

Homework 2: Projected GD, Frank-Wolfe, and Mirror Descent

Submission Guidelines: Your deliverables shall consist of 2 separate files – (i) A PDF file: Please compile all your write-ups and your report into one .pdf file (photos/scanned copies are acceptable; please make sure that the electronic files are of good quality and reader-friendly); (ii) A zip file: Please compress all your source code into one .zip file. Please submit your deliverables via E3.

Problem 1 (Convergence of PGD for Convex and Smooth Problems) (5+5+5+5+5=25 points)

As discussed in Lecture 8, we mentioned that for a convex and L -smooth function, the convergence rate of PGD is

$$f(x_t) - f(x^*) \leq \frac{3L\|x_0 - x^*\|^2 + (f(x_0) - f(x^*))}{t+1}. \quad (1)$$

In this problem, let us prove this result formally in a step-by-step manner.

(a) To begin with, let us show the following lemma: For any $x, z \in C$, let $\bar{x} = \Pi_C(x - \frac{1}{L}\nabla f(x))$ and $g_C(x) = L(x - \bar{x})$ (note that here we basically reuse the same notation as in our lecture slides). Then, we have

$$f(z) \geq f(\bar{x}) + g_C(x)^\top (z - x) + \frac{1}{2L}\|g_C(x)\|^2. \quad (2)$$

(Hint: Consider $f(z) - f(\bar{x}) = (f(z) - f(x)) - (f(\bar{x}) - f(x))$ and then utilize the convexity and smoothness conditions)

(b) By using the result in (a), show that the one-step improvement can be written as

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\|g_C(x_t)\|^2. \quad (3)$$

(c) Next, let us connect $\|g_C(x_t)\|$ to the objective function $f(x_t)$: Show that

$$\|g_C(x_t)\| \geq \frac{f(x_{t+1}) - f(x_t)}{\|x_t - x^*\|}. \quad (4)$$

(Hint: Find a proper way to apply the result in (a) and use Cauchy-Schwarz inequality)

(d) Define the sub-optimality gap at the t -th iteration as $\Delta_t := f(x_t) - f(x^*)$. By using the results in (a)-(c), show that $\Delta_{t+1} - \Delta_t \leq \frac{-\Delta_{t+1}^2}{2L\|x_0 - x^*\|^2}$.

(e) Finally, by using the result in (d), use an induction argument to show the convergence rate of PGD in (1).

Problem 2 (Projection Theorem)

(10+10=20 points)

In this problem, let's prove the fundamental Projection Theorem, which typically involves two very useful properties as follows:

(a) Let C be a convex set. Given some vector $x \in \mathbb{R}^d$, a vector $x_C \in C$ is equal to the projection $\Pi_C(x)$ if and only if

$$(x - x_C)^\top (z - x_C) \leq 0, \quad \forall z \in C. \quad (5)$$

(Hint: You just need to prove this in both directions by leveraging the FONC-C and the basic geometric properties of projection)

- (b) Given a convex set C , the projection operator $\Pi_C : \mathbb{R}^d \rightarrow C$ is non-expansive, i.e.,

$$\|x_C - z_C\| \leq \|x - z\|, \quad \forall x, z \in \mathbb{R}. \quad (6)$$

✓ Problem 3 (Bregman Divergence)

(5+10+10=25 points)

Recall from Lecture 10 that we learned the Bregman divergence, which is a key component in Mirror Descent. In this problem, you will have the opportunity to verify a few useful properties:

✓ (a) $\nabla_y D_\phi(y\|x) = \nabla\phi(y) - \nabla\phi(x)$.

✓ (b) For any strictly convex functions $\phi_1 : X \rightarrow \mathbb{R}, \phi_2 : X \rightarrow \mathbb{R}$ and $\lambda \geq 0$, we have

$$D_{\phi_1 + \lambda\phi_2}(y\|x) = D_{\phi_1}(y\|x) + \lambda D_{\phi_2}(y\|x) \quad (7)$$

✓ (c) As mentioned in Lecture 10, please show that the Bregman divergence satisfies the Generalized Pythagorean Theorem, i.e.,

$$D_\phi(z\|x) \geq D_\phi(z\|\bar{x}) + D_\phi(\bar{x}\|x), \quad (8)$$

where \bar{x} is the Bregman projection of x onto the feasible set C .

Problem 4 (Gurobi Optimization Solver for Frank-Wolfe)

(30 points)

In this homework, you will leverage the Gurobi Optimization Solver again to solve constrained optimization problems. Specifically, let us implement Frank-Wolfe method with the help of Gurobi optimization solver, which can be used through its Python API, and reproduce the performance of Frank-Wolfe in the top-left subfigure of Figure 2 of the paper (<https://arxiv.org/abs/2002.07003>) on a Portfolio Management dataset (provided to you on E3). Portfolio Management can be formulated as a constrained problem can be described as

$$\min_x f(x) := - \sum_{i=1}^n \log(a_i^\top x) \quad (9)$$

$$\text{subject to } \sum_{j=1}^p x_j = 1, x \geq 0, \quad (10)$$

where each a_i is a p -dimensional vector. To facilitate gradient computation, you may write your code in either PyTorch or TensorFlow. If you are a beginner in learning the deep learning framework, please refer to the following tutorials:

- PyTorch: <https://pytorch.org/tutorials/>
- Tensorflow: <https://www.tensorflow.org/tutorials/>

For the introduction to the Python API for Gurobi Optimization Solver, please see:

- To use Gurobi, you need to install Gurobipy (<https://pypi.org/project/gurobipy/>)
- Examples can be found at <https://www.gurobi.com/documentation/>

For the deliverables, please submit the following:

- Technical report: Please summarize all your experimental results in 1 single report (and please be brief)
- All your source code

Problem 2 (a)

$$(a) \Rightarrow x_c = \pi_C(x)$$

Proof $\|x-z\|^2 = \|(x-x_c) + (x_c-z)\|^2$

$$\geq \|x-x_c\|^2 + 2(x-x_c)^T(x_c-z) + \|x_c-z\|^2$$

$$\geq \|x-x_c\|^2 \quad \forall z \in C \text{ by (5).}$$

$$\text{Thus, } \|x-z\| \geq \|x-x_c\| \quad \forall z \in C.$$

$$\text{Hence, } x_c = \operatorname{argmin}_{z \in C} \|x-z\|$$

$$= \pi_C(x) \text{ by the definition of } \pi_C(x). \blacksquare$$

Problem 2 (a)

$$x_c = \pi_C(x) \Rightarrow (a)$$

Proof $x_c = \operatorname{argmin}_{z \in C} \|x-z\|$

$$= \operatorname{argmin}_{z \in C} \frac{1}{2} \|x-z\|^2$$

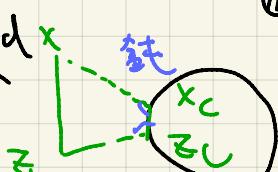
By FONC-C, since x_c is a local minimizer of $f(z) = \frac{1}{2} \|x-z\|^2$,
and $C \subset \mathbb{R}^d$ convex,

$$\nabla f(x_c)^T(z-x_c) = -(x-x_c)^T(z-x_c) \geq 0 \quad \forall z \in C.$$

$$\text{Thus, } (x-x_c)^T(z-x_c) \leq 0 \quad \forall z \in C. \blacksquare$$

Problem 2 (b) Let $x, z \in \mathbb{R}^d$

$$x_c = \pi_C(x) \Rightarrow (b)$$



Proof By (5), $(x-x_c)^T(z_c-x_c) \leq 0$

+ $(z-z_c)(x_c-z_c) \leq 0$.

$$\Rightarrow (z-z_c+x_c-x)^T(x_c-z_c) \leq 0$$

$$\Rightarrow (z-x)^T(x_c-z_c) + \|x_c-z_c\|^2 \leq 0$$

$$\Rightarrow \|x_c-z_c\|^2 \leq (z-x)^T(x_c-z_c) \leq \|z-x\| \|x_c-z_c\| \text{ by 矩阵不等式} \quad \text{⊗}$$

If $x_c = z_c$, we are done

Since $\|x_c - z_c\| = 0 \leq \|x - z\|$.

Assume $x_c \neq z_c$

$$\frac{\textcircled{X}}{\|z_c - x_c\|} : \|x_c - z_c\| \leq \|x - z\| \quad \forall x, z \in \mathbb{R}. \quad \textcircled{1}$$

Problem 3.(a)

By definition, $D_\phi(y||x) = \phi(y) - \phi(x) - \nabla \phi^T(x)(y-x)$.

$$\begin{aligned} \nabla_y D(y||x) &= \nabla \phi(y) - \nabla - \nabla \phi(x) \times | \\ &= \nabla \phi(y) - \nabla \phi(x) \quad \textcircled{2} \end{aligned}$$

Problem 3.(b)

By definition, $D_\phi(y||x) = \phi(y) - \phi(x) - \nabla \phi^T(x)(y-x)$.

$$\begin{aligned} \text{Thus, } D_{\phi_1 + \lambda \phi_2}(y||x) &= (\phi_1 + \lambda \phi_2)(y) - (\phi_1 + \lambda \phi_2)(x) - \nabla(\phi_1 + \lambda \phi_2)^T(x)(y-x) \\ &= \phi_1(y) - \phi_1(x) - \nabla \phi_1^T(x)(y-x) \\ &\quad + \lambda(\phi_2(y) - \phi_2(x) - \nabla \phi_2^T(x)(y-x)) \\ &= D_{\phi_1}(y||x) + \lambda D_{\phi_2}(y||x). \quad \textcircled{3} \end{aligned}$$

Problem 3.(c)

$$D_\phi(z||x) \geq D_\phi(z||\bar{x}) + D_\phi(\bar{x}||x)$$

$$\begin{aligned} \Leftrightarrow \cancel{\phi(z)} - \cancel{\phi(x)} - \phi(x)^T(z-x) &\geq \cancel{\phi(z)} - \cancel{\phi(x)} - \nabla \phi(\bar{x})^T(z-\bar{x}) \\ &\quad + \cancel{\phi(\bar{x})} - \cancel{\phi(x)} - \nabla \phi(x)^T(\bar{x}-x) \end{aligned}$$

by the definition of $D_\phi(\cdot||\cdot)$

$$\Leftrightarrow \phi(x)^T(\bar{x}-x+z-x) \geq \nabla \phi(\bar{x})^T(\bar{x}-z)$$

$$\Leftrightarrow (\phi(x) - \phi(\bar{x}))^T(\bar{x}-z) \geq 0$$

$$\Leftrightarrow (\nabla_x D_\phi(x||y)|_{y=\bar{x}})^T(\bar{x}-z) \geq 0 \text{ by (a).}$$

$$\Leftrightarrow (\nabla_y D_\phi(y||x)|_{y=\bar{x}})^T(z-\bar{x}) \geq 0 \text{ by (a)}$$

Define $f(y) = D_\phi(y \| x)$.

We only need to proof $\nabla f(\bar{x})^T(z - \bar{x}) \geq 0$.

Since $\bar{x} = \arg \min_{y \in C} D_\phi(y \| x)$, and $f(y)$ is convex by Property 3 of Bregman Divergence, by Lec 1 Ch 1, $\nabla f(\bar{x})(z - \bar{x}) \geq 0 \quad \forall z \in C$.

Hence, $D_\phi(z \| x) \geq D_\phi(z \| \bar{x}) + D_\phi(\bar{x} \| x)$. \blacksquare

Problem 1(a)

$$\begin{aligned}
 f(z) - f(x) &= (f(z) - f(x)) - (f(\bar{x}) - f(x)) \\
 &\geq \underbrace{\nabla f(x)(z - x)}_{f \text{ is convex}} - \underbrace{(\nabla f(x)(x - \bar{x}) + \frac{L}{2} \| \bar{x} - x \|^2)}_{L\text{-Smooth}} \\
 &= \nabla f(x)(z - \bar{x}) - \frac{L}{2} \| \bar{x} - x \|^2 \\
 &\geq g_c^T(x)(z - \bar{x}) - \frac{L}{2} \| \bar{x} - x \|^2 \quad \text{by (Claim): } \nabla f(x)(z - \bar{x}) \geq g_c^T(x)(z - \bar{x}) \\
 &= \underbrace{g_c^T(x)(z - x + x - \bar{x})}_{\text{since } g_c(x) = L(x - \bar{x})} - \frac{L}{2} \| \bar{x} - x \|^2 \\
 &= \underbrace{g_c^T(x)(z - x) + L \| x - \bar{x} \|^2}_{\text{by } g_c(x) = L(x - \bar{x})} - \frac{L}{2} \| \bar{x} - x \|^2 \\
 &= g_c^T(x)(z - x) + \frac{1}{2L} \| g_c(x) \|^2 \quad \text{by } g_c(x) = L(x - \bar{x})
 \end{aligned}$$

(Claim): $\nabla f(x)(z - \bar{x}) \geq g_c^T(x)(z - \bar{x})$

By Projection Optimality Condition,

$$(x - \frac{1}{L} \nabla f(x) - \bar{x})^T(z - \bar{x}) \leq 0 \quad \forall z \in C$$

$$\Rightarrow (x - \bar{x} - \frac{1}{L} \nabla f(x))^T(z - \bar{x}) \leq 0$$

$$\Rightarrow (\frac{1}{L} g_c(x) - \frac{1}{L} \nabla f(x))^T(z - \bar{x}) \leq 0$$

Since $g_c(x) = L(x - \bar{x})$

$$\Rightarrow \nabla f(x)^T(z - \bar{x}) \geq g_c^T(x)(z - \bar{x}). \quad \blacksquare$$

Projection Optimality Condition: $(x - \frac{1}{L} \nabla f(x) - \bar{x})^T (z - \bar{x}) \leq 0 \quad \forall z \in C$

It is a direct result by Projection Theorem. \square

Problem 1.(b)

Let $z = x_{t+1}$ into (1).

$$\text{Thus, } f(z) - f(x) + g_c^T(x)(z - x) \leq \frac{-1}{2L} \|g_c(x)\|^2$$

$$\text{Since } x_{t+1} = \bar{x} \text{ and let } x = x_t, f(x_{t+1}) - f(x) \leq \frac{-1}{2L} \|g_c(x_t)\|^2. \quad \square$$

Problem 1.(c)

$$\begin{aligned} \Delta_{t+1} - \Delta_t &= (f(x_{t+1}) - f(x^*)) - (f(x_t) - f(x^*)) \\ &= f(x_{t+1}) - f(x_t) \\ &\leq -\frac{1}{2L} \|g_c(x_t)\|^2 \text{ by (3) - ②} \end{aligned}$$

$$\text{Since } g_c^T(x)(x - z) \geq f(x) - f(z) + \frac{1}{2L} \|g_c(x)\|^2,$$

$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$
 $x_t \quad x^* \quad x_{t+1} \quad x^* \quad x_t$

$$g_c^T(x)(x_t - x^*) \geq f(x_{t+1}) - f(x^*) + \frac{1}{2L} \|g_c(x_t)\|^2 \geq f(x_{t+1}) - f(x^*)$$

$$\begin{aligned} \text{Thus, by ④, } \|g_c^T(x)\| \|x_t - x^*\| &\geq g_c^T(x)(x_t - x^*) \\ &\geq f(x_{t+1}) - f(x^*). \end{aligned}$$

$$\text{Hence, } \|g_c^T(x)\| \geq \frac{f(x_{t+1}) - f(x^*)}{\|x_t - x^*\|}.$$

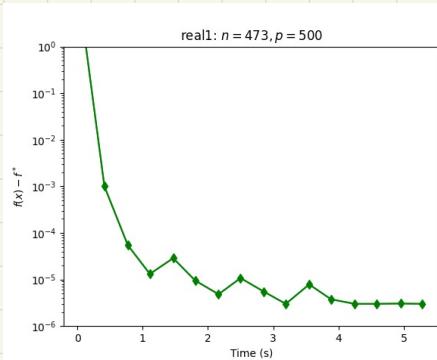
Problem 1.(d)

$$\text{By (4), } \|g_c^T(x)\| \geq \frac{\|f(x_{t+1}) - f(x^*)\|}{\|x_t - x^*\|}. \text{ Since } \|x_t - x^*\| > 0 \text{ and } f(x_{t+1}) \geq f(x^*) - ③$$

$$\begin{aligned} \text{By ②, } \Delta_{t+1} - \Delta_t &\leq -\frac{1}{2L} \|g_c(x_t)\|^2 \\ &\leq \frac{-1}{2L} \frac{\|f(x_{t+1}) - f(x^*)\|^2}{\|x_t - x^*\|^2} \text{ by } ③ \times (-\frac{1}{2L}) \\ &= \frac{-\Delta_{t+1}^2}{2L \|x_t - x^*\|^2} \text{ by the definition of } \Delta_{t+1} \end{aligned}$$

Problem 4.

