COMM 616: Modern Optimization with Applications in ML and OR 2024-25 Fall

Lecture 7: (Sub)Gradient Descent under Different Conditions

Instructor: Jiajin Li   Scribe: Zhuyu Liu

1 Problem Introduction

We are interested in solving the following unconstrained optimization problem:

$$\inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \tag{P}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ and $\inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) > -\infty$.

1.1 Optimization Algorithms

Optimization algorithms are typically iterative procedures. Starting from an initial point $\boldsymbol{x}^0$, they generate a sequence $\{\boldsymbol{x}^k\}_{k\geq 0}$ of iterates designed to converge to a solution, such as a global or local minimum, a stationary point, or a KKT point.

A generic algorithm: A point to set mapping in a subspace of $\mathbb{R}^d$.

Definition 1. *Let $\mathcal{H}$ be an algorithmic mapping defined over $\mathbb{R}^d$ and the sequence $\boldsymbol{x}^0\in\mathbb{R}^d$ starting from a given point $\boldsymbol{x}^0$ be generated from*

$$\boldsymbol{x}^{k+1} = \mathcal{H}(\boldsymbol{x}^k)^{1}.$$

In this class, we focus exclusively on first-order optimization algorithms, where the algorithmic mapping $\mathcal{H}$ relies solely on the (sub)gradient information of f at the current iterates.

1.2 Optimality Conditions and Residual Functions

Definition 2. *We define the residual function $R:\mathbb{R}^d\to\mathbb{R}_+$ with the following properties:*

(i) The function $R(x):\mathbb{R}^d\to\mathbb{R}_+$ is continuous.

(ii) The condition $R(x)=0$ holds if and only if x is the solution.

Typically, we are trying to etablish $R(\boldsymbol{x}^k)\to 0$ in the optimization literature. The conditions outlined above are essential to ensure the validity of this approach, i.e.,

$$\boxed{\lim_{k\to\infty} R(\boldsymbol{x}^k) = 0 \iff_{?} \lim_{k\to\infty} \boldsymbol{x}^k \text{ is the solution.}} \tag{Q}$$

This relationship becomes clearer through the following equation:

$$\lim_{k\to\infty} R(\boldsymbol{x}^k) = R\left(\lim_{k\to\infty} \boldsymbol{x}^k\right) = 0,$$

where the first equality follows from Definition 2 (i) and the second equality can imply that $\lim_{k\to\infty}\boldsymbol{x}^k$ is the solution from Definition 2 (ii).

[1] The algorithmic mapping can be further extended to $\mathcal{H}(\boldsymbol{x}^0, x^1, \cdots, \boldsymbol{x}^k)$.

1.3 Key Questions in the Convergence Analysis

(i) What is the **convergence rate** of $R(x^k)$? e.g.,

$$R(\boldsymbol{x}^k) \leq \mathcal{O}\left(\frac{1}{k}\right), \mathcal{O}\left(\frac{1}{\sqrt{k}}\right), \mathcal{O}\left(\exp(-k)\right), \cdots$$

(ii) Equivalently, how many iterations are required to achieve an ϵ-approximate solution, e.g., $R(\boldsymbol{x}^k) \leq \epsilon$? This is referred to as the **iteration complexity**, e.g.,

$$k = \mathcal{O}\left(\frac{1}{\epsilon^2}\right), \mathcal{O}\left(\frac{1}{\epsilon}\right), \mathcal{O}\left(\log\frac{1}{\epsilon}\right), \cdots$$

When comparing the convergence rates of different optimization algorithms, it is important to ensure that the same residual function is used.

1.4 Structure of the Problem

The structure of the problem is crucial for both algorithm design and convergence analysis. When analyzing a fixed problem, two key factors—convexity and smoothness conditions—play a vital role in establishing convergence. Convexity helps to globally control the lower bound of the function, i.e.,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \quad , \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

while smoothness ensures control over global upper bound or the curvature of the gradient, i.e.,

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \quad , \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

where L is some positive constant.

2 Smoothness and Sufficient Decrease Property

Having spent considerable time on convexity analysis, we will now get into a deeper understanding of smoothness conditions, particularly their relationship with the Sufficient Decrease Property.

Definition 3 (L-Smooth)**.** *A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be L-smooth if its gradient ∇f is L-Lipschitz, i.e.,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$$

where L is a positive constant.

L-smoothness is putting an upper bound on the curvature of the function.

