Optimization for Machine Learning

Ch2: First-Order Methods for Convex Optimization

Dongyoung Lim
UNIST

AI51101, IE55101

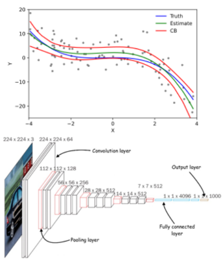
Introduction

- Training an AI model is fundamentally an optimization problem.
- The goal of training is to find parameters that minimize a loss (objective) function:

$$\min_{\theta \in \Theta} L(\theta),$$

where

- L : loss/objective function,
- θ : model parameters (decision variables),
- Θ : feasible set of parameters



Supervised Learning

$$\min_{\theta} \frac{1}{n} \mathcal{L}_\theta(x_i, y_i)$$

Unsupervised Learning

$$\min_{\theta} \frac{1}{n} \mathcal{L}_\theta(x_i)$$

Reinforcement Learning

$$\max_{\theta} \mathbb{E}_{\tau \sim p_\theta}[\mathcal{R}_{total}(\tau)]$$

Introduction

- Example: Supervised learning with training data $\{(x_i, y_i)\}_{i=1}^n$ drawn from $(X, Y) \sim \mathcal{P}$.
- We seek model parameters θ of a predictor $F(\cdot; \theta)$ that minimize the expected loss:

$$\min_{\theta \in \Theta} \mathbb{E}_{(X,Y)\sim\mathcal{P}}\big[L(Y, F(X; \theta))\big].$$

- Because the data distribution $\mathcal{P}$ is unknown, we instead minimize the empirical risk:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, F(x_i; \theta)\big).$$

Introduction

- Generally, an optimization problem can be expressed as

$$\min_{x \in \mathcal{X}} f(x),$$

where

  - $f(x)$: objective (loss) function,
  - x : decision variable or model parameter,
  - $\mathcal{X}$: feasible set (constraints).

- Among various optimization problems, this chapter will focus on a particularly important class called **convex optimization**, for which efficient first-order algorithms and strong theoretical guarantees exist.

Introduction

- To solve an optimization problem, we need to choose an appropriate **optimization method**.
- Optimization algorithms are classified by the type of information they require about f:
  - **Zeroth-order**: use only function values (e.g., bisection).
  - **First-order**: use function values and gradients (e.g., gradient descent).
  - **Second-order**: also use Hessian information (e.g., Newton's method).
- This chapter focuses on **first-order methods**, a fundamental class of algorithms that play a central role in modern AI model training.

Introduction

- **Scope of this chapter:** We study first-order methods for minimizing a **convex** function f with gradient access.
- Different structural assumptions on f lead to different algorithms and rates:
  - **Non-smooth convex:** f is L-Lipschitz (objective Lipschitz).
  - **Smooth convex:** f has β-Lipschitz gradient (β-smooth).
  - **Strongly convex:** f satisfies μ-strong convexity.

Introduction

- **Scope of this chapter:** We treat both **unconstrained** and **constrained** problems:
  - **Unconstrained**: the decision variable x can take any value in $\mathbb{R}^d$; there is no explicit restriction other than the domain of f.

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

  - **Constrained**: x must lie in a specified feasible set $\mathcal{X}$ (assumed convex in this chapter), e.g.,

$$\min_{x \in \mathcal{X}} f(x).$$

- Throughout this chapter, when constraints are present we assume the set $\mathcal{X}$ is convex.

Introduction

- In other words, we will study first-order methods for convex optimization.
- For example, consider the unconstrained problem

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

  where f is a convex function.
- Iterative update (gradient descent):

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

  where $\gamma > 0$ is the step size (learning rate).
- We analyze how these first-order methods converge to the optimal solution under different assumptions on f (non-smooth convex, smooth convex, and strongly convex), and extend the analysis to convex constrained problems.

