Optimization Methods (CS1.404), Spring 2024
Lecture 22

Naresh Manwani

Machine Learning Lab, IIIT-H

April 18th, 2024

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Orthogonal Projection

Projection Theorem

Let C be a closed convex set and $\mathbf{x} \in \mathbb{R}^n$. Let $P_C(\mathbf{x})$ denote the orthogonal projection of $\mathbf{x}$ on the set C. Then $\mathbf{z} = P_C(\mathbf{x})$ if an only if $\mathbf{z} \in C$ and

$$(\mathbf{x} - \mathbf{z})^\top (\mathbf{y} - \mathbf{z}) \leq 0$$

for any $\mathbf{y} \in C$.

- Geometrically, it states that for a given closed and convex set C, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in C$, the angle between $\mathbf{x} - P_C(\mathbf{x})$ and $\mathbf{y} - P_C(\mathbf{x})$ is greater than or equal to 90 degrees.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Properties of Orthogonal Projection Operator

Theorem

Let C be a nonempty and closed convex set. Then

- For any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$

$$(P_C(\mathbf{v}) - P_C(\mathbf{w}))^\top (\mathbf{v} - \mathbf{w}) \geq \|P_C(\mathbf{v}) - P_C(\mathbf{w})\|^2$$

- **Nonexpansiveness:** For any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$

$$\|P_C(\mathbf{v}) - P_C(\mathbf{w})\| \leq \|\mathbf{v} - \mathbf{w})\|$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Representation of Stationarity Using Orthogonal Projection Operator

Theorem

Let f be a continuously differentiable function over a closed and convex set $C \subseteq \mathbb{R}^n$. Let $s > 0$. Then $\mathbf{x}^* \in C$ is a stationary point of

$$(P) \quad \min \ f(\mathbf{x})$$
$$s.t. \ \mathbf{x} \in C$$

if and only if

$$\mathbf{x}^* = P_C(\mathbf{x}^* - s\nabla f(\mathbf{x}^*)).$$

Note that above condition seems to depend on s, but by its equivalence to stationarity, it is essentially independent of s.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

The Gradient Projection Method

- The stationarity condition $\mathbf{x}^* = P_C(\mathbf{x}^* - s\nabla f(\mathbf{x}^*))$ motivates to solve the optimization problem

$$(P) \quad \min \ f(\mathbf{x})$$
$$s.t. \ \mathbf{x} \in C$$

Gradient Projection Method

Input: $\epsilon > 0$ (tolerance parameter)
Initialization: Pick $\mathbf{x}_0 \in C$ arbitrarily
General Steps: For $k = 0, 1, 2, \ldots$ execute the following steps:

1. Pick a stepsize t_k by a line search procedure.
2. Set $\mathbf{x}_{k+1} = P_C(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k))$.
3. If $\|\mathbf{x}_k - \mathbf{x}_{k+1}\| \leq \epsilon$, then STOP and $\mathbf{x}_{k+1}$ is the output.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

The Computation of Projections

- We now address the question, how does one compute the projection onto the convex set C.
- In general this is not a finite process. Nonetheless, for certain important convex sets C it can be done quite efficiently.

Projection onto box constraints

Let us suppose that C is given by $C := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}\}$, where $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$ with $\mathbf{l} \leq \mathbf{u}$. Then P_C can be expressed component-wise as

$$P_C(\mathbf{x})_i = \begin{cases} l_i, & \text{if } x_i \leq l_i \\ x_i, & \text{if } l_i \leq x_i \leq u_i \\ u_i, & \text{if } x_i \geq u_i \end{cases}$$

Projection onto a Polyhedron

- Let C be the polyhedron given by
  $C := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}_i^\top \mathbf{x} \leq b_i,\ i = 1, ..., m_1;\ \mathbf{p}_j^\top \mathbf{x} = q_j,\ j = 1, ..., m_2\}$.
- Then P_C is determined by solving the quadratic program

$$\begin{aligned} &\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{y}\|^2 \\ &s.t.\ \mathbf{a}_i^\top \mathbf{x} \leq b_i,\ i = 1, ..., m_1 \\ &\quad \mathbf{p}_j^\top \mathbf{x} = q_j,\ j = 1, ..., m_2 \end{aligned}$$

