COMP 7070 Advanced Topics in AI and ML January 30, 2024

Lecture 4: Optimization

Instructor: Yifan Chen *Scribes: Xiong Peng, Xuanzhe Xiao* *Proof reader: Zhanke Zhou*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

4.1 Preliminaries

To start this lecture, we will introduce some concepts related to optimization.

4.1.1 Convexity

Definition 4.1 (Convex domain)**.** *Let $\mathcal{D}$ be a domain. For any points $x, y \in \mathcal{D}$ and any constant $\lambda \in [0,1]$, if $\lambda x + (1-\lambda)y \in \mathcal{D}$, then $\mathcal{D}$ is a convex domain.*

The expression $\lambda x + (1-\lambda)y$ is called a convex combination. This definition tells us that the convex combination of points within a convex domain also lies within that convex domain.

The second concept to be introduced next is convex functions.

Definition 4.2 (Convex function)**.** *Assuming $\mathcal{D}$ is a domain, f is a function defined on $\mathcal{D}$. For any points $x, y \in \mathcal{D}$, a function f is a convex function if it satisfies the following condition:*

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda \cdot f(x) + (1-\lambda) \cdot f(y).$$

For a convex function, the value of the convex combination of two points in its domain will be less than or equal to the value of the same convex combination of the function values of these two points. Figure 4.1 is an example of a convex function.

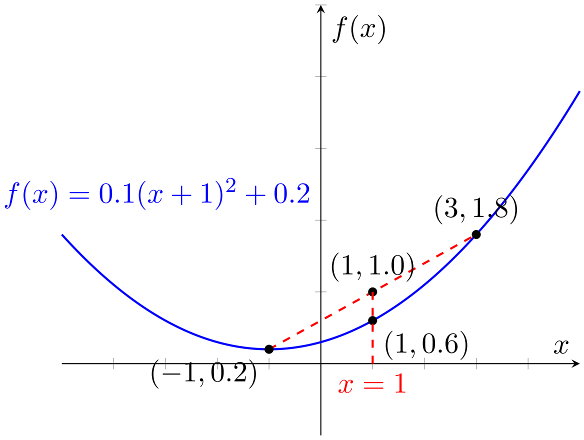

Figure 4.1: An example of a convex function

4.1.2 L-smoothness

Definition 4.3. *Let f be a continuously differentiable function defined on $\mathcal{D}$, meaning its derivative ∇f is continuous. If ∇f is L-Lipschitz, satisfying the following condition for any points $x, y \in \mathcal{D}$, then f is L-smooth:*

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L \cdot \|x - y\|.$$

This implies that the derivative ∇f of a function that satisfies L-smoothness does not change too rapidly. The change in the derivative is bounded by the change in the independent variable of f.

L-smoothness is very common in machine learning. Next, we will take the example of least squares as an illustration.

Example. (Least square) Consider the objective function of least squares: $f(x) = \frac{1}{n}\|Ax - b\|^2$. Next, we will prove that it is L-smooth and provide the corresponding L value. Firstly, we can calculate its derivative:

$$\nabla f(x) = \frac{2}{n}A^\top(Ax - b).$$

Next, according to the definition of L-smoothness, for any x and y, we need to bound $\|\nabla f(x) - \nabla f(y)\|$ with $\|x - y\|$:

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\| &= \frac{2}{n}\left\|A^\top Ax - A^\top Ay\right\| \\ &\leqslant \frac{2}{n}\left\|A^\top A\right\| \cdot \|x - y\| \\ \Rightarrow L = \frac{2}{n}\left\|A^\top A\right\| &= \frac{2}{n}\|A\|^2. \end{aligned}$$

So, f is $\frac{2}{n}\|A\|^2$-smooth.

4.1.3 Difference Quotient

In order to facilitate the subsequent proofs, we will now introduce some relevant concepts and conclusions.

Definition 4.4 (difference quotient)**.** *For a function f and a given point x, the difference quotient is defined as a function of y as follows:*

$$\phi(y) := \frac{f(y) - f(x)}{\|y - x\|}.$$

Next, we will prove an important property of the difference quotient: monotonicity on a line.

Proposition 4.5 (The monotonicity of the difference quotient in a line)**.** *For the difference quotient $\phi(y_1)$ of function f with respect to a given point x, where y_2 is an arbitrary point in the domain of f, and y_1 lies between x and y_2 such that $y_1 = (1-t)x + t \cdot y_2$, with $t \in (0,1)$, it can be deduced that:*

$$\phi(y_1) \leqslant \phi(y_2).$$

Proof.

$$\begin{aligned} \phi(y_1) &= \frac{f(y_1) - f(x) = f((1-t)x + t \cdot y_2) - f(x)}{\|y_1 - x\| = t \cdot \|y_2 - x\|} \\ &\leqslant \frac{(1-t) \cdot f(x) + t \cdot f(y_1) - f(x)}{t \cdot \|y_2 - x\|} \\ &= \frac{f(y_2) - f(x)}{\|y_2 - x\|} = \phi(y_2). \end{aligned}$$

□

4.1.4 Subgradient

If a function is convex and differentiable, then its tangent line at any point always lies below the function, as shown in Figure 4.2. Next, we will prove this proposition.

Proposition 4.6. *If the function f is convex and differentiable on $\mathcal{D}$, then for any $x, y \in \mathcal{D}$, we have:*

$$f(x) + \langle \nabla f(x), y \cdot x \rangle \leqslant f(y).$$

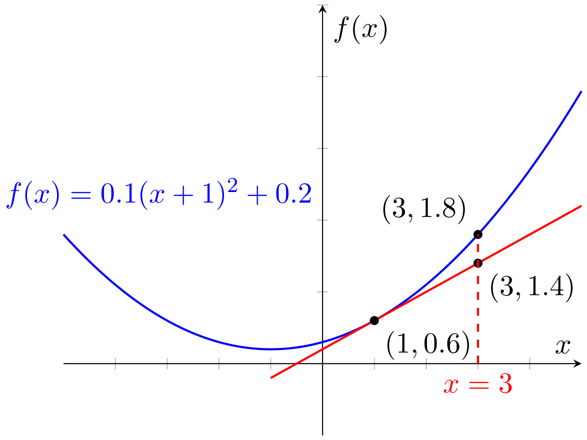

Figure 4.2: The tangent line of a convex function always lies below it.

Proof. According to the definition of a convex function, for any $\alpha \in (0, 1]$, we have:

$$f(x + a(y - x)) \leqslant af(y) + (1 - \alpha)f(x).$$

This inequality can be rearranged as follows:

$$\frac{f(x + a(y - x)) - f(x)}{a} \leqslant f(y) - f(x).$$

Letting $\alpha \to 0$, we obtain:

$$\langle \nabla f(x), y - x \rangle \leqslant f(y) - f(x).$$

□

This proposition is practical. However, regrettably, not all convex functions exhibit differentiability at every point within their domain (for instance, the ReLU function). Nevertheless, we still desire to have similar inequalities in such convex functions. Therefore, subgradients are introduced to extend the aforementioned proposition.

Definition 4.7 (Subgradient)**.** *Let f be a convex function on $\mathcal{D}$. For any interior point $x \in \mathrm{int}(\mathcal{D})$, and for any $y \in \mathcal{D}$, if g_x satisfies the following condition, then it is a subgradient of f:*

$$f(y) \geqslant f(x) + \langle g_x, y - x \rangle .$$

Subgradients are designed for convex functions. Subgradients are not always unique. We can define a set, called the subdifferential set, for a convex function f, which includes all subgradients of f that satisfy the definition.

Besides, we can prove that even for non-differentiable convex functions, such subgradients always exist.

Proposition 4.8 (Existence of subgradients for convex functions)**.** *For a convex function f, the subgradient g_x always exists.*

Before the beginning of the proof, we need to introduce the concept of the epigraph.

Definition 4.9 (epigraph)**.** *For a function $f : \mathbb{R}^n \to \mathbb{R}$, its epigraph is defined as:*

$$\mathrm{epi}(f) = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq t\}.$$

For a convex function f, its epigraph is a convex set. To prove the proposition 4.8, we also need to introduce the Supporting Hyperplane Theorem.

Lemma 4.10 (Supporting Hyperplane Theorem on Epigraph)**.** *Let f be a convex function. For a point x on the boundary of the epigraph* epi(f) *of f, there exist constants a_1, a_2, b such that:*

1. $\langle a_1, x\rangle + a_2 f(x) = b$*;*

2. $\forall (y,t) \in \operatorname{epi}(f), \langle a_1, y\rangle + a_2 t \geqslant b$*.*

Proof of Proposition 4.8. According to Lemma 4.10, it can be concluded that for a convex function f, there exist constants a_1, a_2, b that satisfy both conditions in the lemma. Furthermore, we can infer that $a_2 \geqslant 0$. Otherwise, for the second condition, as t goes to positive infinity, $\langle a_1, y\rangle + a_2 t$ will go to negative infinity, and any value of b will be bigger than it.