Preliminaries

Definition (Lipschitz Continuity)

A real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is *L-Lipschitz continuous* if there exists a constant $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$,

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

- Intuition: the function cannot change faster than slope L; it is "slope-bounded."
- Differentiability is not required. If f is differentiable, then

$$\|\nabla f(x)\| \leq L \quad \forall x \in \mathbb{R}^d.$$

- **Exercise:** Prove the gradient bound above.

Preliminaries

Definition (Convex Combination)

Let $\mathcal{C} = \{x_1, x_2, \ldots, x_n\}$ be a subset of a vector space. A convex combination z of $\mathcal{C}$ is a linear combination of vectors in $\mathcal{C}$ where all coefficients are non-negative and sum to one:

$$z = \sum_{i=1}^{n} \alpha_i x_i, \quad \text{where } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i = 1.$$

Definition (Convex Set)

A set $\mathcal{X}$ is convex if the convex combination of any two points in $\mathcal{X}$ is also in $\mathcal{X}$. That is, for all $x, y \in \mathcal{X}$ and $0 \leq \theta \leq 1$,

$$\theta x + (1 - \theta) y \in \mathcal{X}.$$

Examples of Convex Sets

- **Euclidean Ball**

$$\mathcal{B}(x_0, r) := \{x \in \mathbb{R}^d : \|x - x_0\|_2 \leq r\}.$$

Any line segment between two points in the ball remains inside the ball.

- **Affine Subspace**

$$\mathcal{A} := \{x \in \mathbb{R}^d : Ax = b\}.$$

An affine set is convex because the linear constraint is preserved under convex combinations.

- **Probability Simplex**

$$\Delta_d := \left\{x \in \mathbb{R}^d : x_i \geq 0,\ \sum_{i=1}^{d} x_i = 1\right\}.$$

Convex combinations of probability distributions remain valid distributions.

Preliminaries

Definition (Convex Function)

Let $\mathcal{X}$ be a convex set. A function $f : \mathcal{X} \to \mathbb{R}$ is convex if for all $x, y \in \mathcal{X}$ and $0 \leq \theta \leq 1$,

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

- Intuition: the line segment between $(x, f(x))$ and $(y, f(y))$ lies above the graph of f.
- The function $-f$ is concave.

Preliminaries

- Convexity can also be characterized by first- or second-order conditions.

Lemma (First-Order Condition)

Let $\mathcal{X}$ be convex. A differentiable function f is convex iff for all $x, y \in \mathcal{X}$,

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

- Geometric view: the tangent plane at any point lies below the graph of f.

Property (Monotone Derivative)

Let $f : \mathcal{X} \to \mathbb{R}$ be convex differentiable, and $x, y \in \mathcal{X}$. Define the one–dimensional slice

$$\phi(t) := f\big(x + t(y - x)\big), \qquad t \in [0, 1].$$

Then, ϕ is differentiable and its derivative

$$\phi'(t) = \big\langle \nabla f(x + t(y - x)),\, y - x \big\rangle,$$

is nondecreasing on $[0, 1]$.

Lemma (Second-Order Condition)

Let $\mathcal{X}$ be convex. A twice-differentiable function f is convex iff for all $x \in \mathcal{X}$,

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

- The Hessian being positive semi-definite means curvature is nonnegative in every direction:

$$\text{for any direction } v \in \mathbb{R}^d, \quad g(t) := f(x + tv) \ \Rightarrow \ g''(0) \ = \ v^\top \nabla^2 f(x)\, v \ \geq 0.$$

Examples of Convex Functions

- **Quadratic Function**

$$f(x) = \frac{1}{2}x^\top Qx + b^\top x + c,$$

  where $Q \succeq 0$ (positive semi-definite).

  - Hessian: $\nabla^2 f(x) = Q$.
  - Since $Q \succeq 0$, we have $\nabla^2 f(x) \succeq 0$ for all $x \Rightarrow$ convex.

- **Norms** ($p \geq 1$)

$$f(x) = \|x\|_p.$$

  - Triangle inequality, $\|\theta x + (1-\theta)y\|_p \leq \theta\|x\|_p + (1-\theta)\|y\|_p$, satisfies the definition of a convex function.

Lemma (Monotone Gradient Condition)

Let $\mathcal{X}$ be convex. A differentiable function f is convex iff for all $x, y \in \mathcal{X}$,

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq 0.$$

- Intuition: the gradient changes monotonically along any line segment in $\mathcal{X}$.

Definition (Subgradient)

For a convex function $f : \mathcal{X} \to \mathbb{R}$ (not necessarily differentiable), a vector $g \in \mathbb{R}^d$ is called a *subgradient* of f at $x \in \mathcal{X}$ if

$$f(y) \geq f(x) + \langle g, y - x \rangle \qquad \forall y \in \mathcal{X}.$$

The set of all subgradients of f at x is denoted $\partial f(x)$.

- **Intuition:** The subgradient generalizes the role of the gradient to non-smooth convex functions, acting as a "generalized slope" that supports the graph of f from below.
- If f is differentiable at x, the subgradient set collapses to the usual gradient: $\partial f(x) = \{\nabla f(x)\}$.
- For $g \in \partial f(x)$, the affine function $f(x) + \langle g, y - x \rangle$ lies below $f(y)$ for all y.
- Every convex function admits at least one subgradient at every interior point of $\mathcal{X}$.

Examples of Subgradients

- **Absolute Value**: $f(x) = |x|$.
  - $x > 0$: $\partial f(x) = \{1\}$.
  - $x < 0$: $\partial f(x) = \{-1\}$.
  - $x = 0$: $\partial f(0) = [-1, 1]$ (every slope g with $-1 \leq g \leq 1$ satisfies $f(y) \geq f(0) + g(y - 0)$).
- **Euclidean Norm ($\mathbb{R}^d$)**: $f(x) = \|x\|_2$.
  - $x \neq 0$: $\partial f(x) = \left\{ \dfrac{x}{\|x\|_2} \right\}$.
  - $x = 0$: $\partial f(0) = \{g \in \mathbb{R}^d : \|g\|_2 \leq 1\}$ (the closed unit ball). Why? g should satisfy

$$\|y\|_2 \geq \langle g, y \rangle.$$

By Cauchy-Schwarz, it becomes

$$\langle g, y \rangle \leq \|g\|_2 \|y\|_2 \leq \|y\|_2.$$

Thus, we need $\|g\|_2 \leq 1$.

Preliminaries: Projections onto Convex Sets

- For a closed convex set $\mathcal{C} \subset \mathbb{R}^d$, the **Euclidean projection** of a point $z \in \mathbb{R}^d$ onto $\mathcal{C}$ is

$$\Pi_{\mathcal{C}}(z) := \arg\min_{x \in \mathcal{C}} \|x - z\|.$$

- Intuition: $\Pi_{\mathcal{C}}(z)$ is the point in $\mathcal{C}$ closest to z.

Property I

For any $x \in \mathcal{C}$ and $z \in \mathbb{R}^d$,

$$\langle x - \Pi_{\mathcal{C}}(z),\ z - \Pi_{\mathcal{C}}(z) \rangle \leq 0.$$

Property II

For any $x \in \mathcal{C}$ and $z \in \mathbb{R}^d$,

$$\|\Pi_{\mathcal{C}}(z) - x\| \leq \|z - x\|.$$

- These properties are fundamental in analyzing **projected gradient methods**.

- **Proof of Property I:** Let $p := \Pi_{\mathcal{C}}(z)$ be the unique minimizer of $\min_{u\in\mathcal{C}} \|u-z\|^2$. For any $x \in \mathcal{C}$ and $\theta \in [0,1]$, the convex combination $p + \theta(x-p) \in \mathcal{C}$. By minimality of p,

$$\|z-p\|^2 \le \|z-(p+\theta(x-p))\|^2 = \|z-p\|^2 - 2\theta\langle x-p, z-p\rangle + \theta^2\|x-p\|^2.$$

Define $h(\theta) := \|z-(p+\theta(x-p))\|^2 - \|z-p\|^2 = -2\theta\langle x-p, z-p\rangle + \theta^2\|x-p\|^2$. Then $h(\theta) \ge 0$ for all small $\theta > 0$ and $h(0) = 0$, so the right derivative at 0 satisfies

$$h'_+(0) = -2\langle x-p, z-p\rangle \ \ge 0 \ \Rightarrow \ \langle x-p, z-p\rangle \le 0.$$

- **Proof of Property II:** Let $p := \Pi_{\mathcal{C}}(z)$. Then,

$$\|z - x\|^2 = \|z - p + p - x\|^2 = \|z - p\|^2 + \|p - x\|^2 + 2\langle z - p,\ p - x\rangle.$$

From Property I

$$\langle z - p,\ p - x\rangle \geq 0$$

and $\|z - p\|^2 \geq 0$, we have

$$\|z - x\|^2 \geq \|p - x\|^2.$$

Preliminaries

- **Why Convexity?** For convex functions, *every local minimum is also a global minimum*.

Global Optimality of Local Minima (Unconstrained Case)

Let f be a convex function. If x is a local minimum of f, then x is a global minimum of f. Moreover, this holds if and only if

$$0 \in \partial f(x),$$

where $\partial f(x)$ denotes the subdifferential of f at x.

First-Order Optimality (Constrained Case)

Let $\mathcal{X}$ be a convex set and let f be a differentiable convex function on $\mathcal{X}$. Then $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$ if and only if

$$\langle \nabla f(x^*), x^* - y \rangle \leq 0 \qquad \forall y \in \mathcal{X}.$$

- Intuition: the gradient at x^* points outward or is orthogonal to all feasible directions, so no descent direction exists. That is,

$$\langle \nabla f(x^*), x^* - y \rangle = \left\langle \underbrace{-\nabla f(x^*)}_{\text{descent direction}}, \underbrace{y - x^*}_{\text{feasible direction}} \right\rangle \leq 0.$$

Unconstrained Case: Overview

- We consider the unconstrained problem

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

  where f is a convex objective function.

- We will analyze (sub)gradient descent under the following settings:

  1. f convex and L**-Lipschitz**
  2. f convex and β**-smooth**
  3. f **strongly convex** and L**-Lipschitz**
  4. f **strongly convex** and β**-smooth**

- For each case we will derive the convergence rate of first-order methods.

Unconstrained Case: Convex & L-Lipschitz

Theorem (Subgradient Method Convergence)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that is L-Lipschitz. Consider the subgradient method

$$x_{k+1} = x_k - \eta g_k, \qquad g_k \in \partial f(x_k).$$

Let $R := \|x_0 - x^*\|$ where $x^* \in \arg\min f$. With constant step size

$$\eta = \frac{R}{L\sqrt{T}},$$

the averaged iterate

$$\bar{x}_T := \frac{1}{T} \sum_{k=0}^{T-1} x_k$$

satisfies

$$f(\bar{x}_T) - f(x^*) \leq \frac{LR}{\sqrt{T}}.$$

- Convergence rate: $\mathcal{O}(1/\sqrt{T})$.

UNCONSTRAINED CASE: CONVEX & β-SMOOTH

Definition (β-Smoothness)

A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-*smooth* if its gradient is β-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| \quad \forall x, y.$$

- If f is twice differentiable, $\nabla^2 f(x) \preceq \beta I$ for all x.

Unconstrained Case: Convex & β-Smooth

Lemma

If f is β-smooth,

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{\beta}{2} \|x - y\|^2.$$

Descent Lemma

If f is convex and β-smooth,

$$0 \leq f(x) - f(y) - \nabla f(y)^\top (x - y) \leq \frac{\beta}{2} \|x - y\|^2.$$

This implies

$$f\left(x - \tfrac{1}{\beta} \nabla f(x)\right) \leq f(x) - \frac{1}{2\beta} \|\nabla f(x)\|^2.$$

Lemma

For any x, y,

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2.$$

Unconstrained Case: Convex & β-Smooth

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and β-smooth. Gradient descent

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

with step size $\eta = \frac{1}{\beta}$ satisfies

$$f(x_T) - f(x^*) \leq \frac{2\beta}{T} \|x_0 - x^*\|^2,$$

where $x^* \in \arg\min f$.

- Convergence rate: $\mathcal{O}\big(1/T\big)$

Unconstrained Case: Strongly Convex & L-Lipschitz

Definition (α-Strong Convexity)

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *α-strongly convex* if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^d.$$

- Equivalent conditions:
  - $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.
  - (Strong monotonicity) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|^2$.
  - If f is twice differentiable, $\nabla^2 f(x) \succeq \alpha I$.
- If g is convex and h is α-strongly convex, then $g + h$ is also α-strongly convex.

Unconstrained Case: Strongly Convex & L-Lipschitz

Theorem (Subgradient Method)

Consider the subgradient method

$$x_{k+1} = x_k - \eta_k g_k, \qquad g_k \in \partial f(x_k),$$

where f is α-strongly convex and L-Lipschitz. With step size

$$\eta_k = \frac{2}{\alpha(k+1)},$$

the weighted average

$$\bar{x}_T := \sum_{k=0}^{T-1} \frac{2k}{T(T+1)} x_k$$

satisfies

$$f(\bar{x}_T) - f(x^*) \leq \frac{2L^2}{\alpha(T+1)}.$$

- Convergence rate: $\mathcal{O}(1/T)$ without smoothness.

UNCONSTRAINED CASE: STRONGLY CONVEX & β-SMOOTH

Lemma

If f is β-smooth and α-strongly convex on $\mathbb{R}^d$, then for all $x, y \in \mathbb{R}^d$,

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

- This inequality expresses the *strong monotonicity and co-coercivity* of the gradient mapping when f is both strongly convex and smooth.
- It quantitatively couples the point difference $\|x - y\|$ and the gradient difference $\|\nabla f(x) - \nabla f(y)\|$, ensuring that the gradient grows and aligns with $x - y$ in a controlled way.

Unconstrained Case: Strongly Convex & β-Smooth

Theorem (Gradient Descent Convergence)

Let f be α-strongly convex and β-smooth, and define the condition number $\kappa := \beta/\alpha$. Consider gradient descent

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

with step size

$$\gamma = \frac{2}{\alpha + \beta}.$$

Then the last iterate satisfies

$$f(x_T) - f(x^*) \leq \frac{\beta}{2} \exp\left(-\frac{4T}{\kappa + 1}\right) \|x_0 - x^*\|^2.$$

- Linear (geometric) convergence rate $\mathcal{O}(e^{-4T/(\kappa+1)})$.

Comparison for different function classes

Objective Function	Convergence Rate
convex and L-Lipschitz	$\mathcal{O}(\frac{1}{\sqrt{T}})$
convex and β-smooth	$\mathcal{O}(\frac{1}{T})$
α-strongly convex and L-Lipschitz	$\mathcal{O}(\frac{1}{T})$
α-strongly convex and β-smooth	$\mathcal{O}(\exp\left(-\frac{4T}{\kappa+1}\right))$

Table. Convergence rate of gradient descent for different properties of the objective function.

Constrained Case

- Consider the constrained convex optimization problem

$$\min_{x \in \mathcal{X}} f(x),$$

  where f is convex (possibly non-differentiable) and $\mathcal{X}$ is a closed convex set.

- The **projected subgradient method** is

$$y_{k+1} = x_k - \gamma g_k, \quad g_k \in \partial f(x_k), \qquad x_{k+1} = \Pi_{\mathcal{X}}(y_{k+1}).$$

- **Convergence rate:** With the same step–size choices as in the unconstrained case, this method achieves the *same rates* for all settings:
  - f convex and L-Lipschitz: $\mathcal{O}(1/\sqrt{T})$ (averaged iterate).
  - f convex and β-smooth: $\mathcal{O}(1/T)$ (last iterate).
  - f α-strongly convex and L-Lipschitz: $\mathcal{O}(1/T)$ (weighted average).
  - f α-strongly convex and β-smooth: linear (geometric) rate $\mathcal{O}\left(e^{-\frac{4T}{\kappa+1}}\right)$.

