

# 凸优化笔记

Convex Analysis and Optimization Notes

讲学笔记



RESEARCH NOTE, AT UNIVERSITY OF HOUSTON

GITHUB.COM/LAURETHTEX/CLUSTERING

This is the personal note of learning convex optimization, written by *Yuchen Jin*. Anyone who read this book could share it freely, but any activities of recompilation, modification or reproduction are not permitted, according to the authority of the editor.

*First release, February 9, 2018*



## 目 录

<b>1</b>	<b>引论</b>	<b>5</b>
1.1	<b>探讨矩阵分析</b>	5
1.2	<b>我们要讨论的问题</b>	6
1.3	<b>两个经典问题</b>	6
1.3.1	最小平方问题	6
1.3.2	线性规划问题	8
1.3.3	凸问题浅述	9
<b>2</b>	<b>对偶</b>	<b>11</b>
2.1	<b>探讨矩阵分析</b>	11
2.2	<b>拉格朗日对偶问题</b>	12
2.3	<b>稀疏编码</b>	15
2.3.1	解对偶问题训练字典	16
2.4	<b>对偶条件</b>	<b>20</b>
2.4.1	对偶问题的凹性质	20
2.4.2	弱对偶与强对偶	22
2.4.3	Slater 条件	26





# 1. 引论

**凸优化** (*Convex Optimization*) 是求最优解的一类经典问题, 本章将初探一些简单而经典的凸问题的解法。为了能正确求解出符合优化目标的解, 需要将过往的实数分析推广到矩阵分析。在本书中, 为了体现没有矩阵分析背景的新手上手的思路, 每当需要新的分析知识作为基石的时候, 或将一些基础的推导放在最前面, 或将某些问题的推导作为独立的引理, 随着提出进行求解。

本章主要探讨的问题包括最小平方问题(见节 1.3.1)和线性规划问题(见节 1.3.2)。进一步地, 简要介绍了几种简单的凸问题(见节 1.3.3)。

## 1.1 探讨矩阵分析

矩阵运算推广到数学分析, 需要有以下约定:

- (1) 向量可以看作行/列为 1 的特殊的矩阵;
- (2) 向量一律用形如  $\mathbf{x}$  或  $\vec{x}$  的形式书写, 矩阵一律用形如  $\mathbf{X}$  的形式书写, 用在向量的推导式不可直接推广并应用在矩阵的情形;
- (3) 函数  $f$  对矩阵  $\mathbf{X}$  求导, 所得为矩阵  $\mathbf{F}$ , 它的大小与  $\mathbf{X}$  相同, 且满足  $F_{ij} = \frac{df}{dX_{ij}}$ ;
- (4) 向量  $\mathbf{a}$  对向量  $\mathbf{b}$  求导, 所得为矩阵  $\mathbf{X}$ , 它的大小为  $(A, B)$ , 且满足  $X_{ij} = \frac{da_i}{db_j}$ , 这样的矩阵称为雅可比矩阵 (*Jacobi Matrix*);
- (5) 雅可比矩阵有可能会给出反直觉的结论, 例如  $\frac{da}{db}$  中, 若  $\mathbf{a}$  退化为维数为 1 的实数, 所得的向量形状并非与  $\mathbf{b}$  相同, 而是  $\mathbf{b}$  的转置;
- (6) 向量对矩阵求导, 矩阵对向量求导和矩阵对矩阵求导, 都是存在的, 然而至少到本章为止, 不需要讨论这些复杂的情况。

下述的一些练习, 部分将出现在后文的证明步骤中

■ **例 1.1** 考虑两个向量  $\mathbf{x}, \mathbf{a}$  的内积  $\mathbf{a}^T \mathbf{x}$ , 这是一个常数, 因此, 对其进行向量求导, 有

$$\left( \frac{d\mathbf{a}^T \mathbf{x}}{d\mathbf{x}} \right)_k = \frac{d}{dx_k} \left( \sum_{i=1}^n a_i x_i \right) = a_k. \quad (1.1)$$

同理,  $\frac{d\mathbf{x}^T \mathbf{a}}{d\mathbf{x}}$  得到的结果与上述相同;这两种情况下  $\mathbf{x}$ ,  $\mathbf{a}$  均保持同形,故有

$$\frac{d\mathbf{a}^T \mathbf{x}}{d\mathbf{x}} = \frac{d\mathbf{x}^T \mathbf{a}}{d\mathbf{x}} = \mathbf{a}. \quad (1.2)$$