Lemma 4 (Quadratic Upper Bound)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth on $\mathbb{R}^d$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, one has*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Proof. We construct a function $g : \mathbb{R} \to \mathbb{R}$, defined as:

$$g(t) = f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})).$$

Then, we have $g(0) = f(\boldsymbol{x})$ and $g(1) = f(\boldsymbol{y})$. By Fundamental Theorem of Calculus, we have

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = g(1) - g(0) = \int_0^1 g'(t)\,\mathrm{d}t.$$

Taking the derivative on $g(t)$, we have $g'(t) = \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x})$. Then, we have

$$\begin{aligned}
f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x})\,\mathrm{d}t \\
&= \int_0^1 (\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x})\,\mathrm{d}t + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \\
&\leq \int_0^1 \|\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\|\,\mathrm{d}t + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})\,\mathrm{d}t \\
&\leq \int_0^1 tL\|\boldsymbol{y} - \boldsymbol{x}\|^2 \mathrm{d}t + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})\,\mathrm{d}t \\
&= \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2,
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality is due to Definition 3. ■

The most important result of the L-smoothness property is that when we apply the standard gradient descent step, we can derive the Sufficient Decrease Property as follows:

Proposition 5 (Sufficient Decrease Property for Gradient Descent under L-Smooth Condition)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function, and let the gradient step be defined as*

$$\boldsymbol{x}^+ = \boldsymbol{x} - t \cdot \nabla f(\boldsymbol{x}), \tag{GD}$$

where $t > 0$ is the step size. Then, the following inequality holds:

$$f(\boldsymbol{x}^+) \leq f(\boldsymbol{x}) - \left(\frac{1}{t} - \frac{L}{2}\right) \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2.$$

Proof. By applying Lemma 4, we have

$$\begin{aligned}
f(\boldsymbol{x}^+) &\leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})\top(\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 \\
&= f(\boldsymbol{x}) - \frac{1}{t}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 + \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 \\
&= f(\boldsymbol{x}) - \left(\frac{1}{t} - \frac{L}{2}\right) \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2,
\end{aligned}$$

where the second equality follows from (GD). ■

Observation 6. *When $0 < t < \frac{2}{L}$, we have $f(\boldsymbol{x}^+) < f(\boldsymbol{x})$.*

3 Gradient Descent - Algorithms and Complexity [1, Chapter 3.2]

3.1 Gradient Descent for Convex and L-Smooth Functions

Lemma 7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be both convex and L-smooth. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \leq \nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{y}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2.$$

Proof. Let $\boldsymbol{z} = \boldsymbol{y} - \frac{1}{L}(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))$. Then, one has

$$\begin{aligned} f(\boldsymbol{x}) - f(y) &= f(\boldsymbol{x}) - f(\boldsymbol{z}) + f(\boldsymbol{z}) - f(\boldsymbol{y}) \\ &\leq \nabla f(\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{z}) + \nabla f(\boldsymbol{y})^\top(\boldsymbol{z} - \boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{y}\|^2 \\ &= \nabla f(\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{y}) + (\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top(\boldsymbol{y} - \boldsymbol{z}) + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{y}\|^2 \\ &= \nabla f(\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{y}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \end{aligned}$$

where the first inequality follows from the convexity of f and L-smoothness of f, and the last equality is due to the equation $\boldsymbol{z} = \boldsymbol{y} - \frac{1}{L}(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))$. We complete the proof. ∎

Theorem 8. *Suppose the function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex. Let $t = \frac{1}{L}$, then we gave*

$$f(\boldsymbol{x}^K) - f(\boldsymbol{x}^\star) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2}{K},$$

where $\boldsymbol{x}^\star$ is the optimal solution of Problem (P).

Proof. To start with, we choose the residual function as $R(\boldsymbol{x}) := f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)$. Then, we have

$$\begin{aligned} R(\boldsymbol{x}^{k+1}) - R(\boldsymbol{x}^k) &= (f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^\star)) - (f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star)) \\ &= f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k) \\ &\leq \frac{L}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2, \end{aligned}$$

where the first inequality follows from Proposition 5. It further implies that

$$R(\boldsymbol{x}^{k+1}) \leq R(\boldsymbol{x}^k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2. \tag{1}$$

Moreover, since the function f is convex, we have

$$R(\boldsymbol{x}^k) = f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star) \leq \nabla f(\boldsymbol{x}^k)^\top(\boldsymbol{x}^\star - x^\star) \leq \|\nabla f(\boldsymbol{x}^k)\| \cdot \|\boldsymbol{x}^\star - \boldsymbol{x}^\star\|. \tag{2}$$

Combining (1) and (2) yields

$$R(\boldsymbol{x}^{k+1}) - R(\boldsymbol{x}^k) \leq -\frac{1}{2L}\frac{R(\boldsymbol{x}^k)^2}{\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2}.$$

We have now nearly established the recurrence relation. The remaining task is to demonstrate the boundedness of the sequence.

