# Lecture 15: Projected Gradient Descent

Yudong Chen

Consider the problem

$$\min_{x\in\mathcal{X}} f(x), \tag{P}$$

where $f$ is continuously differentiable and $\mathcal{X} \subseteq \mathrm{dom}(f) \subseteq \mathbb{R}^n$ is a closed, convex, nonempty set.

In this lecture, we further assume $f$ is $L$-smooth (w.r.t. $\|\cdot\|_2$).

## 1 Projected gradient descent and gradient mapping

Recall the first-order condition for $L$-smoothness:

$$\forall x, y: \quad f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|_2^2. \tag{1}$$

For unconstrained problem, recall that each iteration of gradient descent (GD) minimizes the RHS above:

$$\textbf{(GD)} \qquad \begin{aligned} x_{k+1} &= \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}\|y-x_k\|_2^2 \right\} \\ &= x_k - \frac{1}{L}\nabla f(x_k). \end{aligned}$$

**Projected Gradient Descent (PGD)** For constrained problem, we consider PGD, which minimizes the RHS of (1) *over the feasible set* $\mathcal{X}$*:*

$$\textbf{(PGD)} \qquad \begin{aligned} x_{k+1} &= \operatorname*{argmin}_{y\in\mathcal{X}} \Bigg\{ f(x_k) + \underbrace{\langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}\|y-x_k\|_2^2}_{\text{complete this square}} \Bigg\} \\ &= \operatorname*{argmin}_{y\in\mathcal{X}} \left\{ \frac{L}{2}\left\| y - x_k + \frac{1}{L}\nabla f(x_k)\right\|_2^2 \right\} \\ &= P_{\mathcal{X}}\left( x_k - \frac{1}{L}\nabla f(x_k) \right). \end{aligned}$$

As in GD, we can also use some other stepsize $\frac{1}{\eta}$ with $\eta \ge L$:

$$x_{k+1} = P_{\mathcal{X}}\left( x_k - \frac{1}{\eta}\nabla f(x_k) \right).$$

It will be useful later to recall that Euclidean projection is characterized by the minimum principle

$$\forall y \in \mathcal{X}: \quad \langle P_{\mathcal{X}}(x) - x, y - P_{\mathcal{X}}(x)\rangle \ge 0. \tag{2}$$

### 1.1 Gradient mapping

Many results for GD can be generalized to PGD, where the role of the gradient is replaced by the gradient mapping defined below.

**Definition 1** (Gradient Mapping)**.** Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ is closed, convex and nonempty, and $f$ is differentiable. Given $\eta > 0$, the *gradient mapping* $G_\eta : \mathbb{R}^d \to \mathbb{R}^d$ is defined by

$$G_\eta(x) = \eta\left(x - P_{\mathcal{X}}\left(x - \frac{1}{\eta}\nabla f(x)\right)\right) \qquad \text{for } x \in \mathbb{R}^d.$$

Using the above definition, we can write PGD in a form that resembles GD:

$$x_{k+1} = x_k - \frac{1}{\eta}G_\eta(x_k).$$

The fixed points of PGD are those that satisfy $G_\eta(x) = 0$.

*Remark* 1. When $\mathcal{X} = \mathbb{R}^d$, $G_\eta(x) = \nabla f(x)$. Hence the gradient mapping generalizes the gradient.

For constrained problems, gradient mapping acts as a "proxy" for the gradient and has properties similar to the gradient.

- If $G_\eta(x) = 0$, then $x$ is a stationary point, meaning that $-\nabla f(x) \in N_{\mathcal{X}}(x)$. If $\left\|G_\eta(x)\right\|_2 \le \epsilon$, we get a near-stationary point.
- A Descent Lemma holds for PGD: if we use $\eta \ge L$, then $f(x_{k+1}) - f(x_k) \le -\frac{1}{2\eta}\left\|G_\eta(x_k)\right\|_2^2$.

We elaborate below.

