

# 学习中的 优化

---

harry

# 目录

<b>第一章 凸优化简介</b>	<b>9</b>
1.1 数学优化	9
1.2 线性规划与非线性规划	9
1.3 凸优化	9
1.4 最小二乘和线性规划	10
1.4.1 线性规划 (Linear Programming, LP)	10
1.4.2 最小二乘法 (Least Squares Method, LSM)	10
1.5 仿射集合 (Affine Sets)	11
1.5.1 仿射组合与仿射集合	11
1.5.2 仿射集合与线性子空间的关系	11
1.5.3 仿射包 (Affine Hull)	12
1.5.4 典型实例	12
1.6 凸集 (Convex Sets)	12
1.6.1 凸组合与凸集	12
1.6.2 与仿射集合的关系	13
1.6.3 重要概念: 凸包 (Convex Hull)	13
1.6.4 典型实例	13
1.7 锥 (Cones) 与凸锥 (Convex Cones)	13
1.7.1 核心定义: 锥与凸锥	13
1.7.2 重要概念: 凸锥包 (Convex Cone Hull)	14
1.7.3 典型实例	14
1.8 三类集合的核心对比	14
1.9 重要的例子	15
1.9.1 超平面 (Hyperplane) 与半空间 (Halfspace)	15
1.9.2 球 (Ball) 与椭球 (Ellipsoid)	16
1.9.3 范数球 (Norm Ball) 与范数锥 (Norm Cone)	16
1.9.4 多面体 (Polyhedron) 与单纯形 (Simplex)	18
<b>第二章 凸优化的基本性质</b>	<b>20</b>
2.1 凸优化问题的形式	20
2.1.1 一般优化模型	20
2.1.2 凸优化的定义	20
2.2 凸函数的定义以及性质	20
2.2.1 定义与几何意义	20

2.2.2	Jensen 不等式	21
2.2.3	一阶条件 (First-order condition)	21
2.2.4	二阶条件 (Hessian 判定)	22
2.3	全局与局部最优	23
2.3.1	定义回顾	23
2.3.2	凸优化的核心定理	23
2.3.3	强凸函数的唯一最优性	24
2.4	凸优化的几何意义	24
2.4.1	可行域与等高线	24
2.4.2	法向条件 (支撑超平面)	24
2.4.3	凸组合与最优性路径	25
2.5	凸优化的转化以及示例	25
2.5.1	优化问题的等价变换以及实例	25
2.5.2	非凸到凸的重构思路在实例中的延伸 (解决“非凸建模难题”)	27
2.5.3	转化逻辑的核心价值	28
<b>第三章</b>	<b>无约束优化问题</b>	<b>29</b>
3.1	无约束优化问题	29
3.1.1	问题基本形式	29
3.1.2	最优性条件	29
3.1.3	核心求解框架 (下山法迭代格式)	29
3.2	搜索方向的确定	30
3.2.1	视角 1: 线性化与下降条件	30
3.2.2	视角 2: 一般范数下的“最速下降”	30
3.2.3	预条件化的作用	31
3.3	如何确定步长	31
3.3.1	精确线搜索 (Exact Line Search)	32
3.3.2	黄金分割 (Golden Section Search)	32
3.3.3	回溯搜索 (Backtracking Line Search) 与 Armijo-Wolfe 准则	33
3.4	收敛率: 强凸 / PL 条件与“楼梯现象”	33
3.4.1	强凸 + L-光滑: 线性收敛	33
3.4.2	PL (Polyak-Lojasiewicz) 不等式	35
3.4.3	“楼梯现象”: 成因与缓解	36
3.4.4	实践提示	36
3.5	Newton 法: 局部二次近似与两阶段收敛	36
3.5.1	核心思路: 函数的局部二次近似	37
3.5.2	牛顿步 (Newton Step) 的推导	37
3.5.3	牛顿步的下降性	37
3.5.4	局部二次收敛: 牛顿法的核心优势	37
<b>第四章</b>	<b>随机梯度下降 (Stochastic Gradient Descent, SGD)</b>	<b>39</b>
4.1	随机梯度下降基础	39
4.1.1	核心思路: 用“部分数据”估算梯度	39

4.1.2	完整更新流程	39
4.1.3	关键超参数	39
4.2	一个随机估计问题	40
4.2.1	有限样本下的均值计算	40
4.2.2	前后均值的递推关系	40
4.2.3	均值收敛的条件	40
4.3	Robbins-Monro (RM) 算法	41
4.3.1	RM 算法的目标	41
4.3.2	RM 算法的迭代公式	41
4.3.3	RM 算法的收敛条件	41
4.4	SGD 之问：为何能够收敛？	41
4.4.1	设定与记号	42
4.4.2	收敛性证明的核心假设	42
4.4.3	收敛性结论	43
4.4.4	非强凸场景下的 SGD 收敛性（仅凸/一般非凸）	45
4.4.5	两种不同目标下的步长设计及收敛策略差异	46
4.5	从随机估计到动力学	47
4.5.1	从 GD 到 ODE：离散更新是梯度流的“显式欧拉积分”	47
4.5.2	从 SGD 到 SDE：扩散极限与朗之万动力学	48
4.5.3	局部二次近似与 OU 过程：常步长下的方差-曲率权衡	50
4.5.4	学习率、批量与“温度”的定量关系	51
4.5.5	训练策略：将动力学结论落地到实践	51
4.5.6	动力学近似的失效场景	52
4.5.7	核心关系总结	52
4.6	SGD 之问：为什么需要动量？	53
4.6.1	从 SGD 出发：我们到底缺什么？	53
4.6.2	Heavy-Ball (HB)：用“惯性”优化 SGD 的核心痛点	54
4.6.3	Nesterov (NAG)：“前瞻-校正”实现更稳健的加速	55
4.6.4	HB 与 NAG 的噪声鲁棒性对比	55
4.6.5	实践选择：何时用 HB，何时用 NAG？	55
4.6.6	核心总结	56
第五章	无约束优化之动量	57
5.1	符号说明	57
5.2	SGD 的缺点及动量方法改进	57
5.2.1	问题出发点：SGD 的局限	57
5.2.2	引入动量的核心思想：惯性	58
5.2.3	Heavy-Ball (Polyak Momentum)	58
5.2.4	Nesterov 加速梯度 (NAG, 1983)	58
5.2.5	对比总结	59
5.2.6	从 SGD 到动量方法的逻辑链	59
5.3	AdaGrad (Duchi et al., 2011)	59
5.3.1	动机：SGD 学习率“一刀切”的问题	59

5.3.2	算法公式	59
5.3.3	性质与效果	60
5.3.4	优缺点	60
5.3.5	本质理解：累积“几何尺度”的归一化	60
5.4	RMSProp	60
5.4.1	动机：修正 AdaGrad 的“学习率枯竭”	60
5.4.2	算法公式	61
5.4.3	性质与直觉	61
5.4.4	与 AdaGrad 的对比	61
5.4.5	本质理解	61
5.5	Adam (Adaptive Moment Estimation)	62
5.5.1	动机：融合动量与自适应学习率	62
5.5.2	算法核心公式	62
5.5.3	性质与优点	62
5.5.4	缺点与改进方向	63
5.5.5	本质理解	63
5.6	AdamW (Adam with Decoupled Weight)	63
5.6.1	动机：修正 Adam 正则化的逻辑错误	63
5.6.2	核心思想：权重衰减与梯度更新解耦	63
5.6.3	性质与效果	64
5.6.4	本质理解	64
第六章	阻尼牛顿法	65
6.1	牛顿法复习	65
6.2	阻尼牛顿法	65
6.2.1	纯牛顿法的局限性（阻尼牛顿法的必要性）	66
6.2.2	阻尼牛顿法的核心改进：牛顿方向 + 线搜索	66
6.2.3	线搜索：如何确定步长 $\alpha_k$ ?	66
6.2.4	阻尼牛顿法的优势	67
6.2.5	阻尼牛顿法的步骤总结	67
6.3	阻尼牛顿法性质	67
6.3.1	一些假设	67
6.3.2	全局收敛到临界点	68
6.3.3	局部收敛阶段的退化	69
6.4	非显式求逆方法	70
6.4.1	Cholesky 分解 (Cholesky Decomposition)	70
6.4.2	LDL 分解 (LDL Decomposition)	71
第七章	牛顿法和拟牛顿法	72
7.1	符号说明	72
7.2	牛顿法的严格数学建模（复习）	72
7.2.1	核心思路：函数的局部二次近似	73
7.2.2	牛顿步 (Newton Step) 的推导	73

7.2.3	牛顿步的下降性	73
7.2.4	局部二次收敛：牛顿法的核心优势	73
7.3	拟牛顿法	74
7.3.1	拟牛顿法的基本框架	74
7.3.2	BFGS 方法	74
7.3.3	准牛顿法的其他变种	79
7.4	收敛性	79
7.4.1	BFGS 的收敛性	79
7.4.2	BFGS 的收敛速度：介于梯度下降与牛顿法之间，逼近牛顿法	80
7.4.3	BFGS 的最优性：理论性质与实际性能的“最优平衡”	80
7.5	伪代码	81
7.6	L-BFGS（有限记忆二阶近似）	84
7.6.1	推导起点：BFGS 的 G-form 递推关系	84
7.6.2	Step 1: 递推展开 $\mathbf{G}_k \mathbf{v}$	84
7.6.3	Step 2: 反向环（Right Loop）——处理右乘链	85
7.6.4	Step 3: 初始缩放（Initial Scaling）—— $\mathbf{G}_0$ 的作用	85
7.6.5	Step 4: 正向环（Left Loop）——处理左乘链与修正项	85
7.6.6	Step 5: L-BFGS 搜索方向与迭代流程	86
7.6.7	关键性质验证	86
第八章	优化算法的评价	88
8.1	预备知识与符号	88
8.2	基本工具：下降引理（Descent Lemma）	88
8.3	基于 Descent Lemma 的四类典型收敛结果	89
8.3.1	非凸 + L-光滑	89
8.3.2	凸 + L-光滑	90
8.3.3	$\mu$ -强凸 + L-光滑：线性（几何）收敛	91
8.3.4	PL 条件 + L-光滑：线性收敛（无需凸）	92
第九章	从无约束到等式约束：拉格朗日乘子	94
9.1	拉格朗日乘子数学建模	94
9.1.1	1. 无约束优化的数学模型	94
9.1.2	2. 等式约束优化的数学模型	94
9.1.3	3. 拉格朗日函数的建模思想	94
9.1.4	4. 拉格朗日函数的最优性条件（一阶 KKT 条件）	94
9.1.5	5. 几何意义（直观理解）	95
9.2	拉格朗日乘子法的必要条件	95
9.2.1	一阶必要条件及其证明	95
9.2.2	二阶最优性条件及其证明	96
第十章	从等式约束到不等式约束：KKT	99
10.1	为什么需要 KKT	99
10.2	不等式约束建模	99
10.2.1	等式约束优化问题（基础模型）	99

10.2.2 含不等式约束的优化问题（扩展模型）	99
10.3 KKT 条件的推导	100
10.3.1 一阶必要条件（KKT 条件）的严格数学建模	100
10.3.2 约束规格（CQ）的严格数学建模	100
10.3.3 LICQ 下 KKT 条件的严格证明建模	101
10.3.4 Farkas 引理证明	102
10.4 二阶最优性条件	103
10.4.1 临界锥	103
10.4.2 二阶必要条件	104
10.4.3 二阶充分条件	104
<b>第十一章 从等式约束到不等式约束：对偶</b>	<b>105</b>
11.1 从 KKT 条件到对偶问题的严格数学建模	105
11.1.1 原始问题（Primal Problem）建模	105
11.1.2 拉格朗日函数建模	105
11.1.3 拉格朗日对偶函数建模	105
11.1.4 对偶函数的核心性质建模	106
11.1.5 拉格朗日对偶问题建模	107
11.2 强对偶	107
11.2.1 Slater 条件	107
11.2.2 几何解释	108
11.2.3 强对偶定理证明	108
11.3 互补松弛条件	109
11.3.1 KKT 条件的完整构成（含互补松弛）	110
11.4 对偶理论的应用实例	111
11.4.1 线性规划的对偶	111
11.4.2 二次规划的对偶	112
11.4.3 数值例题（二次规划对偶求解）	112
11.4.4 SVM 中的原问题与对偶问题	114
<b>第十二章 约束优化问题解法</b>	<b>116</b>
12.1 外点法	116
12.1.1 核心定义与问题形式	116
12.1.2 数学建模：惩罚函数构造	116
12.1.3 算法流程（严格形式化）	117
12.1.4 收敛性分析（核心结论）	118
12.1.5 优缺点	119
12.1.6 示例：外点法求解	122
12.1.7 外点法的极限满足 KKT	123
12.2 增广拉格朗日（ALM）	124
12.2.1 从外点法到增广拉格朗日法（ALM）	124
12.2.2 拉格朗日函数的基础铺垫	125
12.2.3 增广拉格朗日函数的构造	125

12.2.4 乘子更新规则的推导（基于 KKT 条件）	125
12.2.5 ALM 算法流程	126
12.2.6 ALM 与外点法的核心差异（严谨对比）	127
12.2.7 收敛性核心结论	127
12.2.8 示例：ALM 应用	127
12.3 约束问题的解法之 ADMM	129
12.3.1 第一步：先锁定 ALM 的“可分问题特例”（ADMM 的适用场景）	129
12.3.2 第二步：写出这个特例的 ALM 增广拉格朗日函数	129
12.3.3 第三步：ALM 对这个问题的迭代步骤	129
12.3.4 第四步：关键改进——利用“目标可分”拆分最小化步骤	129
12.3.5 第五步：简化符号——引入 ADMM 的“对偶残差 $\mathbf{u}$ ”	130
12.3.6 最终：从 ALM 拆分得到的 ADMM 迭代（一一对应）	130
12.3.7 示例（与前述问题关联的改写）	130

# 第一章 凸优化简介

## 1.1 数学优化

数学优化问题可以写为如下形式：

$$\begin{aligned} & \text{minimize} && f_0(x) \\ & \text{subject to} && f_i(x) \leq b_i, \quad i = 1, 2, \dots, m \end{aligned} \tag{1.1.0.1}$$

向量  $x = (x_1, x_2, \dots, x_n)$  称为问题的**优化变量**，函数  $f_0$  为目标函数，函数  $f_i$  为约束函数，常数  $b_i$  为约束上限或者约束边界。

最优解  $x^*$  是满足所有约束条件并使目标函数达到最小值的变量取值，即：

$$x^* = \arg \min_{x \in \mathbb{R}^n} f_0(x) \quad \text{满足} \quad f_i(x) \leq b_i, \quad i = 1, 2, \dots, m \tag{1.1.0.2}$$

更具体地说，最优解  $x^*$  满足以下两个条件：

### 1. 可行性 (Feasibility):

$$f_i(x^*) \leq b_i, \quad \forall i = 1, 2, \dots, m \tag{1.1.0.3}$$

即最优解必须满足所有的约束条件。

### 2. 最优性 (Optimality):

$$f_0(x^*) \leq f_0(x), \quad \forall x \in \mathbb{R}^n \text{ 且满足 } f_i(x) \leq b_i, \quad i = 1, 2, \dots, m \tag{1.1.0.4}$$

即最优解在所有可行解中使目标函数达到最小值。

如果存在多个满足上述条件的解，则它们都称为最优解，且此时目标函数的最优值是唯一的。

## 1.2 线性规划与非线性规划

线性规划的核心特征为：目标函数和约束函数均为线性函数。

**定义 1.1 (线性函数)**. 线性函数的定义为：对于任意  $x, y \in \mathbb{R}^n$  和  $\alpha, \beta \in \mathbb{R}$ ，均满足：

$$f_i(\alpha x + \beta y) = \alpha f_i(x) + \beta f_i(y) \tag{1.2.0.1}$$

## 1.3 凸优化

凸优化的核心特征为：目标函数和约束函数均为凸函数。

**定义 1.2 (凸函数).** 凸函数的定义为：对于任意  $x, y \in \mathbb{R}^n$  和  $\alpha, \beta \in \mathbb{R}$ ，且满足  $\alpha + \beta = 1$ 、 $\alpha \geq 0$ 、 $\beta \geq 0$ ，均有：

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y) \quad (1.3.0.1)$$

## 1.4 最小二乘和线性规划

广为人知而且应用广泛的两类凸优化问题：最小二乘和线性规划

### 1.4.1 线性规划 (Linear Programming, LP)

线性规划是一类目标函数与约束函数均为线性的优化问题，是最经典的确定性优化模型之一。

一个函数  $f_i: \mathbb{R}^n \rightarrow \mathbb{R}$  为线性函数，当且仅当对任意  $x, y \in \mathbb{R}^n$  和任意标量  $\alpha, \beta \in \mathbb{R}$ ，满足线性性（齐次性 + 叠加性）：

$$f_i(\alpha x + \beta y) = \alpha f_i(x) + \beta f_i(y) \quad (1.4.1.1)$$

其具体形式可表示为  $f_i(x) = c_i^T x$ ，其中  $c_i \in \mathbb{R}^n$  为常数向量， $x \in \mathbb{R}^n$  为决策变量。

**定义 1.3 (线性规划).** 线性规划的目标是在 *linear* 约束下优化 *linear* 目标函数，标准形式（以最小化为例）为：

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) = c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \quad (1.4.1.2)$$

其中：

- $c \in \mathbb{R}^n$ ：目标函数系数向量；
- $A \in \mathbb{R}^{m \times n}$ ：约束系数矩阵（ $m < n$ ）；
- $b \in \mathbb{R}^m$ ：约束右端项向量；
- $x \geq 0$ ：决策变量非负约束。

### 1.4.2 最小二乘法 (Least Squares Method, LSM)

对于给定的观测数据  $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ ，假设数据满足某种函数关系  $y = f(x; \theta)$ （ $\theta$  为待估参数），最小二乘法的目标是寻找参数  $\theta^*$ ，使得误差平方和最小：

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^m [y_i - f(x_i; \theta)]^2 \quad (1.4.2.1)$$

当待估参数  $\theta$  与函数  $f(x; \theta)$  呈线性关系时，称为线性最小二乘。其典型形式为：已知线性模型  $y = A\theta + \epsilon$ （ $\epsilon$  为误差项），目标函数为：

$$\min_{\theta \in \mathbb{R}^n} f(\theta) = \|A\theta - y\|_2^2 = (A\theta - y)^T (A\theta - y) \quad (1.4.2.2)$$

其中：

- $A \in \mathbb{R}^{m \times n}$ : 设计矩阵 ( $m > n$ , 保证超定系统);
- $y \in \mathbb{R}^m$ : 观测值向量;
- $\theta \in \mathbb{R}^n$ : 待估参数向量。

该目标函数是关于  $\theta$  的二次凸函数, 无约束条件 (或仅含线性约束)。

当然最小二乘还有一些拓展, 如加权最小二乘和正则化最小二乘, 在此不展开。

## 1.5 仿射集合 (Affine Sets)

仿射集合是线性空间中“平移后的线性子空间”, 其核心特征是对仿射组合的封闭性。

### 1.5.1 仿射组合与仿射集合

#### (1) 仿射组合 (Affine Combination)

**定义 1.4.** 设  $x_1, x_2, \dots, x_k \in \mathbb{R}^n$ , 若标量  $\theta_1, \theta_2, \dots, \theta_k \in \mathbb{R}$  满足  $\sum_{i=1}^k \theta_i = 1$ , 则称向量:

$$\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k \quad (1.5.1.1)$$

为  $x_1, x_2, \dots, x_k$  的仿射组合。

特别地, 当  $k = 2$  时, 仿射组合为  $\theta x + (1 - \theta)y$  ( $\theta \in \mathbb{R}$ ), 其几何意义是过点  $x$  和  $y$  的整条直线 (区别于后续凸组合对应的“线段”)。

#### (2) 仿射集合的定义

**定义 1.5.** 一个集合  $A \subseteq \mathbb{R}^n$  被称为仿射集合, 当且仅当对任意  $x, y \in A$  及任意  $\theta \in \mathbb{R}$ ,  $x$  与  $y$  的仿射组合仍属于  $A$ , 即:

$$\theta x + (1 - \theta)y \in A \quad (1.5.1.2)$$

推广到  $k$  个点: 若  $A$  是仿射集合, 则对任意  $x_1, \dots, x_k \in A$  及任意满足  $\sum_{i=1}^k \theta_i = 1$  的  $\theta_1, \dots, \theta_k \in \mathbb{R}$ , 有  $\sum_{i=1}^k \theta_i x_i \in A$  (可通过数学归纳法证明)。

### 1.5.2 仿射集合与线性子空间的关系

仿射集合可通过“线性子空间的平移”来等价描述, 这是理解仿射集合的关键视角。

#### (1) 平移与线性子空间

设  $A \subseteq \mathbb{R}^n$  是仿射集合, 任取  $x_0 \in A$ , 定义集合:

$$L = A - x_0 = \{x - x_0 \mid x \in A\} \quad (1.5.2.1)$$

则  $L$  是  $\mathbb{R}^n$  的线性子空间 (满足对线性组合封闭:  $\forall u, v \in L, \alpha, \beta \in \mathbb{R}, \alpha u + \beta v \in L$ )。证明略

## 1.5.3 仿射包 (Affine Hull)

**定义 1.6.** 对任意集合  $S \subseteq \mathbb{R}^n$ , 包含  $S$  的最小仿射集合称为  $S$  的仿射包, 记为  $\text{aff}(S)$ 。其数学表达式为:

$$\text{aff}(S) = \left\{ \sum_{i=1}^k \theta_i x_i \mid x_1, \dots, x_k \in S, \theta_1, \dots, \theta_k \in \mathbb{R}, \sum_{i=1}^k \theta_i = 1 \right\} \quad (1.5.3.1)$$

直观理解: 仿射包是“由  $S$  中所有点的仿射组合构成的集合”, 例如  $\mathbb{R}^2$  中两个点的仿射包是过这两点的直线, 三个不共线点的仿射包是整个  $\mathbb{R}^2$ 。

## 1.5.4 典型实例

- 线性方程组的解空间: 设  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ , 则  $Ax = b$  的解集合  $X = \{x \in \mathbb{R}^n \mid Ax = b\}$  是仿射集合 (若  $X$  非空)。

证明.

任取  $x, y \in X$ , 则  $A(\theta x + (1 - \theta)y) = \theta Ax + (1 - \theta)Ay = \theta b + (1 - \theta)b = b$ , 故  $\theta x + (1 - \theta)y \in X$ 。

- 单点集  $\{x_0\}$ : 是仿射集合 (仅含自身, 仿射组合仍为自身)。
- 整个空间  $\mathbb{R}^n$ : 是仿射集合 (线性子空间本身, 平移量为 0)。

## 1.6 凸集 (Convex Sets)

凸集是凸优化的核心结构, 其特征是对凸组合的封闭性, 这直接保证了“局部最优即全局最优”的关键性质。

## 1.6.1 凸组合与凸集

## (1) 凸组合 (Convex Combination)

**定义 1.7.** 设  $x_1, x_2, \dots, x_k \in \mathbb{R}^n$ , 若标量  $\theta_1, \theta_2, \dots, \theta_k \in \mathbb{R}$  满足  $\sum_{i=1}^k \theta_i = 1$  且  $\theta_i \geq 0$  ( $i = 1, \dots, k$ ), 则称向量:

$$\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k \quad (1.6.1.1)$$

为  $x_1, x_2, \dots, x_k$  的凸组合。

特别地, 当  $k = 2$  时, 凸组合为  $\theta x + (1 - \theta)y$  ( $\theta \in [0, 1]$ ), 其几何意义是连接  $x$  和  $y$  的线段 (区别于仿射组合的“直线”)。

## (2) 凸集的定义

**定义 1.8.** 一个集合  $C \subseteq \mathbb{R}^n$  被称为凸集, 当且仅当对任意  $x, y \in C$  及任意  $\theta \in [0, 1]$ ,  $x$  与  $y$

的凸组合仍属于  $C$ ，即：

$$\theta x + (1 - \theta)y \in C \quad (1.6.1.2)$$

推广到  $k$  个点：若  $C$  是凸集，则对任意  $x_1, \dots, x_k \in C$  及任意满足  $\sum_{i=1}^k \theta_i = 1$  且  $\theta_i \geq 0$  的  $\theta_1, \dots, \theta_k \in \mathbb{R}$ ，有  $\sum_{i=1}^k \theta_i x_i \in C$ （数学归纳法可证）。

### 1.6.2 与仿射集合的关系

仿射集合是特殊的凸集，但凸集不一定是仿射集合，二者的核心差异在于组合系数的约束范围：

- 仿射组合：系数仅要求和为 1（可正可负、可大于 1）；
- 凸组合：系数要求和为 1 且非负（仅在  $[0, 1]$  内取值）。

因此，仿射集合对更宽松的组合封闭，自然也对凸组合封闭，即：若  $A$  是仿射集合，则  $A$  必是凸集。反之，凸集（如线段、球体）不一定是仿射集合（线段对  $\theta > 1$  的仿射组合不封闭）。

### 1.6.3 重要概念：凸包 (Convex Hull)

**定义 1.9.** 对任意集合  $S \subseteq \mathbb{R}^n$ ，包含  $S$  的最小凸集称为  $S$  的凸包，记为  $\text{conv}(S)$ 。其数学表达式为：

$$\text{conv}(S) = \left\{ \sum_{i=1}^k \theta_i x_i \mid x_1, \dots, x_k \in S, \theta_i \geq 0, \sum_{i=1}^k \theta_i = 1 \right\} \quad (1.6.3.1)$$

直观理解：凸包是“由  $S$  中所有点的凸组合构成的集合”，例如  $\mathbb{R}^2$  中三个不共线点的凸包是三角形，圆上所有点的凸包是闭圆盘。

### 1.6.4 典型例子

- 仿射集合的特例：线性子空间、线性方程组解空间、单点集、 $\mathbb{R}^n$  均为凸集；
- 标准凸集：
  - 闭区间  $[a, b] \subseteq \mathbb{R}$ ；
  - 球体  $\{x \in \mathbb{R}^n \mid \|x - x_0\| \leq r\}$ （ $\|\cdot\|$  为任意范数）；
  - 正象限  $\mathbb{R}_+^n = \{x \in \mathbb{R}^n \mid x_i \geq 0, i = 1, \dots, n\}$ ；
  - 半空间  $\{x \in \mathbb{R}^n \mid a^T x \leq b\}$ （ $a \neq 0$ ，本质是线性不等式约束的解空间）。

## 1.7 锥 (Cones) 与凸锥 (Convex Cones)

锥是一类对“正齐次性”封闭的集合，凸锥则进一步结合了凸性，是锥优化（如半定规划、二次锥规划）的核心结构。

### 1.7.1 核心定义：锥与凸锥

(1) 锥的定义

**定义 1.10.** 一个非空集合  $K \subseteq \mathbb{R}^n$  被称为**锥**，当且仅当对任意  $x \in K$  及任意  $\alpha \geq 0$ （非负标量），有  $\alpha x \in K$ ，即：

$$x \in K, \alpha \geq 0 \implies \alpha x \in K \quad (1.7.1.1)$$

直观理解：锥是“从原点出发的射线族”，若某条射线包含于  $K$ ，则射线的所有非负伸缩段也包含于  $K$ 。

## (2) 凸锥的定义

**定义 1.11.** 一个非空集合  $K \subseteq \mathbb{R}^n$  被称为**凸锥**，当且仅当它既是锥，又是凸集。其等价刻画有两种：

$K$  是凸锥，当且仅当对任意  $x_1, \dots, x_k \in K$  及任意  $\alpha_1, \dots, \alpha_k \geq 0$ ，有  $\sum_{i=1}^k \alpha_i x_i \in K$ （称为对非负线性组合封闭）。

## 1.7.2 重要概念：凸锥包（Convex Cone Hull）

**定义 1.12.** 对任意集合  $S \subseteq \mathbb{R}^n$ ，包含  $S$  的最小凸锥称为  $S$  的凸锥包，记为  $\text{cone}(S)$ 。其数学表达式为：

$$\text{cone}(S) = \left\{ \sum_{i=1}^k \alpha_i x_i \mid x_1, \dots, x_k \in S, \alpha_i \geq 0 \right\} \quad (1.7.2.1)$$

直观理解：凸锥包是“由  $S$  中所有点的非负线性组合构成的集合”，例如  $\mathbb{R}^2$  中两个不共线向量的凸锥包是它们张成的“角形区域”。

## 1.7.3 典型实例

- 非凸锥： $\mathbb{R}^2$  中  $\{(x_1, x_2) \mid x_1 x_2 \geq 0\}$ （第一、三象限的并集，对凸组合不封闭，如  $(1, 0)$  和  $(0, 1)$  的凸组合  $(1/2, 1/2)$  不属于该集合）；
- 凸锥：
  - 原点  $\{0\}$ （平凡凸锥）；
  - 正象限  $\mathbb{R}_+^n = \{x \mid x_i \geq 0\}$ （非负线性组合仍非负）；
  - 线性子空间  $L \subseteq \mathbb{R}^n$ （对任意  $\alpha \in \mathbb{R}$  封闭，自然对  $\alpha \geq 0$  封闭，且是凸集）；
  - 二次锥（冰淇淋锥）： $Q = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\| \leq t\}$ （验证：若  $(x_1, t_1), (x_2, t_2) \in Q$ ， $\alpha_1, \alpha_2 \geq 0$ ，则  $\|\alpha_1 x_1 + \alpha_2 x_2\| \leq \alpha_1 \|x_1\| + \alpha_2 \|x_2\| \leq \alpha_1 t_1 + \alpha_2 t_2$ ，故  $\alpha_1(x_1, t_1) + \alpha_2(x_2, t_2) \in Q$ ）。

# 1.8 三类集合的核心对比

为清晰梳理仿射集合、凸集与凸锥的差异与关联，下表从核心定义、组合类型、关键性质三个维度进行总结：

维度	仿射集合 (Affine Set)	凸集 (Convex Set)	凸锥 (Convex Cone)
核心定义	对仿射组合封闭	对凸组合封闭	对非负线性组合封闭
典型组合形式	$\theta x + (1 - \theta)y$ ( $\theta \in \mathbb{R}$ )	$\theta x + (1 - \theta)y$ ( $\theta \in [0, 1]$ )	$\alpha x + \beta y$ ( $\alpha, \beta \geq 0$ )
与线性子空间关系	平移后的线性子空间	包含线性子空间的子集 (可非平移)	线性子空间是特殊凸锥 (对任意 $\alpha \in \mathbb{R}$ 封闭)
“最小包含集”	仿射包 (aff(S))	凸包 (conv(S))	凸锥包 (cone(S))

## 1.9 重要的例子

### 1.9.1 超平面 (Hyperplane) 与半空间 (Halfspace)

超平面和半空间是由线性函数定义的基本集合，分别对应“线性等式约束”和“线性不等式约束”的解空间，是构建复杂凸集（如多面体）的基石。

#### (1) 超平面 (Hyperplane)

定义

**定义 1.13** (超平面). 设  $a \in \mathbb{R}^n$  且  $a \neq 0$  (非零法向量),  $b \in \mathbb{R}$  (常数), 则  $\mathbb{R}^n$  中的超平面定义为:

$$H = \{x \in \mathbb{R}^n \mid a^T x = b\} \quad (1.9.1.1)$$

**几何意义** 超平面是  $\mathbb{R}^n$  中“维度为  $n - 1$  的仿射集合”，可理解为“与法向量  $a$  垂直且到原点的‘距离’为  $|b|/\|a\|$  的平面”。例如:

- 当  $n = 2$  时,  $H$  是直线 ( $a_1 x_1 + a_2 x_2 = b$ );
- 当  $n = 3$  时,  $H$  是平面 ( $a_1 x_1 + a_2 x_2 + a_3 x_3 = b$ ).

**性质:** 超平面是仿射集合

#### (2) 半空间 (Halfspace)

定义

**定义 1.14** (闭半空间). 设  $a \in \mathbb{R}^n$  且  $a \neq 0$ ,  $b \in \mathbb{R}$ , 则  $\mathbb{R}^n$  中的闭半空间 (Closed Halfspace) 定义为:

$$H_+ = \{x \in \mathbb{R}^n \mid a^T x \geq b\}, \quad H_- = \{x \in \mathbb{R}^n \mid a^T x \leq b\} \quad (1.9.1.2)$$

若将不等式改为严格不等号 ( $>$  或  $<$ ), 则称为开半空间 (Open Halfspace)。

**几何意义** 半空间是超平面将  $\mathbb{R}^n$  分割成的两个“半无限区域”，其中  $H_+$  是法向量  $a$  指向的一侧,  $H_-$  是相反侧。

**性质:** 半空间是凸集 (非仿射集)

### (3) 超平面与半空间的关系

超平面是两个闭半空间的交集： $H = H_+ \cap H_-$ ；反之，每个闭半空间都是超平面的“一侧区域”，二者共同构成线性约束的几何表达。

## 1.9.2 球 (Ball) 与椭球 (Ellipsoid)

球和椭球是基于“距离”定义的凸集，广泛用于建模“变量取值范围的约束”（如稳健优化中的不确定性集合）。

### (1) 球 (Euclidean Ball)

定义

**定义 1.15** (欧氏球). 设  $x_0 \in \mathbb{R}^n$  (中心),  $r > 0$  (半径), 基于欧氏范数 ( $\|x\|_2 = \sqrt{x_1^2 + \dots + x_n^2}$ ) 的闭球定义为:

$$B(x_0, r) = \{x \in \mathbb{R}^n \mid \|x - x_0\|_2 \leq r\} \quad (1.9.2.1)$$

若将不等号改为  $<$ , 则称为开球。

性质: 球是凸集

### (2) 椭球 (Ellipsoid)

椭球是球的“仿射变换”，可描述更一般的“椭圆型区域”，在工程优化中常用于拟合数据分布或刻画变量波动范围。

定义

**定义 1.16** (椭球). 设  $x_0 \in \mathbb{R}^n$  (中心),  $P \in \mathbb{S}_{++}^n$  (正定对称矩阵, 控制椭球的形状与方向),  $r > 0$  (缩放因子), 则椭球定义为:

$$\mathcal{E} = \{x \in \mathbb{R}^n \mid (x - x_0)^T P^{-1} (x - x_0) \leq r^2\} \quad (1.9.2.2)$$

等价表达: 通过仿射变换  $x = x_0 + rP^{1/2}z$  (其中  $P^{1/2}$  是  $P$  的正定平方根,  $z \in \mathbb{R}^n$ ), 椭球可表示为球的像:

$$\mathcal{E} = \{x_0 + rP^{1/2}z \mid \|z\|_2 \leq 1\} \quad (1.9.2.3)$$

当  $P = I$  (单位矩阵) 时, 椭球退化为球  $B(x_0, r)$ 。

性质: 椭球是凸集

## 1.9.3 范数球 (Norm Ball) 与范数锥 (Norm Cone)

范数球和范数锥是基于“一般范数”的扩展集合，将球的“距离约束”与锥的“正齐次性”结合，是范数优化、锥优化的核心结构。

## (1) 范数的回顾

首先明确范数的定义：

**定义 1.17** (范数). 函数  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  称为范数, 若对任意  $x, y \in \mathbb{R}^n$  和  $\alpha \in \mathbb{R}$ , 满足:

1. 非负性:  $\|x\| \geq 0$ , 且  $\|x\| = 0 \iff x = 0$ ;
2. 齐次性:  $\|\alpha x\| = |\alpha| \|x\|$ ;
3. 三角不等式:  $\|x + y\| \leq \|x\| + \|y\|$ .

常见范数包括欧氏范数 ( $\|\cdot\|_2$ )、1-范数 ( $\|x\|_1 = \sum_{i=1}^n |x_i|$ )、无穷范数 ( $\|x\|_\infty = \max_{i=1, \dots, n} |x_i|$ ) 等。

## (2) 范数球 (Norm Ball)

定义

**定义 1.18** (范数球). 设  $\|\cdot\|$  是  $\mathbb{R}^n$  上的范数,  $x_0 \in \mathbb{R}^n$ ,  $r > 0$ , 则范数球定义为:

$$B_{\|\cdot\|}(x_0, r) = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq r\} \quad (1.9.3.1)$$

当  $x_0 = 0$  且  $r = 1$  时, 称为单位范数球 (Unit Norm Ball)。

性质: 范数球是凸集

实例

- 1-范数球 ( $\|x\|_1 \leq 1$ ): 在  $\mathbb{R}^2$  中是菱形,  $\mathbb{R}^3$  中是正八面体;
- 无穷范数球 ( $\|x\|_\infty \leq 1$ ): 在  $\mathbb{R}^2$  中是正方形,  $\mathbb{R}^3$  中是正方体。

## (3) 范数锥 (Norm Cone)

定义

**定义 1.19** (范数锥). 设  $\|\cdot\|$  是  $\mathbb{R}^n$  上的范数, 定义范数锥 (又称“冰淇淋锥”的推广) 为:

$$K_{\|\cdot\|} = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\| \leq t\} \quad (1.9.3.2)$$

其中  $(x, t)$  是  $\mathbb{R}^{n+1}$  中的向量,  $t$  可理解为“范数的上界”。

性质: 范数锥是凸锥

实例

- 二次锥 (Quadratic Cone): 当  $\|\cdot\| = \|\cdot\|_2$  时, 范数锥即为二次锥 (又称 Lorentz 锥):

$$Q = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\|_2 \leq t\} \quad (1.9.3.3)$$

是锥优化中最常用的凸锥之一。

- 1-范数锥：  $K_{\|\cdot\|_1} = \{(x, t) \mid \sum_{i=1}^n |x_i| \leq t\}$ ，在  $\mathbb{R}^2 \times \mathbb{R}$  中呈“四棱锥”形状。

#### 1.9.4 多面体 (Polyhedron) 与单纯形 (Simplex)

多面体是有限个线性等式与不等式约束的交集，是线性规划可行域的抽象；单纯形则是多面体的特殊情况，是“最低维度”的多面体，在数值优化中常用于构建搜索区域。

##### (1) 多面体 (Polyhedron)

定义

**定义 1.20.** 多面体 多面体是有限个线性等式与不等式约束的交集，其数学表达式为：

$$\mathcal{P} = \{x \in \mathbb{R}^n \mid Ax \leq b, Cx = d\} \quad (1.9.4.1)$$

其中：

- $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ ：对应  $m$  个半空间约束 ( $Ax \leq b$  即  $a_i^T x \leq b_i$ ,  $i = 1, \dots, m$ )。
- $C \in \mathbb{R}^{p \times n}$ ,  $d \in \mathbb{R}^p$ ：对应  $p$  个超平面约束 ( $Cx = d$  即  $c_j^T x = d_j$ ,  $j = 1, \dots, p$ )。

性质：多面体是凸集

实例

- 正象限  $\mathbb{R}_+^n = \{x \mid x_i \geq 0, i = 1, \dots, n\}$ ：是多面体 ( $A = -I$ ,  $b = 0$ )；
- 线性规划的可行域：  $X = \{x \mid Ax = b, x \geq 0\}$ ，是多面体。

##### (2) 单纯形 (Simplex)

单纯形是“由  $n+1$  个仿射无关点生成的凸包”，是维度为  $n$  的“最简单”多面体（顶点数量最少的多面体）。

**定义 1.21.** 单纯形

**定义 1 (基于仿射无关点)** 设  $v_0, v_1, \dots, v_n \in \mathbb{R}^n$  是仿射无关的点（即向量  $v_1 - v_0, \dots, v_n - v_0$  线性无关），则由这些点生成的单纯形定义为：

$$\Delta = \text{conv}(v_0, v_1, \dots, v_n) = \left\{ \sum_{i=0}^n \theta_i v_i \mid \theta_i \geq 0, \sum_{i=0}^n \theta_i = 1 \right\} \quad (1.9.4.2)$$

**定义 2 (标准单纯形)** 最常用的是标准单纯形（以单位向量为顶点），定义为：

$$\Delta_n = \{x \in \mathbb{R}^n \mid x_1 + x_2 + \dots + x_n = 1, x_i \geq 0, i = 1, \dots, n\} \quad (1.9.4.3)$$

其顶点为  $e_1 = (1, 0, \dots, 0)^T, e_2 = (0, 1, \dots, 0)^T, \dots, e_n = (0, \dots, 1)^T$  ( $n$  个标准单位向量)，但注意：此处  $n$  维标准单纯形的顶点数量为  $n$ ，与定义 1 中“ $n$  维单纯形需  $n+1$  个顶点”的差异源于“是否包含原点”——若定义为  $\Delta'_n = \{(x, 1 - \sum x_i) \mid x \in \Delta_n\}$ ，则顶点为  $e_1, \dots, e_n, 0$ ，共  $n+1$  个，符合定义 1。

### 关键性质

1. **维度**: 由  $n+1$  个仿射无关点生成的单纯形是  $n$  维的 (与空间维度一致)。
2. **凸性**: 单纯形是有限个点的凸包, 而凸包是凸集 (由凸包定义: 所有凸组合的集合, 自然对凸组合封闭), 故单纯形是凸集。
3. **多面体属性**: 单纯形可表示为有限个线性等式与不等式的交集 (如标准单纯形的约束  $x_1 + \cdots + x_n = 1$  和  $x_i \geq 0$ ), 因此是多面体。

### 实例

- 1 维单纯形:  $\Delta_1 = \{x \in \mathbb{R} \mid x = 1, x \geq 0\}$  即点  $\{1\}$ ; 或扩展为  $\Delta'_1 = \{(x, 1-x) \mid x \geq 0\}$  即线段  $[0, 1]$ ;
- 2 维单纯形 (标准):  $\Delta_2 = \{(x_1, x_2) \mid x_1 + x_2 = 1, x_1, x_2 \geq 0\}$  即连接  $(1, 0)$  和  $(0, 1)$  的线段; 或扩展为  $\Delta'_2 = \{(x_1, x_2, x_3) \mid x_1 + x_2 + x_3 = 1, x_i \geq 0\}$  即三角形 (3 个顶点);
- 3 维单纯形: 扩展形式为四面体 (4 个顶点)。

## 第二章 凸优化的基本性质

### 2.1 凸优化问题的形式

#### 2.1.1 一般优化模型

一个一般优化问题的数学表达式为：

$$\min_{x \in \mathbb{R}^n} f_0(x) \quad \text{s.t.} \quad f_i(x) \leq 0, i = 1, \dots, m; \quad h_j(x) = 0, j = 1, \dots, p \quad (2.1.1.1)$$

其中各部分含义如下：

- 目标函数：  $f_0(x)$ ，即需要最小化的函数；
- 约束函数：
  - 不等式约束：  $f_i(x) \leq 0$ （共  $m$  个， $i$  取值为 1 到  $m$ ）；
  - 等式约束：  $h_j(x) = 0$ （共  $p$  个， $j$  取值为 1 到  $p$ ）；
- 可行域：满足所有约束条件的变量集合，定义为  $X = \{x \mid f_i(x) \leq 0, h_j(x) = 0\}$ 。

#### 2.1.2 凸优化的定义

定义 2.1 (凸优化问题). 当且仅当满足以下三个条件时，一般优化问题是凸优化问题：

- 目标函数  $f_0(x)$  为凸函数；
- 不等式约束函数  $f_i(x)$  ( $i = 1, \dots, m$ ) 均为凸函数；
- 等式约束函数  $h_j(x)$  ( $j = 1, \dots, p$ ) 均为仿射函数，即满足形式  $h_j(x) = a_j^T x + b_j$ （其中  $a_j$  为向量， $b_j$  为常数）。

此时，凸优化问题可简化描述为：**minimize a convex function over a convex set**（在凸集上最小化凸函数）。

### 2.2 凸函数的定义以及性质

#### 2.2.1 定义与几何意义

数学定义

**定义 2.2.** 函数  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  是凸函数，当且仅当对任意  $x_1, x_2 \in \text{dom}(f)$  ( $\text{dom}(f)$  表示函数  $f$  的定义域) 和任意  $\theta \in [0, 1]$ ，满足以下不等式：

$$f(\theta x_1 + (1 - \theta)x_2) \leq \theta f(x_1) + (1 - \theta)f(x_2) \quad (2.2.1.1)$$

### 几何意义

凸函数的曲线始终位于连接两点  $(x_1, f(x_1))$  和  $(x_2, f(x_2))$  的割线之下。

直观理解：在函数定义域内任取两点，两点间的割线不会低于函数曲线本身。

### 2.2.2 Jensen 不等式

Jensen 不等式是凸函数最重要的性质之一，具体表述如下：

**定理 2.1** (Jensen 不等式). 若  $f$  为凸函数， $\{x_i\}$  为  $\text{dom}(f)$  内的任意点集， $\{\theta_i\}$  为非负权重且满足  $\sum_i \theta_i = 1$ ，则：

$$f\left(\sum_i \theta_i x_i\right) \leq \sum_i \theta_i f(x_i) \quad (2.2.2.1)$$

### 应用场景

Jensen 不等式不仅是判定函数凸性的重要依据，还广泛应用于：

- 概率论（如期望相关不等式推导）；
- 信息论（如 KL 散度、熵函数的性质分析）；
- 期望下界的证明。

### 典型示例

若  $f(x) = x^2$ （已知为凸函数），对随机变量  $X$  应用 Jensen 不等式，可得：

$$(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$$

（其中  $\mathbb{E}[X]$  表示  $X$  的期望， $\mathbb{E}[X^2]$  表示  $X^2$  的期望）。

### 2.2.3 一阶条件 (First-order condition)

**适用前提** 函数  $f$  可微（即梯度  $\nabla f(x)$  在定义域内存在）。

### 判定准则

**命题 2.1.** 函数  $f$  是凸函数，当且仅当对任意  $x, y \in \text{dom}(f)$ ，满足：

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad (2.2.3.1)$$

## 几何与直观意义

- 数学层面：函数在任意点  $x$  处的切平面（或切线，当  $n = 1$  时）是函数的全局下界；
- 梯度意义：梯度  $\nabla f(x)$  始终指向函数的上升方向。

## 2.2.4 二阶条件 (Hessian 判定)

适用前提 函数  $f$  二阶可微（即 Hessian 矩阵  $\nabla^2 f(x)$  在定义域内存在）。

## 判定准则

命题 2.2. 函数  $f$  是凸函数，当且仅当对任意  $x \in \text{dom}(f)$ ，其 Hessian 矩阵满足半正定：

$$f \text{ 是凸的} \Leftrightarrow \nabla^2 f(x) \succeq 0, \forall x \quad (2.2.4.1)$$

## 强凸函数的延伸判定

定义 2.3. 当且仅当对任意  $x \in \text{dom}(f)$ ，Hessian 矩阵满足正定 ( $\nabla^2 f(x) \succ 0$ ) 时，函数  $f$  为强凸函数（强凸是凸函数的更强形式）。

## 强凸性定义

定义 2.4 (强凸函数（等价定义）). 函数  $f$  称为强凸函数，若存在常数  $\mu > 0$ ，对所有  $x, y$ ：

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2} \|y - x\|^2 \quad (2.2.4.2)$$

其中  $\mu$  称为强凸系数。

这意味着：函数不仅“向上弯”，而且弯曲程度有下界；强凸函数具有唯一最优解。

- 性质 1：强凸函数有且仅有一个极小点；
- 性质 2：梯度法在强凸问题上以线性速率收敛

证明 证明：强凸函数梯度法线性收敛。

## 证明前提

定义 2.5 (Lipschitz 连续). 梯度 Lipschitz 连续（工程中常用假设）：存在  $L > 0$ ，对所有  $x, y$  有  $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ 。

- 强凸函数  $f$  满足梯度 Lipschitz 连续；
- 梯度法迭代公式： $x_{k+1} = x_k - \alpha \nabla f(x_k)$ ，其中步长  $\alpha$  取  $\alpha = \frac{1}{L}$ （最优步长选择）；
- $x^*$  为强凸函数的唯一极小点，故  $\nabla f(x^*) = 0$ 。

**证明过程（核心推导）** 1. 展开迭代误差范数：对  $x_{k+1} = x_k - \alpha \nabla f(x_k)$ ，两边减去  $x^*$  得：

$$x_{k+1} - x^* = (x_k - x^*) - \alpha \nabla f(x_k)$$

取平方范数（利用  $\|a - b\|^2 = \|a\|^2 - 2a^T b + \|b\|^2$ ）：

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha \nabla f(x_k)^T (x_k - x^*) + \alpha^2 \|\nabla f(x_k)\|^2 \quad (1)$$

2. 利用强凸性放缩梯度项：对  $x = x_k, y = x^*$  应用强凸定义，代入  $\nabla f(x^*) = 0$ ：

$$f(x^*) \geq f(x_k) + \nabla f(x_k)^T (x^* - x_k) + \frac{\mu}{2} \|x^* - x_k\|^2$$

由于  $f(x^*) \leq f(x_k)$  ( $x^*$  是最优解)，整理得：

$$\nabla f(x_k)^T (x_k - x^*) \geq \frac{\mu}{2} \|x_k - x^*\|^2 \quad (2)$$

3. 利用 *Lipschitz* 连续放缩梯度范数：对  $x = x_k, y = x^*$  应用梯度 *Lipschitz* 连续，代入  $\nabla f(x^*) = 0$ ：

$$\|\nabla f(x_k)\| \leq L \|x_k - x^*\| \implies \|\nabla f(x_k)\|^2 \leq L^2 \|x_k - x^*\|^2 \quad (3)$$

