SDSC6015 - Semester A, 2025

Stochastic Optimization for Machine Learning

5. Projected Subgradient Method

1 Motivation

Recall that using gradient descent, one can solve an unconstrained, smooth convex optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ f(\boldsymbol{x}),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable with $\mathbf{dom}(f) = \mathbb{R}^n$. Gradient descent is a line search method in which for each iteration, we do

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \cdot \boldsymbol{d}_k$$

where

- $k \in \mathbb{Z}_+$: the index of iterations
- $\boldsymbol{x}_k$: solution at the k^{th} iteration
- $\boldsymbol{d}_k$: direction at the k^{th} iteration
- α_k: stepsize at the k^{th} iteration

Limitation: What if

- We have constraints?
- The function f is not differentiable?

Answer: Projected Subgradient Method!

- We have constraints? Projection.
- The function f is not differentiable? Subgradient.

2 Subgradients: Definitions and Properties

Definition 1 (Proper function)**.** *A function $f : \mathbb{R}^n \to (-\infty, \infty]$ is called proper if $dom(f) \neq \emptyset$.*

Definition 2 (Subgradient)**.** *Assume $f : \mathbb{R}^n \to (-\infty, \infty]$ is a proper convex function and $\boldsymbol{x}_0 \in \boldsymbol{dom}(f)$. Then, $\boldsymbol{y} \in \mathbb{R}^n$ is called a subgradient of f at $\boldsymbol{x}_0$ if*

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \boldsymbol{y}^\top (\boldsymbol{x} - \boldsymbol{x}_0), \quad \forall \boldsymbol{x} \in \boldsymbol{dom}(f).$$

The set $\partial f(\boldsymbol{x}_0)$ of all subgradients is the subdifferential of f at $\boldsymbol{x}_0$.

Note:

- If f is differentiable at $\boldsymbol{x}_0$, then $\partial f(\boldsymbol{x}_0) = \{\nabla f(\boldsymbol{x}_0)\}$.
- Generally, $\partial f(\boldsymbol{x}_0)$ can be empty or contain one or infinitely many subgradients.

- $\partial f(\boldsymbol{x}_0)$ is a closed convex set.

2.1 Some Examples

Example 1 (Subdifferential of norms at $\mathbf{0}$)**.** *Consider* $f : \mathbb{R}^n \to \mathbb{R}$ *and* $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$*, then*

$$\partial f(\mathbf{0}) = B_{\|\cdot\|_\star}[\mathbf{0}, 1] = \{\boldsymbol{y} \in \mathbb{R}^n \mid \|\boldsymbol{y}\|_\star \leq 1\},$$

where $\|\cdot\|_\star$ *is the dual norm of* $\|\cdot\|$*.*

Example 2 (Subdifferential of the L_1-norm at $\mathbf{0}$)**.** *Consider* $f : \mathbb{R}^n \to \mathbb{R}$ *and* $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$*, then*

$$\partial f(\mathbf{0}) = B_{\|\cdot\|_\infty}[\mathbf{0}, 1] = [-1, 1]^n.$$

Example 3 (Subdifferential of the L_2-norm)**.** *Consider $f : \mathbb{R}^n \to \mathbb{R}$ and $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$, then*

$$\partial f(\mathbf{x}) = \begin{cases} \left\{ \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \right\}, & \boldsymbol{x} \neq \mathbf{0}, \\ B_{\|\cdot\|_2}[\mathbf{0}, 1], & \boldsymbol{x} = \mathbf{0}, \end{cases}$$

Example 4 (Subdifferential of the indicator functions)**.** *Consider $S \subseteq \mathbb{R}^n$ is nonempty and the indicator function*

$$\delta_S(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in S, \\ \infty, & \boldsymbol{x} \notin S. \end{cases}$$

Then

$$\partial \delta_S(\boldsymbol{x}) = N_S(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^n \mid \langle \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{x} \rangle \leq 0,\ \forall \boldsymbol{z} \in S\}, \quad \forall x \in S.$$

For $\boldsymbol{x} \notin S$, $\partial \delta_S(\boldsymbol{x}) = N_S(\boldsymbol{x}) = \emptyset$ by convention. Note that $N_S(\boldsymbol{x})$ is called the normal cone of S at $\boldsymbol{x}$.

