Basic Convex Optimization

Algorithm and Duality

Kyoungman Kim

Sungkyunkwan University

Department of Mathematics

MIMIC Seminar
May 7, 2025

Overviews

- Convex optimization problems emerge across various disciplines.
- There is no general analytical formula for the solution of convex optimization problmes, but there are very effective methods for solving them. In this seminar, we shall restrict our discussion to First-Order Method.
- Lastly, we will take a brief look at the Lagrangian Duality

Definitions

Definition (Convex Set)

A set $\Omega \subseteq \mathbb{R}^n$ is *convex* if for all $x, y \in \Omega$ and for all $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda) y \in \Omega.$$

A point of the form $\lambda x + (1 - \lambda) y$, where $\lambda \in [0, 1]$, is called a *convex combination* of x and y. As λ varies over $[0, 1]$, a line segment is formed between x and y.

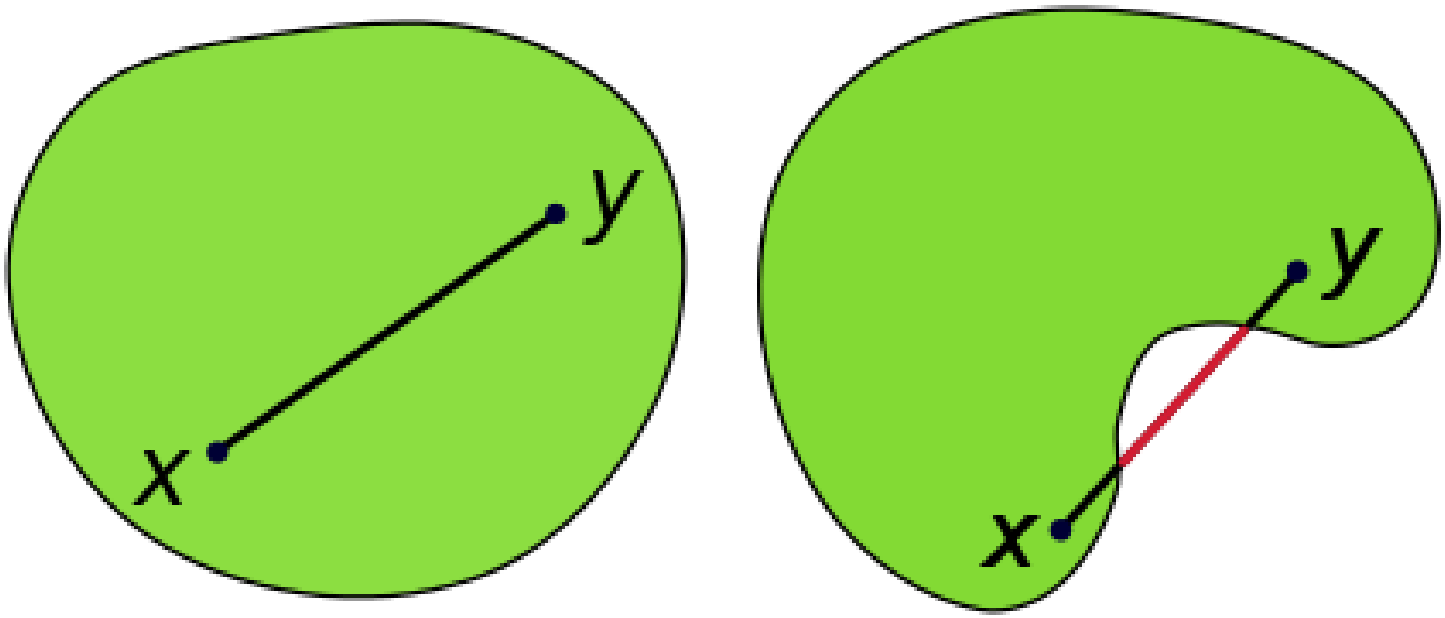

Figure: Convex Set

Definitions

Definition (Convex Function)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if its domain $\mathrm{dom}(f)$ is a convex set and, for all $x, y \in \mathrm{dom}(f)$ and for all $\lambda \in [0, 1]$, we have

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

[Remark]
Geometrically, the line segment connecting

$$(x, f(x)) \quad \text{to} \quad (y, f(y))$$

should lie above the graph of the function.