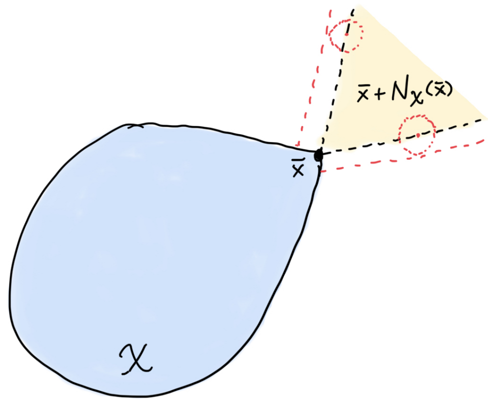
### 1.2 Gradient mapping and stationarity

Let $\mathcal{B}_2(z, r) := \left\{x \in \mathbb{R}^d : \|x - z\|_2 \le r\right\}$ denotes the Euclidean ball of radius $r$ centered at $z$. For two sets $S_1, S_2 \subset \mathbb{R}^d$, let $S_1 + S_2 = \{x + y : x \in S_1, y \in S_2\}$ denote their Minkowski sum.

The first lemma says if $\left\|G_\eta(x)\right\|_2$ is small, then $x$ almost satisfies the first-order optimality condition and can be considered a near-stationary point.

**Lemma 1** (Gradient mapping as a surrogate for stationarity)**.** *Consider (P), where $f$ is $L$-smooth, and $\mathcal{X}$ is closed, convex and nonempty. Denote $\bar{x} = P_{\mathcal{X}}\left(x - \frac{1}{\eta}\nabla f(x)\right)$, so that $G_\eta(x) = \eta(x - \bar{x})$. If $\left\|G_\eta(x)\right\|_2 \le \epsilon$ for some $\epsilon \ge 0$, then:*

$$\begin{aligned}
& -\nabla f(\bar{x}) \in N_{\mathcal{X}}(\bar{x}) + \mathcal{B}_2\left(0, \epsilon\left(\frac{L}{\eta} + 1\right)\right) \\
\iff & \forall u \in \mathcal{X} : \langle -\nabla f(\bar{x}), u - \bar{x}\rangle \le \epsilon\left(\frac{L}{\eta} + 1\right)\|u - \bar{x}\|_2 \\
\implies & \forall u \in \mathcal{X} \cap \mathcal{B}_2(\bar{x}, 1) : \langle -\nabla f(\bar{x}), u - \bar{x}\rangle \le \epsilon\left(\frac{L}{\eta} + 1\right).
\end{aligned}$$




*Proof.* Suppose that $\|G_\eta(x)\|_2 \le \epsilon$. By definition:

$$\bar{x} = P_{\mathcal{X}}\left(x - \frac{1}{\eta}\nabla f(x)\right) = \underset{y\in\mathcal{X}}{\operatorname{argmin}}\left\{\frac{1}{2}\left\|y - \left(x - \frac{1}{\eta}\nabla f(x)\right)\right\|_2^2\right\}.$$

Hence $\bar{x}$ satisfies the optimality condition of the minimization problem above:

$$-\left(\bar{x} - x + \frac{1}{\eta}\nabla f(x)\right) \in N_{\mathcal{X}}(\bar{x}).$$

Adding and subtracting $-\frac{1}{\eta}\nabla f(\bar{x})$:

$$-\frac{1}{\eta}\nabla f(\bar{x}) - \underbrace{\left(\bar{x} - x + \frac{1}{\eta}\nabla f(x) - \frac{1}{\eta}\nabla f(\bar{x})\right)}_{\rho} \in N_{\mathcal{X}}(\bar{x}).$$

