Optimization

Introduction to Optimization for Machine Learning
M1 MLSD/AMSD

October 28, 2025

Roadmap

(1) Optimization Using Gradient Descent

(2) Constrained Optimization and Lagrange Multipliers

(3) Convex Sets and Functions

Summary

- Training machine learning models $=$ finding a good set of parameters
- A good set of parameters $=$ Solution (or close to solution) to some optimization problem
- Directions: Unconstrained optimization, Constrained optimization, Convex optimization
- A necessary condition for the optimal point: $f'(x) = 0$ (stationary point)
    - Gradient will play an important role

Unconstrained Optimization and Gradient Algorithms

- Goal

$$\min f(\boldsymbol{x}), \quad f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}, \quad f \in C^1$$

- Graident-type algorithms

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \gamma_k \boldsymbol{d}_k, \quad k = 0, 1, 2, \ldots$$

- Lemma. Any direction $\boldsymbol{d} \in \mathbb{R}^{n \times 1}$ that satisfies $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{d} < 0$ is a descent direction of f at $\boldsymbol{x}$. That is, if we let $\boldsymbol{x}_\alpha = \boldsymbol{x} + \alpha \boldsymbol{d}$, $\exists \bar{\alpha} > 0$, such that for all $\alpha \in (0, \bar{\alpha}]$, $f(\boldsymbol{x}_\alpha) < f(\boldsymbol{x})$.
- Finding a local optimum $f(\boldsymbol{x}_\star)$, if the step-size γ_k is suitably chosen.

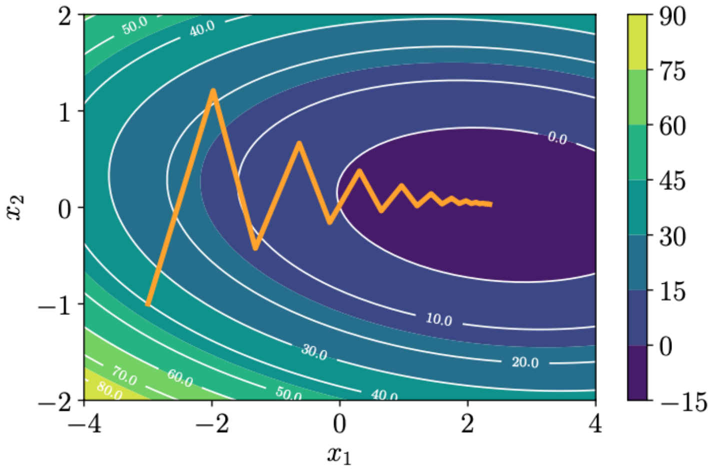
Example

- A quadratic function $f : \mathbb{R}^2 \mapsto \mathbb{R}$.

$$f\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \frac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} 2 & 1 \\ 1 & 20 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 5 \\ 3 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

whose gradient is $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} 2 & 1 \\ 1 & 20 \end{pmatrix} - \begin{pmatrix} 5 \\ 3 \end{pmatrix}^{\mathsf{T}}$

- $\boldsymbol{x}_0 = (-3 - 1)^{\mathsf{T}}$
- constant step size $\alpha = 0.085$
- Zigzag pattern



Taxonomy

- Goal: $\min L(\boldsymbol{\theta})$ for n training data
- Based on the amount of training data used for each iteration
  - Batch gradient descent (the entire n)
  - Mini-batch gradient descent($k < n$ data)
  - Stochastic gradient descent (one sampled data)
- Based on the adaptive method of update
  - Momentum, NAG, Adagrad, RMSprop, Adam, etc
- `https://ruder.io/optimizing-gradient-descent/`

Stochastic Gradient Descent (SGD)

- Assume $L(\boldsymbol{\theta}) = \sum_{i=1}^{n} L_n(\boldsymbol{\theta})$ (which happens in many cases in machine learning, e.g., negative log-likelihood in regression)
- Gradient update

