DSC 211 Introduction to Optimization Winter 2024 Instructor: Jun-Kun Wang
Scribe: David Tu, Can Chen January 16, 2024
Editors/TAs: Merlin Chang, Marialena Sfyraki

Lecture 3: Convexity and Gradient Descent

1 Convexity and Gradient Dominant Condition

Theorem 1. ***(μ-strong convexity implies μ-gradient dominant condition)****:* *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, then it satisfies the μ-gradient dominant condition for any $x, y \in \mathbb{R}^d$, i.e.*

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - \min_{x \in \mathbb{R}^d} f(x))$$

Proof. $\forall x, y \in \mathbb{R}^d$, by definition of μ-strong convexity we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2.$$

Let $h(y)$ be the right-hand side of the above inequality, i.e.

$$h(y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2.$$

Since for all $x, y \in \mathbb{R}^d$ we have $f(y) \geq h(y)$, then the following inequality is always true

$$\min_{y \in \mathbb{R}^d} f(y) \geq \min_{y \in \mathbb{R}^d} h(y). \tag{1}$$

To find $\min\limits_{y \in \mathbb{R}^d} h(y)$, we need to find where the gradient is 0, thus

$$\begin{aligned} &\nabla h(y) = 0 \\ \Leftrightarrow\ &\nabla f(x) + \mu(y - x) = 0 \\ \Leftrightarrow\ &y = x - \frac{1}{\mu}\nabla f(x). \end{aligned}$$

Now that we have the $\arg\min_y h(y)$, we can plug it into h to find $\min_{y \in \mathbb{R}^d} h(y)$ as

$$\begin{aligned} \min_{y \in \mathbb{R}^d} h(y) &= f(x) + \langle \nabla f(x), -\frac{1}{\mu}\nabla f(x) \rangle + \frac{\mu}{2}\|\frac{1}{\mu}\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2. \end{aligned} \tag{2}$$

Therefore, combining (1) and (2), then rearranging, we get

$$\min_{y \in \mathbb{R}^d} f(y) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

$$\Leftrightarrow \|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - \min_{x\in\mathbb{R}^d} f(x)).$$

Thus, we have shown that f satisfies the μ-gradient dominant condition. □

Note: Every stationary point of a function that satisfies gradient dominant condition is a global optimal point. That is because a stationary point x satisfies $\nabla f(x) = 0$. If we plug it into the above inequality, we get

$$0 \geq 2\mu(f(x) - \min_{x\in\mathbb{R}^d} f(x)),$$

where $\mu > 0$. Additionally, since $f(x) - \min_{x\in\mathbf{dom}f} f(x) \geq 0$, by the squeeze theorem we have that $f(x) = \min_{x\in\mathbf{dom}f} f(x)$ and thus x is a global optimal point of f.

2 Dual Norm

Definition 1. ***(Dual Norm)***: *For a norm $\|\cdot\|$ on $\mathbb{R}^d$, its dual norm $\|\cdot\|_*$ is a function $\|\cdot\|_* : \mathbb{R}^d \to \mathbb{R}$ defined as*

$$\|y\|_* := \sup_{x:\|x\|\leq 1} \langle y, x\rangle$$

Fact: For the l_p-norm

$$\|x\|_p := (\sum_{i=1}^{d} |x|^p)^{\frac{1}{p}}$$

its dual norm is the l_q-norm, i.e. $\|x\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$.
Furthermore, the dual norm of the Euclidean norm (l_2-norm) is itself, which can be proven directly with the Cauchy-Schwartz inequality.

3 L-smoothness

Definition 2. ***(L-smoothness)***: *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth w.r.t. $\|\cdot\|$, if for any $x, y \in \mathbb{R}^d$*

$$f(y) \leq f(x) + \langle\nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \tag{3}$$

where $L > 0$.

Definition 3. ***(L-Lipschitz)***: *A function $f : \Omega \to \mathbb{R}$ is L-Lipschitz w.r.t. $\|\cdot\|$ over Ω, if for any $x, y \in \Omega$,*

$$|f(x) - f(y)| \leq L\|x - y\|, \tag{4}$$

where $L > 0$.

Theorem 2. ***(L-Lipschitz gradient implies L-smoothness)***: *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. If the gradient map $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz w.r.t. $\|\cdot\|$, i.e.*

$$\forall x, y : \|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|,$$

then, f is L-smooth, i.e.

$$\forall x, y : f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

Remark: If $f(\cdot)$ is convex, then the converse is also true! See e.g., [3] for the exposition.