4. 代入迭代公式证线性收敛：将 (2)(3) 代入 (1)，并取  $\alpha = \frac{1}{L}$ ：

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2\alpha \cdot \frac{\mu}{2} \|x_k - x^*\|^2 + \alpha^2 L^2 \|x_k - x^*\|^2 \\ &= \|x_k - x^*\|^2 (1 - \alpha\mu + \alpha^2 L^2) \\ &= \|x_k - x^*\|^2 \left(1 - \frac{\mu}{L} + \frac{L^2}{L^2}\right) \\ &= \|x_k - x^*\|^2 \left(1 - \frac{\mu}{L}\right) \end{aligned} \quad (2.2.4.3)$$

其中  $0 < 1 - \frac{\mu}{L} < 1$  (因  $\mu < L$ ，强凸系数小于 *Lipschitz* 常数)。

**结论** 梯度法迭代误差满足  $\|x_{k+1} - x^*\|^2 \leq \gamma \|x_k - x^*\|^2$  ( $\gamma = 1 - \frac{\mu}{L}$  为收敛因子)，即以线性速率收敛。

## 2.3 全局与局部最优

### 2.3.1 定义回顾

对于一般优化问题，局部最优解与全局最优解的定义如下：

- 定义 2.6** (局部与全局最优解).
- **局部最优解**：存在  $\epsilon > 0$ ，使得对所有满足  $\|x - x^*\| < \epsilon$  的可行点  $x$ ，均有  $f(x) \geq f(x^*)$  (即  $x^*$  在自身邻域内是最优的)。
  - **全局最优解**：对所有可行域内的点  $x$ ，均有  $f(x) \geq f(x^*)$  (即  $x^*$  在整个可行域内是最优的)。

### 2.3.2 凸优化的核心定理

**定理 2.2** (凸优化局部最优即全局最优). 若目标函数  $f$  是凸函数，且优化问题的可行域  $\mathcal{X}$  是凸集，则任何局部最优解都是全局最优解。

证明（反证法）。

1. 假设前提：假设  $x^*$  是局部最优解，但非全局最优解。根据全局最优解的定义，此时存在可行点  $x' \in \mathcal{X}$ ，使得  $f(x') < f(x^*)$ 。

2. 构造凸组合：定义凸组合  $x_\theta = (1 - \theta)x^* + \theta x'$ ，其中  $\theta \in (0, 1)$ （即  $x_\theta$  是  $x^*$  与  $x'$  连线上的点）。
3. 利用凸集性质：由于可行域  $\mathcal{X}$  是凸集，根据凸集的定义， $x_\theta \in \mathcal{X}$ （即  $x_\theta$  是可行点）。
4. 利用凸函数性质：由于  $f$  是凸函数，根据凸函数的定义：

$$f(x_\theta) = f((1 - \theta)x^* + \theta x') \leq (1 - \theta)f(x^*) + \theta f(x')$$

5. 推出矛盾：结合假设  $f(x') < f(x^*)$ ，代入上式得：

$$f(x_\theta) \leq (1 - \theta)f(x^*) + \theta f(x') < (1 - \theta)f(x^*) + \theta f(x^*) = f(x^*)$$

即  $f(x_\theta) < f(x^*)$ 。又因为当  $\theta$  足够小时， $x_\theta$  满足  $\|x_\theta - x^*\| = \theta\|x' - x^*\| < \epsilon$ （符合局部最优解的邻域条件），这与  $x^*$  是局部最优解的定义矛盾。

### 2.3.3 强凸函数的唯一最优性

若  $f$  为  $\mu$ -强凸函数（ $\mu > 0$  为强凸系数），则：

- 最优解唯一性：强凸函数的最优解有且仅有一个（不存在多个最优解）；
- 函数值差的二次下界：距离最优点的函数值差满足二次下界关系（具体表现为  $f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2$ ）。

该性质为算法收敛性分析（如梯度下降法、牛顿法等）提供了重要的理论基础。

## 2.4 凸优化的几何意义

### 2.4.1 可行域与等高线

**定义 2.7** (等高线与凸优化几何特征). 凸优化的几何特征具有明确的直观性：目标函数的等高线（*level set*）与凸形可行域（*convex feasible region*）相切的点，即为全局最优解。

几何补充：等高线是目标函数值等于某一常数的点的集合（如二次函数的椭圆等高线）；由于可行域是凸集，其边界呈“凸向外侧”的形态，二者相切时仅存在唯一接触点，该点即为整个可行域内使目标函数最小的点。

### 2.4.2 法向条件（支撑超平面）

**定义 2.8** (支撑超平面). 从几何角度分析，凸优化问题的最优点  $x^*$  处必然存在一个支撑超平面（*supporting hyperplane*），其数学表达式为：

$$\mathbf{a}^T (x - x^*) \geq 0, \quad \forall x \in \mathcal{X} \quad (2.4.2.1)$$

其中：

- $\mathbf{a} = \nabla f(x^*)$ ，即目标函数  $f$  在最优点  $x^*$  处的梯度；
- $\mathcal{X}$  为优化问题的可行域。

支撑超平面的核心意义：

- 几何层面：该超平面与凸集  $\mathcal{X}$  在  $x^*$  处“相切”，且凸集  $\mathcal{X}$  完全位于超平面的一侧；
- 优化层面：超平面的法向量（即梯度  $\nabla f(x^*)$ ）指向目标函数的上升方向，因此不存在从  $x^*$  出发、指向可行域内部且能使目标函数值下降的方向；
- 理论关联：该条件与 KKT 条件（Karush-Kuhn-Tucker 条件）完全对应，本质是“梯度与约束法向一致”的几何体现。

KKT 条件后面会讲到

### 2.4.3 凸组合与最优性路径

在凸优化问题中，全局最优点  $x^*$  常可解释为多个可行极值点的凸组合（convex combination），具体表现为：

以资源分配问题为例：若存在  $k$  种可行的资源配置方案，每种方案对应可行域  $\mathcal{X}$  内的一个点  $x_i$  ( $i = 1, 2, \dots, k$ )，则最终的全局最优解  $x^*$  一定位于这些点的凸包（convex hull）内。

凸包：

$$\text{conv}\{x_1, x_2, \dots, x_k\} = \left\{ \sum_{i=1}^k \theta_i x_i \mid \theta_i \geq 0, \sum_{i=1}^k \theta_i = 1 \right\} \quad (2.4.3.1)$$

关键结论：

凸包是包含所有点  $x_1, x_2, \dots, x_k$  的最小凸集，全局最优解  $x^*$  位于凸包内，体现了凸优化最优解的“折中性质”——最终解是多个局部可行方案的合理权衡，而非极端方案。

几何示例：若两种可行方案对应平面上的两个点，其凸包为两点间的线段，全局最优解即为线段上使目标函数最小的点。

## 2.5 凸优化的转化以及示例

### 2.5.1 优化问题的等价变换以及实例

**定义 2.9** (等价变换条件). 等价变换需满足三个条件：最优值相同、最优解可相互恢复、可行域一一对应。

#### 1. SVM 软间隔原问题 $\rightarrow$ Hinge 损失形式（最典型等价转化）

- 原问题（含松弛变量）：软间隔 SVM 原问题为

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i, \quad \text{s.t.} \quad y_i (w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (2.5.1.1)$$

该问题含变量  $w, b, \xi$ ，约束涉及松弛变量  $\xi_i$ ，求解需同时处理三类变量。

- 等价转化：Hinge 损失形式：通过“松弛变量消去”的等价思路，将原问题转化为仅含  $w$  的无约束问题：

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_i \max(0, 1 - y_i w^T x_i) \quad (2.5.1.2)$$

其中  $\max(0, 1 - y_i w^T x_i)$  为 Hinge 损失，本质是用“损失项”隐式替代松弛变量  $\xi_i$  ( $\xi_i \geq \max(0, 1 - y_i w^T x_i)$ )，最小化目标时二者取等)。

• 等价性验证:

- 最优值相同: 原问题通过  $\xi_i$  控制“间隔违反程度”，Hinge 损失直接量化该程度，最小化目标的核心逻辑一致;
- 最优解可恢复: 从 Hinge 损失的最优解  $w^*$ , 可反向计算  $b^*$  (如通过支持向量满足  $y_i(w^* x_i + b^*) = 1$ ), 松弛变量  $\xi_i^* = \max(0, 1 - y_i w^{*T} x_i)$ ;
- 可行域一一对应: 原问题的可行域 ( $\xi_i \geq 0$  且  $y_i(w^T x_i + b) \geq 1 - \xi_i$ ) 与 Hinge 损失的“隐含可行域”(无显式约束, 但损失项确保等价约束) 完全对应。

- 实例价值: 转化后无需处理松弛变量  $\xi$ , 将带约束问题简化为无约束凸优化, 可直接用梯度法求解, 同时保持 SVM “最大间隔分类”的核心逻辑。

## 2. 逻辑回归: 似然最大化 $\rightarrow$ 负对数似然最小化 (单调变换等价)

- 原问题 (非凸形式): 逻辑回归的核心是“最大化样本似然概率”, 似然函数为  $\prod_{i=1}^m \sigma(y_i w^T x_i)$  ( $\sigma(\cdot)$  为 Sigmoid 函数), 该函数是乘积形式, 非凸且求解困难。
- 等价转化: 凸形式: 利用对数函数的单调性 ( $\log(\cdot)$  单调递增, 最大化  $f$  等价于最大化  $\log f$ ), 再通过“取负”将最大化问题转化为最小化问题, 最终目标函数为:

$$\min_w - \sum_{i=1}^m \log(\sigma(y_i w^T x_i)) \quad (2.5.1.3)$$

该“负对数似然函数为凸函数”。

- 等价性验证: 由于  $\log(\cdot)$  和“取负”均为单调变换, 原问题的最优解  $w^*$  与转化后问题的最优解完全一致, 最优值仅相差常数倍 (对数与负号的影响), 满足等价变换的三个条件。
- 实例价值: 将非凸的乘积似然转化为凸的加法函数, 可通过梯度法、牛顿法高效求解, 且保障解为全局最优。

## 3. 最小二乘回归: 隐含约束 $\rightarrow$ 无约束凸问题 (可行域等价)

- 原问题 (隐含约束): 最小二乘的目标是

$$\min_w \|Xw - y\|^2 \quad (2.5.1.4)$$

其“可行域”本质是“所有使误差有定义的  $w$ ”(即全空间  $\mathbb{R}^n$ ), 无显式约束。

- 等价转化: 无约束凸问题: 无需引入额外变量, 直接利用凸函数判定条件——目标函数的 Hessian 矩阵为  $2X^T X \succeq 0$  (半正定), 因此是凸函数。此时问题等价于“在凸集 (全空间) 上最小化凸函数”, 完全符合凸优化的定义。
- 实例价值: 通过“可行域等价”避免复杂约束处理, 直接用闭式解 ( $w^* = (X^T X)^{-1} X^T y$ ) 或梯度法求解, 是机器学习中最易实现的凸优化实例之一。

### 2.5.2 非凸到凸的重构思路在实例中的延伸（解决“非凸建模难题”）

四类非凸到凸的重构技术：函数松弛、对偶化、变量替换、线性化，这些思路在实例中被灵活应用，让原本非凸的模型具备凸优化的“全局最优性”。

#### 1. 函数松弛：Hinge 损失替代 0-1 损失（SVM 中的非凸近似）

SVM 的 Hinge 损失形式，本质是“函数松弛”技术的落地——用凸上界替代非凸项：

- **非凸痛点**：分类任务的理想损失是“0-1 损失”（正确分类损失为 0，错误分类损失为 1），但 0-1 损失是阶梯函数，非凸且不可微，无法用于凸优化建模。
- **凸松弛方案**：根据“函数松弛”思路，用 Hinge 损失（ $\max(0, 1 - y_i w^T x_i)$ ）作为 0-1 损失的凸上界——对所有  $w$ ，均有  $\max(0, 1 - y_i w^T x_i) \geq \mathbb{I}(y_i w^T x_i < 1)$ （ $\mathbb{I}$  为指示函数，即 0-1 损失）。
- **实例应用**：SVM 选择 Hinge 损失作为目标项，既保留“惩罚间隔违反样本”的核心逻辑，又通过“凸松弛”让目标函数成为凸函数。此时问题转化为凸优化，保障全局最优，避免 0-1 损失导致的“局部最优陷阱”。

#### 2. 对偶化：Lasso 回归的对偶问题（处理不可微凸项）

Lasso 回归的求解，依赖“对偶化”技术——通过拉格朗日对偶将原问题转化为更易求解的凸问题：

- **原问题痛点**：Lasso 回归的目标函数为  $\min_w \|Xw - y\|^2 + \lambda \|w\|_1$ ，其中  $\|w\|_1$  是凸函数但不可微（在  $w_i = 0$  处无梯度），直接用梯度法求解困难。
- **对偶化方案**：根据“对偶化”思路，构造原问题的拉格朗日对偶问题，给出 Lasso 对偶问题为：

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i^T x_j), \quad \text{s.t.} \quad 0 \leq \alpha_i \leq C, \sum_i \alpha_i y_i = 0 \quad (2.5.2.1)$$

该对偶问题是凸二次规划（目标函数为二次函数，Hessian 半正定；约束均为线性，可行域为凸集）。

- **实例价值**：对偶问题虽与原问题变量不同（从  $w$  变为  $\alpha$ ），但满足 Slater 条件（对偶间隙为 0），最优值与原问题相同，且对偶问题可通过成熟的二次规划算法求解（无需处理不可微性）。求解后再通过  $w^* = \sum_i \alpha_i^* y_i x_i$  恢复原问题最优解，同时保留 Lasso “稀疏性”的核心特性。

#### 3. 线性化：投资组合优化的非线性约束处理（Markowitz 模型延伸）

- **基础模型（凸二次规划）**：标准 Markowitz 模型为

$$\min_w \frac{1}{2} w^T \Sigma w - \mu^T w, \quad \text{s.t.} \quad 1^T w = 1, w_i \geq 0 \quad (2.5.2.2)$$

由于  $\Sigma$ （协方差矩阵）半正定，目标函数凸，可行域凸，属于凸二次规划。

- **非凸扩展痛点**：实际投资中常存在非线性约束（如交易成本为  $w_i^2$ ，或最小持仓比例为非线性函数），这些约束会破坏可行域的凸性，导致模型非凸。

- **线性化方案：**根据“线性化”思路，对非线性约束进行“局部线性近似”或“分段线性逼近（PWL）”——例如将交易成本  $w_i^2$  在可行域内分段用线性函数近似，每个分段内约束为线性，整体可行域仍为凸集。
- **实例价值：**线性化后模型仍保持“目标函数凸 + 可行域凸”的特性，保障全局最优，解决了“实际投资约束下模型不可解”的问题，同时让最优解仍落在“有效前沿”上。

### 2.5.3 转化逻辑的核心价值

所有实例均通过“等价变换”或“非凸重构”，最终满足凸优化的定义（目标函数凸 + 可行域凸），进而利用“局部最优即全局最优”的定理，实现“高效求解”与“结果可靠”的双重目标。具体可总结为三类转化路径：

- **复杂凸 → 简洁凸：**如 SVM 原问题 → Hinge 损失（消去松弛变量）、最小二乘 → 无约束问题（简化可行域）；
- **非凸 → 凸近似：**如 0-1 损失 → Hinge 损失（函数松弛）、非线性约束 → 分段线性约束（线性化）；
- **难求解凸 → 易求解凸：**如 Lasso 原问题 → 对偶问题（处理不可微性）。

## 第三章 无约束优化问题

### 3.1 无约束优化问题

#### 3.1.1 问题基本形式

**定义 3.1** (无约束优化问题). 针对凸且二阶可微的目标函数, 无约束优化问题的数学形式定义为:

$$\min f(x) \quad (3.1.1.1)$$

其中,  $f(x)$  满足“凸性”与“二阶可微性”这两个核心前提假设。

#### 3.1.2 最优性条件

**命题 3.1** (最优性条件). 无约束优化问题达到最优解  $x^*$  的核心必要条件为:

$$\nabla f(x^*) = 0 \quad (3.1.2.1)$$

即最优解处函数的梯度（一阶导数）等于零向量；若该等式无法直接求解，则需通过迭代方法逼近最优解。

#### 3.1.3 核心求解框架（下山法迭代格式）

下山法通过迭代构造严格递减的函数值序列以逼近最优解，其数学化迭代流程如下：

1. **初始设定**: 给定初始迭代点  $x^{(0)}$ ，初始化迭代次数  $k = 0$ ；
2. **方向选取**: 确定第  $k$  次迭代的搜索方向  $\Delta x^{(k)}$ （后续需进一步优化方向选取规则）；
3. **步长选取**: 确定第  $k$  次迭代的步长  $\alpha > 0$ （后续需进一步优化步长选取规则）；
4. **迭代更新**: 按以下公式更新迭代点：

$$x^{(k+1)} = x^{(k)} + \alpha \Delta x^{(k)} \quad (3.1.3.1)$$

同时更新迭代次数  $k \leftarrow k + 1$ ，且需满足函数值严格递减条件：

$$f(x^{(0)}) > f(x^{(1)}) > f(x^{(2)}) > \dots \quad (3.1.3.2)$$

综上，无约束优化问题的核心数学转化目标为：通过数学规则确定“搜索方向  $\Delta x^{(k)}$ ”与“步长  $\alpha$ ”，使上述迭代流程满足严格递减性并收敛至最优解。

## 3.2 搜索方向的确定

### 3.2.1 视角 1：线性化与下降条件

**定义 3.2** (梯度  $L$ -Lipschitz 条件). 若函数  $f$  的梯度满足  $L$ -Lipschitz 连续性, 等价于:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad (3.2.1.1)$$

其中  $L$  为 Lipschitz 常数,  $\|\cdot\|_2$  表示欧氏范数。

**命题 3.2** (函数线性化不等式). 利用基本定理与 Cauchy-Schwarz 公式, 可推导出函数在点  $x$  处的线性化上界:

$$f(x + \Delta) \leq f(x) + \nabla f(x)^\top \Delta + \frac{L}{2} \|\Delta\|_2^2 \quad (3.2.1.2)$$

推导: 函数线性化不等式.

梯度  $L$ -Lipschitz 连续性 (推导的基础, 保证梯度变化有界):

$$\|\nabla f(x) - \nabla f(x + t\Delta)\|_2 \leq L \cdot \|t\Delta\|_2 = Lt\|\Delta\|_2 \quad (t \in [0, 1]) \quad (3.2.1.3)$$

多元微积分基本定理:

$$f(x + \Delta) - f(x) = \int_0^1 \nabla f(x + t\Delta)^\top \Delta dt \quad (3.2.1.4)$$

积分拆分与线性项提取:

$$\int_0^1 \nabla f(x + t\Delta)^\top \Delta dt = \nabla f(x)^\top \Delta \cdot \int_0^1 dt + \int_0^1 [\nabla f(x + t\Delta) - \nabla f(x)]^\top \Delta dt \quad (3.2.1.5)$$

其中第一部分积分结果直接为线性项:

$$\nabla f(x)^\top \Delta \cdot \int_0^1 dt = \nabla f(x)^\top \Delta \quad (3.2.1.6)$$

**Cauchy-Schwarz 不等式 +  $L$ -Lipschitz 条件** (控制第二部分积分, 得到 “二次项  $\frac{L}{2} \|\Delta\|_2^2$ ”): 由 Cauchy-Schwarz 不等式:

$$[\nabla f(x + t\Delta) - \nabla f(x)]^\top \Delta \leq \|\nabla f(x + t\Delta) - \nabla f(x)\|_2 \cdot \|\Delta\|_2 \quad (3.2.1.7)$$

代入  $L$ -Lipschitz 条件并积分:

$$\int_0^1 \|\nabla f(x + t\Delta) - \nabla f(x)\|_2 \cdot \|\Delta\|_2 dt \leq \int_0^1 Lt\|\Delta\|_2^2 dt = \frac{L}{2} \|\Delta\|_2^2 \quad (3.2.1.8)$$

合并得到最终不等式:

$$f(x + \Delta) - f(x) \leq \nabla f(x)^\top \Delta + \frac{L}{2} \|\Delta\|_2^2 \implies f(x + \Delta) \leq f(x) + \nabla f(x)^\top \Delta + \frac{L}{2} \|\Delta\|_2^2 \quad (3.2.1.9)$$

**定义 3.3** (下降方向). 令  $\Delta = \alpha p$  ( $\alpha > 0$  为步长,  $p$  为搜索方向), 只要满足特定条件, 就存在足够小的  $\alpha > 0$  使函数值降低——这是下降方向的充要条件。其中, 最自然的搜索方向选择为负梯度方向:

$$p_{gd} = -\nabla f(x) \quad (3.2.1.10)$$

### 3.2.2 视角 2：一般范数下的“最速下降”

**定义 3.4** (最速下降方向). 对任意范数  $\|\cdot\|$  及其对应的对偶范数  $\|\cdot\|_*$ , “最速下降方向”需通过求解以下优化问题得到:

$$p^* = \arg \min_{\|d\|_* \leq 1} \nabla f(x)^\top d \quad (3.2.2.1)$$

上述优化问题的解为:

$$p^* = -\frac{\nabla f(x)}{\|\nabla f(x)\|} \quad (3.2.2.2)$$

**例 3.1** (特殊范数案例). • 当使用欧氏范数时, 对偶范数与原范数一致, 此时  $p^*$  即为负梯度方向 (与前文中结论一致);

- 若使用 **Hessian 度量** (定义为  $\|d\|_{H(x)} = \sqrt{d^\top H(x) d}$ ,  $H(x)$  为函数  $f$  在  $x$  处的 Hessian 矩阵), 则 “最速” 方向等价于 Newton 步 (详见后续 Newton 法相关内容)。

解释: 函数  $f$  在点  $x$  沿方向  $d$  的局部下降速度, 由梯度与方向的内积  $\nabla f(x)^\top d$  决定:

- 内积越小 (越负), 函数沿  $d$  下降越快;
- 因此 “找最速下降方向”, 等价于在约束  $\|d\|_* \leq 1$  (对偶范数单位球) 下, 求解:

$$p^* = \arg \min_{\|d\|_* \leq 1} \nabla f(x)^\top d \quad (3.2.2.3)$$

设  $a = \nabla f(x)$ , 需先确定  $a^\top d$  (即  $\nabla f(x)^\top d$ ) 的最小可能值。根据对偶范数的核心性质: 对任意满足  $\|d\|_* \leq 1$  的  $d$ , 有:

$$a^\top d \geq -\|a\| \quad (3.2.2.4)$$

方向  $p^* = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ , 既满足对偶范数单位球约束, 又能让内积达到最小 (下降最快), 因此它就是单位球上的最速下降方向。

### 3.2.3 预条件化的作用

最速下降方向是 “当前范数下, 使内积  $\nabla f(x)^\top d$  最小 (下降最快) 的方向”。预条件化的本质是改变 “范数度量标准”: 不再用欧氏范数, 而是用 “预条件范数”。

**定义 3.5** (预条件化). 考虑二次目标函数  $f(x) = \frac{1}{2}x^\top Ax - b^\top x$  ( $A \succ 0$ ), 其等高线为椭圆。通过坐标变换  $y = A^{1/2}x$ , 可将原椭圆等高线 “拉成” 圆形 (即消除椭圆的 “细长” 特性), 这一过程称为预条件化。

**直观意义:** 预条件化相当于将优化问题中的 “细长谷” 地形转化为 “圆形洼地”, 从而缓解负梯度下降时容易出现的 “楼梯形” 迂回路径, 提升迭代效率。

## 3.3 如何确定步长

设线搜索的核心目标函数为  $\phi(\alpha) = f(x + \alpha p)$ , 其中  $x$  为当前迭代点,  $p$  为已确定的搜索方向,  $\alpha \geq 0$  为待求解的步长 (需满足函数值下降条件  $f(x + \alpha p) < f(x)$ )。以下分别介绍三种主流线搜索方法:

### 3.3.1 精确线搜索 (Exact Line Search)

**定义 3.6** (精确线搜索). 精确线搜索直接求解“使  $\phi(\alpha)$  最小化”的步长  $\alpha^*$ , 数学表达为:

$$\alpha^* = \arg \min_{\alpha \geq 0} \phi(\alpha) \quad (3.3.1.1)$$

其目标是找到“当前方向下最优的步长”, 理论上能让单次迭代的函数值下降幅度最大。

**例 3.2** (二次函数的封闭解). 若目标函数为二次函数  $f(x) = \frac{1}{2}x^\top Ax - b^\top x$  (其中  $A \succ 0$ , 即  $A$  为正定矩阵), 且搜索方向  $p = -\nabla f(x) = -g$  ( $g = \nabla f(x)$  为当前梯度), 则精确线搜索的步长有封闭解:

$$\alpha^* = \frac{g^\top g}{g^\top Ag} \quad (3.3.1.2)$$

该解的本质是: 在二次函数的近似下, 使“更新后的梯度  $\nabla f(x + \alpha p)$  与搜索方向  $p$  垂直”(即一次更新在二次近似意义上最有效), 等价于求解  $\arg \min_{\alpha} \|g - \alpha Ag\|_2^2$ 。

**与固定步长的比较:** 在函数满足 **L-光滑性** (梯度 L-Lipschitz 连续) 的前提下, 精确线搜索得到的  $\alpha^*$  至少不劣于固定步长  $\alpha = 1/L$  (固定步长仅能保证“函数值下降”, 但无法保证下降幅度最优)。

### 3.3.2 黄金分割 (Golden Section Search)

**适用场景:** 当目标函数  $\phi(\alpha)$  是**单峰函数** (即区间内仅有一个最小值点), 且计算梯度 (或导数  $\phi'(\alpha)$ ) 代价高、难度大时, 采用黄金分割法通过“区间收缩”逼近最优步长  $\alpha^*$ 。

**核心流程:**

1. **初始区间设定:** 确定初始搜索区间  $[a, b]$ , 满足  $\phi(0) < \phi(b)$  (保证最小值点在区间内, 因  $\alpha = 0$  对应当前点, 函数值最大), 初始令  $a = 0$ ;
2. **区间收缩规则:**

- 在区间  $[a, b]$  内选取两个对称点  $t_1$  和  $t_2$  ( $a < t_1 < t_2 < b$ ), 两点间距与区间总长的比例为“黄金分割系数”  $c = \frac{1}{2}(\sqrt{5} - 1) \approx 0.618$ , 即:

$$t_1 = a + (1 - c)(b - a), \quad t_2 = a + c(b - a) \quad (3.3.2.1)$$

- **比较函数值:**

- 若  $\phi(t_2) > \phi(t_1)$ : 说明最小值点  $\alpha^* \in [a, t_2]$ , 令新区间为  $[a, t_2]$ ;
- 若  $\phi(t_1) > \phi(t_2)$ : 说明最小值点  $\alpha^* \in [t_1, b]$ , 令新区间为  $[t_1, b]$ ;

3. **迭代收敛:** 重复步骤 2, 不断收缩区间, 直到区间长度小于预设精度。最终区间内的任意点均可作为  $\alpha^*$  的近似值, 收敛误差上界与  $0.618^k$  成正比 ( $k$  为迭代次数)。

**特点:** 优点是无需计算梯度/导数, 仅通过函数值比较即可迭代; 缺点是收敛速度较慢 (线性收敛), 仅适用于单峰函数。

### 3.3.3 回溯搜索 (Backtracking Line Search) 与 Armijo-Wolfe 准则

回溯搜索通过“先试后调”的方式确定步长，需满足两个核心条件（保证步长既“足够大”以加速收敛，又“足够小”以保证函数值下降）：

**定义 3.7** (Armijo 条件 (充分下降条件))。确保步长能使函数值显著下降，数学表达为：

$$f(x + \alpha p) \leq f(x) + c_1 \alpha \nabla f(x)^\top p \quad (3.3.3.1)$$

其中  $0 < c_1 < 1$  (通常取  $c_1 = 10^{-4}$ )， $\nabla f(x)^\top p < 0$  (因  $p$  为下降方向)，右边项为函数值的“预期下降下限”。

**定义 3.8** (Wolfe 条件 (曲率条件))。确保步长不会过小 (避免收敛过慢)，数学表达为：

$$\nabla f(x + \alpha p)^\top p \geq c_2 \nabla f(x)^\top p \quad (3.3.3.2)$$

其中  $c_1 < c_2 < 1$  (通常取  $c_2 = 0.9$ )，该条件要求“更新后的梯度与搜索方向的内积”不小于“初始梯度与搜索方向内积”的  $c_2$  倍，避免步长停留在“函数值下降缓慢的区域”。

**回溯搜索算法流程：** 给定后退因子  $\beta \in (0, 1)$  (通常取  $\beta = 0.5$  或  $0.8$ )，步骤如下：

1. **初始步长尝试：** 令初始步长  $\alpha \leftarrow 1$  (默认从“单位步长”开始，适配 Newton 法等需要大步长的场景)；
2. **条件判断与步长调整：** 若当前  $\alpha$  不满足 Armijo 条件 (Armijo-Wolfe 联合条件)，则按比例缩小步长： $\alpha \leftarrow \beta \alpha$ ；
3. **终止：** 重复步骤 2，直到  $\alpha$  满足预设条件，输出最终步长  $\alpha$ 。

**关键性质与应用：**

- **终止性：** 由 Descent Lemma 可证明：当  $\alpha$  足够小时，Armijo 条件必成立，因此回溯搜索一定能终止；
- **与 Newton 法的结合 (两阶段收敛)：**
  - 阶段 I (远离最优解时)： $\alpha < 1$ ，通过回溯调整步长进入“可接受域” (满足 Armijo-Wolfe 条件)；
  - 阶段 II (靠近最优解时)：步长会触发  $\alpha = 1$  (单位步长)，此时 Newton 法可实现二次收敛 (收敛速度远快于梯度下降)。

## 3.4 收敛率：强凸 / PL 条件与“楼梯现象”

### 3.4.1 强凸 + L-光滑：线性收敛

**定义 3.9** (强凸与 L-光滑). 目标函数  $f(x)$  需同时满足两大性质:

1. **强凸性**: 存在常数  $\mu > 0$ , 对任意迭代点  $x$ , 其 *Hessian* 矩阵 (二阶导数矩阵) 满足下界约束:

$$\mu I \preceq \nabla^2 f(x) \quad (3.4.1.1)$$

(“强凸”保证函数有唯一最小值点, 且函数形态“下凸程度”可控);

2. **L-光滑性**: 存在常数  $L > 0$ , 对任意迭代点  $x$ , 其 *Hessian* 矩阵满足上界约束:

$$\nabla^2 f(x) \preceq LI \quad (3.4.1.2)$$

(“L-光滑”保证函数曲率不超过阈值, 梯度变化平缓, 避免局部剧烈波动)。

**定理 3.1** (线性收敛性). 当步长取  $\alpha = 1/L$  时, 梯度下降迭代满足严格的线性收敛性质:

1. 函数值下降界 (每次迭代函数值必递减且幅度可控):

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2 \quad (3.4.1.3)$$

2. 迭代点误差界 ( $x^*$  为函数最优解, 与最优解的距离按固定比例缩小):

$$\|x_{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|x_k - x^*\|_2^2 \quad (3.4.1.4)$$

证明: 线性收敛性.

#### 一、明确核心前提与已有结论

在证明前, 需明确 2 个关键性质 (强凸 + L-光滑) 的推论, 及 1 个已证结论:

1. **强凸性的核心推论**: 若函数  $f$  强凸 (存在  $\mu > 0$ , 使  $\mu I \preceq \nabla^2 f(x)$ ), 则对任意迭代点  $x$  与最优解  $x^*$  (满足  $\nabla f(x^*) = 0$ ), 有:

$$f(x) - f(x^*) \geq \frac{\mu}{2} \|x - x^*\|_2^2 \quad (1)$$

(强凸性保证“函数值与最优值的差距”不小于“迭代点与最优解距离平方”的固定倍数, 建立误差与函数值差的关联)

同时, 强凸性还可推出“梯度范数与函数值差的关系”: 因为  $\nabla f(x^*) = 0$ ,

$$f(x^*) \geq f(x) + \nabla f(x)^\top (x^* - x) + \frac{\mu}{2} |x^* - x|^2. \quad (3.4.1.5)$$

移项:

$$f(x) - f(x^*) \leq \nabla f(x)^\top (x - x^*) - \frac{\mu}{2} |x - x^*|^2. \quad (B)$$

由 *Cauchy-Schwarz*:

$$\nabla f(x)^\top (x - x^*) \leq |\nabla f(x)| \cdot |x - x^*|. \quad (3.4.1.6)$$

令右侧关于  $|x - x^*|$  的表达最小化, 可视为二次函数

$$|\nabla f(x)| \cdot |x - x^*| - \frac{\mu}{2} |x - x^*|^2. \quad (3.4.1.7)$$

其最大值出现在  $|x - x^*| = \frac{|\nabla f(x)|}{\mu}$ , 代入得

$$f(x) - f(x^*) \leq \frac{1}{2\mu} |\nabla f(x)|^2. \quad (3.4.1.8)$$

即:

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^*)) \quad (2)$$

(梯度大小能反映函数值与最优值的差距，为后续替换梯度项做准备)

2. 已证的函数值下降界：由  $L$ -光滑性（梯度  $L$ -Lipschitz 连续）及步长  $\alpha = 1/L$ ，已证明函数值满足：

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2 \quad (3)$$

二、步骤 1：推导函数值差的线性衰减关系

将式 (3) 变形为“相邻迭代的函数值差下界”：

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \|\nabla f(x_k)\|_2^2 \quad (3.4.1.9)$$

将强凸性推论式 (2) ( $\|\nabla f(x_k)\|_2^2 \geq 2\mu(f(x_k) - f(x^*))$ ) 代入上式，替换梯度范数项：

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \cdot 2\mu(f(x_k) - f(x^*)) \quad (3.4.1.10)$$

化简后得到“函数值差的衰减关系”：

$$f(x_k) - f(x_{k+1}) \geq \frac{\mu}{L} (f(x_k) - f(x^*)) \quad (3.4.1.11)$$

进一步整理，将函数值差聚焦到“与最优值的差距”：

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \frac{\mu}{L} (f(x_k) - f(x^*)) \quad (3.4.1.12)$$

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right) (f(x_k) - f(x^*)) \quad (4)$$

式 (4) 表明：迭代中“函数值与最优值的差距”按  $(1 - \mu/L)$  的比例线性衰减。

三、步骤 2：将函数值差衰减转化为迭代误差衰减

利用强凸性推论式 (1)，分别对  $x_{k+1}$  和  $x_k$  建立“迭代误差与函数值差的关联”：

- 对  $x_{k+1}$ ：

$$\|x_{k+1} - x^*\|_2^2 \leq \frac{2}{\mu} (f(x_{k+1}) - f(x^*)) \quad (5)$$

(由式 (1) 变形：两边同乘  $2/\mu$ ，不等号方向不变)

- 对  $x_k$ ：

$$f(x_k) - f(x^*) \geq \frac{\mu}{2} \|x_k - x^*\|_2^2 \implies \frac{2}{\mu} (f(x_k) - f(x^*)) \geq \|x_k - x^*\|_2^2 \quad (6)$$

四、步骤 3：合并推导得到迭代误差界

将式 (4) (函数值差衰减) 代入式 (5)，再结合式 (6) (函数值差与  $x_k$  误差的关联)：

$$\|x_{k+1} - x^*\|_2^2 \leq \frac{2}{\mu} \cdot \left(1 - \frac{\mu}{L}\right) (f(x_k) - f(x^*)) \quad (3.4.1.13)$$

由式 (6) 可知  $\frac{2}{\mu} (f(x_k) - f(x^*)) \geq \|x_k - x^*\|_2^2$ ，因此：

$$\|x_{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right) \cdot \frac{2}{\mu} (f(x_k) - f(x^*)) \leq \left(1 - \frac{\mu}{L}\right) \|x_k - x^*\|_2^2 \quad (3.4.1.14)$$

五、结论

最终证得迭代误差界公式：

$$\|x_{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|x_k - x^*\|_2^2 \quad (3.4.1.15)$$

该公式表明：强凸 +  $L$ -光滑条件下，梯度下降的迭代误差按  $(1 - \mu/L)$  的线性因子衰减，收敛速度由条件数  $\kappa = L/\mu$  决定—— $\kappa$  越大， $(1 - 1/\kappa)$  越接近 1，误差衰减越慢，且易因等高线“细长”出现“楼梯形”迂回路径。

**定义 3.10 (条件数)**. 定义条件数  $\kappa = L/\mu$  ( $L$  与  $\mu$  的比值)，线性收敛的“衰减速度”由  $\kappa$  决定：

- $\kappa$  越小 ( $L$  与  $\mu$  接近)：衰减因子  $(1 - 1/\kappa)$  越接近 0，收敛越快；
- $\kappa$  越大 ( $L$  远大于  $\mu$ )：衰减因子越接近 1，收敛越慢，且极易出现“楼梯现象”。

### 3.4.2 PL (Polyak-Lojasiewicz) 不等式

**定义 3.11** (PL 条件). 若存在常数  $\mu > 0$ , 对任意迭代点  $x$ , 函数值差与梯度范数满足以下关系:

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2 \quad (3.4.2.1)$$

( $f(x^*)$  为函数最小值, PL 条件是强凸性的“弱化版本”——无需函数严格强凸, 仅通过“函数值差距”与“梯度大小”的关联约束函数形态)。

**定理 3.2** (PL 条件下的收敛性). 即使  $f(x)$  非强凸, 只要满足 PL 条件, 梯度下降仍能实现线性型收敛, 函数值差的衰减公式为:

$$f(x_k) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k [f(x_0) - f(x^*)] \quad (3.4.2.2)$$

( $x_0$  为初始迭代点,  $k$  为迭代次数)。

**实际意义:** 解释了深度学习训练中的常见现象——深度网络的损失函数通常非强凸, 但可能满足 PL 条件, 因此训练时损失曲线会呈现“近似线性下降”的稳定趋势。

### 3.4.3 “楼梯现象”：成因与缓解

- **现象描述:** 当函数条件数  $\kappa = L/\mu$  过大时, 负梯度下降的迭代路径会呈现“楼梯形”: 沿细长的等高线 (如二次函数的椭圆等高线) 迂回前进, 每次仅沿等高线短轴方向小幅下降, 无法直接逼近最小值点, 迭代效率极低。
- **核心成因:** 条件数  $\kappa$  过大导致函数等高线“细长扁平”, 负梯度方向 (沿等高线法向) 与“最优下降方向” (沿等高线长轴方向) 偏差极大, 梯度下降陷入“来回震荡、缓慢逼近”的困境。
- **缓解方法:** 预条件化 (或输入/参数归一化)——通过线性变换 (如文档中二次函数的坐标变换  $y = A^{1/2}x$ ) 将“细长谷”地形转化为“圆形洼地”, 本质是减小条件数  $\kappa$ , 使负梯度方向更接近最优下降方向, 从而消除楼梯现象。

### 3.4.4 实践提示

1. **预条件化的核心价值:** 通过调整优化空间的度量规则 (如引入预条件矩阵), 显著降低条件数  $\kappa$ , 从根本上改善收敛速度;
2. **迭代停止准则:** 无需迭代至完全收敛, 满足以下任一条件即可终止:
  - 梯度范数足够小 (函数接近平稳):  $\|\nabla f(x_k)\|_2 \leq \varepsilon$  ( $\varepsilon$  为预设精度, 如  $10^{-6}$ );
  - 函数值相对下降不足 (继续迭代收益极低):  $\frac{f(x_{k-1}) - f(x_k)}{\max(1, f(x_{k-1}))} \leq \varepsilon$ 。

## 3.5 Newton 法：局部二次近似与两阶段收敛

Newton 法比梯度下降更高效的优化方法, 核心是通过函数的局部二次近似确定搜索方向, 兼具“局部快速收敛”与“全局有效下降”的特性, 其核心逻辑围绕“二阶展开 → 牛顿步 → 收敛性”展开。

### 3.5.1 核心思路：函数的局部二次近似

梯度下降仅用“一阶信息（梯度）”将函数局部近似为线性函数，而 Newton 法引入“二阶信息（Hessian 矩阵）”，将函数局部近似为二次函数（更贴合非凸函数的局部曲率）。

**定义 3.12** (局部二次近似). 对迭代点  $x$ ，将目标函数  $f(x + \Delta)$  在  $x$  处做二阶泰勒展开 ( $\Delta$  为搜索方向向量)：

$$f(x + \Delta) \approx f(x) + \nabla f(x)^\top \Delta + \frac{1}{2} \Delta^\top H(x) \Delta \quad (3.5.1.1)$$

其中：

- $\nabla f(x)$  是  $f(x)$  的梯度（一阶导数）；
- $H(x) = \nabla^2 f(x)$  是  $f(x)$  的 Hessian 矩阵（二阶导数矩阵），反映函数在  $x$  处的局部曲率。

Newton 法的核心是：最小化上述二次近似函数，直接求解使近似函数最小的搜索方向  $\Delta$ 。

### 3.5.2 牛顿步（Newton Step）的推导

对二阶近似函数关于  $\Delta$  求导，并令导数为 0（二次函数的极值点条件）：

$$\frac{\partial}{\partial \Delta} \left[ f(x) + \nabla f(x)^\top \Delta + \frac{1}{2} \Delta^\top H(x) \Delta \right] = \nabla f(x) + H(x) \Delta = 0 \quad (3.5.2.1)$$

**定义 3.13** (牛顿步). 若 Hessian 矩阵正定 ( $H(x) \succ 0$ ，保证二次近似函数是凸函数，极值点为最小值点)，则可解出唯一的搜索方向——牛顿步：

$$\Delta_{nt} = -H(x)^{-1} \nabla f(x) \quad (3.5.2.2)$$

### 3.5.3 牛顿步的下降性

牛顿步能保证是“下降方向”的前提是  $H(x) \succ 0$ ，证明如下：计算梯度与牛顿步的内积（判断方向是否下降的核心指标，内积  $< 0$  则为下降方向）：

$$\nabla f(x)^\top \Delta_{nt} = \nabla f(x)^\top (-H(x)^{-1} \nabla f(x)) \quad (3.5.3.1)$$

因  $H(x) \succ 0$ ，其逆矩阵  $H(x)^{-1}$  也正定，故对任意非零向量  $\nabla f(x)$ ，有  $\nabla f(x)^\top H(x)^{-1} \nabla f(x) > 0$ ，因此：

$$\nabla f(x)^\top \Delta_{nt} < 0 \quad (3.5.3.2)$$

即牛顿步满足“下降方向”的核心条件。

### 3.5.4 局部二次收敛：牛顿法的核心优势

当迭代点足够靠近最优解  $x^*$  时，Newton 法会呈现二次收敛（收敛速度远快于梯度下降的线性收敛）。

**定理 3.3** (牛顿法局部二次收敛). 若满足以下两个前提:

1. Hessian 矩阵  $H(x)$  在  $x^*$  的邻域内 **Lipschitz** 连续 (曲率变化平缓);
2. 初始迭代点  $x^{(0)}$  足够靠近  $x^*$  (进入 “局部收敛域”).

此时存在常数  $C > 0$ , 使得迭代误差满足:

$$\|x_{k+1} - x^*\| \leq C \cdot \|x_k - x^*\|^2 \quad (3.5.4.1)$$

证明: 局部二次收敛.