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \gamma_k \nabla L(\boldsymbol{\theta}_k)^{\mathsf{T}} = \boldsymbol{\theta}_k - \gamma_k \sum_{n=1}^{N} \nabla L_n(\boldsymbol{\theta}_k)^{\mathsf{T}}$$

  - Batch gradient: $\sum_{n=1}^{N} \nabla L_n(\boldsymbol{\theta}_k)^{\mathsf{T}}$
  - Mini-batch gradient: $\sum_{n \in \mathcal{K}} \nabla L_n(\boldsymbol{\theta}_k)^{\mathsf{T}}$ for a suitable choice of $\mathcal{K}, |\mathcal{K}| < n$
  - Stochastic gradient: $\nabla L_n(\boldsymbol{\theta}_i)^{\mathsf{T}}$ for some (randomly chosen) i. Noisy approximation to the real gradient.
- Tradeoff: computation burden vs. exactness

Adaptivity for Better Convergence: Momemtum

- Step size.
    - Too small: slow update, Too big: overshoot, zig-zag, often fail to converge
- Adaptive update: smooth out the erratic behavior and dampens oscillations
- Gradient descent with momentum

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_i \nabla f(\boldsymbol{x}_k)^{\mathsf{T}} + \alpha \Delta \boldsymbol{x}_k, \quad \alpha \in [0,1]$$

$$\Delta \boldsymbol{x}_k = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$$

- Memory term: $\alpha \Delta \boldsymbol{x}_k$, where α is the degree of how much we remember the past
- Next update = a linear combination of current and previous updates

Standard Constrained Optimization Problem

- An optimization problem in standard form:

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}) \\ &\text{subject to } g_i(\boldsymbol{x}) \leq 0, \quad i = 1, 2, \ldots, m \quad \textit{(Inequality constraints)} \\ &\qquad\qquad\quad h_j(\boldsymbol{x}) = 0, \quad j = 1, 2, \ldots, p \quad \textit{(Equality constraints)} \end{aligned}$$

- Variables: $\boldsymbol{x} \in \mathbb{R}^n$. Assume nonempty feasible set

- Optimal value: p^*. Optimizer: $\boldsymbol{x}^*$

Problem Solving via Langrange Multipliers

- Duality Mentality
  - Bound or solve an optimization problem via a different optimization problem!
  - We'll develop the basic Lagrange duality theory for a general optimization problem, then specialize for convex optimization
- Idea: augment the objective with a weighted sum of constraints
  - Lagrangian:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^{p} \nu_i h_i(\boldsymbol{x})$$

  - Lagrange multipliers (dual variables): $\boldsymbol{\lambda} = (\lambda_i : i = 1, \cdots, m) \succeq 0$, $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_p)$
  - Lagrange dual function:

$$\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

Lower Bound on the Optimal Value

- The dual function $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is the lower bound on the optimal value p^*.
- Theorem. $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*, \quad \forall \boldsymbol{\lambda} \succeq 0, \ \boldsymbol{\nu}$
- Proof. Consider feasible $\tilde{\boldsymbol{x}}$. Then,

  $$\mathcal{L}(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{m} \lambda_i g_i(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{\boldsymbol{x}}) \leq f(\tilde{\boldsymbol{x}})$$

  since $f_i(\tilde{\boldsymbol{x}}) \leq 0$ and $\lambda_i \geq 0$.
  Hence, $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathcal{L}(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\boldsymbol{x}})$ for all feasible $\tilde{\boldsymbol{x}}$. Therefore, $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$.

Lagrangian Dual Problem

- Lower bound from Lagrange dual function depends on $(\boldsymbol{\lambda}, \boldsymbol{\nu})$.
- Question. What's the best lower bound?

