbf{E}}_{i\sim[n]}[w_i'(\langle \theta, a_i\rangle - b_i)^2] + \mathop{\mathbf{E}}_{i\sim[n]}[(1 - w_i')(\langle \theta, x_i\rangle - y_i)^2] \\
&\leqslant \mathop{\mathbf{E}}_{i\sim[n]}[(\langle \theta, a_i\rangle - b_i)^2] + \mathop{\mathbf{E}}_{i\sim[n]}[(1 - w_i')(\langle \theta, x_i\rangle - y_i)^2],
\end{aligned}$$

since $w_i' = (w_i')^2$ and $(\langle \theta, a_i\rangle - b_i)^2$ are squares.

Now, we use the following sum-of-squares version of Hölder's inequality:

Claim 3.2. Let t be a power of 2. Then for indeterminates $u_1, \dots, u_n$ and $f_1, \dots, f_n$,

$$\{u_i^2 = u_i\}_{i\in[n]} \vdash_{O(t)} \left\{\left(\sum_i u_i f_i\right)^t \leqslant \sum_i \left(\sum_i u_i\right)^{t-1}\left(\sum_i f_i^t\right)\right\}$$

Proof. We'll prove the stronger claim,

$$\{u_i^2 = u_i\}_{i\in[n]} \vdash_{O(t)} \left\{\left(\sum_i u_i f_i\right)^t \leqslant \sum_i \left(\sum_i u_i\right)^{t-1}\left(\sum_i u_i f_i^t\right)\right\},$$

by induction, from which the claim follows because the u_i and f_i^t are squares. For the base case, $t = 2$, the statement follows by Cauchy-Schwarz and from the constraints $u_i^2 = u_i$:

$$\left(\sum_i u_i f_i\right)^2 = \left(\sum_i u_i^2 f_i\right)^2 \leqslant \left(\sum_i u_i^2\right)\left(\sum_i u_i^2 f_i^2\right) = \left(\sum_i u_i\right)\left(\sum_i u_i f_i^2\right).$$

Now, assume the statement holds for t, and we will prove it for $2t$. By the induction hypothesis, there is a degree-$O(t)$ proof that

$$\left(\sum_i u_i f_i\right)^t \leqslant \left(\sum_i u_i\right)^{t-1}\left(\sum_i u_i f_i^t\right).$$

Since both sides are non-negative, the inequality holds when we square both sides (which multiplies the degree by 2). From this we have

$$\begin{aligned}\left(\sum_i u_i f_i\right)^{2t} \leqslant \left(\sum_i u_i\right)^{2t-2}\left(\sum_i u_i f_i^t\right)^2 &= \left(\sum_i u_i\right)^{2t-2}\left(\sum_i u_i^2 f_i^t\right)^2 \\ &\leqslant \left(\sum_i u_i\right)^{2t-2}\left(\sum_i u_i^2\right)\left(\sum_i u_i^2 f_i^{2t}\right) = \left(\sum_i u_i\right)^{2t-1}\left(\sum_i u_i f_i^{2t}\right),\end{aligned}$$

where we have used the Booleanity constraints $u_i^2 = u_i$ and Cauchy-Scwarz, increasing the degree by $O(1)$. This concludes the proof. □

Applying this SoS-Hölder's inequality to our expression with $u_i = (1 - w_i')$,

$$\begin{aligned}\left(\mathop{\mathbf{E}}_{i\sim[n]}[(1-w_i')(\langle\theta, x_i\rangle - y_i)^2]\right)^{k/2} &\leqslant \left(\mathop{\mathbf{E}}_{i\sim[n]}(1-w_i')\right)^{(k-2)/2}\left(\mathop{\mathbf{E}}_{i\sim[n]}(\langle\theta, x_i\rangle - y_i)^k\right) \\ &\leqslant (2\varepsilon)^{(k-2)/2}\left(\mathop{\mathbf{E}}_{i\sim[n]}(\langle\theta, x_i\rangle - y_i)^k\right)\end{aligned}$$

since $\sum_i 1 - w_i' \leqslant 2\varepsilon$ by the fact that only εn of the $\mathbf{1}_{(x_i,y_i)\neq(x_i',y_i')}$ are nonzero, and because $\sum_i w_i = (1-\varepsilon)n$. Now, we can introduce the minimizer $\theta^* = \arg\min_\theta \mathrm{err}_{\hat{D}}(\theta)$ as in the proof of Lemma 2.2,

$$= (2\varepsilon)^{(k-2)/2}\left(\mathop{\mathbf{E}}_{i\sim[n]}(\langle\theta-\theta^*, x_i\rangle + \langle\theta^*, x_i\rangle - y_i)^k\right)$$