Theorem 3. ***(Second-order characterization of L-smoothness)*** *A twice differentiable function $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is smooth w.r.t. a norm $\|\cdot\|_2$, if and only if*

$$y^\top \nabla^2 f(x) y \leq L\|y\|_2^2, \ \forall x, y \in \mathbb{R}^d.$$

Remark: See e.g., Section 3.5 of [2] for the proof.

Examples of Smooth Function:

1. $\frac{1}{2}x^2$
2. $\log(1 + exp(-x))$

Examples of Non-smooth Function:

1. $\max\{0, 1 - x\}$ (Hinge-loss function is not differentiable at $x = 1$)
2. $exp(-x)$ (Only smooth in a bounded region $x \in [-c, c], c < \infty$)

4 Gradient Descent

Now let us analyze the iteration complexity of Gradient Descent:

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

Theorem 4. *Assume $f(\cdot)$ is μ-gradient dominant and L-smooth, then gradient descent with $\eta = \frac{1}{L}$ satisfies*

$$f(x_{k+1}) - \min_{x \in \mathbb{R}^d} f(x) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_1) - \min_{x \in \mathbb{R}^d} f(x)\right).$$

Remark: $1-\theta \leq exp(-\theta)$ implies $\left(1-\frac{\mu}{L}\right)^k \leq exp\left(-\frac{\mu}{L}k\right)$

Theorem 3 is also applicable to μ-strongly convex and L-smooth functions.

Proof. We have that

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 && \text{(by L-smoothness)} \\
&= f(x_k) - \eta\|\nabla f(x_k)\|_2^2 + \frac{L\eta^2}{2}\|\nabla f(x_k)\|_2^2 && \text{(by the GD update rule)} \\
&= f(x_k) - \left(\frac{1}{L} - \frac{1}{2L}\right)\|\nabla f(x_k)\|_2^2 \\
&= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \\
&\leq f(x_k) - \frac{\mu}{L}\left(f(x_k) - \min_{x\in\mathbb{R}^d} f(x)\right) && \text{(by Gradient Dominant)}
\end{aligned}$$

Subtracting $\min_{x\in\mathbb{R}^d} f(x)$ from both sides we get

$$f(x_{k+1}) - \min_{x\in\mathbb{R}^d} f(x) \leq \left(1-\frac{\mu}{L}\right)\left(f(x_k) - \min_{x\in\mathbb{R}^d} f(x)\right).$$

Let $\delta_{k+1} := f(x_k) - \min_{x\in\mathbb{R}^d} f(x)$, then

$$\begin{aligned}
\delta_{k+1} &\leq \left(1-\frac{\mu}{L}\right)\delta_k \\
&\leq \left(1-\frac{\mu}{L}\right)\left(1-\frac{\mu}{L}\right)\delta_{k-1} \\
&\leq \left(1-\frac{\mu}{L}\right)\left(1-\frac{\mu}{L}\right)\left(1-\frac{\mu}{L}\right)\delta_{k-2} \\
&\leq \left(1-\frac{\mu}{L}\right)^k\delta_1.
\end{aligned}$$

Thus,

$$\delta_{k+1} \leq \left(1-\frac{\mu}{L}\right)^k\delta_1.$$

□

Fact: We have $L \geq \mu$, i.e., the condition number $\kappa := \frac{L}{\mu} \geq 1$. To make sense of this relation, let us think of μ as μ of the μ-strong convexity (the argument when μ is that of μ-gradient dominance will be more involved). Then, on one hand, by the first-order characterization of the L-smoothness, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^d. \tag{5}$$

On the other hand, by the first-order characterization of the μ-strong convexity, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^d. \tag{6}$$

It is evident that $L \geq \mu$, otherwise the above two inequalities will contradict to each other. We can also deduce $L \geq \mu$ using the second-order characterization of the L-smoothness and the μ-strong convexity when $f(\cdot)$ is twice differentiable:

$$L := \max_{x \in \mathbb{R}^d} \lambda_{max}(\nabla^2 f(x)) \qquad \mu := \min_{x \in \mathbb{R}^d} \lambda_{min}(\nabla^2 f(x))$$

Theorem 5. *Assume $f(\cdot)$ is convex and L-smooth on $\mathbb{R}^d$, then gradient descent with $\eta = \frac{1}{L}$ satisfies*

$$f(x_{k+1}) - \min_{x \in \mathbb{R}^d} f(x) \leq \frac{2LD^2}{k}$$

where $D := \max_k \|x_k - x_*\|_2$, $x_* := \arg\min f(x)$.

Remark: D can be shown to be bounded by the initial distance, i.e., $D = \|x_1 - x_*\|_2$.