Constrained Case: Projections

- Each iteration requires computing the **projection** of a point onto the feasible set $\mathcal{X}$:

$$\Pi_{\mathcal{X}}(z) := \arg\min_{x \in \mathcal{X}} \|x - z\|.$$

- This is itself a convex optimization problem, but often has a closed-form solution for common sets.

Projection Theorem

Let H be a Hilbert space, $x \in H$, and $\mathcal{X}$ a closed subspace. There exists a unique projection $p \in \mathcal{X}$ satisfying

$$x - p \perp \mathcal{X}.$$

Constrained Case: Projection

Setting. Projection onto a line:

$$\mathcal{X} = \{x_0 + t\,u : t \in \mathbb{R}\}, \quad \|u\| = 1, \quad \text{given } x \in \mathbb{R}^d.$$

Solution:

Constrained Case: Projection

Setting. Projection onto a hyperplane:

$$\mathcal{X} = \{z \in \mathbb{R}^d : a^\top z = b\}, \quad a \neq 0, \quad \text{given } x.$$

Solution:

Constrained Case: Projection

Setting. Projection onto a subspace:

$$\mathcal{X} = \mathrm{span}(v_1, \ldots, v_k).$$

where v_i are linearly independent.

Solution:

Lower Bounds on First-Order Methods

- So far we have derived **upper bounds** on convergence rates. Now we ask: what is the **best possible** rate any first-order method can achieve?
- A lower bound shows the fundamental limit: no first-order algorithm can converge faster (in worst case) than this rate.

Theorem (Lower Bound for Gradient Descent)

Let f be α-strongly convex and β-smooth with condition number $\kappa = \beta/\alpha > 1$. For any step size choice and any starting point x_0, gradient descent satisfies

$$f(x_T) - f(x^*) \;\geq\; \frac{\alpha}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T} \|x_0 - x^*\|^2.$$

- For large κ, note that

$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T} \approx \exp\left(-\frac{4T}{\sqrt{\kappa}}\right).$$

Implementation

- Consider

$$\min_x x^2 + 2x + 3$$

where $x^* = -1$.

- Perform gradient descent with $x_1 = 1$ and $\gamma = 0.2$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$ and $x_2 = x_1 - \gamma \nabla f(x_1) = 0.2$.
  - Step 3: $\nabla f(x_2) = 2.4$ and $x_3 = x_2 - \gamma \nabla f(x_2) = -0.28$.
  - Step 4: $\nabla f(x_3) = 1.44$ and $x_4 = x_3 - \gamma \nabla f(x_3) = -0.568$.
  - Step 5: $\nabla f(x_4) = 0.864$ and $x_5 = x_4 - \gamma \nabla f(x_4) = -0.7408$.
  - Step 6: $\nabla f(x_5) = 0.5184$ and $x_6 = x_5 - \gamma \nabla f(x_5) = -0.84448$.

