8.1 Introduction

A quick recap of last lecture: we analyzed a particular optimization algorithm—the method of projected gradient descent (PGD)—as applied to differentiable cost function $L : \mathbb{R}^d \to \mathbb{R}$. For any given step size $\eta > 0$, the PGD algorithm generates a sequence $\{\theta^t\}_{t=0}^{\infty}$ via the recursion

$$
\theta^{t+1} = \arg \min_{\theta \in C} \left\{ L(\theta^t) + \langle \nabla L(\theta^t), \theta \rangle + \frac{1}{2\eta} \| \theta - \theta^t \|^2_2 \right\}.
$$

(8.1)

Let $G_n : C \to C$ denote the PGD updates at the $n$-sample level—that is, as applied to the objective function $L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta; Z_i)$. Similarly, let $\overline{G}$ denote the projected gradient operator as applied at the population level (to the function $\overline{L} = E L_n(\theta)$). Our goal is to bound the error $\| \theta^{t+1} - \theta^* \|_2$.

Last lecture, we showed that if the population loss $\overline{L} : C \to \mathbb{R}$ is $(\gamma, \mu)$-strongly convex, smooth, then the population operator $\overline{G}$ is contractive with parameter $\kappa = \sqrt{1 - \frac{\gamma}{2\mu}}$. Our goal in this lecture is to explore when the sample operator $G_n$ inherits this type of property. In abstract terms, we can view $G_n$ as a perturbed version of $\overline{G}$; if the two operators are “suitably close”, then we expect that updates with $G_n$ will yield an approximate type of contraction. The following condition formalizes this notion:

$(\alpha_n, \epsilon_n)$ perturbation condition: Suppose that for any pair $\Delta, \tilde{\Delta} \in K(\theta^*)$,

$$
|\langle \Delta, G_n(\theta^* + \Delta) - \overline{G}(\theta^* + \Delta) \rangle| \leq \alpha_n \| \Delta \|_2 \| \Delta \|_2 + \epsilon_n \| \Delta \|_2.
$$

(8.2)

Under this condition, Theorem 7.3.1 in last lecture shows that the sequence $\theta^{t+1} = G_n(\theta^t)$ satisfies the bound

$$
\| \theta^{t+1} - \theta^* \|_2 \leq \tilde{\kappa}^{t+1} \| \theta^0 - \theta^* \|_2 + \frac{\epsilon_n}{1 - \tilde{\kappa}} \text{ error floor}
$$

where $\tilde{\kappa} = \kappa + \alpha_n$.

(8.3)

As we have mentioned in last lecture, the perturbation condition (8.2) measures how close the sample version is to the population version. So the challenge is to find the constant $\alpha_n$ and $\epsilon_n$, and doing so requires techniques from empirical process theory and concentration of measure. In today’s lecture, we study some concrete examples that involve proving that the condition (8.2) holds with high probability.
### 8.2 Ordinary least square

Suppose that we observe pairs $Z_i = (y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$ that are linked via the standard linear model

$$y_i = \langle x_i, \theta^* \rangle + w_i,$$

where $x_i \sim N(0, \Sigma)$ and $w_i \sim N(0, \sigma^2)$ are independent.

The finite-sample problem is to minimize the cost function

$$L_n(\theta) := \frac{1}{2n} \|y - X\theta\|_2^2 = \frac{1}{2n} \sum_{i=1}^n (y_i - \langle x_i, \theta \rangle)^2.$$

Under the assumption that $x_i \sim N(0, \Sigma)$, we have

$$\bar{L}(\theta) = \mathbb{E}_{y, X} L_n(\theta) = \frac{1}{2} \{(\theta - \theta^*)^T \Sigma(\theta - \theta^*) + \sigma^2\}.$$

For this quadratic objective, the strong convexity/smoothness conditions hold with $\gamma_\mu = \lambda_{\text{max}}(\Sigma)$ and $\gamma_\ell = \lambda_{\text{min}}(\Sigma)$.