**结论:** 在  $H$  在某邻域内满足 Lipschitz (存在常数  $M$  使得  $|H(x) - H(y)| \leq M|x - y|$ ) 且  $H(x^*)$  非奇异的情形, 从足够近的初值出发, 牛顿迭代局部二次收敛, 即存在常数  $C > 0$  和半径  $r > 0$ , 当  $|x_k - x^*| \leq r$  时

$$|x_{k+1} - x^*| \leq C|x_k - x^*|^2. \quad (3.5.4.2)$$

**证明:** 设误差  $e_k := x_k - x^*$ . 由  $\nabla f(x^*) = 0$  和一维积分形式的泰勒公式:

$$\nabla f(x_k) = \int_0^1 H(x^* + te_k)e_k dt. \quad (3.5.4.3)$$

牛顿更新写作:

$$e_{k+1} = x_k - x^* - H(x_k)^{-1} \nabla f(x_k) = H(x_k)^{-1} \left( H(x_k) - \int_0^1 H(x^* + te_k) dt \right) e_k. \quad (3.5.4.4)$$

取范数并用三角不等式得:

$$|e_{k+1}| \leq |H(x_k)^{-1}| \int_0^1 |H(x_k) - H(x^* + te_k)| dt \cdot |e_k|. \quad (3.5.4.5)$$

利用 Hessian 的 Lipschitz 性质 (常数记为  $M$ ):

$$|H(x_k) - H(x^* + te_k)| \leq M|x_k - (x^* + te_k)| = M(1-t)|e_k|. \quad (3.5.4.6)$$

代入并对  $t$  积分:

$$|e_{k+1}| \leq |H(x_k)^{-1}| \cdot M \left( \int_0^1 (1-t) dt \right) |e_k|^2 = \frac{M}{2} |H(x_k)^{-1}| |e_k|^2. \quad (3.5.4.7)$$

由于  $H$  连续且  $H(x^*)$  非奇异, 存在半径  $r > 0$ , 使得对所有  $|x - x^*| \leq r$ ,  $H(x)$  可逆, 且  $|H(x)^{-1}| \leq B$  ( $B$  为该闭球上逆的上界). 因此当  $|e_k| \leq r$  时:

$$|e_{k+1}| \leq \frac{MB}{2} |e_k|^2. \quad (3.5.4.8)$$

令常数  $C := \frac{MB}{2}$ , 即得局部二次收敛估计。

**直观意义:** 二次收敛意味着 “每次迭代后, 误差的有效位数会翻倍” ——例如: 若第  $k$  步误差为  $10^{-2}$ , 第  $k+1$  步误差可降至  $10^{-4}$ , 第  $k+2$  步可降至  $10^{-8}$ , 接近最优解时收敛极快。

## 第四章 随机梯度下降 (Stochastic Gradient Descent, SGD)

### 4.1 随机梯度下降基础

当面对大规模数据集（数据量记为  $N$ ，单个数据为  $x_i, i = 1, \dots, N$ ），需要优化目标函数  $\min_x \sum_{i=1}^N f_i(x)$  时，若无法一次性获取所有数据  $x_i$  或对应函数  $f_i$ ，则可通过随机梯度下降（SGD）实现优化。

#### 4.1.1 核心思路：用“部分数据”估算梯度

由于无法计算全部数据的完整梯度  $\nabla f$ ，SGD 通过随机选取部分数据（称为“小批量”，记为  $\mathcal{B}^{(k)}$ ，其数据量记为  $|\mathcal{B}^{(k)}|$ ），用这部分数据的梯度平均值近似整体梯度，即：

$$\nabla f \approx \frac{1}{|\mathcal{B}^{(k)}|} \sum_{i \in \mathcal{B}^{(k)}} \nabla \ell_i(x^{(k)}) \quad (4.1.1.1)$$

其中  $\nabla \ell_i(x^{(k)})$  是单个数据  $i$  在当前参数  $x^{(k)}$  下的梯度，近似得到的整体梯度记为  $g^{(k)}$ ，即  $g^{(k)} = \frac{1}{|\mathcal{B}^{(k)}|} \sum_{i \in \mathcal{B}^{(k)}} \nabla \ell_i(x^{(k)})$ 。

#### 4.1.2 完整更新流程

SGD 的优化流程是在传统梯度下降（GD）基础上，修改“梯度计算方式”，具体步骤如下：

1. **初始值设定**：从初始参数  $x^{(0)}$  开始，迭代次数  $k = 0$ ；
2. **确定下降方向**：基于随机选取的小批量数据  $\mathcal{B}^{(k)}$ ，计算近似梯度  $g^{(k)}$ ，下降方向为  $\Delta x^{(k)} = -g^{(k)}$ （负梯度方向，保证函数值下降）；
3. **选择步长（学习率）**：步长  $\alpha^{(k)}$  可设为常数，也可随迭代次数动态调整（如后期逐步减小，避免参数震荡）；
4. **参数更新**：按以下公式更新参数，使新参数对应的函数值更小（即  $f(x^{(k+1)}) < f(x^{(k)})$ ），之后迭代次数  $k = k + 1$ ，重复步骤 2-4：

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} g^{(k)} \quad (4.1.2.1)$$

#### 4.1.3 关键超参数

SGD 的效果依赖两个核心超参数的设置，需根据数据和任务调整：

- **批量大小 (Batch Size):** 即小批量数据  $\mathcal{B}^{(k)}$  包含的数据量  $|\mathcal{B}^{(k)}|$ 。批量越大, 梯度估算越精准 (噪声越小), 但计算速度越慢; 批量越小, 计算越快, 但梯度噪声越大, 参数易震荡。
- **学习率 (Learning Rate):** 即步长  $\alpha^{(k)}$ 。学习率过大可能导致参数“越过”最优解, 函数值不下降反而上升; 学习率过小则参数更新缓慢, 需更多迭代次数才能收敛。

## 4.2 一个随机估计问题

先来看一个随机估计问题。

### 4.2.1 有限样本下的均值计算

- 当我们采样得到  $n$  个样本时, 均值可表示为:

$$f_n(x) = \frac{1}{n} \sum_{i \in [1, n]} f(x_i) \quad (4.2.1.1)$$

- 当继续采样到第  $n+1$  个样本时, 新的均值为:

$$f_{n+1}(x) = \frac{1}{n+1} \sum_{i \in [1, n+1]} f(x_i) \quad (4.2.1.2)$$

### 4.2.2 前后均值的递推关系

通过数学变形, 可建立  $f_n(x)$  与  $f_{n+1}(x)$  的关联, 避免重复计算所有样本:

$$\begin{aligned} f_{n+1}(x) &= \frac{1}{n+1} \left( f(x_{n+1}) + \sum_{i \in [1, n]} f(x_i) \right) \\ &= \frac{1}{n+1} (f(x_{n+1}) + n f_n(x)) \\ &= \left( 1 - \frac{1}{n+1} \right) f_n(x) + \frac{1}{n+1} f(x_{n+1}) \end{aligned} \quad (4.2.2.1)$$

若令步长  $\alpha = \frac{1}{n+1}$ , 则递推式可简化为更通用的形式:

$$f_{n+1}(x) = f_n(x) + \alpha (f(x_{n+1}) - f_n(x)) \quad (4.2.2.2)$$

这意味着新均值 = 旧均值 + 步长  $\times$  (新样本值 - 旧均值), 无需存储所有历史样本, 仅需保留旧均值即可更新。

### 4.2.3 均值收敛的条件

要保证当样本数量  $n \rightarrow \infty$  时, 均值  $f_n(x)$  能稳定收敛到真实期望, 需满足 Robbins-Monro (1951) 提出的步长条件:

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty \quad (4.2.3.1)$$

- 第一个条件  $\sum_{n=1}^{\infty} \alpha_n = \infty$ : 保证步长累积足够大, 均值能持续向真实期望靠近, 避免“半途停滞”;
- 第二个条件  $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$ : 保证步长衰减足够快, 避免后期新样本对均值的干扰过大, 导致结果震荡。

### 4.3 Robbins-Monro (RM) 算法

首先回顾 Robbins-Monro (RM) 算法的基础，它是推导 SGD 的起点。

#### 4.3.1 RM 算法的目标

RM 算法用于求解黑箱函数的根，即找到  $w^*$  满足：

$$g(w^*) = 0 \quad (4.3.1.1)$$

其中  $g: \mathbb{R}^d \rightarrow \mathbb{R}^d$  是未知函数（黑箱），仅能通过带噪声的观测获取信息：

$$\tilde{g}(w, \eta) = g(w) + \eta \quad (4.3.1.2)$$

$\eta$  是观测噪声，满足  $\mathbb{E}[\eta | H_k] = 0$  ( $H_k = \{w_k, w_{k-1}, \dots\}$  为历史信息)，且方差有界  $\mathbb{E}[\eta^2 | H_k] < \infty$ 。

#### 4.3.2 RM 算法的迭代公式

为求解  $g(w) = 0$ ，RM 算法的迭代更新规则为：

$$w_{k+1} = w_k - a_k \cdot \tilde{g}(w_k, \eta_k) \quad (4.3.2.1)$$

其中  $a_k > 0$  是步长序列， $w_k$  是第  $k$  次迭代的估计值。

#### 4.3.3 RM 算法的收敛条件

要保证  $w_k \rightarrow w^*$ （几乎必然收敛），需满足 3 个核心条件：

1. **函数单调性：** $g(w)$  单调递增，且梯度有界  $0 < c_1 \leq \nabla_w g(w) \leq c_2$ （确保根唯一）；
2. **步长条件：** $\sum_{k=1}^{\infty} a_k = \infty$ （步长不收敛太快，保证能逼近根）且  $\sum_{k=1}^{\infty} a_k^2 < \infty$ （步长趋于 0，避免震荡）；
3. **噪声条件：** $\mathbb{E}[\eta_k | H_k] = 0$  且  $\mathbb{E}[\eta_k^2 | H_k] < \infty$ （噪声无偏且方差有界）。

### 4.4 SGD 之问：为何能够收敛？

在此之前，我们先来看一个引理。

**引理 4.1** (Robbins-Siegmund 超鞅收敛引理). 给定非负可测序列  $(X_k)$ ，满足条件：

$$\mathbb{E}[X_{k+1} | \mathcal{F}_k] \leq (1 - a_k)X_k + b_k, \quad (4.4.0.1)$$

其中：

- $0 \leq a_k \leq 1$ ，控制“衰减比例”；
- $b_k \geq 0$ ，表示小的扰动或噪声；
- $\sum a_k = \infty$ ，保证长期衰减足够；
- $\sum b_k < \infty$ ，保证扰动总量有限。

结论：

1.  $(X_k)$  几乎处处收敛；
2.  $\sum a_k X_k < \infty$  几乎处处成立。

直观理解：

- $(X_k)$  类似“衰减量 + 小扰动”的随机过程；
- $(a_k)$  保证每步都有“收敛拉力”，而  $(b_k)$  干扰有限；
- 因此  $(X_k)$  不会发散，最终收敛，并且累计衰减量  $(\sum a_k X_k)$  有限。

Robbins-Siegmund 引理提供了在随机衰减 + 有限扰动下的序列收敛保证，是随机优化与在线算法理论分析的核心工具。

#### 4.4.1 设定与记号

- **数据与参数：**数据（或小批量数据）为  $x^{(k)}$ ，模型参数为  $\theta \in \mathbb{R}^d$ （ $d$  为参数维度）。
- **目标函数：**目标函数定义为期望损失，即  $f(\theta) \triangleq \mathbb{E}_x[L(x, \theta)]$ ，其中  $L(x, \theta)$  是单个数据（或小批量数据）的损失函数。
- **SGD 更新公式：**参数更新遵循  $\theta^{(k+1)} = \theta^{(k)} - \eta_k g^{(k)}$ ，其中  $\eta_k$  是第  $k$  步的学习率， $g^{(k)} \equiv \nabla_{\theta} L(x^{(k)}, \theta^{(k)})$  是第  $k$  步的随机梯度（基于小批量数据计算）。
- **噪声分解：**将随机梯度拆分为“真实梯度”与“噪声”两部分，即  $g^{(k)} = \nabla f(\theta^{(k)}) + \xi^{(k)}$ 。其中  $\nabla f(\theta^{(k)})$  是目标函数在  $\theta^{(k)}$  处的真实梯度， $\xi^{(k)}$  是随机噪声，且满足条件  $\mathbb{E}[\xi^{(k)} | \mathcal{F}_k] = 0$ （ $\mathcal{F}_k$  表示到第  $k$  步的所有观测信息集合，即“自然滤子”）。

这一分解恰好契合 **Robbins-Monro 框架**：该框架旨在寻找方程  $h(\theta) = 0$  的根（即目标函数极小值点，此时  $\nabla f(\theta^*) = 0$ ），但仅能获得带噪声的观测  $H(\theta, x)$ （对应此处的随机梯度  $g^{(k)}$ ），且观测的期望等于真实函数（即  $\mathbb{E}[g^{(k)} | \mathcal{F}_k] = \nabla f(\theta^{(k)})$ ）。令  $h(\theta) = \nabla f(\theta)$ ，即可将 SGD 纳入该框架分析收敛性。

#### 4.4.2 收敛性证明的核心假设

要证明 SGD 收敛，需满足以下 5 个关键假设（记  $\theta^*$  为目标函数极小值点，即  $\nabla f(\theta^*) = 0$ ）：

- **(A1) 无偏噪声：**随机梯度的条件期望等于真实梯度，即  $\mathbb{E}[g^{(k)} | \mathcal{F}_k] = \nabla f(\theta^{(k)})$ 。
- **(A2) 有界二阶矩：**噪声的条件二阶矩有上限，即  $\mathbb{E}[\|\xi^{(k)}\|^2 | \mathcal{F}_k] \leq \sigma^2 + c \|\nabla f(\theta^{(k)})\|^2$ 。其中  $\sigma^2$  是常数， $c$  是系数，该假设限制了噪声的“强度”，避免噪声过大导致参数震荡不收敛。常用特例为“常数方差”，即  $\mathbb{E}[\|\xi^{(k)}\|^2 | \mathcal{F}_k] \leq \sigma^2$ 。
- **(A3) L 平滑：**目标函数的梯度满足 Lipschitz 连续条件，即  $\|\nabla f(\theta) - \nabla f(\phi)\| \leq L \|\theta - \phi\|$ （ $L$  为 Lipschitz 常数）。
- **(A4)  $\mu$ -强凸：**目标函数是  $\mu$ -强凸的，即  $(\nabla f(\theta) - \nabla f(\phi))^\top (\theta - \phi) \geq \mu \|\theta - \phi\|^2$ （ $\mu > 0$  为强凸系数）。强凸性保证目标函数有唯一极小值点  $\theta^*$ ，且参数会“持续向极小值点靠近”，不会在多个局部极小值间徘徊。

• **(A5) Robbins-Monro 步长条件：**学习率序列  $\{\eta_k\}$  需满足两个条件：

1.  $\sum_{k=1}^{\infty} \eta_k = \infty$ （学习率累积和为无穷大）：保证参数有足够的“推进力”，能持续向极小值点靠近，避免因步长过小而“半途停滞”；
2.  $\sum_{k=1}^{\infty} \eta_k^2 < \infty$ （学习率平方的累积和有限）：保证后期步长足够小，避免参数在极小值点附近“来回震荡”。

典型的满足该条件的学习率形式为  $\eta_k = \frac{\alpha}{k+\beta}$  ( $\alpha > 0, \beta \geq 0$ )。

#### 4.4.3 收敛性结论

结论一：几乎处处收敛（基于 Robbins-Siegmund 引理）

**定理 4.1** (a.s. 收敛). 在假设 (A1)-(A5) 成立的前提下, SGD 生成的参数序列满足:

$$\theta^{(k)} \xrightarrow[k \rightarrow \infty]{a.s.} \theta^*, \quad \sum_{k=1}^{\infty} \eta_k \|\nabla f(\theta^{(k)})\|^2 < \infty \quad a.s. \quad (4.4.3.1)$$

其中“a.s.”表示“几乎必然”（即除了概率为 0 的特殊情况外，参数序列一定收敛到  $\theta^*$ ）。

**证明核心思路（关键不等式与引理应用）：**

1. **定义距离变量：**令  $\Delta^{(k)} \triangleq \theta^{(k)} - \theta^*$ （即当前参数与极小值点的距离向量），需证明  $\|\Delta^{(k)}\| \rightarrow 0$ （距离趋近于 0）。
2. **展开距离平方的递推关系：**根据 SGD 更新公式，展开  $\|\Delta^{(k+1)}\|^2$ （第  $k+1$  步的距离平方）：

$$\begin{aligned} \|\Delta^{(k+1)}\|^2 &= \|\theta^{(k+1)} - \theta^*\|^2 \\ &= \|\theta^{(k)} - \eta_k g^{(k)} - \theta^*\|^2 \\ &= \|\Delta^{(k)} - \eta_k g^{(k)}\|^2 \\ &= \|\Delta^{(k)}\|^2 - 2\eta_k \Delta^{(k)\top} g^{(k)} + \eta_k^2 \|g^{(k)}\|^2 \end{aligned} \quad (4.4.3.2)$$

3. **取条件期望并代入假设：**对等式两侧关于  $\mathcal{F}_k$  取条件期望，结合 (A1)（无偏噪声）和 (A2)（有界二阶矩），将  $g^{(k)} = \nabla f(\theta^{(k)}) + \xi^{(k)}$  代入，可化简得到：

$$\begin{aligned} \mathbb{E} [\|\Delta^{(k+1)}\|^2 | \mathcal{F}_k] &\leq \mathbb{E} [\|\Delta^{(k)}\|^2 | \mathcal{F}_k] - 2\eta_k \Delta^{(k)\top} \nabla f(\theta^{(k)}) \\ &\quad + \eta_k^2 (\|\nabla f(\theta^{(k)})\|^2 + \sigma^2 + c \|\nabla f(\theta^{(k)})\|^2) \end{aligned} \quad (4.4.3.3)$$

4. **利用强凸与 L 平滑简化：**由 (A4)（ $\mu$ -强凸）可得  $\Delta^{(k)\top} \nabla f(\theta^{(k)}) \geq \mu \|\Delta^{(k)}\|^2$ ；由 (A3)（L 平滑）可得  $\|\nabla f(\theta^{(k)})\| \leq L \|\Delta^{(k)}\|$ （因  $\nabla f(\theta^*) = 0$ ）。代入上式后，可整理得到：

$$\mathbb{E} [\|\Delta^{(k+1)}\|^2 | \mathcal{F}_k] \leq (1 - 2\mu\eta_k + C\eta_k^2) \|\Delta^{(k)}\|^2 + \sigma^2\eta_k^2 \quad (4.4.3.4)$$

其中  $C \triangleq (1+c)L^2$ （常数）。当  $k$  足够大时， $\eta_k$  足够小，可满足  $1 - 2\mu\eta_k + C\eta_k^2 \leq 1 - \mu\eta_k$ 。

5. **应用 Robbins-Siegmund 引理：**令  $X_k = \|\Delta^{(k)}\|^2$ （待分析的非负序列）， $a_k = \mu\eta_k$ ， $b_k = \sigma^2\eta_k^2$ ，则上述不等式可化为引理要求的形式：

$$\mathbb{E} [X_{k+1} | \mathcal{F}_k] \leq (1 - a_k) X_k + b_k \quad (4.4.3.5)$$

结合 (A5),  $\sum a_k = \mu \sum \eta_k = \infty$ ,  $\sum b_k = \sigma^2 \sum \eta_k^2 < \infty$ , 满足引理条件。根据引理可得出:  $X_k$  几乎必然收敛, 且  $\sum a_k X_k < \infty$ 。再结合  $\sum a_k = \infty$ , 可推出  $\liminf_{k \rightarrow \infty} X_k = 0$ ; 又因目标函数强凸 ((A4)), 最终可得  $X_k \rightarrow 0$  (即  $\theta^{(k)} \rightarrow \theta^*$ ) 几乎必然成立。

## 结论二：强凸下的收敛速率与 Polyak-Ruppert 平均

**定理 4.2** (期望二次误差  $O(1/k)$ )。假设 (A1)-(A4) 成立, 若取学习率  $\eta_k = \frac{\alpha}{k+\beta}$  (其中  $\alpha > \frac{1}{\mu}$ ,  $\beta \geq 1$ ), 则存在常数  $K$ , 使得:

$$\mathbb{E} \left[ \|\theta^{(k)} - \theta^*\|^2 \right] \leq \frac{K}{k+\beta} \quad (4.4.3.6)$$

即参数与极小值点的“期望平方距离”随迭代次数  $k$  增长, 以  $O(1/k)$  的速率衰减。

证明。

将结论一证明中的“距离平方的条件期望递推式”取全期望, 令  $u_k = \mathbb{E} \left[ \|\Delta^{(k)}\|^2 \right]$  (期望平方距离), 代入  $\eta_k = \frac{\alpha}{k+\beta}$  后可得到:

$$u_{k+1} \leq \left( 1 - \frac{2\mu\alpha}{k+\beta} + \frac{C\alpha^2}{(k+\beta)^2} \right) u_k + \frac{\sigma^2\alpha^2}{(k+\beta)^2} \quad (4.4.3.7)$$

通过定义辅助变量  $v_k = (k+\beta)u_k$ , 利用“差分比较法”可证明  $u_k = O(1/k)$ , 进而得到上述期望误差界。

随着迭代次数  $k$  增加, 参数与最优解的“平均距离平方”会以  $1/k$  的速度变小。

**定理 4.3** (Polyak-Ruppert 迭代平均的最优渐近方差)。定义参数的迭代平均为:

$$\bar{\theta}^{(T)} \triangleq \frac{1}{T} \sum_{k=1}^T \theta^{(k)} \quad (4.4.3.8)$$

在假设 (A1)-(A4) 与 (A5) (步长  $\eta_k = \frac{\alpha}{k+\beta}$ ) 成立的前提下, 有:

$$\sqrt{T} \left( \bar{\theta}^{(T)} - \theta^* \right) \Rightarrow \mathcal{N} \left( 0, A^{-1} S A^{-\top} \right) \quad (4.4.3.9)$$

其中  $A \triangleq \nabla^2 f(\theta^*)$  (目标函数在极小值点处的 Hessian 矩阵),  $S$  是噪声协方差的极限值。

该结论表明: 通过对参数序列做“迭代平均”, 可使 SGD 达到随机逼近 (SA) 框架下的“最优渐近效率”——即平均后的参数估计量, 其渐近方差是最小的, 在实践中能显著减小噪声导致的参数波动, 提升收敛稳定性。

普通 SGD 是“每步更新一个参数, 最后用最后一步的参数”; 而 Polyak-Ruppert 方法是“先迭代  $T$  步, 得到  $T$  个参数  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(T)}$ , 再求它们的平均值  $\bar{\theta}^{(T)} = \frac{1}{T} \sum_{k=1}^T \theta^{(k)}$ ”。

用“迭代平均”后的参数  $\bar{\theta}^{(T)}$ , 随着迭代次数  $T$  增加, 它与最优解的差距会服从“正态分布”, 且这个差距的“波动范围 (方差) 是最小的” (即“最优渐近方差”)。

SGD 的收敛性本质上源于 Robbins-Monro 随机逼近框架与 Robbins-Siegmund 超鞅引理的支撑:

1. 强凸目标函数 + 满足 Robbins-Monro 条件的学习率 (如  $\eta_k \propto 1/k$ ), 可保证参数“几乎处处收敛”到极小值点, 且期望二次误差以  $O(1/k)$  速率衰减;
2. 对参数做 Polyak-Ruppert 迭代平均, 能进一步优化渐近方差, 提升收敛精度与稳定性。

#### 4.4.4 非强凸场景下的 SGD 收敛性（仅凸/一般非凸）

在之前的分析中，我们默认目标函数满足“ $\mu$ -强凸”条件（假设 (A4)），但实际场景中很多目标函数不具备强凸性（如仅凸函数、非凸函数），因此需要单独分析这类场景下 SGD 的收敛表现。

##### 1. 仅凸场景（无强凸性，仅满足凸性）

**核心设定：**此时目标函数可能存在“平坦区域”或“多个最优解（构成凸集）”，无法保证参数收敛到唯一极小值点，但可保证“函数值收敛到最优值”。

**收敛性结论：**若调整学习率为  $\eta_k = \frac{1}{k^\alpha}$ （其中  $\alpha \in (1/2, 1]$ ，满足 Robbins-Monro 条件  $\sum \eta_k = \infty$  且  $\sum \eta_k^2 < \infty$ ），则对参数的迭代平均值  $\bar{\theta}^{(T)} = \frac{1}{T} \sum_{k=1}^T \theta^{(k)}$ ，有：

$$f(\bar{\theta}^{(T)}) - f(\theta^*) = o(1) \quad (4.4.4.1)$$

即随着迭代次数  $T$  增大，平均参数对应的函数值会“逐步逼近最优函数值  $f(\theta^*)$ ”，最终趋近于 0。

若进一步量化收敛速率，通常为  $O(\frac{1}{T^{1-\alpha}})$  量级（如  $\alpha = 0.8$  时，速率为  $O(\frac{1}{T^{0.2}})$ ）——相比强凸场景下的  $O(\frac{1}{T})$ ，仅凸场景的收敛更慢，这是因为缺少强凸性带来的“强制向最优解靠近”的约束。

##### 2. 一般非凸场景（无凸性，仅满足 L-平滑）

**核心设定：**“一般非凸”指目标函数既不满足强凸性，也不满足凸性，仅满足 L-平滑条件（假设 (A3)：梯度变化平缓， $\|\nabla f(\theta) - \nabla f(\phi)\| \leq L\|\theta - \phi\|$ ）。这类场景在深度学习中最常见（如神经网络的损失函数），目标函数可能存在大量局部极小值、鞍点，无法保证参数收敛到全局最优解，只能退而求其次——保证参数收敛到“一阶驻点”（即梯度趋近于 0 的点， $\nabla f(\theta) \approx 0$ ，此时参数再更新也难以显著降低函数值）。

**收敛性结论：**在无偏噪声（假设 (A1)）和噪声方差有界（假设 (A2)）的前提下，若采用“分段常数步长”或“ $\eta_k = \frac{1}{\sqrt{k}}$  步长”（注： $\frac{1}{\sqrt{k}}$  不满足 Robbins-Monro 的  $\sum \eta_k^2 < \infty$ ，因此不具备“几乎处处收敛”性质，仅能保证“梯度的期望有界”），则有：

$$\min_{1 \leq k \leq T} \mathbb{E} [\|\nabla f(\theta^{(k)})\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \quad (4.4.4.2)$$

该结论的含义是：在  $T$  次迭代中，至少存在某一步的参数  $\theta^{(k)}$ ，其梯度的期望平方值不超过  $\frac{C}{\sqrt{T}}$ （ $C$  为常数），且随着  $T$  增大，这个“最小梯度期望”会以  $\frac{1}{\sqrt{T}}$  的速率减小，逐步趋近于 0。

需要特别注意：

1. 该速率是“到一阶驻点”的速率，而非“到全局最优解”的速率——最终参数可能停在局部极小值或鞍点，但这些点的梯度已足够小，函数值难以继续下降；
2. 与强凸/仅凸场景不同，非凸场景的收敛结论不涉及“参数是否收敛”或“函数值是否收敛到最优”，仅保证“梯度足够小”，这是因为非凸函数的全局最优解难以通过 SGD 的随机搜索触及，“找到驻点”已是实际能达到的目标。

##### 3. 非强凸场景与强凸场景的核心差异

为了更清晰理解不同场景的收敛特性，可通过下表对比：

场景	目标函数性质	收敛目标	学习率要求	收敛速率（期望）	关键限制
强凸	强凸 + L-平滑	全局最优解 $\theta^*$	$\eta_k \propto \frac{1}{k}$ (满足 RM)	$O(1/T)$	需强凸性，适用场景有限
仅凸	凸 + L-平滑	最优函数值 $f(\theta^*)$	$\eta_k = \frac{1}{k^\alpha}$ ( $\alpha \in (0.5, 1]$ )	$O(1/T^{1-\alpha})$	收敛慢，无唯一最优参数
一般非凸	L-平滑 (无凸性)	一阶驻点 ( $\nabla f \approx 0$ )	分段常数/ $\eta_k \propto \frac{1}{\sqrt{k}}$	$O(1/\sqrt{T})$ (梯度期望)	仅能到驻点，可能是局部最优

表 4.1: 不同场景下的收敛特性对比

#### 4.4.5 两种不同目标下的步长设计及收敛策略差异

##### 1. OGD/Regret（在线学习/长期平均性能）

- 目标：保证长期平均损失接近最优，即 后悔 (regret) 界小。
- 对应公式常见形式：

$$\text{Regret}(T) = \sum_{t=1}^T f_t(x_t) - \min_x \sum_{t=1}^T f_t(x) \leq O(\sqrt{T}) \text{ 或 } O(\log T) \quad (4.4.5.1)$$

- 步长选择：通常用 非递减 或者  $1/\sqrt{t}$  形式，保证平均损失下降快。

##### 2. RM/SA (Robbins-Monro / Stochastic Approximation)

- 目标：保证参数序列几乎处处收敛到最优点 (a.s. convergence)，属于点估计/统计意义。
- 收敛条件：

$$\sum_{k=1}^{\infty} a_k = \infty, \quad \sum_{k=1}^{\infty} a_k^2 < \infty \quad (4.4.5.2)$$

- 常用步长： $a_k = 1/k$  或  $1/k^\gamma$  ( $0.5 < \gamma \leq 1$ )。
- 意义：每步衰减足够慢以保证探索，但衰减快以抑制噪声，满足 Robbins-Siegmund 引理条件。
- 如果使用 OGD/Regret 的步长策略来保证 几乎处处收敛，可能违反 RM/SA 的平方可积条件 ( $\sum < \infty$ )，因此不能保证 a.s. 收敛。
- 反之，如果严格使用 RM/SA 的条件 ( $\sum < \infty$ ) 来优化在线 regret，可能收敛太慢，导致平均损失下降慢。

关键点：两种方法的目标不同，不能直接互换步长策略。

##### 3. 折中 / 统一策略

选用  $a_k = 1/k^\gamma$  ( $0.5 < \gamma < 1$ )

- 满足 RM/SA 条件： $\sum a_k = \infty$  且  $\sum a_k^2 < \infty$ ，保证 a.s. 收敛。
- 同时保持较慢衰减，平均性能也不错（在线学习效果可接受）。

阶段常数步长 + Polyak-Ruppert 平均 + Doubling Trick

- 阶段常数步长：将迭代分阶段，每阶段使用近似最优常数步长，提升该阶段的平均损失性能（降低 regret）。

- **Doubling Trick**: 阶段长度每次加倍 (doubling trick), 保证整体步长衰减满足 RM/SA 条件, 确保 a.s. 收敛。
- **Polyak-Ruppert 平均**: 阶段内取参数平均, 进一步稳定收敛。

注 4.1. *OGD/Regret* 关注长期平均损失; *RM/SA* 关注参数几乎处处收敛。两者步长策略冲突, 但可以通过衰减指数、阶段常数步长和 *Polyak-Ruppert* 平均实现折中, 兼顾在线性能和几乎处处收敛。

## 4.5 从随机估计到动力学

从动力学视角分析 SGD, 核心是将“离散参数更新过程”与“连续的物理运动方程”建立关联——通过极限近似, 把梯度下降 (GD) 对应到确定性的常微分方程 (ODE), 把随机梯度下降 (SGD) 对应到含噪声的随机微分方程 (SDE), 从而用物理运动规律解释 SGD 的收敛行为、噪声影响及参数调整逻辑。

### 4.5.1 从 GD 到 ODE: 离散更新是梯度流的“显式欧拉积分”

梯度下降 (GD) 的参数更新是离散步骤, 而通过“时间标度转换”和“连续极限”, 可将其转化为描述“确定性下降运动”的常微分方程 (ODE), 即“梯度流”。

#### 1.1 离散更新与时间标度定义

GD 的离散更新公式为:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla L(\theta^{(k)}) \quad (4.5.1.1)$$

其中:

- $\theta^{(k)}$ : 第  $k$  步的参数;
- $\eta$ : 步长 (学习率);
- $\nabla L(\theta^{(k)})$ : 目标函数  $L$  在  $\theta^{(k)}$  处的梯度 (确定性, 无噪声)。

为建立连续关联, 定义“连续时间”  $t_k = k \cdot \eta$ ——即把每一步更新的“步长  $\eta$ ”视为“时间增量”, 迭代次数  $k$  越多, 对应的连续时间  $t_k$  越大。

#### 1.2 连续极限: 从离散更新到梯度流 ODE

当步长  $\eta \rightarrow 0$  (时间增量无限小)、且  $t_k \rightarrow t$  (连续时间趋近于某个值) 时, 对 GD 的离散更新公式做“差分近似”: 左边参数增量除以时间增量, 近似为连续时间下的参数变化率 (导数):

$$\frac{\theta^{(k+1)} - \theta^{(k)}}{\eta} \Rightarrow \dot{\theta}(t) \quad (4.5.1.2)$$

右边代入 GD 的更新规则, 可得连续时间下的“梯度流方程” (ODE):

$$\dot{\theta}(t) = -\nabla L(\theta(t)) \quad (4.5.1.3)$$

**物理意义**: GD 的离散更新, 本质是对“梯度流 ODE”的“显式欧拉数值积分”——每一步按当前梯度方向“迈一小步”, 步长越小, 离散参数轨迹越贴近 ODE 描述的“连续下降路径” (类似下山时“小步慢走”更贴近顺滑的山坡轨迹)。

### 1.3 数值稳定性与曲率的关系

GD 的收敛稳定性（是否会“震荡不收敛”），与目标函数的“曲率”直接相关，可通过二次函数案例直观理解：

- 若目标函数为二次形式  $L(\theta) = \frac{1}{2}\theta^\top H\theta$  ( $H \succeq 0$  为 Hessian 矩阵，代表函数曲率)，则 GD 的更新公式可改写为：

$$\theta^{(k+1)} = (I - \eta H)\theta^{(k)} \quad (4.5.1.4)$$

其中  $I$  为单位矩阵。

- 收敛条件：该线性迭代收敛的充要条件是“矩阵  $I - \eta H$  的谱半径  $\rho(I - \eta H) < 1$ ”，等价于步长需满足：

$$0 < \eta < \frac{2}{\lambda_{\max}(H)} \quad (4.5.1.5)$$

( $\lambda_{\max}(H)$  是 Hessian 矩阵的最大特征值，代表函数的“最大曲率”)。

- 一般 L-平滑场景：若目标函数的梯度满足 L-Lipschitz 连续 (L 为平滑常数，可理解为“梯度变化的最大速率”)，则取  $0 < \eta < \frac{2}{L}$  可保证每步更新后函数值下降；若同时满足强凸性，取  $0 < \eta \leq \frac{1}{L}$  还能获得“线性收敛速率”（参数快速靠近最优解）。

**核心启发 (A)：**可将学习率  $\eta$  视为“时间步长”——函数曲率越大 ( $\lambda_{\max}(H)$  或  $L$  越大)，“显式欧拉积分”的稳定范围越窄，GD 需要更小的学习率才能避免震荡；实际中“分段调整学习率”“周期衰减学习率”，本质是通过“细化时间网格”提升数值稳定性，让参数更新更贴合梯度流的顺滑路径。

## 4.5.2 从 SGD 到 SDE：扩散极限与朗之万动力学

SGD 的核心是“用随机小批量梯度近似真实梯度”，存在噪声干扰。通过类似的连续极限，可将其转化为含噪声的随机微分方程 (SDE)，即“朗之万动力学”，从而用“扩散运动”解释 SGD 的噪声探索与收敛平衡。

### 2.1 噪声分解：随机梯度的构成

SGD 的小批量梯度包含“真实梯度”和“噪声”两部分，分解公式为：

$$\nabla L_{\mathcal{B}}(\theta^{(k)}) = \nabla L(\theta^{(k)}) + \xi^{(k)} \quad (4.5.2.1)$$

其中：

- $\nabla L_{\mathcal{B}}(\theta^{(k)})$ ：基于小批量  $\mathcal{B}$  计算的随机梯度；
- $\nabla L(\theta^{(k)})$ ：目标函数的真实梯度（确定性部分）；
- $\xi^{(k)}$ ：小批量采样引入的噪声，满足  $\mathbb{E}[\xi^{(k)}|\theta^{(k)}] = 0$ （无偏噪声），其协方差  $\text{Cov}[\xi^{(k)}] \approx \Sigma(\theta^{(k)})$ （随参数变化的噪声强度）。

基于此，SGD 的离散更新公式可改写为：

$$\theta^{(k+1)} = \theta^{(k)} - \eta (\nabla L(\theta^{(k)}) + \xi^{(k)}) \quad (4.5.2.2)$$

## 2.2 扩散极限：从离散 SGD 到 SDE（欧拉-丸山连续化）

同样定义连续时间  $t_k = k \cdot \eta$ ，当步长  $\eta \rightarrow 0$ （时间增量无限小）、且小批量噪声近似高斯分布时，可将 SGD 的离散更新转化为“随机微分方程（SDE）”：

$$d\theta_t = -\nabla L(\theta_t)dt + G(\theta_t)dW_t \quad (4.5.2.3)$$

其中：

- $\theta_t$ ：连续时间  $t$  下的参数；
- $dt$ ：连续时间增量；
- $dW_t$ ：多维布朗运动（Wiener 过程），代表连续时间下的随机噪声（均值为 0，方差为  $dt$ ）；
- $G(\theta_t)$ ：噪声强度矩阵，满足  $G(\theta_t)G(\theta_t)^\top \approx \eta \cdot \Sigma(\theta_t)$ （将离散噪声的协方差与连续时间的噪声强度关联）。

**规范朗之万形式：**若令噪声强度为“各向同性常数”（即不同参数方向的噪声强度相同），设  $G = \sqrt{2T}$ （ $T$  为“温度”参数，控制噪声整体强度），则 SDE 可简化为标准的“朗之万动力学方程”：

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2T}dW_t \quad (4.5.2.4)$$

其核心性质是：若存在平稳分布（参数长期运动的稳定概率分布），则该分布与目标函数  $L$  的 Gibbs 权重成正比，即  $\propto \exp(-\frac{L}{T})$ ——“温度” $T$  越高，噪声越强，参数探索范围越广（更易跳出局部极小值）； $T$  越低，噪声越弱，参数越容易收敛到目标函数的低价值区域（极小值附近）。

## 2.3 两类极限：消噪极限与扩散极限

SGD 的连续极限存在两种典型场景，对应不同的训练阶段目标：

- **消噪极限：**若步长  $\eta \rightarrow 0$ ，同时小批量大小  $B$  增大（使噪声协方差  $\Sigma \rightarrow 0$ ），则 SDE 中的扩散项（噪声部分） $G(\theta_t)dW_t$  会逐渐消失，SDE 退化为 GD 对应的“梯度流 ODE”——这对应训练后期“增大批量、减小学习率”的策略，目的是“消除噪声，精准收敛到最优解”。
- **扩散极限：**若按比例调整步长  $\eta$  和批量大小  $B$ （如保持  $\frac{\eta}{B}$  为常数），使“有效噪声强度”（ $\eta \cdot \Sigma$ ）保持不变，则 SDE 的扩散项非平凡（噪声持续存在）——这对应训练前期“小批量、稍大学习率”的策略，目的是“保留噪声，通过随机探索找到更优的参数区域”。

## 2.4 Fokker-Planck 视角：参数分布的演化

SGD 的参数在连续时间下的概率密度  $p_t(\theta)$ （即参数在时刻  $t$  处于某个值的概率），满足“Fokker-Planck 方程”：

$$\partial_t p_t = \nabla \cdot (p_t \nabla L) + \frac{1}{2} \sum_{i,j} \partial_i \partial_j ([D(\theta)]_{ij} p_t) \quad (4.5.2.5)$$

其中  $D(\theta) = G(\theta)G(\theta)^\top$  是扩散系数矩阵（代表噪声在不同参数方向的强度）。

该方程的意义是：参数密度的变化由两部分驱动——

1. 确定性漂移项（ $\nabla \cdot (p_t \nabla L)$ ）：由目标函数梯度主导，使参数密度向  $L$  的低价值区域聚集（类似水流向低处）；
2. 随机扩散项（二阶导数项）：由噪声主导，使参数密度向周围扩散（类似墨水在水中扩散）。

若  $D(\theta)$  为常数且各向同性（噪声在所有参数方向强度相同），则平稳密度为 Gibbs 分布；若  $D(\theta)$  随参数变化或各向异性（不同方向噪声强度不同），则平稳密度会偏离简单的 Gibbs 分布——这解释了实际 SGD 中“噪声具有方向性”的现象：某些参数方向的噪声更强，参数在这些方向的探索更活跃，最终收敛位置也会偏向噪声影响更小的“平坦区域”（与“平坦极小值泛化更好”的经验观察一致）。

### 4.5.3 局部二次近似与 OU 过程：常步长下的方差-曲率权衡

在目标函数的极小值点  $\theta^*$  附近，可将函数近似为二次形式（局部二次近似），此时 SGD 的连续极限（SDE）可简化为“Ornstein-Uhlenbeck（OU）过程”——通过分析 OU 过程的平稳分布，能清晰理解“参数曲率”与“噪声方差”的平衡关系。

#### 3.1 局部二次近似

在  $\theta^*$  附近，对目标函数  $L(\theta)$  做泰勒展开并忽略高阶项，得到二次近似：

$$L(\theta) \approx L(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H(\theta - \theta^*) \quad (4.5.3.1)$$

其中  $H = \nabla^2 L(\theta^*)$  是目标函数在  $\theta^*$  处的 Hessian 矩阵（ $H \succ 0$ ，因  $\theta^*$  是极小值点），代表函数在极小值附近的“局部曲率”—— $H$  的特征值越大，对应参数方向的曲率越大（函数在该方向越“陡峭”）。

#### 3.2 OU 过程与平稳协方差

令  $\vartheta_t = \theta_t - \theta^*$ （参数与极小值点的偏差），代入朗之万 SDE，结合局部二次近似（ $\nabla L(\theta_t) \approx H\vartheta_t$ ），可得偏差  $\vartheta_t$  满足的 OU 过程：

$$d\vartheta_t = -H\vartheta_t dt + \sqrt{2T}dW_t \quad (4.5.3.2)$$

OU 过程是“带阻尼的线性随机过程”，其核心性质是存在平稳分布（当时间  $t \rightarrow \infty$  时， $\vartheta_t$  的分布不再变化）：

- 平稳分布为高斯分布  $\mathcal{N}(0, P)$ ，其中  $P$  是协方差矩阵，满足“Lyapunov 方程”：

$$HP + PH = 2TI \quad (4.5.3.3)$$

（ $I$  为单位矩阵， $T$  为温度参数）。

- 若噪声为各向异性常数扩散（ $D = GG^\top$ ，非单位矩阵），则 Lyapunov 方程推广为：

$$HP + PH = D \quad (4.5.3.4)$$

#### 3.3 核心启发（B）：“宽谷偏好”的物理理解

从 Lyapunov 方程可直接推导“曲率”与“平稳方差”的关系：对 Hessian 矩阵  $H$  的某个特征值  $\lambda_i$ （对应第  $i$  个参数方向的曲率），其对应的平稳方差  $P_{ii}$ （参数在该方向的波动范围）满足：

$$P_{ii} = \frac{T}{\lambda_i} \quad (4.5.3.5)$$

这意味着：在相同噪声强度（温度  $T$ ）下，曲率越小（ $\lambda_i$  越小）的参数方向，平稳方差越大——即目标函数的“宽谷区域”（曲率小）对参数的“吸引概率”更高，参数更易在宽谷中稳定下来。

这一结论完美解释了深度学习中的经验观察：“更平坦的极小值泛化性能更好”——因为 SGD 的噪声会使参数自然偏向宽谷区域，而宽谷区域的参数对数据扰动更不敏感，泛化能力更强。

#### 4.5.4 学习率、批量与“温度”的定量关系

通过动力学分析，可建立 SGD 中“学习率 ( $\eta$ )”“批量大小 ( $B$ )”与“温度 ( $T$ , 噪声强度)”的明确关联，为超参数调整提供理论依据。

##### 4.1 小批量梯度的方差尺度

设数据集总大小为  $N$ ，小批量大小为  $B$ ，在“样本独立同分布 (IID)”的近似下，小批量梯度的协方差满足：

$$\text{Cov}[\nabla L_B(\theta)] \approx \left(\frac{1}{B} - \frac{1}{N}\right) C(\theta) \quad (4.5.4.1)$$

其中  $C(\theta)$  是单个样本梯度的协方差（与参数  $\theta$  相关，代表数据本身的梯度波动）。当  $B \ll N$ （小批量远小于总数据量）时， $\frac{1}{N}$  可忽略，协方差近似为  $\frac{1}{B}C(\theta)$ ——即批量越大，随机梯度的噪声越小（方差与批量大小成反比）。

##### 4.2 有效温度与“噪声刻度”

结合 SDE 的扩散系数定义 ( $D \approx \eta \cdot \text{Cov}[\nabla L_B(\theta)]$ ) 和 OU 过程的平稳协方差 ( $P \propto \frac{T}{\lambda}$ )，可推导出“有效温度” $T$  与学习率  $\eta$ 、批量大小  $B$  的关系：

$$T \propto \eta \cdot \left(\frac{1}{B} - \frac{1}{N}\right) \quad (4.5.4.2)$$

（比例系数由问题本身的尺度决定，如  $C(\theta)$  的大小）。

这一关系揭示了“保持温度不变（噪声强度不变）”的两种等效策略：

1. 减小学习率  $\eta$ （降温）：若批量  $B$  不变，减小  $\eta$  会降低有效温度，使参数更稳定收敛；
2. 增大批量  $B$ （降温）：若学习率  $\eta$  不变，增大  $B$  会减小  $\frac{1}{B}$ ，同样降低有效温度，且增大批量更利于并行计算（比减小学习率更高效）。

**核心启发 (C)：等效退火策略** 训练后期需要“降低噪声，精准收敛”，可采用“逐步增大批量”的“等效退火”策略——相比传统的“学习率衰减”，增大批量能在不降低更新速度的前提下减小噪声，同时利用并行计算提升训练效率，是更优的超参数调整方案。

#### 4.5.5 训练策略：将动力学结论落地到实践

基于上述动力学分析，可总结出 5 条切实可行的 SGD 训练策略，直接指导实际调参与优化：

##### 1. 两阶段训练日程：

- **探索阶段（前期）**：采用“小批量 + 稍大学习率”——对应扩散极限，保持较高的有效温度（噪声强度），让参数通过随机探索跳出局部极小值，找到更优的参数区域；
- **精调阶段（后期）**：采用“逐步增大批量或衰减学习率”——对应消噪极限，降低有效温度，使参数在优质区域内精准收敛到极小值点。

2. **学习率的经验上界**：若目标函数满足 L-平滑条件，优先保证学习率  $\eta < \frac{2}{L}$ （避免 GD 的显式欧拉积分不稳定）；若在极小值附近（局部二次区域），可通过 Hessian 矩阵的最大特征值  $\lambda_{\max}$  估算学习率上界 ( $\eta < \frac{2}{\lambda_{\max}}$ )，进一步提升稳定性。

- 常步长 + 迭代平均：**常步长 SGD 在局部二次近似下易形成稳定的平稳分布（OU 过程的平稳态），但参数会因噪声存在波动；对参数序列做“迭代平均”（如 Polyak-Ruppert 平均），可显著降低噪声导致的抖动，同时保留平稳分布的“宽谷偏好”，提升收敛精度与泛化能力。
- 预条件缓解各向异性：**目标函数的各向异性（不同参数方向曲率差异大）会导致 SGD 在大曲率方向震荡、小曲率方向推进缓慢；通过“预条件”（如对不同参数方向设置不同的有效步长，或使用 Adam 等自适应优化器），可平衡不同方向的曲率与噪声强度，缓解“快慢维”问题，加快整体收敛速度。
- 早停并非悖论：**虽然朗之万动力学的平稳分布需要“长时间混合”（参数充分探索），但扩散近似表明：SGD 的“先探索后收束”是宏观规律——训练前期参数快速向优质区域移动，后期若继续训练，参数可能因噪声在平稳区域内波动，反而导致泛化性能下降。因此，结合验证集监控的“早停”策略，本质是在“探索充分”与“收敛稳定”之间找最优平衡点，并非与动力学规律矛盾。

#### 4.5.6 动力学近似的失效场景

需注意，上述 ODE/SDE 近似并非万能，在以下 4 种场景中会失效，需结合实际情况调整策略：

- 大步长或强非线性区域：**步长过大时，显式欧拉积分的误差增大，ODE/SDE 的连续近似失真；目标函数强非线性（如激活函数导致的非光滑区域）会破坏局部二次近似，OU 过程的假设不成立；
- 重尾或异方差噪声：**若小批量梯度的噪声不满足高斯分布（重尾分布），或噪声方差随参数剧烈变化（异方差），扩散近似的偏差会显著增大；
- 非 IID/强自相关采样：**若小批量采样非独立同分布（如时序数据的连续采样），会破坏噪声的无偏性与方差稳定性，噪声项存在“记忆效应”，SDE 的布朗运动假设（无记忆性）失效；
- 强各向异性：**若目标函数的曲率与噪声强度在不同方向差异极大（强各向异性），单一温度参数无法刻画噪声的方向性，需更精细的“随机平均场（SME）”或“Fokker-Planck 方程”分析，或通过预条件技术针对性优化。

#### 4.5.7 核心关系总结

离散算法	连续动力学模型	核心方程	关键参数/概念
梯度下降 (GD)	梯度流 (ODE)	$\dot{\theta}(t) = -\nabla L(\theta(t))$	学习率 $\eta$ (时间步长), L-平滑常数 $L$
随机梯度下降 (SGD)	朗之万动力学 (SDE)	$d\theta_t = -\nabla L dt + \sqrt{2T} dW_t$	有效温度 $T$ (噪声强度), 批量 $B$ , 扩散系数 $D$
极小值附近 SGD	OU 过程	$d\vartheta_t = -H\vartheta_t dt + \sqrt{2T} dW_t$	Hessian $H$ (曲率), 平稳协方差 $P$ , Lyapunov 方程

表 4.2: 离散算法与连续动力学模型的对应关系

**注 4.2** (一些简单的总结). **1.  $GD \rightarrow ODE$**  离散更新:  $\theta_{k+1} = \theta_k - \eta \nabla L(\theta_k)$  是连续方程  $\dot{\theta}(t) = -\nabla L(\theta(t))$  的显式欧拉近似。学习率  $\eta$  对应时间步长。 $\eta$  过大会导致数值不稳定。稳定区间与 Hessian 最大特征值  $\lambda_{\max}$  相关:  $0 < \eta < 2/\lambda_{\max}(H)$ 。

**2.  $SGD \rightarrow SDE$**  随机梯度  $\nabla L_B = \nabla L + \xi$  在  $\eta \rightarrow 0$  时可连续化为  $d\theta_t = -\nabla L(\theta_t)dt +$

$G(\theta_t)dW_t$ ，即朗之万动力学。噪声强度矩阵  $GG^\top \approx \eta\Sigma$ 。

### 3. 两类极限

- 消噪极限 ( $\eta \rightarrow 0, B$  大):  $SDE \rightarrow ODE$ ，收敛。
- 扩散极限 ( $\eta, B$  按比例缩放): 保持噪声强度，探索。

4. **Fokker-Planck** 方程 描述参数分布演化，平衡确定性漂移与随机扩散。噪声方向性解释了“平坦极小值泛化更好”。

5. 局部二次近似  $\rightarrow OU$  过程 在极小值附近:  $d\vartheta_t = -H\vartheta_t dt + \sqrt{2T}dW_t$ 。平稳协方差  $P$  满足  $HP + PH = 2TI$ 。各方向方差  $P_{ii} = T/\lambda_i$ ，曲率越小波动越大  $\rightarrow SGD$  自然偏向宽谷。

6. 温度刻度 小批量方差  $\propto 1/B$ 。有效温度  $T \propto \eta(\frac{1}{B} - \frac{1}{N})$ 。增大  $B$  或减小  $\eta$  均可“降温”，即退火。

### 7. 实践策略

- 前期：小批量 + 大学习率（高温扩散探索）。
- 后期：增大批量或衰减学习率（降温收敛）。
- 常步长 + 迭代平均 降低噪声。
- 预条件处理 各向异性。
- 早停 平衡探索与收敛。

8. 失效条件 步长过大、非高斯噪声、非 IID 采样、强各向异性。此时需改用随机平均场或 Fokker-Planck 分析。

核心洞见：**SGD** 是朗之万扩散在能量地形  $L(\theta)$  上的近似积分。学习率控制时间步，批量控制温度。广义目标是利用噪声探索宽谷、再逐步降温精调。

## 4.6 SGD 之问：为什么需要动量？

### 4.6.1 从 SGD 出发：我们到底缺什么？

SGD 的核心更新公式为：

$$\theta^{(k+1)} = \theta^{(k)} - \eta g^{(k)}, \quad g^{(k)} \equiv \nabla_{\theta} L(x^{(k)}, \theta^{(k)}) \quad (4.6.1.1)$$

其中  $\theta^{(k)}$  是第  $k$  步参数， $\eta$  是学习率， $g^{(k)}$  是基于小批量数据计算的随机梯度。

在实际训练中，纯 SGD 会暴露三个典型“痛点”，这正是动量（如 Heavy-Ball、NAG）要解决的核心问题：

1. **痛点 1：收敛慢** 在  $L$ -平滑凸目标函数上，纯 SGD 即使用最优常数步长，函数值收敛速率也只能达到  $O(1/k)$ ；若目标函数是强凸的，收敛速率还会受“条件数  $\kappa = L/\mu$ ”（ $L$  为平滑常数， $\mu$  为强凸系数）控制——条件数越大（如高维模型），收敛越慢，甚至出现“硬问题”（迭代数千步仍无明显下降）。
2. **痛点 2：“峡谷之字形”震荡** 当目标函数存在“各向异性曲率”（Hessian 矩阵特征值跨度大，即“峡谷地形”）时，纯 SGD 会在峡谷两侧来回震荡：大曲率方向（峡谷壁）的梯度大，迫使步长被“钳制”得很小；而小曲率方向（峡谷底）的梯度小，小步长导致推进缓慢，整体呈现“之字形”路径，严重浪费迭代次数。

3. **痛点 3：噪声底难以突破** 小批量数据带来的梯度噪声，会使纯 SGD 在训练后期陷入“噪声主导”状态——参数围绕极小值点反复波动，无法继续降低函数值，形成“噪声底”，难以收敛到更优解。

