[Lecture 12]

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:* $\mathbb{E}_{x \sim p}[x z^2 x^\top] \preceq \sigma^2 \mathbb{E}_{x \sim p}[x x^\top]$.

As for mean estimation under bounded covariance, our strategy will be to write down a non-convex optimization problem that tries to find small κ and σ, then show that this problem can be approximately solved. Specifically, let

$$F_1(q) = \sup_v \frac{\mathbb{E}_{x \sim q}[\langle x, v \rangle^4]}{\mathbb{E}_{x \sim q}[\langle x, v \rangle^2]^2}, \text{ and} \tag{1}$$

$$F_2(q) = \sup_v \frac{\mathbb{E}_{x \sim q}[\langle x, v \rangle^2 (y - \langle \theta(q), x \rangle)^2]}{\mathbb{E}_{x \sim q}[\langle x, v \rangle^2]}. \tag{2}$$

Then we seek to find a q such that $F_1(q) \le \kappa$, $F_2(q) \le \sigma^2$, and $q \in \Delta_{n,\epsilon}$, where $\Delta_{n,\epsilon}$ is the set of ϵ-deletions of p.

However, there are a couple wrinkles. While with mean estimation we could minimize the objective with gradient descent, in this case we will need to use *quasigradient* descent–following a direction that is not the gradient, but that we can show makes progress towards the optimum. The rough reason for this is that, since p appears on both the left- and right-hand sides of the inequalities above, the gradients become quite messy, e.g. $\nabla F_1(q)$ has a mix of positive and negative terms:

$$\nabla F_1(q)_i = \frac{\langle x_i, v \rangle^4}{\mathbb{E}_{x \sim q}[\langle x, v \rangle^2]^2} - 2\frac{\mathbb{E}_q[\langle x, v \rangle^4] \langle x_i, v \rangle^2}{\mathbb{E}_q[\langle x, v \rangle^2]^3}, \tag{3}$$

and it isn't clear that following them will not land us at bad local minima. To address this, we instead construct a simpler *quasigradient* for F_1 and F_2:

$$g_1(x_i; q) = \langle x_i, v \rangle^4, \qquad \text{where } v = \arg\max_{\|v\|_2 = 1} \frac{\mathbb{E}_q[\langle x, v \rangle^4]}{\mathbb{E}_q[\langle x, v \rangle^2]^2}, \tag{4}$$

$$g_2(x_i; q) = \langle x_i, v \rangle^2 (y_i - \langle \theta^*(q), x_i \rangle)^2, \qquad \text{where } v = \arg\max_{\|v\|_2 = 1} \frac{\mathbb{E}_q[\langle x, v \rangle^2 (y - \langle \theta^*(q), x \rangle)^2]}{\mathbb{E}_q[\langle x, v \rangle^2]}. \tag{5}$$

We will then follow g_1 until F_1 is small, and then follow g_2 until F_2 is small.

The other wrinkle is that computationally, the hypercontractivity condition is difficult to certify, because it involves maximizing $\frac{\mathbb{E}_p[\langle x, v \rangle^4]}{\mathbb{E}_p[\langle x, v \rangle^2]^2}$, which is no longer a straightforward eigenvalue problem as in the mean estimation case. We've already seen this sort of difficulty before–for norms beyond the ℓ_2-norm, we had to use SDP relaxations and Grothendieck's inequality in order to get a constant factor relaxation. Here, there is also an SDP relaxation called the *sum-of-squares* relaxation, but it doesn't always give a constant factor relaxation. We'll mostly ignore this issue and assume that we can find the maximizing v for hypercontractivity.

We are now ready to define our efficient algorithm for linear regression, Algorithm 1. It is closely analogous to the algorithm for mean estimation (Algorithm **??**), but specifies the gradient steps more explicitly.

Analyzing Algorithm 1 enjoys the following loss bound:

Proposition 0.1. *Suppose that a set S of $(1 - \epsilon)n$ of the x_i satisfy:*

$$\mathbb{E}_{p_S}[\langle x, v \rangle^4] \le \kappa \mathbb{E}_{p_S}[\langle x, v \rangle^2]^2 \; \forall v \in \mathbb{R}^d, \text{ and } \mathbb{E}_{p_S}[(y - \langle \theta^*(p_S), x \rangle)^2 x x^\top] \preceq \sigma^2 \mathbb{E}_{p_S}[x x^\top]. \tag{6}$$

Then assuming $\kappa\epsilon \le \frac{1}{80}$, Algorithm 1 terminates and its output has excess loss $L(p_S, \theta^*(q)) \le 40\sigma^2\epsilon$.