Proof. From the proof of Theorem 4, we know gradient descent with the step size $\eta = \frac{1}{L}$ for L-smooth function has

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|_2^2.$$

Denote,

$$f(x_*) := \min_{x \in \mathbb{R}^d} f(x) \qquad \delta_{k+1} := f(x_{k+1}) - f(x_*)$$

Then,

$$\delta_{k+1} - \delta_k \leq -\frac{1}{2L}\|\nabla f(x_k)\|_2^2. \tag{7}$$

Moreover,

$$\begin{aligned}
\delta_k &= f(x_k) - f(x_*) \\
&\leq \langle \nabla f(x_k), x_k - x_* \rangle && \text{(by Convexity)} \\
&\leq \|\nabla f(x_k)\|_2 \|x_k - x_*\|_2 && \text{(by Cauchy-Schwartz inequality)}
\end{aligned}$$

Hence,

$$\|\nabla f(x_k)\|_2 \geq \frac{\delta_k}{\|x_k - x_*\|_2}. \tag{8}$$

Combining (7) and (8), we get

$$\begin{aligned} \delta_{k+1} - \delta_k &\leq -\frac{1}{2L}\frac{{\delta_k}^2}{\|x_k - x_*\|_2^2} \\ \Leftrightarrow \frac{\delta_k - \delta_{k+1}}{\delta_k} &\geq \frac{1}{2L}\frac{\delta_k}{\|x_k - x_*\|_2^2} \geq \frac{\delta_k}{2LD^2} \end{aligned}$$

Therefore,

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} = \frac{\delta_k - \delta_{k+1}}{\delta_k \delta_{k+1}} \geq \frac{\delta_k}{2LD^2 \cdot \delta_{k+1}}.$$

Optimality gap is non-increasing (7), thus

$$\delta_k \geq \delta_{k+1} \Rightarrow \frac{\delta_k}{\delta_{k+1}} \geq 1.$$

Hence,

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \frac{\delta_k}{2LD^2 \cdot \delta_{k+1}} \geq \frac{1}{2LD^2}$$

$$\begin{aligned} \frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} &\geq \frac{1}{2LD^2} \\ \frac{1}{\delta_k} - \frac{1}{\delta_{k-1}} &\geq \frac{1}{2LD^2} \\ \frac{1}{\delta_{k-1}} - \frac{1}{\delta_{k-2}} &\geq \frac{1}{2LD^2} \\ &\vdots \\ \frac{1}{\delta_2} - \frac{1}{\delta_1} &\geq \frac{1}{2LD^2} \end{aligned}$$

By telescoping sum,

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_1} \geq \frac{k}{2LD^2} \tag{9}$$

What is δ_1?

$$\begin{aligned}
\delta_1 &= f(x_1) - f(x_*) \\
&\leq \langle \nabla f(x_1), x_1 - x_* \rangle && \text{(by Convexity)} \\
&\leq \|\nabla f(x_1)\|_2 \|x_1 - x_*\|_2 && \text{(by Cauchy-Schwartz inequality)} \\
&= \|\nabla f(x_1) - \nabla f(x_*)\|_2 \|x_1 - x_*\|_2 && (\nabla f(x_*) = 0) \\
&\leq L\|x_1 - x_*\|_2^2 && \text{(by L-Lipschitz)} \\
&\leq LD^2
\end{aligned}$$

$$\Rightarrow \frac{1}{\delta_1} \geq \frac{1}{LD^2} \tag{10}$$

Combining (9) and (10), we get

$$\begin{aligned}
\frac{1}{\delta_{k+1}} &\geq \frac{k}{2LD^2} + \frac{1}{\delta_1} \\
&\geq \frac{k}{2LD^2} + \frac{1}{LD^2} \\
&= \frac{k+2}{2LD^2}
\end{aligned}$$

$$\begin{aligned}
\Leftrightarrow \delta_{k+1} &\leq \frac{2LD^2}{k+2} \\
\Rightarrow \delta_{k+1} &\leq \frac{2LD^2}{k}
\end{aligned}$$

□

Bibliographic notes

Regarding the smoothness and the gradient Lipschitzness, see Chapter 3 of [2] and [3]. For the iteration complexity of gradient descent, Chapter 6 of [1] provides a nice exposition.

References

[1] Nisheeth K. Vishnoi. Algorithms for Convex Optimization Cambridge University Press, 2021

[2] Aaron Sidford Optimization Algorithms 2023 https://drive.google.com/file/d/1BfMkt2glaZpJGwg7gwsJw9T_XxH3o8gx/view

[3] Xingyu Zhou On the Fenchel Duality between Strong Convexity and Lipschitz Continuous Gradient arXiv:1803.06573, 2018