Choice of Stepsize and Projected Gradient

SIE 449/549: Optimization for Machine Learning

Afrooz Jalilzadeh

Department of Systems and Industrial Engineering
University of Arizona

Gradient Method

- Consider the following optimization problem:

$$\min_x f(x)$$

- Gradient Step: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$
- Assume f has a Lipschitz continuous gradient with constant L
- Constant Stepsize: $\alpha_k = \alpha$, such that $\alpha < 2/L$
- If $\alpha = 1/L$ and f be a convex function, then

$$f(x_T) - f(x^*) \leq \frac{L}{2T} \|x_0 - x^*\|^2, \text{ for all } T \geq 1$$

- How many steps should we take to get to an ϵ-suboptimality?

Choices of Stepsize

1. **Constant Stepsize:** $\alpha_k = \alpha < 2/L$ and if f is convex, $\alpha = 1/L$
   - We need to know Lipschitz constant L
2. **Backtracking for Convex Function:** Consider $\beta \in (0, 1)$, start with an initial step-size $\alpha_k = \alpha$. Then, while

$$\|\nabla f(x_{k+1}) - \nabla f(x_k)\| > \frac{1}{\alpha_k}\|x_{k+1} - x_k\|$$

$\alpha_k \leftarrow \beta\alpha_k$.

Choices of step size

3. **Diminishing Stepsize:** Decrease α_k in each iteration:
   - Intuitively, as the algorithm runs, we will get closer and closer to the optimal point and it might be better to move less in case we miss the optimal point
   - α_k satisfies the following two conditions

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

   - Why we need $\sum_{k=1}^{\infty} \alpha_k = \infty$?
     - Because we may get stuck in certain region and never reach optimum, it allows us to explore the entire space

Choices of step size

4. **Line Search:** Find the “best" α_k that minimize f along the direction of d_k at each iteration as follows
$$\alpha_k \in \text{argmin}_{\alpha > 0} f(x_k + \alpha_k d_k)$$
   - When $d_k = -\nabla f(x_k)$: $\alpha_k \in \text{argmin}_{\alpha > 0} f\left(x_k - \alpha_k \nabla f(x_k)\right)$
   - It can be costly to search for the optimal α

Choices of step size

4. **Line Search:** Find the “best" α_k that minimize f along the direction of d_k at each iteration as follows

$$\alpha_k \in \text{argmin}_{\alpha>0} f(x_k + \alpha_k d_k)$$

- When $d_k = -\nabla f(x_k)$: $\alpha_k \in \text{argmin}_{\alpha>0} f\left(x_k - \alpha_k \nabla f(x_k)\right)$
- It can be costly to search for the optimal α

5. **Backtracking-Armijo:** Iteratively shrink α_k until the decrease in $f(x_k) - f(x_{k+1})$, adequately matches the decrease that is expected to be achieved:

- Given constants $\sigma \in (0, 1)$ and $\beta \in (0, 1)$ and initial stepsize s, while

$$f(x_k) - f(x_k - \alpha_k \nabla f(x_k)) < \sigma \alpha_k \|\nabla f(x_k)\|^2$$

set $\alpha_k \leftarrow \beta \alpha_k$;

Optimization over a Convex Set

- Consider the following constrained optimization problem

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in C \end{aligned} \tag{P}$$

where C is a closed convex subset of $\mathbb{R}^n$ and f is continuously differentiable over C

Definition 1 (Stationary Point)

Let f be a continuously differentiable function over a closed and convex set C. Then x^* is called a **stationary point** of (P) if

$$\nabla f(x^*)^T(x - x^*) \geq 0, \text{ for all } x \in C$$

Stationarity as a Necessary Optimality Condition

Theorem 2

Let f be a continuously differentiable function over a nonempty closed convex set C, and let x^* be a local minimum of (P). Then x^* is a stationary point of (P).

Example

- Show that when $C = \mathbb{R}^n$, then x^* is a stationary point if $\nabla f(x^*) = 0$.

Stationarity in Convex Optimization

- For convex problems, stationarity is a necessary and sufficient condition

Theorem 3

Let f be a continuously differentiable convex function over a nonempty closed and convex set $C \subseteq R^n$. Then x^* is a stationary point of

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in C \end{aligned} \tag{P}$$

iff x^* is an optimal solution of (P).

The Orthogonal Projection Operator

Definition 4

Given a nonempty closed convex set C, the orthogonal projection operator $P_C : \mathbb{R}^n \to C$ is defined by

$$P_C(x) = \text{argmin}\{\|y - x\|^2 : y \in C\}$$

- The first important result is that the orthogonal projection exists and is unique.

Theorem 5 (The First Projection Theorem)

Let $C \subseteq R^n$ be a nonempty closed and convex set. Then for any $x \in \mathbb{R}^n$, the orthogonal projection $P_C(x)$ exists and is unique.

Examples

- $C = \mathbb{R}^n_+$

- A box is a subset of R^n: $C = [\ell_1, u_1] \times \ldots \times [\ell_n, u_n] = \{x \in \mathbb{R}^n : \ell_i \leq x_i \leq u_i\}$

- $C = B[0, r]$

The Second Projection Theorem

Theorem 6 (The Second Projection Theorem)

Let $C \subseteq R^n$ be a nonempty closed and convex set and let $x \in \mathbb{R}^n$. Then, $z = P_C(x)$ if and only if

$$(x - z)^T (y - z) \leq 0, \text{ for any } y \in C$$

Properties of the Orthogonal Projection

Theorem 7

Let C be a nonempty closed and convex set. Then

1. *For any $v, w \in \mathbb{R}^n$:*

$$(P_C(v) - P_C(w))^T (v - w) \geq \|P_C(v) - P_C(w)\|^2$$

2. *(non-expansiveness) For any $v, w \in \mathbb{R}^n$:*

$$\|P_C(v) - P_C(w)\| \leq \|v - w\|$$

Representation of Stationarity via the Orthogonal Projection Operator

Theorem 8

Let f be a continuously differentiable function over the nonempty closed convex set C, and let $s > 0$. Then x^* is a stationary point of

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in C \end{aligned} \tag{P}$$

if and only if

$$x^* = P_C(x^* - s\nabla f(x^*))$$

Projected Gradient Method

- From Theorem 8, x_k is a stationary point iff $\|P_C(x_k - s\nabla f(x_k)) - x_k\| = 0$
- x_k is an ϵ-stationary point iff $\|P_C(x_k - s\nabla f(x_k)) - x_k\| \leq \epsilon$

Algorithm 1 Projected Gradient Method

Initialization: pick $x_0 \in \mathbb{R}^n$ arbitrarily
for $k = 0, 1, 2, \ldots$ **do**
  find a stepsize α_k satisfying $f(x_k + \alpha_k d_k) < f(x_k)$
  set $x_{k+1} = P_C\left(x_k - \alpha_k \nabla f(x_k)\right)$
  if $\|x_{k+1} - x_k\| \leq \epsilon$ then STOP and x_{k+1} is the output
end for

- When f is convex and has a Lipschitz gradient, then choose $\alpha_k = 1/L$
- One can show that $f(x_T) - f(x^*) \leq \mathcal{O}(1/T)$ for all $T \geq 1$