For simplicity, let’s consider the unconstrained case ($\mathcal{C} = \mathbb{R}^d$), so that PGD with stepsize $1 / \gamma_\mu$ reduces to the usual gradient descent update. In particular, we have

$$G_n(\theta) = \theta - \frac{1}{\gamma_\mu} \nabla L_n(\theta), \quad \text{and} \quad \overline{G}(\theta) = \theta - \frac{1}{\gamma_\mu} \nabla \bar{L}(\theta).$$

The following result shows that the operator $G_n$ inherits the good behavior of $\overline{G}$ whenever $n \gtrsim d$:

**Proposition 8.2.1.** For the unconstrained least-squares problem with Gaussian design/noise and sample size $n \gtrsim d$, there are universal constants $(c_0, c_1, c_2)$ such that (w.h.p.) the bound (8.3) holds with

$$\alpha_n = c_0 \sqrt{\frac{d}{n}}, \quad \text{and} \quad \epsilon_n^2 = c_1 \frac{\sigma^2 d}{\lambda_{\text{max}}(\Sigma)} n. \quad (8.4)$$

As a consequence, performing $T \approx \log(n)$ steps of gradient descent yields a solution $\theta^T$ such that (w.h.p.)

$$\|\theta^T - \theta^*\|_2^2 \leq c_2 \frac{\sigma^2 d}{\lambda_{\text{min}}(\Sigma)} n. \quad (8.5)$$

It is worthwhile comparing the guarantee (8.5) to the best possible guarantee that one could achieve. It is easy to show that (as long as $X^TX$ is full rank), the error in the ordinary least-squares estimate $\hat{\theta}$ is given by

$$\hat{\theta} - \theta^* = (X^TX/n)^{-1}X^Tw.$$ Consequently, letting $\hat{\Sigma} = X^TX/n$ denote the sample covariance matrix, we have

$$\mathbb{E}[\|\hat{\theta} - \theta^*\|_2^2] = \mathbb{E}[w^T(X^TX/n)^{-1}w] = \frac{\sigma^2}{n} \mathbb{E}[\text{trace}(\hat{\Sigma}^{-1})] \leq \frac{\sigma^2 d}{\lambda_{\text{min}}(\Sigma) n}.$$

As our proof will show, when $n \gtrsim d$, then $\lambda_{\text{min}}(\hat{\Sigma}) \approx \lambda_{\text{min}}(\Sigma)$, so that this bound matches our guarantee (8.5) up to constants.

**Proof.** Let us first show how the bound (8.5) follows from the guarantee (8.3) with the stated choices (8.4) of $(\alpha_n, \epsilon_n)$. In particular, based on the form of the bound (8.3), it suffices to show that

$$\frac{\epsilon_n^2}{(1 - \tilde{\kappa})^2} \leq c_3 \frac{\sigma^2 d}{\lambda_{\text{min}}(\Sigma) n}.$$
Recall that $\bar{\kappa} = \kappa + \alpha_n$ where $\kappa^2 = 1 - \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}$ (with Nestrov’s acceleration). Therefore, we have

$$1 - \bar{\kappa} = -c_0 \sqrt{\frac{d}{n}} + 1 - \sqrt{1 - \frac{\lambda_{\min}}{\lambda_{\max}}} \geq \frac{1}{2} \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} - c_0 \sqrt{\frac{d}{n}} \geq 4 \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}$$

where in (i), we use $1 - \sqrt{1 - x} \geq x/(1 + \sqrt{1 - x}) \geq x/2$ and in (ii), we assume $\sqrt{\frac{d}{n}} \leq \frac{1}{4} \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}$. As a consequence, the error floor satisfies

$$\epsilon_n \leq c \sqrt{\frac{d \sigma^2}{n \lambda_{\max} \lambda_{\min}}} = c \sqrt{\frac{d \sigma^2}{n \lambda_{\min}}}, \quad (8.6)$$

as claimed.