Projection onto l_2-ball with center $\mathbf{0}$ and radius 1

- Let $C = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq 1\}$.
- if $\mathbf{y} \in C$, then $P_C(\mathbf{y}) = \mathbf{y}$.
- If $\mathbf{y} \notin C$, then $\|\mathbf{y}\| > 1$. In that case, the closest point of $\mathbf{y}$ in the set C is $\frac{\mathbf{y}}{\|\mathbf{y}\|}$ as l_2 norm of $\frac{\mathbf{y}}{\|\mathbf{y}\|}$ is 1.
- Combining the two, we get

$$P_C(\mathbf{y}) = \begin{cases} \mathbf{y}, & \text{if } \|\mathbf{y}\| \leq 1 \\ \frac{\mathbf{y}}{\|\mathbf{y}\|}, & \text{if } \|\mathbf{y}\| > 1 \end{cases} = \mathbf{y} \begin{cases} \frac{1}{1}, & \text{if } \|\mathbf{y}\| \leq 1 \\ \frac{1}{\|\mathbf{y}\|}, & \text{if } \|\mathbf{y}\| > 1 \end{cases} = \frac{\mathbf{y}}{\max(1, \|\mathbf{y}\|)}$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Projection onto l_2-ball with center $\mathbf{0}$ with radius r

- Let $C = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq r\}$.
- if $\mathbf{y} \in C$, then $P_C(\mathbf{y}) = \mathbf{y}$.
- If $\mathbf{y} \notin C$, then $\|\mathbf{y}\| > r$. In that case, the closest point of $\mathbf{y}$ in the set C is $\frac{r\mathbf{y}}{\|\mathbf{y}\|}$ as l_2 norm of $\frac{\mathbf{y}}{\|\mathbf{y}\|}$ is r.
- Combining the two, we get

$$P_C(\mathbf{y}) = \begin{cases} \mathbf{y}, & \text{if } \|\mathbf{y}\| \leq r \\ \frac{r\mathbf{y}}{\|\mathbf{y}\|}, & \text{if } \|\mathbf{y}\| > r \end{cases} = r\mathbf{y} \begin{cases} \frac{1}{r}, & \text{if } \|\mathbf{y}\| \leq 1 \\ \frac{1}{\|\mathbf{y}\|}, & \text{if } \|\mathbf{y}\| > 1 \end{cases} = \frac{r\mathbf{y}}{\max(r, \|\mathbf{y}\|)}$$

Projection onto l_2-ball with center $\mathbf{c}$ with radius 1

- Let $C = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{c}\| \leq 1\}$.
- Let $\mathbf{z} = \mathbf{y} - \mathbf{c}$ and $\mathbf{u} = \mathbf{x} - \mathbf{c}$. Thus,

$$\arg \min_{\mathbf{x}:\|\mathbf{x}-\mathbf{c}\|\leq 1} \|\mathbf{y} - \mathbf{x}\|^2 = \arg \min_{\mathbf{u}:\|\mathbf{u}\|\leq 1} \|\mathbf{z} - \mathbf{u}\|^2 = \frac{\mathbf{z}}{\max(1, \|\mathbf{z}\|)}$$

- Using $\mathbf{y} = \mathbf{z} + \mathbf{c}$, we have $\mathbf{y}_{sol} = \mathbf{z}_{sol} + \mathbf{c}$, we have

$$\arg \min_{\mathbf{x}:\|\mathbf{x}-\mathbf{c}\|\leq 1} \|\mathbf{y} - \mathbf{x}\|^2 = \frac{\mathbf{z}}{\max(1, \|\mathbf{z}\|)} + \mathbf{c} = \frac{\mathbf{y} - \mathbf{c}}{\max(1, \|\mathbf{y} - \mathbf{c}\|)} + \mathbf{c}$$

Projection onto l_2-ball with center $\mathbf{c}$ with radius r

- Let $C = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{c}\| \leq r\}$.
- Let $\mathbf{z} = \mathbf{y} - \mathbf{c}$ and $\mathbf{u} = \mathbf{x} - \mathbf{c}$. Thus,

$$\arg \min_{\mathbf{x}:\|\mathbf{x}-\mathbf{c}\|\leq r} \|\mathbf{y} - \mathbf{x}\|^2 = \arg \min_{\mathbf{u}:\|\mathbf{u}\|\leq r} \|\mathbf{z} - \mathbf{u}\|^2 = \frac{r\mathbf{z}}{\max(r, \|\mathbf{z}\|)}$$

- Using $\mathbf{y} = \mathbf{z} + \mathbf{c}$, we have $\mathbf{y}_{sol} = \mathbf{z}_{sol} + \mathbf{c}$, we have

$$\arg \min_{\mathbf{x}:\|\mathbf{x}-\mathbf{c}\|\leq r} \|\mathbf{y} - \mathbf{x}\|^2 = \frac{r\mathbf{z}}{\max(r, \|\mathbf{z}\|)} + \mathbf{c} = \frac{r(\mathbf{y} - \mathbf{c})}{\max(r, \|\mathbf{y} - \mathbf{c}\|)} + \mathbf{c}$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Sufficient Decrease Lemma for Projected Gradient Descent

Lemma

Suppose that $f \in \mathbb{C}^{1,1}(C)$, where C is a closed convex set. Then for any $\mathbf{x} \in C$ and $t \in (0, 2/L)$, the following inequality will hold.

$$f(\mathbf{x}) - f(P_C(\mathbf{x} - t\nabla f(\mathbf{x}))) \geq t\left(1 - \frac{Lt}{2}\right) \|\frac{1}{t}(\mathbf{x} - P_C(\mathbf{x} - t\nabla f(\mathbf{x})))\|^2.$$

- When $C = \mathbb{R}^n$, the obtained inequality is exactly the same as the one obtained for unconstrained case.