反之,如果  $\exists \mathbf{c} = \mathbf{a}^T$ ,那么代换到(1.2)将求得  $\mathbf{c}^T$ 。 ■

■ **例 1.2** 试求  $\frac{d\mathbf{x}^T \mathbf{A}\mathbf{x}}{d\mathbf{x}}$  的值。

$$\begin{aligned} \because \left( \frac{d}{d\mathbf{x}} (\mathbf{x}^T \mathbf{A}\mathbf{x}) \right)_k &= \frac{d}{dx_k} \left( \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j \right) \\ &= \sum_{j=1}^n A_{kj} x_j + \sum_{i=1}^n A_{ik} x_i. \\ \therefore \frac{d}{d\mathbf{x}} (\mathbf{x}^T \mathbf{A}\mathbf{x}) &= (\mathbf{A} + \mathbf{A}^T)\mathbf{x}. \end{aligned} \quad (1.3)$$

## 1.2 我们要讨论的问题

考虑到如下形式的问题(注意参数和输出均有可能是矩阵)。

**问题 1.1 – 普适优化问题.**

$$\min f_0(x), \quad (1.4-1)$$

$$s.t. f_i(x) \leq b_i, i = 1, \dots, m. \quad (1.4-2)$$

如果对任意的  $i = 0, 1, \dots, m$ ,均有

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y). \quad (1.5)$$

这样的问题称为**凸 (Convex) 问题**,如果(1.5)中的不等号方向反向,这样的问题就称为**凹 (Concave) 问题**。凸/凹问题本质上都可以用相同的方式解出,特别地,如果(1.5)中的不等号退化为等号,这样的问题就称为**线性问题**。

如果进入到离散形式的问题中,我们可以重新看待问题 1.1。其中,  $x$  可以被看作是投资,  $f_0$  可以被看作是耗费 (Cost) 函数,  $f_i$  则是预算 (budget) 限制条件,这样的问题就可以被描述成一个投资与耗费最小化的组合优化 (portfolio optimization) 问题。其他的例子在设备设计、数据拟合、自动决策、嵌入式设备等领域中应用广泛,不一而足。

与**线性规划 (Linear Programming)**、**最小平方 (Least-square)**所一致的是, **凸优化**正是在众多难解、不可解的优化问题中,可以高效求解的例外。

## 1.3 两个经典问题

### 1.3.1 最小平方问题

考虑存在列向量  $x$ ,一个最小平方问题可以描述为:

**问题 1.2 – 最简的最小平方问题.**

$$\min_{\mathbf{x}} f_0(\mathbf{x}) = \min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2. \quad (1.6)$$

如此, 只需对  $f_0$  求取梯度, 即可得到所需的最优解  $x^*$ , 有如下推导

$$\begin{aligned} \frac{df_0}{d\mathbf{x}} &= \frac{d}{d\mathbf{x}} ((\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b})) \\ &= \frac{d}{d\mathbf{x}} ((\mathbf{x}^T \mathbf{A}^T - \mathbf{b}^T)(\mathbf{Ax} - \mathbf{b})) \\ &= \frac{d}{d\mathbf{x}} (\mathbf{x}^T \mathbf{A}^T \mathbf{Ax} - 2\mathbf{b}^T \mathbf{Ax} + \mathbf{b}^T \mathbf{b}) \\ &= 2\mathbf{A}^T \mathbf{Ax} - 2(\mathbf{b}^T \mathbf{A})^T = 2\mathbf{A}^T \mathbf{Ax} - 2\mathbf{A}^T \mathbf{b} = 0. \end{aligned} \quad (1.7)$$

于是,

$$\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}. \quad (1.8)$$

为所求的最优解。注意(1.7)中, 部分推导的详细过程属于矩阵的运算定式, 可在(1.2),(1.3)查看。

**变式: 带权最小平方问题**

现在重新考虑问题 1.2, 考虑一个权重向量  $\mathbf{w}$  与  $\mathbf{b}$  的维数相等, 那么由于方程系数  $\mathbf{A}$  可以拆解成与  $\mathbf{w}$ ,  $\mathbf{b}$  维数相同数目的行向量  $\{\mathbf{a}_i^T\}$ , 亦即:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_n^T \end{bmatrix}. \quad (1.9)$$

如此, (1.6)可以改写成

$$\min_{\mathbf{x}} f_0(\mathbf{x}) = \sum_{i=1}^n \mathbf{a}_i^T \mathbf{x} - b_i. \quad (1.10)$$

