COMP 7070 Advanced Topics in AI and ML Feburary 12, 2025

Lecture 4: Convex Optimization and the Convergence Rate of GD/SGD

Instructor: Yifan Chen *Scribes: Zheng Wu, Runhao Jiang* *Proof reader: Xiong Peng*

4.1 Convexity

4.1.1 Convex Function and Convex Domain

1. **Convex Function:** A function f is called convex, if for all x, $y \in \text{dom}\{f\}$ and for $\forall \lambda \in [0,1]$:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{I}$$

2. **Convex Domain:** A domain $\mathcal{D}$ is called convex, if for $\forall \lambda \in [0,1]$, $\forall x, y \in \mathcal{D}$:

$$\lambda x + (1-\lambda)y \in \mathcal{D}.$$

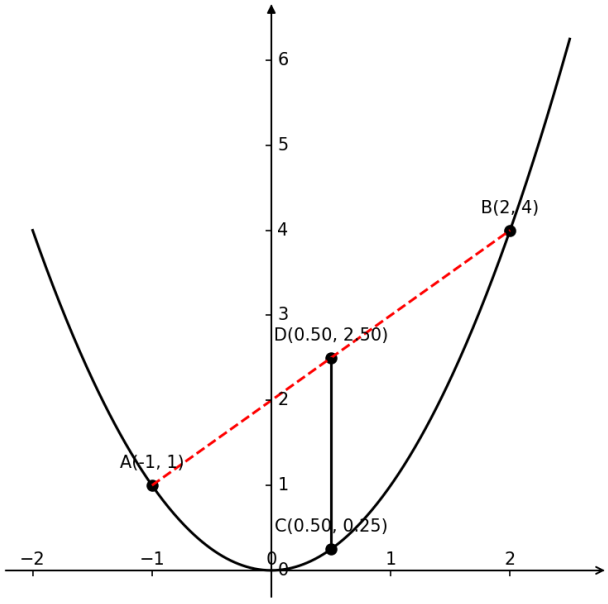

Figure 4.1: An example of a convex function

4.1.2 L-Lipschitz and L-Smoothness

1. **L-Lipschitz:** A function f satisfies $\forall x, y \in \mathcal{D}$,

$$|f(x) - f(y)| \leq L\|x - y\|.$$

2. **L-Smoothness:** For a continuously differentiable function f, it is L-smooth if $\forall x, y \in \mathcal{D}$

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

3. **Example: Least squares**

   - **Function:**
   $$f(x) = \frac{1}{2n}\|Ax - b\|^2$$
   - **Gradient:**
   $$\nabla f(x) = \frac{1}{n}A^\top(Ax - b).$$
   - **Gradient Difference:**
   $$\|\nabla f(x) - \nabla f(y)\| = \frac{1}{n}\|A^\top A(x - y)\| \leq \frac{\|A^\top A\|}{n}\|x - y\|.$$
   - **Conclusion:** The function f is $\frac{\|A\|^2}{n}$-smooth.

4.1.3 Difference Quotient Analysis

1. **Difference Quotient**
   For a function f, define the **difference quotient** as:
   $$\varphi(y; x) = \frac{f(y) - f(x)}{\|y - x\|}.$$

2. **Monotonicity of the Difference Quotient**
   If f is a convex function, then the difference quotient $\varphi(y; x)$ is non-decreasing as y moves further from x along any line. Specifically, for points along the parameterized line: $y_1 = ty_2 + (1 - t)x$, *with* $t \in (0, 1)$.

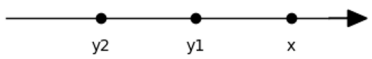

Figure 4.2: An example of points along the parameterized line

   - **Claim:**
   $$\varphi(y_1, x) \leq \varphi(y_2, x).$$
   - **Left-Hand Side (LHS):**
   $$\varphi(y_1, x) = \frac{f(ty_2 + (1 - t)x) - f(x)}{\|ty_2 - tx\|} = \frac{f(ty_2 + (1 - t)x) - f(x)}{t\|y_2 - x\|}.$$
   - **Using Convexity:** By convexity of f:
   $$f(ty_2 + (1 - t)x) \leq tf(y_2) + (1 - t)f(x),$$
   which implies:
   $$\frac{f(ty_2 + (1 - t)x) - f(x)}{t\|y_2 - x\|} \leq \frac{tf(y_2) + (1 - t)f(x) - f(x)}{t\|y_2 - x\|}.$$
   - **Simplifying:**
   $$\frac{t\big(f(y_2) - f(x)\big)}{t\|y_2 - x\|} = \frac{f(y_2) - f(x)}{\|y_2 - x\|} = \varphi(y_2, x).$$
   - **Conclusion:**
   $$\varphi(y_1, x) \leq \varphi(y_2, x) \text{ holds for all } t \in (0, 1)$$
   .

4.1.4 Subgradient Inequality for Convex Functions

For a convex differentiable function f, it satisfies $\forall x, y \in \mathcal{D}$:

$$f(x) + \langle \nabla f(x), y - x \rangle \leq f(y). \tag{II}$$

Proof of Subgradient Inequality

- **Starting Inequality:** From the definition of convexity with $\alpha \in (0, 1)$:

$$f\big(x + \alpha(y - x)\big) \leq \alpha f(y) + (1 - \alpha) f(x).$$

- **Rearranging Terms:** Subtract $f(x)$ and divide by α:

$$\frac{f\big(x + \alpha(y - x)\big) - f(x)}{\alpha} \leq f(y) - f(x).$$

- **Take Limit $\alpha \to 0$:** By the definition of directional derivative:

$$\lim_{\alpha \to 0} \frac{f\big(x + \alpha(y - x)\big) - f(x)}{\alpha} = \langle \nabla f(x), y - x \rangle.$$

  Therefore:

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

This directly yields the subgradient inequality (II).

4.1.5 Subgradient Theory

For a **non-differentiable convex function** f, a vector $g(x)$ is called a subgradient at x if:

$$f(x) + \langle g(x), y - x \rangle \leq f(y), \quad \forall y \in \mathcal{D}.$$

4.1.5.1 Epigraph Representation

The **epigraph** of f, defined as:

$$\text{epi}(f) = \big\{(y, t) \in \mathcal{D} \times \mathbb{R} \mid t \geq f(y)\big\},$$

is always a convex set for convex functions.

4.1.5.2 Supporting Hyperplane Theorem on Epigraph

Let f be a convex function. For a point x on the boundary of the epigraph $\text{epi}(f)$ of f, there exist constants a_1, a_2, b such that:

1. $a_1^\top x + a_2 f(x) = b$;
2. $a_1^\top y + a_2 t \geq b, \quad \forall (y, t) \in \text{epi}(f)$.

- **Non-Negativity of a_2:** If $a_2 < 0$, taking $t \to +\infty$ would violate the inequality.
- **Non-Triviality Condition:** If $a_2 = 0$, then $a^\top y \geq b$ must hold for all $y \in \mathcal{D}$, which generally requires $\mathcal{D}$ to be bounded.

Rescaling the Hyperplane

By normalizing with $a_2 = 1$, the inequality becomes:

$$a_1^\top y + t \geq b = a_1^\top x + f(x).$$

Substituting this back, we get:

$$a_1^\top (y - x) + f(x) \leq f(y), \quad \forall y \in \mathcal{D}.$$

This implies the subgradient $g(x) = a$ satisfies:

$$f(x) + \langle g(x), y - x \rangle \leq f(y), \quad \forall y \in \mathcal{D}.$$

4.1.6 Analysis for L-Smooth Functions

For L-smooth convex functions, the quadratic upper bound holds:

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

4.1.6.1 Derivation of Quadratic Upper Bound

- **Integral Representation:** Start from the fundamental theorem of calculus:

$$f(y) = f(x) + \int_0^1 \big\langle \nabla f(x + \tau(y - x)), y - x \big\rangle d\tau.$$

  Decompose into:

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \big\langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \big\rangle d\tau.$$

- **Applying L-Smoothness:** Using $\|\nabla f(u) - \nabla f(v)\| \leq L\|u - v\|$:

$$\leq f(x) + \langle \nabla f(x), y - x \rangle + \|y - x\| \int_0^1 L\tau \|y - x\| d\tau.$$

  Simplify the integral:

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

4.1.6.2 Convergence Rate of GD for L-smooth and convex objectives

Update Rule:

$$x_{i+1} = x_i - t\nabla f(x_i), \quad t > 0.$$

Key Steps in Convergence Proof

- **Step 1: Subgradient Inequality (II)** From the convexity of f:

$$f(x_i) \leq f(x^*) + \langle \nabla f(x_i), x_i - x^* \rangle. \tag{II-Applied}$$

- **Step 2: Quadratic Upper Bound (III)** Using the L-smoothness property at iteration i:

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

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

$$= f(x_i) - t\|\nabla f(x_i)\|^2 + \frac{L}{2}t^2\|\nabla f(x_i)\|^2.$$

  Simplify under step-size $t \leq 1/L$:

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

- **Step 3: Telescoping Sum** Combine Steps 1-2 and telescope:

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

Substitute $x_{i+1} = x_i - t\nabla f(x_i)$, and manipulate squares:

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

Rearrange to obtain:

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

Thus:

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

- **Final Rate:** Accumulate over k iterations:

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

For the minimal iterate (or average):

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

4.1.7 Convergence Rate Analysis for Gradient Descent

For L-smooth functions with an additional quadratic lower bound (implying strong convexity with parameter M), the convergence rate of gradient descent can be rigorously characterized.

Key Recursive Inequality

Starting from the descent guarantee for gradient descent:

$$\boxed{f(x_{i+1}) \leq f(x_i) - \frac{t}{2}\|\nabla f(x_i)\|^2} \quad \text{(from Step 2 in Section 4).}$$

Quadratic Lower Bound

For μ-strongly convex functions ($\star$), the lower bound holds:

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

- **Minimizer Characterization:** Setting $y = x_i - \frac{1}{\mu}\nabla f(x_i)$, substituting into the lower bound gives:

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

- **Optimality Condition at x^*:** combine with the above:

$$f(x^*) \geq f(x_i) - \frac{1}{2\mu}\|\nabla f(x_i)\|^2 \quad \Longrightarrow \quad \|\nabla f(x_i)\|^2 \geq 2\mu\big[f(x_i) - f(x^*)\big].$$

Linear Convergence Rate

Substitute the gradient norm bound into the descent inequality:

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

Simplifying yields the contraction factor:

$$\boxed{f(x_{i+1}) - f(x^*) \leq (1 - \mu t)\big[f(x_i) - f(x^*)\big]}.$$

Remarks on Contraction Factor

- **Parameter Constraints:** For $\mu \leq L$, the contraction factor $\kappa = 1 - \mu t$ satisfies $\kappa < 1$ when $t < \frac{1}{\mu}$.
- **Geometric Convergence:** Telescoping over k iterations:

$$f(x_k) - f(x^*) \leq (1 - \mu t)^k \big[f(x_0) - f(x^*)\big] \quad \text{(linear rate)}.$$

- **Step Size Requirement:** The step size $t \leq \frac{1}{L}$ ensures both L-smoothness and μ-strong convexity conditions are satisfied.

4.1.8 Stochastic Gradient Descent Convergence

Update Rule:

$$x_{i+1} = x_i - t \cdot v_i \quad \text{where} \quad \mathbb{E}[v_i] = \nabla f(x_i),\ \mathrm{Tr}(\mathrm{Var}(v_i)) \leq \sigma^2.$$

- **Quadratic Upper Bound with SGD:** Using the L-smooth analysis:

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

substitute $x_{i+1} - x_i = -tv_i$:

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

- **Expectation Bound:** Take expectation over v_i:

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

Decompose $\mathbb{E}\|v_i\|^2 = \|\nabla f(x_i)\|^2 + \mathrm{Tr}(\mathrm{Var}(v_i))$:

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

Simplify:

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

- **Telescoping Inequality:** Accumulate over K iterations:

$$\frac{1}{K}\sum_{i=0}^{K-1} \mathbb{E}[f(x_i) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2}{2tK} + t\sigma^2.$$

Choose step size $t = \frac{1}{\sqrt{K}}$:

$$\mathbb{E}f(\bar{x}_K) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\sqrt{K}} + \frac{\sigma^2}{2\sqrt{K}} = O\left(\frac{1}{\sqrt{K}}\right),$$

where $\bar{x}_K = \frac{1}{K}\sum_{i=0}^{K-1} x_i$ (via Jensen's inequality).