7.1 Introduction

In previous lectures, we have analyzed random forms of optimization problems, in which the randomness was injected (via random projection) for algorithmic reasons. On the other hand, in statistical problems—even without considering approximations—the starting point is a random instance of an optimization problem.

To be more concrete, suppose that we are interested in estimating some parameter \( \theta^* \in \mathbb{R}^d \) based on a set of \( n \) samples, say \( \{Z_1, \ldots, Z_n\} \). Many estimators of \( \theta^* \) are based on solving the (random) optimization problem

\[
\hat{\theta} \in \arg\min_{\theta \in C} L_n(\theta),
\]

where \( C \subset \mathbb{R}^d \) is some subset of \( \mathbb{R}^d \), and \( L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta; Z_i) \) decomposes as a sum of terms, one for each data point. Our interest will be in analyzing the sequence \( \{\theta^t\}_{t=0}^{\infty} \) generated by some optimization algorithm. A traditional analysis in (deterministic) optimization involves bounding the optimization error \( \|\theta^t - \hat{\theta}\| \), measuring the distance between the iterates and (some) optimum.

On the other hand, the population version of this problem is defined in terms of the averaged function \( \bar{L}(\theta) := \mathbb{E}[L_n(\theta)] \). If the original problem has been constructed in a reasonable way, then it should be the case that

\[
\theta^* \in \arg\min_{\theta \in C} \bar{L}(\theta),
\]

meaning that the quantity of interest is a global minimizer of the population function.

**Example:** As a concrete example in the linear regression case, suppose that we observe response-covariate pairs \( Z_i = (y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d \) that are linked via the standard linear model \( y = X\theta^* + w \) where \( w \sim N(0, \sigma^2 I_n) \). The finite-sample problem is to minimize cost function

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

Suppose that each covariate vector \( x_i \) is drawn i.i.d. from a zero-mean distribution with covariance matrix \( \Sigma \). We then have

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

As long as \( \Sigma \) is strictly positive definite, it follows that \( \theta^* \) is the unique minimizer of the population objective \( \bar{L} \).
Returning to the general thread, given our set-up thus far, the statistical error corresponds to the distance $\|\hat{\theta} - \theta^*\|$ between the global optimum $\hat{\theta}$ of the finite-sample estimator, and the population truth $\theta^*$. (When the global optimum is not unique, which may happen if the original problem is nonconvex, then this notion is not uniquely defined).

An interesting form of hybrid error, one suitable for analyzing a particular algorithm from the statistical point of view, is the quantity $\|\theta^t - \theta^*\|$, which measures the distance of the iterates from some algorithm to the population truth. In the next few lectures, we develop some techniques that are useful for directly controlling this hybrid error.

### 7.2 Behavior of projected gradient descent

We begin by considering the behavior of projected gradient ascent, as applied to a differentiable cost function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$. It is an iterative algorithm that, for a given step size $\eta > 0$, generates a sequence $\{\theta^t\}_{t=0}^\infty$ via the recursion

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

First-order Taylor approximation

If there is no constraint ($\mathcal{C} = \mathbb{R}^d$), then projected gradient descent reduces to gradient descent with constant stepsize $\eta$—that is, the update

$$
\theta^{t+1} = \theta^t - \eta \nabla \mathcal{L}(\theta^t),
$$

where $-\nabla \mathcal{L}(\theta^t)$ is the steepest descent direction. In the more general constrained setting, the algorithm is based on constrained minimization of the first-order approximation to $\mathcal{L}(\theta)$ around the current iterate $\theta^t$, along with the quadratic penalty term that encourages the next iterate to remain close to $\theta^t$.

We first analyze the oracle version of this problem—meaning in the case that is applied directly to the population problem (so that $\mathcal{L} = \bar{\mathcal{L}}$). We refer this to as the oracle version, since it is not actually something that can be implemented in practice (when we have only finitely many samples). On what should the iterates’ behavior depend on? The following two quantities play an important role in the theory:

- **Strong convexity**: There exists a constant $\gamma_\ell > 0$, such that for any $\theta + \Delta, \Delta \in \mathcal{C}$, we have

$$
\mathcal{L}(\theta + \Delta) - \{\mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), \Delta \rangle \} \geq \frac{\gamma_\ell}{2} \|\Delta\|^2.
$$

Error in first-order Taylor series

In our previous example of linear regression, this condition holds with $\gamma_\ell = \lambda_{\min}(\Sigma)$, corresponding to the smallest eigenvalue of the covariance matrix $\Sigma$.