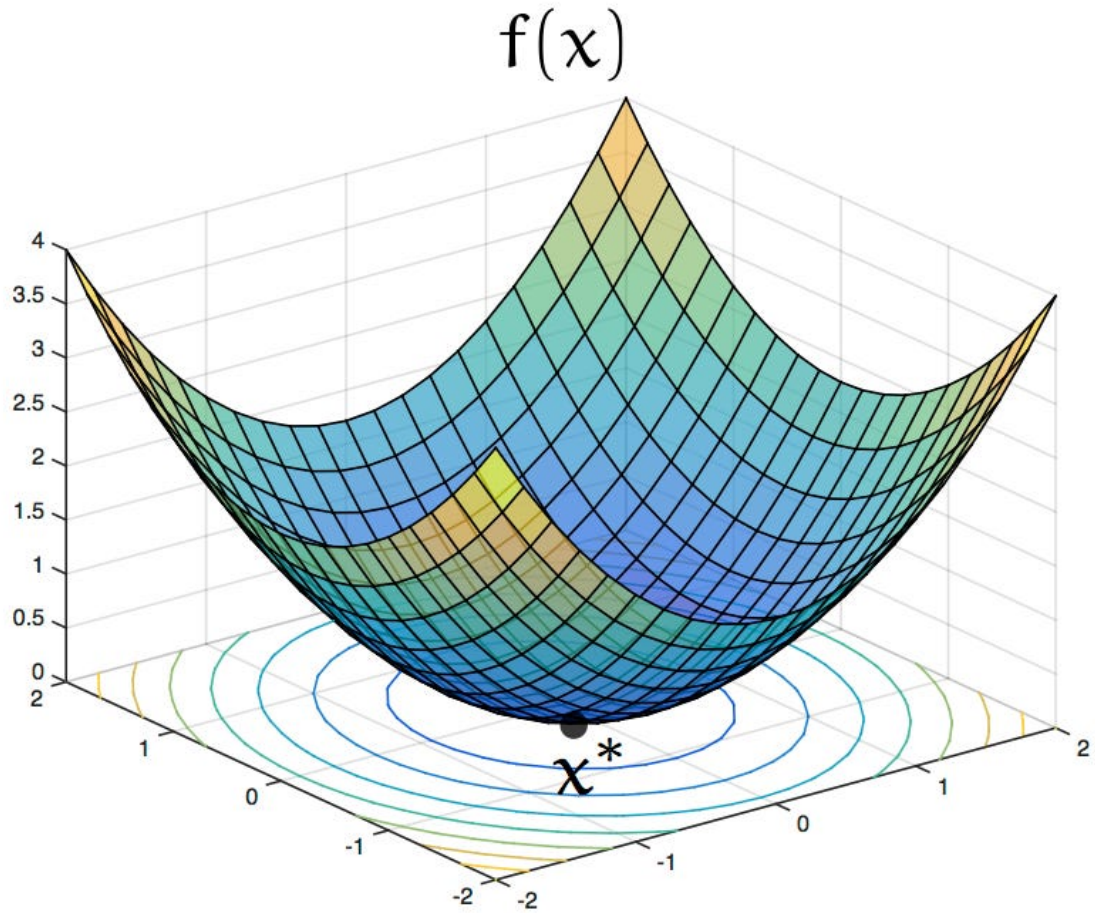

Figure: Convex Function

Definition: Local Optimality

Definition

A feasible solution x^* is **locally optimal** to the optimization problem

$$\text{minimize } f(x) \quad \text{subject to } x \in \Omega$$

if there exists $R > 0$ such that

$$f(x^*) = \min \{ f(x) : x \in \Omega,\ \|x - x^*\| \leq R \}.$$

Nice properties: Local Minimum is Global

Proof.

Suppose x is locally optimal, but there exists a feasible point $y \in \Omega$ such that

$$f_0(y) < f_0(x).$$

Since x is locally optimal, there exists $R > 0$ such that for all feasible z,

$$\|z - x\|_2 \leq R \quad \Rightarrow \quad f_0(z) \geq f_0(x).$$

Consider the point

$$z = \theta y + (1 - \theta)x, \quad \text{where} \quad \theta = \frac{R}{2\|y - x\|_2}.$$

Nice properties: Local Minimum is Global

continued.

- Since $\|y - x\|_2 > R$, we have $0 < \theta < \frac{1}{2}$.
- z is a convex combination of two feasible points, so $z \in \Omega$.
- $\|z - x\|_2 = \frac{R}{2} < R$, so it lies in the neighborhood where x is locally optimal.
- By convexity of f_0,

$$f_0(z) \leq \theta f_0(y) + (1 - \theta) f_0(x) < f_0(x),$$

which contradicts the assumption that x is locally optimal.

Hence, x must be a global minimum. □

Nice properties : Optimality Condition

Theorem

A point $x^* \in \mathbb{R}^n$ is optimal for the unconstrained problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

if and only if

$$\nabla f(x^*) = 0.$$

[Remark] The presence of constraints makes it significantly more challenging to establish general conditions

Optimization Problem in Standard Form

Consider the following constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

- $x \in \mathbb{R}^n$: optimization variable
- $f_0 : \mathbb{R}^n \to \mathbb{R}$: objective function CONVEX
- $f_i : \mathbb{R}^n \to \mathbb{R}$: inequality constraint functions CONVEX
- $h_i : \mathbb{R}^n \to \mathbb{R}$: equality constraint functions AFFINE

The optimal value is denoted by:

$$x^* = \inf \{ f_0(x) \mid f_i(x) \leq 0,\ h_i(x) = 0 \}$$

Example of Convex Optimization(1)

Example

A **least-squares problem** is an unconstrained optimization problem where the objective is a sum of squared terms of the form $a_i^T x - b_i$:

$$\min f_0(x) = \|Ax - b\|_2^2 = \sum_{i=1}^{k} (a_i^T x - b_i)^2,$$

where $A \in \mathbb{R}^{k \times n}$ with $k \geq n$, each row a_i^T is a row of the matrix A, and the optimization variable is $x \in \mathbb{R}^n$.