And using the SoS inequality $(a+b)^k \leqslant 2^{k-1}(a^k + b^k)$ for k a power of 2,

$$\leqslant (2\varepsilon)^{(k-2)/2}2^{k-1}\left(\mathop{\mathbf{E}}_{i\sim[n]}\langle\theta-\theta^*, x_i\rangle^k + (\langle\theta^*, x_i\rangle - y_i)^k\right).$$

The quantities $\mathbf{E}\langle\theta-\theta^*, x_i\rangle^k$ and $\mathbf{E}(\langle\theta^*, x_i\rangle - y_i)^k$ are bounded very much as in Lemma 2.2, except that we must appeal to the SoS-certifiable hypercontractivity. Given this, we have

$$\begin{aligned}&\leqslant (2\varepsilon)^{(k-2)/2}2^{k-1}C_k^{k/2}\left(\left(\mathop{\mathbf{E}}_{i\sim[n]}\langle\theta-\theta^*, x_i\rangle^2\right)^{k/2} + \left(\mathop{\mathbf{E}}_{i\sim[n]}(\langle\theta^*, x_i\rangle - y_i)^2\right)^{k/2}\right) \\ &= (2\varepsilon)^{(k-2)/2}2^{k-1}C_k^{k/2}\left(\left(\mathop{\mathbf{E}}_{i\sim[n]}(((\langle\theta, x_i\rangle - y_i) + (y_i - \langle\theta^*, x_i\rangle))^2\right)^{k/2}\right. \\ &\qquad\left. + \left(\mathop{\mathbf{E}}_{i\sim[n]}(\langle\theta^*, x_i\rangle - y_i)^2\right)^{k/2}\right) \\ &\leqslant (2\varepsilon)^{(k-2)/2}2^{k-1}C_k^{k/2}\left(\left(2\mathop{\mathbf{E}}_{i\sim[n]}(\langle\theta, x_i\rangle - y_i)^2 + 2\mathop{\mathbf{E}}_{i\sim[n]}(y_i - \langle\theta^*, x_i\rangle)^2\right)^{k/2}\right.\end{aligned}$$

$$
\begin{aligned}
&\quad + \left(\mathop{\mathbf{E}}_{i\sim[n]} (\langle \theta^*, x_i \rangle - y_i)^2 \right)^{k/2} \Bigg) \\
&= (2\varepsilon)^{(k-2)/2} 2^{k-1} C_k^{k/2} \left(\left(2 \cdot \mathrm{err}_{\hat{\mathcal{D}}}(\theta) + 2\mathrm{err}_{\hat{\mathcal{D}}}(\theta^*)\right)^{k/2} + \mathrm{err}_{\hat{\mathcal{D}}}(\theta^*)^{k/2} \right) \\
&\leqslant (2\varepsilon)^{(k-2)/2} 2^{k-1} C_k^{k/2} \left(4^{k/2} \cdot \mathrm{err}_{\hat{\mathcal{D}}}(\theta)^{k/2} + (1 + 4^{k/2}) \cdot \mathrm{err}_{\hat{\mathcal{D}}}(\theta^*)^{k/2} \right),
\end{aligned}
$$

where in the last step we have again used that $(a+b)^k \leqslant 2^{k-1}(a^k + b^k)$ is a degree-k sum-of-squares inequality. Putting all of this together, we have a degree-$O(k)$ sum-of-squares proof that

$$
\begin{aligned}
\left(\mathrm{err}_{\hat{\mathcal{D}}}(\theta) - \mathrm{err}_{\mathcal{D}_2}(\theta)\right)^{k/2} &\leqslant (32\varepsilon)^{(k-2)/2} C_k^{k/2} \left(\mathrm{err}_{\hat{\mathcal{D}}}(\theta)^{k/2} + \mathrm{err}_{\hat{\mathcal{D}}}(\theta^*)^{k/2} \right) \\
&= O(C_k \varepsilon^{1-2/k})^{k/2} \cdot \left(\mathrm{err}_{\hat{\mathcal{D}}}(\theta)^{k/2} + \mathrm{err}(\hat{\mathcal{D}})^{k/2} \right)
\end{aligned}
$$

where $\mathcal{D}_2$ denotes the uniform distribution over the variables (a_i, b_i). Finally, we once again apply the inequality $(a+b)^t \leqslant 2^{t-1}(a^t + b^t)$ on the right-hand side, this time with $a+b = \mathrm{err}_{\hat{\mathcal{D}}}(\theta)$, $a = \mathrm{err}_{\hat{\mathcal{D}}}(\theta) - \mathrm{err}_{\mathcal{D}_2}(\theta)$, $b = \mathrm{err}_{\mathcal{D}_2}(\theta)$, and $t = k/2$ to obtain

$$
\leqslant O(C_k \varepsilon^{1-2/k})^{k/2} \left(2^{k/2} \left(\mathrm{err}_{\hat{\mathcal{D}}}(\theta) - \mathrm{err}_{\mathcal{D}_2}(\theta) \right)^{k/2} + 2^{k/2} \mathrm{err}_{\mathcal{D}_2}(\theta)^{k/2} + \mathrm{err}(\hat{\mathcal{D}})^{k/2} \right).
$$