Example 5 (Subgradient of the dual function)**.** *Consider the minimization problem*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\{f(\boldsymbol{x}) \mid \boldsymbol{h}(\boldsymbol{x}) \leq \mathbf{0},\ \boldsymbol{x}\in\mathcal{X}\},$$

where $\emptyset \neq \mathcal{X}$*,* $f:\mathbb{R}^n\to\mathbb{R}$*, and* $\boldsymbol{h}:\mathbb{R}^n\to\mathbb{R}^m$*. The Lagrange dual function is*

$$g(\boldsymbol{\lambda}) = \min_{\boldsymbol{x}\in\mathcal{X}}\left\{L(\boldsymbol{x},\boldsymbol{\lambda}) \equiv f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{h}(\boldsymbol{x})\right\}.$$

2.2 Properties

Recall the following results from our lecture on Convex Sets!

Definition 3 (Supporting hyperplane)**.** *Consider a nonempty set $C \subseteq \mathbb{R}^n$ and a boundary point $\boldsymbol{x}_0 \in \boldsymbol{bd}(C)$. If $\boldsymbol{a} \neq \boldsymbol{0}$ in $\mathbb{R}^n$ satisfies $\boldsymbol{a}^\top \boldsymbol{x} \leq \boldsymbol{a}^\top \boldsymbol{x}_0$, $\forall \boldsymbol{x} \in C$, then $\{\boldsymbol{x} \mid \boldsymbol{a}^\top \boldsymbol{x} = \boldsymbol{a}^\top \boldsymbol{x}_0\}$ is called a supporting hyperplane to C at $\boldsymbol{x}_0$.*

Theorem 1 (Weak Separating Hyperplane Theorem)**.** *Consider any convex set $C \subseteq \mathbb{R}^n$ and a point $\boldsymbol{x}_0 \in \mathbb{R}^n \backslash C$. Then, there exist $\boldsymbol{a} \neq \boldsymbol{0}$ (in $\mathbb{R}^n$) and $b \in \mathbb{R}$ with*

$$\boldsymbol{a}^\top \boldsymbol{x} \leq b, \ \forall \boldsymbol{x} \in C, \qquad \text{and} \qquad \boldsymbol{a}^\top \boldsymbol{x}_0 \geq b.$$

Theorem 2. *For every convex set $C \subseteq \mathbb{R}^n$ and any $\boldsymbol{x}_0 \in \boldsymbol{bd}(C)$ there is a supporting hyperplane to C at $\boldsymbol{x}_0$.*

Theorem 3. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a proper convex function and $\boldsymbol{x}_0 \in \boldsymbol{int}(\boldsymbol{dom}(f))$. Then, $\partial f(\boldsymbol{x}_0) \neq \emptyset$.*

3 Computing Subgradients

There are rules that are useful for computing subgradients and subdifferentials. When the rules/results are useful for computing some of the subgradients in the subdifferential set, they are referred as "*weak results*"; when the rules/results are useful for obtaining the full characterization of the subdifferential set, they are referred as "*strong results*".

3.1 Multiplication by a positive scalar

Theorem 4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper function and let $\alpha > 0$. Then for any $\boldsymbol{x} \in dom(f)$*

$$\partial(\alpha f)(\boldsymbol{x}) = \alpha \partial f(\boldsymbol{x}).$$

Recall that for a scalar $\beta \in \mathbb{R}$ and a set $S \subseteq \mathbb{R}^n$, the set $\beta S = \{\beta \boldsymbol{x} \mid \boldsymbol{x} \in S\}$.

