6FMAI19 Nonlinear Optimization	Spring, 2022

Lecture #5 — Feb. 16th, 2022

Lecturer: Yura Malitsky *Scribe: Oksana Moryakova*

1 Projected gradient descent

During the last lecture, we introduced the optimality condition for a convex function:

- $\begin{cases} f \text{ is convex and differentiable} \\ \mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) \end{cases} \quad \Leftrightarrow \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ for all $\mathbf{x} \in \mathcal{C}$.

where $\mathcal{C}$ is closed and convex, $\mathcal{C} \subset \mathbb{R}^n$. Then, how to solve the problem:

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) \tag{1}$$

1.1 Projection of gradient

If we have a convex closed set $\mathcal{C}$, we define an operator

$$\mathbf{P_C x} = \arg\min_{\mathbf{y}\in\mathcal{C}} \|\mathbf{y} - \mathbf{x}\|. \tag{2}$$

where the function $\|\mathbf{y} - \mathbf{x}\|$ is special - this is coercive function.

Definition 1. *A coercive function is a function $g(\mathbf{y})$ that "grows rapidly", i.e.*

$$if \ \|\mathbf{y}\| \to \infty \Rightarrow \|g(\mathbf{y})\| \to \infty. \tag{3}$$

Why the solution of (1) exists:

- closeness of $\mathcal{C}$;
- convexity of $\mathcal{C}$ that gives uniqueness of the solution;
- coercive function $\|\mathbf{y} - \mathbf{x}\|$.

How to compute it using given set $\mathcal{C}$? There are several cases:

- $\mathcal{C}$ is a unit ball, $\mathcal{C} = B(0,1) \Rightarrow$ solution: $\mathbf{P_C x} = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|} & \text{if } \|\mathbf{x}\| > 1, \\ \mathbf{x} & \text{otherwise.} \end{cases}$
- $\mathcal{C}$ is a hyperplane, $\mathcal{C} : \{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle = b\} \Rightarrow$ solution: $\mathbf{P_C x} = \mathbf{x} - \frac{b - \langle \mathbf{a}, \mathbf{x} \rangle}{\|\mathbf{a}\|^2}\mathbf{a}$

1.2 Characteristic property of projection

$\bar{\mathbf{x}}$ is projection of $\mathbf{x}$ on the $\mathcal{C}$ if and only if $\langle \bar{\mathbf{x}} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \geq 0$, i.e.

$$\bar{\mathbf{x}} = \mathbf{P_C x} \Leftrightarrow \langle \bar{\mathbf{x}} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \geq 0 \tag{4}$$

for any $\mathbf{y} \in \mathcal{C}$. During the previous lectures, we got the gradient descent algorithm

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k), \quad \text{for } k = 0, 1, \cdots, K-1. \tag{5}$$

But now, we are going to consider projection of gradient descent to the set $\mathcal{C}$:

$$\mathbf{x}_{k+1} = \mathbf{P_C x}(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)), \quad \text{for } k = 0, 1, \cdots, K-1. \tag{6}$$

Proof. Let's take an example $f(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2$. For this function $\bar{x} = \arg\min_{y\in\mathcal{C}} f(\mathbf{y})$. Since f is convex $\Rightarrow \quad \langle\nabla f(\bar{\mathbf{x}}), \mathbf{y}-\bar{\mathbf{x}}\rangle \geq 0 \quad \forall \mathbf{y}\in\mathcal{C}$. The gradient of our example: $\nabla f(\mathbf{y}) = \mathbf{y}-\mathbf{x}$. If to substitute $\bar{\mathbf{x}}$ instead of $\mathbf{y}$, we can observe the same equality. □

1.3 Nonexpensive property of projection

We can derive the following inequality:

$$\|P_C x - P_C y\| \leq \|x-y\|, \tag{7}$$

i.e. the distance between images $P_C x$ and $P_C y$ is always less or equal than the distance between preimages x and y.