如果在求和号内加上权值, 则原问题改写成

$$\min_{\mathbf{x}} f_0(\mathbf{x}) = \sum_{i=1}^n w_i (\mathbf{a}_i^T \mathbf{x} - b_i). \quad (1.11)$$

显然这样的形式不再能写成  $l_2$ -norm, 但是若设有对角矩阵  $\mathbf{W} = \text{diag}(\mathbf{w})$ , 则仍可将(1.11)改写成矩阵表示的形式。

**问题 1.3 – 带权的最小平方问题.**

$$\min_{\mathbf{x}} f_0(\mathbf{x}) = (\mathbf{Ax} - \mathbf{b})^T \mathbf{W} (\mathbf{Ax} - \mathbf{b}). \quad (1.12)$$

显然  $\mathbf{W}^T = \mathbf{W}$ , 于是沿用与(1.7)相同的样式, 该问题的求解过程如下:

$$\begin{aligned}\frac{df_0}{d\mathbf{x}} &= \frac{d}{d\mathbf{x}} (\mathbf{x}^T \mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{x} - 2\mathbf{b}^T \mathbf{W} \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{W} \mathbf{b}) \\ &= 2\mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{x} - 2\mathbf{A}^T \mathbf{W} \mathbf{b} = 0.\end{aligned}\quad (1.13)$$

于是,

$$\mathbf{x} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{b}. \quad (1.14)$$

#### 变式: 带正则最小平方问题

有时候, 为了保证求得的  $\mathbf{x}$  足够小, 还需要加上  $\mathbf{x}$  的正则项( $l_2$ -norm)。在该情况下, 问题描述为:

#### 问题 1.4 – 带正则项的最小平方问题.

$$\min_{\mathbf{x}} f_0(\mathbf{x}) = \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \rho \|\mathbf{x}\|_2^2. \quad (1.15)$$

显然, 求解问题 1.4 只需要在梯度后加上  $\rho \frac{d\mathbf{x}^T \mathbf{x}}{d\mathbf{x}}$  即可。显然, 代入(1.3), 该值为  $2\rho\mathbf{x}$ 。于是, 该问题易解, 为

$$\mathbf{x} = (\rho\mathbf{I} + \mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}. \quad (1.16)$$

这表明,  $\rho$  越大,  $\mathbf{x}$  就越倾向于被限定在较小的范围内, 这与最小化  $l_2$ -norm 的目标是一致的。

### 1.3.2 线性规划问题

一个线性规划问题指的是优化目标和限制条件全部为线性方程/不等式的问题。一般地, 该问题可以被描述成如下形式

#### 问题 1.5 – 较一般的线性规划问题.

$$\min_{\mathbf{x}} \mathbf{c}^T \mathbf{x}, \quad (1.17-1)$$

$$s.t. \mathbf{a}_i^T \mathbf{x} \leq b_i, i = 1, \dots, m. \quad (1.17-2)$$

故而, 容易看出, 该问题与问题 1.2 最大的区别在于, 它具有多个限制条件。这意味着, 凭借梯度下降得到的解极有可能不在可行域(Feasible Domain)内(事实上, 这个问题中也不可能在没有限制条件的情况下解出“最优解”)。换言之, 凭入门水平的知识, 暂时还无法解决这个问题, 并且该问题也没有解析形式的解。但是, 存在高效率、相当成熟的算法, 能得到这一类问题的数值解。因此, 在本节我们只讨论如何将一个已知的问题设法转换成一个线性的问题。

#### 问题 1.6 – 切比雪夫近似问题.

$$\min_{\mathbf{x}} \max_{i=1,2,\dots,k} |\mathbf{a}_i^T \mathbf{x} - b_i|. \quad (1.18)$$

该问题需要最小化的是一个最大值函数, 这意味着, 它是一个不可微的函数, 从而难以应用梯度求解。然而, 它可以被近似成线性问题,

问题 1.7 – 切比雪夫近似问题(线性形式).