Lemma 9 (Boundedness of Iterates)**.** *For any $k \geq 0$, we have*

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2.$$

Proof. We have

$$\begin{aligned} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 &= \|\boldsymbol{x}^k - \frac{1}{L}\nabla f(\boldsymbol{x}^k) - \boldsymbol{x}^\star\|^2 \\ &= \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{2}{L}\nabla f(\boldsymbol{x}^k)^\top(\boldsymbol{x}^k - \boldsymbol{x}^\star) + \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^k)\|^2 \\ &\leq \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 + \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^k)\|^2 - \frac{2}{L}(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)) - \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^\star)\|^2 \\ &= \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{2}{L}(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)) \end{aligned}$$

where the first inequality is obtained by applying Lemma 7, i.e.,

$$\nabla f(\boldsymbol{x}^k)^\top(\boldsymbol{x}^k - \boldsymbol{x}^\star) \geq f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*) + \frac{1}{2L}\|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^\star)\|^2,$$

and the last equality follows from $\nabla f(\boldsymbol{x}^\star) = 0$.

■

Armed with Lemma 9, we get

$$R(\boldsymbol{x}^{k+1}) - R(\boldsymbol{x}^k) \leq -\frac{1}{2L}\frac{R(\boldsymbol{x}^k)^2}{\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2},$$

which implies

$$\frac{1}{R(\boldsymbol{x}^{k+1})} - \frac{1}{R(\boldsymbol{x}^k)} \geq -\frac{1}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2}\frac{R(\boldsymbol{x}^k)}{R(\boldsymbol{x}^{k+1})} \geq \frac{1}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2},$$

where the last inequality follows from the monotonicity of the sequence $\{R(\boldsymbol{x}^k)\}_{k\geq 0}$. Thus, we have

$$\frac{1}{R(\boldsymbol{x}^{k+1})} \geq \frac{k}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2}.$$

We finished the proof. ■

Remark 10. *To speed up the convergence rate, the key is to control the right-hand side of*

$$R(\boldsymbol{x}^{k+1}) - R(\boldsymbol{x}^k) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2.$$

By the convexity of f, we can only bound via

$$R(\boldsymbol{x}^k) \leq \|\nabla f(\boldsymbol{x}^k)\| \cdot \|\boldsymbol{x}^\star - \boldsymbol{x}^\star\|.$$

For instance, if we have $R(\boldsymbol{x}^k) \leq \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}^k)\|^2$, for some positive constant μ. Then, we have

$$R(\boldsymbol{x}^{k+1}) - R(\boldsymbol{x}^k) \leq -\frac{\mu}{L}R(\boldsymbol{x}^k)$$

and we achieve the linear convergence. The condition $R(\boldsymbol{x}^k) \leq \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}^k)\|^2$ is precisely the Polyak-Lojasiewicz (PL) condition studied in the literature [2].

3.2 Gradient Descent for μ-Strongly Convex and L-Smooth Functions

A stronger condition than the PŁ condition is strong convexity, see

Definition 11 (Strongly Convex Functions)**.** *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if we have*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} in \mathbb{R}^d.$$

We observe that μ-strong convexity provides a tighter lower bound compared to convexity.

Proposition 12. *A μ-strong convex function is also a μ-PL function.*

Please see [3, Theorem 3.1] for further details.

Remark 13. *Another related regularity condition is called slope (Luo-Tseng) error bound condition [4, 3], i.e.,*

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\| \leq \|\nabla f(x^k)\|.$$

With a stronger lower bound on f, we can derive a stronger version of Lemma 7.

Lemma 14. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be both μ-strongly convex and L-smooth. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have*

$$f(\boldsymbol{x}^+) - f(\boldsymbol{y}) \le \nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{y}) + \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|^2 - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

where $\boldsymbol{x}^+ = \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})$.