Algorithm 1 `QuasigradientDescentLinReg`

```
Input: $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$.
Initialize $q \in \Delta_{n,\epsilon}$ arbitrarily.
while $F_1(q) \geq 2\kappa$ or $F_2(q) \geq 4\sigma^2$ do
    if $F_1(q) \geq 2\kappa$ then
        Find the unit vector $v$ that maximizes $\mathbb{E}_q[\langle x, v\rangle^4]/\mathbb{E}_q[\langle x, v\rangle^2]^2$.
        Take a projected gradient step in the direction $(g_1)_i = \langle x_i, v\rangle^4$.
    else
        Compute the empirical least squares regressor: $\theta^*(q) = (\sum_{i=1}^n q_i x_i x_i^\top)^{-1}(\sum_{i=1}^n q_i x_i y_i)$.
        Find the unit vector $v$ that maximizes $\mathbb{E}_q[\langle x, v\rangle^2(y - \langle \theta^*(q) x\rangle)^2]/\mathbb{E}_q[\langle x, v\rangle^2]$.
        Take a projected gradient step in the direction $(g_2)_i = \langle x_i, v\rangle^2(y_i - \langle \theta^*(q), x_i\rangle)^2$.
    end if
end while
Output $\theta^*(q)$.
```

To prove Proposition 0.1, we need a few ideas. The first is a result from optimization justifying the use of the quasigradients:

Lemma 0.2 (Informal)**.** *Asymptotically, the iterates of Algorithm 1 (or any other low-regret algorithm) satisfy the conditions*

$$\mathbb{E}_{X\sim q}[g_j(X; q)] \leq \mathbb{E}_{X\sim p_S}[g_j(X; q)] \text{ for } j = 1, 2. \tag{7}$$

We will establish this more formally later, and for now assume that (7) holds. Then, asuming this, we will show that q is both hypercontractive and has bounded noise with constants κ', σ' that are only a constant factor worse than κ and σ.

First, we will show this for hypercontractivity:

Lemma 0.3. *Suppose that $\mathbb{E}_{X\sim q}[g_1(X; q)] \leq \mathbb{E}_{X\sim p_S}[g_1(X; q)]$ and that $\kappa\epsilon \leq \frac{1}{80}$. Then q is hypercontractive with parameter $\kappa' = 1.5\kappa$.*

Proof. Take the maximizing v such that $g_1(x_i; q) = \langle x_i, v\rangle^4$. To establish hypercontractivity, we want to show that $\mathbb{E}_q[\langle x, v\rangle^4]$ is small while $\mathbb{E}_q[\langle x, v\rangle^2]^2$ is large. Note the quasigradient condition gives us that $\mathbb{E}_q[\langle x, v\rangle^4] \leq \mathbb{E}_{p_S}[\langle x, v\rangle^4]$; so we will mainly focus on showing that $\mathbb{E}_q[\langle x, v\rangle^2]$ is large, and in fact not much smaller than $\mathbb{E}_{p_S}[\langle x, v\rangle^2]$. Specifically, by the fact that $\mathsf{TV}(q, p_S) \leq \frac{\epsilon}{1-\epsilon}$, together with resilience applied to $\langle x, v\rangle^2$, we have

$$\begin{aligned}
|\mathbb{E}_q[\langle x, v\rangle^2] - \mathbb{E}_{p_S}[\langle x, v\rangle^2]| &\leq \Big(\frac{\epsilon}{(1-2\epsilon)^2}(\mathbb{E}_q[\langle x, v\rangle^4] + \mathbb{E}_{p_S}[\langle x, v\rangle^4])\Big)^{\frac{1}{2}} &\quad (8)\\
&\overset{(i)}{\leq} \Big(\frac{2\epsilon}{(1-2\epsilon)^2}\mathbb{E}_{p_S}[\langle x, v\rangle^4]\Big)^{\frac{1}{2}} &\quad (9)\\
&\overset{(ii)}{\leq} \Big(\frac{2\kappa\epsilon}{(1-2\epsilon)^2}\Big)^{\frac{1}{2}}\mathbb{E}_{p_S}[\langle x, v\rangle^2], &\quad (10)
\end{aligned}$$