#### 4.6.2 Heavy-Ball (HB)：用“惯性”优化 SGD 的核心痛点

Heavy-Ball 是最经典的动量方法，核心是给 SGD 加入“惯性记忆”，通过累积历史更新方向，实现“抑制震荡、加快收敛、抵抗噪声”的效果。

##### 两种等价实现形式

- **位移形式（经典 Polyak 公式）**：直接通过历史参数差引入惯性

$$\theta^{(k+1)} = \theta^{(k)} - \eta g^{(k)} + \beta (\theta^{(k)} - \theta^{(k-1)}) \quad (4.6.2.1)$$

其中  $\beta \in [0, 1)$  是动量系数， $\theta^{(k)} - \theta^{(k-1)}$  是上一步的参数更新量（历史方向）， $\beta$  越大，惯性越强。

- **速度-EMA 形式（深度学习常用）**：通过“指数移动平均（EMA）”维护一个“速度”变量，间接引入惯性

$$v^{(k+1)} = \beta v^{(k)} + (1 - \beta)g^{(k)}, \quad \theta^{(k+1)} = \theta^{(k)} - \eta v^{(k+1)} \quad (4.6.2.2)$$

其中  $v^{(k)}$  是速度变量（可理解为“加权平均后的梯度”）， $(1 - \beta)$  是当前梯度的权重， $\beta$  是历史速度的权重——本质是对梯度做平滑，降低噪声影响。

##### Heavy-Ball 为什么有效？（基于一维/特征方向的直觉）

以“二次目标函数”（最易理解的凸函数）为例，设  $L(\theta) = \frac{1}{2}\lambda(\theta - \theta^*)^2$ （ $\theta^*$  是最优解， $\lambda$  是曲率），此时 HB 的误差（ $e^{(k)} = \theta^{(k)} - \theta^*$ ）满足二阶递推关系：

$$e^{(k+1)} = (1 - \eta\lambda + \beta)e^{(k)} - \beta e^{(k-1)} \quad (4.6.2.3)$$

通过选择合适的  $(\eta, \beta)$ （如  $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ ， $\kappa = L/\mu$  为条件数），可实现三大优化：

- **加速收敛**：将强凸二次函数上的收敛复杂度从纯 SGD 的  $O(\kappa \log \frac{1}{\epsilon})$ （ $\epsilon$  为精度要求）降低到  $O(\sqrt{\kappa} \log \frac{1}{\epsilon})$ ——条件数越大，加速效果越明显；
- **抑制“之字形”震荡**：动量对“高频反向梯度”（如峡谷壁的来回震荡方向）提供阻尼（历史方向与当前方向相反时，惯性会抵消部分更新），对“低频一致梯度”（如峡谷底的前进方向）做累积放大，使参数沿谷底顺滑推进；
- **抵抗噪声**：EMA 形式的速度变量会将梯度噪声的方差按系数  $\frac{1-\beta}{1+\beta}$  压低——例如  $\beta = 0.9$  时，噪声方差仅为纯 SGD 的约 5%，轨迹更平滑，后期更易突破噪声底。

##### 局限性

HB 在“二次函数”或“局部强凸目标”上效果显著，但对“一般凸目标”（无强凸性）的全局加速速率，缺乏像 NAG 那样的普适理论保证——在非强凸场景下，HB 的加速效果可能不稳定。

### 4.6.3 Nesterov (NAG): “前瞻-校正” 实现更稳健的加速

Nesterov 动量（简称 NAG）是对 HB 的改进，核心是加入“前瞻步骤”：先根据历史惯性“预判”下一步的参数位置，再用该位置的梯度做校正，避免 HB 可能出现的“过冲”问题，实现更稳健的全局加速。

#### 3.1 两种常见实现形式

- 原始 Nesterov 加速梯度（FISTA/FGM 形态）：先计算前瞻点，再用前瞻点的梯度更新

$$y^{(k)} = \theta^{(k)} + \beta (\theta^{(k)} - \theta^{(k-1)}), \quad \theta^{(k+1)} = y^{(k)} - \eta \nabla L(y^{(k)}) \quad (4.6.3.1)$$

其中  $y^{(k)}$  是“前瞻点”（基于历史惯性预判的下一步参数）， $\nabla L(y^{(k)})$  是前瞻点的梯度——相比 HB 直接用当前点梯度，NAG 用前瞻点梯度能更精准地捕捉“下一步的真实坡度”。

- 深度学习 NAG（look-ahead 梯度形态）：结合速度变量的前瞻更新

$$v^{(k+1)} = \beta v^{(k)} + g(\theta^{(k)} - \eta \beta v^{(k)}), \quad \theta^{(k+1)} = \theta^{(k)} - \eta v^{(k+1)} \quad (4.6.3.2)$$

其中  $\theta^{(k)} - \eta \beta v^{(k)}$  是前瞻点（用历史速度预判的位置）， $g(\cdot)$  是前瞻点的梯度——本质与原始形态一致，只是用速度变量简化了计算。

Nesterov 为什么更优？（核心是“前瞻-校正”）

- 凸问题的全局加速保证：在 L-平滑凸目标函数上，NAG 能通过“前瞻-校正”实现  $O(1/k^2)$  的函数值收敛速率（纯 SGD 是  $O(1/k)$ ）；在强凸目标上，能达到与 HB 相同的线性收敛率（最坏收敛因子  $\propto 1 - \frac{1}{\sqrt{\kappa}}$ ），但全局理论保证更普适。
- 减少“过冲”与误判：在强各向异性的“峡谷地形”中，HB 的惯性可能导致参数“冲过”谷底（因用当前点梯度判断方向，未考虑下一步的坡度变化）；而 NAG 的前瞻点梯度更贴近“下一步真正会到的位置”的曲率，能提前校正惯性方向，减少反复横跳，稳定性更强。

### 4.6.4 HB 与 NAG 的噪声鲁棒性对比

在相同步幅（有效更新量）下，HB 和 NAG 都能抵抗梯度噪声，但机制与适用场景略有差异：

- **HB 的优势**：EMA 形式的速度平滑对“低噪声、强曲率”场景更友好——例如小批量较大（噪声小）的强凸任务，HB 的惯性能快速累积前进方向，且实现简单、调参成本低；
- **NAG 的优势**：前瞻-校正对“高噪声、强非线性”场景更稳健——例如小批量较小（噪声大）的深度学习任务，NAG 能减少“陈旧梯度”（当前点梯度与下一步实际梯度的偏差）导致的误判，在弯曲剧烈的噪声场中轨迹更稳定。

两者的共同局限是：若训练后期不“降温”（减学习率或增批量），都会因噪声存在“噪声底”——因此需配合“迭代平均”或“等效退火”策略，进一步提升收敛精度。

### 4.6.5 实践选择：何时用 HB，何时用 NAG？

基于目标函数特性与工程需求，可按以下原则选择：

场景特征	优先选择	理由
有理论保证需求 (凸/强凸任务)	NAG	能提供 $O(1/k^2)$ 或线性收敛的全局理论保证，结果更可控
明显“峡谷地形”、过冲严重	NAG	前瞻-校正能减少惯性过冲，抑制横跳
调参简洁、兼容现有 SGD 流程	HB	EMA 形式易集成到现有代码，仅需新增一个动量系数 $\beta$ ，调参成本低
大批量、低噪声、强凸任务	HB	噪声小，无需复杂的前瞻校正，HB 的惯性加速更直接
极大条件数、低噪声 (确定性任务)	两者均可	都能实现 $\sqrt{\kappa}$ 级加速，NAG 理论保证更优，HB 实现更简单

表 4.3: HB 与 NAG 的选择指南

4.6.6 核心总结

动量的本质是给 SGD 加入“历史方向记忆”，解决纯 SGD “慢、晃、抖”的痛点：

- HB 通过“惯性累积”实现加速与震荡抑制，适合简单场景与低噪声任务；
- NAG 通过“前瞻-校正”实现更稳健的全局加速，适合复杂场景与高噪声任务；
- 无论选择哪种动量，训练后期都需配合“降温”（减学习率/增批量）或“迭代平均”，才能突破噪声底，实现精准收敛。

## 第五章 无约束优化之动量

### 5.1 符号说明

符号	含义
$x_t$	第 $t$ 步的参数向量
$\nabla f(x_t)$	目标函数在 $x_t$ 处的梯度
$\eta$	学习率（步长）
$\beta, \beta_1, \beta_2$	动量/滑动平均系数（通常取 0.9, 0.999 等）
$v_t$	第 $t$ 步的动量项（速度）
$G_{t,i}$	第 $i$ 个参数到第 $t$ 步的历史梯度平方和（AdaGrad）
$E[g^2]_t$	梯度平方的指数滑动平均（RMSProp）
$m_t$	一阶动量（梯度的指数滑动平均，Adam）
$v_t$	二阶动量（梯度平方的指数滑动平均，Adam）
$\hat{m}_t, \hat{v}_t$	偏置校正后的一阶/二阶动量（Adam）
$\varepsilon$	数值稳定项（防止除以零，通常取 $10^{-8}$ ）
$\lambda$	权重衰减系数（正则化强度）
$\nabla_i f_t(x_t)$	目标函数在 $x_t$ 处关于第 $i$ 个参数的偏导数
$\text{diag}(G_t)$	以 $G_t$ 为对角线元素的对角矩阵
$I$	单位矩阵

表 5.1: 符号说明

### 5.2 SGD 的缺点及动量方法改进

#### 5.2.1 问题出发点：SGD 的局限

标准随机梯度下降（SGD）更新：

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

问题：

- 在狭长峡谷形损失面（即特征方向尺度差异大）中，梯度方向不断剧烈摆动；
- 沿陡峭方向振荡，沿平缓方向进展缓慢；
- 收敛速率接近  $O(1/t)$ ，远慢于二阶方法。

### 5.2.2 引入动量的核心思想：惯性

想法：像物理中有质量的粒子那样，让优化“带惯性”。动量项记为速度  $v_t$ ，模拟动能积累。

$$v_{t+1} = \beta v_t + (1 - \beta)(-\nabla f(x_t))$$

$$x_{t+1} = x_t + \eta v_{t+1}$$

这就是 Polyak' s Heavy Ball (1964)。

### 5.2.3 Heavy-Ball (Polyak Momentum)

算法 5.1 (Heavy-Ball 更新规则). 形式：

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

解释：

- 当前步沿梯度下降；
- 再加上前一步的“惯性”；
- 像滚动的重球在势场中前进，惯性帮助越过小坑和振荡区。

优点：

- 加速收敛；
- 缓解振荡。

缺点：

- 对非凸问题容易过冲；
- 需要精心调节  $\beta$  与  $\eta$ 。

### 5.2.4 Nesterov 加速梯度 (NAG, 1983)

Nesterov 注意到 Heavy-Ball 更新滞后：你先算完梯度，再加惯性，但惯性早就改变了位置。他提出：提前感知未来位置。

算法 5.2 (Nesterov 加速梯度 (NAG)).

$$v_{t+1} = \beta v_t - \eta \nabla f(x_t + \beta v_t)$$

$$x_{t+1} = x_t + v_{t+1}$$

解释：

- 先“预测”下一个位置；
- 在预测点计算梯度；
- 因此能提前修正方向。

直觉：Heavy-Ball 是“被动加速”，Nesterov 是“前瞻修正”。

### 5.2.5 对比总结

特性	Heavy-Ball	Nesterov
物理意义	惯性滚动	预判修正
梯度计算点	当前点	预测点
稳定性	易过冲	更平稳
收敛速度	$O(1/k)$	$O(1/k^2)$ 在凸情形

表 5.2: Heavy-Ball 与 Nesterov 对比

### 5.2.6 从 SGD 到动量方法的逻辑链

SGD  $\rightarrow$  抖动严重  
 $\rightarrow$  引入“惯性”平滑更新 (Heavy-Ball)  
 $\rightarrow$  进一步在“预测点”计算梯度 (Nesterov)  
 $\rightarrow$  演化出现代动量优化器（如 Adam, RMSProp, AdaBelief）中融合动量思想的分支。

## 5.3 AdaGrad (Duchi et al., 2011)

### 5.3.1 动机：SGD 学习率“一刀切”的问题

SGD 使用固定学习率：

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

缺陷：

1. 各参数维度梯度尺度不同，统一学习率不合理。
2. 稀疏特征学习慢（小梯度参数被忽略）。
3. 学习率难以手动调整。

核心想法：

让每个参数拥有独立的、自适应的学习率。

梯度大的维度  $\rightarrow$  下降步长变小；

梯度小的维度  $\rightarrow$  下降步长变大。

### 5.3.2 算法公式

**算法 5.3** (AdaGrad 算法). 设第  $i$  个参数的历史梯度平方和为：

$$G_{t,i} = \sum_{\tau=1}^t (\nabla_i f_{\tau}(x_{\tau}))^2$$

更新规则为：

$$x_{t+1,i} = x_{t,i} - \frac{\eta}{\sqrt{G_{t,i}} + \varepsilon} \nabla_i f_t(x_t)$$

或向量形式：

$$x_{t+1} = x_t - \eta \cdot D_t^{-1/2} \nabla f_t(x_t)$$

其中：

$$D_t = \text{diag}(G_t) + \varepsilon I$$

### 5.3.3 性质与效果

1. 方向自适应：梯度大（噪声多）的维度衰减快，梯度小的维度保持较大学习率。
2. 天然适用于稀疏数据（如 NLP 中的 embedding 训练），罕见词梯度少 → 步长较大。
3. 单调递减学习率：  
因为  $G_{t,i}$  累积增长， $\sqrt{G_{t,i}}$  也持续增大。  
结果是学习率不断衰减，最终趋近于 0。

### 5.3.4 优缺点

优点：

- 不需要手动调节学习率；
- 稀疏特征训练效果突出；
- 理论上可证明收敛率  $O(1/\sqrt{t})$ 。

缺点：

- 学习率衰减过快（在非凸问题上几乎停止更新）；
- 对密集梯度任务表现差（如深度网络）。

### 5.3.5 本质理解：累积“几何尺度”的归一化

从几何角度看，AdaGrad 在梯度空间中进行各向异性缩放（anisotropic scaling）：

$$\Delta x_t = -\eta (G_t)^{-1/2} \nabla f_t(x_t)$$

即在每个方向上使用“与历史梯度能量成反比”的缩放。

可以理解为在一个逐步扭曲的黎曼度量（Riemannian metric）下优化。

AdaGrad 相当于在每一步都重新定义“距离”的概念，让常被更新的方向走得更谨慎，少被更新的方向走得更大胆。

## 5.4 RMSProp

### 5.4.1 动机：修正 AdaGrad 的“学习率枯竭”

AdaGrad 的问题核心在于

$$G_{t,i} = \sum_{\tau=1}^t (\nabla_i f_{\tau})^2$$

不断累加，使得分母

$$\sqrt{G_{t,i}}$$

持续增大 → 学习率单调下降 → 在训练后期几乎不动。

**RMSProp 的核心改进：**不再无限累积，而是使用**指数滑动平均（EMA）**仅保留“近期”梯度信息。

### 5.4.2 算法公式

**算法 5.4** (RMSProp 算法). 定义平方梯度的滑动平均：

$$E[g^2]_t = \rho E[g^2]_{t-1} + (1 - \rho)(\nabla f_t(x_t))^2$$

其中  $\rho \in [0, 1)$  通常取 0.9。

更新规则：

$$x_{t+1} = x_t - \frac{\eta}{\sqrt{E[g^2]_t + \varepsilon}} \odot \nabla f_t(x_t)$$

或分量形式：

$$x_{t+1,i} = x_{t,i} - \frac{\eta}{\sqrt{E[g_i^2]_t + \varepsilon}} \cdot \nabla_i f_t(x_t)$$

### 5.4.3 性质与直觉

1. **滑动窗口的能量归一化：**不再记住所有历史，而只记住“近几步”的梯度能量。这样学习率不会无限衰减。
2. **自适应但平稳：**对于方差大的维度（梯度震荡），分母变大 → 学习率下降；对于稳定的维度，学习率保持相对较大。
3. **鲁棒性好：**对非平稳损失（如深度网络早期阶段的抖动）有缓冲作用。

### 5.4.4 与 AdaGrad 的对比

特性	AdaGrad	RMSProp
梯度记忆	全历史累积	指数滑动平均
学习率	单调递减至零	稳定在某范围
稀疏特征适配	强	一般
深度网络表现	弱	强

表 5.3: AdaGrad 与 RMSProp 对比

### 5.4.5 本质理解

RMSProp 相当于让优化器在一个“动态调整的、局部平滑”的度量空间中前进。在几何意义上，它不是单调放大的尺度，而是根据当前梯度方差**实时自适应地拉伸或压缩**参数空间。

换句话说：

AdaGrad 是“记仇”的学生（过去的错误全记着）；  
 RMSProp 是“善忘”的学生（只记得最近的错误）。

## 5.5 Adam (Adaptive Moment Estimation)

### 5.5.1 动机：融合动量与自适应学习率

前两者的优劣：

- 动量法（Heavy-Ball / Nesterov）平滑方向，加速收敛。
- **RMSProp** 通过平方梯度的滑动平均调节学习率，抑制震荡。

Adam 结合两者：

“动量提供方向一致性，RMSProp 提供步长自适应。”

同时引入偏置校正（bias correction）解决初始化期估计偏小的问题。

### 5.5.2 算法核心公式

算法 5.5 (Adam 算法). 定义梯度：

$$g_t = \nabla f_t(x_t)$$

1. 一阶动量（梯度均值）

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

2. 二阶动量（平方梯度均值）

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

3. 偏置校正（*bias correction*）在初期  $m_t, v_t$  向 0 偏移，故修正为：

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

4. 更新规则：

$$x_{t+1} = x_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

### 5.5.3 性质与优点

1. 方向平滑 + 步长自适应：

- $m_t$  平滑梯度方向，避免震荡；
- $v_t$  控制各维度学习率，防止陡峭方向过冲。

2. 偏置校正保证早期稳定性：使早期梯度统计不再被低估。

3. 无需手动调参：在大部分任务中默认参数即可工作良好。

4. 适用于非平稳目标、稀疏特征与深度网络。

### 5.5.4 缺点与改进方向

1. 可能欠收敛：在部分凸问题中，Adam 不一定收敛到最优点（Reddi et al., 2018 指出）。
2. 过度自适应导致步长不稳定：解决方案有 AMSGrad、AdamW、AdaBelief 等。

### 5.5.5 本质理解

Adam 在几何意义上是时间加权的各向异性梯度法：

- $m_t$  代表一阶动量场，提供“惯性”；
- $v_t$  则定义一个时变的度量张量（metric tensor），自适应调整每个方向的步长；
- 整体行为等价于在动态曲率修正的黎曼空间中作平滑梯度下降。

直观比喻：

SGD 是盲目走路的人，

RMSProp 是谨慎地根据地形调整步伐的人，

Adam 是既看惯性又看地形的“自动巡航”行者。

## 5.6 AdamW (Adam with Decoupled Weight)

### 5.6.1 动机：修正 Adam 正则化的逻辑错误

Adam 原版通常通过 L2 正则项实现权重衰减：

$$\min_x f(x) + \frac{\lambda}{2}|x|^2$$

SGD 中这等价于在梯度中加项  $\lambda x$ ：

$$x_{t+1} = x_t - \eta(\nabla f_t(x_t) + \lambda x_t)$$

但 Adam 的更新包含自适应缩放：

$$x_{t+1} = x_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

此时把  $+\lambda x_t$  加进梯度会被  $(\sqrt{\hat{v}_t} + \varepsilon)^{-1}$  缩放，导致正则化强度与梯度统计耦合，不再是纯粹的权重衰减。

### 5.6.2 核心思想：权重衰减与梯度更新解耦

AdamW 的关键修改：不再把  $\lambda x_t$  加入梯度，而是直接对参数施加衰减：

算法 5.6 (AdamW 算法).

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$x_{t+1} = x_t - \eta \left( \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} + \lambda x_t \right)$$

注意：衰减项  $\lambda x_t$  不再乘以自适应比例因子。

### 5.6.3 性质与效果

特性	Adam	AdamW
权重衰减	通过梯度项实现，受自适应缩放影响	与梯度更新解耦，恒定衰减率
正则化一致性	不稳定	稳定且可控
理论收敛性	弱	改善（更接近 SGD 行为）
实践效果	对超参数敏感	更鲁棒，普遍优于原版 Adam

表 5.4: Adam 与 AdamW 对比

### 5.6.4 本质理解

AdamW 将优化器的两个任务分离：

1. **梯度驱动更新**：由  $\hat{m}_t/\sqrt{\hat{v}_t}$  决定方向与步长。
2. **权重衰减**：独立控制参数幅度，起到正则作用。

几何视角：

- Adam 在一个**动态加权的度量空间**中做下降；
- AdamW 额外加入一个**欧式长度惩罚**，保持参数范数稳定；
- 两者独立，因此正则化强度与自适应缩放无关。

## 第六章 阻尼牛顿法

### 6.1 牛顿法复习

这是牛顿法区别于梯度下降的核心，用二次函数拟合目标函数局部形态：

$$f(x + \Delta) \approx f(x) + \nabla f(x)^\top \Delta + \frac{1}{2} \Delta^\top H(x) \Delta$$

对二次近似函数求极值（导数为 0），解出的搜索方向即“牛顿步”：

$$\Delta_{nt} = -H(x)^{-1} \nabla f(x)$$

- **作用：**直接给出使局部二次函数最小的方向，无需像梯度下降那样手动调整步长（理论上步长为 1 时局部最优）。
- **前提：** $H(x) \succ 0$ （Hessian 正定），保证解唯一且为最小值点。

通过梯度与牛顿步的内积，证明牛顿步是“有效下降方向”：

$$\nabla f(x)^\top \Delta_{nt} = -\nabla f(x)^\top H(x)^{-1} \nabla f(x) < 0$$

- **作用：**确保沿牛顿步迭代时，目标函数值会减小（内积  $< 0$ ，方向与梯度相反且符合曲率）。
- **关键：**因  $H(x) \succ 0$ ，其逆矩阵也正定，故  $\nabla f(x)^\top H(x)^{-1} \nabla f(x) > 0$ ，最终内积为负。

靠近最优解  $x^*$  时，迭代误差呈“平方级”减小，收敛速度远快于梯度下降：

$$\|x_{k+1} - x^*\| \leq C \cdot \|x_k - x^*\|^2 \quad (C > 0)$$

- **作用：**体现牛顿法的核心优势——一旦进入“局部收敛域”，迭代会快速收敛到最优解。
- **前提：** $H(x)$  在  $x^*$  邻域 Lipschitz 连续，且初始点  $x^{(0)}$  足够靠近  $x^*$ 。

### 6.2 阻尼牛顿法

阻尼牛顿法（Damped Newton Method）是为解决纯牛顿法在远离最优解时可能不下降、发散的问题而提出的改进方法，核心思路是：**保留牛顿方向的优势，通过引入线搜索（Line Search）确定合适的步长**，而非纯牛顿法中默认的步长 1，从而保证每次迭代都能使目标函数值下降，兼顾“局部快速收敛”与“全局有效下降”。

### 6.2.1 纯牛顿法的局限性（阻尼牛顿法的必要性）

纯牛顿法的迭代公式为  $x_{k+1} = x_k + \Delta_{nt}$ （步长固定为 1），但存在两个关键问题：

1. **Hessian 矩阵非正定**：当  $H(x_k)$  不正定时，牛顿步  $\Delta_{nt} = -H(x_k)^{-1}\nabla f(x_k)$  可能不是下降方向（甚至是上升方向），导致  $f(x_{k+1}) > f(x_k)$ 。
2. **步长 1 过大**：即使  $H(x_k)$  正定（牛顿步是下降方向），但远离最优解时，二次近似的误差较大，步长 1 可能导致迭代点“迈过”最优解，反而使函数值上升。

### 6.2.2 阻尼牛顿法的核心改进：牛顿方向 + 线搜索

阻尼牛顿法保留“牛顿步”作为搜索方向（利用二阶信息的高效性），但通过**线搜索**动态调整步长  $\alpha_k$  ( $\alpha_k > 0$ )，确保每次迭代满足  $f(x_{k+1}) < f(x_k)$ 。

迭代公式：

$$x_{k+1} = x_k + \alpha_k \cdot \Delta_{nt}$$

其中：

- $\Delta_{nt} = -H(x_k)^{-1}\nabla f(x_k)$  是牛顿方向（与纯牛顿法一致）；
- $\alpha_k$  是通过线搜索确定的步长（核心改进点，替代固定步长 1）。

### 6.2.3 线搜索：如何确定步长 $\alpha_k$ ？

线搜索的目标是找到最小的  $\alpha_k$ （通常从 1 开始尝试），使得目标函数值“充分下降”。常用准则为 **Armijo 准则**（保证下降性的同时避免步长过小）：

**定义 6.1** (Armijo 准则). *Armijo 准则 (Armijo Criterion)* 是**线搜索 (Line Search)** 中用于确定合适步长的经典准则。

设当前迭代点为  $x_k$ ，搜索方向为  $d_k$ （需满足“下降方向”，即  $\nabla f(x_k)^\top d_k < 0$ ），步长为  $\alpha > 0$ 。*Armijo 准则*要求步长  $\alpha$  满足：

$$f(x_k + \alpha \cdot d_k) \leq f(x_k) + c \cdot \alpha \cdot \nabla f(x_k)^\top d_k$$

其中：

- $f(\cdot)$  是目标函数；
- $\nabla f(x_k)$  是  $f$  在  $x_k$  处的梯度；
- $c$  是预设常数，满足  $0 < c < 1$ （通常取  $c = 10^{-4}$ ，控制“最小可接受的下降量”）。

实际计算步骤（回溯线搜索）：

1. 初始尝试步长  $\alpha = 1$ （纯牛顿法的默认步长，靠近最优解时通常有效）；
2. 检查是否满足 Armijo 准则：若  $f(x_k + \alpha d_k) \leq f(x_k) + c \cdot \alpha \cdot \nabla f(x_k)^\top d_k$ ，则接受该  $\alpha$ ；
3. 若不满足，缩小步长（如乘以收缩因子  $\beta$ ， $0 < \beta < 1$ ，常用  $\beta = 0.5$ ），即  $\alpha = \beta \cdot \alpha$ ，重复步骤 2；

4. 直到找到满足准则的  $\alpha$  (理论上, 因  $d_k$  是下降方向, 当  $\alpha \rightarrow 0$  时准则必然满足, 故一定能找到)。

对于阻尼牛顿法, 步长  $\alpha_k$  需满足:

$$f(x_k + \alpha_k \Delta_{nt}) \leq f(x_k) + c \cdot \alpha_k \cdot \nabla f(x_k)^\top \Delta_{nt}$$

说明:

- 不等式右边是函数值的“预期下降量”(基于一阶近似), 左边是实际下降量。
- 因牛顿步是下降方向 (若  $H$  正定, 则  $\nabla f(x_k)^\top \Delta_{nt} < 0$ ),  $c \cdot \alpha_k \cdot (\text{负数})$  会使右边小于  $f(x_k)$ , 从而强制左边 (实际函数值) 必须下降。
- 若  $\alpha = 1$  满足 Armijo 准则, 则直接使用 (接近最优解时通常成立, 保持二次收敛); 否则按比例缩小  $\alpha$  (如乘以 0.5), 直到满足条件。

### 6.2.4 阻尼牛顿法的优势

1. **全局下降保证:** 通过线搜索, 无论  $H(x_k)$  是否正定 (即使牛顿步不是下降方向, 线搜索会筛选出使函数下降的步长), 都能确保  $f(x_{k+1}) < f(x_k)$ , 避免发散。
2. **保留局部快速收敛:** 当迭代点靠近最优解  $x^*$  时, 二次近似误差很小,  $\alpha_k$  会趋近于 1, 此时阻尼牛顿法退化为纯牛顿法, 仍保持二次收敛速度。
3. **适用性更广:** 相比纯牛顿法, 对初始点的要求更低 (无需“足够靠近最优解”), 在非凸问题或远离最优解的场景中更稳定。

### 6.2.5 阻尼牛顿法的步骤总结

**算法 6.1** (阻尼牛顿法). 1. 初始化迭代点  $x_0$ , 设置精度阈值  $\epsilon > 0$ ;

2. 计算梯度  $\nabla f(x_k)$ , 若  $\|\nabla f(x_k)\| < \epsilon$ , 停止迭代 (已收敛);

3. 计算 Hessian 矩阵  $H(x_k)$ , 求解牛顿步  $\Delta_{nt} = -H(x_k)^{-1} \nabla f(x_k)$ ;

4. 通过 Armijo 准则线搜索确定步长  $\alpha_k$ ;

5. 更新迭代点:  $x_{k+1} = x_k + \alpha_k \Delta_{nt}$ , 返回步骤 2。

简言之, 阻尼牛顿法通过“方向用牛顿 (高效), 步长靠搜索 (稳定)”的策略, 完美弥补了纯牛顿法的缺陷, 是实际中更常用的牛顿类优化方法。

## 6.3 阻尼牛顿法性质

### 6.3.1 一些假设

- (A1)  $f \in C^2(\mathbb{R}^n)$ , 且梯度 **Lipschitz 连续**: 存在常数  $L > 0$  使

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

等价于  $f$  为  $L$ -光滑 ( $L$ -smooth)。

- (A2)  $f$  下有界:  $\inf_x f(x) > -\infty$ 。
- (A3) 方向矩阵 一致有界且正定: 存在常数  $0 < m \leq M < \infty$ , 对所有  $k$  有

$$mI \preceq H_k \preceq MI.$$

(若直接用 Hessian, 则这相当于全局强凸与曲率上界; 在一般非凸情形, 可通过正定化保证。)

- (A4) 步长  $\alpha_k$  由回溯满足 Armijo 条件 (或更强的 Wolfe 条件)。

### 6.3.2 全局收敛到临界点

要证明阻尼牛顿法的全局收敛性 ( $\|\nabla f(x_k)\| \rightarrow 0$ ), 我们基于上述两个引理和假设 (A1)-(A4), 分步骤推导:

**引理 6.1** (牛顿方向的下降性与有界性). 设  $g_k = \nabla f(x_k)$ , 牛顿方向  $p_k = -H_k^{-1}g_k$  (由  $H_k p_k = -g_k$  定义)。

1. 下降方向证明: 因  $H_k \succ 0$ , 其逆矩阵  $H_k^{-1} \succ 0$ , 故

$$-g_k^\top p_k = g_k^\top H_k^{-1} g_k \geq \frac{1}{\|H_k\|} \|g_k\|^2 \geq \frac{1}{M} \|g_k\|^2 > 0$$

这说明  $p_k$  是下降方向 (梯度与方向的内积为正, 即方向与负梯度同向)。

2. 方向有界性与夹角估计: 由  $H_k \succeq mI$ , 得  $\|H_k^{-1}\| \leq \frac{1}{m}$ , 因此

$$\|p_k\| = \|H_k^{-1} g_k\| \leq \|H_k^{-1}\| \|g_k\| \leq \frac{1}{m} \|g_k\|$$

方向与负梯度的夹角余弦为:

$$\cos \theta_k = \frac{-g_k^\top p_k}{\|g_k\| \|p_k\|} \geq \frac{1/M}{1/m} = \frac{m}{M} > 0$$

即方向与负梯度夹角远离  $90^\circ$ , 是“有效下降方向”。

**引理 6.2** (回溯线搜索的步长下界). 设 (A1)-(A3) 成立, 定义

$$\alpha_0 := \frac{2(1-c_1)m^2}{LM}$$

则回溯法接受的步长  $\alpha_k$  满足

$$\alpha_k \geq \underline{\alpha} := \min\{1, \beta\alpha_0\} > 0$$

证明.

由  $f$  是  $L$ -光滑 (A1), 其泰勒展开满足上界:

$$f(x_k + \alpha p_k) \leq f(x_k) + \alpha g_k^\top p_k + \frac{L}{2} \alpha^2 \|p_k\|^2$$

Armijo 条件要求:

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha g_k^\top p_k$$

结合上式，只需：

$$(1 - c_1)\alpha g_k^\top p_k + \frac{L}{2}\alpha^2 \|p_k\|^2 \leq 0$$

令  $\delta_k = -g_k^\top p_k > 0$  (因  $p_k$  是下降方向)，上式等价于：

$$-(1 - c_1)\alpha\delta_k + \frac{L}{2}\alpha^2 \|p_k\|^2 \leq 0$$

当  $\alpha \leq \frac{2(1-c_1)\delta_k}{L\|p_k\|^2}$  时成立。

由引理 1,  $\delta_k \geq \frac{1}{M}\|g_k\|^2$  且  $\|p_k\| \leq \frac{1}{m}\|g_k\|$ ，代入得：

$$\frac{2(1-c_1)\delta_k}{L\|p_k\|^2} \geq \frac{2(1-c_1)(1/M)\|g_k\|^2}{L(1/m^2)\|g_k\|^2} = \frac{2(1-c_1)m^2}{LM} = \alpha_0$$

因此，当  $\alpha \leq \alpha_0$  时 Armijo 条件必成立。

回溯法从  $\alpha = 1$  开始按因子  $\beta$  ( $0 < \beta < 1$ ) 递减，首次满足条件的步长不少于  $\beta\alpha_0$ ，故  $\alpha_k \geq \min\{1, \beta\alpha_0\} = \underline{\alpha} > 0$ 。

**定理 6.1 (全局收敛性).** 在假设 (A1)-(A4) 下，阻尼牛顿法生成的序列满足  $\|\nabla f(x_k)\| \rightarrow 0$ ，即全局收敛。

证明.

1. 函数值序列的单调性与有界性：由 Armijo 条件 (A4),  $f(x_{k+1}) < f(x_k)$ ，即  $\{f(x_k)\}$  单调递减。又  $f$  下有界 (A2)，故  $\{f(x_k)\}$  收敛，即

$$\lim_{k \rightarrow \infty} (f(x_k) - f(x_{k+1})) = 0$$

2. 结合步长下界与下降量的关系：由引理 1,  $-g_k^\top p_k \geq \frac{1}{M}\|g_k\|^2$ ；由引理 2,  $\alpha_k \geq \underline{\alpha} > 0$ 。结合 Armijo 条件的下降量：

$$f(x_k) - f(x_{k+1}) \geq c_1 \alpha_k (-g_k^\top p_k) \geq c_1 \underline{\alpha} \cdot \frac{1}{M} \|g_k\|^2$$

3. 级数收敛推导出梯度范数收敛：由于  $\sum_{k=0}^{\infty} (f(x_k) - f(x_{k+1}))$  收敛，其通项必须趋于 0，即

$$\lim_{k \rightarrow \infty} \|g_k\|^2 = \lim_{k \rightarrow \infty} \|\nabla f(x_k)\|^2 = 0$$

因此

$$\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$$

### 6.3.3 局部收敛阶段的退化

这部分内容是在分析阻尼牛顿法在“局部收敛阶段”的行为，核心是证明：当迭代点足够靠近最优解时，阻尼牛顿法的步长会恢复为  $\alpha_k = 1$  (即“单位步”，退化为纯牛顿法)，从而恢复局部二次收敛的快速速率。

**引理 6.3.** 目的是给出一个充分条件，使得步长  $\alpha_k = 1$  满足 Armijo 条件 (即无需缩小步长，直接用纯牛顿步迭代)。

- 条件：若牛顿步  $p_k = -H_k^{-1}g_k$  ( $g_k = \nabla f(x_k)$ ) 满足

$$\|p_k\| \leq \frac{3(1-2c_1)}{L_H} \lambda_{\min}(H_k), \quad c_1 \in (0, \frac{1}{2})$$

则  $\alpha_k = 1$  满足 Armijo 条件。

- 意义：当该条件成立时，阻尼牛顿法的步长不再需要“回溯缩小”，直接用纯牛顿步迭代，从而继承纯牛顿法的局部二次收敛速率。

通过带三阶余项的泰勒展开分析函数值变化，验证  $\alpha = 1$  时 Armijo 条件成立：

- 对  $f(x_k + p_k)$  做三阶泰勒展开（余项由 Hessian 的 Lipschitz 常数  $L_H$  控制）：

$$f(x_k + p_k) \leq f(x_k) + g_k^\top p_k + \frac{1}{2} p_k^\top H_k p_k + \frac{L_H}{6} \|p_k\|^3$$

- 代入牛顿步的定义  $g_k = -H_k p_k$ ，化简得：

$$f(x_k + p_k) \leq f(x_k) - \frac{1}{2} p_k^\top H_k p_k + \frac{L_H}{6} \|p_k\|^3$$

- 对比 Armijo 条件的要求（ $f(x_k + p_k) \leq f(x_k) - c_1 p_k^\top H_k p_k$ ），推导得：当

$$\left(\frac{1}{2} - c_1\right) \lambda_{\min}(H_k) \|p_k\|^2 \geq \frac{L_H}{6} \|p_k\|^3$$

时，Armijo 条件成立。整理后即得到引理中的条件  $\|p_k\| \leq \frac{3(1-2c_1)}{L_H} \lambda_{\min}(H_k)$ 。

当迭代点足够靠近最优解  $x^*$  时：

- 牛顿步的长度  $\|p_k\| = O(\|e_k\|)$ （ $e_k = x_k - x^*$  是迭代误差），且 Hessian 的最小特征值  $\lambda_{\min}(H_k) \rightarrow \lambda_{\min}(H^*)$ （因 Hessian 在邻域 Lipschitz 连续）。
- 此时引理的条件会被满足，回溯线搜索将以  $\alpha_k = 1$  终止，迭代退化为纯牛顿步： $x_{k+1} = x_k + p_k$ 。
- 进而恢复纯牛顿法的局部二次收敛速率，误差满足：

$$\|e_{k+1}\| \leq C \|e_k\|^2, \quad C = \frac{L_H}{2m}$$

这部分内容的核心是桥接阻尼牛顿法的“全局收敛稳定性”与纯牛顿法的“局部二次收敛快速性”：通过证明“局部邻域内步长恢复为 1”，说明阻尼牛顿法在全局收敛后，会快速进入二次收敛阶段，从而兼具“全局稳定下降”和“局部快速收敛”的优势。

## 6.4 非显式求逆方法

### 6.4.1 Cholesky 分解（Cholesky Decomposition）

定义：若矩阵  $A \in \mathbb{R}^{n \times n}$  是对称正定矩阵（symmetric positive definite），则存在唯一的下三角矩阵  $L$  满足：

$$A = LL^T$$

其中  $L$  的对角元素全为正。

含义：

- 把正定矩阵看作“平方”出来的结果。
- 等价于高斯消元的稳定版本。
- 数值优化中，常用于求解  $Ax = b$  时避免直接求逆：

$$LL^T x = b \Rightarrow Ly = b, \quad L^T x = y$$

计算方式：对每个元素递推：

$$L_{ii} = \sqrt{A_{ii} - \sum_{k=1}^{i-1} L_{ik}^2}, \quad L_{ij} = \frac{1}{L_{jj}} \left( A_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk} \right) \quad (i > j)$$

条件：必须是正定矩阵，否则根号项出现负数，分解失败。

### 6.4.2 LDL 分解 (LDL Decomposition)

定义：若矩阵  $A \in \mathbb{R}^{n \times n}$  是对称矩阵（不要求正定），则可以分解为：

$$A = LDL^T$$

其中：

- $L$  是单位下三角矩阵（对角元为 1），
- $D$  是对角矩阵（可以含正或负元素）。

含义：

- 是对称矩阵的广义 Cholesky 分解。
- 适用于半正定或不定矩阵，不会出现在 Cholesky 分解中那样的平方根问题。
- 若  $A$  正定，则  $D$  全为正，退化为标准 Cholesky：

$$A = (L\sqrt{D})(L\sqrt{D})^T$$

算法形式：递推式：

$$D_{jj} = A_{jj} - \sum_{k=1}^{j-1} L_{jk}^2 D_{kk}, \quad L_{ij} = \frac{1}{D_{jj}} \left( A_{ij} - \sum_{k=1}^{j-1} L_{ik} D_{kk} L_{jk} \right)$$

优点：

- 不需要取平方根，数值更稳定。
- 可处理不定矩阵。

## 第七章 牛顿法和拟牛顿法

### 7.1 符号说明

符号	类型	定义与用途
$f(\mathbf{x})$	标量函数	无约束优化问题的目标函数，输入为 $n$ 维优化变量 $\mathbf{x}$ ，输出为实数值。
$\nabla f(\mathbf{x})$	向量函数	目标函数 $f(\mathbf{x})$ 的梯度，输入为 $\mathbf{x}$ ，输出为 $n$ 维梯度向量。
$\mathbf{x}_k$	向量	第 $k$ 步迭代点， $\mathbf{x}_k \in \mathbb{R}^n$ ； $\mathbf{x}_0$ 为初始迭代点。
$\mathbf{x}^*$	向量	目标函数 $f(\mathbf{x})$ 的最优解（近似）。
$\mathbf{g}_k$	向量	第 $k$ 步迭代点的梯度，即 $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$ ， $\mathbf{g}_k \in \mathbb{R}^n$ 。
$\mathbf{H}$	矩阵	目标函数 $f(\mathbf{x})$ 的 Hessian 矩阵（二阶导数矩阵）， $\mathbf{H} \in \mathbb{R}^{n \times n}$ ； $\mathbf{H}_k = \nabla^2 f(\mathbf{x}_k)$ 为第 $k$ 步迭代点的 Hessian 矩阵。
$\mathbf{B}_k$	矩阵	BFGS 方法中第 $k$ 步的 Hessian 矩阵近似， $\mathbf{B}_k \in \mathbb{R}^{n \times n}$ ，满足拟牛顿条件 $\mathbf{B}_{k+1}\mathbf{s}_k = \mathbf{y}_k$ 。
$\mathbf{G}_k$	矩阵	$\mathbf{B}_k$ 的逆矩阵近似（即 $\mathbf{G}_k \approx \mathbf{B}_k^{-1}$ ）， $\mathbf{G}_k \in \mathbb{R}^{n \times n}$ ，用于直接计算搜索方向，避免矩阵求逆。
$\mathbf{p}_k$	向量	第 $k$ 步的搜索方向（下降方向），牛顿法中 $\mathbf{p}_k = -\mathbf{H}_k^{-1}\mathbf{g}_k$ ，BFGS 中 $\mathbf{p}_k = -\mathbf{G}_k\mathbf{g}_k$ 。
$\alpha_k$	标量	第 $k$ 步的迭代步长，通过 Wolfe 条件（Armijo 条件 + 曲率条件）确定， $\alpha_k > 0$ 。
$\mathbf{s}_k$	向量	第 $k$ 步的变量增量，即 $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ ，反映迭代点的变化量。
$\mathbf{y}_k$	向量	第 $k$ 步的梯度增量，即 $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ ，反映梯度的变化量。

表 7.1: 符号说明

### 7.2 牛顿法的严格数学建模（复习）

Newton 法是比较梯度下降更高效的优化方法，核心是通过函数的局部二次近似确定搜索方向，兼具“局部快速收敛”与“全局有效下降”的特性，其核心逻辑围绕“二阶展开  $\rightarrow$  牛顿步  $\rightarrow$  收敛性”展开：

### 7.2.1 核心思路：函数的局部二次近似

梯度下降仅用“一阶信息（梯度）”将函数局部近似为线性函数，而 Newton 法引入“二阶信息（Hessian 矩阵）”，将函数局部近似为二次函数（更贴合非凸函数的局部曲率），具体如下：

对迭代点  $x_k$ ，将目标函数  $f(x_k + p)$  在  $x_k$  处做二阶泰勒展开（ $p$  为搜索方向向量）：

$$f(x_k + p) \approx f(x_k) + g_k^\top p + \frac{1}{2} p^\top H_k p$$

其中：

- $g_k = \nabla f(x_k)$  是  $f(x)$  在  $x_k$  处的梯度（一阶导数）；
- $H_k = \nabla^2 f(x_k)$  是  $f(x)$  在  $x_k$  处的 Hessian 矩阵（二阶导数矩阵），反映函数在  $x_k$  处的局部曲率。

Newton 法的核心是：最小化上述二次近似函数，直接求解使近似函数最小的搜索方向  $p$ 。

### 7.2.2 牛顿步（Newton Step）的推导

对二阶近似函数关于  $p$  求导，并令导数为 0（二次函数的极值点条件）：

$$\frac{\partial}{\partial p} \left[ f(x_k) + g_k^\top p + \frac{1}{2} p^\top H_k p \right] = g_k + H_k p = 0$$

若 Hessian 矩阵正定（ $H_k \succ 0$ ，保证二次近似函数是凸函数，极值点为最小值点），则可解出唯一的搜索方向——牛顿步：

$$p_k = -H_k^{-1} g_k$$

### 7.2.3 牛顿步的下降性

牛顿步能保证是“下降方向”的前提是  $H_k \succ 0$ ，证明如下：计算梯度与牛顿步的内积（判断方向是否下降的核心指标，内积  $< 0$  则为下降方向）：

$$g_k^\top p_k = g_k^\top (-H_k^{-1} g_k)$$

因  $H_k \succ 0$ ，其逆矩阵  $H_k^{-1}$  也正定，故对任意非零向量  $g_k$ ，有  $g_k^\top H_k^{-1} g_k > 0$ ，因此：

$$g_k^\top p_k < 0$$

即牛顿步满足“下降方向”的核心条件。

### 7.2.4 局部二次收敛：牛顿法的核心优势

当迭代点足够靠近最优解  $x^*$  时，Newton 法会呈现二次收敛（收敛速度远快于梯度下降的线性收敛），严格定义如下：

#### 收敛条件

若满足以下两个前提：

1. Hessian 矩阵  $H(x)$  在  $x^*$  的邻域内 **Lipschitz 连续**（曲率变化平缓）；
2. 初始迭代点  $x_0$  足够靠近  $x^*$ （进入“局部收敛域”）。

### 二次收敛公式

此时存在常数  $C > 0$ ，使得迭代误差满足：

$$\|x_{k+1} - x^*\| \leq C \cdot \|x_k - x^*\|^2$$

### 直观意义

二次收敛意味着“每次迭代后，误差的有效位数会翻倍”——例如：若第  $k$  步误差为  $10^{-2}$ ，第  $k+1$  步误差可降至  $10^{-4}$ ，第  $k+2$  步可降至  $10^{-8}$ ，接近最优解时收敛极快。

牛顿法也有缺陷，数学层面的问题包括：

#### 1. Hessian 构造与求解成本高：

- 构造 Hessian 矩阵  $H_k$  需计算  $n(n+1)/2$  个二阶偏导数（复杂度  $O(n^2)$ ）；
- 求解线性方程组  $H_k p_k = -g_k$  需  $O(n^3)$  复杂度（如 LU 分解），当  $n$  较大（如大规模优化）时计算成本不可承受。

#### 2. Hessian 不定导致方向非下降：若 $H_k$ 不定（非正定时），牛顿方向 $p_k = -H_k^{-1}g_k$ 可能不满足“下降方向”条件（即 $g_k^\top p_k \geq 0$ ），导致迭代点 $x_{k+1}$ 处 $f(x_{k+1}) > f(x_k)$ ，算法不稳定。

## 7.3 拟牛顿法

BFGS（Broyden-Fletcher-Goldfarb-Shanno）是最常用的拟牛顿方法之一，主要用于求解无约束优化问题。拟牛顿法的目标是通过逐步逼近目标函数的 Hessian 矩阵来优化目标函数，而无需显式计算二阶导数。其核心思想是通过逐步更新一个近似 Hessian 矩阵，使得每一步的更新尽可能接近真实的 Hessian 矩阵。

### 7.3.1 拟牛顿法的基本框架

考虑一个无约束优化问题：

$$\min_x f(x),$$

其中  $f(x)$  是一个可微的目标函数， $x \in \mathbb{R}^n$  是优化变量。拟牛顿法的基本思路是通过计算一系列梯度来逼近目标函数的 Hessian 矩阵，而不是直接计算 Hessian 矩阵。

牛顿法的更新公式为：

$$x_{k+1} = x_k - H_k^{-1}g_k,$$

其中  $H_k$  是  $f(x)$  在  $x_k$  处的 Hessian 矩阵。由于计算 Hessian 矩阵代价较高，拟牛顿法通过构造一个近似矩阵  $B_k$  来代替真实的  $H_k$ ，从而避免显式计算它。

### 7.3.2 BFGS 方法

BFGS 是一种常用的拟牛顿方法，给出了如何通过一系列梯度信息更新一个 Hessian 近似矩阵  $B_k$  的规则。其核心思想是通过引入变量  $s_k = x_{k+1} - x_k$  和  $y_k = g_{k+1} - g_k$  来更新近似 Hessian 矩阵。

## BFGS 更新公式

BFGS 方法通过以下更新公式计算  $B_{k+1}$ :

$$B_{k+1} = B_k + \frac{y_k y_k^\top}{y_k^\top s_k} - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k},$$

其中:

- $B_k$  是第  $k$  步的 Hessian 近似。
- $s_k = x_{k+1} - x_k$  是变量的变化。
- $y_k = g_{k+1} - g_k$  是梯度的变化。

该公式满足割线方程  $B_{k+1} s_k = y_k$ , 这是对 Hessian 矩阵  $H$  性质  $Hs \approx y$  的近似。

公式中两个修正项的直观解释:

- 第一项  $\frac{y_k y_k^\top}{y_k^\top s_k}$ : 引入了最新的梯度变化信息  $y_k$  和步长信息  $s_k$ , 是对当前 Hessian 近似的主要校正。
- 第二项  $-\frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k}$ : 移除了旧的、与新步长  $s_k$  方向相关的信息, 为新的信息腾出空间。