3.2 Summation

Theorem 5. *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be proper convex functions, and let $\boldsymbol{x} \in dom(f_1) \cap dom(f_2)$.*

(a) The following inclusion holds:

$$\partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x}) \subseteq \partial(f_1 + f_2)(\boldsymbol{x}).$$

(b) If $\boldsymbol{x} \in int(dom(f_1)) \cap int(dom(f_2))$, then

$$\partial(f_1 + f_2)(\boldsymbol{x}) = \partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x}).$$

Recall that for two sets $A, B \subseteq \mathbb{R}^n$ that reside in the same space, the sum $A + B$ stands for the Minkowski sum given by $A + B = \{\boldsymbol{a} + \boldsymbol{b} \mid \boldsymbol{a} \in A,\ \boldsymbol{b} \in B\}$.

Example 6 (L_1-norm function)**.** *Consider the function* $f : \mathbb{R}^n \to \mathbb{R}$ *and* $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$*. What is the subdifferential? Could you compute any subgradient?*

3.3 Affine Transformation

Theorem 6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper convex function and $h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ for some $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Assume that h is proper, meaning that*

$$dom(h) = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \in dom(f)\} \neq \emptyset.$$

(a) For any $\boldsymbol{x} \in dom(h)$,

$$\boldsymbol{A}^\top(\partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})) \subseteq \partial h(\boldsymbol{x}).$$

(b) If $\boldsymbol{x} \in int(dom(h))$ and $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \in int(dom(f))$, then

$$h(\boldsymbol{x}) = \boldsymbol{A}^\top(\partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})).$$

Note that $\boldsymbol{A}^\top S = \{\boldsymbol{A}^\top \boldsymbol{x} \mid \boldsymbol{x} \in S\}$.

Example 7. *Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ and $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\|_1$. What is the subdifferential?*

3.4 Composition

Theorem 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $g : \mathbb{R} \to \mathbb{R}$ be a nondecreasing convex function. Let $\boldsymbol{x} \in \mathbb{R}^n$ and suppose that g is differentiable at the point $f(\boldsymbol{x})$. Let $h = g \circ f$. Then*

$$\partial h(\boldsymbol{x}) = g'(f(\boldsymbol{x}))\partial f(\boldsymbol{x}).$$

Example 8. *Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ and $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_1^2$.*

3.5 Maximization

Theorem 8. *Let $f_1, f_2, f_3, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ be proper convex functions, and define*

$$f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\}.$$

Let $\boldsymbol{x} \in \cap_{i=1}^m int(dom(f_i))$. Then

$$\partial f(\boldsymbol{x}) = conv(\cup_{i \in I(\boldsymbol{x})} \partial f_i(\boldsymbol{x})),$$

where $I(\boldsymbol{x}) = \{i \in [m] \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x})\}$

Example 9. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be given by* $f(\boldsymbol{x}) = \max\{x_1, x_2, \ldots, x_n\}$.

Example 10. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be given by* $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$.

4 Orthogonal Projection

To handle constraints, we make use of the orthogonal projection.

Definition 4. *Let $C \subseteq \mathbb{R}^n$ be nonempty closed and convex set. The orthogonal projection mapping*

$$P_C(\boldsymbol{x}) \equiv \arg\min_{\boldsymbol{y} \in C} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

Theorem 9. *Let $C \subseteq \mathbb{R}^n$ be nonempty, closed, and convex. Then, for every $\boldsymbol{x}_0 \in \mathbb{R}^n$, there exists a unique point $\boldsymbol{x}^\star \in C$ that is closest (in the Euclidean norm) to $\boldsymbol{x}_0$.*

4.1 Examples

1. If $C = \mathbb{R}^n_+$, $P_C(\boldsymbol{x}) = \max\{\boldsymbol{x}, \boldsymbol{0}\} = [x]_+$

2. If $C = \mathrm{Box}[\boldsymbol{\ell}, \boldsymbol{u}] = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{\ell} \leq \boldsymbol{x} \leq \boldsymbol{u}\}$, $P_C(\boldsymbol{x}) = (\min\{\max\{x_i, \ell_i\}, u_i\})_{i=1}^n$