Result



of epoch = 4500

1 point / 300 epochs

When epoch \rightarrow ,

$$f(x) - f(x^*) \downarrow.$$

Problem 1.(e)

We use the following Lemma proven in Reference P.b

$$\text{Lemma } \sum_{t=0}^k (f(x_{t+1}) - f(x^*)) \leq \frac{L}{2} \|x_0 - x^*\|^2$$

And $f(x_t) \geq f(x_{t+1})$, since by (2),

$$\text{put } z = x = x_t, \text{ we have } f(x_t) \geq f(x_{t+1}) + \frac{1}{2L} \|g_t(x)\|^2.$$

$$\text{Thus, } f(x_t) - f(x^*) \geq f(x_{t+1}) - f(x^*). \quad \textcircled{1}$$

$$\text{By Lemma } \sum_{t=0}^k (f(x_{t+1}) - f(x^*)) + f(x_k) - f(x^*) \leq \frac{L}{2} \|x_0 - x^*\|^2 + (f(x_k) - f(x^*)). \quad \textcircled{2}$$

$$\text{By } \textcircled{1} \text{ and } \textcircled{2}, (t+2)(f(x_{t+1}) - f(x^*)) \leq \frac{L}{2} \|x_0 - x^*\|^2 + (f(x_k) - f(x^*)).$$

$$\text{Thus, } f(x_{t+1}) - f(x^*) \leq \frac{3L \|x_0 - x^*\|^2 + (f(x_k) - f(x^*))}{t+2}.$$

$$\text{Hence, } f(x_t) - f(x^*) \leq \frac{3L \|x_0 - x^*\|^2 + (f(x_k) - f(x^*))}{t+1}. \quad \textcircled{3}$$

Lecture 15: Projected Gradient Descent

Yudong Chen

Consider the problem

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

where f is continuously differentiable and $\mathcal{X} \subseteq \text{dom}(f) \subseteq \mathbb{R}^n$ 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:

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

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

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

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

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

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

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

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

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

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* $G_\eta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined by

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

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 $\|G_\eta(x)\|_2 \leq \epsilon$, we get a near-stationary point.
- A Descent Lemma holds for PGD: if we use $\eta \geq L$, then $f(x_{k+1}) - f(x_k) \leq -\frac{1}{2\eta} \|G_\eta(x_k)\|_2^2$.

We elaborate below.

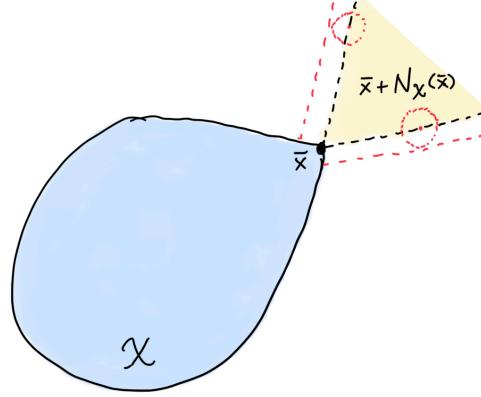
1.2 Gradient mapping and stationarity

Let $\mathcal{B}_2(z, r) := \{x \in \mathbb{R}^d : \|x - z\|_2 \leq r\}$ 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.

The first lemma 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 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 $\|G_\eta(x)\|_2 \leq \epsilon$ for some $\epsilon \geq 0$, then:

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



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

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

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

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

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

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

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

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

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

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

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),

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

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

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

by the minimum principle (2). \square

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:

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

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

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

Hence

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

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

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

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

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

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

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

(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

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

The first-order optimality condition gives

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

Taking $y = x^*$ gives

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

Plugging into (3), we get

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

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

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

But $\text{LHS} \geq (k+1)(f(x_{k+1}) - f(x^*))$ due to monotonicity $f(x_{k+1}) \leq f(x_k) \leq \dots \leq f(x_0)$. 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 $P_{\mathcal{X}}(x^* - \frac{1}{L}\nabla f(x^*)) = x^*$ (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

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

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

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

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and smooth, and $g : \mathbb{R}^d \rightarrow \bar{\mathbb{R}}$ 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.