注 7.1 (从矩阵秩的角度理解这个式子). 1. 秩 1 矩阵的基本性质: 若  $a$  是一个非零向量, 则矩阵  $aa^\top$  是秩 1 矩阵 (因为其列向量都可由  $a$  线性表示, 秩最多为 1)。除以一个非零标量 (内积  $a^\top b$  是标量) 不会改变其秩, 因此形如  $\frac{aa^\top}{a^\top b}$  的矩阵仍为秩 1 矩阵。

2. 分解式子中的秩 1 项: 对于更新式  $B_{k+1} = B_k + \frac{y_k y_k^\top}{y_k^\top s_k} - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k}$ , 我们分别分析两个修正项:

- 项 1:  $\frac{y_k y_k^\top}{y_k^\top s_k}$  是秩 1 矩阵 (因  $y_k$  是非零向量, 且  $y_k^\top s_k \neq 0$ )。
- 项 2: 令  $v_k = B_k s_k$  ( $v_k$  是向量), 则项 2 可表示为  $\frac{v_k v_k^\top}{v_k^\top s_k}$ , 也是秩 1 矩阵 (因  $v_k$  非零且  $v_k^\top s_k \neq 0$ )。

3. 秩的变化: “秩 2 修正”: 整个更新式可看作对  $B_k$  进行两个秩 1 矩阵的加减操作, 即:

$$B_{k+1} = B_k + (\text{秩 1 矩阵}) - (\text{秩 1 矩阵})$$

根据矩阵秩的不等式: 若  $A, C$  是秩分别为  $r_A, r_C$  的矩阵, 则  $\text{rank}(A+C) \leq r_A + r_C$ , 且  $\text{rank}(A-C) \geq |r_A - r_C|$ 。这里两个修正项都是秩 1 矩阵, 且在拟牛顿法的背景下 (满足拟牛顿条件  $B_{k+1} s_k = y_k$ ), 这两个秩 1 矩阵线性无关, 因此对  $B_k$  的修正属于秩 2 修正 (即  $B_{k+1}$  与  $B_k$  的秩差最多为 2)。

4. 拟牛顿法中的秩意义: 在拟牛顿法中,  $B_k$  是 Hessian 矩阵的近似。通过这种秩 2 更新, 可以在保持矩阵对称性 (或正定性, 在一定条件下) 的同时, 逐步调整  $B_k$  的秩结构, 使其逼近真实的 Hessian 矩阵 (在算法收敛时)。初始时  $B_0$  通常取满秩矩阵 (如单位矩阵), 经过每次秩 2 修正后,  $B_k$  的秩会逐渐适应 Hessian 的秩特性, 从而实现高效的梯度近似迭代。

综上，从秩的角度看，该式子是对  $B_k$  进行秩 2 修正（由两个秩 1 矩阵的加减构成），通过这种修正来近似 Hessian 矩阵，同时满足拟牛顿条件，保证迭代的有效性。

### 推导过程

- 目标：构造一个新的  $B_{k+1}$  来近似 Hessian 矩阵，使得在每一步迭代中，拟牛顿方法的更新步长和真实牛顿法的更新步长尽量相似。
- 确保更新的对称性和正定性：
  - 对称性：若  $B_k$  对称，则  $B_{k+1}$  也是对称的，因为  $y_k y_k^\top$  和  $B_k s_k s_k^\top B_k$  都是对称矩阵。
  - 正定性：可以证明，如果初始矩阵  $B_0$  是正定的，并且线搜索满足 Wolfe 条件，那么后续所有的  $B_k$  都会保持正定。

注 7.2 ( $B_k \rightarrow B_{k+1}$  构造推导 (BFGS 框架))。一、前提与核心定义（统一符号）首先明确推导所需的基础记号、约束条件与优化目标：

#### 1. 迭代核心变量：

- 位移向量： $\mathbf{s} = \mathbf{x}_{k+1} - \mathbf{x}_k$ （第  $k$  到  $k+1$  步的迭代位移）；
- 梯度差分： $\mathbf{y} = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) = \mathbf{g}_{k+1} - \mathbf{g}_k$ （对应 Hessian 作用于位移的近似）；

#### 2. 初始近似矩阵：

- 已知  $\mathbf{B}_k \succ 0$ （对称正定的 Hessian 近似矩阵），其逆矩阵为  $\mathbf{G}_k = \mathbf{B}_k^{-1}$ （对称正定的逆 Hessian 近似）；

#### 3. 加权范数（度量“矩阵接近度”）：为量化 $\mathbf{B}_{k+1}$ 与 $\mathbf{B}_k$ 的“改变量”，定义基于 $\mathbf{G}_k$ 的加权内积与范数（保证更新对线性变换（单位缩放、坐标变化）不变，优于仅对正交变换不变的 Frobenius 范数 $\|\cdot\|_F$ ）：

$$\langle \mathbf{X}, \mathbf{Z} \rangle_{\mathbf{G}_k} = \text{tr}(\mathbf{G}_k \mathbf{X} \mathbf{G}_k \mathbf{Z}), \quad \|\mathbf{X}\|_{\mathbf{G}_k}^2 = \langle \mathbf{X}, \mathbf{X} \rangle_{\mathbf{G}_k}$$

其中  $\text{tr}(\cdot)$  为矩阵迹算子；

#### 4. 优化目标：寻找对称矩阵 $\mathbf{B}_{k+1}$ ，满足两大核心约束：

- 割线约束： $\mathbf{B}_{k+1} \mathbf{s} = \mathbf{y}$ （拟合“平均曲率”，即拟牛顿条件）；
- 最小改变量：在满足割线约束的所有对称矩阵中， $\mathbf{B}_{k+1}$  与  $\mathbf{B}_k$  在  $\|\cdot\|_{\mathbf{G}_k}$  范数下最接近。

二、第一步：“先忘”——构造子空间最优矩阵  $\mathbf{B}'$  首先在子空间  $\mathcal{S}_0 = \{\mathbf{B} \mid \mathbf{B} = \mathbf{B}^\top, \mathbf{B}\mathbf{s} = 0\}$  中，找到与  $\mathbf{B}_k$  最接近的矩阵  $\mathbf{B}'$ （即“擦除” $\mathbf{B}_k$  中沿  $\mathbf{s}$  方向的曲率信息，使其满足  $\mathbf{B}'\mathbf{s} = 0$ ）。

2.1 优化问题转化（白化变换）直接求解  $\mathbf{B}$  的优化问题较复杂，通过“白化变换”将其转化为 Frobenius 范数下的简化问题（利用  $\mathbf{G}_k = \mathbf{B}_k^{-1}$  的正定性）：

1. 白化矩阵定义：令  $\mathbf{C} = \mathbf{G}_k^{1/2} \mathbf{B} \mathbf{G}_k^{1/2}$ ，其中  $\mathbf{G}_k^{1/2}$  是  $\mathbf{G}_k$  的对称平方根（因  $\mathbf{G}_k \succ 0$ ，平方根存在且对称正定）；

2. 初始白化矩阵：代入  $\mathbf{G}_k = \mathbf{B}_k^{-1}$ ，得初始白化矩阵：

$$\mathbf{C}_k = \mathbf{G}_k^{1/2} \mathbf{B}_k \mathbf{G}_k^{1/2} = \mathbf{G}_k^{1/2} \mathbf{G}_k^{-1} \mathbf{G}_k^{1/2} = \mathbf{I}$$

其中  $\mathbf{I}$  为单位矩阵；

3. 位移白化：令  $\tilde{\mathbf{s}} = \mathbf{G}_k^{-1/2} \mathbf{s}$ （将位移向量转化到白化空间）。

(1) 范数等价转化 利用迹的循环不变性 ( $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ )，可证明  $\mathbf{B}$  与  $\mathbf{B}_k$  的加权距离等价于白化矩阵的 *Frobenius* 距离：

$$\|\mathbf{B} - \mathbf{B}_k\|_{\mathbf{G}_k}^2 = \text{tr}(\mathbf{G}_k(\mathbf{B} - \mathbf{B}_k)\mathbf{G}_k(\mathbf{B} - \mathbf{B}_k))$$

将  $\mathbf{B} - \mathbf{B}_k = \mathbf{G}_k^{-1/2}(\mathbf{C} - \mathbf{C}_k)\mathbf{G}_k^{-1/2}$ （由  $\mathbf{C} = \mathbf{G}_k^{1/2}\mathbf{B}\mathbf{G}_k^{1/2}$  变形得）代入，展开后利用迹的循环不变性抵消  $\mathbf{G}_k^{1/2}$  与  $\mathbf{G}_k^{-1/2}$ ，最终得：

$$\|\mathbf{B} - \mathbf{B}_k\|_{\mathbf{G}_k}^2 = \|\mathbf{C} - \mathbf{I}\|_F^2$$

(2) 约束等价转化 割线约束  $\mathbf{B}\mathbf{s} = 0$  可转化为白化空间的约束：

$$\mathbf{B}\mathbf{s} = 0 \implies \mathbf{G}_k^{1/2}\mathbf{B}\mathbf{G}_k^{1/2} \cdot \mathbf{G}_k^{-1/2}\mathbf{s} = \mathbf{G}_k^{1/2}\mathbf{B}\mathbf{s} = 0 \implies \mathbf{C}\tilde{\mathbf{s}} = 0$$

(3) 转化后的优化问题 原问题（加权范数下的约束优化）等价于 *Frobenius* 范数下的简化问题：

$$\min_{\mathbf{C}=\mathbf{C}^\top} \frac{1}{2} \|\mathbf{C} - \mathbf{I}\|_F^2 \quad s.t. \quad \mathbf{C}\tilde{\mathbf{s}} = 0$$

2.2 求解白化空间优化问题（正交相似变换）通过构造正交基对齐约束，简化  $\mathbf{C}$  的结构并求解最小值：

1. 构造正交矩阵  $\mathbf{Q}$ ：取正交矩阵  $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n]$ （满足  $\mathbf{Q}^\top \mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$ ），其中：

- 第一列  $\mathbf{q}_1 = \frac{\tilde{\mathbf{s}}}{\|\tilde{\mathbf{s}}\|}$ （ $\tilde{\mathbf{s}}$  的单位向量，因  $\mathbf{s} \neq 0$  且  $\mathbf{G}_k^{-1/2}$  可逆，故  $\|\tilde{\mathbf{s}}\| \neq 0$ ）；
- 其余列  $\mathbf{q}_2, \dots, \mathbf{q}_n$  为  $\mathbf{q}_1$  正交补空间的标准正交基。

此时  $\mathbf{Q}^\top \tilde{\mathbf{s}} = \|\tilde{\mathbf{s}}\| \mathbf{e}_1$ （ $\mathbf{e}_1 = (1, 0, \dots, 0)^\top$  为标准基向量），因  $\mathbf{q}_1^\top \tilde{\mathbf{s}} = \|\tilde{\mathbf{s}}\|$ ， $\mathbf{q}_j^\top \tilde{\mathbf{s}} = 0$ （ $j \geq 2$ ）。

2. 分块对角化  $\mathbf{C}$ ：令  $\mathbf{D} = \mathbf{Q}^\top \mathbf{C} \mathbf{Q}$ （正交相似变换），则：

- 对称性：因  $\mathbf{C} = \mathbf{C}^\top$  且  $\mathbf{Q}$  正交，故  $\mathbf{D} = \mathbf{D}^\top$ ；
- 范数不变性：*Frobenius* 范数在正交变换下不变，即  $\|\mathbf{C} - \mathbf{I}\|_F^2 = \|\mathbf{D} - \mathbf{I}\|_F^2$ ；
- 约束转化： $\mathbf{C}\tilde{\mathbf{s}} = 0 \implies \mathbf{D}(\mathbf{Q}^\top \tilde{\mathbf{s}}) = 0 \implies \mathbf{D}(\|\tilde{\mathbf{s}}\| \mathbf{e}_1) = 0 \implies \mathbf{D}\mathbf{e}_1 = 0$ （因  $\|\tilde{\mathbf{s}}\| \neq 0$ ）。

3. 最小化目标函数：由  $\mathbf{D} = \mathbf{D}^\top$  且  $\mathbf{D}\mathbf{e}_1 = 0$ ，可知  $\mathbf{D}$  的第一列全为 0，对称性导致第一行也全为 0，故  $\mathbf{D}$  可分块为：

$$\mathbf{D} = \begin{bmatrix} 0 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{M} \end{bmatrix}, \quad \mathbf{M} = \mathbf{M}^\top \in \mathbb{R}^{(n-1) \times (n-1)}$$

代入目标函数：

$$\|\mathbf{D} - \mathbf{I}\|_F^2 = \left\| \begin{bmatrix} -1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{M} - \mathbf{I}_{n-1} \end{bmatrix} \right\|_F^2 = 1 + \|\mathbf{M} - \mathbf{I}_{n-1}\|_F^2$$

要最小化该式，需  $\|\mathbf{M} - \mathbf{I}_{n-1}\|_F^2 = 0$ ，即  $\mathbf{M} = \mathbf{I}_{n-1}$ 。因此，白化空间的最优解为：

$$\mathbf{D}^* = \begin{bmatrix} 0 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{I}_{n-1} \end{bmatrix} = \mathbf{I} - \mathbf{e}_1 \mathbf{e}_1^\top$$

### 2.3 反变换回 $\mathbf{B}'$ （从白化空间到原空间）

1. 从  $\mathbf{D}^*$  到  $\mathbf{C}^*$ ：由  $\mathbf{D}^* = \mathbf{Q}^\top \mathbf{C}^* \mathbf{Q}$ ，得  $\mathbf{C}^* = \mathbf{Q} \mathbf{D}^* \mathbf{Q}^\top$ 。代入  $\mathbf{D}^* = \mathbf{I} - \mathbf{e}_1 \mathbf{e}_1^\top$ ，并利用  $\mathbf{Q} \mathbf{e}_1 = \mathbf{q}_1 = \frac{\tilde{\mathbf{s}}}{\|\tilde{\mathbf{s}}\|}$ ，得：

$$\mathbf{C}^* = \mathbf{I} - \mathbf{Q} \mathbf{e}_1 \mathbf{e}_1^\top \mathbf{Q}^\top = \mathbf{I} - \frac{\tilde{\mathbf{s}} \tilde{\mathbf{s}}^\top}{\|\tilde{\mathbf{s}}\|^2}$$

2. 从  $\mathbf{C}^*$  到  $\mathbf{B}'$ ：由  $\mathbf{C}^* = \mathbf{G}_k^{1/2} \mathbf{B}' \mathbf{G}_k^{1/2}$ ，得  $\mathbf{B}' = \mathbf{G}_k^{-1/2} \mathbf{C}^* \mathbf{G}_k^{-1/2}$ 。代入  $\mathbf{C}^*$  并分拆两项展开：

- 第一项： $\mathbf{G}_k^{-1/2} \mathbf{I} \mathbf{G}_k^{-1/2} = \mathbf{G}_k^{-1} = \mathbf{B}_k$ （因  $\mathbf{G}_k = \mathbf{B}_k^{-1}$ ）；
- 第二项： $\mathbf{G}_k^{-1/2} \cdot \frac{\tilde{\mathbf{s}} \tilde{\mathbf{s}}^\top}{\|\tilde{\mathbf{s}}\|^2} \cdot \mathbf{G}_k^{-1/2}$ 。

进一步计算第二项的分子与分母：

- 分子： $\mathbf{G}_k^{-1/2} \tilde{\mathbf{s}} = \mathbf{G}_k^{-1/2} \cdot \mathbf{G}_k^{-1/2} \mathbf{s} = \mathbf{G}_k^{-1} \mathbf{s} = \mathbf{B}_k \mathbf{s}$ ，故分子为  $(\mathbf{B}_k \mathbf{s})(\mathbf{B}_k \mathbf{s})^\top = \mathbf{B}_k \mathbf{s} \mathbf{s}^\top \mathbf{B}_k$ ；
- 分母： $\|\tilde{\mathbf{s}}\|^2 = \tilde{\mathbf{s}}^\top \tilde{\mathbf{s}} = (\mathbf{G}_k^{-1/2} \mathbf{s})^\top (\mathbf{G}_k^{-1/2} \mathbf{s}) = \mathbf{s}^\top \mathbf{G}_k^{-1} \mathbf{s} = \mathbf{s}^\top \mathbf{B}_k \mathbf{s}$ 。

因此，“先忘”步骤的结果为：

$$\mathbf{B}' = \mathbf{B}_k - \frac{\mathbf{B}_k \mathbf{s} \mathbf{s}^\top \mathbf{B}_k}{\mathbf{s}^\top \mathbf{B}_k \mathbf{s}}$$

验证约束： $\mathbf{B}' \mathbf{s} = \mathbf{B}_k \mathbf{s} - \frac{\mathbf{B}_k \mathbf{s} \mathbf{s}^\top \mathbf{B}_k \mathbf{s}}{\mathbf{s}^\top \mathbf{B}_k \mathbf{s}} = \mathbf{B}_k \mathbf{s} - \mathbf{B}_k \mathbf{s} = \mathbf{0}$ ，满足子空间约束。

三、第二步：“再写”——构造修正项  $\Delta^+$  在  $\mathbf{B}'$  的基础上，添加最小改动的对称矩阵  $\Delta^+$ ，使  $\mathbf{B}_{k+1} = \mathbf{B}' + \Delta^+$  满足割线约束  $\mathbf{B}_{k+1} \mathbf{s} = \mathbf{y}$ 。

3.1 白化空间的修正项  $\Delta_C^+$  因  $\mathbf{B}' \mathbf{s} = \mathbf{0}$ ，故割线约束等价于  $\Delta^+ \mathbf{s} = \mathbf{y}$ 。转化到白化空间：

1. 梯度差分白化：令  $\tilde{\mathbf{y}} = \mathbf{G}_k^{1/2} \mathbf{y}$ （与位移白化对应）；
2. 约束转化： $\Delta_C^+ \tilde{\mathbf{s}} = \tilde{\mathbf{y}}$ （推导同 2.1.2，利用  $\Delta^+ = \mathbf{G}_k^{-1/2} \Delta_C^+ \mathbf{G}_k^{-1/2}$ ）。

为满足约束且最小化改动，选择对称秩-1 矩阵作为  $\Delta_C^+$ ：

$$\Delta_C^+ = \frac{\tilde{\mathbf{y}} \tilde{\mathbf{y}}^\top}{\tilde{\mathbf{y}}^\top \tilde{\mathbf{s}}}$$

验证约束： $\Delta_C^+ \tilde{\mathbf{s}} = \frac{\tilde{\mathbf{y}} (\tilde{\mathbf{y}}^\top \tilde{\mathbf{s}})}{\tilde{\mathbf{y}}^\top \tilde{\mathbf{s}}} = \tilde{\mathbf{y}}$ ，完全满足。

3.2 反变换回  $\Delta^+$ （原空间修正项）由  $\Delta^+ = \mathbf{G}_k^{-1/2} \Delta_C^+ \mathbf{G}_k^{-1/2}$ ，代入  $\Delta_C^+$  展开：

- 分子： $\mathbf{G}_k^{-1/2} \tilde{\mathbf{y}} = \mathbf{G}_k^{-1/2} \cdot \mathbf{G}_k^{1/2} \mathbf{y} = \mathbf{y}$ ，故分子为  $\mathbf{y} \mathbf{y}^\top$ ；
- 分母： $\tilde{\mathbf{y}}^\top \tilde{\mathbf{s}} = (\mathbf{G}_k^{1/2} \mathbf{y})^\top (\mathbf{G}_k^{-1/2} \mathbf{s}) = \mathbf{y}^\top \mathbf{G}_k^{1/2} \mathbf{G}_k^{-1/2} \mathbf{s} = \mathbf{y}^\top \mathbf{s}$ 。

因此，原空间的修正项为：

$$\Delta^+ = \frac{\mathbf{y}\mathbf{y}^\top}{\mathbf{y}^\top \mathbf{s}}$$

验证约束： $\Delta^+ \mathbf{s} = \frac{\mathbf{y}(\mathbf{y}^\top \mathbf{s})}{\mathbf{y}^\top \mathbf{s}} = \mathbf{y}$ ，满足割线约束要求。

四、最终构造： $\mathbf{B}_{k+1}$  的表达式与正定性验证 4.1  $\mathbf{B}_{k+1}$  的最终公式 合并“先忘”步骤的  $\mathbf{B}'$  与“再写”步骤的  $\Delta^+$ ，得到 BFGS 方法中 Hessian 近似矩阵的更新公式 (B-form)：

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{\mathbf{B}_k \mathbf{s} \mathbf{s}^\top \mathbf{B}_k}{\mathbf{s}^\top \mathbf{B}_k \mathbf{s}} + \frac{\mathbf{y}\mathbf{y}^\top}{\mathbf{y}^\top \mathbf{s}}$$

4.2 正定性验证（保证下降方向）若满足曲率条件  $\mathbf{s}^\top \mathbf{y} > 0$ （强 Wolfe 线搜索可保证），则  $\mathbf{B}_{k+1} \succ 0$ （对称正定），证明如下：对任意非零向量  $\mathbf{z}$ ，代入  $\mathbf{B}_{k+1}$  的表达式得：

$$\mathbf{z}^\top \mathbf{B}_{k+1} \mathbf{z} = \mathbf{z}^\top \mathbf{B}_k \mathbf{z} - \frac{(\mathbf{z}^\top \mathbf{B}_k \mathbf{s})^2}{\mathbf{s}^\top \mathbf{B}_k \mathbf{s}} + \frac{(\mathbf{z}^\top \mathbf{y})^2}{\mathbf{y}^\top \mathbf{s}}$$

由 Cauchy-Schwarz 不等式， $\frac{(\mathbf{z}^\top \mathbf{B}_k \mathbf{s})^2}{\mathbf{s}^\top \mathbf{B}_k \mathbf{s}} \leq \mathbf{z}^\top \mathbf{B}_k \mathbf{z}$ （等号仅当  $\mathbf{z}$  与  $\mathbf{s}$  线性相关时成立）。结合  $\mathbf{s}^\top \mathbf{y} > 0$ ，第二项  $\frac{(\mathbf{z}^\top \mathbf{y})^2}{\mathbf{y}^\top \mathbf{s}} \geq 0$ ，且仅当  $\mathbf{z}^\top \mathbf{y} = 0$  时为 0。

- 若  $\mathbf{z}^\top \mathbf{y} \neq 0$ ： $\mathbf{z}^\top \mathbf{B}_{k+1} \mathbf{z} > 0$ ；
- 若  $\mathbf{z}^\top \mathbf{y} = 0$ ： $\mathbf{z}^\top \mathbf{B}_{k+1} \mathbf{z} = \mathbf{z}^\top \mathbf{B}_k \mathbf{z} - \frac{(\mathbf{z}^\top \mathbf{B}_k \mathbf{s})^2}{\mathbf{s}^\top \mathbf{B}_k \mathbf{s}} \geq 0$ ，且仅当  $\mathbf{z} = 0$  时取等号（因  $\mathbf{B}_k \succ 0$ ）。

综上， $\mathbf{B}_{k+1} \succ 0$ ，保证后续搜索方向  $\mathbf{p}_k = -\mathbf{G}_{k+1} \mathbf{g}_k$  ( $\mathbf{G}_{k+1} = \mathbf{B}_{k+1}^{-1}$ ) 为下降方向。

### 7.3.3 准牛顿法的其他变种

除了 BFGS，拟牛顿法还有其他变种，例如 DFP（Davidon-Fletcher-Powell）方法。DFP 的更新规则与 BFGS 相似，但它直接更新 Hessian 的逆矩阵近似  $\mathbf{G}_k$ 。DFP 方法的更新公式为：

$$\mathbf{G}_{k+1} = \mathbf{G}_k + \frac{\mathbf{s}_k \mathbf{s}_k^\top}{\mathbf{s}_k^\top \mathbf{y}_k} - \frac{\mathbf{G}_k \mathbf{y}_k \mathbf{y}_k^\top \mathbf{G}_k}{\mathbf{y}_k^\top \mathbf{G}_k \mathbf{y}_k}$$

DFP 与 BFGS 实际上是对偶关系。BFGS 更新 Hessian 的近似  $\mathbf{B}_k$ ，而 DFP 更新其逆的近似  $\mathbf{G}_k$ 。

## 7.4 收敛性

### 7.4.1 BFGS 的收敛性

BFGS 的收敛性需分目标函数类型（二次/非二次）和线搜索策略（精确/不精确）讨论，核心依赖“拟牛顿条件”和“近似 Hessian 矩阵的正定性”。

#### 1. 二次函数下的收敛性

若目标函数为严格凸二次函数（即 Hessian 矩阵  $\mathbf{H}$  正定且恒定），且采用精确线搜索（即每次步长选择使目标函数沿搜索方向最小化），BFGS 具有以下收敛性质：

- **有限终止性**：对于  $n$  维二次函数，BFGS 最多迭代  $n$  步即可收敛到全局最优解。  
原因：二次函数的 Hessian 恒定，BFGS 通过迭代更新的  $\mathbf{B}_k$  会逐步逼近真实  $\mathbf{H}$ ，当  $\mathbf{B}_k = \mathbf{H}$  时，一步即可达到最优，而理论上最多  $n$  步可完成逼近。

- **正定性保持**：若初始近似矩阵  $B_0$  正定，则所有迭代过程中的  $B_k$  均保持正定，确保搜索方向始终为”下降方向”（即  $-B_k^{-1}g_k$  与梯度反向），避免迭代发散。

## 2. 非二次函数下的收敛性

实际优化问题多为非二次函数（如机器学习中的损失函数），此时需假设目标函数满足**光滑性和凸性条件**（如二阶导数 Lipschitz 连续、Hessian 在最优解附近正定），BFGS 的收敛性质为：

- **局部收敛性**：若初始点  $x_0$  足够接近全局最优解  $x^*$ ，且采用精确/不精确线搜索（如 Wolfe 条件），BFGS 会收敛到  $x^*$ 。  
关键前提：最优解  $x^*$  处的 Hessian  $H^* = \nabla^2 f(x^*)$  正定，确保迭代过程中梯度方向的”有效性”。
- **全局收敛性（凸函数下）**：若目标函数为**严格凸函数**，且采用精确线搜索或满足 Wolfe 条件的不精确线搜索，BFGS 可实现**全局收敛**（即无论初始点  $x_0$  如何，最终均收敛到全局最优解）。  
非凸函数下：仅能保证**局部收敛**（可能收敛到局部最优解），这是多数无约束优化方法的共性（除非结合全局优化策略，如随机初始化）。

### 7.4.2 BFGS 的收敛速度：介于梯度下降与牛顿法之间，逼近牛顿法

收敛速度衡量迭代序列  $\{x_k\}$  趋近最优解  $x^*$  的快慢，常用”收敛阶”定义（如线性收敛、超线性收敛、二次收敛）。BFGS 的收敛速度需结合函数类型分析：

#### 1. 收敛阶的定义（参考基准）

- **线性收敛**：存在常数  $0 < c < 1$ ，使  $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = c$ （如梯度下降法）。
- **超线性收敛**： $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$ （比线性快）。
- **二次收敛**：存在常数  $c > 0$ ，使  $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} = c$ （如牛顿法，最快）。

#### 2. BFGS 的收敛速度

- **二次函数下**：若采用精确线搜索，BFGS 对  $n$  维二次函数是**有限收敛**（最多  $n$  步），本质上比二次收敛更快（无需无限迭代）。
- **非二次函数下**：在最优解  $x^*$  附近（满足 Hessian 正定且 Lipschitz 连续），BFGS 是**超线性收敛**。

BFGS 的近似 Hessian  $B_k$  满足  $\lim_{k \rightarrow \infty} \frac{\|(B_k - H^*)p_k\|}{\|p_k\|} = 0$ （Dennis-Moré 条件），使得迭代步长逐渐接近牛顿法的最优步长，因此收敛速度接近牛顿法，但无需显式计算 Hessian。

### 7.4.3 BFGS 的最优性：理论性质与实际性能的”最优平衡”

BFGS 的”最优性”并非指它在所有场景下都是”最好”的优化方法，而是指它在**计算成本、数值稳定性、收敛性能**三者间达到了工程应用中的”最优权衡”，具体体现在以下方面：

#### 1. 理论最优性：满足拟牛顿法的核心目标

拟牛顿法的核心目标是”用梯度信息逼近 Hessian，以降低牛顿法的计算成本”，BFGS 在这一目标下满足：

- **拟牛顿条件的严格满足：**每次更新的  $B_{k+1}$  均严格满足  $B_{k+1}s_k = y_k$ （这是逼近 Hessian 的核心条件），确保  $B_k$  对真实 Hessian 的逼近是“有效”的。
- **正定性的严格保持：**若  $B_0$  正定且线搜索满足“充分下降条件”（如 Wolfe 条件），则所有  $B_k$  均正定，避免搜索方向变为“上升方向”（这是 DFP 等方法有时难以保证的，BFGS 的数值稳定性更优）。

## 2. 实际应用中的最优性：兼顾效率与稳定性

- **计算成本最优：**
  - 每次迭代仅需计算 1 次梯度（成本  $O(n)$ ），更新  $B_k$  的成本为  $O(n^2)$ （无需计算 Hessian 的  $O(n^3)$  成本）。
  - 存储成本为  $O(n^2)$ （仅需存储  $B_k$  或其逆矩阵  $G_k$ ），适用于中大规模问题（ $n$  从几百到几万）。
- **数值稳定性最优：**  
相比 DFP、SR1 等其他拟牛顿法，BFGS 对“线搜索误差”和“梯度噪声”的容忍度更高，即使采用不精确线搜索（实际应用中常用），也不易出现  $B_k$  奇异或迭代发散的情况。
- **收敛性能最优：**  
在相同计算成本下，BFGS 的收敛速度远快于梯度下降（线性收敛），且接近牛顿法（二次收敛），同时避免了牛顿法计算 Hessian 和求解线性方程组的高昂成本，因此在机器学习、工程优化等领域成为“首选方法”之一。

**定义 7.1** (Wolfe 条件). Wolfe 条件是不精确线搜索的核心准则。设目标函数为  $f(x)$ ，梯度为  $g(x)$ ，搜索方向为  $p_k$ ，则步长  $\alpha_k$  需满足：

1. **Armijo 条件（充分下降条件）：**

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k g_k^\top p_k$$

2. **曲率条件（步长不太小条件）：**

$$g(x_k + \alpha_k p_k)^\top p_k \geq c_2 g_k^\top p_k$$

其中  $0 < c_1 < c_2 < 1$ （典型值  $c_1 = 10^{-4}, c_2 = 0.9$ ）。

## 7.5 伪代码

**算法 7.1** (BFGS 拟牛顿优化算法). **输入：**目标函数  $f(x)$ ，梯度函数  $\nabla f(x)$ ，初始点  $x_0 \in \mathbb{R}^n$ ，收敛阈值  $\epsilon > 0$ ，最大迭代次数  $T$

**输出：**最优解近似  $x^*$

1. **初始化：**

- 令  $k = 0$ ,  $x_k = x_0$
  - 计算梯度  $g_k = \nabla f(x_k)$
  - 初始化 *Hessian* 逆矩阵近似  $G_k = I_n$  ( $n \times n$  单位矩阵)
2. 收敛判断: 若  $\|g_k\| < \epsilon$ , 则输出  $x^* = x_k$  并终止
3. 迭代主循环 (当  $k < T$  时):
- (a) 计算搜索方向:  $p_k = -G_k g_k$  (下降方向)
- (b) 线搜索: 寻找步长  $\alpha_k > 0$ , 使其满足 *Wolfe* 条件:
- $$\begin{cases} f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k g_k^\top p_k \\ \nabla f(x_k + \alpha_k p_k)^\top p_k \geq c_2 g_k^\top p_k \end{cases}$$
- (c) 更新迭代点:  $x_{k+1} = x_k + \alpha_k p_k$
- (d) 更新梯度:  $g_{k+1} = \nabla f(x_{k+1})$
- (e) 计算增量:
- 变量增量:  $s_k = x_{k+1} - x_k$
  - 梯度增量:  $y_k = g_{k+1} - g_k$
- (f) 检查正定性条件: 若  $y_k^\top s_k \leq 0$  (不满足曲率条件), 则令  $G_{k+1} = G_k$  (跳过更新); 否则, 按 *BFGS* 公式更新 *Hessian* 逆近似:
- $$G_{k+1} = \left( I - \frac{s_k y_k^\top}{y_k^\top s_k} \right) G_k \left( I - \frac{y_k s_k^\top}{y_k^\top s_k} \right) + \frac{s_k s_k^\top}{y_k^\top s_k}$$
- (g) 迭代更新:  $k = k + 1$ , 返回步骤 2
4. 终止: 若达到最大迭代次数, 输出  $x^* = x_k$

注 7.3 (逆 *Hessian* 近似形式 (G-form) 的推导与使用原因). 在 *BFGS* 拟牛顿法中, 逆 *Hessian* 近似形式 (*G-form*) 指直接对逆 *Hessian* 近似矩阵  $\mathbf{G}_k = \mathbf{B}_k^{-1}$  ( $\mathbf{B}_k$  为 *Hessian* 近似矩阵) 进行更新, 而非对  $\mathbf{B}_k$  (*B-form*) 更新。以下先推导 *G-form* 的数学表达式, 再详细说明为何优先使用逆 *BFGS* 形式。

一、*G-form* (逆 *Hessian* 近似) 的数学推导 *G-form* 的更新公式需从 *B-form* ( $\mathbf{B}_{k+1}$  的更新) 出发, 利用 *Woodbury* 矩阵求逆公式 (适用于秩修正矩阵的逆计算) 推导  $\mathbf{G}_{k+1} = \mathbf{B}_{k+1}^{-1}$  的表达式, 核心步骤如下:

#### 1. 已知前提与工具

- *B-form* 更新公式 (已推导的 *Hessian* 近似更新):

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{\mathbf{B}_k \mathbf{s}_k \mathbf{s}_k^\top \mathbf{B}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k}$$

其中  $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$  (位移),  $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$  (梯度差分), 且  $\mathbf{B}_k \succ 0$  (对称正定, 故  $\mathbf{G}_k = \mathbf{B}_k^{-1}$  存在)。

- **Woodbury 公式** (秩- $r$  修正矩阵的逆): 若矩阵  $\mathbf{A}$  可逆,  $\mathbf{U}, \mathbf{V}$  为  $n \times r$  矩阵,  $\mathbf{C}$  为  $r \times r$  可逆矩阵, 则:

$$(\mathbf{A} + \mathbf{UCV}^\top)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{V}^\top\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^\top\mathbf{A}^{-1}$$

本推导中  $\mathbf{B}_{k+1} = \mathbf{B}_k + \mathbf{UMU}^\top$  (秩-2 修正,  $\mathbf{U}$  为  $n \times 2$  矩阵,  $\mathbf{M}$  为  $2 \times 2$  对角矩阵), 符合 Woodbury 公式适用场景。

2. 步骤 1: 将 *B-form* 改写为“原矩阵 + 秩修正”形式 令:

- 秩修正矩阵的列向量:  $\mathbf{U} = [\mathbf{y}_k, \mathbf{B}_k \mathbf{s}_k]$  ( $n \times 2$  矩阵);
- 对角系数矩阵:  $\mathbf{M} = \text{diag}\left(\frac{1}{\mathbf{y}_k^\top \mathbf{s}_k}, -\frac{1}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\right)$  ( $2 \times 2$  矩阵)。

则 *B-form* 可改写为:

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \mathbf{UMU}^\top$$

(验证:  $\mathbf{UMU}^\top = \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\mathbf{B}_k \mathbf{s}_k \mathbf{s}_k^\top \mathbf{B}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}$ , 与 *B-form* 一致)。

3. 步骤 2: 应用 Woodbury 公式求  $\mathbf{G}_{k+1} = \mathbf{B}_{k+1}^{-1}$  将  $\mathbf{A} = \mathbf{B}_k$ 、 $\mathbf{V} = \mathbf{U}$ 、 $\mathbf{C} = \mathbf{M}$  代入 Woodbury 公式, 且  $\mathbf{G}_k = \mathbf{B}_k^{-1}$ , 展开计算:

1. 计算  $\mathbf{C}^{-1} = \mathbf{M}^{-1}$  (对角矩阵逆为对角元素倒数):

$$\mathbf{M}^{-1} = \text{diag}(\mathbf{y}_k^\top \mathbf{s}_k, -\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k)$$

2. 计算  $\mathbf{V}^\top \mathbf{A}^{-1} \mathbf{U} = \mathbf{U}^\top \mathbf{G}_k \mathbf{U}$  ( $2 \times 2$  矩阵):

$$\mathbf{U}^\top \mathbf{G}_k \mathbf{U} = \begin{bmatrix} \mathbf{y}_k^\top \mathbf{G}_k \mathbf{y}_k & \mathbf{y}_k^\top \mathbf{s}_k \\ \mathbf{s}_k^\top \mathbf{y}_k & \mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k \end{bmatrix}$$

(因  $\mathbf{G}_k \mathbf{B}_k = \mathbf{I}$ , 故  $\mathbf{G}_k \mathbf{B}_k \mathbf{s}_k = \mathbf{s}_k$ )。

3. 计算  $\mathbf{C}^{-1} + \mathbf{U}^\top \mathbf{G}_k \mathbf{U}$  (记为  $\mathbf{S}$ ,  $2 \times 2$  矩阵): 令  $\rho_k = \frac{1}{\mathbf{y}_k^\top \mathbf{s}_k}$  (简化符号), 则  $\mathbf{y}_k^\top \mathbf{s}_k = \frac{1}{\rho_k}$ , 代入得:

$$\mathbf{S} = \begin{bmatrix} \mathbf{y}_k^\top \mathbf{G}_k \mathbf{y}_k + \frac{1}{\rho_k} & \frac{1}{\rho_k} \\ \frac{1}{\rho_k} & 0 \end{bmatrix}$$

求  $\mathbf{S}^{-1}$  (2 阶矩阵逆公式), 并代入 Woodbury 公式展开、化简后, 最终得到 *G-form* 的紧凑更新公式:

$$\mathbf{G}_{k+1} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^\top) \mathbf{G}_k (\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top) + \rho_k \mathbf{s}_k \mathbf{s}_k^\top$$

4. *G-form* 的关键性质验证

- 割线条件:  $\mathbf{G}_{k+1} \mathbf{y}_k = \mathbf{s}_k$  (与 *B-form* 的  $\mathbf{B}_{k+1} \mathbf{s}_k = \mathbf{y}_k$  等价, 保证曲率拟合);
- 对称正定: 若  $\mathbf{G}_k \succ 0$  且  $\mathbf{s}_k^\top \mathbf{y}_k > 0$  (曲率条件), 则  $\mathbf{G}_{k+1} \succ 0$  (保证搜索方向为下降方向)。

二、总结 逆 BFGS 形式 (*G-form*) 的核心价值在于“以更低的计算/存储成本, 保留牛顿方向的高质量性”:

1. 推导上，通过 Woodbury 公式从  $B$ -form 转化而来，严格满足拟牛顿的割线条件与正定性；
2. 实践上，其搜索方向计算仅需矩阵-向量乘法，且  $L$ -BFGS 基于  $G$ -form 实现了“有限记忆 + 低复杂度”，成为大规模无约束优化（如深度学习、科学计算）的默认基线方法。

相比之下， $B$ -form 因需解线性方程组、存储成本高，仅在小规模问题（ $n \ll 10^3$ ）中偶尔使用，工程价值远低于  $G$ -form。

## 7.6 L-BFGS（有限记忆二阶近似）

L-BFGS（Limited-Memory BFGS）的核心是放弃显式存储  $n \times n$  的逆 Hessian 近似  $\mathbf{G}_k$ ，仅保留最近  $m$  对曲率信息  $(\mathbf{s}_i, \mathbf{y}_i)$ （ $m \in [5, 20]$ ，远小于变量维度  $n$ ），通过“两环递推”在线构造  $\mathbf{G}_k \mathbf{v}$ （ $\mathbf{v}$  为任意向量，通常取梯度  $\mathbf{g}_k$ ）的结果，实现“低内存 + 低复杂度”的二阶优化。

### 7.6.1 推导起点：BFGS 的 $G$ -form 递推关系

L-BFGS 源于 BFGS 的逆 Hessian 更新公式（ $G$ -form）。回顾已推导的  $G$ -form：

$$\mathbf{G}_{k+1} = \underbrace{(\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^\top)}_{V_k} \mathbf{G}_k \underbrace{(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top)}_{W_k} + \underbrace{\rho_k \mathbf{s}_k \mathbf{s}_k^\top}_{R_k} \quad (1)$$

其中：

- $V_k, W_k$  为“投影矩阵”（秩- $n-1$ ，用于消除旧曲率中与  $\mathbf{s}_k$  相关的冗余信息）；
- $R_k$  为“秩-1 修正项”（用于注入新曲率  $(\mathbf{s}_k, \mathbf{y}_k)$ ）。

关键观察： $\mathbf{G}_k$  的“乘向量算子”属性 L-BFGS 不直接存储  $\mathbf{G}_k$ ，而是关注  $\mathbf{G}_k$  对任意向量  $\mathbf{v}$  的作用（记为  $\mathbf{G}_k \mathbf{v}$ ）。对式 (1) 两边同时右乘  $\mathbf{v}$ ，展开得：

$$\mathbf{G}_{k+1} \mathbf{v} = V_k \cdot (\mathbf{G}_k \cdot (W_k^\top \mathbf{v})) + R_k \mathbf{v} \quad (2)$$

上式揭示： $\mathbf{G}_{k+1} \mathbf{v}$  可由  $\mathbf{G}_k$  对“预处理后的  $\mathbf{v}$ ”（即  $W_k^\top \mathbf{v}$ ）的作用，再经  $V_k$  投影和  $R_k$  修正得到。递推展开该式，即可用历史  $\{(\mathbf{s}_i, \mathbf{y}_i)\}$  和初始  $\mathbf{G}_0$  表示  $\mathbf{G}_k \mathbf{v}$ ，无需显式  $\mathbf{G}_k$ 。

### 7.6.2 Step 1：递推展开 $\mathbf{G}_k \mathbf{v}$

假设保留最近  $m$  对曲率信息  $(\mathbf{s}_{k-m}, \mathbf{y}_{k-m}), \dots, (\mathbf{s}_{k-1}, \mathbf{y}_{k-1})$ ，对式 (2) 从  $\mathbf{G}_k$  反向递推至  $\mathbf{G}_{k-m}$ （初始逆 Hessian 近似，通常取缩放单位矩阵  $\mathbf{G}_0 = \gamma_k \mathbf{I}$ ）：

1. 对  $\mathbf{G}_k$ ： $\mathbf{G}_k \mathbf{v} = V_{k-1} \cdot (\mathbf{G}_{k-1} \cdot (W_{k-1}^\top \mathbf{v})) + R_{k-1} \mathbf{v}$
2. 对  $\mathbf{G}_{k-1}$ ： $\mathbf{G}_{k-1} \cdot (W_{k-1}^\top \mathbf{v}) = V_{k-2} \cdot (\mathbf{G}_{k-2} \cdot (W_{k-2}^\top \cdot W_{k-1}^\top \mathbf{v})) + R_{k-2} \cdot (W_{k-1}^\top \mathbf{v})$
3. 以此类推，直到  $\mathbf{G}_{k-m}$ ：

$$\mathbf{G}_k \mathbf{v} = V_{k-1} V_{k-2} \dots V_{k-m} \cdot (\mathbf{G}_{k-m} \cdot (W_{k-m}^\top \dots W_{k-2}^\top W_{k-1}^\top \mathbf{v})) + \text{秩-1 修正项总和} \quad (3)$$

式 (3) 可拆分为两部分：

- 右乘链： $W_{k-m}^\top \dots W_{k-1}^\top \mathbf{v}$ （对应“反向环”，处理所有  $W_i^\top$  对  $\mathbf{v}$  的预处理）；
- 左乘链 + 修正项： $V_{k-m} \dots V_{k-1} \cdot (\mathbf{G}_0 \cdot \text{右乘结果}) + \text{修正项}$ （对应“正向环”，处理  $V_i$  投影和  $R_i$  修正）。

### 7.6.3 Step 2: 反向环 (Right Loop) ——处理右乘链

反向环的目标是计算“预处理后的向量” $\mathbf{q}$ ，即式 (3) 中  $\mathbf{G}_0$  的输入： $\mathbf{q} = W_{k-m}^\top \dots W_{k-1}^\top \mathbf{v}$ 。

**3.1 单个  $W_i^\top$  的作用** 由  $W_i = \mathbf{I} - \rho_i \mathbf{y}_i \mathbf{s}_i^\top$ ，其转置为  $W_i^\top = \mathbf{I} - \rho_i \mathbf{s}_i \mathbf{y}_i^\top$  (因  $\mathbf{s}_i \mathbf{y}_i^\top$  的转置为  $\mathbf{y}_i \mathbf{s}_i^\top$ )。对任意向量  $\mathbf{z}$ ， $W_i^\top \mathbf{z}$  的计算为：

$$W_i^\top \mathbf{z} = \mathbf{z} - \rho_i \mathbf{s}_i (\mathbf{y}_i^\top \mathbf{z}) \quad (4)$$

但结合递推顺序 (从  $i = k-1$  到  $i = k-m$ )，需定义中间系数  $\alpha_i$  简化计算：令  $\alpha_i = \rho_i (\mathbf{s}_i^\top \mathbf{z})$ ，则式 (4) 可改写为：

$$\mathbf{z} \leftarrow \mathbf{z} - \alpha_i \mathbf{y}_i \quad (5)$$

(注： $\alpha_i$  记录了  $\mathbf{s}_i$  与当前  $\mathbf{z}$  的内积信息，后续正向环需复用该系数，避免重复计算。)

**3.2 反向环完整流程初始化**  $\mathbf{q} = \mathbf{v}$  (初始向量，若计算搜索方向则  $\mathbf{v} = \mathbf{g}_k$ )，从最近的曲率对开始，自后向前迭代 ( $i = k-1, k-2, \dots, k-m$ )：

$$\begin{cases} \alpha_i = \rho_i \cdot \mathbf{s}_i^\top \mathbf{q} \\ \mathbf{q} = \mathbf{q} - \alpha_i \cdot \mathbf{y}_i \end{cases} \quad (6)$$

作用：通过  $m$  次迭代，将所有  $W_i^\top$  的作用“吸收”到  $\mathbf{q}$  中，得到  $\mathbf{q} = W_{k-m}^\top \dots W_{k-1}^\top \mathbf{v}$ ，为后续  $\mathbf{G}_0$  作用做准备。

### 7.6.4 Step 3: 初始缩放 (Initial Scaling) —— $\mathbf{G}_0$ 的作用

L-BFGS 的初始逆 Hessian 近似  $\mathbf{G}_0$  不直接取  $\mathbf{I}$  (单位矩阵)，而是取缩放单位矩阵，目的是让  $\mathbf{G}_0$  的尺度接近真实逆 Hessian  $\nabla^2 f(\mathbf{x}_k)^{-1}$  的尺度，提升方向质量。

**4.1 缩放系数  $\gamma_k$  的选择** 缩放系数  $\gamma_k$  由最近一次的曲率对  $(\mathbf{s}_{k-1}, \mathbf{y}_{k-1})$  计算，满足“模拟 Hessian 的对角尺度”：

$$\gamma_k = \frac{\mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}}{\mathbf{y}_{k-1}^\top \mathbf{y}_{k-1}} \quad (7)$$

物理意义： $\mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}$  是“平均曲率”的近似 ( $\mathbf{y}_{k-1} \approx \nabla^2 f(\bar{\mathbf{x}}) \mathbf{s}_{k-1}$ ，故  $\mathbf{s}_{k-1}^\top \mathbf{y}_{k-1} \approx \mathbf{s}_{k-1}^\top \nabla^2 f(\bar{\mathbf{x}}) \mathbf{s}_{k-1}$ )， $\gamma_k$  相当于  $\nabla^2 f(\bar{\mathbf{x}})^{-1}$  的“对角平均”，确保  $\mathbf{G}_0 = \gamma_k \mathbf{I}$  的尺度合理。

**4.2 初始缩放计算** 对反向环得到的  $\mathbf{q}$ ，施加  $\mathbf{G}_0$  的作用：

$$\mathbf{r} = \gamma_k \cdot \mathbf{q} \quad (8)$$

此时  $\mathbf{r} = \mathbf{G}_0 \cdot \mathbf{q} = \gamma_k W_{k-m}^\top \dots W_{k-1}^\top \mathbf{v}$ ，对应式 (3) 中  $\mathbf{G}_{k-m}$  (右乘结果)。

### 7.6.5 Step 4: 正向环 (Left Loop) ——处理左乘链与修正项

正向环的目标是将式 (3) 中的“左乘链  $V_{k-m} \dots V_{k-1}$ ”和“秩-1 修正项总和”融入  $\mathbf{r}$ ，最终得到  $\mathbf{G}_k \mathbf{v}$ 。

