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:

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

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

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

$y - (x_k - \frac{1}{L} \nabla f(x_k))$

($\frac{1}{2L} \|\nabla f(x_k)\|_2^2$ is a const when x_k is fixed)

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

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

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 $G_\eta(x) = 0$.

Then $x_{k+1} = x_k$

We just want this form, so we defined $G_\eta(x)$.

That is, $x_k - \frac{1}{\eta} G_\eta(x_k) = P_\mathcal{X}(x_k - \frac{1}{\eta}\nabla f(x_k))$

$$\Rightarrow G_\eta(x_k) = \eta(x_k - P_\mathcal{X}(x_k - \frac{1}{\eta}\nabla f(x_k))).$$

Remark 1. When $\boxed{\mathcal{X} = \mathbb{R}^d}$, $G_\eta(x) = \nabla f(x)$. 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.

① • $\boxed{\text{If } G_\eta(x) = 0}$, then x is a stationary point, meaning that $-\nabla f(x) \in N_\mathcal{X}(x)$. If $\|G_\eta(x)\|_2 \le \epsilon$, we get a near-stationary point.

② • A Descent Lemma holds for PGD: if we use $\eta \ge L$, then $f(x_{k+1}) - f(x_k) \le -\frac{1}{2\eta}\|G_\eta(x_k)\|_2^2$.

$\frac{1}{\eta} \le \frac{1}{L}$

We elaborate below.

1.2 Gradient mapping and stationarity

that is, $-\nabla f(x^*) \in N_\mathcal{X}(x^*)$

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

Lemma 1 (Wright-Recht Prop 7.8). *Consider (P), where f is $\boxed{L\text{-smooth,}}$ 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, $G_\eta(x^*) = 0$).*

Pf: $(\Leftarrow)$ $x^* = P_\mathcal{X}(x^* - \frac{1}{\eta}\nabla f(x^*)) = \arg\min_{y\in\mathcal{X}} \{\frac{1}{2}\|y - (x^* - \frac{1}{\eta}\nabla f(x^*))\|_2^2\}$.

By 1-st order necessary optimality cond, let $h(y) = \frac{1}{2}\|y - (x^* - \frac{1}{\eta}\nabla f(x^*))\|_2^2$,

$-\nabla h(x^*) \in N_\mathcal{X}(x^*)$. $\nabla h(y) = y - (x^* - \frac{1}{\eta}\nabla f(x^*))$. $\Rightarrow \nabla h(x^*) = \frac{1}{\eta}\nabla f(x^*)$.

Hence $-\frac{1}{\eta}\nabla f(x^*) \in N_\mathcal{X}(x^*)$. which is equivalent to $-\nabla f(x^*) \in N_\mathcal{X}(x^*)$ (property of a cone)

$(\Rightarrow)$. $-\nabla f(x^*) \in N_\mathcal{X}(x^*)$. $\Rightarrow \forall y \in \mathcal{X}, \langle -\nabla f(x^*), y - x^*\rangle \le 0$.

$\therefore \frac{1}{\eta}\langle -\nabla f(x^*), y - x^*\rangle = \langle x^* - \frac{1}{\eta}\nabla f(x^*) - x^*, y - x^*\rangle \le 0$.

$\Leftrightarrow \langle x^* - (x^* - \frac{1}{\eta}\nabla f(x^*)), y - x^*\rangle \ge 0$, $\forall y$. $\Rightarrow x^* = P_\mathcal{X}(x^* - \frac{1}{\eta}\nabla f(x^*))$

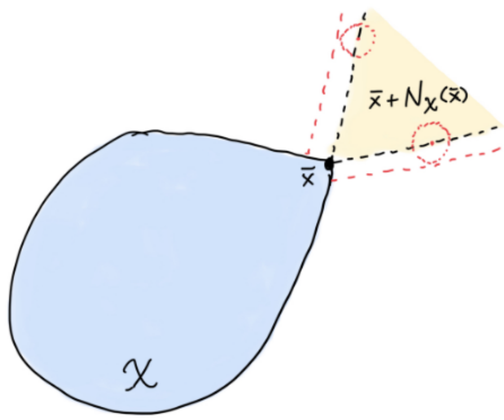
To state the next lemma, we need some notations. Let $\mathcal{B}_2(z,r) := \{x \in \mathbb{R}^d : \|x-z\|_2 \le r\}$ denotes the Euclidean ball of radius r centered at z. For two sets $S_1, S_2 \subset \mathbb{R}^d$, let $S_1 + S_2 = \{x+y : x \in S_1, y \in S_2\}$ denote their Minkowski sum.