- **Smoothness**: There exists a constant $\gamma_\mu > 0$, for any $\theta + \Delta, \Delta \in \mathcal{C}$,

$$
\mathcal{L}(\theta + \Delta) - \{\mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), \Delta \rangle \} \leq \frac{\gamma_\mu}{2} \|\Delta\|^2.
$$

In our previous example of linear regression, this condition holds with $\gamma_\mu = \lambda_{\max}(\Sigma)$, corresponding to the largest eigenvalue of the covariance matrix $\Sigma$. 


The ratio $\frac{\gamma_\ell}{\gamma_\mu} \in (0, 1]$ is the condition number, which determines how fast PGD method converges to $\theta^*$.

Let’s now state a classical result on the convergence rate of PGD.

**Theorem 7.2.1.** For any loss function $L : C \to \mathbb{R}$ that is $(\gamma_\ell, \gamma_\mu)$ strongly convex, smooth, the projected gradient descent method with stepsize $\eta = 1/\gamma_\mu$ satisfies

$$\|\theta^{t+1} - \theta^*\|^2 \leq (1 - \frac{\gamma_\ell}{\gamma_\mu})\|\theta^t - \theta^*\|^2. \quad (7.4)$$

In words, we are guaranteed that the iterates $\theta^t$ converges to $\theta^*$ geometrically with contraction factor $\kappa := \sqrt{1 - \frac{\gamma_\ell}{\gamma_\mu}} \in [0, 1)$.

We note that there are methods that can be used to improve the geometric rate of convergence: for instance, the accelerated gradient method of Nesterov [2] yields contraction with rate $\sqrt{1 - \frac{\gamma_\ell}{\gamma_\mu}}$. This type of geometric convergence can be shown to be optimal, via a matching lower bound.

**Proof.** Our proof hinges upon the following auxiliary lemma:

**Lemma 7.2.1.** When applied to any $(\gamma_\ell, \gamma_\mu)$ strongly convex and smooth function, projected gradient descent with stepsize $\eta = 1/\gamma_\mu$ satisfies

$$\gamma_\mu (\theta^t - \theta^{t+1}, \theta^t - \theta^*) \geq \frac{\gamma_\mu}{2}\|\theta^t - \theta^{t+1}\|^2 + \frac{\gamma_\ell}{2}\|\theta^t - \theta^*\|^2.$$  

To clarify, this lemma is important because it controls the decrease afforded by taking the PGD step from $\theta^t \to \theta^{t+1}$. This is easiest to see in the unconstrained case, in which $\theta^{t+1} - \theta^t = -(1/\gamma_\mu)\nabla L(\theta^t)$, so that Lemma 7.2.1 gives a bound bound on the quantity $\langle \nabla L(\theta^t), \theta^t - \theta^* \rangle \approx L(\theta^t) - L(\theta^*)$.

Taking Lemma 7.2.1 as given, the proof of the bound (7.4) is straightforward. By definition of the updates, we have

$$\|\theta^{t+1} - \theta^*\|^2 = \|\theta^{t+1} - \theta^t + \theta^t - \theta^*\|^2$$

$$= \|\theta^{t+1} - \theta^t\|^2 + \|\theta^t - \theta^*\|^2 + 2\langle \theta^{t+1} - \theta^t, \theta^t - \theta^* \rangle$$

$$\leq (1 - \frac{\gamma_\ell}{\gamma_\mu})\|\theta^t - \theta^*\|^2$$

where the last inequality uses the lemma.

Accordingly, it remains to to prove Lemma 7.2.1.

**Proof.** For notational convenience, let us define the function

$$\psi_t(\theta) = L(\theta^t) + \langle \nabla L(\theta^t), \theta - \theta^t \rangle + \frac{\gamma_\mu}{2}\|\theta - \theta^t\|^2.$$  

[https://blogs.princeton.edu/imabandit/2013/03/28/smoothfunctions/](https://blogs.princeton.edu/imabandit/2013/03/28/smoothfunctions/)
Note that by definition of the projected gradient descent update, we have \( \theta^{t+1} = \arg\min_{\theta \in C} \psi_t(\theta) \).