**5.1 单个  $V_i$  与  $R_i$  的作用** 由  $V_i = \mathbf{I} - \rho_i \mathbf{s}_i \mathbf{y}_i^\top$  和  $R_i = \rho_i \mathbf{s}_i \mathbf{s}_i^\top$ ，结合式 (2) 的递推逻辑，对任意向量  $\mathbf{z}$ ， $V_i \mathbf{z} + R_i \mathbf{v}$  的计算为：

$$V_i \mathbf{z} + R_i \mathbf{v} = \mathbf{z} - \rho_i \mathbf{s}_i (\mathbf{y}_i^\top \mathbf{z}) + \rho_i \mathbf{s}_i (\mathbf{s}_i^\top \mathbf{v}) \quad (9)$$

利用反向环中已存储的  $\alpha_i = \rho_i (\mathbf{s}_i^\top \mathbf{v})$  (式 (6))，定义中间系数  $\beta_i = \rho_i (\mathbf{y}_i^\top \mathbf{z})$ ，则式 (9) 可简化为：

$$\mathbf{z} \leftarrow \mathbf{z} + \mathbf{s}_i (\alpha_i - \beta_i) \quad (10)$$

推导验证:

$$\mathbf{z} - \beta_i \mathbf{s}_i + \alpha_i \mathbf{s}_i = \mathbf{z} + \mathbf{s}_i (\alpha_i - \beta_i)$$

完全匹配式 (9), 且复用了反向环的  $\alpha_i$ , 避免重复计算  $\mathbf{s}_i^\top \mathbf{v}$ 。

**5.2 正向环完整流程**从最早保留的曲率对开始, 自前向后迭代 ( $i = k - m, k - m + 1, \dots, k - 1$ ):

$$\begin{cases} \beta_i = \rho_i \cdot \mathbf{y}_i^\top \mathbf{r} \\ \mathbf{r} = \mathbf{r} + \mathbf{s}_i \cdot (\alpha_i - \beta_i) \end{cases} \quad (11)$$

作用: 通过  $m$  次迭代, 将所有  $V_i$  的左乘作用和  $R_i$  的秩-1 修正融入  $\mathbf{r}$ , 最终  $\mathbf{r}$  即为  $\mathbf{G}_k \mathbf{v}$  的结果:

$$\mathbf{r} = \mathbf{G}_k \mathbf{v} \quad (12)$$

### 7.6.6 Step 5: L-BFGS 搜索方向与迭代流程

当  $\mathbf{v} = \mathbf{g}_k$  (当前梯度) 时, 由式 (12) 得  $\mathbf{r} = \mathbf{G}_k \mathbf{g}_k$ , 因此 L-BFGS 的搜索方向为:

$$\mathbf{p}_k = -\mathbf{r} = -\mathbf{G}_k \mathbf{g}_k \quad (13)$$

(与 BFGS 方向一致, 保证是下降方向, 因  $\mathbf{G}_k \succ 0$  且  $\mathbf{g}_k^\top \mathbf{p}_k = -\mathbf{g}_k^\top \mathbf{G}_k \mathbf{g}_k < 0$ )。

**算法 7.2** (L-BFGS 完整迭代流程). 1. 初始化:

- 初始点  $\mathbf{x}_0$ , 最大记忆数  $m$ , 线搜索参数 (强 Wolfe), 容忍度  $tol$ ;
- 清空存储队列  $\mathcal{S} = []$  (存  $\mathbf{s}_i$ )、 $\mathcal{Y} = []$  (存  $\mathbf{y}_i$ ),  $k = 0$ ;
- 计算  $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$ , 若  $\|\mathbf{g}_0\| \leq tol$ , 停止。

2. 方向计算:

- 若  $k = 0$  (无历史曲率): 取  $\mathbf{p}_0 = -\gamma_0 \mathbf{g}_0$  ( $\gamma_0$  取 1 或经验值);
- 若  $k \geq 1$ : 执行“反向环 (式 6)  $\rightarrow$  初始缩放 (式 8)  $\rightarrow$  正向环 (式 11)”, 得  $\mathbf{p}_k = -\mathbf{r}$ 。

3. 线搜索: 用强 Wolfe 条件求步长  $\alpha_k > 0$ , 满足:

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \alpha_k \mathbf{g}_k^\top \mathbf{p}_k, \quad |\mathbf{g}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)^\top \mathbf{p}_k| \leq c_2 |\mathbf{g}_k^\top \mathbf{p}_k|$$

( $c_1 \in (0, 1), c_2 \in (c_1, 1)$ , 强 Wolfe 保证  $\mathbf{s}_k^\top \mathbf{y}_k > 0$ , 即曲率条件成立)。

4. 更新与存储管理:

- 计算  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$ ,  $\mathbf{s}_k = \alpha_k \mathbf{p}_k$ ,  $\mathbf{g}_{k+1} = \nabla f(\mathbf{x}_{k+1})$ ,  $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ ;
- 若  $\mathbf{s}_k^\top \mathbf{y}_k > 0$  (曲率条件): 将  $\mathbf{s}_k$  加入  $\mathcal{S}$ ,  $\mathbf{y}_k$  加入  $\mathcal{Y}$ ; 若队列长度  $> m$ , 删除最早的  $\mathbf{s}_{k-m}$  和  $\mathbf{y}_{k-m}$ ;
- 计算  $\gamma_{k+1} = \mathbf{s}_k^\top \mathbf{y}_k / (\mathbf{y}_k^\top \mathbf{y}_k)$  (供下次初始缩放)。

5. 终止判断: 若  $\|\mathbf{g}_{k+1}\| \leq tol$ , 停止; 否则  $k = k + 1$ , 返回步骤 2。

### 7.6.7 关键性质验证

1. 割线条件保持: L-BFGS 的两环递推严格继承 BFGS 的割线条件  $\mathbf{G}_k \mathbf{y}_i = \mathbf{s}_i$  ( $i = k - m, \dots, k - 1$ ), 确保曲率拟合的准确性;

2. 正定性保证：若所有保留的  $\mathbf{s}_i^\top \mathbf{y}_i > 0$ ，则  $\mathbf{G}_k \succ 0$ ，搜索方向  $\mathbf{p}_k$  必为下降方向；

3. 复杂度优势：

- 内存复杂度：仅存储  $2m$  个  $n$  维向量（ $\mathcal{S}$  和  $\mathcal{Y}$ ），为  $O(nm)$ （标准 BFGS 为  $O(n^2)$ ）；
- 时间复杂度：每次方向计算需  $2m$  次向量内积和  $2m$  次向量加法，为  $O(nm)$ （标准 BFGS 为  $O(n^2)$ ），适配大规模问题（ $n \gtrsim 10^5$ ）。

L-BFGS 的推导核心是“将 **G-form** 的矩阵递推转化为向量操作”：通过反向环吸收右乘投影、正向环融合左乘投影与秩-1 修正，仅用  $m$  对历史曲率对在线模拟  $\mathbf{G}_k \mathbf{v}$  的计算，既保留了 BFGS 的二阶收敛性，又解决了标准 BFGS 的内存瓶颈。其本质是“用少量历史信息近似逆 Hessian”，是大规模无约束优化（如深度学习、科学计算）的默认基线方法。

## 第八章 优化算法的评价

### 8.1 预备知识与符号

- 目标函数:  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , 可微。
- L-光滑 (Lipschitz 梯度): 存在  $L > 0$  使

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

- 凸/强凸:  $f$  凸; 若存在  $\mu > 0$  使

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

则  $f$  为  $\mu$ -强凸。

- PL (Polyak-Łojasiewicz) 不等式: 存在  $\mu > 0$  使

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*), \quad \forall x.$$

注: PL  $\nRightarrow$  凸, 但常见于过参数模型/深网络的局部区域。

- GD 更新:  $x_{k+1} = x_k - \eta_k \nabla f(x_k)$ , 其中步长  $\eta_k > 0$ 。

### 8.2 基本工具: 下降引理 (Descent Lemma)

引理 8.1 (下降引理). 若函数  $f$  为  $L$ -光滑, 则对任意  $x, d$  与  $\eta > 0$ , 有:

$$f(x + \eta d) \leq f(x) + \eta \nabla f(x)^\top d + \frac{L}{2} \eta^2 \|d\|^2$$

证明.

由  $L$ -光滑性的定义, 对任意  $y$  有:

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

令  $y = x + \eta d$  (即沿方向  $d$  步长  $\eta$  移动), 代入后即得下降引理的不等式。

推论 8.1 (沿负梯度下降). 取  $d = -\nabla f(x)$  (沿负梯度方向更新, 梯度下降算法的核心思想), 代入下降引理得:

$$f(x - \eta \nabla f(x)) \leq f(x) - \left(\eta - \frac{L}{2}\eta^2\right) \|\nabla f(x)\|^2$$

进一步, 当  $0 < \eta \leq \frac{1}{L}$  时, 函数值单调下降, 且有:

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2} \|\nabla f(x_k)\|^2$$

- 该推论明确了梯度下降的下降性保证: 只要步长  $\eta$  满足  $0 < \eta \leq \frac{1}{L}$ , 每一步迭代后目标函数值必严格减小, 确保算法的”下降”特性。
- 为步长选择提供了理论依据 (步长不超过  $\frac{1}{L}$  时算法稳定下降), 也为后续收敛速率分析奠定了基础。

## 8.3 基于 Descent Lemma 的四类典型收敛结果

### 8.3.1 非凸 + L-光滑

由推论 2.2, 当  $0 < \eta \leq \frac{1}{L}$  时, 梯度下降的单步迭代满足:

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2} \|\nabla f(x_k)\|^2$$

将  $k = 0, 1, \dots, K-1$  的不等式依次展开并累加:

$$\begin{aligned} f(x_1) &\leq f(x_0) - \frac{\eta}{2} \|\nabla f(x_0)\|^2 \\ f(x_2) &\leq f(x_1) - \frac{\eta}{2} \|\nabla f(x_1)\|^2 \\ &\vdots \\ f(x_K) &\leq f(x_{K-1}) - \frac{\eta}{2} \|\nabla f(x_{K-1})\|^2 \end{aligned}$$

将这些不等式左右两边分别相加, 中间项  $f(x_1), f(x_2), \dots, f(x_{K-1})$  会相互抵消, 最终得到:

$$f(x_K) \leq f(x_0) - \frac{\eta}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2$$

已知  $f$  下有界 (即  $f_{\inf} = \inf_x f(x) > -\infty$ ), 因此对任意  $K$ , 有  $f(x_K) \geq f_{\inf}$ 。将其代入上式:

$$f_{\inf} \leq f(x_0) - \frac{\eta}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2$$

整理得梯度范数平方和的上界:

$$\sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_{\inf})}{\eta}$$

将上式两边同时除以  $K$ , 得到平均梯度范数平方的上界:

$$\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_{\inf})}{\eta K}$$

进一步分析最小梯度范数的量级: 由于平均值不小于”集合中的最小值”(即  $\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 \geq \min_{0 \leq k < K} \|\nabla f(x_k)\|^2$ ), 因此:

$$\min_{0 \leq k < K} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_{\inf})}{\eta K} = O\left(\frac{1}{K}\right)$$

该定理在非凸深度学习场景中具有关键价值:

1. 即使损失函数非凸，只要满足 L-光滑且下有界，梯度下降的平均梯度范数会随迭代次数增加而趋于 0，说明算法能逐步逼近”驻点”（最优解的必要条件）。
2. 收敛速率为  $O(1/K)$ ，明确了”迭代次数越多，平均梯度越小”的量化规律，为训练过程的收敛性分析提供了理论依据。
3. 解释了”为什么深度网络在梯度下降训练中能逐步收敛”——即使损失函数非凸，L-光滑性和下有界性确保了梯度的整体衰减趋势。

### 8.3.2 凸 + L-光滑

对”凸且 L-光滑函数”的梯度下降（GD）算法进行收敛性分析，核心是推导”函数值次优量”与”解的平方距离差分”的关联不等式（即望远镜技巧的核心步骤）

- 函数设定：  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  是凸且 L-光滑的（即满足  $\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$ ），最优解为  $x^* \in \arg \min f$ 。
- 算法设定：梯度下降取固定步长  $\eta = \frac{1}{L}$ ，迭代规则为  $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ 。
- 基础工具：
  - 凸函数的一阶性质（3.2.a）：  $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ （凸函数的定义式，用于将”梯度内积”转化为”函数值差距”）。
  - 向量范数展开（3.2.c）：  $\|a - c\|^2 = \|b - c\|^2 + 2\langle a - b, b - c \rangle + \|a - b\|^2$ （用于对”解的距离”  $\|x_{k+1} - x^*\|^2$  做代数展开）。

#### 步骤 1：解的距离展开（3.2.d）

令  $a = x_{k+1}$ 、 $b = x_k$ 、 $c = x^*$ ，结合梯度下降的更新式  $x_{k+1} - x_k = -\frac{1}{L}\nabla f(x_k)$ ，代入范数展开式（3.2.c）得：

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \frac{2}{L}\langle \nabla f(x_k), x_k - x^* \rangle + \frac{1}{L^2}\|\nabla f(x_k)\|^2$$

#### 步骤 2：结合凸性，转化梯度内积为函数值差距

由凸函数的性质（3.2.b）：  $f(x) - f^* \leq \langle \nabla f(x), x - x^* \rangle$ ，代入（3.2.d）得：

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{2}{L}(f(x_k) - f^*) + \frac{1}{L^2}\|\nabla f(x_k)\|^2$$

#### 步骤 3：消去梯度范数，关联函数值下降量

利用 L-光滑函数的”下降性质”，推导得：

$$\frac{1}{L^2}\|\nabla f(x_k)\|^2 \leq \frac{2}{L}(f(x_k) - f(x_{k+1}))$$

#### 步骤 4：关键不等式（望远镜差分，3.2.g）

将（3.2.f）代回（3.2.e）并整理，最终得到：

$$\frac{2}{L}(f(x_{k+1}) - f^*) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2$$

意义：这是”望远镜技巧”的核心——它将”函数值到最优的差距”与”解的距离的差分”直接关联。后续对  $k$  累加该不等式，可消去中间项，从而推导出收敛速率（如函数值次优量的  $O(1/K)$  速率）。

**定理 8.1** ( $O(1/k)$  次优率). 若  $f$  凸且  $L$ -光滑, 取  $\eta = \frac{1}{L}$ , 则

$$f(x_k) - f^* \leq \frac{L}{2k} \|x_0 - x^*\|^2, \quad k \geq 1.$$

证明 (望远镜技巧).

对  $k = 0, \dots, T-1$  求和, 右侧望远镜展开:

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

因此

$$\frac{1}{T} \sum_{k=1}^T (f(x_k) - f^*) \leq \frac{L}{2T} \|x_0 - x^*\|^2. \quad (3.2.h)$$

又知  $\{f(x_k)\}$  单调不减, 故

$$f(x_T) - f^* \leq \frac{1}{T} \sum_{k=1}^T (f(x_k) - f^*) \leq \frac{L}{2T} \|x_0 - x^*\|^2.$$

令  $T = k$  即得结论。

一般步长  $\eta \in (0, 1/L]$ :

完全同样的推导 (把上文中的  $1/L$  换为一般  $\eta$ ) 给出

$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\eta k}, \quad \text{取 } \eta = \frac{1}{L} \text{ 恢复定理常数。} \quad (3.2.i)$$

### 8.3.3 $\mu$ -强凸 + $L$ -光滑: 线性 (几何) 收敛

强凸与光滑结合下的梯度下降” 线性收敛”

**定理 8.2.** 若  $f$  为  $\mu$ -强凸且  $L$ -光滑, 取  $0 < \eta \leq \frac{2}{\mu+L}$ , 则

$$\|x_{k+1} - x^*\|^2 \leq \rho^2 \|x_k - x^*\|^2, \quad \rho := \max\{1 - \eta\mu, |1 - \eta L|\} < 1,$$

进而

$$f(x_k) - f^* \leq \frac{L}{2} \rho^{2k} \|x_0 - x^*\|^2.$$

证明: 从”梯度与 Hessian”到”几何收缩”。

我们先引入误差项  $e_k := x_k - x^*$ , 梯度下降的更新式为  $x_{k+1} = x_k - \eta \nabla f(x_k)$ 。

**第一步: Hessian 的”平均积分表示”** 由一维积分形式的平均 Hessian, 梯度的差可表示为:

$$\nabla f(x_k) - \nabla f(x^*) = \left( \int_0^1 \nabla^2 f(x^* + t(x_k - x^*)) dt \right) (x_k - x^*) =: H_k e_k,$$

其中  $H_k$  是对称矩阵, 且满足  $\mu I \preceq H_k \preceq LI$  (强凸和  $L$ -光滑的核心体现: Hessian 的特征值被  $\mu$  和  $L$  夹逼)。

**注 8.1** (Hessian 积分表示的推导). 令  $g(t) = \nabla f(x^* + t(x_k - x^*))$ , 其中  $t \in [0, 1]$ 。

- 当  $t = 0$  时,  $g(0) = \nabla f(x^*)$ ;
- 当  $t = 1$  时,  $g(1) = \nabla f(x_k)$ 。

对  $g(t)$  关于  $t$  求导 (利用 Hessian 的定义:  $\nabla f$  的导数是 Hessian 矩阵  $\nabla^2 f$ ):

$$g'(t) = \nabla^2 f(x^* + t(x_k - x^*)) \cdot (x_k - x^*)$$

根据牛顿-莱布尼茨公式（微积分基本定理），有：

$$g(1) - g(0) = \int_0^1 g'(t) dt$$

将  $g(1) = \nabla f(x_k)$ 、 $g(0) = \nabla f(x^*)$  和  $g'(t)$  的表达式代入，即可得到：

$$\nabla f(x_k) - \nabla f(x^*) = \left( \int_0^1 \nabla^2 f(x^* + t(x_k - x^*)) dt \right) (x_k - x^*)$$

这一步的价值在于将”梯度差”转化为”Hessian 的积分平均形式”，从而可以利用” $\mu$ -强凸（Hessian 下界  $\mu I$ ）”和” $L$ -光滑（Hessian 上界  $LI$ ）”的条件，分析误差项  $x_k - x^*$  的收缩性（即后续的几何收敛）。

简单来说，它是”强凸 +  $L$ -光滑”场景下梯度下降线性收敛证明的”第一块基石”——通过把梯度差和 Hessian 联系起来，我们才能量化误差的几何收缩速率。

**第二步：误差项的迭代收缩** 将更新式代入误差项  $e_{k+1} = x_{k+1} - x^*$ ，得：

$$e_{k+1} = e_k - \eta(\nabla f(x_k) - \nabla f(x^*)) = (I - \eta H_k) e_k.$$

由于  $H_k$  可正交对角化，矩阵  $I - \eta H_k$  的谱范数（即最大特征值的绝对值）由  $H_k$  的特征值范围决定。结合  $\mu \leq \lambda(H_k) \leq L$ ，得：

$$\|e_{k+1}\| \leq \max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \cdot \|e_k\|.$$

记  $\rho := \max\{1 - \eta\mu, |1 - \eta L|\}$ ，当  $0 < \eta \leq \frac{2}{\mu+L}$  时，可验证  $1 - \eta\mu < 1$  且  $|1 - \eta L| < 1$ ，故  $\rho < 1$ 。因此误差项的平方满足几何收缩：

$$\|e_{k+1}\|^2 \leq \rho^2 \|e_k\|^2.$$

**第三步：函数值次优量的收敛** 迭代展开误差项得  $\|e_k\|^2 \leq \rho^{2k} \|e_0\|^2$ 。再结合”上二次界” $f(x) - f^* \leq \frac{L}{2} \|x - x^*\|^2$ ，即可推出函数值次优量的几何收敛：

$$f(x_k) - f^* \leq \frac{L}{2} \rho^{2k} \|x_0 - x^*\|^2.$$

**最优步长的”黄金选择”**

当取  $\eta^* = \frac{2}{\mu+L}$  时，两端点的绝对值相等： $|1 - \eta^*\mu| = |1 - \eta^*L| = \frac{L-\mu}{L+\mu}$ ，此时最优收缩因子  $\rho^* = \frac{L-\mu}{L+\mu} = \frac{\kappa-1}{\kappa+1}$ （其中  $\kappa = \frac{L}{\mu}$  是”条件数”，刻画强凸与光滑的平衡）。

之前凸光滑场景是  $O(1/k)$  的次线性收敛，而这里强凸 + 光滑下是  $\rho^{2k}$  的几何收敛（也叫线性收敛）——前者是”慢慢悠悠逼近”，后者是”指数级收缩”，效率天差地别。这解释了为什么”强凸假设”能让优化算法”跑起来”。

强凸（ $\mu > 0$ ）让函数”长得更陡”， $L$ -光滑（ $L < \infty$ ）让函数”长得不突兀”。两者结合时，Hessian 的特征值被”夹在  $\mu$  和  $L$  之间”，这才使得误差项能通过矩阵谱范数实现”几何收缩”。这种”刚柔并济”的结构，是线性收敛的核心密码。

虽然深度学习中损失函数大多非强凸，但这个结论是正则化、fine-tuning 阶段的理论参照——当模型接近收敛时，局部可能近似满足强凸性，此时梯度下降的收敛会呈现”加速特性”。同时，”最优步长  $\eta^* = \frac{2}{\mu+L}$ ”也为自适应步长设计提供了灵感。

### 8.3.4 PL 条件 + $L$ -光滑：线性收敛（无需凸）

这部分聚焦非凸场景下的线性收敛分析，核心是通过”PL 条件（替代凸性）+  $L$ -光滑（保证梯度平滑）”的组合，证明梯度下降在非凸函数上也能实现线性收敛

**定理 8.3.** 若  $f$  满足  $PL$  条件且  $L$ -光滑, 取步长  $\eta = \frac{1}{L}$ , 则

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right) (f(x_k) - f^*).$$

$PL$  (Polyak-Łojasiewicz) 不等式定义: 存在  $\mu > 0$ , 使得对所有  $x$ ,

$$\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu (f(x) - f^*), \quad \forall x.$$

**注 8.2** (PL 条件的意义). 它把”梯度的大小”和”函数离最优值的差距 (次优量  $f(x) - f^*$ )”直接关联, 是”非凸场景下替代凸性”的关键——无需函数全局凸, 只要局部满足该不等式, 就能量化梯度与优化程度的关系。

由  $L$ -光滑的”沿负梯度下降”推论, 当  $0 < \eta \leq \frac{1}{L}$  时, 梯度下降单步迭代满足:

$$f(x - \eta \nabla f(x)) \leq f(x) - \frac{\eta}{2} \|\nabla f(x)\|^2.$$

将步长  $\eta = \frac{1}{L}$  代入不等式, 得:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

再结合  $PL$  不等式  $\frac{1}{2} \|\nabla f(x_k)\|^2 \geq \mu (f(x_k) - f^*)$ , 两边同乘  $\frac{1}{L}$  得:

$$\frac{1}{2L} \|\nabla f(x_k)\|^2 \geq \frac{\mu}{L} (f(x_k) - f^*).$$

将其代入函数值下降的不等式, 整理后得到:

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right) (f(x_k) - f^*).$$

- **非凸场景的线性收敛突破:** 无需凸性假设, 仅靠” $PL$  (梯度与次优量挂钩) +  $L$ -光滑 (梯度变化平滑)”, 就能让函数次优量以几何级数 (线性速率) 收缩——这解释了”为什么非凸的深度模型训练能收敛到较好效果” (过参数化网络局部常满足  $PL$  条件)。
- **条件的普适性:**  $PL$  条件比凸性弱, 更贴合深度学习中损失函数的非凸特性, 是分析非凸优化收敛性的核心工具之一。

# 第九章 从无约束到等式约束：拉格朗日乘子

## 9.1 拉格朗日乘子数学建模

### 9.1.1 1. 无约束优化的数学模型

无约束优化的基本模型为：

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

其中  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  是可微函数。其一阶最优性条件（极值点的必要条件）为：

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

即目标函数在最优解  $x^*$  处的梯度为零（可微函数的无约束极值点必是梯度为零的点）。

### 9.1.2 2. 等式约束优化的数学模型

当引入等式约束后，模型变为：

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & h_i(x) = 0, \quad i = 1, \dots, m \end{cases}$$

其中  $h_i: \mathbb{R}^n \rightarrow \mathbb{R}$  是可微的约束函数， $m$  是约束个数（通常  $m < n$ ）。

此时，变量  $x$  被限制在由  $m$  个等式约束定义的可行域上，无法直接应用无约束的“梯度为零”条件——因为可行域是约束的“限制空间”，需考虑约束对优化的“限制作用”。

### 9.1.3 3. 拉格朗日函数的建模思想

为了将等式约束“融入”目标函数，我们引入拉格朗日乘子  $\lambda = (\lambda_1, \dots, \lambda_m)^\top \in \mathbb{R}^m$ ，构造拉格朗日函数：

$$\mathcal{L}(x, \lambda) = f(x) + \lambda^\top h(x)$$

其中  $h(x) = (h_1(x), \dots, h_m(x))^\top$  是约束函数的向量形式， $\lambda^\top h(x) = \sum_{i=1}^m \lambda_i h_i(x)$  是“约束项与乘子的耦合项”。

### 9.1.4 4. 拉格朗日函数的最优性条件（一阶 KKT 条件）

对拉格朗日函数  $\mathcal{L}(x, \lambda)$  分别关于  $x$  和  $\lambda$  求偏导，并令其为零，得到一阶最优性条件（等式约束下的 KKT 条件核心）：

- 对  $x$  求梯度并令其为零：

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + \sum_{i=1}^p \lambda_i^* \nabla h_i(x^*) = 0$$

即目标函数的梯度可表示为约束函数梯度的线性组合，组合系数就是拉格朗日乘子  $\lambda_i^*$ 。这一条件体现了“目标函数与约束的梯度平衡”——在最优解处，目标函数的梯度被约束的梯度“抵消”，使得在可行域的切空间内无下降方向。

- 对  $\lambda$  求梯度并令其为零：

$$\nabla_\lambda \mathcal{L}(x^*, \lambda^*) = h(x^*) = 0$$

这就是原问题的等式约束条件，保证最优解满足约束。

### 9.1.5 5. 几何意义（直观理解）

等式约束  $h_i(x) = 0$  定义了一个可行域（流形），其在最优解  $x^*$  处的切空间由“与所有  $\nabla h_i(x^*)$  正交的方向”构成。

目标函数的梯度  $\nabla f(x^*)$  若要使  $x^*$  是极值点，必须“无法在切空间内找到下降方向”，即  $\nabla f(x^*)$  必须位于可行域的法空间中（法空间由  $\nabla h_1(x^*), \dots, \nabla h_m(x^*)$  张成）。因此， $\nabla f(x^*)$  可表示为这些法向量的线性组合，这正是  $\nabla_x \mathcal{L} = 0$  所表达的“梯度线性组合”关系。

综上，拉格朗日函数的建模是通过引入乘子变量  $\lambda$ ，将等式约束转化为无约束优化的梯度条件，从而把“带约束的优化”转化为“对  $(x, \lambda)$  的无约束优化（在一阶条件意义下）”，实现了从无约束到等式约束优化的数学衔接。

## 9.2 拉格朗日乘子法的必要条件

### 9.2.1 一阶必要条件及其证明

**定理 9.1** (一阶必要条件). 设  $x^*$  是等式约束优化问题的局部极小点, 且约束梯度  $\nabla g_1(x^*), \dots, \nabla g_m(x^*)$  线性无关 (即约束规格满足)。则存在唯一的拉格朗日乘子  $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)^T$  使得：

$$\begin{cases} \nabla_x \mathcal{L}(x^*, \lambda^*) = 0 \\ \nabla_\lambda \mathcal{L}(x^*, \lambda^*) = 0 \end{cases}$$

**证明**

证明.

考虑等式约束优化问题：

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & h_i(x) = 0, \quad i = 1, \dots, m \end{cases}$$

其中  $f, h_i$  连续可微,  $x^*$  是局部极小点, 且约束梯度线性无关 (即约束规格满足):  $\nabla h_1(x^*), \dots, \nabla h_m(x^*)$  线性无关。

记约束函数的向量形式为  $h(x) = (h_1(x), \dots, h_m(x))^T$ , 其雅可比矩阵为  $J_h(x) \in \mathbb{R}^{m \times n}$ , 第  $i$  行为  $\nabla h_i(x)^T$ 。由约束梯度线性无关,  $J_h(x^*)$  的秩为  $m$  (列满秩)。

将变量  $x$  分块为  $x = (y, z)$ , 其中  $y \in \mathbb{R}^m, z \in \mathbb{R}^{n-m}$  (通过变量重排, 可假设  $J_h(x^*)$  的前  $m$  列构成的子矩阵  $\nabla_y h(x^*)$  是可逆的  $m \times m$  矩阵)。

根据隐函数定理, 存在  $x^*$  的邻域和可微函数  $g: \mathbb{R}^{n-m} \rightarrow \mathbb{R}^m$ , 使得在该邻域内, 约束  $h(y, z) = 0$  可唯一表示为  $y = g(z)$ , 且  $g(z^*) = y^*$  (即  $x^* = (y^*, z^*)$ )。

原约束问题可转化为关于  $z$  的无约束优化问题:

$$\min_{z \in \mathbb{R}^{n-m}} f(g(z), z)$$

由于  $x^*$  是原问题的局部极小点,  $z^*$  是上述无约束问题的局部极小点。根据无约束优化的一阶必要条件, 对  $z$  的梯度为 0:

$$\nabla_z f(x^*) + \nabla_y f(x^*) \cdot \nabla_z g(z^*) = 0 \quad (1)$$

对  $h(g(z), z) = 0$  关于  $z$  求导, 由链式法则得:

$$\nabla_y h(x^*) \cdot \nabla_z g(z^*) + \nabla_z h(x^*) = 0$$

由于  $\nabla_y h(x^*)$  可逆, 解得:

$$\nabla_z g(z^*) = -(\nabla_y h(x^*))^{-1} \nabla_z h(x^*) \quad (2)$$

将式 (2) 代入式 (1):

$$\nabla_z f(x^*) - \nabla_y f(x^*) (\nabla_y h(x^*))^{-1} \nabla_z h(x^*) = 0$$

定义拉格朗日乘子  $\lambda^* = (\nabla_y h(x^*))^{-T} \nabla_y f(x^*)^T$  (转置是因为矩阵逆的转置等于转置的逆)。

拉格朗日函数为  $\mathcal{L}(x, \lambda) = f(x) + \lambda^T h(x)$ , 其对  $x$  的梯度为:

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla f(x) + J_h(x)^T \lambda$$

将  $\lambda^*$  代入, 分块验证:

- $y$  分量:  $\nabla_y f(x^*) + \nabla_y h(x^*)^T \lambda^*$ 。代入  $\lambda^* = (\nabla_y h(x^*))^{-T} \nabla_y f(x^*)^T$ , 得  $\nabla_y f(x^*) + \nabla_y f(x^*) = 0$  (转置后等式成立)。
- $z$  分量: 结合式 (1)(2) 的推导, 可验证  $\nabla_z f(x^*) + \nabla_z h(x^*)^T \lambda^* = 0$ 。

因此,  $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ ; 同时,  $\nabla_\lambda \mathcal{L}(x^*, \lambda^*) = h(x^*) = 0$  (因  $x^*$  是可行点)。

唯一性:

假设存在两个乘子  $\lambda^*$  和  $\lambda'^*$ , 满足:

$$\nabla f(x^*) + J_h(x^*)^T \lambda^* = 0, \quad \nabla f(x^*) + J_h(x^*)^T \lambda'^* = 0$$

两式相减得  $J_h(x^*)^T (\lambda^* - \lambda'^*) = 0$ 。由于  $J_h(x^*)$  列满秩,  $J_h(x^*)^T$  的零空间仅含零向量, 故  $\lambda^* = \lambda'^*$ , 唯一性得证。

### 9.2.2 二阶最优性条件及其证明

**定理 9.2** (二阶必要条件). 设  $x^*$  是等式约束优化问题的局部极小点,  $f$  和  $g_i$  在  $x^*$  处二阶连续可微, 且约束梯度  $\nabla g_1(x^*), \dots, \nabla g_m(x^*)$  线性无关。则存在  $\lambda^*$  使得一阶条件成立, 且对任意满足  $J_g(x^*)d = 0$  的  $d \in \mathbb{R}^n$ , 有:

$$d^T \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d \geq 0$$

其中  $\nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*)$  是拉格朗日函数关于  $x$  的 Hessian 矩阵。

**定理 9.3** (二阶充分条件). 设  $f$  和  $g_i$  在  $x^*$  处二阶连续可微, 存在  $\lambda^*$  满足一阶条件, 且对任意非零向量  $d$  满足  $J_g(x^*)d = 0$ , 有:

$$d^T \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d > 0$$

则  $x^*$  是严格局部极小点。

二阶必要条件证明.

要证明等式约束优化的二阶必要条件, 我们通过泰勒展开结合一阶必要条件推导, 步骤如下:

考虑等式约束优化问题:

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & h_i(x) = 0, \quad i = 1, \dots, m \end{cases}$$

其中  $x^*$  是局部极小点,  $f, h_i$  二阶连续可微, 且约束梯度  $\nabla h_1(x^*), \dots, \nabla h_m(x^*)$  线性无关 (约束规格满足)。

根据一阶必要条件 (拉格朗日条件), 存在拉格朗日乘子  $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)^\top$ , 使得:

$$\begin{cases} \nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) = 0 \\ h(x^*) = 0 \end{cases}$$

其中拉格朗日函数  $\mathcal{L}(x, \lambda) = f(x) + \lambda^\top h(x)$ 。

考虑切空间中的方向  $d \in \mathbb{R}^n$ , 即满足约束雅可比矩阵  $J_h(x^*)$  零空间的方向:

$$J_h(x^*)d = 0 \implies \nabla h_i(x^*)^\top d = 0, \quad \forall i = 1, \dots, m$$

(该方向  $d$  是“不违反约束的微小移动方向”, 即当  $t$  充分小时,  $x^* + td$  是可行点)。

对  $f(x^* + td)$  做二阶泰勒展开:

$$f(x^* + td) = f(x^*) + t \nabla f(x^*)^\top d + \frac{t^2}{2} d^\top \nabla^2 f(x^*) d + o(t^2)$$

对每个约束  $h_i(x^* + td)$  做一阶泰勒展开 (因  $h_i(x^*) = 0$  且  $\nabla h_i(x^*)^\top d = 0$ , 一阶项为 0):

$$h_i(x^* + td) = 0 + t \nabla h_i(x^*)^\top d + \frac{t^2}{2} d^\top \nabla^2 h_i(x^*) d + o(t^2) = 0 \quad (\text{可行点})$$

由一阶条件  $\nabla f(x^*) = -\sum_{i=1}^m \lambda_i^* \nabla h_i(x^*)$ , 代入  $\nabla f(x^*)^\top d$  得:

$$\nabla f(x^*)^\top d = -\sum_{i=1}^m \lambda_i^* \nabla h_i(x^*)^\top d = 0 \quad (\text{因 } \nabla h_i(x^*)^\top d = 0)$$

因此,  $f(x^* + td) - f(x^*)$  的展开式可简化为:

$$f(x^* + td) - f(x^*) = \frac{t^2}{2} d^\top \left( \nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(x^*) \right) d + o(t^2)$$

注意到拉格朗日函数关于  $x$  的二阶 **Hessian** 矩阵为:

$$\nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) = \nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(x^*)$$

因此, 上式可写为:

$$f(x^* + td) - f(x^*) = \frac{t^2}{2} d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d + o(t^2)$$

由于  $x^*$  是局部极小点, 存在  $\delta > 0$ , 使得对所有  $t \in (0, \delta)$ , 若  $x^* + td$  可行, 则  $f(x^* + td) \geq f(x^*)$ 。

因此, 对充分小的  $t > 0$ , 有:

$$\frac{t^2}{2} d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d + o(t^2) \geq 0$$

两边除以  $\frac{t^2}{2}$  ( $t > 0$ , 故  $\frac{t^2}{2} > 0$ ), 并令  $t \rightarrow 0$ , 得:

$$d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d \geq 0$$

对任意满足  $J_h(x^*)d = 0$  的  $d \in \mathbb{R}^n$ , 有  $d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d \geq 0$ , 即二阶必要条件成立。

二阶充分条件证明.

证明二阶充分条件, 通过泰勒展开和严格局部极小的定义推导, 步骤如下:

考虑等式约束优化问题:

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & h_i(x) = 0, \quad i = 1, \dots, m \end{cases}$$

已知:

1.  $f, h_i$  在  $x^*$  处二阶连续可微;
2. 存在拉格朗日乘子  $\lambda^*$ , 满足一阶拉格朗日条件:  $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$  且  $h(x^*) = 0$  (其中  $\mathcal{L}(x, \lambda) = f(x) + \lambda^\top h(x)$  是拉格朗日函数);
3. 对任意非零向量  $d \in \mathbb{R}^n$  满足  $J_h(x^*)d = 0$  (即  $\nabla h_i(x^*)^\top d = 0, \forall i = 1, \dots, m$ ), 有  $d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*)d > 0$  (拉格朗日函数关于  $x$  的 Hessian 矩阵在切空间方向上正定)。

**证明目标** 需证明  $x^*$  是严格局部极小点, 即存在邻域  $\mathcal{N}(x^*)$ , 使得对所有可行点  $x \in \mathcal{N}(x^*)$  且  $x \neq x^*$ , 有  $f(x) > f(x^*)$ 。

**证明步骤** 步骤 1: 可行点的局部参数化 (切空间方向)

设  $d \in \mathbb{R}^n$  满足  $J_h(x^*)d = 0$  (称此类  $d$  为切空间方向, 沿该方向移动不违反约束)。对充分小的  $t$ , 定义  $x(t) = x^* + td$ 。

对约束  $h_i(x(t))$  做二阶泰勒展开:

$$h_i(x(t)) = h_i(x^*) + t \nabla h_i(x^*)^\top d + \frac{t^2}{2} d^\top \nabla^2 h_i(x^*) d + o(t^2)$$

由  $h_i(x^*) = 0$  且  $\nabla h_i(x^*)^\top d = 0$ , 得:

$$h_i(x(t)) = \frac{t^2}{2} d^\top \nabla^2 h_i(x^*) d + o(t^2)$$

当  $t$  充分小时,  $h_i(x(t)) = 0$  (高阶小量可忽略), 故  $x(t)$  是可行点。

步骤 2: 拉格朗日函数的二阶泰勒展开

拉格朗日函数  $\mathcal{L}(x, \lambda^*) = f(x) + (\lambda^*)^\top h(x)$ , 对  $x(t)$  做二阶泰勒展开:

$$\mathcal{L}(x(t), \lambda^*) = \mathcal{L}(x^*, \lambda^*) + t \nabla_x \mathcal{L}(x^*, \lambda^*)^\top d + \frac{t^2}{2} d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d + o(t^2)$$

由一阶拉格朗日条件  $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ , 上式简化为:

$$\mathcal{L}(x(t), \lambda^*) = \mathcal{L}(x^*, \lambda^*) + \frac{t^2}{2} d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d + o(t^2) \quad (1)$$

步骤 3: 结合可行点的目标函数

因  $x(t)$  是可行点 ( $h(x(t)) = 0$ ), 故  $\mathcal{L}(x(t), \lambda^*) = f(x(t))$ 。同时,  $\mathcal{L}(x^*, \lambda^*) = f(x^*)$  (因  $h(x^*) = 0$ )。

将式 (1) 改写为:

$$f(x(t)) = f(x^*) + \frac{t^2}{2} d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d + o(t^2)$$

步骤 4: 利用二阶正定性推导严格不等式

由已知条件, 对非零  $d$ , 有  $d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d > 0$ 。因此, 存在  $t_0 > 0$ , 当  $0 < |t| < t_0$  时:

$$\frac{t^2}{2} d^\top \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) d + o(t^2) > 0$$

即  $f(x(t)) > f(x^*)$ 。

步骤 5: 验证严格局部极小

由隐函数定理, 可行域在  $x^*$  附近的局部结构可由所有切空间方向  $d$  生成的  $x(t)$  覆盖。因此, 存在  $x^*$  的邻域  $\mathcal{N}(x^*)$ , 使得对所有可行点  $x \in \mathcal{N}(x^*)$  且  $x \neq x^*$ , 必有  $f(x) > f(x^*)$ , 即  $x^*$  是严格局部极小点。

# 第十章 从等式约束到不等式约束：KKT

## 10.1 为什么需要 KKT

等式约束优化（拉格朗日乘子法）仅能处理光滑边界、无需区分约束是否“起作用”的场景；而不等式约束的可行集有“内部/边界（活跃约束）”，可行方向呈锥形，等式方法无法区分活跃约束、也没法处理约束梯度需非负权重的需求。KKT 正是补上这一缺口，将“目标在可行方向不下降”的几何要求，转化为含“活跃约束区分、非负乘子、互补松弛”的代数条件，从而能解不等式约束的优化问题。

## 10.2 不等式约束建模

### 10.2.1 等式约束优化问题（基础模型）

设优化目标为最小化目标函数  $f(\mathbf{x})$ ，仅受等式约束限制，数学建模如下：

$$\begin{cases} \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \\ \text{subject to } g_i(\mathbf{x}) = 0 \quad (i = 1, 2, \dots, m) \end{cases}$$

其中：

- $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  是  $n$  维决策变量（ $\mathbb{R}^n$  表示  $n$  维实数空间）；
- $f: \mathbb{R}^n \rightarrow \mathbb{R}$  是连续可微的目标函数（映射到实数域）；
- $g_i: \mathbb{R}^n \rightarrow \mathbb{R}$  ( $i = 1, \dots, m$ ) 是连续可微的等式约束函数， $m < n$ （约束数量少于变量维度，保证可行集非空）。

### 10.2.2 含不等式约束的优化问题（扩展模型）

在等式约束基础上引入不等式约束，形成更通用的优化模型：

$$\begin{cases} \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \\ \text{subject to } \begin{aligned} g_i(\mathbf{x}) &= 0 \quad (i = 1, 2, \dots, m) \\ h_j(\mathbf{x}) &\leq 0 \quad (j = 1, 2, \dots, p) \end{aligned} \end{cases}$$

其中：

- 新增  $p$  个连续可微的不等式约束函数  $h_j: \mathbb{R}^n \rightarrow \mathbb{R}$  ( $j = 1, \dots, p$ )，约束满足  $h_j(\mathbf{x}) \leq 0$ ；
- 可行集定义为  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0, h_j(\mathbf{x}) \leq 0, \forall i, j\}$ ，最优解需在  $\mathcal{X}$  内使  $f(\mathbf{x})$  最小；

- 需通过积极约束集  $\mathcal{A}(\mathbf{x}^*) = \{j \in \{1, \dots, p\} \mid h_j(\mathbf{x}^*) = 0\}$  区分”起作用”（边界）与”不起作用”（内部）的不等式约束，为后续 KKT 条件奠基。

## 10.3 KKT 条件的推导

### 10.3.1 一阶必要条件（KKT 条件）的严格数学建模

前提假设

设约束优化问题为：

$$\mathcal{P} : \begin{cases} \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \\ \text{s.t. } g_i(\mathbf{x}) = 0 \quad (i = 1, 2, \dots, m) \\ h_j(\mathbf{x}) \leq 0 \quad (j = 1, 2, \dots, p) \end{cases}$$

其中：

- $f, g_i, h_j \in C^1(\mathbb{R}^n)$ （均为一阶连续可微函数）；
- $\mathbf{x}^* \in \mathcal{X}$ （ $\mathcal{X}$  为可行集，即  $\mathcal{X} = \{\mathbf{x} \mid g_i(\mathbf{x}) = 0, h_j(\mathbf{x}) \leq 0\}$ ），且  $\mathbf{x}^*$  是  $\mathcal{P}$  的局部极小点；
- 满足约束规格（Constraint Qualification, CQ）（后续定义）。

数学结论

存在拉格朗日乘子  $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_m^*)^T \in \mathbb{R}^m$  和  $\mu^* = (\mu_1^*, \mu_2^*, \dots, \mu_p^*)^T \in \mathbb{R}^p$ ，使得以下 4 个条件同时成立：

1. 平稳性条件（目标梯度与约束梯度平衡）：

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0} \in \mathbb{R}^n$$

2. 原始可行性条件（解满足所有约束）：

$$h_j(\mathbf{x}^*) \leq 0 \quad \forall j \in \{1, 2, \dots, p\}$$

（注：等式约束  $g_i(\mathbf{x}^*) = 0$  已隐含在  $\mathbf{x}^* \in \mathcal{X}$  中，此处补充不等式约束的显式条件）

3. 对偶可行性条件（不等式约束乘子非负）：

$$\mu_j^* \geq 0 \quad \forall j \in \{1, 2, \dots, p\}$$

4. 互补松弛条件（非活跃约束乘子为 0）：

$$\mu_j^* h_j(\mathbf{x}^*) = 0 \quad \forall j \in \{1, 2, \dots, p\}$$

### 10.3.2 约束规格（CQ）的严格数学建模

约束规格是保证”局部极小点满足 KKT 条件”的关键假设，以下为两类核心 CQ 的严格定义：

### 线性无关约束规格 (Linear Independence Constraint Qualification, LICQ)

设  $\mathbf{x}^* \in \mathcal{X}$ , 记积极约束集  $\mathcal{A}(\mathbf{x}^*) = \{j \in \{1, \dots, p\} \mid h_j(\mathbf{x}^*) = 0\}$  (即不等式约束中“起作用”的集合)。

若向量集合:

$$\mathcal{G}(\mathbf{x}^*) = \{\nabla g_i(\mathbf{x}^*) \mid i = 1, \dots, m\} \cup \{\nabla h_j(\mathbf{x}^*) \mid j \in \mathcal{A}(\mathbf{x}^*)\}$$

满足线性无关 (即不存在不全为零的常数  $\alpha_1, \dots, \alpha_m, \beta_j (j \in \mathcal{A}(\mathbf{x}^*))$ , 使得  $\sum_{i=1}^m \alpha_i \nabla g_i(\mathbf{x}^*) + \sum_{j \in \mathcal{A}(\mathbf{x}^*)} \beta_j \nabla h_j(\mathbf{x}^*) = \mathbf{0}$ ), 则称在  $\mathbf{x}^*$  处满足 LICQ。

### Mangasarian-Fromovitz 约束规格 (Mangasarian-Fromovitz Constraint Qualification, MFCQ)

设  $\mathbf{x}^* \in \mathcal{X}$ , 若满足以下两个条件:

1. 等式约束梯度集  $\{\nabla g_i(\mathbf{x}^*) \mid i = 1, \dots, m\}$  线性无关;
2. 存在可行方向  $\mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$  (非零方向), 使得:

$$\nabla g_i(\mathbf{x}^*)^T \mathbf{d} = 0 \quad \forall i \in \{1, \dots, m\}$$

$$\nabla h_j(\mathbf{x}^*)^T \mathbf{d} < 0 \quad \forall j \in \mathcal{A}(\mathbf{x}^*)$$

则称在  $\mathbf{x}^*$  处满足 MFCQ。

注 10.1. MFCQ 是弱于 LICQ 的约束规格, 即 LICQ 成立可推出 MFCQ 成立, 但反之不成立。

### 10.3.3 LICQ 下 KKT 条件的严格证明建模

#### 第一步: 定义线性化可行方向锥

设  $\mathbf{x}^*$  是  $\mathcal{P}$  的局部极小点, 且在  $\mathbf{x}^*$  处满足 LICQ。定义线性化可行方向锥:

$$\mathcal{F}(\mathbf{x}^*) = \{\mathbf{d} \in \mathbb{R}^n \mid \nabla g_i(\mathbf{x}^*)^T \mathbf{d} = 0 \ (\forall i = 1, \dots, m), \nabla h_j(\mathbf{x}^*)^T \mathbf{d} \leq 0 \ (\forall j \in \mathcal{A}(\mathbf{x}^*))\}$$

几何意义:  $\mathcal{F}(\mathbf{x}^*)$  是在  $\mathbf{x}^*$  处“沿该方向移动, 线性近似下仍可行”的所有方向集合。

#### 第二步: 证明核心引理

引理 10.1. 对所有  $\mathbf{d} \in \mathcal{F}(\mathbf{x}^*)$ , 有  $\nabla f(\mathbf{x}^*)^T \mathbf{d} \geq 0$ 。

证明 (反证法)。

假设存在  $\mathbf{d}_0 \in \mathcal{F}(\mathbf{x}^*)$  使得  $\nabla f(\mathbf{x}^*)^T \mathbf{d}_0 < 0$ 。

由 LICQ 成立, 根据隐函数定理, 可构造可行曲线:

$$\mathbf{x}(t) = \mathbf{x}^* + t\mathbf{d}_0 + o(t) \quad (t \geq 0)$$

其中  $o(t)$  是高阶无穷小 (满足  $\lim_{t \rightarrow 0+} \frac{\|o(t)\|}{t} = 0$ ), 且曲线满足:

1. 对等式约束:  $g_i(\mathbf{x}(t)) = 0 + t\nabla g_i(\mathbf{x}^*)^T \mathbf{d}_0 + o(t) = o(t)$  (因  $\mathbf{d}_0 \in \mathcal{F}(\mathbf{x}^*)$ , 故  $\nabla g_i(\mathbf{x}^*)^T \mathbf{d}_0 = 0$ );