Gradient Mapping

- We define gradient mapping as

$$G_M(\mathbf{x}) = M\left[\mathbf{x} - P_C\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right)\right]$$

  where $M > 0$.
- When $C = \mathbb{R}^n$, we see that $G_M(\mathbf{x}) = \nabla f(\mathbf{x})$. So, the gradient mapping is an extension of usual gradient operation.
- We see that $G_M(\mathbf{x}) = \mathbf{0}$ if and only $\mathbf{x} = P_C\left(\mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})\right)$. This happens if and only if $\mathbf{x}$ is a stationary point of optimization problem (P).
- Thus, we can use $\|G_M(\mathbf{x})\|$ as an optimality measure.
- Thus, the sufficient decrease property of projected gradient method essentially states that

$$f(\mathbf{x}) - f(P_C(\mathbf{x} - t\nabla f(\mathbf{x}))) \geq t\left(1 - \frac{Lt}{2}\right)\|G_{\frac{1}{t}}(\mathbf{x})\|^2.$$

Monotonicity Property of Norm of Gradient Mapping $G_M(\mathbf{x})$

Lemma

Let f be a continuously differentiable function defined on a nonempty closed convex set C. Suppose that $L_1 \geq L_2 > 0$. Then,

$$\|G_{L_1}(\mathbf{x})\| \geq \|G_{L_2}(\mathbf{x})\|$$

and

$$\frac{\|G_{L_1}(\mathbf{x})\|}{L_1} \leq \frac{\|G_{L_2}(\mathbf{x})\|}{L_2}$$

for any $\mathbf{x} \in \mathbb{R}^n$.

See Lemma 9.12 of **Amir Beck's** book.

Backtracking

- The procedure requires three parameters s, α, β where $s > 0, \alpha \in (0,1)$ and $\beta \in (0,1)$.
- The choice of t_k is done as follows. First, t_k is set to be equal to the initial guess s.
- Then while

$$f(\mathbf{x}_k) - f(P_C(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k))) < t_k \| G_{\frac{1}{t_k}}(\mathbf{x}_k) \|^2.$$

  we set $t_k = \beta t_k$.
- In other words, the step size is chosen as $t_k = s\beta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$f(\mathbf{x}_k) - f(P_C(\mathbf{x}_k - s\beta^{i_k} \nabla f(\mathbf{x}))) \geq s\beta^{i_k} \| G_{\frac{1}{s\beta^{i_k}}}(\mathbf{x}) \|^2.$$

  is satisfied.

Backtracking for $f \in \mathbb{C}_L^{1,1}(C)$

- If $f \in \mathbb{C}_L^{1,1}(C)$, then for $\mathbf{x} = \mathbf{x}_k$, we can write

$$f(\mathbf{x}_k) - f(P_C(\mathbf{x}_k - t\nabla f(\mathbf{x}_k))) \geq t\left(1 - \frac{Lt}{2}\right)\|G_{\frac{1}{t}}(\mathbf{x}_k)\|^2.$$

- For any $t \leq \frac{2(1-\alpha)}{L}$, the following inequality holds.

$$f(\mathbf{x}_k) - f(P_C(\mathbf{x}_k - t\nabla f(\mathbf{x}_k))) \geq \alpha t_k\|G_{\frac{1}{t}}(\mathbf{x}_k)\|^2.$$

- This implies that the backtracking procedure ends when t_k is smaller or equal to $\frac{2(1-\alpha)}{L}$.
- Note that the backtracking procedure is finite for $\mathbb{C}^{1,1}$ functions.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Lower bound on t_k in Backtracking

- We can find lower bound on t_k as follows.
- Either t_k is equal to s or the backtracking procedure is invoked, meaning that the stepsize $\frac{t_k}{\beta}$ did not satisfy the backtracking condition.
- Using the property that $t_k \leq \frac{2(1-\alpha)}{L}$, we have $\frac{t_k}{\beta} > \frac{2(1-\alpha)}{L}$. Thus, $t_k > \frac{2\beta(1-\alpha)}{L}$.
- To summarize, in the backtracking procedure, the chosen stepsize t_k satisfies

$$t_k \geq \min\left(s, \frac{2\beta(1-\alpha)}{L}\right).$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD