ECE 273 Convex Optimization and Applications — Instructor: Jun-Kun Wang
Scribe: Trevor Kam, Nam Do, Siddhanth Ramani — April 9, 2024
Editor/TA: Marialena Sfyraki

Lecture 3: (Continue) Convex Analysis I and Gradient Descent

1 Review of Lecture 2

1.1 Convex Functions

Definition 1. *(**Zero Order Characterization of Convex Functions**): A function $f : C \to \mathbb{R}$ defined over a convex set C is called convex if, for any $\mathbf{x}, \mathbf{y} \in C$ and any $\alpha \in [0, 1]$, the following inequality holds*

$$f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}\right) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).$$

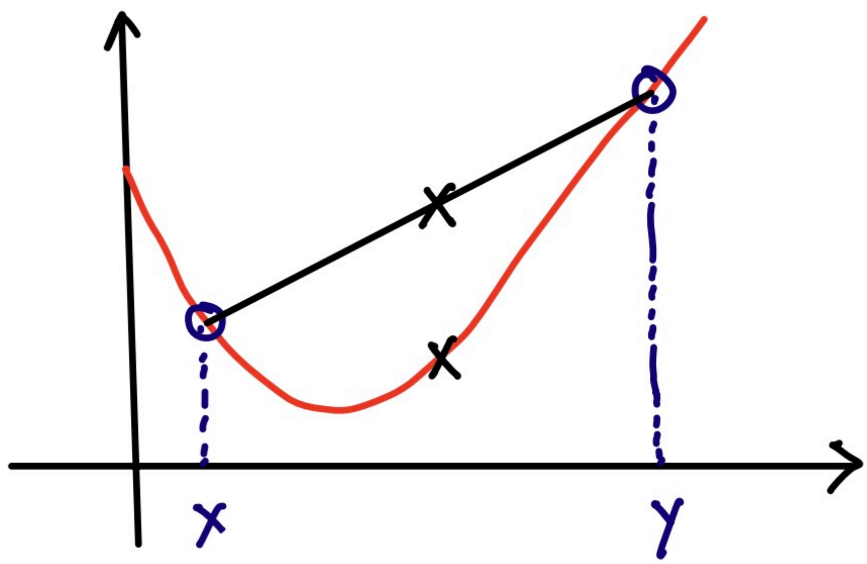

Figure 1: An illustration of zero-order characterization of convexity

Figure 1 shows that for a convex function f, for any two points $\mathbf{x}, \mathbf{y}$, the function f evaluated at any convex combination of $\mathbf{x}, \mathbf{y}$ should be no larger than the same convex combination of $f(\mathbf{x})$ and $f(\mathbf{y})$.

Definition 2. *(**First Order Characterization of Convex Functions**): A differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is called convex **if and***

only if, *for any* $\mathbf{x}, \mathbf{y} \in C$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), (\mathbf{y} - \mathbf{x}) \rangle.$$

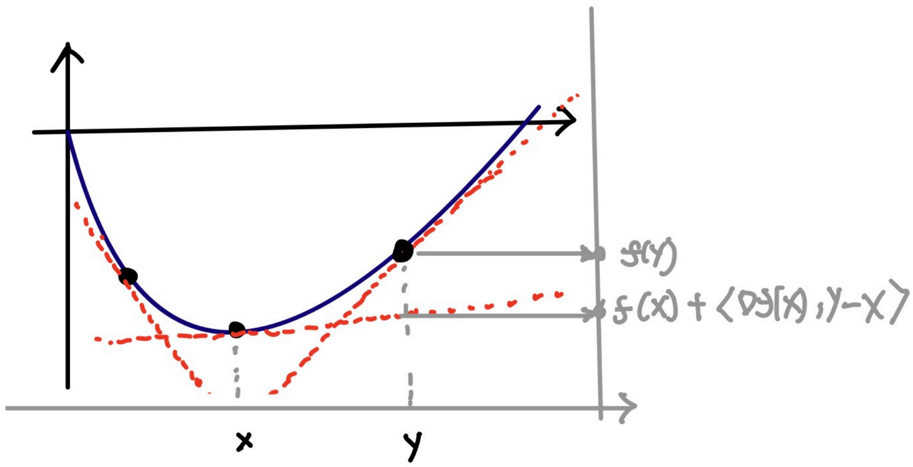

Figure 2: An illustration of first-order characterization of convexity

Figure 2 shows that the function always dominates its first-order (linear) Taylor approximation.

Definition 3. *(**Second Order Characterization of Convex Functions**): A twice-differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is convex **if and only if**, for any $\mathbf{x} \in C$, the Hessian matrix evaluated at $\mathbf{x}$ is positive semi-definite, i.e.*

$$\nabla^2 f(\mathbf{x}) \succeq 0.$$

Definition 4. *(**Equivalency of convexity**): For any $\mathbf{x} \in C \subseteq \mathbb{R}^d$ and $\mathbf{y} \in C \subseteq \mathbb{R}^d$, and any $\alpha \in [0, 1]$:*

$$f\big((1-\alpha)\mathbf{x} + \alpha\mathbf{y}\big) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y})$$
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$
$$\nabla^2 f(\mathbf{x}) \succeq 0$$

Definition 5. *(**Equivalency of strong convexity**): A function $f : C \to \mathbb{R}$ defined over a convex set C is μ-strongly convex w.r.t. a norm $|| \cdot ||$ if, for any $\mathbf{x} \in C \subseteq \mathbb{R}^d$, $\mathbf{y} \in C \subseteq \mathbb{R}^d$, $\mathbf{z} \in \mathbb{R}^d$, and any $\alpha \in [0, 1]$:*

$$f\big((1-\alpha)\mathbf{x} + \alpha\mathbf{y}\big) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y} - \mathbf{x}||^2.$$
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}||\mathbf{y} - \mathbf{x}||^2.$$
$$\mathbf{z}^\top \nabla^2 f(\mathbf{x})\mathbf{z} \geq \mu\|\mathbf{z}\|^2$$

for some $\mu > 0$.

Remark:
If $\|\cdot\| \equiv \|\cdot\|_2$ and $\mathbf{z}$ is an eigenvector of $\nabla^2 f(\mathbf{x})$, then

$$\begin{aligned}
&\nabla^2 f(\mathbf{z}) = \lambda \mathbf{z}, \text{ for some } \lambda \\
&\Rightarrow \mathbf{z}^\top \nabla^2 f(\mathbf{x}) \mathbf{z} = \lambda \|\mathbf{z}\|^2 \geq \mu \|\mathbf{z}\|^2 \\
&\Leftrightarrow \lambda \geq \mu
\end{aligned}$$

Question: What happens if the norm is not l_2?
Answer: In that case, $\lambda \geq 0$ but the inequality between λ and μ will be related by some constants.

Definition 6. *(**Gradient Dominant** or **Polyak-Lojasiewicz (PL) Condition**): We say a function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the "Gradient Dominance" condition if*

$$||\nabla f(\mathbf{x})||_2^2 \geq 2\mu \left(f(\mathbf{x}) - \min_{\mathbf{x}} f(\mathbf{x}) \right) , \quad \textit{for some} \ \ \mu > 0.$$

Remark: For any function satisfying the PL condition, every stationary point is a global minimum.

Gradient Flow: The Gradient Flow is defined as:

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla f \left(\mathbf{x}(t) \right)$$

Question: Can an optimal solution also be a maximum?
Answer: We usually consider the minima in an optimization problem. Observe that the problem of maximizing a function f is equivalent to the problem of minimizing $-f$.

Theorem 1. *Assume $f(\cdot)$ satisfies μ-gradient dominance condition. Gradient Flow for $\min_w f(w)$ satisfies:*

$$f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x}) \leq \exp\left(-2\mu t\right) \left(f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x}) \right)$$

Remark: Denote $a_k = f(\mathbf{x}_t) - \min_{\mathbf{x}} f(\mathbf{x})$.

Definition 7. *(**Linear Rate**): We say a_k converges linearly if there exist constants $c > 0, q \in (0, 1]$ satisfying*

$$a_k \leq c(1-q)^k \text{ for all } k. \tag{1}$$

In this case, we call $1 - q$ the linear rate of convergence.

Equivalency of strong convexity: For any $\mathbf{x} \in C \subseteq \mathbb{R}^d$, $\mathbf{y} \in C \subseteq \mathbb{R}^d$, $\mathbf{z} \in \mathbb{R}^d$, and any $\alpha \in [0, 1]$:

$$f\big((1-\alpha)\mathbf{x} + \alpha\mathbf{y}\big) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)||\mathbf{y} - \mathbf{x}||^2. \tag{2}$$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}||\mathbf{y} - \mathbf{x}||^2. \tag{3}$$

$$\mathbf{z}^\top \nabla^2 f(\mathbf{x})\mathbf{z} \geq \mu\|\mathbf{z}\|^2 \tag{4}$$

for some $\mu > 0$, while convexity is when $\mu = 0$.

Question: For a given problem, which inequality should be used to prove convexity and strong convexity?
Answer: If the function is continuous, then we can use the zero-order characterization. If the function is additionally (once) continuously differentiable, then we can also use the first-order characterization. If, additionally, the function is continuously differentiable up to the second-order, then we can also use the second-order characterization.

1.2 Proof: First Order Def. (3) $\rightarrow$ Second Order Def. (4)

Denote $\mathbf{x}_\alpha := \mathbf{x} + \alpha\mathbf{z}$, and denote $g(\alpha) := f(\mathbf{x}_\alpha)$.
Then, by chain rule,

$$\begin{aligned} g^{'}(\alpha) &= \frac{\partial f(\mathbf{x}_\alpha)}{\partial \alpha} \\ &= \sum_{i=1}^{d} \frac{\partial f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i]} \frac{\partial \mathbf{x}_\alpha[i]}{\partial \alpha} \\ &= \sum_{i=1}^{d} \frac{\partial f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i]} \mathbf{z}[i] \\ &= \langle \nabla f(\mathbf{x}_\alpha), \mathbf{z}\rangle. \end{aligned}$$

Also,

$$
\begin{aligned}
g^{''}(\alpha) &= \sum_{i=1}^{d} \frac{\partial}{\partial \alpha} \left(\frac{\partial f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i]} \right) \mathbf{z}[i] \\
&= \sum_{i=1}^{d} (\sum_{j=1}^{d} \frac{\partial^2 f(\mathbf{x}_\alpha)}{\partial \mathbf{x}_\alpha[i] \partial \mathbf{x}_\alpha[j]} \frac{\partial \mathbf{x}_\alpha[j]}{\partial \alpha}) \mathbf{z}[i] \\
&= \sum_{i=1}^{d} (\sum_{j=1}^{d} \nabla^2 f(\mathbf{x}_\alpha[i,j]) \mathbf{z}[j]) \mathbf{z}[i] \\
&= \mathbf{z}^\top \nabla^2 f(\mathbf{x}_\alpha) \mathbf{z}.
\end{aligned}
$$

Continuing, we have that

$$
\begin{aligned}
g'(\alpha) &= \langle \nabla f(\mathbf{x}_\alpha), \mathbf{z} \rangle \\
g'(0) &= \langle \nabla f(\mathbf{x}), \mathbf{z} \rangle
\end{aligned}
$$

and

$$
g''(\alpha) = \mathbf{z}^T \nabla^2 f(\mathbf{x}_\alpha) \mathbf{z} \tag{5}
$$

$$
\begin{aligned}
\mathbf{z}^T \nabla^2 f(\mathbf{x}) \mathbf{z} = g''(0) &= \lim_{\alpha \to 0} \frac{g'(\alpha) - g'(0)}{\alpha} && (6) \\
&= \lim_{\alpha \to 0} \frac{\langle \nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x}), \mathbf{z} \rangle}{\alpha} && (7) \\
&= \lim_{\alpha \to 0} \frac{\langle \nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x}), \mathbf{x}_\alpha - \mathbf{x} \rangle}{\alpha^2} && (8)
\end{aligned}
$$

Remark: We get from (7) to (8) by subbing in $\mathbf{z} = \frac{\mathbf{x}_\alpha - \mathbf{x}}{\alpha}$

We further lower-bound $\langle \nabla f((x_\alpha) - \nabla f(\mathbf{x}), \mathbf{x}_\alpha - \mathbf{x} \rangle$ as follows:
By strong convexity:

$$
f(\mathbf{x}_\alpha) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}, \mathbf{x}_\alpha - \mathbf{x} \rangle + \frac{\mu}{2} ||\mathbf{x}_\alpha - \mathbf{x}||^2 \tag{9}
$$

$$
f(\mathbf{x}) \geq f(\mathbf{x}_\alpha) + \langle \nabla f(\mathbf{x}_\alpha, \mathbf{x} - \mathbf{x}_\alpha \rangle + \frac{\mu}{2} ||\mathbf{x}_\alpha - \mathbf{x}||^2 \tag{10}
$$

Remark: Here, $f(\mathbf{x}_\alpha)$ and $f(\mathbf{x})$ cancel each other out.

Adding the above two, (9) and (10), we get

$$
\langle \nabla f(\mathbf{x}_\alpha) - \nabla f(\mathbf{x}), \mathbf{x}_\alpha - \mathbf{x} \rangle \geq \mu \alpha^2 ||\mathbf{z}||^2 \tag{11}
$$

Combining (5), (10), (11) yields

$$
\mathbf{z}^T \nabla^2 f(\mathbf{x}) \mathbf{z} \geq \mu ||\mathbf{z}||^2
$$

1.3 Proof First Order Def. (4) to Second Order Def. (3)

We denote $\mathbf{x}_\alpha := \mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})$

Lemma:

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \int_0^\theta (\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x}_\alpha)(\mathbf{y} - \mathbf{x}) d\alpha d\theta$$

see e.g., Lemma 3.11 of [Sidford (2024)] for the proof. It can be shown by (the variants of) the Fundamental Theorem of Calculus that we saw in Lecture 1.
Starting from

$$\mathbf{z}^T \nabla^2 f(\mathbf{x}) \mathbf{z} \geq \mu ||\mathbf{z}||^2, \forall \mathbf{z} \in \mathbb{R}^d.$$

Plugging in $\mathbf{z} \leftarrow \mathbf{y} - \mathbf{x}$ and $\mathbf{x} \leftarrow \mathbf{x}_\alpha$ we get,

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \int_0^\theta (\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x}_\alpha)(\mathbf{y} - \mathbf{x}) d\alpha d\theta \\ &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \int_0^\theta \mu ||\mathbf{y} - \mathbf{x}||^2 d\alpha d\theta \quad \text{, by second-order characterization} \\ &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} ||\mathbf{y} - \mathbf{x}||^2, \text{ since } \int_0^\theta \mu ||\mathbf{y} - \mathbf{x}||^2 d\alpha = \mu ||\mathbf{y} - \mathbf{x}||^2 \theta \\ &\text{and } \int_0^1 \theta d\theta = \frac{1}{2}. \end{aligned}$$

1.4 Examples of functions satisfying the "Gradient Dominant" condition

Example 1: Squared loss

$$\frac{1}{2}\mathbf{x}^2$$

Example 2: Negative Entropy over the simplex $\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}[i] \geq 0, \sum_{i=1}^d \mathbf{x}[i] = 1\}$

$$f(\mathbf{x}) = \sum_{i=1}^d \mathbf{x}[i] log \mathbf{x}[i]$$

Example 3: Strongly convex functions

Theorem 2. *The μ-strong convexity implies the μ-Gradient Dominant condition, i.e., $||\nabla f(\boldsymbol{x})||_2^2 \geq 2\mu(f(\boldsymbol{x}) - min_x f(\boldsymbol{x}))$, for some $\mu > 0$.*

Definition 8. *(μ-strong convexity):*
For $\forall \boldsymbol{x}, \boldsymbol{y}$

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

Proof. Let $h_{\mathbf{x}}(\mathbf{y}) := f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} ||\mathbf{y} - \mathbf{x}||_2^2$, then we get can get

$$\min_{\mathbf{y} \in \mathbb{R}^d} f(\mathbf{y}) \geq \min_{\mathbf{y} \in \mathbb{R}^d} h_{\mathbf{x}}(\mathbf{y}) \tag{12}$$

Solving for $\min_{\mathbf{y}} h_{\mathbf{x}}(\mathbf{y})$ we get,

$$\min_{\mathbf{y}} h_{\mathbf{x}}(\mathbf{y}) \equiv \min_{\mathbf{y} \in \mathbb{R}^d} f(\mathbf{x}) + \langle \nabla f(\mathbf{x}, \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} ||\mathbf{y} - \mathbf{x}||_2^2$$

Let

$$\mathbf{y}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} h_{\mathbf{x}}(\mathbf{y})$$

Then,

$$\nabla h(\mathbf{y}^*) = \nabla f(\mathbf{x}) + \mu(\mathbf{y}^* - \mathbf{x}) = 0 \in \mathbb{R}^d$$

$$\Leftrightarrow \mathbf{y}^* - \mathbf{x} = -\frac{\nabla f(\mathbf{x})}{\mu}$$

Using this we get,

$$\begin{aligned}\min_{\mathbf{y}} h_{\mathbf{x}}(\mathbf{y}) &\equiv f(\mathbf{x}) - \frac{||\nabla f(\mathbf{x})||_2^2}{\mu} + \frac{1}{2\mu} ||\nabla f(\mathbf{x})||_2^2 \\ &= f(\mathbf{x}) - \frac{1}{2\mu} ||\nabla f(\mathbf{x})||_2^2\end{aligned}$$

Solving for $\min_{\mathbf{y}} f(\mathbf{y})$ we get,

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

$$\Leftrightarrow ||\nabla f(\mathbf{x})||_2^2 \geq 2\mu(f(\mathbf{x}) - \min_{\mathbf{y}} f(\mathbf{y}))$$

$\square$

2 Upper and Lower Bound of a function *f(y)*

2.1 L-smoothness and u-strong convexity

Definition 9. *(**L-smoothness of a function**): A differentiable function f is L-smooth w.r.t. a norm $\|\cdot\|$, **if** $\forall \mathbf{x}, \mathbf{y}$*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \tag{13}$$

where $L > 0$.

Remark: L is a finite number otherwise if L is infinite this becomes a trivial upper bound of $f(\mathbf{y})$.

Definition 10. *(μ-**strong convexity**): A differentiable function $f : C \to \mathbb{R}$ defined over a convex set C is μ-strongly convex w.r.t. a norm $|| \cdot ||$ **if and only if** for any $\mathbf{x}, \mathbf{y} \in C$ we have*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}||\mathbf{y} - \mathbf{x}||^2 \tag{14}$$

for some $\mu > 0$.

Remark: Smoothness inequality provides an upper bound. Strong convexity inequality provides a lower bound.

If a function $f(\mathbf{y})$ satisfies both conditions, i.e.,

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

then the condition number of the Hessian is $\kappa = \dfrac{L}{\mu} \geq 1$.

In Figure 3 $f(\mathbf{y})$ is lower bounded by the red curve found by applying μ-strong convexity and is upper bounded by the green curve found by applying the L-smoothness.

Definition 11. ***(Second-order characterization of L-smoothness)***[1] *A twice differentiable function $f(\cdot) : C \to \mathbb{R}$ defined over a set $C \subseteq \mathbb{R}^d$ is smooth w.r.t. a norm $\| \cdot \|_2$, if and only if $\forall \mathbf{x} \in C$,*

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

[1] See e.g., Section 3.5 of Aaron Sidford "Optimization Algorithms" for the proof.

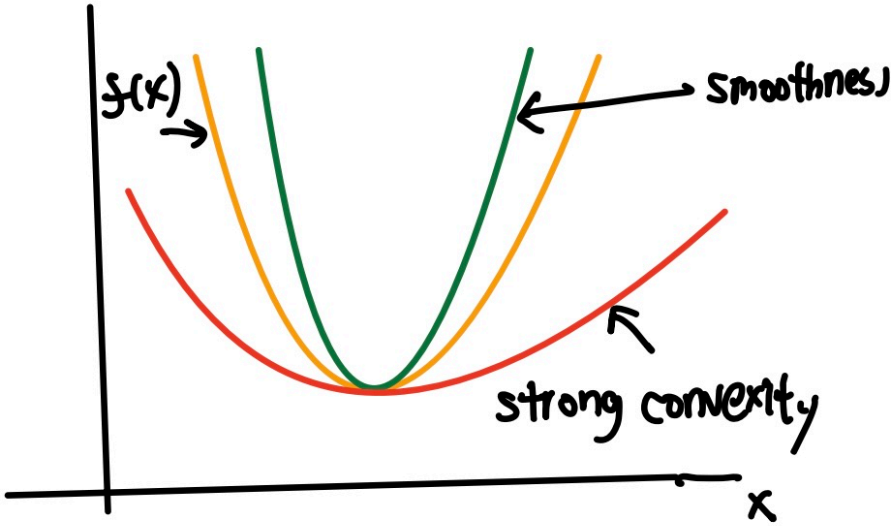

Figure 3: Visualization of lower and upper bounds

Remark: As $\nabla^2 f(\mathbf{x})$ is positive semi-definite, all of its eigenvalues λ are non-negative. The maximum eigenvalue satisfies $\lambda_{max}(\nabla^2 f(\mathbf{x})) \leq L, \forall \mathbf{x} \in C$. Also, if a function $f(\mathbf{x})$ satisfies second-order μ-strong convexity e.g., $\mathbf{z}^\top \nabla^2 f(\mathbf{x})\mathbf{z} \geq \mu \|\mathbf{z}\|_2^2$, then the condition number of $\nabla^2 f(\mathbf{x})$ is $\kappa = \dfrac{L}{\mu} \geq 1$.

Example 1:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - b^\top \mathbf{x}, \text{ where } A \succ 0.$$

- Denote $\lambda_{\max}(A)$ the largest eigenvalue of A is also the smoothness constant L, i.e. $L = \lambda_{\max}(A)$.
- Denote $\lambda_{\min}(A) > 0$ the smallest eigenvalue of A is also the μ-strong convexity constant μ, i.e. $\mu = \lambda_{\min}(A)$.

Example 2: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^2$ is a smooth function as $\nabla^2 f(\mathbf{x}) = 1$ the Hessian is upper bounded by 1.

Example 3: $\log(1+exp(-x))$ this is called logistic function and is a smooth function. The first derivative is

$$f'(x) = -\frac{\exp(-x)}{1+\exp(-x)} = -\frac{1}{1+\exp(x)}$$

And then, the second derivative $f''(x)$ is

$$f''(x) = \frac{\exp(x)}{(1+\exp(x))^2} \leq 0$$

because the denominator approaches infinity faster than the numerator does as $x \to \infty$.

Example 4: $max\{0, 1-x\}$ is called Hinge Loss, and it is not a smooth function as it is not differentiable at 1. Verifying by computing the derivative at 1. $f'(1) = -1$ when $x \to 1^-$ while $f'(1) = 0$ when $x \to 1^+$.

Example 5: $f(x) = exp(-x)$ has $f''(x) = exp(-x)$ which is not bounded thus not a smooth function.

2.2 L-Lipschitz gradients

Theorem 3. *Suppose that $f(\cdot)$ has L-Lipschitz gradients w.r.t. l_2 norm, i.e.,*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x}-\mathbf{y}\|_2 \tag{15}$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Then, L-Lipschitz gradients implies L-smoothness, i.e.,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \tag{16}$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Remark: When the function is convex equation, the inverse is also true, i.e. L-smoothness (16) implies L-Lipschitz gradients (15) (verifying this in homework 1, problem 5)

3 The Upper Bound of Optimality Gap in Gradient Descent

Consider the problem of minimizing f using Gradient Descent i.e. $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

Algorithm 1 GRADIENT DESCENT

1: Input: an initial point $\mathbf{x}_0 \in \mathbf{dom}\ f$ and step size η.
2: **for** $k = 1$ to K **do**
3: $\quad\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta\nabla f(\mathbf{x_k})$
4: **end for**
5: Return $\mathbf{x}_{k+1}$.