2. 对积极不等式约束 ( $j \in \mathcal{A}(\mathbf{x}^*)$ ):  $h_j(\mathbf{x}(t)) = 0 + t\nabla h_j(\mathbf{x}^*)^T \mathbf{d}_0 + o(t) \leq 0 + o(t)$  (因  $\mathbf{d}_0 \in \mathcal{F}(\mathbf{x}^*)$ , 故  $\nabla h_j(\mathbf{x}^*)^T \mathbf{d}_0 \leq 0$ )。

取充分小的  $t_0 > 0$ , 当  $t \in (0, t_0)$  时:

- $|o(t)| < t \cdot \min \{1, |\nabla h_j(\mathbf{x}^*)^T \mathbf{d}_0|\} (\forall j \in \mathcal{A}(\mathbf{x}^*))$ , 故  $g_i(\mathbf{x}(t)) \approx 0$ 、 $h_j(\mathbf{x}(t)) \leq 0$ , 即  $\mathbf{x}(t) \in \mathcal{X}$  (可行);
  - 目标函数泰勒展开:  $f(\mathbf{x}(t)) = f(\mathbf{x}^*) + t\nabla f(\mathbf{x}^*)^T \mathbf{d}_0 + o(t) < f(\mathbf{x}^*)$  (因  $\nabla f(\mathbf{x}^*)^T \mathbf{d}_0 < 0$ , 且  $t$  充分小)。
- 这与  $\mathbf{x}^*$  是局部极小点矛盾, 故假设不成立, 引理得证。

### 第三步: 应用 Farkas 引理推导乘子存在性

**引理 10.2** (Farkas 引理). 设  $\mathbf{A} \in \mathbb{R}^{k \times n}$ ,  $\mathbf{b} \in \mathbb{R}^n$ , 则以下两个系统有且仅有一个有解:

- 系统 1:  $\mathbf{A}\mathbf{d} \leq \mathbf{0}, \mathbf{b}^T \mathbf{d} > 0$  ( $\mathbf{d} \in \mathbb{R}^n$ );
- 系统 2:  $\mathbf{A}^T \mathbf{y} = \mathbf{b}, \mathbf{y} \geq \mathbf{0}$  ( $\mathbf{y} \in \mathbb{R}^k$ )。

将第二步引理转化为 Farkas 引理的”系统 1 无解”场景:

构造矩阵  $\mathbf{A}$  和向量  $\mathbf{b}$  如下:

- $\mathbf{A}$  的行由  $\nabla g_i(\mathbf{x}^*)^T$ 、 $-\nabla g_i(\mathbf{x}^*)^T$  (对应  $\nabla g_i(\mathbf{x}^*)^T \mathbf{d} = 0$  拆分为  $\leq 0$  和  $\geq 0$ )、 $\nabla h_j(\mathbf{x}^*)^T$  ( $j \in \mathcal{A}(\mathbf{x}^*)$ ) 组成;
- $\mathbf{b} = -\nabla f(\mathbf{x}^*)$ 。

由第二步引理, ”系统 1:  $\mathbf{A}\mathbf{d} \leq \mathbf{0}, \mathbf{b}^T \mathbf{d} > 0$ ” 无解, 故由 Farkas 引理, ”系统 2:  $\mathbf{A}^T \mathbf{y} = \mathbf{b}, \mathbf{y} \geq \mathbf{0}$ ” 有解。

整理系统 2 的解, 可得:

- 存在  $\lambda_i^* \in \mathbb{R}$  (对应等式约束的乘子, 由  $\nabla g_i(\mathbf{x}^*)^T$  和  $-\nabla g_i(\mathbf{x}^*)^T$  的系数合成);
- 存在  $\mu_j^* \geq 0$  (对应积极不等式约束的乘子, 由  $\nabla h_j(\mathbf{x}^*)^T$  的系数给出);
- 对非积极约束 ( $j \notin \mathcal{A}(\mathbf{x}^*)$ ), 令  $\mu_j^* = 0$ , 则平稳性条件成立。

同时, 互补松弛条件  $\mu_j^* h_j(\mathbf{x}^*) = 0$  自然满足 (非积极约束  $\mu_j^* = 0$ , 积极约束  $h_j(\mathbf{x}^*) = 0$ ), 对偶可行性条件  $\mu_j^* \geq 0$  由 Farkas 引理的  $\mathbf{y} \geq \mathbf{0}$  保证。

综上, KKT 条件在 LICQ 下得证。

### 10.3.4 Farkas 引理证明

**定理 10.1** (Farkas 引理). 设  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{b} \in \mathbb{R}^n$ , 则以下两个系统有且仅有一个有解:

- 系统 1:  $\mathbf{A}\mathbf{x} \leq \mathbf{0}, \mathbf{b}^T \mathbf{x} > 0$  ( $\mathbf{x} \in \mathbb{R}^n$ );
- 系统 2:  $\mathbf{A}^T \mathbf{y} = \mathbf{b}, \mathbf{y} \geq \mathbf{0}$  ( $\mathbf{y} \in \mathbb{R}^m$ )。

证明 (分两步核心逻辑)。

**第一步：证明”两系统不能同时有解”（矛盾法）** 假设系统 1、系统 2 同时存在解，即存在  $x \in \mathbb{R}^n$  满足  $Ax \leq 0$  且  $b^T x > 0$ ，同时存在  $y \in \mathbb{R}^m$  满足  $A^T y = b$  且  $y \geq 0$ 。

对  $b^T x$  做代数变形：

由  $A^T y = b$ ，两边转置得  $b^T = y^T A$ ，因此  $b^T x = y^T (Ax)$ 。

结合已知条件分析：

- 因  $Ax \leq 0$ （系统 1）且  $y \geq 0$ （系统 2），向量内积  $y^T (Ax) \leq 0$ ，即  $b^T x \leq 0$ ；
- 但系统 1 要求  $b^T x > 0$ ，二者矛盾。故两系统不能同时有解。

**第二步：证明”若系统 1 无解，则系统 2 必有解”（凸集分离定理 + 矛盾法）**

1. 定义闭凸锥：设集合  $C = \{A^T y \mid y \geq 0\}$ ，易证  $C$  是  $\mathbb{R}^n$  中的闭凸锥（闭性由线性映射连续性 + 非负锥闭性保证，凸性由线性映射凸性 + 非负锥凸性保证）。

2. 反证假设与凸集分离：

假设系统 1 无解，且系统 2 也无解（即  $b \notin C$ ，若  $b \in C$  则存在  $y \geq 0$  使  $A^T y = b$ ，系统 2 有解）。

因  $C$  是闭凸集且  $b \notin C$ ，由凸集分离定理，存在非零向量  $x \in \mathbb{R}^n$  和实数  $\alpha \in \mathbb{R}$ ，使得：

$$b^T x > \alpha \quad \text{且} \quad (A^T y)^T x \leq \alpha \quad \forall y \geq 0$$

3. 推导系统 1 有解（矛盾）：

- 令  $y = 0$ （因  $0 \in \{y \mid y \geq 0\}$ ，故  $A^T 0 = 0 \in C$ ），代入右边不等式得  $0^T x \leq \alpha$ ，即  $\alpha \geq 0$ ，因此  $b^T x > \alpha \geq 0$ ，即  $b^T x > 0$ 。
- 若存在某个分量  $(Ax)_k > 0$ ，取  $y = te_k$ （ $e_k$  为第  $k$  个单位向量， $t > 0$ ），则  $(A^T y)^T x = t(Ax)_k$ 。当  $t \rightarrow +\infty$  时， $t(Ax)_k \rightarrow +\infty$ ，与  $(A^T y)^T x \leq \alpha$ （ $\alpha$  是固定实数）矛盾。故必须  $Ax \leq 0$ 。

此时  $x$  满足  $Ax \leq 0$  且  $b^T x > 0$ ，即系统 1 有解——与”系统 1 无解”的初始假设矛盾。故系统 2 必有解。

## 10.4 二阶最优性条件

在学习优化问题时，一阶 KKT 条件告诉我们”最优解处的梯度要平衡约束”（相当于”坡度为零”），但这只够判断”可能是极值点”——就像走到平地上，分不清是山顶、山谷还是半山腰的平台。而二阶最优性条件是在一阶条件基础上，通过判断”曲率”（相当于地面的弯曲方向），明确这个”平地”到底是不是真正的极小点。

要判断”曲率”，首先得明确：在 KKT 点  $x^*$  处，还有可能让目标函数下降的方向有哪些？这个方向集合就是”临界锥”。

### 10.4.1 临界锥

**定义 10.1** (临界锥). 在 KKT 点  $x^*$  处，临界锥定义为：

$$\mathcal{C}(x^*, \mu^*) = \left\{ d \in \mathbb{R}^n : \begin{array}{l} \nabla g_i(x^*)^T d = 0, \quad i = 1, \dots, m \\ \nabla h_j(x^*)^T d = 0, \quad j \in \mathcal{A}(x^*) \text{ 且 } \mu_j^* > 0 \\ \nabla h_j(x^*)^T d \leq 0, \quad j \in \mathcal{A}(x^*) \text{ 且 } \mu_j^* = 0 \end{array} \right\}$$

临界锥包含了在  $x^*$  处所有可能的”候选下降方向”，即”在不违反任何约束的前提下，可能让目标函数下降的所有方向”。

## 10.4.2 二阶必要条件

**定理 10.2** (二阶必要条件). 设  $x^*$  是优化问题的局部极小点, 且满足线性无关约束规格 (LICQ),  $(\lambda^*, \mu^*)$  是对应的 KKT 乘子, 则对所有方向  $d \in \mathcal{C}(x^*, \mu^*)$ , 有:

$$d^T \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*) d \geq 0$$

其中:

- 拉格朗日函数:  $\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x)$ , 其中  $\lambda \in \mathbb{R}^m$ 、 $\mu \in \mathbb{R}^p$  为拉格朗日乘子;
- 拉格朗日 Hessian 矩阵:  $\nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*)$  是拉格朗日函数在  $(x^*, \lambda^*, \mu^*)$  处关于  $x$  的二阶偏导数矩阵 (Hessian 矩阵), 严格定义为:

$$\nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*) = \nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 g_i(x^*) + \sum_{j=1}^p \mu_j^* \nabla^2 h_j(x^*)$$

其中  $\nabla^2 f(x^*)$ 、 $\nabla^2 g_i(x^*)$ 、 $\nabla^2 h_j(x^*)$  分别表示  $f$ 、 $g_i$ 、 $h_j$  在  $x^*$  处的 Hessian 矩阵 ( $n \times n$  对称矩阵)。

**注 10.2.** ”必要条件”的意思是: 如果  $x^*$  是局部极小点, 那么在所有”可能下降的方向” (临界锥内), 综合曲率必须  $\geq 0$ 。

## 10.4.3 二阶充分条件

**定理 10.3** (二阶充分条件). 设  $x^*$  是优化问题的可行点 (满足  $g_i(x^*) = 0$ 、 $h_j(x^*) \leq 0$ ), 存在乘子  $(\lambda^*, \mu^*)$  使得  $(x^*, \lambda^*, \mu^*)$  满足 KKT 条件, 且对所有非零方向  $d \in \mathcal{C}(x^*, \mu^*)$ , 有:

$$d^T \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*, \mu^*) d > 0$$

则  $x^*$  是优化问题的严格局部极小点。

**注 10.3.** ”充分条件”的意思是: 只要在所有”可能下降的方向” (临界锥内的非零方向), 综合曲率都  $> 0$ , 那么  $x^*$  一定是局部极小点。

# 第十一章 从等式约束到不等式约束：对偶

## 11.1 从 KKT 条件到对偶问题的严格数学建模

### 11.1.1 原始问题（Primal Problem）建模

定义（原始约束优化问题）

给定变量空间  $\mathbb{R}^n$ ，目标函数与约束函数满足：

- 目标函数：  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ （从  $n$  维欧氏空间到实数域的映射）
- 等式约束函数：  $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$  ( $i = 1, 2, \dots, m$ )，共  $m$  个等式约束
- 不等式约束函数：  $h_j : \mathbb{R}^n \rightarrow \mathbb{R}$  ( $j = 1, 2, \dots, p$ )，共  $p$  个不等式约束

原始问题的数学表达式为：

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{subject to} & g_i(x) = 0, \quad i = 1, 2, \dots, m \\ & h_j(x) \leq 0, \quad j = 1, 2, \dots, p \end{cases}$$

其中,  $x = (x_1, x_2, \dots, x_n)^T$  为原始问题的优化变量, 原始问题的最优值记为  $p^* = \inf\{f(x) \mid x \text{ 满足所有约束}\}$ 。

### 11.1.2 拉格朗日函数建模

定义（拉格朗日函数）

引入拉格朗日乘子：

- 等式约束乘子：  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)^T \in \mathbb{R}^m$ （无符号限制）
- 不等式约束乘子：  $\mu = (\mu_1, \mu_2, \dots, \mu_p)^T \in \mathbb{R}^p$ （后续将限定非负）

拉格朗日函数定义为：

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x)$$

其中,  $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ （从  $n + m + p$  维乘积空间到实数域的映射）。

### 11.1.3 拉格朗日对偶函数建模

**定义 11.1** (拉格朗日对偶函数). 拉格朗日对偶函数是拉格朗日函数关于原始变量  $x$  的逐点下确界, 定义为:

$$d(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu)$$

其中,  $d: \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \cup \{-\infty\}$  (对偶函数的值域可包含负无穷, 当下确界不存在时取  $-\infty$ ).

#### 11.1.4 对偶函数的核心性质建模

**定理 11.1** (对偶函数的凹性). 对偶函数  $d(\lambda, \mu)$  是关于  $(\lambda, \mu)$  的凹函数。

证明.

1. 对任意固定的  $x \in \mathbb{R}^n$ , 构造函数  $\phi_x(\lambda, \mu) = \mathcal{L}(x, \lambda, \mu)$ 。由拉格朗日函数的定义可知:

$$\phi_x(\lambda, \mu) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x)$$

其中  $f(x), g_i(x), h_j(x)$  均与  $(\lambda, \mu)$  无关, 因此  $\phi_x(\lambda, \mu)$  是关于  $(\lambda, \mu)$  的仿射函数 (线性函数加常数项)。

2. 对偶函数  $d(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \phi_x(\lambda, \mu)$ , 即对偶函数是一族仿射函数  $\{\phi_x(\lambda, \mu) \mid x \in \mathbb{R}^n\}$  的逐点下确界。

3. 由凸分析基本性质: 一族仿射函数的逐点下确界是凹函数, 因此  $d(\lambda, \mu)$  是凹函数。

**定理 11.2** (对偶函数的下界性质). 对任意  $\lambda \in \mathbb{R}^m$  和  $\mu \geq 0$  (即  $\mu_j \geq 0$  对所有  $j = 1, 2, \dots, p$  成立), 有:

$$d(\lambda, \mu) \leq p^*$$

其中  $p^*$  是原始问题的最优值。

证明.

1. 设  $x$  是原始问题的任意可行点, 即满足:

$$g_i(x) = 0 \quad (i = 1, \dots, m), \quad h_j(x) \leq 0 \quad (j = 1, \dots, p)$$

2. 由于  $\mu \geq 0$  且  $h_j(x) \leq 0$ , 可得:

$$\sum_{j=1}^p \mu_j h_j(x) \leq 0$$

又因为  $g_i(x) = 0$ , 故:

$$\sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x) = 0 + \sum_{j=1}^p \mu_j h_j(x) \leq 0$$

3. 代入拉格朗日函数得:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x) \leq f(x)$$

4. 由对偶函数的定义 (下确界性质), 对任意  $x \in \mathbb{R}^n$  有:

$$d(\lambda, \mu) = \inf_{y \in \mathbb{R}^n} \mathcal{L}(y, \lambda, \mu) \leq \mathcal{L}(x, \lambda, \mu)$$

5. 结合步骤 3 和步骤 4, 得:

$$d(\lambda, \mu) \leq \mathcal{L}(x, \lambda, \mu) \leq f(x)$$

6. 由于该不等式对所有原始可行点  $x$  成立, 而  $p^*$  是所有可行点对应的  $f(x)$  的下确界, 因此:

$$d(\lambda, \mu) \leq p^*$$

### 11.1.5 拉格朗日对偶问题建模

#### 定义（对偶问题）

基于对偶函数的下界性质，对偶问题的目标是最大化对偶函数的下界，同时满足乘子约束  $\mu \geq 0$ 。其数学表达式为：

$$\begin{cases} \max_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p} d(\lambda, \mu) \\ \text{subject to } \mu_j \geq 0, \quad j = 1, 2, \dots, p \end{cases}$$

#### 关键定义补充

1. 对偶问题的最优值：  $d^* = \sup\{d(\lambda, \mu) \mid \mu \geq 0\}$ （上确界，因对偶函数是凹函数，最大值若存在则上确界等于最大值）。
2. 对偶间隙：原始问题最优值与对偶问题最优值的差值，即  $\Delta = p^* - d^*$ 。
3. 弱对偶性：由定理 4.2 直接可得  $d^* \leq p^*$ ，即对偶间隙非负（ $\Delta \geq 0$ ），该性质对所有原始-对偶问题对恒成立。

## 11.2 强对偶

当对偶间隙消失（即原始问题与对偶问题的最优值相等）时，称**强对偶成立**，这是对偶理论中“原始-对偶等价”的关键条件。

**定义 11.2** (强对偶). 若对偶间隙  $\Delta = 0$ ，即：

$$p^* = d^*$$

则称原始问题与对偶问题满足强对偶性。

需特别说明：强对偶并非对所有约束优化问题恒成立，仅在满足特定条件（如 Slater 条件）时可被保证。

### 11.2.1 Slater 条件

**定理 11.3** (Slater 条件). 若原始问题是凸问题，且存在严格可行点  $x_{strict} \in \mathbb{R}^n$ ，满足：

$$\begin{cases} g_i(x_{strict}) = 0, & i = 1, \dots, m \quad (\text{等式约束仍严格满足}) \\ h_j(x_{strict}) < 0, & j = 1, \dots, p \quad (\text{不等式约束严格满足, 无“紧约束”}) \end{cases}$$

则原始问题与对偶问题满足强对偶性（ $p^* = d^*$ ）。

**注 11.1.** Slater 条件是强对偶成立的充分条件，而非必要条件——即满足 Slater 条件一定有强对偶，但强对偶成立时未必满足 Slater 条件。

### 11.2.2 几何解释

构造三维乘积空间  $\mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}$  中的集合  $\mathcal{G}$ , 其元素为满足”约束与目标函数不等式”的三元组  $(u, v, t)$ , 数学定义为:

$$\mathcal{G} = \{(u, v, t) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \mid \exists x \in \mathbb{R}^n, g_i(x) = u_i \ (i = 1, \dots, m), h_j(x) \leq v_j \ (j = 1, \dots, p), f(x) \leq t\}$$

其中:

- $u = (u_1, \dots, u_m)^T$  (等式约束  $g_i(x)$  的取值);
- $v = (v_1, \dots, v_p)^T$  (不等式约束  $h_j(x)$  的上界);
- $t$  (目标函数  $f(x)$  的上界)。

原始问题的最优值  $p^*$  是”使  $(0, 0, t) \in \mathcal{G}$  的最小  $t$ ”——即当等式约束取  $u = 0$ 、不等式约束取  $v = 0$  (满足原始约束) 时, 目标函数上界  $t$  的下确界, 数学表达式为:

$$p^* = \inf \{t \mid (0, 0, t) \in \mathcal{G}\}$$

几何意义:  $\mathcal{G}$  中所有”第一分量为 0、第二分量为 0”的点, 其第三分量的最小值即为  $p^*$ 。

对偶问题的最优值  $d^*$  是”对  $\mu \geq 0$ , 在  $\mathcal{G}$  上最小化  $t + \lambda^T u + \mu^T v$  的最大值”, 数学表达式为:

$$d^* = \sup_{\mu \geq 0} \inf_{(u, v, t) \in \mathcal{G}} \{t + \lambda^T u + \mu^T v\}$$

其中, 内层下确界对应偶函数  $d(\lambda, \mu) = \inf_{(u, v, t) \in \mathcal{G}} \{t + \lambda^T u + \mu^T v\}$ , 外层上界对应偶问题的最大化目标。

强对偶成立 ( $p^* = d^*$ ) 的几何意义是: 原始最优值的”下确界”与对偶最优值的”上确界-下确界”相等, 即  $\mathcal{G}$  的极值特性满足”对偶无间隙”。

### 11.2.3 强对偶定理证明

**定理 11.4** (强对偶定理). 若原始问题是凸问题 ( $f, h_j$  凸,  $g_i$  仿射), 且满足 Slater 条件 (存在严格可行点  $x_{strict}$ ), 则强对偶成立, 即  $p^* = d^*$ 。

证明.

定义集合  $A = \mathcal{G}$  (即前述几何建模中的凸集, 因原始问题是凸问题,  $A$  是凸集); 定义集合  $B = \{(0, 0, s) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \mid s < p^*\}$  (所有”前两分量为 0、第三分量小于  $p^*$ ”的点,  $B$  是凸集)。

**关键引理:**  $A$  与  $B$  不相交。

若存在  $(0, 0, s) \in A \cap B$ , 则  $s < p^*$  且  $(0, 0, s) \in \mathcal{G}$ ——由  $\mathcal{G}$  的定义, 存在  $x$  满足  $g_i(x) = 0, h_j(x) \leq 0, f(x) \leq s < p^*$ , 这与  $p^*$  是原始问题最优值 (最小  $f(x)$ ) 矛盾, 故  $A \cap B = \emptyset$ 。

由于  $A, B$  是  $\mathbb{R}^{m+p+1}$  中的不相交凸集, 根据凸集分离定理, 存在非零向量  $(\tilde{\lambda}, \tilde{\mu}, \nu) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}$  和实数  $\alpha \in \mathbb{R}$ , 使得对所有  $(u, v, t) \in A, (0, 0, s) \in B$ , 有:

$$\begin{cases} \tilde{\lambda}^T u + \tilde{\mu}^T v + \nu t \geq \alpha & (\text{超平面上方包含 } A) \\ \tilde{\lambda}^T \cdot 0 + \tilde{\mu}^T \cdot 0 + \nu s \leq \alpha & (\text{超平面下方包含 } B) \end{cases}$$

**第一步: 证明  $\nu \geq 0$**  假设  $\nu < 0$ , 对任意  $(u, v, t) \in A$ , 当  $t \rightarrow +\infty$  时,  $\tilde{\lambda}^T u + \tilde{\mu}^T v + \nu t \rightarrow -\infty$ , 与” $\geq \alpha$ ”矛盾, 故  $\nu \geq 0$ 。

**第二步: 证明  $\nu \neq 0$**  假设  $\nu = 0$ , 则分离不等式变为  $\tilde{\lambda}^T u + \tilde{\mu}^T v \geq \alpha$  (对所有  $(u, v, t) \in A$ )。由 Slater 条件, 存在严格可行点  $x_{strict}$ , 使得  $g_i(x_{strict}) = 0$ ,  $h_j(x_{strict}) = v_j < 0$ , 即  $(0, v, t) \in A$ 。代入得  $\tilde{\mu}^T v \geq \alpha$ 。若令  $v_j \rightarrow -\infty$  (通过调整  $x$ ), 则  $\tilde{\mu}^T v \rightarrow -\infty$ , 与 “ $\geq \alpha$ ” 矛盾, 故  $\nu \neq 0$ 。

由于  $\nu > 0$ , 可对分离向量  $(\tilde{\lambda}, \tilde{\mu}, \nu)$  进行缩放 (不影响分离性质), 令  $\nu = 1$ 。此时:

1. 分离不等式变为  $\tilde{\lambda}^T u + \tilde{\mu}^T v + t \geq \alpha$  (对所有  $(u, v, t) \in A$ )。对原始问题的任意可行点  $x$ , 取  $u_i = g_i(x) = 0$ ,  $v_j = h_j(x) \leq 0$ ,  $t = f(x)$ , 代入得:

$$\tilde{\lambda}^T \cdot 0 + \tilde{\mu}^T h(x) + f(x) \geq \alpha$$

2. 对  $B$  的不等式 ( $\nu s \leq \alpha$ ,  $s < p^*$ ), 令  $s \rightarrow p^*$ , 得  $p^* \leq \alpha$  (因  $\nu = 1$ )。

3. 结合 1 和 2, 得  $f(x) + \tilde{\mu}^T h(x) \geq \alpha \geq p^*$ 。对所有可行  $x$  取下确界 (即对偶函数定义):

$$d(\tilde{\lambda}, \tilde{\mu}) = \inf_x \{f(x) + \tilde{\lambda}^T g(x) + \tilde{\mu}^T h(x)\} \geq p^*$$

4. 由弱对偶性 ( $d(\lambda, \mu) \leq p^*$ ), 得  $d(\tilde{\lambda}, \tilde{\mu}) = p^*$ 。

假设存在  $j$  使得  $\tilde{\mu}_j < 0$ , 则选择  $x$  使  $h_j(x)$  足够大 (正值), 可令  $\tilde{\lambda}^T g(x) + \tilde{\mu}^T h(x) + f(x) \rightarrow -\infty$ , 与 “ $\geq p^*$ ” 矛盾, 故  $\tilde{\mu} \geq 0$ 。

$(\tilde{\lambda}, \tilde{\mu})$  是对偶问题的可行解 ( $\tilde{\mu} \geq 0$ ), 且  $d(\tilde{\lambda}, \tilde{\mu}) = p^*$ 。由对偶最优值  $d^* = \sup\{d(\lambda, \mu) \mid \mu \geq 0\}$ , 得  $d^* \geq p^*$ ; 结合弱对偶性  $d^* \leq p^*$ , 故  $d^* = p^*$ , 强对偶成立。

## 11.3 互补松弛条件

设原始问题与对偶问题满足:

1. 原始问题最优解  $x^*$  存在, 对偶问题最优解  $(\lambda^*, \mu^*)$  存在;
2. 强对偶成立 ( $p^* = d^*$ )。

则必有以下两个等价结论:

1. 拉格朗日函数极值等式:  $\mathcal{L}(x^*, \lambda^*, \mu^*) = f(x^*)$ ;
2. 乘子-约束乘积为零: 对所有不等式约束的下标  $j = 1, 2, \dots, p$ , 有

$$\mu_j^* \cdot h_j(x^*) = 0$$

(注: 结论 2 是互补松弛条件的”核心量化形式”, 也是实际应用中最常用的表述。)

证明.

由强对偶成立 ( $p^* = d^*$ ), 结合原始最优值与对偶函数的定义, 可得:

$$f(x^*) = p^* = d^* = d(\lambda^*, \mu^*) \quad (1)$$

根据对偶函数的定义 ( $d(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu)$ ), 对任意  $x \in \mathbb{R}^n$ , 对偶函数是拉格朗日函数的下确界, 因此:

$$d(\lambda^*, \mu^*) \leq \mathcal{L}(x^*, \lambda^*, \mu^*) \quad (2)$$

结合式 (1) 与式 (2), 得:

$$f(x^*) \leq \mathcal{L}(x^*, \lambda^*, \mu^*) \quad (3)$$

将拉格朗日函数在  $(x^*, \lambda^*, \mu^*)$  处展开:

$$\mathcal{L}(x^*, \lambda^*, \mu^*) = f(x^*) + \sum_{i=1}^m \lambda_i^* g_i(x^*) + \sum_{j=1}^p \mu_j^* h_j(x^*) \quad (4)$$

由原始可行性 ( $x^*$  满足等式约束), 对所有  $i = 1, \dots, m$ , 有  $g_i(x^*) = 0$ , 因此式 (4) 中的等式约束乘子项消失:

$$\mathcal{L}(x^*, \lambda^*, \mu^*) = f(x^*) + \sum_{j=1}^p \mu_j^* h_j(x^*) \quad (5)$$

由对偶可行性 ( $\mu^* \geq 0$ ), 对所有  $j = 1, \dots, p$ , 有  $\mu_j^* \geq 0$ ; 由原始可行性 ( $x^*$  满足不等式约束), 对所有  $j = 1, \dots, p$ , 有  $h_j(x^*) \leq 0$ 。

因此, 对每个  $j$ ,  $\mu_j^* \cdot h_j(x^*) \leq 0$  (非正数), 求和后仍为非正数:

$$\sum_{j=1}^p \mu_j^* h_j(x^*) \leq 0 \quad (6)$$

将式 (6) 代入式 (5), 得:

$$\mathcal{L}(x^*, \lambda^*, \mu^*) \leq f(x^*) \quad (7)$$

结合式 (3) ( $f(x^*) \leq \mathcal{L}(x^*, \lambda^*, \mu^*)$ ) 与式 (7) ( $\mathcal{L}(x^*, \lambda^*, \mu^*) \leq f(x^*)$ ), 所有不等式变为等式:

$$\mathcal{L}(x^*, \lambda^*, \mu^*) = f(x^*) \quad (8)$$

将式 (8) 代入式 (5), 得:

$$f(x^*) = f(x^*) + \sum_{j=1}^p \mu_j^* h_j(x^*) \implies \sum_{j=1}^p \mu_j^* h_j(x^*) = 0 \quad (9)$$

由步骤 3 可知, 每个  $\mu_j^* h_j(x^*) \leq 0$  (非正数), 而非正数的和为零, 当且仅当每个非正数均为零。因此:

$$\mu_j^* h_j(x^*) = 0 \quad \forall j = 1, 2, \dots, p \quad (10)$$

式 (8) 与式 (10) 共同构成互补松弛条件。

**注 11.2.** 互补松弛条件  $\mu_j^* h_j(x^*) = 0$  的本质是判断不等式约束对最优解的“活性”, 可分为两种互斥情况, 直观反映约束是否影响最优解:

- 情况 1:  $\mu_j^* > 0$  (对偶乘子为正)。对应约束  $h_j(x) \leq 0$  是紧约束 (起作用的约束)。最优解  $x^*$  恰好落在约束边界上 ( $h_j(x^*) = 0$ ), 该约束限制了目标函数的进一步优化。
- 情况 2:  $\mu_j^* = 0$  (对偶乘子为零)。对应约束  $h_j(x) \leq 0$  是非紧约束 (不起作用的约束)。最优解  $x^*$  落在约束内部 ( $h_j(x^*) < 0$ ), 即使移除该约束, 最优解也不会改变。

(注: 等式约束  $g_i(x) = 0$  始终为“紧约束”, 无互补松弛判断, 因其对偶乘子  $\lambda_i^*$  无符号限制, 无需通过乘积为零判断活性。)

互补松弛条件是 KKT (Karush-Kuhn-Tucker) 最优性条件的重要组成部分。在凸问题 + 强对偶成立 + 函数可微的前提下, KKT 条件是原始-对偶最优解的充要条件, 其结构如下:

### 11.3.1 KKT 条件的完整构成 (含互补松弛)

设原始问题为凸问题 ( $f, h_j$  凸,  $g_i$  仿射),  $f, g_i, h_j$  可微,  $x^*$  为原始最优解,  $(\lambda^*, \mu^*)$  为对偶最优解, 则:

1. 平稳性:  $\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0$  (拉格朗日函数在  $x^*$  处梯度为零, 即无改进方向);
2. 原始可行性:  $g_i(x^*) = 0$  ( $i = 1, \dots, m$ ),  $h_j(x^*) \leq 0$  ( $j = 1, \dots, p$ );
3. 对偶可行性:  $\mu_j^* \geq 0$  ( $j = 1, \dots, p$ );

4. 互补松弛:  $\mu_j^* h_j(x^*) = 0$  ( $j = 1, \dots, p$ )。

可见, 互补松弛条件是 KKT 条件的“闭环环节”——它连接了原始约束的可行性 ( $h_j(x^*) \leq 0$ ) 与对偶乘子的可行性 ( $\mu_j^* \geq 0$ ), 确保原始-对偶最优解的一致性。

## 11.4 对偶理论的应用实例

### 11.4.1 线性规划的对偶

考虑线性规划问题 (原始问题):

$$\begin{cases} \min_x & c^T x \\ \text{subject to} & Ax = b \\ & x \geq 0 \end{cases}$$

其中,  $x \in \mathbb{R}^n$  为原始优化变量,  $c \in \mathbb{R}^n$ 、 $A \in \mathbb{R}^{m \times n}$ 、 $b \in \mathbb{R}^m$  为已知参数。

引入拉格朗日乘子:

- 等式约束  $Ax = b$  对应的乘子:  $\lambda \in \mathbb{R}^m$  (无符号限制);
- 不等式约束  $x \geq 0$  对应的乘子:  $\mu \in \mathbb{R}^n$  (满足  $\mu \geq 0$ )。

拉格朗日函数定义为:

$$\mathcal{L}(x, \lambda, \mu) = c^T x + \lambda^T (b - Ax) - \mu^T x$$

对偶函数  $d(\lambda, \mu)$  是拉格朗日函数关于  $x$  的下确界, 即:

$$d(\lambda, \mu) = \inf_x [c^T x + \lambda^T (b - Ax) - \mu^T x]$$

将函数整理为关于  $x$  的线性形式:

$$d(\lambda, \mu) = \inf_x [(c - A^T \lambda - \mu)^T x + \lambda^T b]$$

为使下确界有限 (避免取值为  $-\infty$ ), 需满足线性项系数为零:

$$c - A^T \lambda - \mu = 0 \implies \mu = c - A^T \lambda$$

结合  $\mu \geq 0$  的约束, 可得:

$$A^T \lambda \leq c$$

将  $\mu = c - A^T \lambda$  代入拉格朗日函数, 对偶函数简化为:

$$d(\lambda) = \lambda^T b$$

对偶问题的目标是最大化对偶函数  $d(\lambda)$ , 同时满足对偶可行性约束, 即:

$$\begin{cases} \max_{\lambda} & b^T \lambda \\ \text{subject to} & A^T \lambda \leq c \end{cases}$$

此为线性规划的标准对偶形式。

### 11.4.2 二次规划的对偶

考虑二次规划问题（原始问题）：

$$\begin{cases} \min_x & \frac{1}{2}x^T Qx + c^T x \\ \text{subject to} & Ax = b \\ & x \geq 0 \end{cases}$$

其中,  $x \in \mathbb{R}^n$  为原始优化变量,  $Q \in \mathbb{R}^{n \times n}$  为半正定矩阵 ( $Q \succeq 0$ , 保证目标函数凸),  $c \in \mathbb{R}^n$ 、 $A \in \mathbb{R}^{m \times n}$ 、 $b \in \mathbb{R}^m$  为已知参数。

引入拉格朗日乘子：

- 等式约束  $Ax = b$  对应的乘子： $\lambda \in \mathbb{R}^m$ （无符号限制）；
- 不等式约束  $x \geq 0$  对应的乘子： $\mu \in \mathbb{R}^n$ （满足  $\mu \geq 0$ ）。

拉格朗日函数定义为：

$$\mathcal{L}(x, \lambda, \mu) = \frac{1}{2}x^T Qx + c^T x + \lambda^T (b - Ax) - \mu^T x$$

对偶函数  $d(\lambda, \mu)$  是拉格朗日函数关于  $x$  的下确界，即：

$$d(\lambda, \mu) = \inf_x \left[ \frac{1}{2}x^T Qx + (c - A^T \lambda - \mu)^T x + \lambda^T b \right]$$

由于目标函数是关于  $x$  的二次函数，且  $Q \succeq 0$ （凸函数），当  $Q$  正定 ( $Q \succ 0$ ) 时，函数存在唯一极小值。对  $x$  求导并令梯度为零，得最优  $x$ ：

$$\nabla_x \mathcal{L} = Qx + (c - A^T \lambda - \mu) = 0 \implies x = -Q^{-1}(c - A^T \lambda - \mu)$$

将最优  $x$  代入拉格朗日函数，化简得对偶函数：

$$d(\lambda, \mu) = -\frac{1}{2} (c - A^T \lambda - \mu)^T Q^{-1} (c - A^T \lambda - \mu) + \lambda^T b$$

对偶问题的目标是最大化对偶函数  $d(\lambda, \mu)$ ，同时满足对偶可行性约束，即：

$$\begin{cases} \max_{\lambda, \mu} & -\frac{1}{2} (c - A^T \lambda - \mu)^T Q^{-1} (c - A^T \lambda - \mu) + \lambda^T b \\ \text{subject to} & \mu \geq 0 \end{cases}$$

### 11.4.3 数值例题（二次规划对偶求解）

以具体二次规划问题为例，验证对偶理论的应用及强对偶性。

$$\begin{cases} \min_{x_1, x_2} & \frac{1}{2}(x_1^2 + x_2^2) \\ \text{subject to} & x_1 + x_2 = 1 \quad (\text{等式约束}) \\ & x_1 \geq 0, x_2 \geq 0 \quad (\text{不等式约束}) \end{cases}$$

**步骤 1：构造拉格朗日函数**

引入乘子：

- 等式约束  $x_1 + x_2 = 1$  对应  $\lambda \in \mathbb{R}$ ;
- 不等式约束  $x_1 \geq 0$ 、 $x_2 \geq 0$  对应  $\mu_1 \geq 0$ 、 $\mu_2 \geq 0$ 。

拉格朗日函数：

$$\mathcal{L}(x, \lambda, \mu) = \frac{1}{2}(x_1^2 + x_2^2) + \lambda(1 - x_1 - x_2) - \mu_1 x_1 - \mu_2 x_2$$

**步骤 2：列写 KKT 条件**

1. 平稳性： $\nabla_x \mathcal{L} = 0$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial x_1} &= x_1 - \lambda - \mu_1 = 0 \\ \frac{\partial \mathcal{L}}{\partial x_2} &= x_2 - \lambda - \mu_2 = 0\end{aligned}$$

2. 原始可行性： $x_1 + x_2 = 1$ ,  $x_1 \geq 0$ ,  $x_2 \geq 0$ ;

3. 对偶可行性： $\mu_1 \geq 0$ ,  $\mu_2 \geq 0$ ;

4. 互补松弛： $\mu_1 x_1 = 0$ ,  $\mu_2 x_2 = 0$ 。

**步骤 3：分析最优解**

假设  $x_1 > 0$  且  $x_2 > 0$ ，由互补松弛条件得  $\mu_1 = \mu_2 = 0$ 。代入平稳性条件：

$$x_1 = \lambda, \quad x_2 = \lambda$$

结合等式约束  $x_1 + x_2 = 1$ ，得  $\lambda = 0.5$ ,  $x_1 = x_2 = 0.5$ 。

验证所有约束均满足，因此原始最优解为  $x^* = (0.5, 0.5)$ ，原始最优值  $p^* = \frac{1}{2}(0.5^2 + 0.5^2) = 0.25$ 。

**步骤 4：对偶问题求解**

对偶函数是拉格朗日函数关于  $x_1, x_2$  的下确界：

$$d(\lambda, \mu) = \inf_{x_1, x_2} \left[ \frac{1}{2}(x_1^2 + x_2^2) + \lambda(1 - x_1 - x_2) - \mu_1 x_1 - \mu_2 x_2 \right]$$

对  $x_1, x_2$  求导并令梯度为零，得：

$$x_1 = \lambda + \mu_1, \quad x_2 = \lambda + \mu_2$$

将  $x_1, x_2$  代入拉格朗日函数，展开化简：

$$\begin{aligned}d(\lambda, \mu) &= \frac{1}{2} [(\lambda + \mu_1)^2 + (\lambda + \mu_2)^2] + \lambda(1 - 2\lambda - \mu_1 - \mu_2) \\ &\quad - \mu_1(\lambda + \mu_1) - \mu_2(\lambda + \mu_2) \\ &= -\lambda^2 - \lambda\mu_1 - \lambda\mu_2 - \frac{1}{2}\mu_1^2 - \frac{1}{2}\mu_2^2 + \lambda\end{aligned}$$

对偶问题为：

$$\begin{cases} \max_{\lambda, \mu_1, \mu_2} & -\lambda^2 - \lambda\mu_1 - \lambda\mu_2 - \frac{1}{2}\mu_1^2 - \frac{1}{2}\mu_2^2 + \lambda \\ \text{subject to} & \mu_1 \geq 0, \mu_2 \geq 0 \end{cases}$$

由原始最优解的互补松弛条件 ( $\mu_1 = \mu_2 = 0$ )，代入对偶函数：

$$d(\lambda, 0, 0) = -\lambda^2 + \lambda$$

对  $\lambda$  求导并令导数为零，得  $\lambda = 0.5$ ，此时对偶最优值  $d^* = -(0.5)^2 + 0.5 = 0.25$ 。

原始最优值  $p^* = 0.25$ ，对偶最优值  $d^* = 0.25$ ，满足  $p^* = d^*$ ，强对偶成立。

#### 11.4.4 SVM 中的原问题与对偶问题

支持向量机 (SVM) 是对偶理论在机器学习中的典型应用，核心是通过对偶问题简化高维特征空间中的优化求解。

**关键结论 (原对偶关系)**

- 原始问题 (Primal)：优化变量为模型参数  $w$  (权重向量)、 $b$  (偏置)，目标是最大化几何间隔；
- 对偶问题 (Dual)：优化变量为对偶乘子  $\alpha_i$  (对应每个样本)，目标是最大化对偶函数；
- 等价性：通过 KKT 条件联结原对偶解，满足  $w = \sum_i \alpha_i y_i x_i$ 、 $\sum_i \alpha_i y_i = 0$ 、 $0 \leq \alpha_i \leq C$  ( $C$  为罚系数)；
- 支持向量特性：仅  $\alpha_i > 0$  的样本 (支持向量) 决定  $w$  与几何间隔  $1/\|w\|$ ；
- 核化能力：对偶问题仅依赖样本内积  $\langle x_i, x_j \rangle$ ，可直接替换为核函数  $K(x_i, x_j)$ ，实现高维空间映射。

#### 软间隔 SVM 的原始问题

目标：在允许样本“软违约” (用  $\xi_i \geq 0$  表示违约程度) 的前提下，最小化  $\frac{1}{2}\|w\|^2$  (等价于最大化几何间隔)，同时控制违约惩罚。

$$\begin{cases} \min_{w, b, \xi} & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - \xi_i \quad (i = 1, \dots, n) \\ & \xi_i \geq 0 \quad (i = 1, \dots, n) \end{cases}$$

其中， $C > 0$  为罚系数 (平衡间隔大小与违约惩罚)， $y_i \in \{+1, -1\}$  为样本标签， $\xi_i$  为违约变量。

#### 拉格朗日函数与 KKT 条件

引入对偶乘子：

- 约束  $y_i(w^T x_i + b) \geq 1 - \xi_i$  对应  $\alpha_i \geq 0$ ；
- 约束  $\xi_i \geq 0$  对应  $\beta_i \geq 0$ 。

拉格朗日函数:

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i(w^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^n \beta_i \xi_i$$

对  $w, b, \xi_i$  求导并令梯度为零:

- 对  $w$ :  $\nabla_w \mathcal{L} = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \implies w = \sum_{i=1}^n \alpha_i y_i x_i$ ;
- 对  $b$ :  $\nabla_b \mathcal{L} = -\sum_{i=1}^n \alpha_i y_i = 0 \implies \sum_{i=1}^n \alpha_i y_i = 0$ ;
- 对  $\xi_i$ :  $\nabla_{\xi_i} \mathcal{L} = C - \alpha_i - \beta_i = 0 \implies \alpha_i + \beta_i = C$ 。

结合对偶可行性  $\alpha_i \geq 0, \beta_i \geq 0$ , 得  $0 \leq \alpha_i \leq C$ 。

### 软间隔 SVM 的对偶问题

消去原始变量  $w, b, \xi_i$ , 代入拉格朗日函数, 最终对偶问题为:

$$\begin{cases} \max_{\alpha} & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{subject to} & 0 \leq \alpha_i \leq C \quad (i = 1, \dots, n) \\ & \sum_{i=1}^n \alpha_i y_i = 0 \end{cases}$$

**核技巧应用** 将内积  $\langle x_i, x_j \rangle$  替换为核函数  $K(x_i, x_j)$  (如线性核  $K(x_i, x_j) = x_i^T x_j$ 、RBF 核  $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ ), 即可得到核 SVM 的对偶形式, 解决高维特征空间的优化问题。

### 原对偶的意义与决策函数

#### 原对偶的核心价值

- **维度优势**: 当特征维度  $d$  极大 (如文本分类)、样本数  $n$  较小时, 对偶问题 (变量数为  $n$ ) 比原始问题 (变量数为  $d+1$ ) 更易求解;
- **可解释性**:  $\alpha_i$  可视为“样本重要性权重” (支持向量的  $\alpha_i > 0$ , 非支持向量的  $\alpha_i = 0$ );
- **最优性判据**: 原始问题值  $P(w, b, \xi)$  与对偶问题值  $D(\alpha)$  的对偶间隙  $P - D \geq 0$ , 可作为算法收敛的停机准则。

#### 决策函数与偏置计算

- **决策函数**: 已知对偶最优解  $\alpha^*$ , 结合核函数  $K$ , 模型对新样本  $x$  的预测为:

$$f(x) = \text{sign} \left( \sum_{i=1}^n \alpha_i^* y_i K(x_i, x) + b^* \right)$$

- **偏置  $b^*$** : 取任意满足  $0 < \alpha_i^* < C$  的支持向量  $x_i$ , 代入约束  $y_i(w^* x_i + b^*) = 1$ , 得:

$$b^* = y_i - \sum_{j=1}^n \alpha_j^* y_j K(x_j, x_i)$$

## 第十二章 约束优化问题解法

本章系统介绍常见的带约束优化问题求解思想与典型算法，包括外点法（惩罚函数法）、增广拉格朗日法（ALM）以及在可分结构下广泛应用的 ADMM（交替方向乘子法）。

### 12.1 外点法

外点法是惩罚函数法（Penalty Function Method）的核心分支，属于约束优化的间接解法——通过将约束条件转化为目标函数的惩罚项，将原约束问题转化为一系列无约束优化子问题，迭代求解无约束子问题的最优解，使其逐步逼近原约束问题的最优解。其核心特征是：迭代点始终位于可行域外部，通过惩罚项迫使迭代点向可行域边界收敛。

#### 12.1.1 核心定义与问题形式

定义 12.1 (原约束优化问题（标准形式）). 设优化变量  $\mathbf{x} \in \mathbb{R}^n$ ，原问题定义为：

$$\begin{cases} \min_{\mathbf{x}} & f(\mathbf{x}) \quad (\text{目标函数, 连续可微}) \\ \text{s.t.} & g_i(\mathbf{x}) \leq 0 \quad (i = 1, 2, \dots, m) \quad (\text{不等式约束}) \\ & h_j(\mathbf{x}) = 0 \quad (j = 1, 2, \dots, p) \quad (\text{等式约束}) \end{cases}$$

其中：

- 可行域  $\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0\}$ ;
- 假设  $\Omega \neq \emptyset$  (可行域非空)，且原问题存在最优解  $\mathbf{x}^*$ 。

#### 2. 外点法的核心思想

- 对于可行域外部的点（违反约束的点），通过惩罚项施加“惩罚”，使其目标函数值增大；
- 惩罚强度由惩罚参数  $\mu > 0$  控制，且  $\mu$  随迭代逐步增大 ( $\mu_k \rightarrow +\infty$ )；
- 当  $\mu$  足够大时，无约束子问题的最优解会“被迫”靠近可行域，最终收敛到原问题的最优解  $\mathbf{x}^*$ 。

#### 12.1.2 数学建模：惩罚函数构造

外点法的核心是设计惩罚函数  $P(\mathbf{x}, \mu)$ ，其通用形式为：

$$P(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu \cdot \Phi(\mathbf{x})$$

其中：

- $f(\mathbf{x})$  为原目标函数；
- $\mu > 0$  为惩罚参数（迭代中满足  $\mu_{k+1} > \mu_k$ ，且  $\mu_k \rightarrow +\infty$ ）；
- $\Phi(\mathbf{x})$  为约束违反度量函数（非负、连续可微），用于量化点  $\mathbf{x}$  对约束的违反程度，满足：

$$\Phi(\mathbf{x}) = 0 \iff \mathbf{x} \in \Omega \quad (\text{可行点无惩罚})$$

$$\Phi(\mathbf{x}) > 0 \iff \mathbf{x} \notin \Omega \quad (\text{不可行点有惩罚，违反越严重惩罚越大})$$

### 3. 约束违反度量函数的具体形式

根据约束类型（不等式/等式）， $\Phi(\mathbf{x})$  通常分解为两部分：

$$\Phi(\mathbf{x}) = \sum_{i=1}^m \phi(g_i(\mathbf{x})) + \sum_{j=1}^p \psi(h_j(\mathbf{x}))$$

其中：

(1) 不等式约束惩罚项  $\phi(g_i(\mathbf{x}))$  最常用的是二次惩罚（连续可微，便于无约束优化求解）：

$$\phi(t) = \max(0, t)^2 = \begin{cases} t^2 & t > 0 \quad (\text{违反约束，施加惩罚}) \\ 0 & t \leq 0 \quad (\text{满足约束，无惩罚}) \end{cases}$$

- 其他形式：一次惩罚  $\phi(t) = \max(0, t)$ （不可微，仅用于简单问题）、指数惩罚  $\phi(t) = e^{\alpha t} - 1$ （ $\alpha > 0$ ，惩罚增长更快）。

(2) 等式约束惩罚项  $\psi(h_j(\mathbf{x}))$  等式约束无“满足/违反”的中间状态，直接惩罚偏差：

$$\psi(t) = t^2 \quad (\text{二次惩罚，最常用})$$

- 其他形式： $\psi(t) = |t|$ （不可微）、 $\psi(t) = t^4$ （惩罚增长更快）。

### 4. 完整惩罚函数示例（二次惩罚）

结合上述形式，二次惩罚的外点法惩罚函数为：

$$P(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu \left[ \sum_{i=1}^m \max(0, g_i(\mathbf{x}))^2 + \sum_{j=1}^p h_j(\mathbf{x})^2 \right]$$

#### 12.1.3 算法流程（严格形式化）

**算法 12.1** (外点法算法流程). 输入：

- 原问题目标函数  $f(\mathbf{x})$ 、约束  $g_i(\mathbf{x})$ 、 $h_j(\mathbf{x})$ ；
- 初始参数：初始点  $\mathbf{x}_0 \in \mathbb{R}^n$ （可在可行域外部）、初始惩罚参数  $\mu_1 > 0$ 、惩罚参数增长因子  $\beta > 1$ （通常取 10）、收敛精度  $\epsilon > 0$ 。

迭代步骤：

1. 初始化：令迭代次数  $k = 1$ ；

2. 构造惩罚函数：针对当前  $\mu_k$ ，构造  $P(\boldsymbol{x}, \mu_k)$ ;
3. 求解无约束子问题：以  $\boldsymbol{x}_{k-1}$  为初始点，求解  $\min_{\boldsymbol{x}} P(\boldsymbol{x}, \mu_k)$ ，得到最优解  $\boldsymbol{x}_k$ ;
4. 收敛判断：若满足以下任一收敛条件，停止迭代，输出  $\boldsymbol{x}_k \approx \boldsymbol{x}^*$ ：
  - 约束违反度量足够小：  $\Phi(\boldsymbol{x}_k) < \epsilon$ ;
  - 目标函数变化足够小：  $\|f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k-1})\| < \epsilon$ ;
  - 迭代点变化足够小：  $\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\| < \epsilon$ ;
5. 更新惩罚参数：令  $\mu_{k+1} = \beta \cdot \mu_k$ ， $k = k + 1$ ，返回步骤 2。
- 输出：原约束问题的近似最优解  $\boldsymbol{x}_k$ 。

注 12.1 (无约束子问题求解).

