EE5239: Nonlinear Optimization Notes

目录

1 Lecture 1: Unconstrained Optimization

Outline

- Definition
- Existence of optimal solution
- Necessary and sufficient conditions for local minima
- How to use these conditions
- Convexity
- All local minima are global minima for convex problems

1.1 Unconstrained Optimization Problem

We consider the problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.

Local minimum. A point x^* is called a *local minimum* if there exists $\varepsilon > 0$ such that

$$f(x) \geq f(x^*) \quad \forall x,\ \|x - x^*\| \leq \varepsilon.$$

Global minimum. A point x^* is a *global minimum* if

$$f(x) \geq f(x^*) \quad \forall x \in \mathbb{R}^n.$$

1.2 Existence of an Optimal Solution

By the **Bolzano–Weierstrass theorem**, every continuous function f attains its infimum on a compact set X. That is, there exists $x^* \in X$ such that

$$f(x^*) = \inf_{x \in X} f(x).$$

Alternatively, if the *level set*

$$\{x \mid f(x) \leq f(x^0)\}$$

is compact for some x^0, then f achieves a global minimum on this set.

1.3 Checkable Conditions for Local Minima

The following are the **necessary conditions:**

$$\nabla f(x^*) = 0 \qquad \text{(first-order condition)}$$
$$\nabla^2 f(x^*) \succeq 0 \qquad \text{(second-order condition).}$$

Such points x^* are called *stationary points*.

Sufficient condition:

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succ 0 \implies x^* \text{ is a strict local minimum.}$$

1.4 Mean Value Theorem (MVT)

For a differentiable scalar function f,

$$f(a) - f(b) = f'(c)(a - b),$$

for some $c \in (a, b)$.

Alternatively,

$$f(a) - f(b) = f'(b)(a - b) + \frac{1}{2}(a - b)^2 f''(\xi_c),$$

where ξ_c lies between a and b.

The MVT suggests that there exists $\alpha \in (0, 1)$ such that

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2}(x - x^*)^2 f''(x^* + \alpha(x - x^*)).$$

Applying the sufficient condition for local minimum: for any x sufficiently close to x^*,

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^2 f''(x^* + \alpha(x - x^*)) \geq 0.$$

This holds because $f''(x^*) > 0$, so there exists a neighborhood around x^* in which $f''(x) > 0$.

1.5 Use of Optimality Conditions

The optimality conditions provide a systematic way to determine whether a stationary point of a differentiable function corresponds to a local minimum, local maximum, or saddle point.

Consider the unconstrained optimization problem:

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable.

First-Order Necessary Condition. If x^* is a local minimum of f, then the gradient must vanish:

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

Such a point x^* is called a **stationary point** (or critical point). At this point, the directional derivative of f in any direction is zero, indicating that no first-order descent direction exists.

Second-Order Necessary Condition. If x^* is a local minimum of f and f is twice differentiable, then the Hessian must be positive semidefinite:

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

This condition is *necessary* but not sufficient; a positive semidefinite Hessian only ensures that no direction yields negative curvature locally, but does not rule out flat or saddle regions.

Second-Order Sufficient Condition. If $\nabla f(x^*) = 0$ and the Hessian is positive definite,

$$\nabla^2 f(x^*) \succ 0,$$

then x^* is a **strict local minimum**. Analogously:

- If $\nabla^2 f(x^*) \prec 0$, then x^* is a **strict local maximum**.
- If $\nabla^2 f(x^*)$ has both positive and negative eigenvalues, then x^* is a **saddle point**.

Summary.

$$\begin{aligned} &\text{Necessary for local minimum:} && \nabla f(x^*) = 0,\ \nabla^2 f(x^*) \succeq 0, \\ &\text{Sufficient for strict local minimum:} && \nabla f(x^*) = 0,\ \nabla^2 f(x^*) \succ 0. \end{aligned}$$

These optimality conditions form the foundation for analyzing stationary points in both unconstrained and constrained optimization problems.

1.6 Convexity

Convex functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if, for all $x, y \in \mathbb{R}^n$ and all $\lambda \in [0, 1]$,

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

Convex sets. A set $S \subseteq \mathbb{R}^n$ is convex if, for any $x, y \in S$ and $\lambda \in [0, 1]$,

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

Properties of convex functions.

- If $f(x)$ is convex, then $-f(x)$ is concave.
- If $f_1(x), f_2(x)$ are convex, then $f_1(x) + f_2(x)$ is convex.
- If $f_1(x), f_2(x)$ are convex and $a, b > 0$, then $g(x) = af_1(x) + bf_2(x)$ is convex.
- For a smooth scalar function f, it is convex if and only if its second-order derivative is nonnegative:

  $$f''(x) \geq 0.$$

  For multivariate problems, convexity is equivalent to a positive semidefinite Hessian:

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

2 Lecture 2: Gradient Methods

Overview

- Basic steps of gradient methods
- How to choose stepsizes
- Descent Lemma
- Estimate the descent of steepest descent with constant stepsize
- Analysis of steepest gradient descent algorithm with $1/L$ stepsize
- General analysis of convergence
- Strong convexity
- Definition of condition number
- Convergence rate analysis steps
- Gradient–type algorithm also depends on condition # (condition number)

2.1 Gradient Descent Method

If $\nabla f(x^*) = 0$, then x^* is a candidate solution. If $\nabla f(x) \neq 0$, there exists an interval $(0, \delta)$ such that

$$f\big(x - \alpha \nabla f(x)\big) < f(x), \quad \forall\, \alpha \in (0, \delta).$$

More generally, for a given direction d^r that is with obtuse angle w.r.t. $\nabla f(x)$ (i.e. $\langle \nabla f(x), d\rangle < 0$), there exists an interval $(0, \delta)$ of stepsizes with

$$f(x + \alpha d) < f(x), \quad \forall\, \alpha \in (0, \delta).$$

Iterative form.

$$x^{r+1} = x^r + \alpha_r d^r, \qquad r = 0, 1, 2, \ldots$$

where, if $\nabla f(x^r) \neq 0$, the direction d^r satisfies $\nabla f(x^r)^\top d^r < 0$ and $\alpha_r > 0$ is a stepsize.

General case (gradient–type methods).

$$x^{r+1} = x^r - \alpha_r D^r \nabla f(x^r), \quad r = 0, 1, \ldots,$$

with $D^r \succ 0$ a positive definite matrix (*called scaling matrix*).

Special cases.

Steepest descent: $x^{r+1} = x^r - \alpha_r \nabla f(x^r)$.

Newton's method: $x^{r+1} = x^r - \alpha_r \big[\nabla^2 f(x^r)\big]^{-1} \nabla f(x^r)$. （二次收敛很快；每步代价高且数值稳定性要注

2.2 Choice of Stepsize (*step*)

- **Constant stepsize:** $\alpha_r = \alpha$.
- **Minimization rule:** $\alpha_r = \arg\min_{\alpha>0} f(x^r + \alpha d^r)$ (*maximum reduction, but expensive*).
- **Limited minimization rule:** $\alpha_r = \arg\min_{\alpha\in[0,s]} f(x^r + \alpha d^r)$.
- **Diminishing stepsize:** $\alpha_r \to 0$, $\sum_{r=0}^{\infty} \alpha_r = \infty$.
  （接近最优点时步长变得很小；常见如 $\alpha_r = \frac{1}{r}$、$\alpha_r = \frac{c}{r}$、或 $\alpha_r = \frac{1}{\sqrt{r}}$；保证最终收敛）
- **Armijo rule (backtracking).** Let $0 < \sigma < \frac{1}{2}$ and $0 < \beta < 1$ be constants. Starting from a trial α, keep shrinking by $\alpha \leftarrow \beta\alpha$ (即 $\beta, \beta^2, \beta^3, \ldots$) until

$$f(x^r + \alpha d^r) \le f(x^r) + \sigma\, \alpha\, \nabla f(x^r)^\top d^r.$$

  （能保证足够下降 *sufficient descent*，但可能需要测试多次）

2.3 Analysis of GD Method & Descent Lemma

Assume f has L-Lipschitz gradient:

$\|\nabla f(x) - \nabla f(y)\| \le L\,\|x - y\|, \quad \forall x, y.$ （梯度 *Lipschitz*；意味着曲率有上界/是 *bounded* 的）

Then for all x, y,

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 =: u(x; y). \qquad \textbf{(Descent Lemma)}$$