Implementation

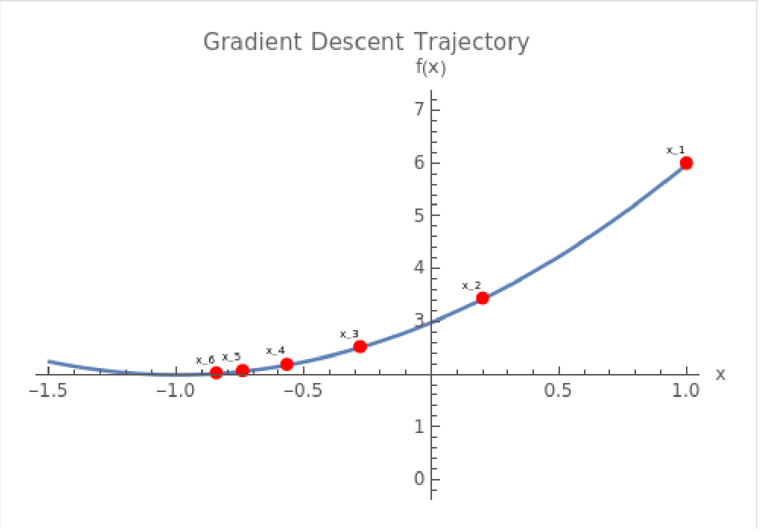


Figure. $\gamma = 0.1$

Implementation

- What if the step size is too large?
- Perform gradient descent with $x_1 = 1$ and $\gamma = 2$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$ and $x_2 = x_1 - \gamma \nabla f(x_1) = -7$.
  - Step 3: $\nabla f(x_2) = -12$ and $x_3 = x_2 - \gamma \nabla f(x_2) = 17$.
  - Step 4: $\nabla f(x_3) = 36$ and $x_4 = x_3 - \gamma \nabla f(x_3) = -55$.
  - Step 5: $\nabla f(x_4) = -108$ and $x_5 = x_4 - \gamma \nabla f(x_4) = 161$.
  - Step 6: $\nabla f(x_5) = 324$ and $x_6 = x_5 - \gamma \nabla f(x_5) = -487$.

Implementation

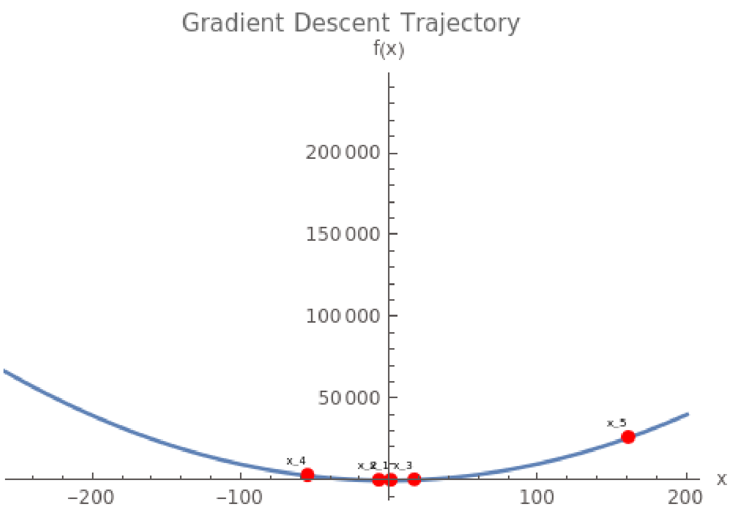


Figure. $\gamma = 2$

- The value of x_n changes very drastically due to a large learning rate.

Implementation

- One can calculate $\alpha = \beta = 2$.
- Perform gradient descent with $x_1 = 1$ and $\gamma = \frac{2}{\alpha+\beta} = 0.5$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$ and $x_2 = x_1 - \gamma \nabla f(x_1) = -1$.

Implementation

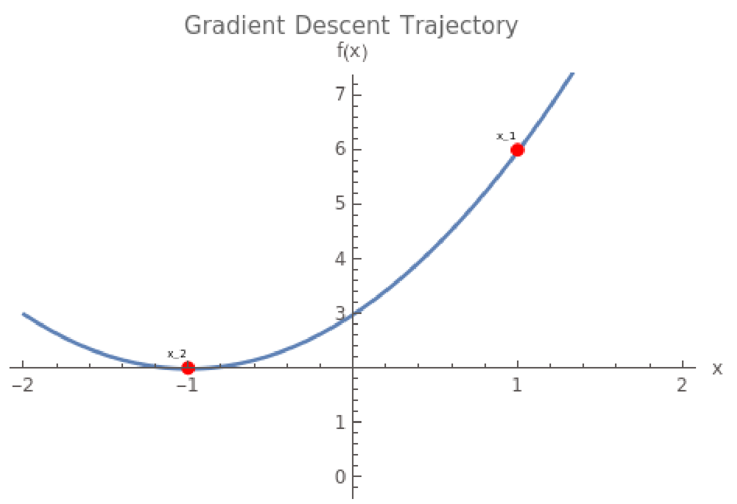


Figure. $\gamma = 0.5$

Implementation

- Consider the following constrained optimization

$$\min_{x \in [0,1]} x^2 + 2x + 3$$

  where $x^* = 0$.