Using strong convexity at \( \theta^t \), we have
\[
\mathcal{L}(\theta^*) - \frac{\gamma_\ell}{2} \| \theta^t - \theta^* \|^2 \geq \mathcal{L}(\theta^t) + (\nabla \mathcal{L}(\theta^t), \theta^* - \theta^t)
= \mathcal{L}(\theta^t) + (\nabla \mathcal{L}(\theta^t), \theta^{t+1} - \theta^t) + (\nabla \mathcal{L}(\theta^t), \theta^* - \theta^{t+1}).
\] (7.5)

On the other hand, since \( \{\theta^{t+1} \text{ and } \theta^*\} \) are \{optimal and feasible\} for the problem defining the PGD update, we have \( (\nabla \psi_t(\theta^{t+1}), \theta^* - \theta^{t+1}) \geq 0 \), from which some algebra yields that
\[
(\nabla \mathcal{L}(\theta^t) + \gamma_\mu (\theta^{t+1} - \theta^t), \theta^* - \theta^{t+1}) \geq 0,
\] and thus
\[
(\nabla \mathcal{L}(\theta^t), \theta^* - \theta^{t+1}) \geq \gamma_\mu (\theta^t - \theta^{t+1}, \theta^* - \theta^{t+1}).
\] (7.6)

Combining inequalities (7.5) and (7.6) yields
\[
\mathcal{L}(\theta^*) - \frac{\gamma_\ell}{2} \| \theta^t - \theta^* \|^2 \geq \psi_t(\theta^{t+1}) - \frac{\gamma_\mu}{2} \| \theta^{t+1} - \theta^t \|^2 + \gamma_\mu (\theta^t - \theta^{t+1}, \theta^* - \theta^{t+1}).
\] (7.7)

Finally, by adding and subtracting a term involving \( \theta^t \) from inside the inner product, we find that
\[
\mathcal{L}(\theta^*) - \frac{\gamma_\ell}{2} \| \theta^t - \theta^* \|^2 \geq \psi_t(\theta^{t+1}) + \frac{\gamma_\mu}{2} \| \theta^{t+1} - \theta^t \|^2 + \gamma_\mu (\theta^t - \theta^{t+1}, \theta^* - \theta^t).
\]

By the smoothness assumption, we have
\[
\psi_t(\theta^{t+1}) = \mathcal{L}(\theta^t) + (\nabla \mathcal{L}(\theta^t), \theta^{t+1} - \theta^t) + \frac{\gamma_\mu}{2} \| \theta^{t+1} - \theta^t \|^2 \geq \mathcal{L}(\theta^{t+1}) \geq \mathcal{L}(\theta^*)
\]
where the last inequality follows from the optimality of \( \theta^* \). Plugging into inequality (7.7), some simple algebra then yields the claim of the lemma.

We have thus finished the proof of the Theorem 7.2.1.

To give some intuition, the strong convexity condition plus the smoothness condition is essentially saying that the object function is sandwiched between a lower quadratic function and a upper quadratic function. At each iterate, we minimize the upper approximation of the object function, we expect a linear convergence rate which is similar to that of minimizing a quadratic objective function with condition number \( \gamma_\ell/\gamma_\mu \).

### 7.3 Projected Gradient Descent on finite-sample problem

It is frequently the case that it is reasonable to expect that the population objective \( \bar{\mathcal{L}} \) satisfies the convexity/smoothness assumptions stated above. However, in finite-sample settings, these conditions may or may
not hold for the objective $\mathcal{L}_n$. One route is to directly try and prove that convexity/smoothness conditions—or appropriate relaxations thereof—hold for the finite-sample objective $\mathcal{L}_n$. (For instance, see the paper [1] for an analysis of this type in application to a certain class of high-dimensional problems.) An alternative route is to view an algorithm, when applied to the finite-sample problem, as a perturbed version of the population operator. Let us pursue this avenue here.

In particular, suppose that we let $\overline{G} : \mathcal{C} \to \mathcal{C}$ denote the projected gradient operator as applied at the population level (to the function $\overline{L}$), and let $G_n$ denote the PGD updates at the $n$-sample level (as applied to the objective function $\mathcal{L}_n$.) In abstract terms, Theorem 7.2.1 guarantees that the operator $\overline{G}$ is a contraction with parameter $\kappa = \sqrt{1 - \frac{\gamma}{\gamma_n}}$—i.e., it satisfies the bound

$$\|\overline{G}(\theta) - \theta^*\|_2 \leq \kappa \|\theta - \theta^*\|_2$$

Now consider the finite-sample operator $G_n$, generating the sequence $\{\theta^t\}_{t=0}^\infty$ from the finite-sample problem. If we think of the finite-sample operator $G_n$ as a perturbed version of this contraction, then these iterates should also behave “well” as long as the difference $G_n(\theta) - \overline{G}(\theta)$ is uniformly controlled.