（$u(x;y)$ 是以 y 为中心的二次上界 *upper model*）

Minimizing the upper model. Minimizing $u(x; y)$ w.r.t. x gives

$$x^{r+1} = x^r - \frac{1}{L}\nabla f(x^r).$$

In particular,

$$f\Big(x^r - \frac{1}{L}\nabla f(x^r)\Big) \le f(x^r) - \frac{1}{2L}\|\nabla f(x^r)\|^2,$$

which shows *sufficient descent*. （每次至少下降 $\frac{1}{2L}\|\nabla f(x^r)\|^2$）

Proof sketch on paper reproduced in steps. Using Lipschitz gradient and integrating the directional derivative along the segment $y \to x$:

$$\begin{aligned} f(x) &= f(y) + \int_0^1 \langle \nabla f\big(y + t(x-y)\big),\, x - y\rangle \, dt \\ &\le f(y) + \langle \nabla f(y), x - y\rangle + \int_0^1 \|\nabla f(y + t(x-y)) - \nabla f(y)\| \, \|x - y\| \, dt \\ &\le f(y) + \langle \nabla f(y), x - y\rangle + \int_0^1 Lt\|x - y\|^2 \, dt = f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2. \end{aligned}$$

2.4 General Analysis for Convergence

Prove convergence to points that satisfy first-order optimality (*stationary*). A *gradient-related* condition: for any sequence $\{x^r\}$ that converges to a nonstationary point, the corresponding directions $\{d^r\}$ are bounded and satisfy

$$\lim_{r\to\infty} \langle \nabla f(x^r), d^r\rangle < 0.$$

This holds for $d^r = -D^r \nabla f(x^r)$ with $D^r \succ 0$. If $d^r = -\nabla f(x^r)$ and $\alpha_r \in (0, 2/L)$, then the sufficient-descent inequality above holds; 因此我们可以选 $\alpha_r = \dfrac{1}{L}$.

Summary: GD converges to first-order optimality as the iteration number goes to infinity.

2.5 Strong Convex Function (SC)

A function f is *strongly convex* iff there exists $\delta > 0$ such that

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

(整个函数有足够的“弯曲”——没有“平坦”区域)

If f is twice continuously differentiable, then there exists $\delta > 0$ such that

$$\delta I \preceq \nabla^2 f(x) \preceq LI, \quad \forall x. \quad (\text{从 } \textit{Hessian} \text{ 角度看：} \delta I \text{ 为下界，} LI \text{ 为上界})$$

For symmetric matrices, $A \succeq B$ means $A - B \succeq 0$ (semidefinite). E.g. $f(x) = \frac{1}{2}x^\top A x$ with strictly positive definite A implies $A \succeq \delta I$ (这里 δ 是 A 的最小特征值).

2.6 GD for Strongly Convex Functions

Step 1 (sufficient descent). With $\alpha = 1/L$,

$$f(x^{r+1}) = f\big(x^r - \tfrac{1}{L}\nabla f(x^r)\big) \le f(x^r) - \frac{1}{2L}\|\nabla f(x^r)\|^2.$$

Step 2 (bound the error by gradient). Let $e(x) := f(x) - f(x^*)$. Strong convexity yields

$$\|\nabla f(x)\|^2 \geq 2\delta \left(f(x) - f(x^*)\right) = 2\delta\, e(x).$$

Combining with Step 1 gives

$$e(x^{r+1}) \leq \left(1 - \frac{\delta}{L}\right) e(x^r) =: \beta\, e(x^r), \quad \beta = 1 - \frac{\delta}{L} \in (0,1),$$

i.e. *linear convergence.* （同上每步至少按同一比例减少）

2.7 Condition Number

$$\kappa = \frac{L}{\delta} \quad (\textit{condition number}).$$

Here δ is the smallest eigenvalue of the Hessian of f (and L is the largest eigenvalue, e.g. for quadratic problems).

- Large $\kappa \Rightarrow$ large β (*ill-conditioned / slow convergence*).
- Small $\kappa \Rightarrow$ small β (*well-conditioned / fast convergence*).

2.8 Convergence Rate Analysis

Let $e^r := e(x^r) = f(x^r) - f(x^*)$.

- **Linear convergence** means: $\exists\, \beta \in (0,1)$ such that $\limsup_{r\to\infty} \dfrac{e^{r+1}}{e^r} \leq \beta$. Equivalently, if it holds for all r,

$$e^{r+1} \leq \beta e^r \quad \Longrightarrow \quad \ln e^{r+1} \leq \ln\beta + \ln e^r.$$

- **Superlinear convergence** means: $\limsup_{r\to\infty} \dfrac{e^{r+1}}{(e^r)^p} < \beta$ for some constant $p > 1$. （误差的“阶数”大于线性，极限意义上满足的）

2.9 Convergence Rate for SC

If $e(x^0) = D_0$, then $e(x^r) \leq \beta^r D_0 \leq \varepsilon$, so

$$r \;\geq\; \frac{\ln(D_0/\varepsilon)}{\ln(1/\beta)}.$$

（通过对数变换把迭代次数 r 表达出来；只要右边的 ε 足够小，就能保证获得 ε-*optimal* 解，$x_\varepsilon := \{x :\, f(x) - f(x^*) \leq \varepsilon\}$）

以上针对 *strong convex*；对一般 *convex*，收敛为 *sublinear*，粗略量级 $r \gtrsim \frac{1}{\varepsilon}$.

2.10 Scaling Variable

Suppose we have data sets $\{(a_i, b_i)\}_{i=1}^M$ with $a_i \in \mathbb{R}^k$. Arrange them into (A, b) with $A \in \mathbb{R}^{M\times k}$.

(1) **For each feature (column):** standardize

$$A'_k \;=\; \frac{A_k - \bar{A}_k}{\sigma_{A_k}}.$$

(2) **For b:** center

$$b' \;=\; b - \bar{b} \qquad \textit{(centering).}$$

3 Lecture 3: First-Order Methods

Overview

- Conjugate direction
- Conjugate gradient method
- Q-conjugate vectors, linearly independent vectors, and orthogonal vectors
- How scaled steepest descent works in theory
- Incremental Gradient Method
- Coordinate Gradient Method

3.1 Conjugate Direction

Directions $d^0, d^1, \ldots, d^r$ are **Q-conjugate** if

$$(d^i)^\top Q d^j = 0, \quad \text{for } i \neq j.$$

These conjugate directions are linearly independent. （几何上，它们在 Q 加权内积下正交；在优化问题中代表“互不干扰”的下降方向）

对于二次型函数：

$$f(x) = \tfrac{1}{2}x^\top Q x - b^\top x,$$

其中 Q 是对称正定矩阵（SPD matrix）。

—

3.2 Conjugate Gradient Method (CG)

用于求解二次优化问题：

$$x^{r+1} = x^r + \alpha_r d^r,$$

其中 d^r 为 Q-共轭方向，α_r 由 **line minimization（线搜索）**得到。

思想： The CG method is a conjugate direction method where the search direction is generated using gradients.

从理论上，CG 可以视为对梯度进行 Gram-Schmidt 正交化的结果：

$$d^r = -g^r + \sum_{j=0}^{r-1} \beta_j d^j, \quad \text{其中} g^r = \nabla f(x^r).$$

—

Algorithm Summary

Let $g^r = \nabla f(x^r) = Qx^r - b$. Then:

$$\begin{aligned}
d^0 &= -g^0, \\
\alpha_r &= \frac{(g^r)^\top d^r}{(d^r)^\top Q d^r}, \\
x^{r+1} &= x^r + \alpha_r d^r, \\
g^{r+1} &= Qx^{r+1} - b = g^r + \alpha_r Q d^r, \\
\beta_r &= \frac{(g^{r+1})^\top Q d^r}{(d^r)^\top Q d^r}, \\
d^{r+1} &= -g^{r+1} + \beta_r d^r.
\end{aligned}$$

最终：

$$x^{r+1} = x^r + \alpha_r d^r.$$

（每次迭代生成一个新的共轭方向；n 步后可精确解出二次型问题的最优解）

—

3.3 Scaled Steepest Descent

$$x^{r+1} = x^r - \alpha_r D^r \nabla f(x^r),$$

其中 $D^r \succ 0$ 是某个正定矩阵（常见选择为对角矩阵）。

若取 $D^r = D$（即常量矩阵，所有迭代中相同），则算法为

$$x^{r+1} = x^r - \alpha_r D \nabla f(x^r).$$

—

变量变换解释： 令 $x = Sy$, 其中 $S = (D)^{1/2}$。那么在 y 空间中，优化问题变为

$$\min_y h(y) = f(Sy).$$

对 y 应用普通的最速下降法：

$$y^{r+1} = y^r - \alpha_r \nabla h(y^r).$$

两边同时乘 S：

$$Sy^{r+1} = Sy^r - \alpha_r S \nabla h(y^r)$$

等价于

$$x^{r+1} = x^r - \alpha_r D \nabla f(x^r),$$

即原来的 scaled steepest descent。

（几何意义：通过线性变换重新刻画不同坐标方向的“拉伸”，从而改善收敛速度。）

—

3.4 Incremental Gradient Method

考虑问题：

$$\min_x f(x) = \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}\sum_{i=1}^{m}(a_i^\top x - b_i)^2 = \frac{1}{m}\sum_{i=1}^{m} g_i(x),$$

其中 $g_i(x)$ 表示样本 i 的损失项。

—

算法步骤： 在第 r 次外层迭代中：

初始化：$\psi_0 = x^r$.