where (i) uses the quasigradient condition and (ii) uses hypercontractivity for p_S. Now assuming that $\kappa\epsilon \leq \frac{1}{80}$ (and hence also $\epsilon \leq \frac{1}{80}$), the coefficient on the right-hand-side is at most $\sqrt{(1/40)/(1-1/40)^2} < \frac{1}{6}$. Consequently $|\mathbb{E}_q[\langle x, v\rangle^2] - \mathbb{E}_{p_S}[\langle x, v\rangle]^2| \leq \frac{1}{6}\mathbb{E}_{p_S}[\langle x, v\rangle^2]$, and re-arranging then yields

$$\mathbb{E}_q[\langle x, v\rangle^2] \geq \frac{5}{6}\mathbb{E}_{p_S}[\langle x, v\rangle^2]. \tag{11}$$

But we already have $\mathbb{E}_q[\langle x, v\rangle^4] \leq \mathbb{E}_{p_S}[\langle x, v\rangle^2]$, and so the ratio $\mathbb{E}_q[\langle x, v\rangle^2]/\mathbb{E}_q[\langle x, v\rangle^2]^2$ is at most $(6/5)^2$ the same ratio under p_S, and in particular at most $(6/5)^2\kappa \leq 1.5\kappa$. □

Next, we will show this for bounded noise assuming that hypercontractivity holds:

Lemma 0.4. *Suppose that $F_1(q) \leq 2\kappa$ and that $\mathbb{E}_{X\sim q}[g_2(X;q)] \leq \mathbb{E}_{X\sim p_S}[g_2(X;q)]$, and that $\kappa\epsilon \leq \frac{1}{80}$. Then q has bounded noise with parameter $(\sigma')^2 = 4\sigma^2$, and furthermore satisfies $L(p_S, \theta^*(q)) \leq 40\sigma^2\epsilon$.*

Proof. Again take the maximizing v such that $g_2(x_i;q) = \langle x_i, v\rangle^2(y_i - \langle\theta^*(q), x_i\rangle)^2$. We want to show that q has bounded noise, or in other words that $\mathbb{E}_q[g_2(x;q)]$ is small relative to $\mathbb{E}_q[\langle x, v\rangle^2]$. By the quasigradient assumption, we have

$$\begin{aligned}&\mathbb{E}_q[g_2(x;q)] + \mathbb{E}_q[\langle x, v\rangle^2(y - \langle\theta^*(q), x\rangle)^2] && (12)\\ &\quad\leq \mathbb{E}_{p_S}[\langle x, v\rangle^2(y - \langle\theta^*(q), x\rangle)^2]. && (13)\end{aligned}$$

Intuitively, we want to use the bounded noise condition for p_S to upper-bound the right-hand-side. The problem is that the term inside the expectation contains $\theta^*(q)$, rather than $\theta^*(p_S)$. But we can handle this using the AM-RMS inequality. Specifically, we have

$$\mathbb{E}_{p_S}[\langle x, v\rangle^2(y - \langle\theta^*(q), x\rangle)^2] \leq 2\big(\underbrace{\mathbb{E}_{p_S}[\langle x, v\rangle^2(y - \langle\theta^*(p_S), x\rangle)^2]}_{(a)} + \underbrace{\mathbb{E}_{p_S}[\langle x, v\rangle^2\langle\theta^*(q) - \theta^*(p_S), x\rangle^2]}_{(b)}\big). \quad (14)$$

We will bound (a) and (b) in turn. To bound (a) note that by the bounded noise condition we simply have $(a) \leq \sigma^2\mathbb{E}_{p_S}[\langle x, v\rangle^2]$.