3. If $C = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ ($\boldsymbol{A}$ has full row rank), $P_C(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$

4. If $C = B_{\|\cdot\|_2}[\boldsymbol{c}, r]$, $P_C(\boldsymbol{x}) = \boldsymbol{c} + \dfrac{r}{\max\{\|\boldsymbol{x} - \boldsymbol{c}\|_2, r\}}(\boldsymbol{x} - \boldsymbol{c})$

5. If $C = \{\boldsymbol{x} \mid \boldsymbol{a}^\top\boldsymbol{x} \leq \alpha\}$, $P_C(\boldsymbol{x}) = \boldsymbol{x} - \dfrac{[\boldsymbol{a}^\top\boldsymbol{x} - \alpha]_+}{\|\boldsymbol{a}\|^2}\boldsymbol{a}$

5 Projected Subgradient Method

Consider the following optimization model

$$\min_{\boldsymbol{x}\in C} f(\boldsymbol{x}),$$

where (we assume)

- $f : \mathbb{R}^n \to \mathbb{R}$ is proper closed and convex.
- $C \subseteq \mathbb{R}^n$ is nonempty close and convex.
- $C \subseteq \text{int}(\text{dom}(f))$
- The optimal set is nonempty and denoted by $\mathcal{X}^\star$. The optimal value of the problem is denoted by $f^\star$.
- There exists a constant $L_f > 0$ for which $\|\boldsymbol{y}\| \leq L_f$ for all $\boldsymbol{y} \in \partial f(\boldsymbol{x})$, $\boldsymbol{x} \in C$.

Projected Subgradient Method

Initialization: $\boldsymbol{x}_0 \in C$

General step: For $k = 0, 1, 2, \ldots$

(a) pick a stepsize $\alpha_k > 0$ and a subgradient $f'(\boldsymbol{x}_k) \in \partial f(\boldsymbol{x}_k)$.

(b) set $\boldsymbol{x}_{k+1} = P_C(\boldsymbol{x}_k - \alpha_k f'(\boldsymbol{x}_k))$.

The sequene generated by the projected subgradient method is $\{\boldsymbol{x}_k\}_{k\geq 0}$ while the sequence of function values generated by the method is $\{f(\boldsymbol{x}_k)\}_{k\geq 0}$. This sequence of functions values is not necessarily monotone. Therefore, we would be interested in

$$f^k_{\text{best}} \equiv \min_{n=0,\ldots,k} f(\boldsymbol{x}_n).$$

Lemma 1. *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the projected subgradient method. Then for any $\boldsymbol{x}^\star \in \mathcal{X}^\star$ and $k \geq 0$,*

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|^2 - 2\alpha_k(f(\boldsymbol{x}_k) - f^\star) + \alpha_k^2\|f'(\boldsymbol{x}_k)\|^2.$$

5.1 Polyak's Stepsize

The Polyak's stepsize rule takes stepsize α_k that minimize the RHS of the result in Lemma 1, and so

$$\alpha_k = \begin{cases} \dfrac{f(\boldsymbol{x}_k) - f^\star}{\|f'(\boldsymbol{x}_k)\|^2}, & f'(\boldsymbol{x}_k) \neq \mathbf{0}, \\ 1, & f'(\boldsymbol{x}_k) = \mathbf{0}. \end{cases}$$

Theorem 10. *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the projected subgradient method with Polyak's stepsize rule. Then*

- $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|^2$ *for any* $k \geq 0$ *and* $\boldsymbol{x}^\star \in \mathcal{X}^\star$*;*
- $f(\boldsymbol{x}_k) \to f^\star$ *as* $k \to \infty$*;*
- $f_{best}^k - f^\star \leq \dfrac{L_f d_{\mathcal{X}^\star}(\boldsymbol{x}_0)}{\sqrt{k+1}}$ *for any* $k \geq 0$*. [$d_{\mathcal{X}^\star}$ is the distance function.]*

5.2 Dynamic Stepsizes

Theorem 11. *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the projected subgradient method with positive stepsize $\{\alpha_k\}_{k\geq 0}$. If*

$$\frac{\sum_{n=0}^k \alpha_n^2}{\sum_{n=0}^k \alpha_n} \to 0 \quad \textit{as} \quad k \to \infty,$$

then

$$f_{best}^k \to f^\star \quad \textit{as} \quad k \to \infty.$$

Possible choice: $\alpha_k = 1/\sqrt{k+1}$.