Proof. Let's consider two points x and y, for which $\langle P_{cx}-x, P_{cy}-P_{cx}\rangle \geq 0$ and $\langle P_{cy}-y, P_{cx}-P_{cy}\rangle \geq 0$. After summarization and simplification the result correspond to (7). □

1.4 Projected gradient descent algorithm

Let's consider projection of gradient descent method (6). Assume that f is differentiable and convex, $\mathcal{C}$ is closed. We're going to apply the main characterization property (4):

$$\langle \mathbf{x}_{k+1}-\mathbf{x}_k+\alpha\nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_{k+1}\rangle \geq 0. \quad \forall \mathbf{x}\in\mathcal{C} \tag{8}$$

Split it into two parts

$$\langle \mathbf{x}_{k+1}-\mathbf{x}_k, \mathbf{x}-\mathbf{x}_{k+1}\rangle + \alpha\langle\nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_{k+1}\rangle \geq 0. \tag{9}$$

Using the definition of squared norm of sum

$$2\langle\mathbf{a},\mathbf{b}\rangle = \|\mathbf{a}+\mathbf{b}\|^2 - \|\mathbf{a}\|^2 - \|\mathbf{b}\|^2, \tag{10}$$

we obtain

$$\|\mathbf{x}_k-\mathbf{x}\|^2 - \|\mathbf{x}_{k+1}-\mathbf{x}\|^2 - \|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 + 2\alpha\langle\nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_{k+1}\rangle \geq 0. \tag{11}$$

Let's concentrate on the last term of (11) and split the inner product into two ones

$$\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}-\mathbf{x}_{k+1}\rangle = \langle\nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \langle\nabla f(\mathbf{x}_k), \mathbf{x}_k-\mathbf{x}_{k+1}\rangle, \tag{12}$$

where the term $\langle\nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_{k+1}\rangle$ is constrained from above, the term $\langle\nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle \leq f(\mathbf{x})-f(\mathbf{x}_k)$ according to the main inequality of convexity.
Considering Decent lemma definition

$$0 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle\nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2, \tag{13}$$

we can get

$$f(x_{\mathbf{x}+1}) - f(\mathbf{x}_k) - \langle\nabla f(\mathbf{x}_k), \mathbf{x}_{k+1}-\mathbf{x}_k\rangle \leq \frac{L}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2. \tag{14}$$

Rewrite (12) in the form of $\langle\nabla f(\mathbf{x}_{k+1}, \mathbf{x}-\mathbf{x}_{k+1}\rangle \leq f(\mathbf{x}) - f(\mathbf{x}_k) + \frac{L}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 - f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)$. Consequently,

$$\langle\nabla f(\mathbf{x}_{k+1}, \mathbf{x}-\mathbf{x}_{k+1}\rangle \leq f(\mathbf{x}) - f(\mathbf{x}_{k+1}) + \frac{L}{2}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2. \tag{15}$$

Now, when we estimated all terms of (12), rewrite (11) in the form of

$$\|\mathbf{x}_k - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + 2\alpha(f(\mathbf{x}) - f(\mathbf{x}_{k+1})) + \alpha L \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \geq 0. \quad (16)$$

Considering that $\alpha \leq \frac{1}{L}$,

$$\|\mathbf{x}_k - \mathbf{x}\|^2 + 2\alpha(f(\mathbf{x}_{k+1}) - f(\mathbf{x})) \leq \|\mathbf{x}_k - \mathbf{x}\|^2. \quad (17)$$

Until this moment x was an arbitrary point. However, it to say that $\mathbf{x} = \mathbf{x}^*$ that is solution, then we obtain

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 + 2\alpha(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)) \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2. \quad (18)$$

Assume that $\mathbf{x}_k \to \mathbf{x}^* \quad \Rightarrow$, we obtain the equation of projected gradient descent:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2\alpha}. \quad (19)$$

2 Subgradient and subdifferentials

2.1 Subgradient

Let's allow a function to take value $+\infty$, i.e. $f : \mathbb{R}^n \to \bar{\mathbb{R}}$.

Definition 2. *Indicator function is defined as*

$$\delta_c(x) = \begin{cases} 0 & x, \in \mathcal{C}, \\ +\infty, & x \notin \mathcal{C}, \end{cases} \quad (20)$$

where $\mathcal{C}$ is closed set.