$$\min_{\mathbf{x}} \xi, \quad (1.19-1)$$

$$s.t. \mathbf{a}_i^T \mathbf{x} - \xi \leq b_i, i = 1, \dots, m, \quad (1.19-2)$$

$$-\mathbf{a}_i^T \mathbf{x} - \xi \leq -b_i, i = 1, \dots, m. \quad (1.19-3)$$

与问题 1.5 形式相同, 它可以求解。这种近似的含义是, 存在一个绝对距离  $\xi$ , 使得所有  $|\mathbf{a}_i^T \mathbf{x} - b_i|$  小于这个距离, 并确定满足这一条件的最小距离。实质上, 它相当于将  $|\mathbf{a}_i^T \mathbf{x} - b_i|$  代换成了  $\xi$ , “最大值” 隐含在了所有约束条件均满足不等式这一设定中。

### 1.3.3 凸问题浅述

一个**凸问题**, 已经存在较为成熟的方法去求解它了。尽管这可能不及最小平方, 抑或是**线性规划**的求解一般成熟, 但也仅此而已。我们有一套非常准确的方法得到最优解, 而不需要担心陷入局部最优、或者是不能弥合的误差之中。

因此, **凸优化**的关键步骤不再是求解的细节, 而是如何将一个已知的问题建模成**凸问题**, 后续的章节会对此进行理论和应用的探讨。

那么, 这是否是说, **凸优化**止步于那些能转换成**凸问题**的问题上呢? 并非如此。我们知道在人有穷的智慧和有穷的计算能力之下, 很难、甚至不可解一些需要巨额计算资源的问题, 这些问题, 往往都在**非线性优化** (*Non-linear optimization*) 的框架下。但是, 这并不妨碍我们分情况讨论我们的解, 或者做一些折中的方案。

第一种情况是局部最优问题。局部最优不再将视野集中在真正的最优解上, 它的解与全局最优解的区别就和极小值与最小值的区别一样。局部最优能非常迅速地给出解。

求解这样的问题时, 选定合适的初始值至关重要, 否则就容易陷入次优解的泥潭中。考虑图 1.1 的问题形式, 在蓝色的曲线需要找到最小值, 如果选取一个点梯度下降, 有可能有两个收敛结果; 红色的曲线代表作二次曲线拟合的结果, 这一拟合的过程完全是一个**凸问题(最小平方)**, 由于二次曲线本身是凸/凹的, 在它上求最小值是解析的。这一解析解的位置, 由紫色的直线标出。由于曲线的不对称(左侧的解更优), 这一值更接近左侧的最优解, 选定它为初值梯度下降, 容易得到左侧最优解的结果。

第二种情况是全局最优问题, 全局最优方法的显著代价是, 牺牲了效率, 有时达到指数级复杂度的水平。但是在特定的应用下, 例如, 对于参数不多的**最坏预期分析** (*Worst-case analysis or verification*), 求解全局最优是可以理解和接受的, 因为局部最优解只能确定系统不安全, 却不能保证系统必定是安全的。

**凸优化**在为求解提供初值、为启发式算法求解最优参数分布、为一个复杂的问题确定最优解的边界等问题上, 均有应用。后续将详细探讨这些应用, 以及它们的理论背景。

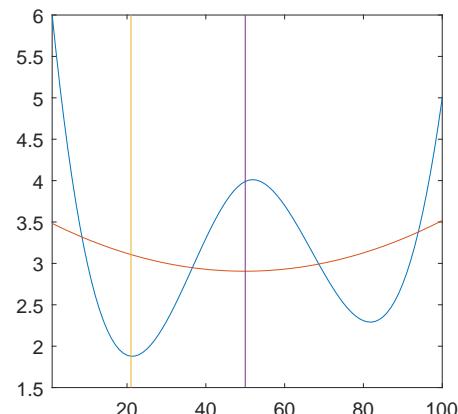


图 1.1: 求解凸问题对求解非线性问题的帮助。





## 2. 对偶

对偶问题, 是解具有标准形式的限制条件的凸问题的一个甚为有效的方法。本章从拉格朗日对偶问题 (*Lagrange Dual Problem*)入手, 逐步讨论有关对偶性的一些理论和例证。

### 2.1 探讨矩阵分析

解拉格朗日对偶问题时, 有时会涉及到矩阵的 *Frobenius norm*, 而这一范数可以改写成矩阵的迹, 因此, 需要推导、理解一些对迹的运算。

■ 例 2.1 考虑矩阵  $\mathbf{X}$  有  $[M, N]$  的大小, 且存在一个矩阵  $\mathbf{A}$  大小为  $[N, M]$ , 于是试求  $\frac{d\text{Tr}(\mathbf{XA})}{d\mathbf{X}}$  的值。