Example of Convex Optimization (1)

Example (Analytical Solution of L-S Problem)

The optimal solution to the least-squares problem is given analytically by:

$$x^* = (A^T A)^{-1} A^T b,$$

assuming $A^T A$ is invertible (i.e., A has full column rank).

Example of Convex Optimization (2)

Example (Linear Programming (LP))

Consider the linear program:

$$\min c^T x \quad \text{subject to } a_i^T x \leq b_i, \quad i = 1, \ldots, m,$$

where $c, a_1, \ldots, a_m \in \mathbb{R}^n$ are vectors and $b_1, \ldots, b_m \in \mathbb{R}$ are scalars.

Gradient Descent: Basic Iteration

Algorithm:Gradient Descent

- **Step 1:** Choose an initial point $x_1 \in \mathbb{R}^n$.
- **Step 2:** For $t = 1, 2, \ldots, N$ (for some large $N \in \mathbb{N}$), compute

$$x_{t+1} = x_t - \eta \nabla f(x_t),$$

where $\eta > 0$ is the step size (also called learning rate).

[Remark] We will focus on proving the convergence of the algorithm rather than implementing it through coding.

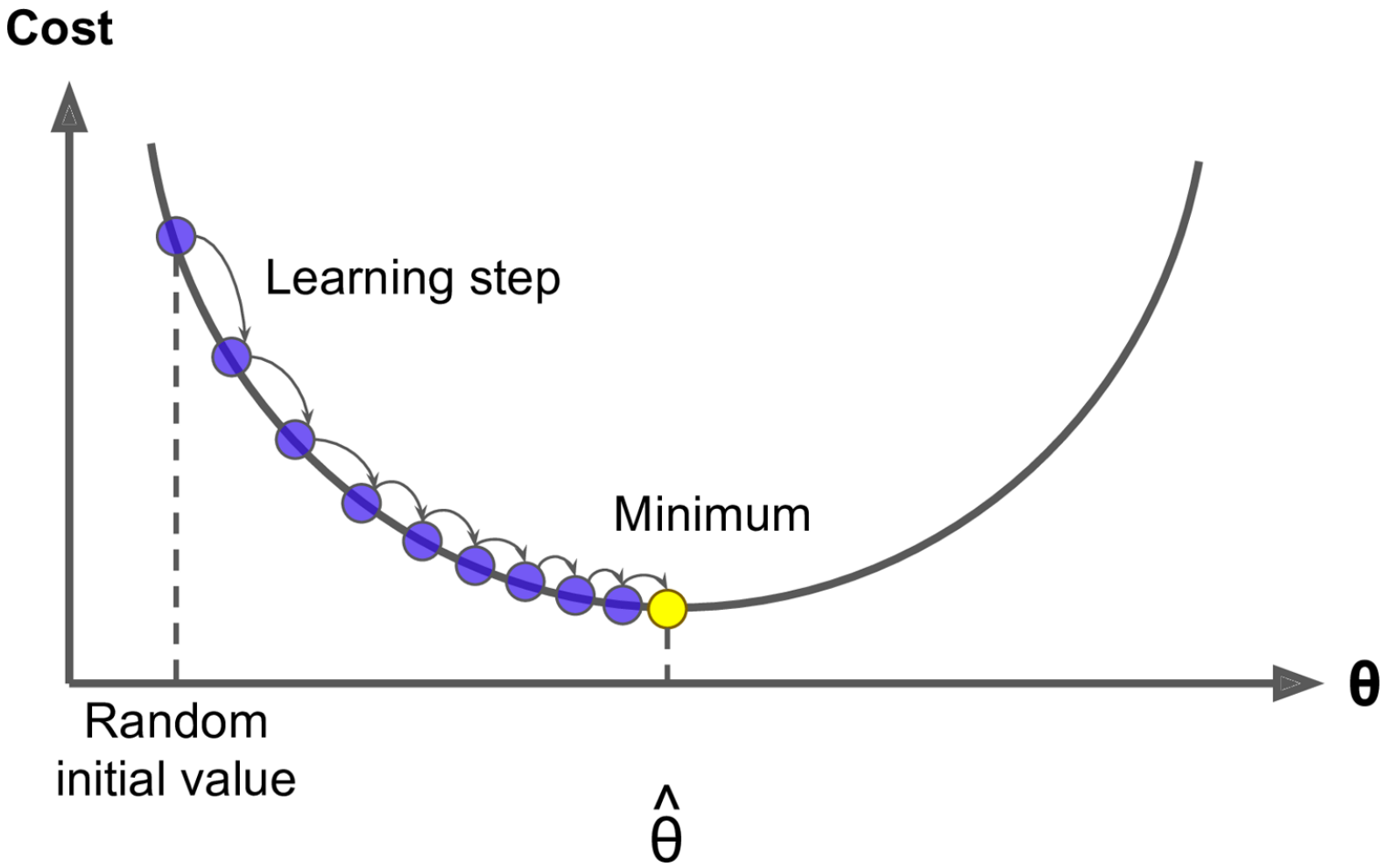

Figure: Gradient Descent in 2D

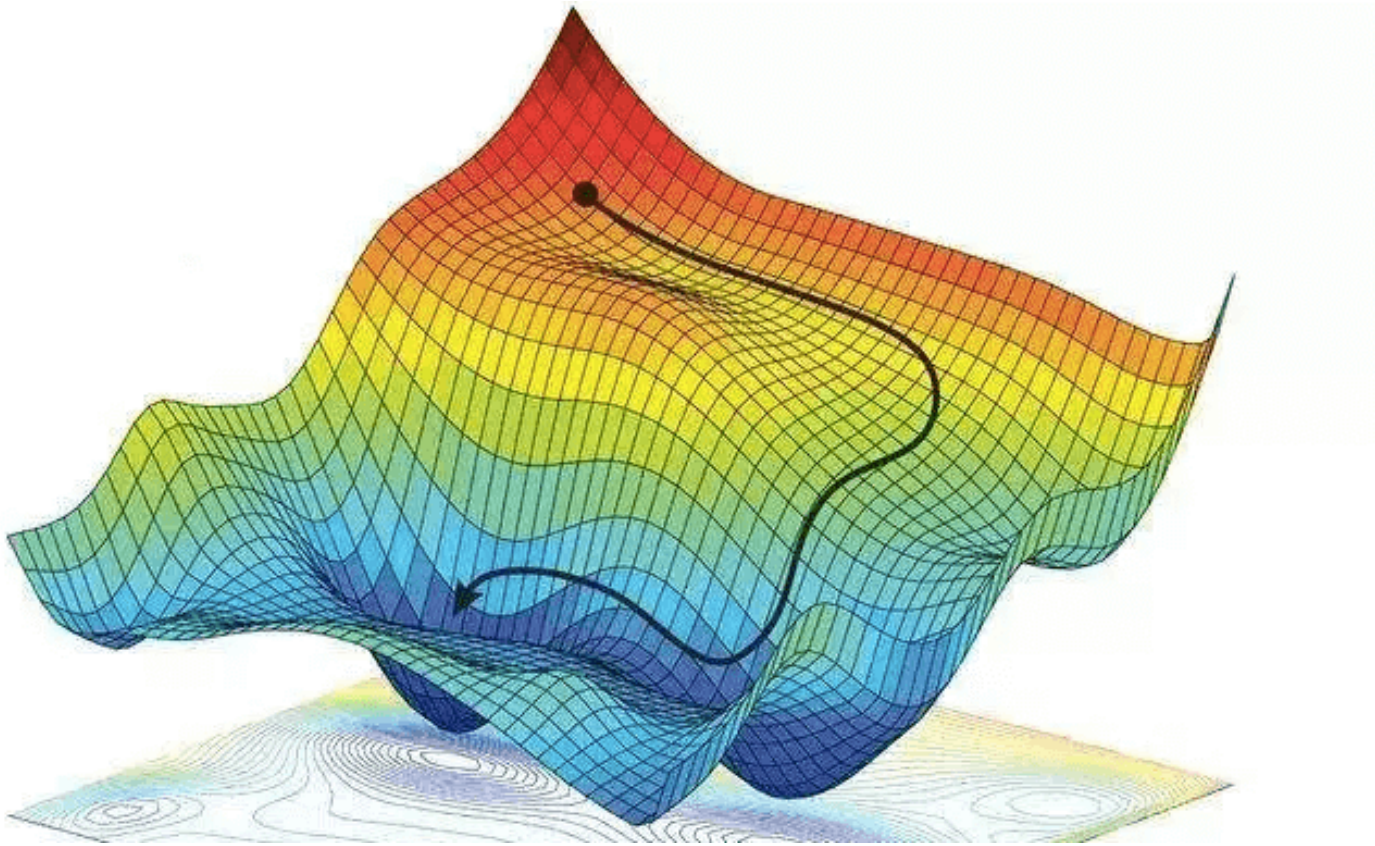

Figure: Gradient Descent in 3D

First-Order Condition for Convexity

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable over an open domain. Then the following statements are equivalent:

(i) *f is convex.*

(ii) *For all $x, y \in \mathrm{dom}(f)$,*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

Statement (ii) is known as the *first-order condition* for convexity. It means that the tangent plane at any point lies below the graph of f.