Additionally, a_2 cannot be equal to 0. If $a_2 = 0$, then according to the first condition, $\langle a_1, x\rangle = b$, and we only need to find a y that is smaller than x to make the second condition invalid.

From this, we can deduce that $a_2 > 0$. Now, let's complete the proof. First, for any $y \in \mathcal{D}$, $(y, f(y)) \in \operatorname{epi}(f)$, so $\langle a_1, y\rangle + a_2 f(y) \geqslant b$. Dividing both sides by a_2, we have:

$$\left\langle \frac{a_1}{a_2}, y \right\rangle + f(y) \geqslant \frac{b}{a_2}.$$

Next, x is an interior point of $\mathcal{D}$, so it lies on the boundary of $\operatorname{epi}(f)$. Therefore, $\langle a_1, x\rangle + a_2 f(x) = b$. Similarly, dividing both sides by a_2, we have:

$$\left\langle \frac{a_1}{a_2}, x \right\rangle + f(x) = \frac{b}{a_2}.$$

Combining these two equations, we have:

$$\begin{aligned} \left\langle \frac{a_1}{a_2}, y \right\rangle + f(y) &\geqslant \left\langle \frac{a_1}{a_2}, x \right\rangle + f(x) \\ f(y) &\geqslant f(x) + \left\langle -\frac{a_1}{a_2}, y - x \right\rangle. \end{aligned}$$

Thus, we have constructed a subgradient of f at x, denoted as $g_x = -\dfrac{a_1}{a_2}$, which proves its existence. □

Finally, we prove that when the function f possesses L-smoothness property, it also satisfies a similar inequality in the opposite direction.

Proposition 4.11 (Indication of L-smoothness)**.** *If f is L-smooth, then*

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

Proof.

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + \tau(y-x)), y - x\rangle d\tau \\ &= f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x\rangle d\tau \\ &\leqslant f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \|\nabla f(x + \tau(y-x)) - \nabla f(x)\| \cdot \|y - x\| d\tau \\ &\leqslant f(x) + \langle \nabla f(x), y - x\rangle + \|y - x\| \int_0^1 L\tau \, \|y - x\| d\tau \\ &= f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \end{aligned}$$

The first inequality comes from the Cauchy-Schwarz inequality, and the second inequality comes from the property of L-smoothness. Therefore, the proposition is proven. □

4.2 GD - Setting 1

Consider the iterative update rule in Gradient Descent (GD):

$$x_{i+1} = x_i - t\nabla f(x_i),$$

where t is a fixed step size satisfying $t \leq \frac{1}{L}$, for an L-smooth and convex function f.

(I) From the convexity of f, we have:

$$f(x_i) \leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle.$$

(II) Using the L-smoothness property:

$$f(x_{i+1}) \leq f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \frac{L}{2}\|x_{i+1} - x_i\|^2.$$

Substituting $x_{i+1} - x_i = -t\nabla f(x_i)$:

$$f(x_{i+1}) \leq f(x_i) - t\|\nabla f(x_i)\|^2 + \frac{L}{2}t^2\|\nabla f(x_i)\|^2.$$

Since $t \leq \frac{1}{L}$, it simplifies to:

$$f(x_{i+1}) \leq f(x_i) - \frac{t}{2}\|\nabla f(x_i)\|^2.$$

Monotonicity of $f(x_{i+1})$ for GD:

(III) From the convexity of f:

$$f(x_{i+1}) \leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle - \frac{t}{2}\|\nabla f(x_i)\|^2.$$

Expanding $\nabla f(x_i) = \frac{1}{t}(x_i - x_{i+1})$:

$$f(x_{i+1}) \leq f(x^*) + \frac{1}{t}\langle x_i - x_{i+1}, x_i - x^* \rangle - \frac{1}{2t}\|x_i - x_{i+1}\|^2.$$

Rewriting the inner product term:

$$f(x_{i+1}) \leq f(x^*) + \frac{1}{2t}\|x_i - x^*\|^2 - \frac{1}{2t}\|x_{i+1} - x^*\|^2.$$

Summing over k iterations and simplifying:

$$\sum_{i=0}^{k-1} f(x_{i+1}) - f(x^*) \leq \frac{1}{2t}\|x_0 - x^*\|^2 - \frac{1}{2t}\|x_k - x^*\|^2.$$

Thus, the averaged sum of function values converges to the optimal value:

$$\frac{1}{k}\sum_{i=0}^{k-1} f(x_{i+1}) - f(x^*) \leq \frac{1}{2t}\frac{1}{k}\|x_0 - x^*\|^2,$$

which gives us the convergence rate of:

$$f(x_k) - f(x^*) \leq \frac{1}{2t}\frac{1}{k}\|x_0 - x^*\|^2 = O\left(\frac{1}{k}\right).$$

4.3 GD - Setting 2

For an L-smooth and μ-strongly convex function, the Gradient Descent (GD) algorithm with step size $t \leq \frac{1}{L}$ is considered.