考虑展开迹成为元素的形式,

$$\left( \frac{d}{d\mathbf{X}} (\text{Tr}(\mathbf{XA})) \right)_{kl} = \frac{d}{dx_{kl}} \left( \sum_{i=1}^M \sum_{j=1}^N x_{ij} a_{ji} \right) = a_{lk}. \quad (2.1)$$

由于求导得到的矩阵形状和  $\mathbf{X}$  相同, 易

$$\frac{d\text{Tr}(\mathbf{XA})}{d\mathbf{X}} = \mathbf{A}^T. \quad (2.2)$$

同时, 由于  $\text{Tr}(A) = \text{Tr}(A^T)$ , 可以推断出  $\frac{d\text{Tr}(\mathbf{AX}^T)}{d\mathbf{X}} = \mathbf{A}_\circ$  ■

■ 例 2.2 考虑矩阵  $\mathbf{X}$  有  $[M, N]$  的大小, 且存在一个矩阵  $\mathbf{A}$  大小为  $[N, N]$ , 于是试求  $\frac{d\text{Tr}(\mathbf{XAX}^T)}{d\mathbf{X}}$  的值。

考虑展开迹成为元素的形式,

$$f(\mathbf{X}, \mathbf{A}) = (\text{Tr}(\mathbf{XAX}^T))_{kl} = \sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^N x_{ij} x_{ik} a_{jk}. \quad (2.3)$$

可见, 如果对(2.3)求取  $x_{pq}$  的导数, 讨论  $j, k$  可以分成三种情况:

- (1)  $i = p, j = k = q: \frac{df}{dx_{pq}}|_{j=k=q} = \frac{da_{qq}x_{pq}^2}{dx_{pq}} = 2x_{pq}a_{qq}.$
- (2)  $i = p, j = q, k \neq q: \frac{df}{dx_{pq}}|_{j=q, k \neq q} = \frac{d\sum_{k=1, k \neq q}^N x_{pq}x_{pk}a_{qk}}{dx_{pq}} = \sum_{k=1, k \neq q}^N x_{pk}a_{qk}.$
- (3)  $i = p, j \neq q, k = q: \frac{df}{dx_{pq}}|_{j \neq q, k=q} = \frac{d\sum_{j=1, j \neq q}^N x_{pj}x_{pq}a_{jq}}{dx_{pq}} = \sum_{j=1, j \neq q}^N x_{pj}a_{jq}.$

合并(1)(2)(3)的值, 可得:

$$\frac{df}{dx_{pq}} = \sum_{i=1}^N x_{pi}(a_{iq} + a_{qi}). \quad (2.4)$$

这恰恰是  $\mathbf{X}(\mathbf{A} + \mathbf{A}^T)$  位于的  $p, q$  的元素, 于是

$$\frac{d\text{Tr}(\mathbf{X}\mathbf{A}\mathbf{X}^T)}{d\mathbf{X}} = \mathbf{X}(\mathbf{A} + \mathbf{A}^T). \quad (2.5)$$

类似的推导还可以得到

$$\frac{d\text{Tr}(\mathbf{X}^T\mathbf{A}\mathbf{X})}{d\mathbf{X}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{X}. \quad (2.6)$$

注意  $\mathbf{X}$  和  $\mathbf{A}$  的尺寸, (2.6)并不等价于2.5, 但是推导过程相同, 这是因为, (2.6)唯一的区别是, 解元素所为契合的尺寸相当于(2.5)中的  $\mathbf{X}^T$ 。

## 2.2 拉格朗日对偶问题

考虑一个优化问题的标准形式,

**问题 2.1 – 普通优化问题.**

$$\min_x f_0(x), \quad (2.7-1)$$

$$s.t. f_i(x) \leq 0, i = 1, \dots, m, \quad (2.7-2)$$

$$h_i(x) = 0, i = 1, \dots, p. \quad (2.7-3)$$

显然, 它不但是非线性的, 还有可能是非凸的, 甚至还有不止一种限制条件, 这甚至为采用梯度下降法求取局部最优解都带来了困难。然而, 如果我们设存在这样的函数:

$$I_-(u) = \begin{cases} 0, & u \leq 0, \\ +\infty, & u > 0. \end{cases} \quad (2.8)$$

以及这样的函数

$$I_0(u) = \begin{cases} 0, & u = 0, \\ +\infty, & u \neq 0. \end{cases} \quad (2.9)$$

就可以将原问题的限制条件写入到目标中, 从而改写成

$$\min_x f_0(x) + \sum_i I_-(f_i(x)) + \sum_i I_0(h_i(x)). \quad (2.10)$$

显然, 不在可行域<sup>1</sup>内, (2.10)必定导出  $+\infty$ 。因此, 该问题的最小值解必定限定在可行域内(尽管未必是最优解)。

然而, (2.10)并无实用价值, 因为  $I_-$ ,  $I_0$  均是不可微的函数, 从而无法优化。为了解决这个问题, 可以考虑对它们进行松弛, 设存在松弛变量  $\lambda_i$ ,  $\mu_i$ , 于是有拉格朗日函数:

$$\mathcal{L}(x, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x), \quad (2.11-1)$$

$$\text{s.t. } \lambda_i \geq 0, \mu_i \in \mathbb{R}. \quad (2.11-2)$$

可见, 假设  $x$  已经确定,  $\mathcal{L}$  势必成为仅仅和  $\boldsymbol{\lambda}$ ,  $\boldsymbol{\mu}$  有关的函数, 此时所有的  $f_i$  和  $h_i$  也都成为了定值, 从而易知  $\mathcal{L}$  成为对  $\boldsymbol{\lambda}$ ,  $\boldsymbol{\mu}$  的线性函数。

因此, 考虑若  $f_i(x) > 0$ , 必定  $\exists \lambda_i = +\infty$  使得  $\mathcal{L} = +\infty$ ; 同理, 若  $h_i(x) \neq 0$ , 必定  $\exists \mu_i = \text{sign}(h_i(x)) \times (+\infty)$  使得  $\mathcal{L} = +\infty$ 。

反之, 若有  $\forall i$ ,  $f_i(x) \leq 0$ ,  $h_i(x) = 0$ , 则此时  $\max\{\lambda_i f_i(x)\} = 0$ , 且  $\mu_i h_i(x) \equiv 0$ 。

故而, 可以想到,

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} f_0(x), & \forall i, f_i(x) \leq 0, h_i(x) = 0, \\ +\infty, & \exists i, f_i(x) > 0 \text{ or } h_i(x) \neq 0. \end{cases} \quad (2.12)$$

可见, 记(2.12)的值为  $p(x)$ ,

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= p(x), \\ p(x) &= f_0(x) + \sum_i I_-(f_i(x)) + \sum_i I_0(h_i(x)). \end{aligned} \quad (2.13)$$

$p(x)$  反映了, 对确定的  $x$ ,  $\mathcal{L}$  的最大值即理想问题(2.10)的值。因此, 可以用  $\min_x p(x)$  来代替原问题的解。我们称这个解为  $p^*$ 。

上述推导表明,  $\mathcal{L}$  可以通过先最大化参数  $\boldsymbol{\lambda}$ ,  $\boldsymbol{\mu}$ , 后最小化自变量  $x$  这样的顺序, 达到对问题 2.1 的严格等价。然而, 它仍然是一个不可微的问题, 因为  $\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \mathcal{L}$  完全有可能取到  $+\infty$ 。因此, 我们需要重新考虑对  $\mathcal{L}$  的处理。

首先, 需要假设的是, 可行域的限制条件都是有效的。什么是有效的呢? 一言以蔽之, 就是这些限制条件本身要有全部成立的可能, 亦即记可行域为  $\mathcal{D}$ , 则  $\forall x_0 \in \mathcal{D}$ ,  $\forall i$ ,  $f_i(x_0) \leq 0$ ,  $h_i(x_0) = 0$ 。否则, 就永远找不到可行域内的解, 这样问题就变成了无解的问题。

既然这些限制条件满足有效, 就表明,

$$\because \forall \boldsymbol{\lambda}, \boldsymbol{\mu}, f_0(x_0) + \sum_{i=1}^m \lambda_i f_i(x_0) + \sum_{i=1}^p \mu_i h_i(x_0) = f_0(x_0) + \sum_{i=1}^m \lambda_i f_i(x_0) \leq f_0(x_0), \quad (2.14-1)$$

$$\therefore \forall \boldsymbol{\lambda}, \boldsymbol{\mu}, \min_x f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x) \leq \min_{x \in \mathcal{D}} f_0(x) = p^*. \quad (2.14-2)$$

---

<sup>1</sup>可行域: 满足  $x$  符合限制条件的取值范围

$x$  不限于在  $\mathcal{D}$ , 但因为  $x \in \mathcal{D}$  时(2.14-2)就必然成立, 故而(2.14-2)恒成立。这表明, 如果先对  $\mathcal{L}$  最小化自变量  $x$ , 再对  $\mathcal{L}$  最大化参数  $\lambda, \mu$ , 必然有:

$$\max_{\lambda, \mu} \min_x \mathcal{L}(x, \lambda, \mu) \leq \min_{x \in \mathcal{D}} f_0(x) = \min_x \max_{\lambda, \mu} \mathcal{L}(x, \lambda, \mu). \quad (2.15)$$

这个不等式左侧是一个优化问题, 我们称之为**拉格朗日对偶问题**; 而右侧是考虑限制条件下, 优化目标的最小值。在可行域内, (2.14-2)恒成立, 但可行域外, (2.14-2)也有可能成立(因为  $\lambda, \mu$  需要在求对  $x$  的最小时先给定)。故而, **对偶问题**的解既必定可以取可行域内最小值的解, 也可能取可行域外的、某个比可行域内最小值还要小的解, 这是因为, 这个解本身是在全域的  $x$  上定义的。

综上, 为了便于识读, 定义**对偶问题**的解的函数值为  $d^*$ , 于是可以写成如下的对偶形式:

### 问题 2.2 – 拉格朗日对偶问题.

$$d^* = \max_{\lambda, \mu} \min_x \mathcal{L}(x, \lambda, \mu) \leq \min_x \max_{\lambda, \mu} \mathcal{L}(x, \lambda, \mu) = p^*. \quad (2.16-1)$$

$$s.t. \mathcal{L}(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x), \quad (2.16-2)$$

$$\forall i, \lambda_i \geq 0, \mu_i \in \mathbb{R}. \quad (2.16-3)$$

(2.16-1)显示了, **对偶问题**昭示了原问题的下界。它的最小值  $d^*$  一定小于等于可行域内的最小值  $p^*$ , 但不一定等于全域的最小值。换言之,  $d^*$  在全域最小值和可行域最小值之间, 它的解不一定在**可行域内**。

图 2.1 分别展示了两种情形下**拉格朗日对偶问题**的求解过程。在该情形下, 只有一个  $f_1(x) \leq 0$  的限制条件, 且该限制条件是一个将可行域限制在约  $x \in [25, 75]$  之间的开口向上的二次函数。其中图 (a) 收敛到了最优解, 但图 (b) 没有收敛到最优解, 下面将会探讨在什么情况下, **对偶问题**能收敛到最优解, 即  $d^* = p^*$ 。

(a) 下界最优

(b) 下界非最优

图 2.1: 拉格朗日对偶问题下界收敛范例。

### 2.3 稀疏编码

稀疏编码 (*Sparse Coding*) 是一个典型的对偶问题求解过程。作为广泛应用于图像处理领域的一个工具, 形式最简单的稀疏编码问题也需要进行一些预处理工作。为了以便人能理解它的物理含义, 尽管它可能和问题本身没有太大的联系, 但本节还是对其进行了叙述, 它可以描述成如下过程:

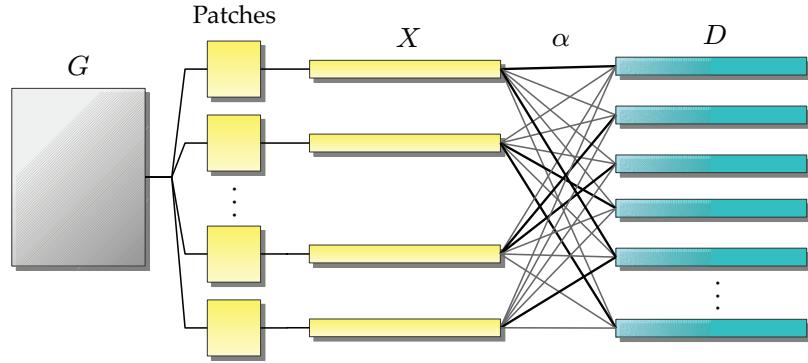


图 2.2: 稀疏编码的预处理过程。

- (1) 如图 2.2 所示, 首先将原图  $G$  按位置、通过滑动窗口的方法拆解成片 (*Patch*) 的集合;
- (2) 对每个 Patch, 将其分别展平成一条维数等于像素总数的向量 (向量化)。这样的向量集合称为输入集  $\mathbf{X}$ ;
- (3) 构建  $K$  个字典向量, 这些向量的集合称为字典  $\mathbf{D}$ , 每条向量的维数和 Patch 的维数相同, 用随机取出的 Patch 或其他随机数等方式初始化数值。字典向量的数目  $K$  人为决定;
- (4) 构建字典-Patch 编码关系, 每个 Patch 对应一组编码, 编码向量记为  $\alpha_i$ , 向量的维数为  $K$ ;
- (5) 至此, 开始本节所进行的训练过程。