To bound (b), let $R = \mathbb{E}_{p_S}[\langle\theta^*(q) - \theta^*(p_S), x\rangle^2] = L(p_S, \theta^*(q))$ be the excess loss of $\theta^*(q)$ under p_S. We will upper-bound (b) in terms of R, and then apply resilience to get a bound for R in terms of itself. Solving the resulting inequality will provide an absolute bound on R and hence also on (b).

More specifically, we have

$$\begin{aligned}\mathbb{E}_{p_S}[\langle x, v\rangle^2\langle\theta^*(q) - \theta^*(p_S), x\rangle^2] &\overset{(i)}{\leq} \Big(\mathbb{E}_{p_S}[\langle x, v\rangle^4\Big)^{1/4}\Big(\mathbb{E}_{p_S}[\langle\theta^*(q) - \theta^*(p_S), x\rangle^4]\Big)^{1/2} && (15)\\ &\overset{(ii)}{\leq} \kappa\Big(\mathbb{E}_{p_S}[\langle x, v\rangle^2]\Big)\Big(\mathbb{E}_{p_S}[\langle\theta^*(q) - \theta^*(p_S), x\rangle^2\Big) && (16)\\ &= \kappa R\mathbb{E}_{p_S}[\langle x, v\rangle^2]. && (17)\end{aligned}$$

Here (i) is Cauchy-Schwarz and (ii) invokes hypercontractivity of p_S. Combining the bounds on (a) and (b) and plugging back in to (14), we obtain

$$\mathbb{E}_{p_S}[\langle x, v\rangle^2(y - \langle\theta^*(q), x\rangle)^2] \leq 2(\sigma^2 + \kappa R)\mathbb{E}_{p_S}[\langle x, v\rangle^2]. \quad (18)$$

Remember that we would like a bound such as the above, but with expectations taken with respect to q instead of p_S. For the left-hand-side, we can directly move to $\mathbb{E}_q[\cdot]$ using the quasigradient assumption. For the right-hand-side, since $F_1(q) \leq 2\kappa$, the same argument as in (8)-(11) yields (with modified constants) that $\mathbb{E}_q[\langle x, v\rangle^2] \geq \frac{4}{5}\mathbb{E}_{p_S}[\langle x, v\rangle^2]$. Applying both of these, we have that

$$\mathbb{E}_q[\langle x, v\rangle^2(y - \langle\theta^*(q), x\rangle)^2] \leq 2.5(\sigma^2 + \kappa R)\mathbb{E}_q[\langle x, v\rangle^2]. \quad (19)$$

This establishes bounded noise with parameter $(\sigma')^2 = 2.5(\sigma^2 + \kappa R)$. By assumption, we also have hypercontractivity with parameter $\kappa' = 2\kappa$. We are not yet done, because we do not know R. But recall that R is the excess loss, and so by the resilience conditions for linear regression (Proposition **??**) we have

$$R \leq 5\rho(\kappa', \sigma') \leq 10(\sigma')^2\epsilon = 25(\sigma^2 + \kappa R)\epsilon, \quad (20)$$

as long as $\epsilon \leq \frac{1}{8}$ and $\kappa'\epsilon = 2\kappa\epsilon \leq \frac{1}{6}$. Re-arranging, we have

$$R \leq \frac{25\sigma^2\epsilon}{1 - 25\kappa\epsilon} \leq 40\sigma^2\epsilon, \quad (21)$$

since $\kappa\epsilon \leq \frac{1}{80}$. Plugging back into σ', we also have $(\sigma')^2 \leq 2.5\sigma^2(1 + 40\kappa\epsilon) \leq 4\sigma^2$, as claimed. □

Quasigradient bounds via low-regret algorithms. Combining Lemmas 0.2, 0.3, and 0.4 together yields Proposition 0.1. However, we still need to formalize Lemma 0.2, showing that we can drive the quasigradients to be small. We can do so with the idea of low-regret *online optimization algorithms.*

An online optimization algorithm is one that takes a sequence of losses $\ell_1(\cdot), \ell_2(\cdot)$ rather than a fixed loss ℓ. In traditional optimization, we have a single ℓ with parameter w, and seek to produce iterates $w_1, w_2, \dots$ such that $\ell(w_T) - \ell(w^*) \to 0$ as $T \to \infty$. In online optimization algorithms, we instead consider the regret, defined as

$$\mathrm{Regret}_T = \max_w \frac{1}{T} \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(w). \tag{22}$$

This is the average excess loss compared to the best fixed w, picked in hindsight. We then seek to produce iterates w_t such that $\mathrm{Regret}_T \to 0$ as $T \to \infty$. Note that when $\ell_t = \ell$ is fixed for all t, this is exactly the same as traditional optimization.