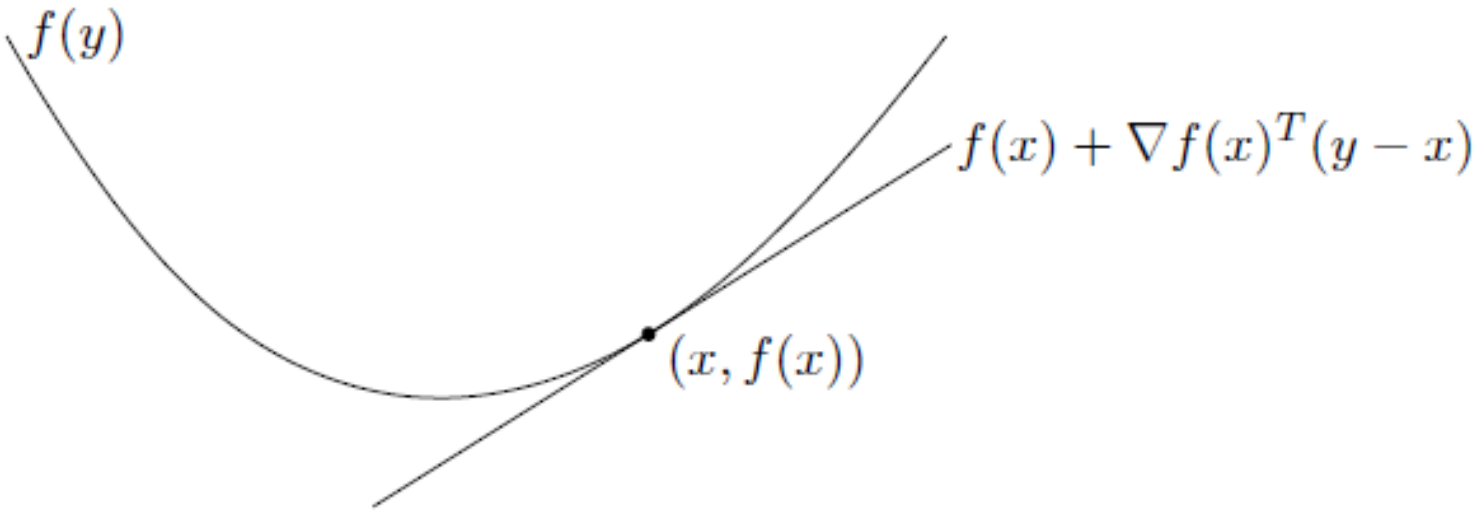

Figure: First Order Condition

Convergence of GD (Convex + Lipschitz)

Theorem

Assume that $\|x_1 - x^*\| \leq R$, and f is convex and L-Lipschitz continuous. Then, gradient descent with step size

$$\eta_t = \frac{R}{L\sqrt{t}}$$

satisfies the bound

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}.$$

Definition: Smoothness

Definition

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be β-**smooth** if its gradient is β-Lipschitz. That is,

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n.$$

Lemma (Smoothness Upper Bound)

Let f be a β-smooth function on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$,

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

Convergence of GD (Convex + Smooth)

Theorem

Let f be convex and β-smooth on $\mathbb{R}^n$. Then, gradient descent with constant step size

$$\eta = \frac{1}{\beta}$$

satisfies the bound

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_{s+1}\right) - f(x^*) \leq \frac{\beta\|x_1 - x^*\|^2}{2t}.$$

Proof of Smooth + Convex GD Convergence (1)

Proof.

Since f is β-smooth, we have:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2 \tag{1}$$

Convexity of f gives:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{2}$$

Apply (1) with $x = x_s, y = x_{s+1}$, and (2) with $x = x_s, y = x^*$:

$$f(x_{s+1}) \leq f(x_s) + \langle \nabla f(x_s), x_{s+1} - x_s \rangle + \frac{\beta}{2}\|x_{s+1} - x_s\|^2 \tag{3}$$

$$f(x_s) + \langle \nabla f(x_s), x^* - x_s \rangle \leq f(x^*) \tag{4}$$

Proof of Smooth + Convex GD Convergence (2)

continued.

Adding (3) and (4):

$$f(x_{s+1}) + \langle \nabla f(x_s), x^* - x_s \rangle \leq f(x^*) + \langle \nabla f(x_s), x_{s+1} - x_s \rangle + \frac{\beta}{2}\|x_{s+1} - x_s\|^2$$

Rearranging:

$$f(x_{s+1}) \leq f(x^*) + \langle \nabla f(x_s), x_{s+1} - x^* \rangle + \frac{\beta}{2}\|x_{s+1} - x_s\|^2$$

With GD update: $x_{s+1} = x_s - \eta \nabla f(x_s)$, plug in:

$$f(x_{s+1}) \leq f(x^*) - \frac{1}{\eta}\langle x_{s+1} - x^*, x_{s+1} - x_s \rangle + \frac{\beta}{2}\|x_{s+1} - x_s\|^2$$

Proof of Smooth + Convex GD Convergence (3)

continued.