- Perform projected gradient descent with $x_1 = 0.5$ and $\gamma = 0.5$:
  - Step 1: $x_1 = 1$
  - Step 2: $\nabla f(x_1) = 4$. $y_2 = x_1 - \gamma \nabla f(x_1) = -1$ and $x_2 = \Pi_{[0,1]}(y_2) = 0$.

IMPLEMENTATION

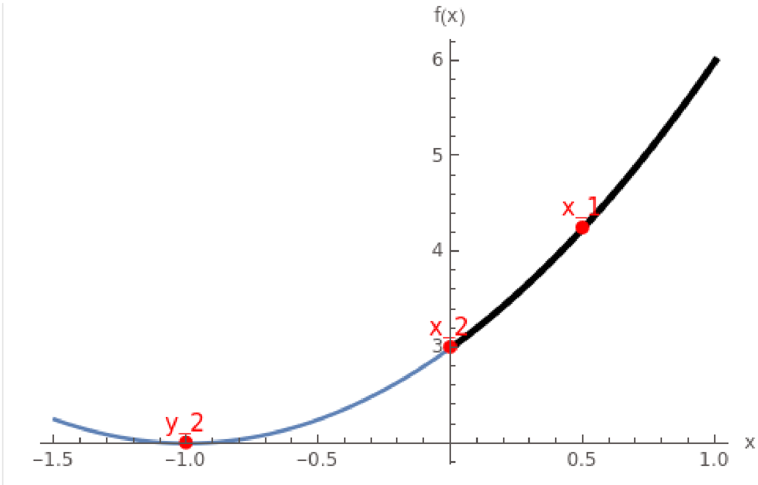


Figure. $\gamma = 0.5$

Example: Support Vector Machine

- Goal: classify each email as "Spam" or "Not Spam." Each email is represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ with label $y_i \in \{-1, +1\}$.
- A natural formulation seeks the separating hyperplane with the fewest misclassifications:

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \mathbf{1}_{\mathrm{sign}(\theta^\top \mathbf{x}_i) \neq y_i}.$$

- This 0–1 loss leads to a non-convex, NP-hard problem.

Example: Soft-Margin SVM (Convex Formulation)

- Replace the 0–1 loss with the convex **hinge loss**:

$$\ell(\theta; \mathbf{x}_i, y_i) = \max\{0,\ 1 - y_i \mathbf{x}_i^\top \theta\}.$$

- Add ℓ_2 regularization to control margin size:

$$\min_{\theta \in \mathbb{R}^d} f(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; \mathbf{x}_i, y_i) + \frac{\lambda}{2} \|\theta\|^2.$$

- This is an **unconstrained, non-smooth, strongly convex** problem! Therefore, subgradient descent achieves the rate $\mathcal{O}(1/T)$.

Example: Soft-Margin SVM (Convex Formulation)

Let

$$L(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i, y_i), \qquad \ell(\theta; x_i, y_i) = \max\{0,\, 1 - y_i x_i^\top \theta\},$$

and define

$$f(\theta) = L(\theta) + \frac{\lambda}{2} \|\theta\|^2, \qquad \lambda > 0.$$

Proof: L is convex since it is an average of convex hinge losses (each is a pointwise max of affine maps). In addition, $q(\theta) := \frac{\lambda}{2}\|\theta\|^2$ is λ-strongly convex because $\nabla^2 q(\theta) \succeq \lambda$. Since sum of a convex function and an λ-strongly convex function is λ-strongly convex, f is λ-strongly convex.