Remarkably, for most "nice" loss functions ℓ_t, it is possible to ensure that $\mathrm{Regret}_T \to 0$; in fact, projected gradient descent with an appropriate step size will achieve this.

How does this relate to quasigradients? For any quasigradient $g(X; q)$, define the loss function $\ell_t(q) = \mathbb{E}_{x\sim q}[g(x; q_t)]$. Even though this loss depends on q_t, the regret is well-defined:

$$\mathrm{Regret}_T = \max_{q'} \frac{1}{T} \mathbb{E}_{x\sim q_t}[g(x; q_t)] - \mathbb{E}_{x\sim q'}[g(x; q_t)] \geq \frac{1}{T} \mathbb{E}_{x\sim q_t}[g(x; q_t)] - \mathbb{E}_{x\sim p_S}[g(x; q_t)]. \tag{23}$$

In particular, as long as $\mathrm{Regret}_T \to 0$, we asymptotically have that $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x\sim q_t}[g(x; q_t)] \leq \sum_{t=1}^{T} \mathbb{E}_{x\sim p_S}[g(x; q_t)]$, so eventually the quasigradient bound $\mathbb{E}_{x\sim q}[g(x; q)] \leq \mathbb{E}_{x\sim p_S}[g(x; q)]$ must (almost) hold for one of the q_t.

This is enough to show that, for any fixed quasigradient, a statement such as Lemma 0.2 holds[1]. However, we want *both* g_1 and g_2 to be small simultaneously.

There are two ways to handle this. The first, crude way is to first use g_1 until we have hypercontractivity, then take the resulting q as our new $\tilde{p}$ (so that we restrict to ϵ-deletions of q) and running gradient descent with g_2 until we have bounded noise. This uses the fact that ϵ-deletions of hypercontractive distributions are still hypercontractive, but yields worse constants (since we need 2ϵ-deletions instead of ϵ-deletions).

A slicker approach is to alternate between g_1 and g_2 as in Algorithm 1. Note that we then still (asymptotically) have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{x\sim q_t}[g_{j_t}(x; q_t)] \leq \sum_{t=1}^{T} \mathbb{E}_{x\sim p_S}[g_{j_t}(x; q_t)], \tag{24}$$

where $j_t \in \{1, 2\}$ denotes the choice of quasigradient at iteration t of Algorithm 1.

Next note that asymptotically, only a vanishingly small fraction of j_t can equal 1, since we only take quasigradient steps in g_1 when $F_1(q) \geq 2\kappa$, and in these cases $\mathbb{E}_{q_t}[g_1(x; q_t)]$ is quite a bit larger than $\mathbb{E}_{p_S}[g_1(x; q_t)]$, since if they were equal we would have $F_1(q) \leq 1.5\kappa$. Therefore, eventually almost all of the quasigradient steps are with respect to g_2, and so low regret of the entire sum implies low regret of g_2. We therefore both have $F_1(q_t) \leq 2\kappa$ and the quasigradient condition for g_2:

Lemma 0.5 (Formal version of Lemma 0.2)**.** *Suppose that $|g_1(x_i, q)| \leq B$ and $|g_2(x_i, q)| \leq B$ for all i, where B is at most polynomially-large in the problem parameters. Then for any δ, within polynomially many steps Algorithm 1 generates an iterate q_t such that $F_1(q_t) \leq 2\kappa$ and $\mathbb{E}_{g_t}[g_2(x, q_t)] \leq \mathbb{E}_{p_S}[g_2(x, q_t)] + \delta$.*

Combining Lemma 0.2 with Lemma 0.4 then yields the desired Proposition 0.1.

[1] We have to be a bit careful because outliers could make $g(x; q)$ arbitrarily large, which violates the assumptions needed to achieve low regret. This can be addressed with a pre-filtering step that removes data points that are obviously too large to be inliers, but we will not worry about this here.