In order to make this intuition precise, suppose that there exists constants $\alpha_n, \epsilon_n \geq 0$ such that for every $\Delta, \overline{\Delta} \in \mathcal{C} - \{\theta^*\} := \{\theta - \theta^*, \theta \in \mathcal{C}\}$, we have

$$\|\Delta, G_n(\theta^t + \overline{\Delta}) - \overline{G}(\theta^t + \overline{\Delta})\| \leq \alpha_n \|\Delta\|_2 \|\overline{\Delta}\|_2 + \epsilon_n \|\Delta\|_2.$$  \hfill (7.8)

In lectures to follow, we will give examples for which a bound of this form holds with useful choices of $(\alpha_n, \epsilon_n)$.

**Theorem 7.3.1.** Suppose that the projected gradient descent operator $\overline{G}$ is a contraction on $\mathcal{C}$ with parameter $\kappa$, and that condition (7.8) holds with parameters $(\alpha_n, \epsilon_n)$ such that $\overline{\kappa} := \kappa + \alpha_n < 1$. Then the sequence $\theta^{t+1} = G_n(\theta^t)$ satisfies the bound

$$\|\theta^{t+1} - \theta^*\|_2 \leq \frac{\epsilon_n}{1 - \overline{\kappa}} + \frac{\epsilon_n}{1 - \kappa}$$  \hfill (7.9)

for all iterations $t = 0, 1, \ldots$.

The presence of an error floor in this result is to be expected, and in fact—by statistical minimax theory—is unavoidable. We cannot expect that have an estimator based on a finite sample size $n$ that yields an arbitrarily good approximation to $\theta^*$. On the other hand, the bound (7.9) is meaningless unless we can establish that it holds with $\epsilon_n$ sufficiently “small”—in the ideal case, of the same order as the minimax error.

The proof to this theorem is quite straight-forward. The challenge in its use is in verifying that condition (7.8) holds with “good” choices of the constants $\alpha_n$ and $\epsilon_n$, and doing so requires techniques from empirical process theory and concentration of measure.

To conclude today’s lecture, let’s prove this theorem.

**Proof.** We introduce the convenient shorthand notation $\Delta^{t+1} := \theta^{t+1} - \theta^* = G_n(\theta^t) - \theta^*$. Using triangle inequality, we have

$$\|\Delta^{t+1}\|_2^2 = \langle \Delta^{t+1}, G_n(\theta^t) - \overline{G}(\theta^t) + \overline{G}(\theta^t) - \theta^* \rangle \leq \|\Delta^{t+1}, G_n(\theta^t) - \overline{G}(\theta^t)\|_2 + \|\Delta^{t+1}, \overline{G}(\theta^t) - \theta^*\|_2.$$
On one hand, by our assumption (7.8), we have
\[ T_1 \leq \alpha_n \| \Delta^{t+1} \|_2 \| \Delta^t \|_2 + \epsilon_n \| \Delta^{t+1} \|_2. \]

On the other hand, by the Cauchy-Schwarz inequality, we have
\[ T_2 \leq \| \Delta^{t+1} \|_2 \| \mathcal{G}(\theta^t) - \theta^* \|_2 \overset{(i)}{\leq} \kappa \| \Delta^{t+1} \|_2 \| \Delta^t \|_2 \]

where step (i) follows from the \( \kappa \)-contractivity of \( \mathcal{G} \). Putting together the pieces, we find that
\[ \| \Delta^{t+1} \|_2 \leq (\alpha_n + \epsilon_n) \| \Delta^t \|_2 + \epsilon_n. \]

Iterating this recursion yields the bound
\[ \| \Delta^{t+1} \|_2 \leq (\bar{\kappa})^{t+1} \| \Delta^0 \|_2 + \sum_{t=0}^{T} \bar{\kappa}^t \epsilon_n \overset{(ii)}{\leq} \bar{\kappa}^{t+1} \| \Delta^0 \|_2 + \frac{\epsilon_n}{1 - \bar{\kappa}}, \]

where step (ii) follows from the assumption that \( \bar{\kappa} = \alpha_n + \epsilon_n < 1 \), which allows to sum the geometric series.

\[ \square \]
Bibliography