We use the identity:

$$\langle p, q \rangle = \frac{1}{2} \left(\|p\|^2 + \|q\|^2 - \|p - q\|^2 \right)$$

Apply this with $a = x_{s+1}$, $b = x_s$, $c = x^*$, we get:

$$\langle x_{s+1} - x^*, x_{s+1} - x_s \rangle = \frac{1}{2} \left(\|x_s - x^*\|^2 - \|x_{s+1} - x_s\|^2 - \|x_{s+1} - x^*\|^2 \right)$$

Substituting into the previous inequality:

$$f(x_{s+1}) \leq f(x^*) + \frac{1}{2\eta} \left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2 \right)$$

$$+ \left(\frac{\beta}{2} - \frac{1}{2\eta} \right) \|x_{s+1} - x_s\|^2$$

Proof of Smooth + Convex GD Convergence (4)

continued.

Now set $\eta = \frac{1}{\beta}$. Then the second term vanishes:

$$f(x_{s+1}) - f(x^*) \leq \frac{\beta}{2}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right)$$

Summing both sides over $s = 1$ to t:

$$\sum_{s=1}^{t}(f(x_{s+1}) - f(x^*)) \leq \frac{\beta}{2}\|x_1 - x^*\|^2$$

Using Jensen's inequality:

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_{s+1}\right) - f(x^*) \leq \frac{\beta\|x_1 - x^*\|^2}{2t}$$

Definition: Strong Convexity

Definition

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be α**-strongly convex** if for all $x, y \in \mathbb{R}^n$,

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

Convergence of GD (Strongly Convex + Smooth)

Theorem

Let f be α-strongly convex and β-smooth on $\mathbb{R}^n$. Then gradient descent with step size

$$\eta = \frac{1}{\beta}$$

satisfies the following for all $t \geq 0$:

$$\|x_{t+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^t \|x_1 - x^*\|^2.$$

Stochastic Gradient Descent (SGD)

Minimizing Finite Sum

$$F(x) = \frac{1}{m}\sum_{k=1}^{m} f_k(x), \quad x \in \mathbb{R}^n.$$

When m is large, computing the full gradient becomes expensive.

Algorithm: Stochastic Gradient Descent

- **Step 1:** Choose an initial point $x_1 \in \mathbb{R}^n$.
- **Step 2:** For $t = 1, 2, \ldots, N$
  - Randomly select $a_t \in \{1, \ldots, m\}$ and update

$$x_{t+1} = x_t - \eta \nabla f_{a_t}(x_t)$$

Unbiased Stochastic Gradient

Definition (Unbiased Stochastic Gradient)

A stochastic gradient $\tilde{g}(x_t)$ is called **unbiased** if it satisfies:

$$\mathbb{E}[\tilde{g}(x_t) \mid x_t] = \nabla F(x_t).$$

This condition ensures that, in expectation, we are descending along the true gradient of the objective function F.

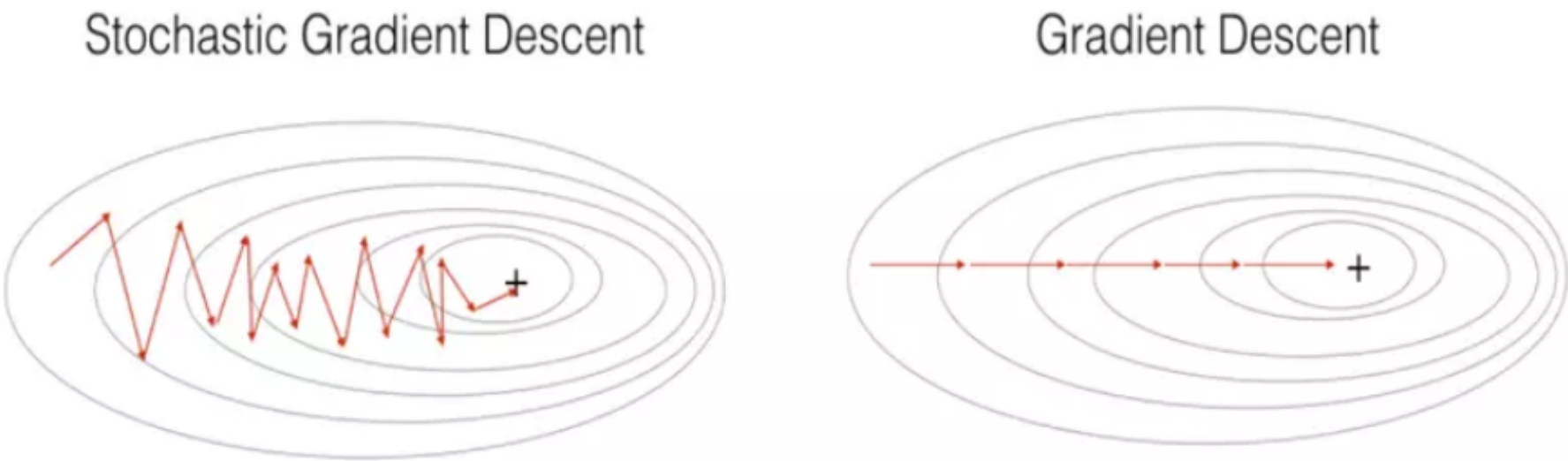

Figure: SGD vs GD

Convergence of SGD (Convex + Bounded Gradient)

Theorem

Let f be convex, and suppose $\mathbb{E}[\|\tilde{g}(x)\|^2] \leq B^2$. Let $R = \|x_1 - x^*\|$, and use step size

$$\eta_t = \frac{R}{B\sqrt{t}}.$$

Then stochastic gradient descent satisfies

$$\mathbb{E}\left[f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*)\right] \leq \frac{RB}{\sqrt{t}}.$$

Convergence of SGD (Convex + Smooth + Noisy Gradient)

Theorem

Suppose f is convex and β-smooth, and assume

$$\mathbb{E}[\|\nabla f(x) - \tilde{g}(x)\|^2] \leq \sigma^2.$$

Let $R = \|x_1 - x^*\|$, and use fixed step size

$$\eta = \frac{R}{\sigma}.$$

Then

$$\mathbb{E}\left[f\left(\frac{1}{t}\sum_{s=1}^{t} x_{s+1}\right) - f(x^*)\right] \leq \frac{\beta R^2}{2t} + \frac{\sigma R}{\sqrt{t}}.$$

Convergence of SGD (Strong Convexity)

Theorem

Suppose f is α-strongly convex and β-smooth. Assume

$$\mathbb{E}\left[\|\nabla f(x) - \tilde{g}(x)\|^2\right] \leq \sigma^2.$$

Then SGD with step size η_t satisfies the recurrence:

$$\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] \leq (1 - c\eta_t)\,\mathbb{E}\left[\|x_t - x^*\|^2\right] + \sigma^2\eta_t^2,$$

Proximal Point Method

Algorithm: Proximal Point Method

- **Step 1:** Choose an initial point $x_1 \in \mathbb{R}^n$.
- **Step 2:** For $k = 1, 2, \ldots, N$, compute:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\rho} \|x - x_k\|^2 \right\},$$

where $\rho > 0$ is a fixed parameter.

Proximal Point Method

Algorithm: Proximal Point Method

- **Step 1:** Choose an initial point $x_1 \in \mathbb{R}^n$.
- **Step 2:** For $k = 1, 2, \ldots, N$, compute:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\rho} \|x - x_k\|^2 \right\},$$

where $\rho > 0$ is a fixed parameter.

- Each step solves a regularized version of the original problem.
- This method is especially useful when f is convex but not necessarily smooth.

Convergence of PPM (convex)

Theorem

Let f be convex. Then the iterates x_k generated by the Proximal Point Method satisfy:

$$f\left(\frac{1}{T}\sum_{k=1}^{T} x_{k+1}\right) - f(x^*) \leq \frac{\|x_1 - x^*\|^2}{2\rho T} \quad \text{for any } T \in \mathbb{N}.$$

- The convergence is $\mathcal{O}(1/T)$ in terms of average iterate suboptimality.
- No smoothness assumption is required—convexity suffices.

Proof Sketch of Proximal Point Convergence

By convexity of f, we have:

$$f(x_{k+1}) - f(x) \leq \langle \nabla f(x_{k+1}), x_{k+1} - x \rangle.$$

From the optimality condition of the proximal step:

$$\nabla f(x_{k+1}) + \frac{1}{\rho}(x_{k+1} - x_k) = 0,$$

hence,

$$f(x_{k+1}) - f(x) \leq -\frac{1}{\rho}\langle x_{k+1} - x, x_{k+1} - x_k \rangle.$$

Using the identity

$$\langle p, q \rangle = \frac{1}{2}\left(\|p\|^2 + \|q\|^2 - \|p - q\|^2 \right),$$

the result follows by telescoping summation and averaging.

Convergence of PPM (Strong Convex)

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and α-strongly convex. Then the iterates $\{x_k\}$ generated by the proximal point method

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\}$$

satisfy the following linear convergence bound:

$$\|x_{t+1} - x^*\|^2 \leq \frac{1}{(1+\eta\alpha)^2} \|x_t - x^*\|^2.$$

Lagrangian and Lagrange Dual Function

Consider the following minimization problem:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \ldots, m \\ & \ell_j(x) = 0, \quad j = 1, \ldots, r \end{aligned}$$

The **Lagrangian** is defined as:

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

The **Lagrange dual function** is:

$$g(u, v) = \inf_x L(x, u, v)$$

Lagrange Dual Problem and Properties

The **dual problem** corresponding to the Lagrangian is:

$$\max_{u,v} \quad g(u,v)$$
$$\text{subject to} \quad u \geq 0$$

Important properties:

- The dual problem is always convex:

  $g(u,v)$ is concave (even if the primal is not convex).