场景分类	推荐方法	工具/实现	核心注意事项
低维 ( $n \leq 10$ ) + 光滑 $P(\boldsymbol{x}, \mu)$	解析法(求梯度 =0 解方程组)	手动推导	仅适合简单问题/教学验证
中高维 ( $n \leq 1000$ ) + 光滑 $P(\boldsymbol{x}, \mu)$ (工程主流)	拟牛顿法 (L-BFGS/BFGS), 梯度下降法等等	<i>Scipy.fmin_bfgs</i> , <i>Matlab.fminunc</i>	1. 传前一轮 $\boldsymbol{x}_{k-1}$ 作初始点; 2. 配合线搜索; 3. 病态时优先 L-BFGS
高维 ( $n > 1000$ ) + 光滑 $P(\boldsymbol{x}, \mu)$	共轭梯度法	<i>Scipy.fmin_cg</i>	低内存消耗, 适合大规模问题
不可微 $P(\boldsymbol{x}, \mu)$ (一次惩罚等)	梯度自由法 (Nelder-Mead)	<i>Scipy.fmin</i>	收敛慢, 仅用于简单非光滑问题

注 12.2 (惩罚参数更新方法).

更新方法	公式/逻辑	优点	缺点	适用场景
经典倍增法则 (基础)	$\mu_{k+1} = \beta \mu_k$ ( $\beta = 5 \sim 10$ )	实现最简单, 无需额外计算	易病态, 需调 $\beta$	快速验证、简单问题
自适应更新法 (首选)	阈值触发: $\Phi(\boldsymbol{x}_k) > \eta_k$ 则 $\mu_{k+1} = \beta \mu_k$ , 否则保持	平衡收敛速度与稳定性, 鲁棒性强	需计算约束违反度量 $\Phi(\boldsymbol{x}_k)$	工程实践、中高维问题 (优先选)
线性增长法	$\mu_{k+1} = \mu_k + c$ ( $c > 0$ )	数值稳定性最好, 不易病态	收敛慢	高维等式约束、易病态问题
指数增长法	$\mu_{k+1} = \beta^k \mu_1$	收敛极快	极易病态	约束违反程度下降快的简单问题
KKT 残量法	按 KKT 残量 $r_k$ 动态调整 $\mu$	精准, 收敛速率高	复杂, 需估计拉格朗日乘子	高精度需求、学术研究

12.1.4 收敛性分析（核心结论）

**定理 12.1** (外点法收敛性). 外点法的收敛性依赖于惩罚参数  $\mu_k \rightarrow +\infty$ , 关键结论如下 (严格证明需用到变分不等式或闭映射理论):

1. **序列有界性**: 若原问题最优解存在, 且惩罚函数  $P(\mathbf{x}, \mu)$  是强制函数 (当  $\|\mathbf{x}\| \rightarrow +\infty$  时  $P(\mathbf{x}, \mu) \rightarrow +\infty$ ), 则迭代序列  $\{\mathbf{x}_k\}$  有界;
2. **收敛性**: 设  $\{\mathbf{x}_k\}$  是迭代序列, 其任一聚点  $\mathbf{x}^*$  都是原约束问题的最优解;
3. **收敛速率**: 二次惩罚外点法的收敛速率为 **线性收敛** (当  $\mu_k$  按指数增长时, 可达到超线性收敛)。

**关键直观解释** 当  $\mu_k$  增大时, 惩罚项权重越来越大:

- 若  $\mathbf{x}_k$  仍在可行域外部, 惩罚项会主导  $P(\mathbf{x}, \mu_k)$ , 迫使  $\mathbf{x}_{k+1}$  向可行域靠近;
- 当  $\mu_k \rightarrow +\infty$  时, 可行域外部的点会被赋予无穷大惩罚, 因此无约束子问题的最优解必须“落在”可行域边界或内部, 即收敛到原问题最优解。

### 12.1.5 优缺点

#### 优点

1. **初始点灵活**: 无需初始点在可行域内 (内点法必须初始点可行), 尤其适合可行域难以构造初始点的问题;
2. **构造简单**: 惩罚函数形式直观, 无约束子问题可直接用梯度下降、牛顿法等成熟算法求解;
3. **兼容性强**: 可同时处理不等式约束和等式约束, 无需单独设计逻辑。

#### 缺点

1. **惩罚病 (Penalty Ill-Conditioning)**: 当  $\mu_k$  过大时, 惩罚函数  $P(\mathbf{x}, \mu_k)$  的 Hessian 矩阵会呈现“病态” (条件数极大), 导致无约束子问题求解困难 (梯度下降步长过小、收敛变慢);
2. **仅收敛到可行域边界**: 对于不等式约束, 最优解若在可行域内部 (内点), 外点法仍会收敛到边界 (需结合其他准则修正);
3. **线性收敛速率**: 相比内点法 (超线性收敛), 收敛速度较慢, 适合中小规模约束优化问题。

**深入分析:**  $\mu_k$  过大会让 Hessian 病态?

1. **Gauss-Newton 近似下的 Hessian 结构**: 在 Gauss-Newton 近似 (忽略约束的二阶项) 下, 罚函数的 Hessian 可以拆成两部分:

$$H_{\mu_k} := \nabla^2 \Phi_{\mu_k}(\mathbf{x}) \approx \underbrace{\nabla^2 f(\mathbf{x})}_{H_f} + \mu_k \underbrace{J_h(\mathbf{x})^T J_h(\mathbf{x})}_{H_h}$$

通俗解释:

- $H_f$  是原目标函数的 Hessian: 反映原目标在当前点的“曲率”;

- $H_h$  是等式约束惩罚项的 **Hessian**：由约束的雅可比矩阵外积得到，反映约束对惩罚项的“影响强度”；
- $\mu_k$  是惩罚参数：控制约束惩罚项的权重。

2. 特征值的变化（矩阵“伸缩能力”的改变）：当  $\mu_k$  很大时， $H_{\mu_k}$  的特征值会出现“两极分化”：

$$\lambda_{\max}(H_{\mu_k}) \approx \lambda_{\max}(H_f) + \mu_k \sigma_{\max}^2 \quad (\text{最大特征值被 } \mu_k \text{ 放大})$$

$$\lambda_{\min}(H_{\mu_k}) \approx \lambda_{\min}(H_f) \quad (\text{最小特征值基本不变})$$

3. 条件数暴增  $\rightarrow$  **Hessian 病态**：矩阵的“条件数”是 最大特征值  $\div$  最小特征值，用来衡量矩阵的“病态程度”：

$$\kappa(H_{\mu_k}) = \frac{\lambda_{\max}(H_{\mu_k})}{\lambda_{\min}(H_{\mu_k})} \approx \frac{\lambda_{\max}(H_f) + \mu_k \sigma_{\max}^2}{\lambda_{\min}(H_f)}$$

当  $\mu_k \rightarrow +\infty$  时，条件数  $\kappa(H_{\mu_k}) \rightarrow +\infty$  ——这就是“Hessian 病态”。

4. 病态的后果：

- 线性系统求解困难：迭代法收敛慢，直接法数值不稳定。
- 线搜索步长受限：病态 Hessian 对应的二次模型会变成“很尖的山谷”，导致线搜索只能小步慢挪。

5. 结论：不要一开始把  $\mu_k$  设太大，应逐步增大。

### 注 12.3. 深入分析：等式约束惩罚项的本质

等式二次罚函数形式  $\frac{\mu_k}{2} \|h(x)\|^2 = \frac{\mu_k}{2} \sum_i h_i(x)^2$ ，是等式约束下二次罚函数的标准形式（乘以  $\frac{1}{2}$  是为了推导方便）。

对罚函数  $\frac{\mu_k}{2} \|h(x)\|^2$  求梯度：根据链式法则，单个  $h_i(x)^2$  的梯度是  $2h_i(x)\nabla h_i(x)$ ，汇总后为  $\mu_k \sum_i h_i(x)\nabla h_i(x)$ 。而雅可比矩阵  $J_h(x)$  的定义是“每行对应  $\nabla h_i(x)^T$ ”，因此  $J_h(x)^T$  的每列对应  $\nabla h_i(x)$ ，乘以  $h(x)$ （列向量）恰好得到  $\sum_i h_i(x)\nabla h_i(x)$ 。最终梯度  $\nabla \left( \frac{\mu_k}{2} \|h\|^2 \right) = \mu_k J_h(x)^T h(x)$ 。

对梯度  $\mu_k J_h(x)^T h(x)$  求 Hessian（二阶导数），需用乘积求导法则：

$$\nabla^2 (\mu_k J_h^T h) = \mu_k (\nabla(J_h^T) \cdot h + J_h^T \cdot \nabla h)$$

- 第一项  $\nabla(J_h^T) \cdot h$  对应“雅可比矩阵的导数乘以  $h$ ”，即  $\sum_i h_i(x) \nabla^2 h_i(x)$ （因为  $J_h$  的元素是  $\partial h_i / \partial x_j$ ，其导数是  $\partial^2 h_i / \partial x_j \partial x_k$ ，即  $\nabla^2 h_i$  的元素）；
- 第二项  $J_h^T \cdot \nabla h$  对应“ $J_h^T$  乘以  $h$  的雅可比（即  $J_h$ ）”，即  $J_h^T J_h$ 。

因此 Hessian  $\nabla^2 \left( \frac{\mu_k}{2} \|h\|^2 \right) = \mu_k (J_h^T J_h + \sum_i h_i \nabla^2 h_i)$ 。

当迭代点处于可行邻域 ( $h(x) \approx 0$ ) 时， $\sum_i h_i \nabla^2 h_i \approx 0$ ；而 **Gauss-Newton** 近似本身就是“忽略残差项的二阶导数（即  $\sum_i h_i \nabla^2 h_i$ ）”，因此简化后得到  $\nabla^2 \Phi_{\mu_k(x)} \approx \nabla^2 f(x) + \mu_k J_h^T J_h$ ，是合理的近似（工程中广泛使用）。

- $\mu_k J_h^T J_h$  是正半定矩阵：对任意向量  $d$ ，有  $d^T (\mu_k J_h^T J_h) d = \mu_k \|J_h d\|^2 \geq 0$ ，符合“正则项需正半定”的要求；
- 特征值与方向的关系：

- 沿等式法向 ( $J_h$  的行空间, 即不可行方向):  $J_h d \neq 0$ , 正则项贡献  $\mu_k \|J_h d\|^2$ , 特征值随  $\mu_k$  线性增大, 实现“强回拉”;
- 沿切空间 ( $J_h d = 0$ , 即可行方向): 正则项贡献为 0, 由原目标函数的 Hessian  $\nabla^2 f$  主导, 不干扰可行方向的搜索。

深入分析:  $(g(x)^+)^2$  只惩罚越界方向

一维标量情形 对于一维变量  $t$ , 定义惩罚函数:

$$r(t) = (t_+)^2 = \begin{cases} t^2, & t > 0 \\ 0, & t \leq 0 \end{cases}$$

其中  $t_+ = \max(0, t)$ , 表示只取  $t$  的非负部分 (即“越界”部分)。

导数分析 (一阶性质) 该惩罚函数的导数为:

$$r'(t) = \begin{cases} 0, & t \leq 0 \\ 2t, & t > 0 \end{cases}$$

关键性质: 在  $t = 0$  处, 左导数 ( $t \rightarrow 0^-$ ) 为 0, 右导数 ( $t \rightarrow 0^+$ ) 也为 0, 因此  $r(t)$  在  $t = 0$  处可导 (光滑过渡)。

直观理解:

- 当  $t \leq 0$  (未越界) 时, 导数为 0, 惩罚函数无变化趋势, 即“不惩罚”;
- 当  $t > 0$  (越界) 时, 导数为  $2t$  (随越界程度增大而增大), 惩罚函数随  $t$  增大而快速增长, 即“只惩罚越界方向”。

推广到多维不等式约束分量  $g_j(x)$  对于优化问题中的第  $j$  个不等式约束  $g_j(x) \leq 0$  (可行域要求  $g_j(x) \leq 0$ , 越界即  $g_j(x) > 0$ ), 其惩罚项  $(g_j(x)^+)^2$  的梯度为:

$$\nabla(g_j(x)^+)^2 = 2g_j(x)^+ \nabla g_j(x) = \begin{cases} 0, & g_j(x) \leq 0 \\ 2g_j(x) \nabla g_j(x), & g_j(x) > 0 \end{cases}$$

核心结论:

- 当约束满足时 ( $g_j(x) \leq 0$ ), 梯度为 0, 惩罚项对优化方向无影响 (不惩罚);
- 当约束被违反时 ( $g_j(x) > 0$ ), 梯度非零 (与  $\nabla g_j(x)$  同向), 推动迭代点向  $g_j(x)$  减小的方向移动 (即可向可行域回归, 只惩罚越界方向)。

Hessian 矩阵分析 (二阶性质) 惩罚项  $(g_j(x)^+)^2$  的二阶导数 (Hessian 矩阵) 为:

$$\nabla^2(g_j(x)^+)^2 = 2\nabla g_j(x) \nabla g_j(x)^T + 2g_j(x) \nabla^2 g_j(x)$$

可行边界附近的近似: 当迭代点接近可行边界 ( $g_j(x) \approx 0^+$ , 即刚越界时),  $g_j(x) \approx 0$ , 第二项可忽略, 因此 Hessian 近似为:

$$\nabla^2(g_j(x)^+)^2 \approx 2\nabla g_j(x) \nabla g_j(x)^T$$

意义: 近似后的 Hessian 是半正定矩阵 (外积形式), 其“增强方向”与约束的梯度  $\nabla g_j(x)$  一致 (即越界方向)。这意味着在可行边界附近, 惩罚项的二阶特性会“抬升”越界方向的曲率, 进一步阻止迭代点向越界方向移动, 强化对越界方向的惩罚。

**总结**  $(g_j(x)^+)^2$  作为不等式约束的惩罚项，通过一阶导数（梯度）和二阶导数（Hessian）的设计，实现了“只在约束被违反时生效”的特性：

- 可行域内 ( $g_j(x) \leq 0$ )：惩罚项及其导数均为 0，不干扰优化；
- 可行域外 ( $g_j(x) > 0$ )：惩罚项随越界程度增大而增长，梯度和 Hessian 引导迭代向可行域回归，精准惩罚越界方向。

这种特性使其成为不等式约束外点法中最常用的惩罚形式（连续可微且惩罚针对性强）。

### 12.1.6 示例：外点法求解

我们用外点法求解该约束优化问题，步骤如下：

#### 一、问题定义

目标函数：  $f(x) = x_1^2 + 2x_2^2 - 2x_1 - 2x_2$  约束：

- 等式约束：  $h(x) = x_1 + x_2 - 1 = 0$
- 不等式约束：  $g(x) = x_1 - 0.6 \leq 0$

#### 二、外点法惩罚函数构造

外点法通过惩罚项将约束转化为无约束问题：

- 等式约束用二次惩罚：  $h(x)^2$
- 不等式约束仅惩罚越界部分：  $(\max(0, g(x)))^2$

因此，惩罚函数为：

$$P(x, \mu_k) = f(x) + \mu_k [h(x)^2 + (\max(0, x_1 - 0.6))^2]$$

其中  $\mu_k > 0$  是惩罚参数，需逐步增大 ( $\mu_k \rightarrow +\infty$ )。

#### 三、求解无约束子问题（对固定 $\mu_k$ ）

对  $P(x, \mu_k)$  求偏导并令其为 0，分不等式约束越界 ( $x_1 > 0.6$ ) 和不越界 ( $x_1 \leq 0.6$ ) 两种情况：

**情况 1：**  $x_1 \leq 0.6$ （惩罚项中  $\max(0, x_1 - 0.6) = 0$ ） 惩罚函数简化为：

$$P(x, \mu_k) = x_1^2 + 2x_2^2 - 2x_1 - 2x_2 + \mu_k(x_1 + x_2 - 1)^2$$

求偏导并令其为 0：

$$\begin{cases} \frac{\partial P}{\partial x_1} = 2x_1 - 2 + 2\mu_k(x_1 + x_2 - 1) = 0 \\ \frac{\partial P}{\partial x_2} = 4x_2 - 2 + 2\mu_k(x_1 + x_2 - 1) = 0 \end{cases}$$

两式相减得  $x_1 = 2x_2$ ，代入后解得：

$$x_1 = \frac{2(1 + \mu_k)}{2 + 3\mu_k}, \quad x_2 = \frac{1 + \mu_k}{2 + 3\mu_k}$$

但当  $\mu_k > 0$  时， $x_1 = \frac{2(1 + \mu_k)}{2 + 3\mu_k} > 0.6$ （验证： $2(1 + \mu_k) > 0.6(2 + 3\mu_k)$  恒成立），因此情况 1 不成立。

情况 2:  $x_1 > 0.6$  (惩罚项中  $\max(0, x_1 - 0.6) = x_1 - 0.6$ ) 惩罚函数为:

$$P(x, \mu_k) = x_1^2 + 2x_2^2 - 2x_1 - 2x_2 + \mu_k [(x_1 + x_2 - 1)^2 + (x_1 - 0.6)^2]$$

求偏导并令其为 0:

$$\begin{cases} \frac{\partial P}{\partial x_1} = 2x_1 - 2 + 2\mu_k(x_1 + x_2 - 1) + 2\mu_k(x_1 - 0.6) = 0 \\ \frac{\partial P}{\partial x_2} = 4x_2 - 2 + 2\mu_k(x_1 + x_2 - 1) = 0 \end{cases}$$

两式相减消去  $2\mu_k(x_1 + x_2 - 1)$ , 得:

$$x_1(1 + \mu_k) - 2x_2 = 0.6\mu_k \implies x_2 = \frac{x_1(1 + \mu_k) - 0.6\mu_k}{2}$$

代入偏导方程, 最终解得:

$$x_1 = \frac{0.6\mu_k^2 + 3.2\mu_k + 2}{\mu_k^2 + 5\mu_k + 2}, \quad x_2 = \frac{0.4\mu_k^2 + 2\mu_k + 1}{\mu_k^2 + 5\mu_k + 2}$$

四、令  $\mu_k \rightarrow +\infty$ , 求极限解

当  $\mu_k$  足够大时, 分子分母的高次项主导, 因此:

$$\lim_{\mu_k \rightarrow +\infty} x_1 = \frac{0.6\mu_k^2}{\mu_k^2} = 0.6, \quad \lim_{\mu_k \rightarrow +\infty} x_2 = \frac{0.4\mu_k^2}{\mu_k^2} = 0.4$$

五、验证最优解

- 约束满足:  $x_1 + x_2 = 0.6 + 0.4 = 1$  (等式约束),  $x_1 - 0.6 = 0$  (不等式约束)。
- 目标函数值:  $f(0.6, 0.4) = 0.6^2 + 2 \times 0.4^2 - 2 \times 0.6 - 2 \times 0.4 = -1.32$ 。

最终结果: 原问题的最优解为  $\mathbf{x}^* = (0.6, 0.4)$ , 最优目标函数值为  $f(\mathbf{x}^*) = -1.32$ 。

### 12.1.7 外点法的极限满足 KKT

#### 1. 核心定理/引理 (基于外点法的推导)

先将惩罚函数中的  $\frac{\rho}{2}$  替换为  $\mu_k$ , 对应惩罚函数为:

$$\Phi_{\mu_k}(x) = f(x) + \mu_k \|h(x)\|^2 + \mu_k \|g(x)^+\|^2 \quad (g(x)^+ \text{ 是 } g \text{ 的非负部分})$$

**引理 12.1** (可行性残差必趋零). 外点法的迭代点  $x^{(k)}$  满足  $h(x^{(k)}) \rightarrow 0$ 、 $g(x^{(k)})^+ \rightarrow 0$ 。即迭代点会逐渐趋近原问题的可行域 (约束违反程度趋近于 0)。

**引理 12.2** (构造候选乘子与近似平稳). 定义候选乘子  $v^{(k)} = \mu_k h(x^{(k)})$ 、 $\lambda_j^{(k)} = \mu_k g_j(x^{(k)})^+$  ( $\lambda^{(k)} \geq 0$ ), 则迭代点满足“近似平稳条件”:

$$\nabla f(x^{(k)}) + J_h(x^{(k)})^T v^{(k)} + \sum_{j=1}^p \lambda_j^{(k)} \nabla g_j(x^{(k)}) = r^{(k)}$$

其中  $r^{(k)} \rightarrow 0$  (近似误差趋近于 0)。

**引理 12.3** (MFCQ 推出乘子有界). 若 MFCQ (约束规范) 成立, 则候选乘子  $v^{(k)}$ 、 $\lambda^{(k)}$  是有界的; 且非活跃约束 ( $g_j(x^*) < 0$ ) 对应的  $\lambda_j^{(k)} \rightarrow 0$ 。

**定理 12.2** (外点法的极限满足 KKT). 外点法的迭代点极限  $x^*$  是原问题的 KKT 点。

## 2. 为什么能证明外点法极限满足 KKT?

通过三个引理逐步“补全”KKT 条件:

1. 引理 1 保证极限点满足约束:  $h(x^*) = 0$ 、 $g(x^*) \leq 0$ ;
2. 引理 2 构造了候选乘子, 得到近似平稳条件 (接近 KKT 的梯度条件);
3. 引理 3 通过 MFCQ 让乘子有界, 从而乘子存在收敛子列;
4. 对近似平稳条件取极限 ( $r^{(k)} \rightarrow 0$ 、乘子收敛), 最终得到完整的 KKT 条件。

## 3. 外点法极限满足的 KKT 条件描述

外点法的迭代点极限  $x^*$  满足以下 KKT 条件:

1. 约束条件:  $h(x^*) = 0$ ,  $g(x^*) \leq 0$  (处于可行域内);
2. 梯度平稳条件: 存在乘子  $v^*$ 、 $\lambda^* \geq 0$ , 使得:

$$\nabla f(x^*) + J_h(x^*)^T v^* + J_g(x^*)^T \lambda^* = 0$$

( $J_h, J_g$  是  $h, g$  的雅可比矩阵);

3. 互补松弛条件: 对每个约束  $j$ ,  $\lambda_j^* g_j(x^*) = 0$  (非活跃约束的乘子为 0)。

# 12.2 增广拉格朗日 (ALM)

## 12.2.1 从外点法到增广拉格朗日法 (ALM)

外点法的核心缺陷是 惩罚参数  $\mu \rightarrow +\infty$  导致的数值病态。增广拉格朗日法 (Augmented Lagrangian Method, ALM) 的核心改进的是: 引入拉格朗日乘子近似反映约束的“优先级”, 将“单纯增大惩罚”改为“乘子调整 + 适度惩罚”, 使得惩罚参数  $\mu$  无需趋于无穷大即可实现收敛, 从根本上解决了外点法的数值病态问题。

### 12.2.2 拉格朗日函数的基础铺垫

原约束优化问题的拉格朗日函数定义为：

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

其中：

- $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)^T \geq \mathbf{0}$  (不等式约束的拉格朗日乘子)；
- $\boldsymbol{\nu} = (\nu_1, \dots, \nu_p)^T \in \mathbb{R}^p$  (等式约束的拉格朗日乘子)。

外点法的本质是“忽略乘子，仅用惩罚强制满足约束”，而 ALM 则是“用乘子近似 KKT 条件中的对偶信息，用惩罚修正约束偏差”。

### 12.2.3 增广拉格朗日函数的构造

ALM 的核心是增广拉格朗日函数，其构造逻辑是：在拉格朗日函数基础上，加入与外点法一致的二次惩罚项。

#### 1. 标准增广拉格朗日函数

$$\mathcal{L}_A(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mu) = f(\mathbf{x}) + \sum_{i=1}^m \left[ \lambda_i g_i(\mathbf{x}) + \frac{\mu}{2} \max \left( 0, g_i(\mathbf{x}) + \frac{\lambda_i}{\mu} \right)^2 \right] + \sum_{j=1}^p \left[ \nu_j h_j(\mathbf{x}) + \frac{\mu}{2} h_j(\mathbf{x})^2 \right]$$

#### 2. 简化形式与物理意义

通过代数变形，可将不等式约束的惩罚项简化为更直观的形式（利用  $\max(0, a)^2 = \max(0, a^2 + 2a \cdot 0)^2$ ，但核心是保持“违反约束时惩罚生效”）：

$$\max \left( 0, g_i(\mathbf{x}) + \frac{\lambda_i}{\mu} \right)^2 = \max(0, g_i(\mathbf{x}))^2 + 2 \frac{\lambda_i}{\mu} g_i(\mathbf{x}) + \frac{\lambda_i^2}{\mu^2} \quad (\text{仅当 } g_i(\mathbf{x}) > 0)$$

代入增广拉格朗日函数后，常数项  $\sum_{i=1}^m \frac{\lambda_i^2}{2\mu}$  不影响无约束优化的最优解（对  $\mathbf{x}$  求导时消失），因此可简化为：

$$\mathcal{L}_A(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mu) = f(\mathbf{x}) + \mu \left[ \sum_{i=1}^m \frac{1}{2} \max(0, g_i(\mathbf{x}))^2 + \sum_{j=1}^p \frac{1}{2} h_j(\mathbf{x})^2 \right] + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x})$$

核心物理意义：

- 当迭代点  $\mathbf{x}$  违反约束时，惩罚项  $\mu \cdot \Phi(\mathbf{x})$  生效，强制迭代点向可行域靠近；
- 拉格朗日乘子  $\boldsymbol{\lambda}, \boldsymbol{\nu}$  随迭代更新，逐步逼近最优乘子  $\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ ，其作用是“引导”惩罚项的方向。

### 12.2.4 乘子更新规则的推导（基于 KKT 条件）

ALM 的关键是乘子迭代更新，其更新规则源于无约束子问题的最优性条件和 KKT 条件的一致性。

### 1. 无约束子问题的最优性条件

对于固定的  $\lambda_k, \nu_k, \mu_k$ , 求解无约束子问题:

$$\mathbf{x}_{k+1} = \arg \min_{\mathbf{x}} \mathcal{L}_A(\mathbf{x}, \lambda_k, \nu_k, \mu_k)$$

其最优性条件为梯度为零:

$$\nabla_{\mathbf{x}} \mathcal{L}_A(\mathbf{x}_{k+1}, \lambda_k, \nu_k, \mu_k) = \mathbf{0}$$

### 2. 对不等式约束乘子 $\lambda_i$ 的更新

展开梯度条件中关于  $g_i(\mathbf{x})$  的项:

$$\nabla f(\mathbf{x}_{k+1}) + \sum_{i=1}^m [\lambda_{k,i} + \mu_k \cdot \max(0, g_i(\mathbf{x}_{k+1}))] \nabla g_i(\mathbf{x}_{k+1}) + \sum_{j=1}^p [\nu_{k,j} + \mu_k h_j(\mathbf{x}_{k+1})] \nabla h_j(\mathbf{x}_{k+1}) = \mathbf{0}$$

根据 KKT 条件, 原问题最优解  $\mathbf{x}^*$  满足:

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}$$

且互补松弛条件  $\lambda_i^* g_i(\mathbf{x}^*) = 0$  (可行点  $g_i(\mathbf{x}^*) \leq 0$ , 故  $\lambda_i^* > 0 \implies g_i(\mathbf{x}^*) = 0$ )。

对比两式, 为使  $\lambda_k$  逐步逼近  $\lambda_i^*$ , 自然得到乘子更新规则:

$$\lambda_{k+1,i} = \max(0, \lambda_{k,i} + \mu_k \cdot g_i(\mathbf{x}_{k+1}))$$

- 当  $\mathbf{x}_{k+1}$  满足约束 ( $g_i(\mathbf{x}_{k+1}) \leq 0$ ) 且  $\lambda_{k,i} + \mu_k g_i(\mathbf{x}_{k+1}) \geq 0$  时,  $\lambda_{k+1,i} = \lambda_{k,i} + \mu_k g_i(\mathbf{x}_{k+1})$ , 逐步逼近最优乘子;
- 当  $\lambda_{k,i} + \mu_k g_i(\mathbf{x}_{k+1}) < 0$  时,  $\lambda_{k+1,i} = 0$ , 满足 KKT 条件中  $\lambda_i \geq 0$  的要求。

### 3. 对等式约束乘子 $\nu_j$ 的更新

同理, 展开梯度条件中关于  $h_j(\mathbf{x})$  的项, 由于等式约束无互补松弛的非负性要求, 直接得到:

$$\nu_{k+1,j} = \nu_{k,j} + \mu_k \cdot h_j(\mathbf{x}_{k+1})$$

## 12.2.5 ALM 算法流程

**算法 12.2** (ALM 算法流程). 输入:

- 原问题目标函数  $f(\mathbf{x})$ 、约束  $g_i(\mathbf{x})$ 、 $h_j(\mathbf{x})$ ;
- 初始参数: 初始点  $\mathbf{x}_0$ ; 初始乘子  $\lambda_0 = \mathbf{0}$ 、 $\nu_0 = \mathbf{0}$ ; 初始惩罚参数  $\mu_1 > 0$  (无需过大, 通常取 1 或 10); 惩罚参数增长因子  $\beta > 1$ ; 收敛精度  $\epsilon > 0$ 。

迭代步骤:

1. 初始化: 令迭代次数  $k = 1$ ;
2. 构造增广拉格朗日函数: 针对当前  $\lambda_{k-1}, \nu_{k-1}, \mu_k$ , 构造  $\mathcal{L}_A(\mathbf{x}, \lambda_{k-1}, \nu_{k-1}, \mu_k)$ ;
3. 求解无约束子问题: 以  $\mathbf{x}_{k-1}$  为初始点, 求解  $\min_{\mathbf{x}} \mathcal{L}_A$ , 得到最优解  $\mathbf{x}_k$ ;

## 4. 更新拉格朗日乘子:

- 不等式约束乘子:  $\lambda_{k,i} = \max(0, \lambda_{k-1,i} + \mu_k \cdot g_i(\mathbf{x}_k))$ ;
- 等式约束乘子:  $\nu_{k,j} = \nu_{k-1,j} + \mu_k \cdot h_j(\mathbf{x}_k)$ ;

5. 收敛判断: 若满足以下所有收敛条件, 停止迭代, 输出  $\mathbf{x}_k \approx \mathbf{x}^*$ :

- 约束违反度量足够小:  $\Phi(\mathbf{x}_k) < \epsilon$ ;
- 乘子变化足够小:  $\|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k-1}\| < \epsilon$  且  $\|\boldsymbol{\nu}_k - \boldsymbol{\nu}_{k-1}\| < \epsilon$ ;
- 迭代点变化足够小:  $\|\mathbf{x}_k - \mathbf{x}_{k-1}\| < \epsilon$ ;

6. 更新惩罚参数: 若未收敛, 令  $\mu_{k+1} = \beta \cdot \mu_k$  (或保持  $\mu_k$  不变),  $k = k + 1$ , 返回步骤 2。

输出: 原约束问题的近似最优解  $\mathbf{x}_k$  及对应的最优乘子  $\boldsymbol{\lambda}_k, \boldsymbol{\nu}_k$ 。

## 12.2.6 ALM 与外点法的核心差异 (严谨对比)

对比维度	外点法 (Penalty Function Method)	增广拉格朗日法 (ALM)
核心函数	惩罚函数 $P(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu\Phi(\mathbf{x})$	增广拉格朗日函数 $\mathcal{L}_A(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mu)$
关键参数	仅惩罚参数 $\mu$ (需 $\mu \rightarrow +\infty$ )	惩罚参数 $\mu$ (可固定或适度增大) + 拉格朗日乘子 $\boldsymbol{\lambda}, \boldsymbol{\nu}$
数值稳定性	差 ( $\mu$ 过大导致 Hessian 条件数恶化)	好 (乘子引导惩罚, $\mu$ 无需无穷大)
收敛速度	慢 (依赖 $\mu$ 逐步增大, 迭代后期收敛迟缓)	快 (乘子自适应调整约束权重, 迭代前期即可快速靠近最优解)
最优性条件满足	仅满足原始可行性 ( $\mathbf{x}_k \rightarrow \Omega$ ), 对偶信息无保障	同时逼近原始最优和对偶最优 ( $\boldsymbol{\lambda}_k \rightarrow \boldsymbol{\lambda}^*, \boldsymbol{\nu}_k \rightarrow \boldsymbol{\nu}^*$ )

## 12.2.7 收敛性核心结论

假设原问题满足:  $f(\mathbf{x})$  凸、 $g_i(\mathbf{x})$  凸、 $h_j(\mathbf{x})$  仿射, 且 Slater 条件成立 (存在严格可行点), 则 ALM 的迭代序列  $\{\mathbf{x}_k, \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k\}$  满足:

1.  $\mathbf{x}_k$  强收敛到原问题的最优解  $\mathbf{x}^*$ ;
2.  $\boldsymbol{\lambda}_k$  强收敛到最优拉格朗日乘子  $\boldsymbol{\lambda}^*$ ,  $\boldsymbol{\nu}_k$  强收敛到  $\boldsymbol{\nu}^*$ ;
3. 惩罚参数  $\mu$  可固定为某个常数 (无需增大), 仍能保证收敛 (这是 ALM 相对于外点法的本质优势)。

若原问题非凸, 在适当的约束品性下, ALM 仍能保证迭代序列的聚点是原问题的 KKT 点。

## 12.2.8 示例: ALM 应用

$$\begin{cases} \min_{x_1, x_2} & f(x) = x_1^2 + 2x_2^2 - 2x_1 - 2x_2 \\ \text{s.t.} & h(x) = x_1 + x_2 - 1 = 0 \quad (\text{等式约束}) \\ & g(x) = x_1 - 0.6 \leq 0 \quad (\text{不等式约束}) \end{cases}$$

- 初始点:  $\mathbf{x}_0 = (0, 0)$
- 初始乘子:  $\lambda_0 = 0, \nu_0 = 0$
- 初始惩罚参数:  $\mu_1 = 1$
- 惩罚参数增长因子:  $\beta = 2$

### 第一轮迭代 ( $k = 1$ )

步骤 1: 构造增广拉格朗日函数 代入  $\lambda_0 = 0, \nu_0 = 0, \mu_1 = 1$ , 简化得:

$$\mathcal{L}_A(\mathbf{x}) = x_1^2 + 2x_2^2 - 2x_1 - 2x_2 + \frac{1}{2} \max(0, x_1 - 0.6)^2 + \frac{1}{2}(x_1 + x_2 - 1)^2$$

步骤 2: 求解无约束子问题 对  $x_1, x_2$  求偏导并令其为 0。若  $x_1 > 0.6$ , 偏导为:

$$\begin{cases} \frac{\partial \mathcal{L}_A}{\partial x_1} = 4x_1 + x_2 - 3.6 = 0 \\ \frac{\partial \mathcal{L}_A}{\partial x_2} = x_1 + 5x_2 - 3 = 0 \end{cases}$$

解方程组得:

$$x_1 \approx 0.7895, x_2 \approx 0.4421 \quad (\text{满足 } x_1 > 0.6)$$

即第一轮迭代点:  $\mathbf{x}_1 = (0.7895, 0.4421)$ 。

### 步骤 3: 更新乘子

- 不等式约束乘子:  $\lambda_1 = \max(0, 0 + 1 \cdot (0.7895 - 0.6)) = 0.1895$
- 等式约束乘子:  $\nu_1 = 0 + 1 \cdot (0.7895 + 0.4421 - 1) = 0.2316$

### 第二轮迭代 ( $k = 2$ )

步骤 1: 更新惩罚参数 & 构造增广拉格朗日函数  $\mu_2 = 2 \times 1 = 2$ 。代入  $\lambda_1 = 0.1895, \nu_1 = 0.2316, \mu_2 = 2$ , 增广拉格朗日函数为:

$$\mathcal{L}_A(\mathbf{x}) = x_1^2 + 2x_2^2 - 2x_1 - 2x_2 + 0.1895(x_1 - 0.6) + \max(0, x_1 - 0.50525)^2 + 0.2316(x_1 + x_2 - 1) + (x_1 + x_2 - 1)^2$$

步骤 2: 求解无约束子问题 对  $x_1, x_2$  求偏导并令其为 0, 解得:

$$x_1 \approx 0.6251, x_2 \approx 0.4197 \quad (\text{接近 } x_1 = 0.6 \text{ 的约束边界})$$

即第二轮迭代点:  $\mathbf{x}_2 = (0.6251, 0.4197)$ 。

### 步骤 3: 更新乘子

- 不等式约束乘子:  $\lambda_2 = \max(0, 0.1895 + 2 \cdot (0.6251 - 0.6)) = 0.2397$
- 等式约束乘子:  $\nu_2 = 0.2316 + 2 \cdot (0.6251 + 0.4197 - 1) = 0.3212$

## 迭代结果总结

迭代轮次	迭代点 $\mathbf{x}_k$	不等式乘子 $\lambda_k$	等式乘子 $\nu_k$	惩罚参数 $\mu_k$
初始	(0, 0)	0	0	1
1	(0.7895, 0.4421)	0.1895	0.2316	1
2	(0.6251, 0.4197)	0.2397	0.3212	2

可以看到： $\mathbf{x}_2$  已靠近不等式约束边界  $x_1 = 0.6$ ，等式约束偏差  $h(\mathbf{x}_2) \approx 0.0448$  也显著减小，说明迭代在向可行域收敛。

## 12.3 约束问题的解法之 ADMM

## 12.3.1 第一步：先锁定 ALM 的“可分问题特例”（ADMM 的适用场景）

我们从 ALM 能处理的一般约束问题中，挑一个目标可分 + 线性等式约束的特例（这正是 ADMM 专门解决的问题）：

$$\begin{cases} \min_{\mathbf{x}, \mathbf{z}} & \underbrace{f(\mathbf{x})}_{\text{仅依赖 } \mathbf{x}} + \underbrace{g(\mathbf{z})}_{\text{仅依赖 } \mathbf{z}} & (\text{目标可分}) \\ \text{s.t.} & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c} & (\text{线性等式约束}) \end{cases}$$

（注：ADMM 只处理这种“目标拆成两个独立部分 + 线性等式约束”的问题，这是它和 ALM 的第一个区别——ALM 处理更一般的约束）

## 12.3.2 第二步：写出这个特例的 ALM 增广拉格朗日函数

$$\mathcal{L}_A^{\text{ALM}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\nu}, \rho) = f(\mathbf{x}) + g(\mathbf{z}) + \boldsymbol{\nu}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\rho}{2} \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|_2^2$$

其中：

- $\boldsymbol{\nu}$  是等式约束的拉格朗日乘子；
- $\rho > 0$  是惩罚参数（和 ALM 的  $\mu$  是同一个东西）。

## 12.3.3 第三步：ALM 对这个问题的迭代步骤

ALM 处理这个问题时，迭代是“同时最小化  $\mathbf{x}, \mathbf{z}$ ，再更新乘子  $\boldsymbol{\nu}$ ”：

1. 最小化增广拉格朗日： $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) = \arg \min_{\mathbf{x}, \mathbf{z}} \mathcal{L}_A^{\text{ALM}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\nu}^k, \rho)$
2. 更新乘子： $\boldsymbol{\nu}^{k+1} = \boldsymbol{\nu}^k + \rho(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{z}^{k+1} - \mathbf{c})$

## 12.3.4 第四步：关键改进——利用“目标可分”拆分最小化步骤

问题来了：同时最小化  $\mathbf{x}, \mathbf{z}$  可能很难（比如  $\mathbf{x}, \mathbf{z}$  维度很大时）。

但我们的目标是“可分的”： $f(\mathbf{x})$  只和  $\mathbf{x}$  有关， $g(\mathbf{z})$  只和  $\mathbf{z}$  有关。所以可以交替最小化  $\mathbf{x}$  和  $\mathbf{z}$ （先固定  $\mathbf{z}$  求  $\mathbf{x}$ ，再固定  $\mathbf{x}$  求  $\mathbf{z}$ ）——这就是 ADMM 的核心！

拆分 1: 固定  $z^k$ , 先求  $x^{k+1}$  (ADMM 的  $x$ -步)

把  $z = z^k$  代入增广拉格朗日, 此时只有  $x$  是变量:

$$x^{k+1} = \arg \min_x \left[ f(x) + \nu^{kT}(Ax + Bz^k - c) + \frac{\rho}{2} \|Ax + Bz^k - c\|_2^2 \right]$$

拆分 2: 固定  $x^{k+1}$ , 再求  $z^{k+1}$  (ADMM 的  $z$ -步)

把  $x = x^{k+1}$  代入增广拉格朗日, 此时只有  $z$  是变量:

$$z^{k+1} = \arg \min_z \left[ g(z) + \nu^{kT}(Ax^{k+1} + Bz - c) + \frac{\rho}{2} \|Ax^{k+1} + Bz - c\|_2^2 \right]$$

### 12.3.5 第五步: 简化符号——引入 ADMM 的“对偶残差 $u$ ”

为了让式子更简洁, ADMM 把 ALM 的乘子  $\nu$  缩放一下: 令  $u = \frac{\nu}{\rho}$  ( $u$  就是 ADMM 里的“对偶变量”).

把  $\nu = \rho u$  代入上面的式子, 就能消掉  $\nu$  的线性项, 最终得到 ADMM 的标准步骤:

### 12.3.6 最终: 从 ALM 拆分得到的 ADMM 迭代 (一一对应)

ALM 的步骤 (可分特例)	对应 ADMM 的步骤 (符号简化后)
1. 固定 $z^k$ , 求 $x^{k+1}$	1. $x$ -步: $x^{k+1} = \arg \min_x \left[ f(x) + \frac{\rho}{2} \ Ax + Bz^k - c + u^k\ _2^2 \right]$
2. 固定 $x^{k+1}$ , 求 $z^{k+1}$	2. $z$ -步: $z^{k+1} = \arg \min_z \left[ g(z) + \frac{\rho}{2} \ Ax^{k+1} + Bz - c + u^k\ _2^2 \right]$
3. 更新乘子 $\nu^{k+1}$	3. $u$ -步 (对应乘子更新): $u^{k+1} = u^k + Ax^{k+1} + Bz^{k+1} - c$

ADMM 就是“可分目标 + 线性等式约束”问题下的 ALM——它把 ALM “同时最小化所有变量”的步骤, 拆成了“交替最小化可分变量块”的简单步骤, 同时缩放乘子符号让式子更简洁。

### 12.3.7 示例 (与前述问题关联的改写)

要对这个问题用 ADMM 迭代, 首先得把它转化为 ADMM 的标准形式 (可分目标 + 线性等式约束)。

步骤 1: 转化为 ADMM 标准形式

原问题:

$$\begin{cases} \min_{x_1, x_2} & f(x_1, x_2) = x_1^2 + 2x_2^2 - 2x_1 - 2x_2 \\ \text{s.t.} & x_1 + x_2 = 1 \quad (\text{等式约束}) \\ & x_1 \leq 0.6 \quad (\text{不等式约束}) \end{cases}$$

拆分变量 + 转化约束:

- 令变量块 1:  $x = x_2$  (仅依赖  $x_2$ ), 目标项为  $f(x) = 2x^2 - 2x$ ;
- 令变量块 2:  $z = x_1$  (仅依赖  $x_1$ ), 目标项为  $g(z) = z^2 - 2z + I(z \leq 0.6)$  ( $I(\cdot)$  是指示函数);
- 约束转化为线性等式:  $z + x = 1$  (对应 ADMM 的标准约束  $Az + Bx = c$ , 这里  $A = B = 1, c = 1$ )。

**步骤 2: ADMM 的增广拉格朗日函数**

$$\mathcal{L}_A(x, z, u, \rho) = f(x) + g(z) + \frac{\rho}{2} (z + x - 1 + u)^2 - \frac{\rho}{2} u^2$$

其中  $u$  是对偶变量,  $\rho = 1$ 。

**步骤 3: ADMM 迭代流程 (3 步循环)**

初始化:  $x^0 = 0, z^0 = 0, u^0 = 0$ 。

**第 1 轮迭代 ( $k = 0 \rightarrow k = 1$ )**

1. **x-步**: 固定  $z^0 = 0, u^0 = 0$ , 最小化  $\mathcal{L}_A$  关于  $x$ :

$$\min_x 2x^2 - 2x + \frac{1}{2} (0 + x - 1 + 0)^2$$

求导并令其为 0:  $4x - 2 + (x - 1) = 5x - 3 = 0 \implies x^1 = 0.6$ 。

2. **z-步**: 固定  $x^1 = 0.6, u^0 = 0$ , 最小化  $\mathcal{L}_A$  关于  $z$ :

$$\min_z \underbrace{z^2 - 2z + I(z \leq 0.6)}_{\text{含约束}} + \frac{1}{2} (z + 0.6 - 1 + 0)^2$$

无约束解:  $2z - 2 + (z - 0.4) = 3z - 2.4 = 0 \implies z = 0.8$  (违反约束); 约束下最优解:  $z^1 = 0.6$  (约束边界)。

3. **u-步**: 更新对偶变量:

$$u^1 = u^0 + z^1 + x^1 - 1 = 0 + 0.6 + 0.6 - 1 = 0.2$$

**第 2 轮迭代 ( $k = 1 \rightarrow k = 2$ )**

1. **x-步**: 固定  $z^1 = 0.6, u^1 = 0.2$ :

$$\min_x 2x^2 - 2x + \frac{1}{2} (0.6 + x - 1 + 0.2)^2$$

求导得:  $4x - 2 + (x - 0.2) = 5x - 2.2 = 0 \implies x^2 = 0.44$ 。

2. **z-步**: 固定  $x^2 = 0.44, u^1 = 0.2$ :

$$\min_z z^2 - 2z + I(z \leq 0.6) + \frac{1}{2} (z + 0.44 - 1 + 0.2)^2$$

无约束解仍违反  $z \leq 0.6$ , 故  $z^2 = 0.6$ 。

3. **u-步**:

$$u^2 = 0.2 + 0.6 + 0.44 - 1 = 0.24$$

第 3 轮迭代 ( $k = 2 \rightarrow k = 3$ )

1. **x-步**: 固定  $z^2 = 0.6, u^2 = 0.24$ :

$$\min_x 2x^2 - 2x + \frac{1}{2}(0.6 + x - 1 + 0.24)^2$$

求导得:  $5x - 2.16 = 0 \implies x^3 = 0.432$ 。

2. **z-步**: 约束下仍取  $z^3 = 0.6$ 。

3. **u-步**:

$$u^3 = 0.24 + 0.6 + 0.432 - 1 = 0.272$$

## 迭代趋势

迭代轮次	$x_k = x_2$	$z_k = x_1$	对偶变量 $u_k$	约束满足度 $z_k + x_k - 1$
0	0	0	0	-1
1	0.6	0.6	0.2	0.2
2	0.44	0.6	0.24	0.04
3	0.432	0.6	0.272	0.032

可以看到:  $z_k = x_1$  稳定在约束边界 0.6,  $x_k = x_2$  逐渐靠近最优值 0.4 (原问题最优解为  $x_1 = 0.6, x_2 = 0.4$ )。