Note that

$$\begin{aligned}
\|\rho\|_2 &= \Big\| \underbrace{\bar{x} - x}_{-\frac{1}{\eta}G_\eta(x)} + \frac{1}{\eta}\left(\nabla f(x) - \nabla f(\bar{x})\right)\Big\|_2 \\
&\le \frac{1}{\eta}\|G_\eta(x)\|_2 + \frac{1}{\eta}\underbrace{\|\nabla f(x) - \nabla f(\bar{x})\|_2}_{\le L\|x-\bar{x}\|_2 = \frac{L}{\eta}\|G_\eta(x)\|_2} \\
&\le \frac{1}{\eta}\left(1 + \frac{L}{\eta}\right)\|G_\eta(x)\|_2 \\
&\le \frac{\epsilon}{\eta}\left(1 + \frac{L}{\eta}\right).
\end{aligned}$$

Hence

$$\begin{aligned}
& -\frac{1}{\eta}\nabla f(\bar{x}) \in N_{\mathcal{X}}(\bar{x}) + \rho \\
\iff & -\nabla f(\bar{x}) \in N_{\mathcal{X}}(\bar{x}) + \eta\rho \\
\implies & -\nabla f(\bar{x}) \in N_{\mathcal{X}}(\bar{x}) + \mathcal{B}_2\left(0, \epsilon\left(1 + \frac{L}{\eta}\right)\right).
\end{aligned}$$

□

The next lemma shows that $x^*$ is a stationary point of (P) if and only if $G_\eta(x^*) = 0$.

**Lemma 2** (Wright-Recht Prop 7.8)**.** *Consider (P), where $f$ is $L$-smooth, and $\mathcal{X}$ is closed, convex and nonempty. Then, $x^* \in \mathcal{X}$ satisfies the first-order condition $-\nabla f(x^*) \in N_{\mathcal{X}}(x^*)$ if and only if $x^* = P_{\mathcal{X}}\left(x^* - \frac{1}{\eta}\nabla f(x^*)\right)$ (equivalently, $G_\eta(x^*) = 0$).*

*Proof.* **"if" part:** Suppose $G_\eta(x^*) = 0$. Applying Lemma 1 with $\epsilon = 0$ and noting that $\bar{x} = x^*$, we get $-\nabla f(x^*) \in N_{\mathcal{X}}(x^*)$.

- Explicit proof: $G_\eta(x^*) = 0$ means

$$x^* = P_{\mathcal{X}}\left(x^* - \frac{1}{\eta}\nabla f(x^*)\right) = \underset{y\in\mathcal{X}}{\operatorname{argmin}}\left\{\frac{1}{2}\left\|y - \left(x^* - \frac{1}{\eta}\nabla f(x^*)\right)\right\|_2^2\right\}.$$

By first-order optimality condition applied to the above minimization problem, we have

$$N_{\mathcal{X}}(x^*) \ni -\nabla\left[\frac{1}{2}\left\|y - \left(x^* - \frac{1}{\eta}\nabla f(x^*)\right)\right\|_2^2\right]\Bigg|_{y=x^*} = -\frac{1}{\eta}\nabla f(x^*),$$

which is equivalent to $N_{\mathcal{X}}(x^*) \ni -\frac{1}{\eta}\nabla f(x^*)$.

**"only if" part:** Suppose $-\nabla f(x^*) \in N_{\mathcal{X}}(x^*)$. By definition of $N_{\mathcal{X}}(x^*)$, we have

$$\begin{aligned}\forall y \in \mathcal{X}: \qquad 0 &\geq \frac{1}{\eta}\langle -\nabla f(x^*), y - x^*\rangle \\ &= \left\langle x^* - \frac{1}{\eta}\nabla f(x^*) - x^*, y - x^*\right\rangle.\end{aligned}$$

By the minimum principle (2) with $x = x^* - \frac{1}{\eta}\nabla f(x^*)$, the above inequality implies

$$x^* = P_{\mathcal{X}}(x) = P_{\mathcal{X}}\left(x^* - \frac{1}{\eta}\nabla f(x^*)\right).$$

□

### 1.3 Sufficient descent property/descent lemma

The gradient mapping also inherits the descent lemma.