Proof. We have

$$\begin{aligned} f(\boldsymbol{x}^+) - f(\boldsymbol{y}) &= f(\boldsymbol{x}^+) - f(\boldsymbol{x}) + f(\boldsymbol{x}) - f(\boldsymbol{y}) \\ &\le \nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 + \nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{y}) - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \\ &= \nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{y}) + \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|^2 - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \end{aligned}$$

where the first inequality follows from the μ-strongly convexity of f and L-smoothness of f, and the second equality is due to the fact that $\boldsymbol{x}^+ = \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})$. ■

Theorem 15. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth. Then, (GD) with $t = \frac{1}{L}$ satisfies the following for $K \ge 0$:*

$$\|\boldsymbol{x}^{K+1} - \boldsymbol{x}^\star\|^2 \le \exp\left(-\frac{K}{\kappa}\right)\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2$$

where κ is the condition number defined as $\kappa = \frac{L}{\mu}$.

Proof. We can follow the proof of Lemma 9. Now, we obtain a tighter bound for the inner product term, i.e.,

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 = \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{2}{L}\nabla f(\boldsymbol{x}^k)^\top (\boldsymbol{x}^k - \boldsymbol{x}^\star) + \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^k)\|^2.$$

By applying Lemma 14, we get

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 \le \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 \le \cdots \le \left(1 - \frac{\mu}{L}\right)^k \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2 \le \exp\left(-\frac{k}{\kappa}\right)\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2,$$

where the last inequality follows from the inequality $(1 - x) \le \exp(-x)$.

■

3.3 Gradient Descent for Smooth Nonconvex Functions

Without the convexity, either $f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)$ or $\|\boldsymbol{x} - \boldsymbol{x}^\star\|$ is not a suitable residual criterion. As we discussed in the last lecture, an alternative optimality condition is based on gradient information, which we consider as follows:

$$R(\boldsymbol{x}) = \|\nabla f(\boldsymbol{x})\|.$$

Still, from the sufficient decrease property in Proposition 5, we have

$$f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k) \le -\frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|^2. \tag{3}$$

By summing (3) from $k = 0$ to $k = K$, we obtain

$$f(\boldsymbol{x}^K) - f(\boldsymbol{x}^\star) \le -\frac{1}{2L}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}^k)\|^2.$$

Thus, we conclude

$$\min_{k\in[K]}\|\nabla f(\boldsymbol{x}^k)\|^2 \le \frac{2L}{K}(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^\star)).$$

Remark 16. *Without convexity, we cannot ensure last-iterate convergence. Instead, we can only guarantee the existence of an index $k \in [K]$ such that $R(\boldsymbol{x}^k)$ is sufficiently small.*

4 SubGradient Method - Algorithms and Complexity

In this section, we primarily focus on nonsmooth optimization problems, without assuming the L-smooth condition.

4.1 Subgradient Method for Convex and L-Lipschitz Functions

The iterative scheme is given by:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - t_k \boldsymbol{g}_k \quad \text{where} \quad \boldsymbol{g}_k \in \partial f(x^k). \tag{SubG}$$

Here, $\partial f(\boldsymbol{x})$ is well defined due to the convexity of f. Moreover, However, if we continue to use a constant step size strategy, such as $t_k = 1/L$, the subgradient method may diverge; for example, consider $f(x) = |x|$.

To conduct the analysis, we make the blanket assumptions (can be further relaxed) as below:

Assumption 17. *The following assumptions hold:*

(i) The condition $\|g\|_2 \le L$ holds for all $g \in \partial f$, meaning f is L-Lipschitz.

(ii) $\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\| \le D$

Theorem 18. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-Lipschitz. Then, for $K \ge 0$, (SubG) satisfies the following:*

$$\min_{k \in [K]} f(\boldsymbol{x}^k) - f^\star \le \frac{D^2 + L^2 \sum_{k=0}^K t_k^2}{2 \sum_{k=0}^K t_k}.$$

Remark 19. *The above theorem statement suggests us to apply the following step size strategy:*

$$\sum_{k=0}^\infty t_k = \infty, \quad \sum_{k=0}^\infty t_k^2 < \infty.$$

Proof. We can follow the proof of Lemma 9. However, we cannot impose the L-smooth condition anymore to bound the gradient term, i.e.,

$$\begin{aligned} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 &= \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - 2t_k \boldsymbol{g}_k^\top (\boldsymbol{x}^k - \boldsymbol{x}^\star) + t_k^2 \|\boldsymbol{g}_k\|^2 \\ &\le \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - 2t_k (f(\boldsymbol{x}^k) - f^\star) + t_k^2 L^2, \end{aligned}$$

where the inequality follows from the convexity of f and L-Lipschitz of f.