Theorem 12. *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the projected subgradient method with positive stepsizes $\alpha_k = 1/(\|f'(\boldsymbol{x}_k)\|\sqrt{k+1})$ if $f'(\boldsymbol{x}_k) \neq 0$ and $\alpha_k = 1/L_f$ otherwise. Then for any $k \geq 1$*

$$f_{best}^k - f^\star \leq \frac{L_f}{2} \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 + 1 + \log(k+1)}{\sqrt{k+1}}$$

and

$$f(\boldsymbol{x}_{(k)}) - f^\star \leq \frac{L_f}{2} \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 + 1 + \log(k+1)}{\sqrt{k+1}}$$

where $\boldsymbol{x}_{(k)} = \dfrac{1}{\sum_{n=0}^k \alpha_n} \displaystyle\sum_{n=0}^k \alpha_n \boldsymbol{x}_n$.

6 Special Case: Convex Feasibility Problem

Let $S_1, S_2, \ldots, S_m \in \mathbb{R}^n$ be close and convex sets. Assume that

$$S \equiv \cap_{i=1}^m S_i \neq \emptyset.$$

The *convex feasibility problem* is the problem of finding a point $\boldsymbol{x}$ in the intersection $\cap_{i=1}^m S_i$. One can formulate this problem as

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \left\{ f(\boldsymbol{x}) \equiv \max_{i \in [m]} d_{S_i}(\boldsymbol{x}) \right\},$$

where $d_{S_i}(\boldsymbol{x}) = \min_{\boldsymbol{y} \in \mathbb{R}^n} \{ \|\boldsymbol{x} - \boldsymbol{y}\| \mid \boldsymbol{y} \in S_i \}$, for all $i \in [m]$. We assume that the intersection is nonempty, and we have $f^\star = 0$. One can show that $L_f = 1$.

Greedy Projection Algorithm

Input: m nonempty closed and convex sets $S_1, S_2, \ldots, S_m$

Initialization: pick $\boldsymbol{x}_0 \in \mathbb{R}^n$.

General step: For $k = 0, 1, 2, \ldots$, compute

$$\boldsymbol{x}_{k+1} = P_{S_{i_k}}(\boldsymbol{x}_k),$$

where $i_k \in \arg\max_{i \in [m]} d_{S_i}(\boldsymbol{x}_k)$.

When $m = 2$, the algorithm amounts the alternating projection method.

Alternating Projection Method

Input: two nonempty closed and convex sets S_1, S_2.

Initialization: pick $\boldsymbol{x}_0 \in S_2$ arbitrarily.

General step: For $k = 0, 1, 2, \ldots$, compute

$$\boldsymbol{x}_{k+1} = P_{S_2}(P_{S_1}(\boldsymbol{x}_k)).$$

Special Case: Consider the system of linear equalities and inequalities (standard form of LP):

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0},$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has full row rank and $\boldsymbol{b} \in \mathbb{R}^m$. Define

$$S_1 = \{\boldsymbol{x} \in \mathbb{R}^m \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}, \quad S_2 = \mathbb{R}^n_+.$$

As we know, we have

$$P_{S_1}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}), \quad P_{S_2}(\boldsymbol{x}) = [\boldsymbol{x}]_+.$$

Alternating Projection Method (System of Linear Equalities and Inequalities)

Initialization: pick $\boldsymbol{x}_0 \in \mathbb{R}^n_+$ arbitrarily.

General step: For $k = 0, 1, 2, \ldots$, compute

$$\boldsymbol{x}_{k+1} = \Big[\boldsymbol{x}_k - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b})\Big]_+$$