**Lemma 3** (Thm 2.2.13 in Nes'18)**.** *Consider (P), where $f$ is an $L$-smooth function. If $\eta \geq L$ and $\bar{x} = x - \frac{1}{\eta}G_\eta(x)$, then:*

$$f(\bar{x}) \leq f(x) - \frac{1}{2\eta}\|G_\eta(x)\|_2^2.$$

*Proof.* From the first-order condition for $L$-smoothness (Lecture 4, Lemma 1),

$$\begin{aligned}f(\bar{x}) &\leq f(x) + \langle\nabla f(x), \bar{x} - x\rangle + \frac{\eta}{2}\left\|\bar{x} - x\right\|_2^2 \\ &= f(x) - \frac{1}{\eta}\langle\nabla f(x), G_\eta(x)\rangle + \frac{1}{2\eta}\|G_\eta(x)\|_2^2 && \bar{x} - x = -\frac{1}{\eta}G_\eta(x) \\ &= f(x) - \frac{1}{2\eta}\|G_\eta(x)\|_2^2 + \frac{1}{\eta}\langle G_\eta(x) - \nabla f(x), G_\eta(x)\rangle. && \text{add/subtract } \frac{1}{\eta}\langle G_\eta(x), G_\eta(x)\rangle = \frac{1}{\eta}\|G_\eta(x)\|_2^2\end{aligned}$$

It remains to show that $\langle G_\eta(x) - \nabla f(x), G_\eta(x)\rangle \le 0$. Plugging in the definition of $G_\eta(x)$, we have

$$\begin{aligned}
&\langle G_\eta(x) - \nabla f(x), G_\eta(x)\rangle \\
=&\left\langle \eta\left[x - P_{\mathcal{X}}\left(x - \frac{1}{\eta}\nabla f(x)\right)\right] - \nabla f(x), \eta\left[x - P_{\mathcal{X}}\left(x - \frac{1}{\eta}\nabla f(x)\right)\right]\right\rangle \\
=&\eta^2\left\langle \underbrace{x - \frac{1}{\eta}\nabla f(x)}_{y} - P_{\mathcal{X}}\left(x - \frac{1}{\eta}\nabla f(x)\right), x - P_{\mathcal{X}}\left(x - \frac{1}{\eta}\nabla f(x)\right)\right\rangle \\
=&\eta^2\left\langle y - P_{\mathcal{X}}(y), x - P_{\mathcal{X}}(y)\right\rangle \\
\le&0
\end{aligned}$$

by the minimum principle (2). □

## 2 Convergence guarantees for projected gradient descent

Consider the PGD update

$$x_{k+1} = P_{\mathcal{X}}\left(x_k - \frac{1}{L}\nabla f(x_k)\right) = x_k - \frac{1}{L}G_L(x_k),$$

where we fix the stepsize to be $\frac{1}{L}$, with $L$ being the smoothness parameter of $f$.

The convergence guarantees of PGD parallel those of GD.

### 2.1 Nonconvex case

Suppose $f$ is $L$-smooth.

By the Descent Lemma 3:

$$f(x_{k+1}) - f(x_k) \le -\frac{1}{2L}\|G_L(x_k)\|_2^2.$$

Summing up over $k$ and noting that the LHS telescopes:

$$f(x_{k+1}) - f(x_0) \le -\frac{1}{2L}\sum_{i=0}^{k}\|G_L(x_i)\|_2^2.$$

If $\bar{f} := \inf_{x\in\mathcal{X}} f(x) > -\infty$, then

$$\frac{1}{2L}\sum_{i=0}^{k}\|G_L(x_k)\|_2^2 \le f(x_0) - \bar{f}.$$

Hence

$$\min_{0\le i\le k}\|G_L(x_i)\|_2 \le \sqrt{\frac{2L\left(f(x_0) - \bar{f}\right)}{k+1}}.$$

Equivalently, after at most $k = \frac{8L\left(f(x_0)-\bar{f}\right)}{\epsilon^2}$ iterations of PGD, we have

$$\begin{aligned}
&\min_{0\le i\le k}\|G_L(x_i)\|_2 \le \frac{\epsilon}{2} \\
\implies &\exists i \in \{1, \dots, k+1\} : -\nabla f(x_i) \in N_{\mathcal{X}}(x_i) + \mathcal{B}_2(0, \epsilon)
\end{aligned}$$

where the last line follows from Lemma 1.