- **Weak duality:** The optimal values f^* (primal) and g^* (dual) satisfy:

  $$f^* \geq g^*$$

- **Slater's condition (for strong duality):**
  If the primal problem is convex and there exists x such that

  $$h_i(x) < 0 \text{ for all } i, \quad \ell_j(x) = 0 \text{ for all } j,$$

  then strong duality holds.

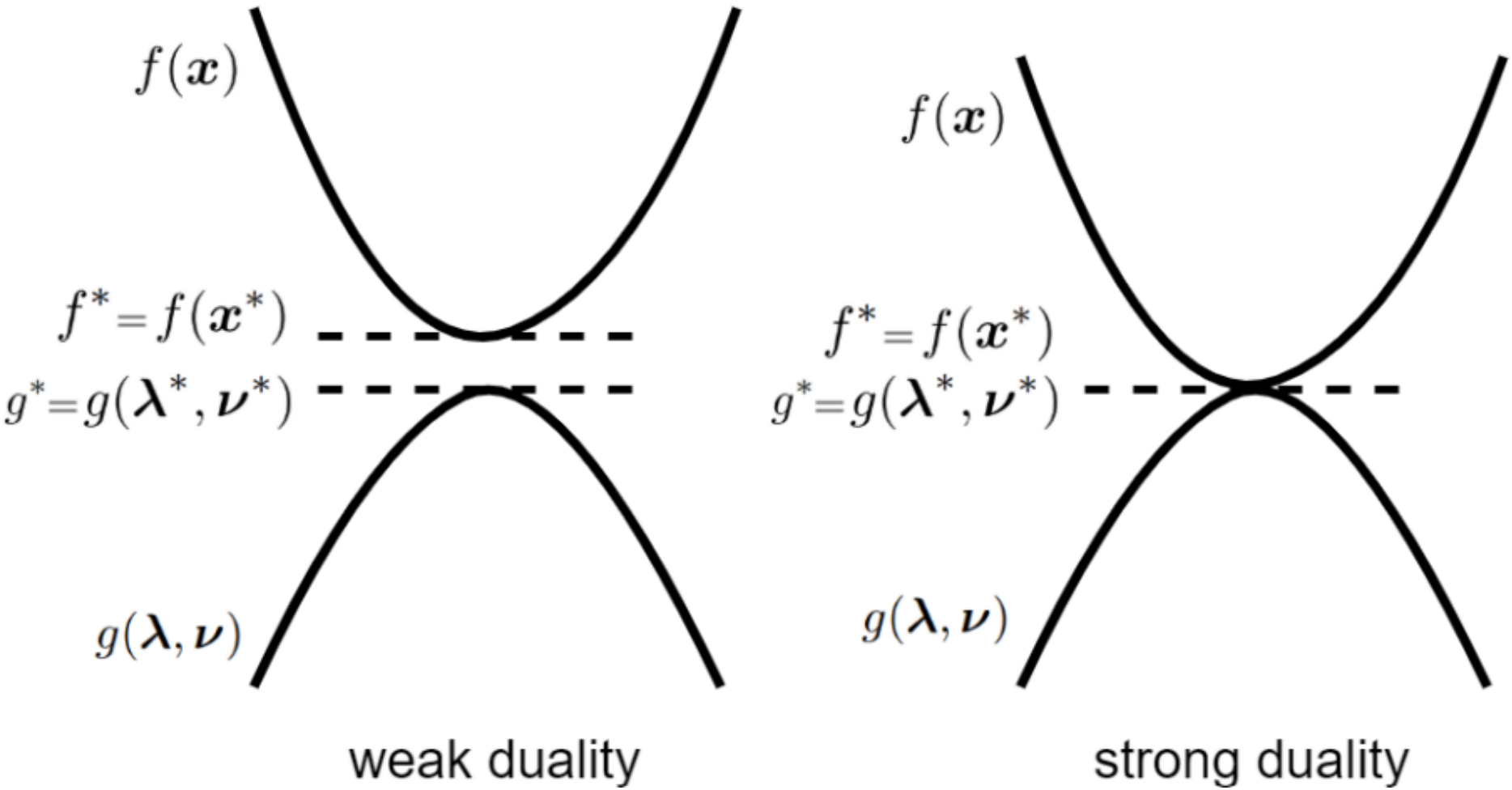

Figure: Weak duality vs Strong dulaity

Strong Duality

Exception: Trust Region Subproblem (TRS)

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2} x^T Q x + b^T x \quad \text{subject to } \|x\| \leq \Delta$$

- This could be **non-convex** problem when Q has negative eigenvalues,
- It is well-known that **strong duality still holds**, as shown by Moré and Sorensen (1983) and Sorensen (1982).

- **Strong duality** typically holds for convex optimization problems under mild regularity conditions (e.g., Slater's condition).
- In contrast, strong duality often **fails** for non-convex problems.

The End

Questions? Comments?