Again, we sum the above inequality from $k = 0$ to $k = K$ to obtain:

$$\|\boldsymbol{x}^{K+1} - \boldsymbol{x}^\star\|^2 \le \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|^2 - \sum_{k=0}^K 2t_k (f(\boldsymbol{x}^k) - f^\star) + \sum_{k=0}^K t_k^2 L^2.$$

Rearranging both sides yields

$$\min_{k \in [K]} f(\boldsymbol{x}^k) - f^\star \le \frac{D^2 + L^2 \sum_{k=0}^K t_k^2}{2 \sum_{k=0}^K t_k}.$$

■

Remark 20. *If we choose $t_k = \mathcal{O}(1/\sqrt{k})$, we have $\min_{k \in [K]} f(\boldsymbol{x}^k) - f^\star \le \mathcal{O}(\log K/\sqrt{K})$. As we observed, Subgradient methods is not a descent method.*

4.2 Gradient Descent for μ-Strongly Convex and L-Lipschitz Functions

Theorem 21. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-Lipschitz continuous, then with $t_k = \frac{2}{\mu(k+1)}$, we have*

$$f\left(\sum_{k=1}^{K} \frac{2k}{K(K+1)} \boldsymbol{x}_k\right) - f^\star \le \frac{2L^2}{\mu(K+1)}.$$

Proof. Similar with the convex case, we have

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 = \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - 2t_k \boldsymbol{g}_k^\top (\boldsymbol{x}^k - \boldsymbol{x}^\star) + t_k^2 \|\boldsymbol{g}_k\|^2.$$

By the strong convexity of f, we have

$$\begin{aligned} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 &\le \|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - 2t_k \left(f(\boldsymbol{x}^k) - f^\star + \frac{\mu}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2\right) + t_k^2 L^2 \\ &= \frac{k-1}{k+1}\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{4}{\mu(k+1)}(f(\boldsymbol{x}^k) - f^\star) + t_k^2 L^2, \end{aligned}$$

where the equality follows from $t_k = \frac{2}{\mu(k+1)}$. Rearranging both sides leads to

$$f(\boldsymbol{x}_k) - f^\star \le \frac{\mu(k-1)}{4}\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{\mu(k+1)}{4}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 + \frac{t_k}{2} L^2.$$

Then, we can derive an inequality that allows us to perform a telescoping sum later:

$$k(f(\boldsymbol{x}_k) - f^\star) \le \frac{\mu k(k-1)}{4}\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{\mu k(k+1)}{4}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 + \frac{L^2}{\mu}.$$

In the final step, we construct the point $\sum_{k=1}^{K} \frac{2k}{K(K+1)} \boldsymbol{x}_k$ and apply Jensen's inequality due to the convexity of f:

$$\begin{aligned} f\left(\sum_{k=1}^{K} \frac{2k}{K(K+1)} \boldsymbol{x}_k\right) &\le \frac{2}{K(K+1)} \sum_{k=1}^{K} k f(\boldsymbol{x}^k) \\ &\le \frac{2}{K(K+1)} \sum_{k=1}^{K} \left(k f^\star + \frac{\mu k(k-1)}{4}\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|^2 - \frac{\mu k(k+1)}{4}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^\star\|^2 + \frac{L^2}{\mu}\right) \\ &= f^\star - \frac{\mu}{2}\|\boldsymbol{x}^{K+1} - \boldsymbol{x}^\star\|^2 + \frac{2L^2}{\mu(K+1)}. \end{aligned}$$

Then, we have

$$f\left(\sum_{k=1}^{K} \frac{2k}{K(K+1)} \boldsymbol{x}_k\right) - f^\star \le \frac{2L^2}{\mu(K+1)}.$$

■

References

[1] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. 3

[2] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016. 5

[3] Feng-Yi Liao, Lijun Ding, and Yang Zheng. Error bounds, pl condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods. In *6th Annual Learning for Dynamics & Control Conference*, pages 993–1005. PMLR, 2024. 5

[4] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993. 5