### 2.2 Convex case

Suppose $f$ is $L$-smooth and convex, with a global minimizer $x^*$.

1) From HW 4: $\|G_L(x_k)\|_2 \leq \|G_L(x_{k-1})\|_2, \forall k$. (In HW3 we proved a similar monotonicity property for the gradient.) The result above thus implies

$$\|G_L(x_k)\|_2 \leq \sqrt{\frac{2L\left(f(x_0)-\bar{f}\right)}{k+1}}.$$

2) From Descent Lemma 3:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|G_L(x_k)\|_2^2 \leq f(x_k),$$

so the function value is non-increasing in $k$.

3) Convexity gives the lower bound

$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle,$$

whence

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x^* - x_k \rangle \\ &= f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \langle \nabla f(x_k), x_{k+1} - x^* \rangle. \end{aligned} \tag{3}$$

(In the analysis of GD, we then use $\nabla f(x_k) = L(x_k - x_{k+1})$ and the 3-point identity). Recall that

$$x_{k+1} = \operatorname*{argmin}_{y\in\mathcal{X}} \left\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|_2^2 \right\}.$$

The first-order optimality condition gives

$$\forall y \in \mathcal{X} : \langle \nabla f(x_k) + L\left(x_{k+1} - x_k\right), y - x_{k+1} \rangle \geq 0.$$

Taking $y = x^*$ gives

$$\begin{aligned} \langle \nabla f(x_k), x_{k+1} - x^* \rangle &\leq L\langle x_{k+1} - x_k, x^* - x_{k+1} \rangle \\ &= \frac{L}{2}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 - \|x_{k+1} - x_k\|_2^2\right). \qquad \text{3-point identity} \end{aligned}$$

Plugging into (3), we get

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq \underbrace{f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle - \frac{L}{2}\|x_{k+1} - x_k\|_2^2}_{\leq 0 \text{ by } L\text{-smoothness}} + \frac{L}{2}\|x_k - x^*\|_2^2 - \frac{L}{2}\|x_{k+1} - x^*\|_2^2 \\ &\leq \frac{L}{2}\|x_k - x^*\|_2^2 - \frac{L}{2}\|x_{k+1} - x^*\|_2^2. \end{aligned}$$

We then follow the same steps as in the analysis of GD, summing up and telescoping the above inequality:

$$\sum_{i=0}^{k}\left(f(x_{i+1}) - f(x^*)\right) \leq \frac{L}{2}\|x_0 - x^*\|_2^2 - \frac{L}{2}\|x_{k+1} - x^*\|_2^2 \leq \frac{L}{2}\|x_0 - x^*\|_2^2.$$

But LHS $\geq (k+1)\left(f(x_{k+1}) - f(x^*)\right)$ due to monotonicity $f(x_{k+1}) \leq f(x_k) \leq \cdots \leq f(x_0)$. It follows that

$$f(x_{k+1}) - f(x^*) \leq \frac{L\|x_0 - x^*\|_2^2}{2(k+1)}.$$

### 2.3 Strongly convex case

Suppose $f$ is $m$-strongly convex and $L$-smooth, with a unique global minimizer $x^*$.