内层循环：$\psi_i = \psi_{i-1} - \alpha_r \nabla g_i(\psi_{i-1}), \quad i = 1, 2, \ldots, m.$

更新外层：$x^{r+1} = \psi_m$.

（每次只使用一部分样本更新梯度，可理解为 *mini-batch* 或 *online gradient* 方法。）

3.5 Coordinate Descent Method （坐标下降法）

不同于优化所有变量，而是一次仅优化一个变量（或一组变量），其余变量保持不变。只沿着坐标轴的方向搜索最小值。

—

Version I: Exact minimization 在第 $t+1$ 次迭代中，选择第 i 个坐标：

$$i = \mathrm{rem}(t, n) + 1 \quad (\text{序数 } index,\ \text{变量编号})$$

然后：

$$x_i^{t+1} = \arg\min_{x_i} f(x_1^t, \ldots, x_i, \ldots, x_n^t), \qquad x_j^{t+1} = x_j^t,\ j \neq i.$$

—

Version II: Gradient step 从精确最小化（Version I）改为使用梯度更新（gradient update）：

$$i = \mathrm{rem}(t, n) + 1, \qquad x_i^{t+1} = x_i^t - \alpha_i\, \partial_i f(x^t), \qquad x_j^{t+1} = x_j^t,\ j \neq i.$$

—

Stepsize 选择 步长 α_i 与第 i 个子问题的局部曲率（curvature）相关：

$$\alpha_i \sim \frac{1}{L_i}, \quad \text{其中} L_i \ll L \Rightarrow \text{曲率小、步长大。}$$

—

Pick coordinate 的方法

1. **Cyclic:** $1, 2, \ldots, n, 1, 2, \ldots, n$
2. **Randomized:** 每个坐标被随机选择，概率相同
3. **Permuted:** 每一轮随机打乱顺序后依次选择（sample without replacement）

最小二乘问题优化示例

$$\min_{\mathbf{x}} \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|^2$$

How to do coordinate descent?

$$\min_{\mathbf{x}} \frac{1}{2} \left\| \sum_{i=1}^{n} A_i x_i - \mathbf{b} \right\|^2$$

where A_i represents the ith column of A, and x_i is the ith element of $\mathbf{x}$.

Minimize with the ith coordinate:

$$\min_{x_i} \frac{1}{2} \left\| A_i x_i + \sum_{j \neq i}^{n} A_j x_j^{(r)} - \mathbf{b} \right\|^2$$

Solution is very simple (no matrix inversion needed!):

$$x_i^{(r+1)} = \frac{A_i^\top \left(\mathbf{b} - \sum_{j \neq i}^{n} A_j x_j^{(r)} \right)}{A_i^\top A_i}$$

—

4 Lecture 4: First-Order Methods —Examples

Overview

- k-means formulation
- Apply coordinate descent to k-means
- Implement the Perceptron algorithm

4.1 k-Means Clustering

Formulation.

$$L(r, \mu) = \sum_{m=1}^{M} \sum_{p=1}^{P} r_{mp} \|a_m - \mu_p\|^2$$

Decision variables are $r_{mp} \in \{0, 1\}$ and cluster means μ_p.

—

k-Means clustering idea. Optimization is performed by using **block coordinate descent**:

- First block coordinates: $\{r_{mp}\}$
- Second block coordinates: $\{\mu_p\}$

Initialize with arbitrary $\{\mu_p\}$. Iteratively update until convergence.

—

Clustering steps.

- **Update for r_{mp}:** Assign a_m to the nearest cluster mean μ_p:

$$r_{mp} = \begin{cases} 1, & \text{if } p = \arg\min_j \|a_m - \mu_j\|^2, \\ 0, & \text{else.} \end{cases}$$

(每个样本点分配给距离最近的聚类中心)

- **Update for μ_p:** Optimize $L(r, \mu)$ w.r.t. μ_p 得到:

$$\mu_p = \frac{\sum_m r_{mp} a_m}{\sum_m r_{mp}} = \frac{\sum_{m \in C_p} a_m}{|C_p|},$$

where $|C_p|$ denotes the number of elements in cluster p.
(即每个聚类中心等于该簇所有点的平均值)

4.2 Perceptron Algorithm

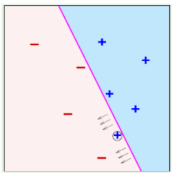
A simple learning algorithm (cont.)

- At iteration $t = 1, 2, \cdots$, continue pick misclassified point from

$$(\mathbf{a}_1, b_1), (\mathbf{a}_2, b_2), \cdots$$

and keep running the algorithm

- If the training data is linearly separable, then a correct separation plane will be found after finite number of iterations



A simple learning algorithm "Perceptron"

- Suppose we start at an arbitrary solution $\mathbf{x}$
- Pick a misclassified point: $\text{sgn}((\mathbf{x})^T \mathbf{a}_n) \neq b_n$, or $b_n(\mathbf{x})^T \mathbf{a}_n < 0$
- Update the weight vector by

$$\mathbf{x} \leftarrow \mathbf{x} + b_n \mathbf{a}_n$$

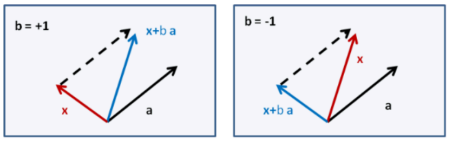


Figure 0.1: The update rules (Y. Abu-Mostafa).

The Linear Classification Training: Revisited

- Let's consider the objective function (why I am minimizing?)

$$L(\mathbf{x}) = \sum_{i=1}^{m} \max\{-b_i \mathbf{x}^T \mathbf{a}_i, 0\} := \sum_{i=1}^{m} g_i(\mathbf{x})$$

- For each miss-classified data point i, $-b_i \mathbf{x}^T \mathbf{a}_i > 0$; the gradient

$$\nabla g_i(\mathbf{x}) = -b_i \mathbf{a}_i$$

- For each correct data point i, $-b_i \mathbf{x}^T \mathbf{a}_i < 0$; gradient is zero!
- Directly applying the incremental method?

$$\mathbf{x}^{r+1} = \mathbf{x}^r + \alpha^r b_i \mathbf{a}_i$$

- **Good News**: $\alpha^r = 1$!

The Convergence

- Define $\mathbf{u}$, with $\|\mathbf{u}\| = 1$, be an optimal solution, i.e.,

$$b_i \mathbf{u}^T \mathbf{a}_i \geq \gamma > 0, \ \forall\, i$$

where γ is the largest constant that satisfies the above inequality

- Let $R := \max_i \|\mathbf{a}_i\|$
- **Claim**: If the data points are separable, and we start with $\mathbf{x}_0 = 0$, then training stops after at most $\left(\frac{R}{\gamma}\right)^2$ iterations
- Exact Convergence in Finite Steps!