Let us now prove the claim (8.4). By a direct calculation, we find that the left hand side of condition (8.2) is given by

$$\left| \langle \Delta, \frac{1}{\gamma_\mu} (\nabla L_n(\theta + \Delta) - \nabla \bar{L}(\theta + \Delta)) \rangle \right| = \left| \langle \Delta, \frac{1}{\gamma_\mu} \left( \frac{X^T X}{n} \Delta - \frac{X^T w}{n} - \Sigma \Delta \right) \rangle \right|$$

$$\leq \frac{1}{\gamma_\mu} \left| \Delta^T \left( \frac{X^T X}{n} - \Sigma \right) \Delta \right| + \frac{1}{\gamma_\mu} \left| \Delta \right| \left| \frac{X^T w}{n} \right| \left( T_1 \right)$$

$$\leq \frac{1}{\gamma_\mu} \left( \| \Sigma - \frac{X^T X}{n} \|_{op} \| \Delta \|_2 \| \Delta \|_2 \right) \left( T_1 \right) + \frac{1}{\gamma_\mu} \left( \| \Delta \|_2 \right) \left( \frac{X^T w}{n} \right) \left( T_2 \right)$$

where (i) is due to the triangle inequality and (ii) is by definition of the operator norm and Hölder’s inequality.

Let’s bound these two parts above separately. In particular, we claim that

$$T_1 \leq c \sqrt{\frac{d}{n}} \| \Delta \|_2, \quad \text{and} \quad T_2 \leq c \frac{\sigma}{\sqrt{\lambda_{\max}}} \sqrt{\frac{d}{n}}, \quad (8.7a)$$

both with high probability.

To prove the first bound, we recall that standard results in non-asymptotic random matrix theory guarantee that the operator norm $\| \Sigma - \frac{X^T X}{n} \|_{op}$ is bounded as

$$\| \Sigma - \frac{X^T X}{n} \|_{op} \leq c \lambda_{\max}(\Sigma) \sqrt{\frac{d}{n}} \quad \text{w.h.p. when } n \geq d.$$

The claimed bound thus follows after some algebra.

Turning to the second bound, on the other hand, for any fixed $X$, the function $w_{\frac{\sigma}{\sqrt{n}}} \to \| \frac{X^T w}{n} \|_2$ is a $L$-Lipschitz function of standard Gaussian vector $w_{\frac{\sigma}{\sqrt{n}}}$ and $L \leq \frac{\sigma}{\sqrt{n}} \| \frac{X}{\sqrt{n}} \|_{op}$. Thus, by Borell’s concentration for Lipschitz function (see Lecture #2 for details), we have

$$P(\| \frac{X^T w}{n} \|_2 - E \| \frac{X^T w}{n} \|_2 \geq \delta) \leq e^{-\delta^2/2L^2}.$$
The expectation can be bounded as
\[ E \left\| \frac{X^T w}{n} \right\|_2^2 = \frac{1}{n} \mathbb{E} \left( \frac{w^T XX^T}{n} - w \right) \leq \frac{\sigma^2}{n} \mathbb{E} \text{trace} (X^T X) \leq d \lambda_{\text{max}} (\Sigma) \frac{\sigma^2}{n} \]
where in the last inequality, we use a uniform bound for all the eigenvalues of \( \Sigma \). Putting together the pieces, for any fixed \( X \), we have with high probability
\[ \left\| \frac{X^T w}{n} \right\|_2 \leq \sqrt{d \lambda_{\text{max}} (\Sigma) \frac{\sigma^2}{n} + c L}, \quad L \leq \frac{\sigma}{\sqrt{n}} \left\| X \right\|_{\text{op}}. \]

From our earlier results, the operator norm \( \left\| X \sqrt{n} \right\|_{\text{op}} \) is upper bounded by \( 2 \sqrt{\lambda_{\text{max}} (\Sigma)} \) with high probability, and therefore \( L \leq 2 \frac{\sigma}{\sqrt{n}} \sqrt{\lambda_{\text{max}} (\Sigma)} \). Conditioning on this event, we have
\[ \left\| \frac{X^T w}{n} \right\|_2 \leq c \sqrt{d \lambda_{\text{max}} (\Sigma) \frac{\sigma^2}{n}} \quad (8.8) \]
which implies the claimed bound on \( T_2 \).

\[ \square \]