$-\nabla f(x) \in N_{\mathcal{X}}(x)$

Our next Lemma 2 says if $\|G_\eta(x)\|_2$ is small, then x almost satisfies the first-order optimality condition and can be considered a near-stationary point. Lemma 2 is a generalization of the "if" part of Lemma 1.

Lemma 2 (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 $\|G_\eta(x)\|_2 \le \epsilon$ for some $\epsilon \ge 0$, then:*

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



pf: Suppose $\|G_\eta(x)\|_2 \le \varepsilon$.

$\bar{x} = P_{\mathcal{X}}(x - \frac{1}{\eta}\nabla f(x)) = \underset{y \in \mathcal{X}}{\arg\min} \left\{\frac{1}{2}\|y - (x - \frac{1}{\eta}\nabla f(x))\|_2^2\right\}$ $\Rightarrow$ $\bar{x}$ satisfies $-\nabla h(\bar{x}) \in N_{\mathcal{X}}(\bar{x})$

Let $h(y) = \frac{1}{2}\|y - (x - \frac{1}{\eta}\nabla f(x))\|_2^2$. $\nabla h(y) = y - (x - \frac{1}{\eta}\nabla f(x))$

$\nabla h(\bar{x}) = \bar{x} - (x - \frac{1}{\eta}\nabla f(x)) \Rightarrow -(\bar{x} - (x - \frac{1}{\eta}\nabla f(x))) \in N_{\mathcal{X}}(\bar{x})$

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

$$\|\rho\|_2 = \|\underbrace{(\bar{x} - x)}_{= -\frac{1}{\eta}G_\eta(x)} + \frac{1}{\eta}(\nabla f(x) - \nabla f(\bar{x}))\|_2 \le \frac{1}{\eta}\|G_\eta(x)\|_2 + \frac{1}{\eta}\underbrace{\|\nabla f(x) - \nabla f(\bar{x})\|_2}_{\text{Apply smoothness}}$$

$$\leq \frac{1}{\eta}\|G_\eta(x)\| + \frac{L}{\eta}\|\bar{x}-x\|_2 = \left(\frac{1}{\eta}+\frac{L}{\eta^2}\right)\|G_\eta(x)\|_2 = \frac{1}{\eta}\left(1+\frac{L}{\eta}\right)\|G_\eta(x)\|_2$$

$$\leq \frac{\varepsilon}{\eta}\left(1+\frac{L}{\eta}\right).$$

Hence $\quad -\frac{1}{\eta}\nabla f(\bar{x}) \in N_\mathcal{X}(\bar{x}) + \rho$

$$\Leftrightarrow \quad -\nabla f(\bar{x}) \in N_\mathcal{X}(\bar{x}) + \eta\rho$$

$$\Rightarrow \quad -\nabla f(\bar{x}) \in N_\mathcal{X}(x) + B_2\left(0, \varepsilon\left(1+\frac{L}{\eta}\right)\right)$$

1.3 Sufficient descent property/descent lemma

The gradient mapping also inherits the descent lemma.

Lemma 3 (Theorem 2.2.13 in Nesterov's 2018 textbook)**.** *Consider (P), where f is an L-smooth function. If $\boxed{\eta \geq L}$ and $\bar{x} = x - \frac{1}{\eta}G_\eta(x)$, then:*

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

Pf: Directly apply smoothness, also $\underline{\eta \geq L}$.

$$f(\bar{x}) \leq f(x) + \langle \nabla f(x), \bar{x}-x\rangle + \frac{\eta}{2}\|x-\bar{x}\|^2$$

$$= f(x) - \frac{1}{\eta}\langle \nabla f(x), G_\eta(x)\rangle + \frac{1}{2\eta}\|G_\eta(x)\|_2^2$$

$$= f(x) - \frac{1}{2\eta}\|G_\eta(x)\|_2^2 + \frac{1}{\eta}\langle G_\eta(x) - \nabla f(x), G_\eta(x)\rangle$$

Now let's show $\langle G_\eta(x) - \nabla f(x), G_\eta(x)\rangle \leq 0$. Plug in def of $G_\eta(x) := \eta\left(x - P_\mathcal{X}\left(x - \frac{1}{\eta}\nabla f(x)\right)\right)$

$$\langle G_\eta(x) - \nabla f(x), G_\eta(x)\rangle$$

$$= \eta^2\left\langle \underbrace{x - \frac{1}{\eta}\nabla f(x)}_{= y} - P_\mathcal{X}\left(x - \frac{1}{\eta}\nabla f(x)\right), x - P_\mathcal{X}\left(x - \frac{1}{\eta}\nabla f(x)\right)\right\rangle$$

Identity: (of $P_\mathcal{X}(y)$)

$$\underline{\langle P_\mathcal{X}(y) - y, x - P_\mathcal{X}(y)\rangle \geq 0.} \quad \forall x$$

$$= \eta^2\langle y - P_\mathcal{X}(y), x - P_\mathcal{X}(y)\rangle \leq 0.$$

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.