5 Lecture 5: Case Study: Logistic Regression

Overview

- What causes different practical behavior for algorithms
- The basic idea behind logistic regression
- Strong convexity and strict convexity
- Why different datasets yield different behavior for GD algorithm

5.1 When Does Gradient Descent Converge?

若步长（step size）满足：

$$\alpha < \frac{2}{L}, \quad \text{with } L \text{ being Lipschitz constant.}$$

使用 line search 方法，不固定步长 α，最终每步寻求最优下降量。

5.2 Logistic Regression

预测一个数值量（numerical quantity）使用 $x^\top w$。预测类别 $(-1, 1)$ 时，使用：

$$\hat{y} = \operatorname{sgn}(x^\top w).$$

新的模型用于预测概率（probability）：

$$h(x) = \theta(x^\top w),$$

其中 $\theta(z)$ 是 logistic 函数：

$$\theta(z) = \frac{e^z}{1+e^z}, \quad \theta(z) \in (0, 1).$$

学习目标是概率：

$$h(x) = P(y = 1 \mid x),$$

其对应的后验分布为：

$$P(y \mid x) = \begin{cases} h(x), & y = +1, \\ 1 - h(x), & y = -1. \end{cases}$$

或等价地写作：

$$P(y \mid x) = \theta(yx^\top w).$$

学习目标是最大化：

$$\frac{1}{n}\sum_{i=1}^{n}\ln P(y_i \mid x_i),$$

即最小化 loss function：

$$L(w)=\frac{1}{n}\sum_{i=1}^{n}\ln\Big(1+e^{-y_i x_i^\top w}\Big).$$

5.3 Gradient of the Loss

梯度为：

$$\nabla L(w)=-\frac{1}{n}\sum_{i=1}^{n}\frac{y_i x_i}{1+e^{y_i x_i^\top w}}=-\frac{1}{n}\sum_{i=1}^{n}(1-\theta(y_i x_i^\top w))y_i x_i.$$

其中 $\theta(t)=\frac{1}{1+e^{-t}}$。

5.4 Strict Convexity 和 Strong Convexity

(1) 普通凸性 (Convexity): 对所有 x,y 及 $t\in[0,1]$，若

$$f(tx+(1-t)y)\le tf(x)+(1-t)f(y),$$

则称 f 为凸函数。若 $\nabla^2 f(x)\succeq 0$，则 f 凸。

(2) 严格凸性 (Strict Convexity): 对所有 $x\neq y$ 及 $t\in(0,1)$，若

$$f(tx+(1-t)y)<tf(x)+(1-t)f(y),$$

则称 f 为严格凸函数。

对于 logistic regression:

$$f''(w)=\frac{e^{x_i^\top w}}{(1+e^{x_i^\top w})^2}>0,$$

因此为严格凸函数。

(3) 强凸性 (Strong Convexity): 若存在 $\delta>0$，使得对任意 x,y：

$$f(y)\ge f(x)+\nabla f(x)^\top(y-x)+\frac{\delta}{2}\|y-x\|^2,$$

则称 f 为强凸函数 (strongly convex)。

换句话说，除了凸性外，曲率还必须有下界：

$$\nabla^2 f(x)\succeq \delta I.$$

5.5 Stop Criteria

两个常用的终止指标：

$$\nabla f(x^r), \quad f(x^r) - f(x^{r+1}).$$

后者不适合用于 diminishing step size 的情况。

停止条件 (任一满足即可):

$$\|\nabla f(w^r)\|_2 \le \varepsilon, \quad \|f(w^r) - f(w^{r+1})\| \le \varepsilon, \quad \|w^{r+1} - w^r\| \le \varepsilon.$$

通常：

$$\varepsilon = 10^{-3} \text{ (low precision)}, \quad \varepsilon = 10^{-10} \text{ (high precision)}.$$

5.6 Different Data Settings

还是这个分类问题 (binary classification problem)。

(1) Separable Case 数据可被某超平面完全分开。在这种情况下，GD 会沿着可分边界快速下降：

$$O\left(\frac{1}{r}\right),$$

但最终参数 w 会无限增大（趋于无界），例如：

$$y_i x_i^\top w \to +\infty.$$

(2) Non-separable Case 上界样本杂糅在一起，此时：

$$L(w) = \text{有限值}, \quad w \text{ 有界}.$$

此时收敛速率为：

$$O\left(\frac{1}{r}\right),$$

并达到一个有限最优点。

—

6 Lecture 6a: Optimization Over Constraints （有约束优化）

6.1 有无约束的差异

- 无约束（unc.）：x 不受限。
- 有约束（constrained）：$x \in C$，其中 C 是可行域。

—

6.2 The Problem We Are Analyzing

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

Suppose X is a convex set and f is continuously differentiable.

(a) 若 x^* 是局部最优解（local minimum），则

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in X.$$

这是**必要条件**(necessary condition);若 f 是凸函数,则该条件也是充分条件(sufficient condition)。

(b) 若 f 是凸函数，则该条件等价于最优性；若 f 非凸，则该条件仅对应**驻点**（stationary point）。

Convex set:

$$g_i(x) \leq 0 \text{ 是凸函数}, \quad h(x) = Cx + d = 0 \text{ 是仿射函数（affine function）}.$$

—

6.3 凸优化问题要求可行域是凸集

几何上：如果可行域 X 不是凸的，那么 x^* 前进或变动时，可能会中途离开可行域。对凸集 X，若 x^* 是最优点，则 $\nabla f(x^*)$ 与边界的法向方向一致。

—

6.4 Simple Scalar Quadratic Problem

(1) Case 1:

$$\min_{x\geq 0}\ \frac{1}{2}x^2 + a(a-x).$$

其约束可写为 $g(x) = -x \leq 0$。

拉格朗日函数为：

$$L(x,\lambda) = \frac{1}{2}x^2 + a(a-x) + \lambda(-x) = \frac{1}{2}x^2 - ax + a^2 - \lambda x.$$

KKT 条件：

$$\begin{cases} x \geq 0,\ \lambda \geq 0, & \text{(原始与对偶可行性)} \\ \lambda x = 0, & \text{(互补松弛)} \\ \nabla_x L = x - a - \lambda = 0. & \text{(站立性条件)} \end{cases}$$

由此得：

$$\begin{cases} a \geq 0 \Rightarrow x^* = a,\ \lambda^* = 0, \\ a < 0 \Rightarrow x^* = 0,\ \lambda^* = -a. \end{cases}$$

因此最优解可写为：

$$x^* = \max\{a, 0\}.$$

—

(2) Case 2:

$$\min_x \frac{1}{2}x^2, \quad \text{s.t. } a \leq x \leq b.$$

Lagrangian:

$$L(x,\lambda_1,\lambda_2) = \frac{1}{2}x^2 + \lambda_1(a-x) + \lambda_2(x-b).$$

KKT 条件：

$$\begin{cases} x \geq a,\ b \geq x,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0, \\ \lambda_1(a-x) = 0,\ \lambda_2(x-b) = 0, \\ x - \lambda_1 + \lambda_2 = 0. \end{cases}$$

由此得：

$$\begin{cases} a < 0 < b \Rightarrow x^* = 0,\ \lambda_1 = \lambda_2 = 0, \\ a > 0 \Rightarrow x^* = a,\ \lambda_1 = a, \\ b < 0 \Rightarrow x^* = b,\ \lambda_2 = -b. \end{cases}$$

—

6.5 Projection 投影方法

如果点 z 落在可行域 X 之外，可通过**投影（projection）**将其拉回：

$$\text{找到} x^* \in X \text{ 使} \|x^* - z\|^2 \text{ 最小化。}$$

$$x^* = \text{proj}_X(z) = \arg\min_{x \in X} \frac{1}{2}\|x - z\|^2.$$

几何性质：

$$\langle z - \text{proj}_X(z), x - \text{proj}_X(z) \rangle \leq 0, \quad \forall x \in X.$$

投影梯度法（projected gradient method）：

$$x^{r+1} = \text{proj}_X(x^r - \alpha_r \nabla f(x^r)),$$

其中 $\alpha_r > 0$ 为步长。

—

6.6 Compute Simple Projections

(1) Box constraint: 若 $a \leq x \leq b$，则

$$\text{proj}_{[a,b]}(z) = \min(\max(z, a), b).$$

(2) Euclidean ball: 若 $\|x\| \leq 1$，则

$$\text{proj}_{\|x\| \leq 1}(z) = \begin{cases} z, & \|z\| \leq 1, \\ \dfrac{z}{\|z\|}, & \|z\| > 1. \end{cases}$$

(3) Sphere: 若 $\|x\| = 1$，则

$$\text{proj}_{\|x\|=1}(z) = \frac{z}{\|z\|}.$$

—

6.7 Example 2.1.2: Optimization Over a Simplex

可行域：

$$X = \{x \mid x \geq 0,\ \sum_i x_i = r\}.$$

局部最优点的必要条件：

$$\nabla f(x^*)^\top (x - x^*) \geq 0, \quad \forall x \geq 0 \text{ 且} \sum_i x_i = r.$$

设 $x_i = 0,\ x_j, x_m > 0$，则：

$$\frac{\partial f(x^*)}{\partial x_i} \geq \frac{\partial f(x^*)}{\partial x_m}, \quad \forall i, j.$$

因此，所有在最优点处取正的变量，其对应的偏导数必须最小（partial cost derivative 最小）。

—

6.8 Example 2.1.1: Optimization Subject to Bounds

(1) For $X = \{x \mid x_i \geq 0\}$:

$$\frac{\partial f(x^*)}{\partial x_i} = \begin{cases} 0, & x_i^* > 0, \\ \geq 0, & x_i^* = 0. \end{cases}$$

几何意义：若位于内部，梯度为零；若在边界上，梯度分量必须指向外侧（nonnegative）。

—

(2) For $X = \{x \mid \alpha_i \leq x_i \leq \beta_i\}$:

$$\frac{\partial f(x^*)}{\partial x_i} = \begin{cases} > 0, & x_i^* = \alpha_i, \\ < 0, & x_i^* = \beta_i, \\ = 0, & \alpha_i < x_i^* < \beta_i. \end{cases}$$

7 Lecture 6b: Optimization Over Constraints (II)

Topics

- Feasible direction method

- Gradient Projection (GP) methods and their relationship with gradient descent
- Convergence result statements
- Perform projection on nonnegative set
- Convergence rate result
- Frank–Wolfe step and its geometric interpretation

—

7.1 Feasible Directions

A feasible direction at a point $x \in X$ is a vector $d \neq 0$ such that

$$x + \alpha d \in X, \quad \forall \text{ sufficiently small } \alpha > 0.$$

即：从 x 出发沿方向 d，对于足够小的步长 α，新点仍在可行域 X 内。可行方向集合定义为：

$$\mathcal{D}(x) = \{d = z - x \mid z \in X, z \neq x\}.$$

—

7.2 Feasible Direction Method

更新规则为：

$$x^{r+1} = x^r + \alpha_r d^r,$$

其中：

$$d^r \text{ 是一个 feasible descent direction，满足} \nabla f(x^r)^\top d^r < 0,$$

且 $\alpha_r > 0$ 使得 $x^{r+1} \in X$。

若 $\nabla f(x^r)^\top (x - x^r) \geq 0,\ \forall x \in X$，则 x^r 为 **stationary point**。

—

Step Size Rule:

(a) **Line minimization:**

$$\alpha_r = \arg\min_{\alpha > 0} f(x^r + \alpha d^r),$$

即在可行方向上直接最小化（计算量较大）。

(b) **Armijo Rule:**

$$\alpha_r = \beta^{m_r} s_r,$$

其中 m_r 为第一个非负整数，使得：

$$f(x^r + \beta^{m_r} s_r d^r) \leq f(x^r) + \sigma \beta^{m_r} s_r \nabla f(x^r)^\top d^r,$$

该条件保证下降方向的充分性（sufficient descent）。

(c) **固定步长（constant step size）：**

$$\alpha_r = 1.$$

—

7.3 Gradient Projection Methods (GP)

Idea: Find a feasible direction that is gradient-related.

更新公式：

$$x^{r+1} = x^r + \alpha_r(\bar{x}^r - x^r),$$

其中

$$\bar{x}^r = \mathrm{proj}_X[x^r - s_r \nabla f(x^r)], \quad \alpha_r \in (0, 1],\ s_r > 0.$$

—

Special Case I: 先沿负梯度方向找 s_r，即

$$s_r = \arg\min_{s>0} f(x^r - s\nabla f(x^r)),$$

然后再将结果投影回 X。(step-size rule for α_r)

Special Case II: 固定 s_r，先执行梯度下降，再将结果投影回 X：

$$x^{r+1} = \mathrm{proj}_X[x^r - s_r \nabla f(x^r)].$$

—

7.4 Convergence Result

Proposition 2.3.1: 固定 s_r，若 α_r 由 limited minimization rule 或 Armijo rule 选取，则序列 $\{x^r\}$ 的极限点为 stationary point。

Proof: $x^{r+1}-x^r$ 是 gradient-related，满足

$$\nabla f(x^r)^\top (x^{r+1}-x^r)<0.$$

注意：

$$\bar{x}^r-x^r=\mathrm{proj}_X[x^r-s_r\nabla f(x^r)]-x^r.$$

代入得：

$$\langle \nabla f(x^r),\bar{x}^r-x^r\rangle \le 0.$$

展开：

$$\langle \nabla f(x^r),x^{r+1}-x^r\rangle=\alpha_r\langle \nabla f(x^r),\bar{x}^r-x^r\rangle\le 0.$$

因此下降方向与梯度相反，算法保证下降。

—

Proposition 2.3.2: 若 $\alpha_r=1$ 且 s_r 满足下列条件：

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|,$$

则若 $0<s_r<2/L$，可保证 $\{x^r\}$ 收敛且极限点为 stationary 点。

—

7.5 Convergence Rate Analysis

考虑 $f(x)=\frac{1}{2}x^\top Ax+b^\top x$，其中 $A\succ 0$。

存在最优解 x^*，并且

$$x^{r+1}=\mathrm{proj}_X[x^r-s\nabla f(x^r)].$$

则：

$$\|x^{r+1}-x^*\|=\|\,\mathrm{proj}_X[x^r-s\nabla f(x^r)]-\mathrm{proj}_X[x^*-s\nabla f(x^*)]\|.$$

利用非扩张性（non-expansiveness）：

$$\|x^{r+1}-x^*\|\le\|(I-sA)(x^r-x^*)\|.$$

由此得到：

$$\|x^{r+1}-x^*\|\le\|I-sA\|\|x^r-x^*\|.$$

进一步估计：

$$\max\{|1-s\lambda_{\min}|,|1-s\lambda_{\max}|\}\Rightarrow\rho=\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}.$$

选择最优步长：

$$s = \frac{2}{\lambda_{\max} + \lambda_{\min}},$$

可得最优线性收敛率。

收敛速度取决于条件数 $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$，并且与无约束情形类似，达到 ε-精度大约需 $O(\kappa \ln(1/\varepsilon))$ 次迭代。

其中，$\lambda_{\max}$ 和 $\lambda_{\min}$ 分别表示矩阵 A 的最大与最小特征值（eigenvalues）。

—

7.6 Frank–Wolfe Method

定义更新规则：

$$x^{r+1} = x^r + \alpha_r(\bar{x}^r - x^r),$$

其中

$$\bar{x}^r = \arg\min_{x \in X} \nabla f(x^r)^\top (x - x^r).$$

假设 X 是紧集（compact），则 $\bar{x}^r$ 一定存在。

该方法收敛较慢，即使在 strongly convex 问题下仍如此。但其优点是**不需要投影（no projection needed）**。

8 Lecture 7: Lagrangian Multipliers

Main Topics

- Definition of Lagrangian function
- Dual function and dual problem
- Solve quadratic example in the "dual" domain
- Equality constrained problem
- Two interpretations of the theorem
- Lagrangian theorem for linear constrained problems (statement of result)
- Sensitivity: measure the change of objective by using multiplier

—

8.1 Problem to Solve

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t. } h_i(x) = 0, \quad i = 1, \ldots, m, \\ &\quad g_j(x) \leq 0, \quad j = 1, \ldots, r. \end{aligned}$$

The problem is convex if:

$$f(x) \text{ is convex}, \quad h_i(x) \text{ are affine (i.e. } h_i(x) = A_i x + b_i), \quad g_j(x) \text{ are convex.}$$

—

8.2 Lagrangian Function

The Lagrangian can be formed using multipliers $\lambda_j \geq 0$ and $\nu_i \in \mathbb{R}$:

$$L(x, \lambda, \nu) = f(x) + \sum_{j=1}^{r} \lambda_j g_j(x) + \sum_{i=1}^{m} \nu_i h_i(x).$$

Here: - λ_j: multipliers for inequality constraints, - ν_i: multipliers for equality constraints.

They can be viewed as *prices for violating the constraints.*

—

8.3 Dual Function and Dual Problem

Define the **dual function**:

$$L^*(\lambda, \nu) = \inf_{x \in X} L(x, \lambda, \nu) = \inf_{x \in X} \left[f(x) + \sum_{j=1}^{r} \lambda_j g_j(x) + \sum_{i=1}^{m} \nu_i h_i(x) \right].$$

The **dual problem** is:

$$\max_{\lambda, \nu} L^*(\lambda, \nu) \quad \text{s.t. } \lambda \geq 0.$$

- $L^*(\lambda, \nu)$ is a concave function (even if f is not convex).

- For $\lambda \geq 0$, we have $L^*(\lambda, \nu) \leq f^*$, since $f(x) + \lambda^\top g(x) + \nu^\top h(x) \leq f(x)$ for any feasible x, implying the weak duality property.

—

8.4 Quadratic Example

$$\min_x \|x\|^2 \quad \text{s.t. } Ax = b.$$

$$L(x, \nu) = \|x\|^2 + \nu^\top (Ax - b).$$

Dual function:

$$\begin{aligned} L^*(\nu) &= \inf_x L(x, \nu) \\ &= \inf_x \big(\|x\|^2 + \nu^\top (Ax - b) \big). \end{aligned}$$

Set gradient $= 0$:

$$\nabla_x L = 2x + A^\top \nu = 0 \quad \Rightarrow \quad x = -\frac{1}{2} A^\top \nu.$$

Substitute back:

$$L^*(\nu) = -\frac{1}{4} \nu^\top A A^\top \nu - \nu^\top b,$$

which is concave.

—

8.5 Weak and Strong Duality

Weak duality:

$$d^* \leq f^*.$$

Strong duality:

$$d^* = f^*,$$

holds under **Slater' s condition**:

$$f \text{ is convex}, \quad \exists x \in X \text{ s.t. } h_i(x) = 0,\ g_j(x) < 0.$$

—

8.6 Equality Constrained Problem

Lagrange Multiplier Theorem Consider

$$\min f(x) \quad \text{s.t. } Ax = b.$$

Lagrangian:

$$L(x, \lambda) = f(x) + \lambda^\top (Ax - b).$$

If x^* is a local minimum and a regular point (i.e. $\{\nabla h_i(x^*)\}$ linearly independent), then there exist unique scalars $\lambda_1^*, \ldots, \lambda_m^*$ such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0.$$

This characterizes a set of necessary conditions for a local minimum.

—

解释 I： At a local optimal solution, the gradient of f can be expressed as a linear combination of the gradients of the constraints:

$$\nabla f(x^*) = -\sum_i \lambda_i^* \nabla h_i(x^*).$$

几何意义：$\nabla f(x^*)$ 位于约束可行方向的法向空间上。

解释 II： The cost gradient $\nabla f(x^*)$ is orthogonal to the subspace of first-order feasible variations:

$$\nabla f(x^*)^\top \Delta x = \sum_i \lambda_i^* \nabla h_i(x^*)^\top \Delta x = 0.$$

即每个可行微扰 Δx 都满足 $h_i(x^* + \Delta x) = 0$，因此 $\nabla f(x^*)$ 与可行集的切空间正交。

—

8.7 Sensitivity Analysis

Consider the linearly constrained problem:

$$\min_x f(x) \quad \text{s.t. } a^\top x = b.$$

If b changes to $b + \Delta b$, the minimum changes from x^* to $x^* + \Delta x$.

Because

$$b + \Delta b = a^\top (x^* + \Delta x) = a^\top x^* + a^\top \Delta x \Rightarrow \Delta b = a^\top \Delta x.$$

Thus:

$$\nabla f(x^*) = -\lambda^* a.$$

Then:

$$\begin{aligned}\Delta\text{cost} &= f(x^* + \Delta x) - f(x^*) \\ &= \nabla f(x^*)^\top \Delta x + O(\|\Delta x\|^2) \\ &= -\lambda^* a^\top \Delta x \\ &= -\lambda^* \Delta b + O(\|\Delta x\|^2).\end{aligned}$$

Hence:

$$\lambda^* = -\frac{\Delta\text{cost}}{\Delta b},$$

表示约束右端 b 的微小变化对最优目标值的线性影响（敏感度）。其中当约束右端 b 增大时，可行域放宽，最优目标值（如成本）下降。

9 Lecture 8: Inequality Constraints and Duality

Main Topics

- How to use Lagrangian on inequality–constrained problems
- KKT condition and when it is sufficient
- Use KKT to decide optimal solutions
- General sufficiency condition
- The dual problem and duality theorem

—

9.1 Inequality–Constrained Problem

$$\begin{aligned}&\min_x f(x) \\ &\text{s.t. } h(x) = 0, \quad g(x) \le 0,\end{aligned}$$

where

$$h = (h_1, \ldots, h_m), \quad g = (g_1, \ldots, g_r).$$

Consider the set of active inequality constraints:

$$A(x) = \{j \mid g_j(x) = 0\}.$$

If x^* is a local minimum, then the active inequality constraints at x^* can be treated as equalities, and the inactive ones at x^* do not matter.

Thus, there exist multipliers λ_i^*, μ_j^* such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*) = 0,$$

where $\mu_j^* = 0$ for $j \notin A(x^*)$.

Extra property: $\mu_j^* \geq 0$ for all j.

Intuitive reason: Relax the j-th constraint $g_j(x) \leq q_j$, and notice

$$\mu_j^* = -\frac{\text{cost due to } g_j}{q_j}.$$

—

9.2 KKT Conditions

Let x^* be a local minimum and a regular point. Then there exist unique Lagrange multipliers

$$\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*), \quad \mu^* = (\mu_1^*, \ldots, \mu_r^*)$$

such that

$$\begin{cases} \nabla_x L(x^*, \lambda^*, \mu^*) = 0, \\ \mu_j^* \geq 0, \quad j = 1, \ldots, r, \\ \mu_j^* = 0, \ j \notin A(x^*). \end{cases} \tag{1}$$

The condition $\mu_j^* g_j(x^*) = 0$ can be compactly written as

$$\mu_j^* g_j(x^*) = 0, \quad \forall j = 1, \ldots, r, \text{(Complementarity Condition)} \tag{2}$$

and also

$$g_i(x^*) \leq 0, \quad h_j(x^*) = 0, \quad \forall i, j. \tag{3}$$

Equations (1)–(3) together are called the **KKT conditions**.

必要性: 当 x^* 为局部最优且满足 regularity 条件时，必存在一组 λ^*, μ^* 使 KKT 成立。

充分性: 若 $f(x)$ 是凸函数，$g_i(x)$ 为凸函数，$h_j(x)$ 为仿射函数，则 KKT 条件亦为充分条件。

—

9.3 How to Use KKT Conditions to Determine Optimal Solutions

Consider the general constrained optimization problem:

$$\min_x f(x) \quad \text{s.t.} \quad g_i(x) \le 0,\ i = 1, \dots, m, \quad h_j(x) = 0,\ j = 1, \dots, p.$$

The **Karush–Kuhn–Tucker (KKT) conditions** provide a set of necessary conditions for a local optimum x^*, assuming regularity (constraint qualification) holds:

$$\begin{cases} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) = 0, & \text{(Stationarity)} \\ g_i(x^*) \le 0, \quad h_j(x^*) = 0, & \text{(Primal Feasibility)} \\ \lambda_i^* \ge 0, & \text{(Dual Feasibility)} \\ \lambda_i^* \, g_i(x^*) = 0, & \text{(Complementary Slackness)} \end{cases}$$

KKT 条件构成了一组方程与不等式的系统。要找到满足条件的解 (x^*, λ^*, μ^*)，关键在于系统地分析其中的 **互补松弛条件**：

$$\lambda_i^* \, g_i(x^*) = 0, \quad i = 1, \dots, m.$$

这一条件意味着每个不等式约束只能有两种可能情况：约束要么“紧”（即达到边界），要么“松”（不活跃）。因此，通常需要对每个约束进行分类讨论。

逐约束情形分析 对于每个不等式约束 $i \in \{1, \dots, m\}$，我们需要考察以下两种互斥的情况：

- **情形 A（约束非活跃）**：假设 $\lambda_i^* = 0$。此时第 i 个约束并未起作用。在这种假设下，求解其余 KKT 方程组，并验证所得解是否满足

  $$g_i(x^*) < 0,$$

  即满足原始可行性（Primal Feasibility）。
- **情形 B（约束活跃）**：假设 $g_i(x^*) = 0$。此时第 i 个约束在最优点上“紧绑定”，相当于一个新的等式约束。在此假设下求解 KKT 方程组，并检查相应的乘子是否满足

  $$\lambda_i^* \ge 0,$$

  即对偶可行性（Dual Feasibility）。

对 m 个不等式约束而言，理论上共有 2^m 种可能的活跃/非活跃组合。在实际求解中，很多组合由于不满足可行性条件会被迅速排除。凡是能同时满足所有 KKT 条件的 (x^*, λ^*, μ^*) 都称为 **KKT 点（KKT point）**，是潜在的最优解候选。