Re-arranging we have the degree-$O(k)$ sum-of-squares inequality

$$
\begin{aligned}
\left(\mathrm{err}_{\hat{\mathcal{D}}}(\theta) - \mathrm{err}_{\mathcal{D}_2}(\theta)\right)^{k/2} &\leqslant \frac{O(C_k \varepsilon^{1-2/k})^{k/2}}{1 - O(C_k \varepsilon^{1-2/k})^{k/2}} \left(\mathrm{err}_{\mathcal{D}_2}(\theta)^{k/2} + \mathrm{err}(\hat{\mathcal{D}})^{k/2} \right) \\
&= O(C_k \varepsilon^{1-2/k})^{k/2} \left(\mathrm{err}_{\mathcal{D}_2}(\theta)^{k/2} + \mathrm{err}(\hat{\mathcal{D}})^{k/2} \right)
\end{aligned}
$$

Applying the pseudoexpectation on both sides as well as the SoS version of Jensen's inequality ($\widetilde{\mathbf{E}}[(a+b)^k] \geqslant (\widetilde{\mathbf{E}}[a+b])^k$ for k a power of 2), from this we get,

$$
\left(\widetilde{\mathbf{E}}[\mathrm{err}_{\hat{\mathcal{D}}}(\theta)] - \widetilde{\mathbf{E}}[\mathrm{err}_{\mathcal{D}_2}(\theta)]\right)^{k/2} \leqslant \widetilde{\mathbf{E}}[(\mathrm{err}_{\hat{\mathcal{D}}}(\theta) - \mathrm{err}_{\mathcal{D}_2}(\theta))^{k/2}] \leqslant O(C_k \varepsilon^{1-2/k})^{k/2} \cdot \left(\widetilde{\mathbf{E}}[\mathrm{err}_{\mathcal{D}_2}(\theta)^{k/2}] + \mathrm{err}(\hat{\mathcal{D}})^{k/2} \right).
$$

Now, $\widetilde{\mathbf{E}}[\mathrm{err}_{\mathcal{D}_2}(\theta)^{k/2}]$ is the quantity we were minimizing, and in particular it is at most $\mathrm{err}(\hat{\mathcal{D}})^{k/2}$ because $\hat{\mathcal{D}}$ and θ^* were feasible solutions to our polynomial optimization problem. Similarly, $\widetilde{\mathbf{E}}[\mathrm{err}_{\mathcal{D}_2}(\theta)] \leqslant \widetilde{\mathbf{E}}[\mathrm{err}_{\mathcal{D}_2}(\theta)^{k/2}]^{2/k} \leqslant \mathrm{err}_{\hat{\mathcal{D}}}(\theta^*) = \mathrm{err}(\hat{\mathcal{D}})$. Hence, we have that

$$
\mathrm{err}_{\hat{\mathcal{D}}}(\widetilde{\mathbf{E}}[\theta]) \leqslant \widetilde{\mathbf{E}}[\mathrm{err}_{\hat{\mathcal{D}}}(\theta)] \leqslant (1 + O(C_k \varepsilon^{1-2/k})) \cdot \mathrm{err}(\hat{\mathcal{D}}),
$$

concluding the proof. □

4 Conclusion

Bibliographic remarks. This lecture is based on the work of Klivans, Kothari, and Meka, who gave algorithms for robust linear regression with both mean squared loss and ℓ_1 loss in [KKM18]. Also of interest is the follow-up work of Bakshi and Prasad [BP21], who obtain optimal statistical rates for regression under the stronger assumption that there is some negative correlation with the noise; their proof differs from the above in that they exploit the condition that the minimizing θ^* satisfies $\nabla_\theta \mathrm{err}_{\mathcal{D}}(\theta^*) = 0$. The condition of SoS-certifiable hypercontractivity is studied extensively in [KS17].

Contact. Comments are welcome at tselil@stanford.edu.

References

[BP21] Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021. 7

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018. 7

[KS17] Pravesh K Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *arXiv preprint arXiv:1711.07465*, 2017. 4, 7