Similar res as GD:

2.1 Nonconvex case

Suppose f is L-smooth.

By the Descent Lemma 3:

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

Summing up over k and noting that the LHS telescopes:

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

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

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

Hence

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

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

$$\min_{0\leq i\leq k}\|G_L(x_i)\|_2 \leq \frac{\epsilon}{2}$$
$$\implies \exists i \in \{1,\ldots,k+1\} : -\nabla f(x_i) \in N_{\mathcal{X}}(x_i) + \mathcal{B}_2(0,\epsilon)$$

where the last line follows from Lemma 2.

$\eta = L$ here.

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.

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

$$f(x_{k+1}) - f(x^*) \leq f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x^* - x_k \rangle$$

$$= f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \langle \nabla f(x_k), x_{k+1} - x^* \rangle$$

(In GD, we next use $\nabla f(x_k) = L(x_k - x_{k+1})$).

Recall $x_{k+1} = \underset{y \in \mathcal{X}}{\arg\min} \left\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|_2^2 \right\} = \underset{y \in \mathcal{X}}{\arg\min} \left\{ \frac{1}{2}\|y - (x_k - \frac{1}{L}\nabla f(x_k))\|_2^2 \right\}$

By 1st-order optimality cond, $-h(y)|_{y = x_{k+1}} \in N_{\mathcal{X}}(x_{k+1})$

$\nabla h(y) = y - x_k + \frac{1}{L}\nabla f(x_k)$. $\quad -h(y)|_{y = x_{k+1}} = -(x_{k+1} - x_k + \frac{1}{L}\nabla f(x_k)) \in N_{\mathcal{X}}(x_{k+1})$

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

Take $y = x^*$.

$$\langle \nabla f(x_k), x_{k+1} - x^* \rangle \leq L \langle x_{k+1} - x_k, x^* - x_{k+1} \rangle$$

3-pt identity.

$$= \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|x_{k+1} - x_k\|^2 \right)$$

Hence, plug this back, we have

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

$$+\frac{L}{2}\|x_{k+1}-x_k\|_2^2 + \frac{L}{2}\left(\|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2-\|x_{k+1}-x_k\|^2\right)$$

$$=\frac{L}{2}\left(\|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2\right)$$

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

$$\le \frac{L}{2}\|x_0-x^*\|^2$$

By monotosity of $f(x_k)$, LHS $\ge (k+1)\left(f(x_{k+1})-f(x^*)\right)$

$(f(x_{k+1})\le f(x_k))$

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

2.3. Strongly cvx case. (f: L-smooth, m-strongly-cvx)

$x^*=P_X(x^*-\frac{1}{L}\nabla f(x^*))$, by lemma 1.

$$\|x_{k+1}-x^*\|^2 = \|P_X(x_k-\tfrac{1}{L}\nabla f(x_k)) - P_X(x^*-\tfrac{1}{L}\nabla f(x^*))\|_2^2$$

$$\le \|x_k-\tfrac{1}{L}\nabla f(x_k)-(x^*-\tfrac{1}{L}\nabla f(x^*))\|_2^2$$ $P_X(\cdot)$, non-expansive

$$=\|x_k-x^*\|_2^2+\frac{1}{L^2}\|\nabla f(x_k)-\nabla f(x^*)\|_2^2-\frac{2}{L}\langle\nabla f(x_k)-\nabla f(x^*),\, x_k-x^*\rangle$$

HW2, coerciveness of L-smooth function,

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

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

m- Strongly - cvx:

$$f(x_k) \ge f(x^*) + \langle \nabla f(x^*), x_k - x^* \rangle + \frac{m}{2}\|x_k - x^*\|_2^2$$

$$f(x^*) \ge f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle + \frac{m}{2}\|x_k - x^*\|_2^2$$

strong mono

Add up, $m\|x_k - x^*\|_2^2 \le \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle$ Plug back

$$\|x_{k+1} - x^*\|^2 \le \left(1 - \frac{m}{L}\right)\|x_k - x^*\|^2$$

$$\Rightarrow \|x_{k+1} - x^*\|^2 \le \left(1 - \frac{m}{L}\right)^{k+1}\|x_0 - x^*\|^2$$

Ex: Extend to $\eta \ge L$ case.

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.