We're going to minimize

$$F(\mathbf{x}) = \min_x [f(\mathbf{x}) + \delta_c(\mathbf{x})], \quad (21)$$

where $F : \mathbb{R}^n \to \bar{\mathbb{R}}$.

Definition 3. *Domain of a function is*

$$\mathrm{dom(f)} = \{\mathbf{x} : f(x) < +\infty\}. \quad (22)$$

Before we considered differentiable functions. Now, our function f is non-smooth, i.e. is not differentiable (at two points in our case).

Definition 4. $f : \mathbb{R}^n \to \bar{\mathbb{R}}, \mathbf{x} \in \mathrm{dom(f)}, \mathbf{u}$ *is a subgradient of function f at x, if*

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle \forall y \in \mathbf{R}^n. \quad (23)$$

2.2 Subdifferential

Definition 5. *Subdifferential of f at* $\mathbf{x} : \partial f(\mathbf{x}) = \{\mathbf{u}$ *is subgradient*$\}$.

Remark 6. *If f is differentiable at* $\mathbf{x} \Rightarrow \partial f(\mathbf{x}) = \{\nabla f(\mathbf{x}\}$.

Let's consider an example $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$, $f_i(\mathbf{x})$ is diff.

$$\partial f(\mathbf{x}) = \begin{cases} \nabla f_1(\mathbf{x}) & \text{if } f_1(\mathbf{x}) > f_2(\mathbf{x}), \\ \nabla f_2(\mathbf{x}) & \text{if } f_1(\mathbf{x}) < f_2(\mathbf{x}), \\ \text{conv}\{\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x})\} & \text{otherwise,} \end{cases}$$

where $\text{conv}\{\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x})\}$ means convex combinations, i.e. $\{\mathbf{y} : \mathbf{y} = \alpha \nabla f_1(\mathbf{x}) + (1 - \alpha)\nabla f_2(\mathbf{x})\}$. We're going to minimize function

$$\min_{\mathbf{x}} f(\mathbf{x}),$$

where f is convex.

$$\mathbf{x} \in \arg\min_{\mathbf{x}} f(\mathbf{x}) \Leftrightarrow \mathbf{0} \in \partial f(\mathbf{x})$$

Proof. Using the definition of a differential of a subgradient, we can obtain

$$\mathbf{0} \in \partial f(\mathbf{x}) \Rightarrow f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{0}, \mathbf{y} - \mathbf{x} \rangle = \mathbf{0}, \quad \forall \mathbf{x}, \mathbf{y}.$$

□

2.3 Property of subdifferentials - Monotonicity

Suppose we have two subgradients

$$\begin{cases} \mathbf{g_x} \in \partial f(\mathbf{x}), \\ \mathbf{g_y} \in \partial f(\mathbf{y}), \end{cases} \quad \Leftrightarrow \quad \langle \mathbf{g_x} - \mathbf{g_y}, \mathbf{x} - \mathbf{y} \rangle \geq 0. \tag{24}$$

Proof. Since $\mathbf{g_x}$ is a subgradient, then

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{g_x}, \mathbf{y} - \mathbf{x} \rangle.$$

Since $\mathbf{g_y}$ is a subgradient as well, then

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \langle \mathbf{g_y}, \mathbf{x} - \mathbf{y} \rangle.$$

After summarization of the last two inequalities, we can obtain (24). □

2.4 Subgradient method

To solve the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

we can use subgradient method or subgradient method with constrain.

Definition 7. *Subgradient method*

$$\begin{cases} \mathbf{g_k} \in \partial f(\mathbf{x}_k), \\ \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g_k}. \end{cases} \tag{25}$$

Definition 8. *Subgradient method with constrain*

$$\begin{cases} \mathbf{g_k} \in \partial f(\mathbf{x}_k), \\ \mathbf{x}_{k+1} = \mathbf{P_C x}(\mathbf{x}_k - \alpha_k \mathbf{g_k}). \end{cases} \tag{26}$$