### 8.3 Sparse least squares

Let us now turn to a second example, one in which the constraints enter in an interesting way. Consider again the linear regression model \( y_i = \langle x_i, \theta^* \rangle + w_i \) and \( w_i \sim N(0, \sigma^2) \), but now let suppose that the unknown regression vector \( \theta^* \) is sparse with \( \| \theta^* \|_0 = k \). For simplicity, let’s consider the case where each \( x_i \) comes from a mean-zero Gaussian distribution with \( I_d \) covariance matrix. Then the population function equals
\[ \bar{L}(\theta) = \mathbb{E}_{y,X} L_n(\theta) = \frac{1}{2} \{ \| \theta - \theta^* \|_2^2 + \sigma^2 \}. \]
which implies \( \gamma_\mu = \gamma_\ell = 1 \). So projected gradient descent with unit stepsize on \( \bar{L} \) converges in a single step.

On the other hand, consider the projected gradient updates on the sample function \( L_n \): they are given by the recursion
\[ \theta^{t+1} = \arg \min_{\| \theta \|_1 \leq \| \theta^* \|_1} \left\{ L_n(\theta) + \langle \nabla L_n(\theta^t), \theta \rangle + \frac{1}{2} \| \theta - \theta^t \|_2^2 \right\}. \]
Notice that here, we cheat by setting the constraint parameter equal to \( \| \theta^* \|_1 \). The update can be computed efficiently by a form of soft-thresholding. The following result characterizes its behavior.

**Theorem 8.3.1.** Suppose \( n \geq k \log(\frac{d}{k}) \). Then there are universal constants such that PGD updates with \( \ell_1 \)-constraint satisfy
\[ \| \theta^{t+1} - \theta^* \|_2 \leq c_0 \left( \sqrt{k \log(\frac{d}{k})} \right)^{t+1} \| \theta^0 - \theta^* \|_2 + c_1 \sigma \sqrt{\frac{k \log(\frac{d}{k})}{n}}. \quad (8.9) \]

From this theorem, we can see that the convergence rate depends on \( r := \sqrt{\frac{d \log(\frac{d}{k})}{n}} \), the smaller \( r \) is, the faster convergence rate we will obtain. So the quantity \( r \) measures how hard the optimization problem is.
On the other hand, smaller $r$ also results in a smaller error floor—easier statistics problem. This fact means, when a statistics problem is easier the corresponding optimization should also be easier. We illustrate this phenomenon with Fig 8.3 below, where in three different cases, we plot the convergence behavior versus iteration for parameter $r = 1, 0.5$ or $0.1$.

\[ r = \sqrt{\frac{k \log(\frac{d}{k})}{n}} \]

**Figure 8.1:** Illustration of the convergence behavior versus iteration that is predicted by Theorem 8.3.1

We can compare this result with the minimax rate we have for sparse regression. By known results [2], for this particular ensemble of sparse problems, any estimator $\hat{\theta}$ satisfies the minimax lower bound

\[ \sup_{\|\theta^*\|_0 = k} E\|\hat{\theta} - \theta^*\|_2^2 \geq \sigma^2 k \log(\frac{ed}{k}) n. \] (8.10)

Consequently, the error floor exhibited in Theorem 8.3.1 is unimprovable.

To conclude, let's sketch the proof of this theorem.

**Proof.** (sketch) By Lagrangian duality, the update corresponds to minimizing a function of the form $\langle \nabla \mathcal{L}_n(\theta^t), \theta \rangle + \frac{1}{2} \| \theta - \theta^t \|_2^2 + \lambda_n \| \theta \|_1$ for some $\lambda_n > 0$. Consequently, the updates satisfy the equation

\[ G_n(\theta) - \theta + \nabla \mathcal{L}_n(\theta) + \lambda_n z = 0 \]

where $z$ is the sub-gradient and $\| z \|_\infty \leq 1$. Then the difference is $G_n(\theta) - \nabla \bar{\mathcal{L}}(\theta) = \nabla \mathcal{L}_n(\theta) - \nabla \bar{\mathcal{L}}(\theta) + \lambda_n (z - \bar{z})$ and by some algebra,

\[ |\langle \hat{\Delta}, G_n(\theta^* + \Delta) - \nabla \bar{\mathcal{L}}(\theta^* + \Delta) \rangle| \leq 2\lambda_n \| \Delta \|_1 + |\langle \hat{\Delta}, \nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \bar{\mathcal{L}}(\theta^* + \Delta) \rangle|. \]