Since $x^*$ satisfies the first-order optimality condition, we have $P_{\mathcal{X}}\left(x^* - \frac{1}{L}\nabla f(x^*)\right) = x^*$ (Lemma 2). By nonexpansiveness of $P_{\mathcal{X}}$, we have

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \left\| P_{\mathcal{X}}\left(x_k - \frac{1}{L}\nabla f(x_k)\right) - P_{\mathcal{X}}\left(x^* - \frac{1}{L}\nabla f(x^*)\right) \right\|_2^2 \\
&\leq \left\| \left(x_k - \frac{1}{L}\nabla f(x_k)\right) - \left(x^* - \frac{1}{L}\nabla f(x^*)\right) \right\|_2^2 \\
&= \|x_k - x^*\|_2^2 + \frac{1}{L^2}\|\nabla f(x_k) - \nabla f(x^*)\|_2^2 - \frac{2}{L}\langle x_k - x^*, \nabla f(x_k) - \nabla f(x^*)\rangle .
\end{aligned}$$

But

$$\|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \leq L\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^*\rangle$$

by HW2 Q1, hence

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - \frac{1}{L}\langle x_k - x^*, \nabla f(x_k) - \nabla f(x^*)\rangle . \tag{4}$$

By strong convexity of $f$:

$$\begin{aligned}
f(x_k) &\geq f(x^*) + \langle \nabla f(x^*), x_k - x^*\rangle + \frac{m}{2}\|x_k - x^*\|_2^2, \\
f(x^*) &\geq f(x_k) + \langle \nabla f(x_k), x^* - x_k\rangle + \frac{m}{2}\|x_k - x^*\|_2^2 .
\end{aligned}$$

Adding up the two inequalities gives

$$\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^*\rangle \geq m\|x_k - x^*\|_2^2 .$$

(this is called the *strong monotonicity* property of the gradient.) Plugging into (4), we obtain

$$\begin{aligned}
&\|x_{k+1} - x^*\|_2^2 \leq \left(1 - \frac{m}{L}\right)\|x_k - x^*\|_2^2 \\
\implies &\|x_{k+1} - x^*\|_2^2 \leq \left(1 - \frac{m}{L}\right)^{k+1}\|x_0 - x^*\|_2^2 .
\end{aligned}$$

**Exercise 1.** Generalize the above results to PGD with a general stepsize $\frac{1}{\eta}$, where $\eta \geq L$.

## 3 Extensions

### 3.1 Acceleration (optional)

Nesterov's acceleration scheme can be extended to PGD:

$$\begin{aligned}
y_k &= x_k + \beta_k(x_k - x_{k-1}), && \text{momentum step} \\
x_{k+1} &= P_{\mathcal{X}}(y_k - \alpha_k \nabla f(y_k)) . && \text{projected gradient step}
\end{aligned}$$

This is a special case of the *accelerated proximal gradient method* (a.k.a. fast iterative shrinkage-thresholding algorithm, FISTA), which applies to problems of the form

$$\min_{x\in\mathbb{R}^d} f(x) + g(x), \tag{5}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and smooth, and $g : \mathbb{R}^d \to \bar{\mathbb{R}}$ is convex and lower semicontinuous with a computable proximal operator. Equation (5) is called a *composite problem*. As discussed in Lecture 1–2, the constrained problem (P) corresponds to a special case of the composite problem (5) with $g(x) = I_{\mathcal{X}}(x)$ being the indicator function of $\mathcal{X}$.

For details see the chapter from Beck's book.

### 3.2 Other search direction?

Recall that for unconstrained problems, we may use some other search direction $p_k$ instead of the negative gradient direction and still guarantee descent in function value (Lecture 7–8).

For constrained problem, can we use some other direction $p_k \neq -\nabla f(x_k)$ in the update $x_{k+1} = P_{\mathcal{X}}\left(x_k + \frac{1}{\eta} p_k\right)$? In general, doing so does *not* guarantee the descent property $f(x_{k+1}) < f(x_k)$, even when $p_k$ satisfies $\langle p_k, -\nabla f(x_k)\rangle > 0$. See below for an illustration.