Langrangian dual problem
$$\begin{array}{ll} \text{maximize} & \mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} & \boldsymbol{\lambda} \succeq 0 \end{array}$$

- Dual variables: $(\boldsymbol{\lambda}, \boldsymbol{\nu})$
- Always a convex optimization, because $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is always concave over $\boldsymbol{\lambda}, \boldsymbol{\nu}$.
  - Infimum over $\boldsymbol{x}$ of a family of affine functions in $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ (we will see this later)
- Denote the optimal value of Lagrange dual problem by d^*.

Weak Duality

- What's the relationship between d^* and p^*?

Weak Duality

$d^* \leq p^*$

- Weak duality always hold (even if the primal problem is not convex):
- Optimal duality gap: $p^* - d^*$
- Efficient generation of the lower bounds through the dual problem

Convex Optimization

- Convex optimization problem

  $$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{X}, \end{aligned}$$

  where $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is a convex function, and $\mathcal{X}$ is a convex set.
- The watershed between easily solvable problem and intractable ones is not ‘linearity’, but ‘convexity’
- Let’s overview the background of convex functions, convex sets, and their basic properties.

Convex Set

- Set $\mathcal{C}$ is a convex set if the line segment between any two points in $\mathcal{C}$ lies in $\mathcal{C}$, i.e., if for any $x_1, x_2 \in \mathcal{C}$ and any $\theta \in [0, 1]$, we have $\theta x_1 + (1 - \theta)x_2 \in \mathcal{C}$

- Convex hull of $\mathcal{C}$ is the set of all convex combinations of points in $\mathcal{C}$:

  $$\left\{ \sum_{i=1}^{k} \theta_i x_i \mid x_i \in \mathcal{C}, \theta_i \geq 0, i = 1, 2, \ldots, k, \sum_{i=1}^{k} \theta_i = 1 \right\}$$

  - What is k? For all k? For some k?

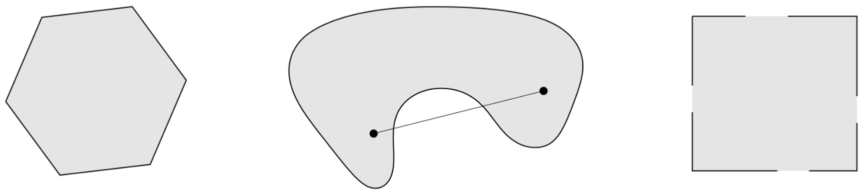
- Generalize to infinite sums and integrals:

  $$\sum_{i=1}^{\infty} \theta_i x_i \in \mathcal{C}, \quad \int_{\mathcal{C}} p(x) x dx \in \mathcal{C},$$

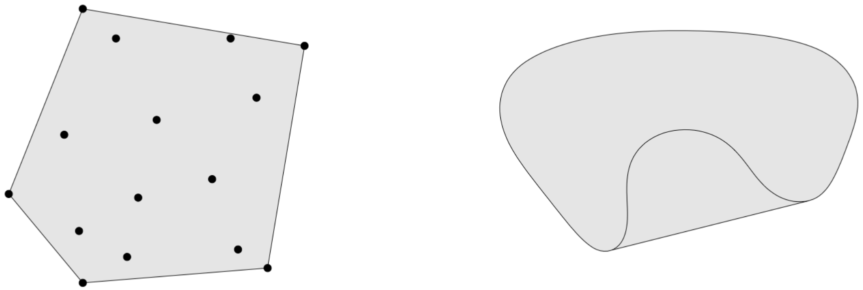
  where $\sum_{i=1}^{\infty} \theta_i = 1$ and $p(x)$ is a pdf of some random variable.

Examples

- Convex and Non-convex sets



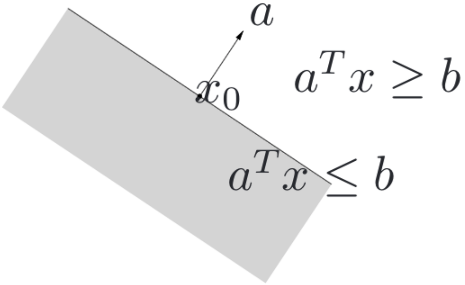
- Convex hulls



Examples of Convex Sets

- Hyperplane in $\mathbb{R}^n$ is a set: $\{x \mid a^{\mathsf{T}} x = b\}$ where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$
  In other words, $\{x \mid a^{\mathsf{T}}(x - x_0) = 0\}$, where x_0 is any point in the hyperplane, i.e., $a^{\mathsf{T}} x_0 = b$.

- Divides $\mathbb{R}^n$ into two halfspaces:
  $\{x|a^{\mathsf{T}}x \leq b\}$ and $\{x|a^{\mathsf{T}}x > b\}$

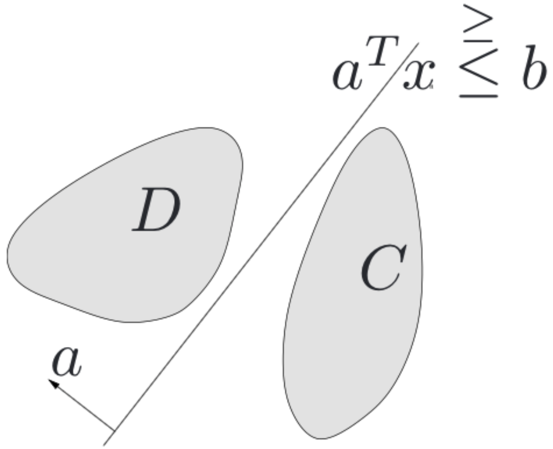


- Polyhedron is the solution set of a finite number of linear equalities and inequalities (intersection of finite number of halfspaces and hyperplanes)

  $$\mathcal{P} = \{x \mid a_j^{\mathsf{T}} x \leq b_j, j = 1, \ldots, m, c_j^{\mathsf{T}} x = d_j, j = 1, \ldots, p\} = \{x \mid Ax \leq b, Cx = d\}$$

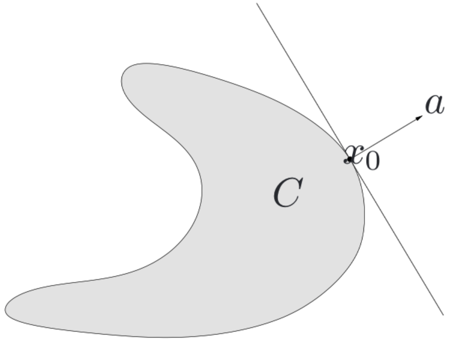
- Polytope: a bounded polyhedron

Separating Hyperplane Theorem



- $\mathcal{C}$ and $\mathcal{D}$: non-intersecting convex sets, i.e., $\mathcal{C} \bigcap \mathcal{D} = \phi$.
- Then, there exist $a \neq 0$ and b such that $a^{\mathsf{T}} x \leq b$ for all $x \in \mathcal{C}$ and $a^{\mathsf{T}} x \geq b$ for all $x \in \mathcal{D}$.

Supporting Hyperplane Theorem



- Given a set $\mathcal{C} \in \mathbb{R}^n$ and a point x_0 on its boundary, if $a \neq 0$ satisfies $a^\mathsf{T} x \leq a^\mathsf{T} x_0$ for all $x \in \mathcal{C}$, then $\{x | a^\mathsf{T} x = a^\mathsf{T} x_0\}$ is called a supporting hyperplane to $\mathcal{C}$ at x_0
- For any nonempty convex set $\mathcal{C}$ and any x_0 on boundary of $\mathcal{C}$, there exists a supporting hyperplane to $\mathcal{C}$ at x_0
- What happens if $\mathcal{C}$ is non-convex?

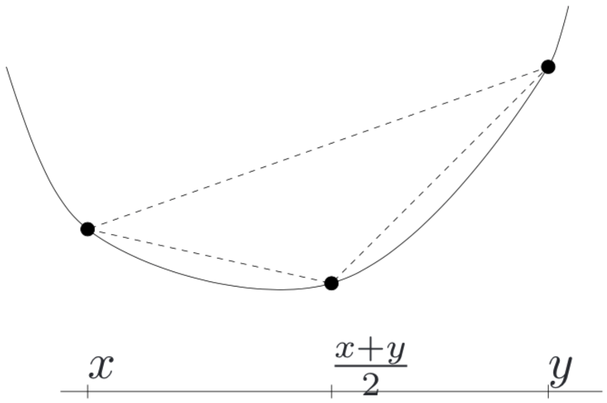
Convex Functions

- $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a convex function if dom f is a convex set and for all $x, y \in$ dom f and $\theta \in [0, 1]$, we have

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

- f is strictly convex if the strict inequality in the above holds for all $x \neq y$ and $0 < \theta < 1$.

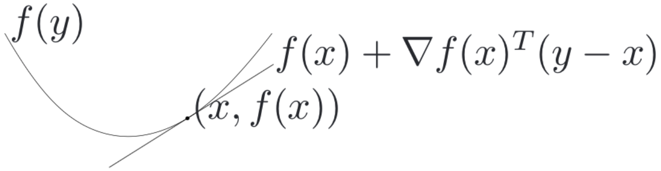
- f is concave if $-f$ is convex
- Affine functions are convex and concave
- Jensen's inequality. For a rv X, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.



Conditions of Convex Functions (1)

- First-order condition. For differentiable functions, f is convex iff

  $f(y) - f(x) \geq \nabla f(x)^\mathsf{T}(y - x), \quad \forall x, y \in \text{dom } f$, and dom f is convex



- Example. $f(y) = y^2$.
- $f(y) \geq \tilde{f}_x(y)$ where $\tilde{f}_x(y)$ is the first order Taylor expansion of $f(y)$ at x.
- Local information (first order Taylor approximation) about a convex function provides global information (global underestimator).
- If $\nabla f(x) = 0$, then $f(y) \geq f(x)$, $\forall y$. Thus, x is a global minimizer of f

Conditions of Convex Functions (2)

- Second-order condition. For twice differentiable functions, f is convex iff

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

  for all $x \in \operatorname{dom} f$ (upward slope) and $\operatorname{dom} f$ is convex
- Example: $f(x) = x^2$.
- Meaning: The graph of the function have positive (upward) curvature at x.

Examples of Convex or Concave Functions

- e^{ax} is convex on $\mathbb{R}$, for any $a \in \mathbb{R}$
- x^a is convex on $\mathbb{R}_{++}$ when $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$
- $|x|^p$ is convex on $\mathbb{R}$ for $p \geq 1$
- $\log x$ is concave on $\mathbb{R}_{++}$
- $x \log x$ is strictly convex on $\mathbb{R}_{++}$
- Every norm on $\mathbb{R}^n$ is convex
- $f(x) = \max\{x_1, \ldots, x_n\}$ is convex on $\mathbb{R}^n$
- $f(x) = \log \sum_{i=1}^n e^{x_i}$ is convex on $\mathbb{R}^n$
- $f(x) = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$ is concave on $\mathbb{R}^n_{++}$

Convexity-Preserving Operations

- $f = \sum_{i=1}^{n} w_i f_i$ convex if f_i are all convex and $w_i \geq 0$
- $g(x) = f(Ax + b)$ is convex iff $f(x)$ is convex
- $f(x) = \max\{f_1(x), f_2(x)\}$ convex if f_i convex, e.g., sum of r largest components is convex
- $f(x) = h(g(x))$, where $h : \mathbb{R}^k \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \mapsto \mathbb{R}^k$.

  If $k = 1$: $f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$. So, f is convex if h is convex and nondecreasing and g is convex, or if h is convex and nonincreasing and g is concave ...