Setting the parameter $\lambda_n \leq c\sigma \sqrt{\frac{k \log(\frac{ed}{k})}{n}}$ takes care of the first part, so that it remains to bound the
remaining part. We decompose it into two terms, namely
\[ |⟨ \tilde{\Delta}, \nabla L_n(\theta^* + \Delta) − \nabla \tilde{L}(\theta^* + \Delta)⟩| = |⟨ \tilde{\Delta}, \frac{X^T X}{n} w⟩ + ⟨ \tilde{\Delta}, (\frac{X^T X}{n} - I_d) \Delta⟩| \]
\[ \leq |⟨ \tilde{\Delta}, \frac{X^T X}{n} w⟩| + |⟨ \tilde{\Delta}, (\frac{X^T X}{n} - I_d) \Delta⟩|. \]

If \( \Delta \) and \( \tilde{\Delta} \) are arbitrary, \( T_2 \) can be as large as \( \|X^T X_n - I_d\|_2 \|\Delta\|_2 \) which concentrates on \( \sqrt{d/n} \|\tilde{\Delta}\|_2 \|\Delta\|_2 \) according to the random matrix theory. Therefore, if we want to the estimation error to vanish, we need \( n \gtrsim d \). So it is important to use the fact that \( \Delta = \theta^t - \theta^* \) and \( \tilde{\Delta} = \theta^{t+1} - \theta^* \) lie in the cone
\[ \mathcal{K}(\theta^*) = \{ t(\theta - \theta^*) \mid \|\theta\|_1 \leq \|\theta^*\|_1, t \geq 0 \}. \] (8.11)

There only remains two steps before reaching the final conclusion.

(a) We claim that if \( n \gtrsim k \log(\frac{ed}{k}) \), then with high probability,
\[ \sup_{\Delta, \tilde{\Delta} \in \mathcal{K}(\theta^*) \cap S^{d-1}} |⟨ \tilde{\Delta} (I_d - \frac{X^T X}{n}) \Delta⟩| \leq c \sqrt{\frac{k \log(\frac{ed}{k})}{n}}. \]

This claim is similar to the results we showed in the first few lectures about Gaussian random sketch. We need to argue that the Gaussian sketch preserves inner product approximately.

(b) Also, we claim with high probability,
\[ \sup_{\tilde{\Delta} \in \mathcal{K}(\theta^*) \cap S^{d-1}} |⟨ \tilde{\Delta}, \frac{X^T X}{n} w⟩| \leq c\sigma \sqrt{\frac{k \log(\frac{ed}{k})}{n}}. \]

The proof of these claims requires a geometric lemma which is of interest in its own right, so we state it here. Let \( \mathbb{B}_p(R) = \{ x \in \mathbb{R}^d \mid \|x\|_p \leq R \} \) denote the \( \ell_p \)-norm ball of radius \( R \), and let \( \text{clconv} \) denote the closed convex hull.

Lemma 8.3.1. For any vector \( \theta^* \) that is \( k \)-sparse, define the cone \( \mathcal{K}(\theta^*) \) as in (8.11), then we have
\[ \mathcal{K}(\theta^*) \cap S^{d-1} \overset{(a)}{=} \mathbb{B}_1(2\sqrt{k}) \cap \mathbb{B}_2(1) \overset{(b)}{=} \text{clconv} \left\{ \mathbb{B}_0(4k) \cap \mathbb{B}_2(1) \right\}. \] (8.12)

We proved part (a) at the end of Lecture # 3, and we leave the proof of part (b) as an exercise for the student. See also the paper [1].

Hint for part (b): For any two closed convex sets \( A, B \) of \( \mathbb{R}^d \), we have \( A \subseteq B \) if and only if \( \phi_A(u) \leq \phi_B(u) \) for all \( u \in \mathbb{R}^d \). Here \( \phi_A(u) := \sup_{v \in A} \langle v, u \rangle \) denotes the support function of the set \( A \), with a similar definition for \( \phi_B \).
Bibliography