(I) From the definition of μ-strong convexity, we have:

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

(II) By taking $y = x_i - \frac{1}{\mu}\nabla f(x_i)$, we get:

$$f(y) \geq f(x_i) - \frac{1}{2\mu}\|\nabla f(x_i)\|^2.$$

(III) By the L-smoothness property (from Setting 1), we also have:

$$f(x_{i+1}) \leq f(x_i) - \frac{t}{2}\|\nabla f(x_i)\|^2.$$

Combining the two properties:

$$f(x_{i+1}) - f(x^*) \leq f(x_i) - f(x^*) - \frac{t}{2}\|\nabla f(x_i)\|^2.$$

(IV) From the μ-strong convexity again:

$$\|\nabla f(x_i)\|^2 \geq 2\mu(f(x_i) - f(x^*)).$$

(V) Combining **III** and **IV**:

$$f(x_{i+1}) - f(x^*) \leq (1 - t\mu)(f(x_i) - f(x^*))$$

Since $\mu \leq L$, we have $t\mu \leq tL \leq 1$, ensuring the convergence of the sequence $\{f(x_i)\}$ to $f(x^*)$. The condition number κ is defined as $\kappa = \frac{L}{\mu} \geq 1$, which influences the convergence rate.

4.4 SGD

Consider the Stochastic Gradient Descent (SGD) update rule:

$$x_{i+1} = x_i - t \cdot V_i,$$

where $\mathbb{E}[V_i] = \nabla f(x_i)$ and $\text{tr}(\text{Var}(V_i)) = \sigma^2 < \infty$, and t is a fixed step size satisfying $t \leq \frac{1}{L}$, for an L-smooth and convex function f.

L-smoothness: Using the L-smoothness of f, we have:

$$f(x_{i+1}) \leq f(x_i) + \langle \nabla f(x_i), x_{i+1} - x_i \rangle + \frac{L}{2}\|x_{i+1} - x_i\|^2.$$

Taking expectations conditioned on x_i and using the SGD update rule:

$$\mathbb{E}[f(x_{i+1})] \leq f(x_i) - t\|\nabla f(x_i)\|^2 + \frac{t^2 L}{2}\mathbb{E}[\|V_i\|^2].$$

Since $\mathbb{E}[\|V_i\|^2] = \|\nabla f(x_i)\|^2 + \sigma^2$ and $Lt \leq 1$, the inequality simplifies to:

$$\mathbb{E}[f(x_{i+1})] \leq f(x_i) - \frac{t}{2}\|\nabla f(x_i)\|^2 + \frac{t}{2}\sigma^2.$$

Convexity: From convexity, we have:

$$f(x_i) \leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle.$$

Combining this with the previous expectation, we obtain:

$$\begin{aligned}
\mathbb{E}[f(x_{i+1})] &\leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle - \frac{t}{2}\|\nabla f(x_i)\|^2 + \frac{t}{2}\sigma^2 && (4.1)\\
&\leq f(x^*) + \mathbb{E}[\langle V_i, x_i - x^* \rangle - \frac{t}{2}\|V_i\|^2 + \frac{t}{2}\sigma^2] + \frac{t}{2}\sigma^2 && (4.2)\\
&= f(x^*) + \frac{1}{2t}\mathbb{E}[-\|x_{i+1} - x^*\|^2 + \|x_i - x^*\|^2] + t\sigma^2. && (4.3)
\end{aligned}$$

$$\mathbb{E}[f(x_{i+1}) - f(x^*)] \leq \frac{1}{2t}\mathbb{E}[-\|x_{i+1} - x^*\|^2 + \|x_i - x^*\|^2] + t\sigma^2.$$

Summing over k iterations and rearranging terms gives the convergence rate:

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \frac{1}{2tk}\|x_0 - x^*\|^2 + t\sigma^2,$$

where $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ is the average of the iterates.

This implies that the convergence rate is $O\left(\frac{1}{\sqrt{k}}\right)$ for a constant step size t satisfying $t \leq \frac{1}{L}$, and $t \approx \frac{1}{\sqrt{k}}$ for diminishing step size, which is slower compared to the rate of convergence for Gradient Descent in a non-stochastic setting.