计算实践中的说明 上述“逐案分析”的方法在低维问题或教学演示中十分有效；但在高维或大规模优化中，显然无法穷举全部 2^m 种情况。此时通常采用数值算法（如**主动集法（Active-Set Method）**或 **内点法（Interior-Point Method）**）自动识别活跃约束集合。

最终判定（根据凸性） 在找到全部 KKT 点之后，如何判定它们是否为最优解，取决于问题的性质：

- **若问题为凸优化问题**：即 $f(x)$ 及所有 $g_i(x)$ 均为凸函数，$h_j(x)$ 为仿射函数，且满足约束资格条件（如 **Slater 条件**），则 KKT 条件是**充要条件**。任意满足 KKT 条件的点即为**全局最优解**。
- **若问题为非凸优化问题**：此时 KKT 条件仅是**必要条件**。所得到的 KKT 点可能对应局部极小点、鞍点或局部极大点。因此需要在所有 KKT 点处计算目标函数 $f(x)$ 的值，并可进一步通过**二阶充分条件**验证其是否为局部最小点：

$$d^\top \nabla^2_{xx} L(x^*, \lambda^*, \mu^*)\, d > 0, \quad \forall d \in \mathcal{T}(x^*),$$

其中 $\mathcal{T}(x^*)$ 表示活跃约束的切空间。

综上，利用互补松弛条件进行分类讨论，是理解 KKT 条件结构和手工求解低维优化问题的关键步骤；而在数值优化中，求解器会自动识别这些活跃约束，从而等价地实现同样的逻辑。

—

9.4 General Sufficiency Condition

Consider the problem

$$\min f(x) \quad \text{s.t. } x \in X,\ g_i(x) \le 0,\ i = 1, \dots, r.$$

Let (x^*, μ^*) be feasible and satisfy

$$\mu_j^* \ge 0, \quad j = 1, \dots, r, \quad \nabla f(x^*) + \sum_j \mu_j^* \nabla g_j(x^*) = 0.$$

Then x^* is the **global minimum** of the problem.

Proof sketch:

$$f(x^*) = f(x^*) + \mu^{*\top} g(x^*) = \min_{x \in X} \big\{ f(x) + \mu^{*\top} g(x) \big\} \le f(x), \quad \forall x \in X.$$

—

9.5 Dual Problem

Define the dual function:

$$q(\mu) = \inf_{x \in X} \Big\{ f(x) + \sum_{j=1}^{r} \mu_j (a_j^\top x - b_j) \Big\}, \qquad q : \mathbb{R}^r \to [-\infty, \infty).$$

Basically this corresponds to minimizing the Lagrangian function. Note: introduce multipliers only for a subset of constraints.

Define the dual problem:

$$\max_{\mu \geq 0} q(\mu).$$

The effective constraint set of the dual is

$$Q = \{\mu \mid \mu \geq 0,\ q(\mu) > -\infty\}.$$

10 Lecture 9: Duality –Examples

Topics

- Dual problem for linear programs (from profit maximization to cost minimization)
- Solve SVM step by step

—

10.1 Primal–Dual Linear Program (LP)

Primal Problem (maximize profit):

$$\max_{x_1, x_2, x_3} \quad C_1 x_1 + C_2 x_2 + C_3 x_3$$

$$\text{s.t.} \begin{cases} a_{11} x_1 + a_{12} x_2 + a_{13} x_3 \leq b_1, & \text{(raw material 1 limited)} \\ a_{21} x_1 + a_{22} x_2 + a_{23} x_3 \leq b_2, & \text{(raw material 2 limited)} \\ x_1, x_2, x_3 \geq 0. \end{cases}$$

Here:

- C_j: profit per unit of product j produced,
- b_i: total amount of raw material i,
- a_{ij}: units of raw material i needed to produce one unit of product j.

—

Compact form:

$$\max_x c^\top x \quad \text{s.t. } Ax \le b,\ x \ge 0.$$

Dual form (cost minimization):

$$\min_y b^\top y \quad \text{s.t. } A^\top y \ge c,\ y \ge 0.$$

此处的 y 即为原问题中不等式约束 $Ax \le b$ 对应的 **拉格朗日乘子 (Lagrange multipliers)**, 在经济意义上可理解为每种原料的**影子价格 (shadow price)** 或**边际价值 (marginal value)**。

—

Economic interpretation: Suppose there is a buyer who wants to purchase all the raw materials (b_1 and b_2 units) from the firm.

Let y_1, y_2 denote the prices the buyer is willing to pay for the two materials.

当买家愿意支付的价格 y_1, y_2 超过转化为产品的利润价格 C_i 时，买家会亏损，因此需满足如下优化问题：

$$\min_{y_1,y_2}\ b_1y_1 + b_2y_2$$

$$\text{s.t. } \begin{cases} a_{11}y_1 + a_{21}y_2 \ge C_1, \\ a_{12}y_1 + a_{22}y_2 \ge C_2, \\ a_{13}y_1 + a_{23}y_2 \ge C_3, \\ y_1, y_2 \ge 0. \end{cases}$$

Example: Primal–Dual Linear Programming

考虑一个简单的生产计划问题。某工厂生产两种产品 x_1, x_2，其单位利润分别为 $C_1 = 3$、$C_2 = 2$。生产每种产品都需要两种原材料，其资源消耗如下表所示：

原料	产品 1 所需	产品 2 所需	资源上限
原料 1	1	2	8
原料 2	1	1	6

Primal Problem (利润最大化)

$$\begin{aligned} \max_{x_1,x_2} \quad & 3x_1 + 2x_2, \\ \text{s.t.} \quad & x_1 + 2x_2 \le 8, \\ & x_1 + x_2 \le 6, \\ & x_1, x_2 \ge 0. \end{aligned}$$

解： 由约束方程

$$\begin{cases} x_1 + 2x_2 = 8, \\ x_1 + x_2 = 6, \end{cases} \Rightarrow x_1 = 4,\ x_2 = 2.$$

计算目标函数:

$$f(4,2) = 3(4) + 2(2) = 16.$$

再比较边界点:

$$(6,0) : f = 18, \quad (0,4) : f = 8,$$

可得最优解为

$$x_1^* = 6, \quad x_2^* = 0, \quad f^* = 18.$$

Dual Problem (成本最小化) 根据标准形式:

$$\max_x \ c^\top x \quad \text{s.t. } Ax \le b,\ x \ge 0$$

的对偶为

$$\min_y \ b^\top y \quad \text{s.t. } A^\top y \ge c,\ y \ge 0.$$

代入

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 8 \\ 6 \end{bmatrix}, \quad c = \begin{bmatrix} 3 \\ 2 \end{bmatrix},$$

得

$$\begin{aligned} \min_{y_1,y_2} \quad & 8y_1 + 6y_2, \\ \text{s.t.} \quad & y_1 + y_2 \ge 3, \\ & 2y_1 + y_2 \ge 2, \\ & y_1, y_2 \ge 0. \end{aligned}$$

求解： 由第一个约束 $y_2 = 3 - y_1$，代入目标函数：

$$8y_1 + 6y_2 = 8y_1 + 6(3 - y_1) = 2y_1 + 18.$$

为最小化该式，取 $y_1 = 0$，得

$$y_1^* = 0, \quad y_2^* = 3, \quad b^\top y^* = 8(0) + 6(3) = 18.$$

结果验证：

$$f^*_{\text{primal}} = 18, \qquad f^*_{\text{dual}} = 18.$$

因此满足**强对偶性（Strong Duality）**，即

$$\boxed{f^*_{\text{primal}} = f^*_{\text{dual}}.}$$

经济学解释： y_1, y_2 分别代表原料 1 与原料 2 的**影子价格（shadow prices）**或**资源价值**。其中 $y_2^* = 3$ 表示：若额外增加 1 单位原料 2 的供给量，最多可使总利润增加 3 个单位。

—

10.2 Apply KKT to SVM –Separable Case

Primal:

$$\min_x \frac{1}{2}\|x\|^2 \quad \text{s.t. } b_i(a_i^\top x) \geq 1, \quad \forall i.$$

To simplify, absorb bias x_0 into x by augmenting $a_i \leftarrow [a_i, 1]$ and $x \leftarrow [x, x_0]$.

Lagrangian:

$$L(x, \lambda) = \frac{1}{2}\|x\|^2 - \sum_i \lambda_i [b_i(a_i^\top x) - 1],$$

where $\lambda_i \geq 0$.

Set gradient w.r.t. x to zero:

$$\nabla_x L = 0 \quad \Rightarrow \quad x = \sum_i \lambda_i b_i a_i.$$

Substitute back to obtain dual:

$$L^*(\lambda) = \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j b_i b_j a_i^\top a_j,$$

$$\text{s.t. } \lambda_i \geq 0, \forall i.$$

Dual problem:

$$\max_{\lambda} \quad \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j b_i b_j a_i^\top a_j \quad \text{s.t. } \lambda_i \geq 0.$$

After solving λ:

$$x = \sum_i \lambda_i b_i a_i.$$

If $\lambda_i > 0$, then a_i is a **support vector**. If $\lambda_i = 0$, then a_i is not a support vector.

Hence, the optimal model is a nonnegative combination of data points.

—

10.3 SVM –Non-Separable Case

Primal:

$$\begin{aligned} \min_{x,\xi} \quad & \frac{1}{2}\|x\|^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & b_i(a_i^\top x) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ \forall i. \end{aligned}$$

Lagrangian:

$$L(x, \lambda, \mu, \xi) = \frac{1}{2}\|x\|^2 + C\sum_i \xi_i - \sum_i \lambda_i [b_i(a_i^\top x) - 1 + \xi_i] - \sum_i \mu_i \xi_i.$$

Set gradient w.r.t. x:

$$\nabla_x L = 0 \Rightarrow x = \sum_i \lambda_i b_i a_i.$$

Set gradient w.r.t. ξ_i:

$$\nabla_{\xi_i} L = 0 \Rightarrow C - \lambda_i - \mu_i = 0.$$

Substitute to obtain the dual:

$$\begin{aligned} L^*(\lambda) = \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j b_i b_j a_i^\top a_j, \\ \text{s.t. } 0 \leq \lambda_i \leq C. \end{aligned}$$

If $0 < \lambda_i < C$, then a_i is a support vector. If $\lambda_i = 0$ or $\lambda_i = C$, it is not a support vector.

Primal feasibility gives

$$b_i(a_i^\top x) - 1 + \xi_i = 0 \Rightarrow \xi_i = 1 - b_i(a_i^\top x).$$

This tells us the **training error**.

—

10.4 Solving the SVM Dual Problem –Coordinate Descent (CCD)

For SVM' s dual, a popular method is the **Coordinate Descent** approach.

Fix all λ_j except one λ_i, and update λ_i at a time. Each time we are effectively processing one data point.

Gradient w.r.t. λ_i:

$$\nabla_{\lambda_i} L^*(\lambda) = 1 - b_i a_i^\top \left(\sum_j \lambda_j b_j a_j \right) = 1 - b_i a_i^\top x.$$

Gradient projection update:

$$\lambda_i^+ = \text{proj}_{[0,C]} \left[\lambda_i - \frac{1}{H_{ii}} \nabla_{\lambda_i} L^*(\lambda) \right],$$

where $H_{ii} = a_i^\top a_i$ (Hessian diagonal).

Projection means:

$$\text{proj}_{[0,C]}(x) = \max\{\min\{x, C\}, 0\}.$$

这就是 SVM Dual 的常见优化算法（coordinate descent）。

Appendix A: Worked Examples of Gradient Descent and Projection

A.1 Gradient Descent and Variants

Consider the quadratic least-squares problem:

$$\min_{x \in \mathbb{R}^2} f(x) = \frac{1}{2} \|Ax - b\|^2, \quad A = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}.$$

(1) Gradient Descent (GD)

$$\nabla f(x) = A^\top (Ax - b), \quad x^{r+1} = x^r - \alpha \nabla f(x^r).$$

Let $x^0 = (0, 0)^\top,\ \alpha = 0.05$.

$$g^0 = A^\top (Ax^0 - b) = (-11, -5)^\top, \quad x^1 = x^0 - \alpha g^0 = (0.55, 0.25).$$

(2) Incremental Gradient Descent (Incre. GD) Each sample loss:

$$g_i(x) = \tfrac{1}{2}(a_i^\top x - b_i)^2, \quad \nabla g_i(x) = (a_i^\top x - b_i)a_i.$$

Let $\psi_0 = x^0 = (0,0)^\top,\ \alpha = 0.05$:

$$\psi_1 = \psi_0 - \alpha\nabla g_1(\psi_0) = (0,0) - 0.05(-1)\begin{bmatrix}1\\1\end{bmatrix} = (0.05, 0.05),$$

$$\psi_2 = \psi_1 - \alpha\nabla g_2(\psi_1) = (0.05,0.05) - 0.05((2,1)^\top(0.05,0.05) - 2)(2,1) = (0.235, 0.1425),$$

$$\psi_3 = \psi_2 - \alpha\nabla g_3(\psi_2) = (0.235,0.1425) - 0.05((3,1)^\top(0.235,0.1425) - 2)(3,1) = (0.407875, 0.200125).$$

Hence $x^1 = \psi_3 = (0.407875, 0.200125)$.

(3) Coordinate Descent (Version I: Exact minimization). 记两列为 $A_1 = [1,2,3]^\top,\ A_2 = [1,1,1]^\top$。Version I 在第 t 次迭代中，按选定坐标 i 做精确一维最小化：

$$x_i^{\text{new}} = \arg\min_{z\in\mathbb{R}} \ \tfrac{1}{2}\Big\|A_i z + \sum_{j\neq i} A_j x_j^{(\text{latest})} - b\Big\|^2 = \frac{A_i^\top\big(b - \sum_{j\neq i} A_j x_j^{(\text{latest})}\big)}{A_i^\top A_i}.$$

Step 1（更新 x_1）. 此时 $x_2 = 0$，故

$$x_1^{(1)} = \frac{A_1^\top b}{A_1^\top A_1} = \frac{1\cdot 1 + 2\cdot 2 + 3\cdot 2}{1^2+2^2+3^2} = \frac{11}{14}.$$

Step 2（更新 x_2）. 采用 Gauss–Seidel 方式，用最新的 $x_1^{(1)}$：

$$r := b - A_1 x_1^{(1)} = \begin{bmatrix}1\\2\\2\end{bmatrix} - \frac{11}{14}\begin{bmatrix}1\\2\\3\end{bmatrix} = \begin{bmatrix}\frac{3}{14}\\ \frac{3}{7}\\ -\frac{5}{14}\end{bmatrix}, \qquad x_2^{(1)} = \frac{A_2^\top r}{A_2^\top A_2} = \frac{\frac{3}{14}+\frac{3}{7}-\frac{5}{14}}{3} = \frac{2}{21}.$$

因此，一次循环后的迭代点为

$$\boxed{x^{(1)} = \begin{bmatrix}\frac{11}{14}\\ \frac{2}{21}\end{bmatrix} \approx \begin{bmatrix}0.785714\\ 0.095238\end{bmatrix}.}$$

—

A.2 Projection Examples

(1) Box constraint projection

$$\min \tfrac{1}{2}x^2 \quad \text{s.t. } a \le x \le b.$$

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

$$x^* = \Pi_{[a,b]}(0) = \min\{\max\{0, a\}, b\} = \begin{cases} a, & 0 < a, \\ 0, & a \le 0 \le b, \\ b, & b < 0. \end{cases}$$

Projection operator:

$$\Pi_{[a,b]}(x) = \arg\min_{y \in [a,b]} \|y - x\|.$$

(2) Nonnegative constraint projection

$$\min \tfrac{1}{2}\|x - y\|^2 \quad \text{s.t. } x \ge 0.$$

Solution:

$$x_i^* = \begin{cases} y_i, & y_i \ge 0, \\ 0, & y_i < 0, \end{cases} \quad \forall i.$$

Or compactly written as

$$x^* = [y]^+.$$

For example:

$$[-2, 3]^\top \;\mapsto\; [0, 3]^\top.$$