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

$$f(\mathbf{x}_{k+1}) - \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(\mathbf{x}_1) - \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})\right).$$

Remark: Under the assumptions of theorem (4), the convergence rate of Gradient Descent is $(1 - \frac{1}{\kappa})$.

Proof. Starting from L-smoothness inequality:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \tag{17}$$

Assigning:

$$\mathbf{y} \leftarrow \mathbf{x}_{k+1}$$

$$\mathbf{x} \leftarrow \mathbf{x}_k$$

and from gradient descent update step, we have:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \eta\nabla f(\mathbf{x}_k) \\ \Leftrightarrow \mathbf{x}_{k+1} - \mathbf{x}_k &= -\eta\nabla f(\mathbf{x}_k) \end{aligned}$$

Equation (17) becomes:

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), -\eta\nabla f(\mathbf{x}_k) \rangle + \frac{L}{2}\| - \eta\nabla f(\mathbf{x}_k)\|^2 \\ &= f(\mathbf{x}_k) - \eta\|\nabla f(\mathbf{x}_k)\|^2 + \frac{L\eta^2}{2}\|\nabla f(\mathbf{x}_k)\|^2 \\ &= f(\mathbf{x}_k) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla f(\mathbf{x}_k)\|^2 \\ &= f(\mathbf{x}_k) - \left(\frac{1}{L} - \frac{L}{2L^2}\right)\|\nabla f(\mathbf{x}_k)\|^2, \quad \text{as } \eta = \frac{1}{L} \\ &= f(\mathbf{x}_k) - \left(\frac{1}{2L}\right)\|\nabla f(\mathbf{x}_k)\|^2 \\ &\leq f(\mathbf{x}_k) - \left(\frac{2\mu}{2L}\right)(f(\mathbf{x}_k) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})), \quad \text{by } \textbf{PL-condition} \end{aligned}$$

Remark: The last inequality is obtained by manipulating the gradient dominant or PL condition as:

$$||\nabla f(\mathbf{x}_k)||_2^2 \geq 2\mu\left(f(\mathbf{x}_k) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})\right) \iff -||\nabla f(\mathbf{x}_k)||_2^2 \leq -2\mu\left(f(\mathbf{x}_k) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})\right)$$

Thus:

$$
\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \left(\frac{\mu}{L}\right)(f(\mathbf{x}_k) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})) \\
\iff f(\mathbf{x}_{k+1}) - \min_{\mathbf{x}\in R^d} f(\mathbf{x}) &\leq (f(\mathbf{x}_k) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})) - \left(\frac{\mu}{L}\right)(f(\mathbf{x}_k) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})) \\
&\leq \left(1 - \frac{\mu}{L}\right)(f(\mathbf{x}_k) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})) \\
&\leq \left(1 - \frac{\mu}{L}\right)^2 (f(\mathbf{x}_{k-1}) - \min_{\mathbf{x}\in R^d} f(\mathbf{x})) \\
&\vdots \\
&\leq \left(1 - \frac{\mu}{L}\right)^k (f(\mathbf{x}_1) - \min_{\mathbf{x}\in R^d} f(\mathbf{x}))
\end{aligned}
$$

□

Remark: The optimality gap at the next iteration $k+1$ is bounded by $\left(1-\frac{\mu}{L}\right)$ times the current optimality gap at iteration k, and is bounded by $\left(1-\frac{\mu}{L}\right)^k$ times the gap at iteration 1.

Bibliographic notes

More to read on Chapter 3 and Chapter 4 of [Drusvyatskiy (2020)] and Chapter 3 and Chapter 4.2 of [Sidford (2024)] and Chapter 6 of [Vishnoi (2021)]

References

[Drusvyatskiy (2020)] Dmitriy Drusvyatskiy. Convex Analysis and Nonsmooth Optimization. 2020.

[Sidford (2024)] Aaron Sidford. Optimization Algorithms. 2024.

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