首先, 通过以上步骤, 我们可以了解到该问题包括哪些物理量:

表 2.1: 稀疏编码的参数说明

符号	描述	维数
$D$	Patch 向量化后的维数, 只与 Patch 的大小有关	-
$K$	构建的字典向量数目, 人为设定	-
$N$	Patch 的数目, 与输入图像 $G$ 的大小有关	-
$\mathbf{X}$	一次训练的输入集, 不一定来自同一原图 $G$	$D \times N$
$\mathbf{x}_i$	输入集的一个样本(Patch 向量)	$D \times 1$
$\mathbf{D}$	包括 $K$ 个字典向量的字典总集(注意 $K \gg N$ )	$D \times K$
$D(:, k)$	第 $k^{th}$ 条字典向量	$D \times 1$
$\alpha_i$	与第 $i^{th}$ 个样本 $\mathbf{x}_i$ 对应的稀疏编码	$K \times 1$
$\lambda$	控制稀疏性的罚参数, 同时影响到保真度	-

可见, 稀疏编码实质上是将  $D$  维数的 Patch 转移到  $K$  维数的过程。看起来, 这似乎可以理解成用字典压缩编码数据, 但实际上这种理解是似是而非的。因为, 实际操作中为了保证能恢复出原图, 需要取较大的  $K$ , 这个数额至少会大于  $D$ 。可见, 由于  $N$  一般很大, 一个原图

产生的所有  $\alpha_i$  集合本身就是一个巨大的矩阵了, 如果它不是稀疏的, 在计算复杂度上都会带来不可忽视的开销(哪怕是多项式时间的算法)。

训练稀疏编码的算法有很多种, 在本节中, 介绍基于凸优化和对偶问题的方法。

### 2.3.1 解对偶问题训练字典

训练字典的过程, 可以按照表 2.1 描述成如下问题:

#### 问题 2.3 – 稀疏编码的训练.

$$\min_{\mathbf{D}, \{\alpha_i\}_{i=1}^N} \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1, \quad (2.17-1)$$

$$\text{s.t. } \|D(:, k)\|_2 \leq 1, \forall k \in \{1, 2, \dots, K\}. \quad (2.17-2)$$

如此相似, 不是吗? 它的形式, 和节 1.3.1 介绍的最小平方形式相若, 所不同的是, 问题 1.2 中的  $\mathbf{A}$ (此处对应的是  $\mathbf{D}$ ) 是已知的定值, 但此处  $\mathbf{D}$  同样需要训练。

此外, 还追加了一个限制条件(2.17-2), 这个限制条件是什么意思呢? 原来, 如果不限制单个字典向量的  $l_2$ -norm, 就有无数组  $\mathbf{D}$  能和  $\alpha_i$  形成相同的矩阵积; 而由于  $l_2$ -norm 的作用, 为了使  $\alpha_i$  无限制地小, 会使  $\mathbf{D}$  无限制地大, 这就不是我们所期望看到的了! 因此, 加上了这样一个限制  $l_2$ -norm 的条件。但是, 由于  $\alpha_i$  总是倾向于足够小, 实际操作中  $\mathbf{D}$  的单个字典向量的  $l_2$ -norm 往往为 1(我们所设定的最大值)。

#### Lasso 问题的解析解

由于该问题中存在两个互相耦合的自变量  $\mathbf{D}$ ,  $\{\alpha_i\}_{i=1}^N$ , 常规的做法是: 在一个步长里, 先固定其中一个, 训练另外一个变量; 接下来固定另外一个, 训练之前固定的这个变量。我们先考虑简单的场合, 即固定  $\mathbf{D}$  的情况:

#### 问题 2.4 – Lasso 问题.

$$\min_{\{\alpha_i\}_{i=1}^N} \sum_{i=1}^N \left( \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right). \quad (2.18-1)$$

注意由于  $\mathbf{D}$  的固定, 限制条件(2.17-2)变得与  $\{\alpha_i\}_{i=1}^N$  无关, 从而转换成一个无限制条件的问题, 直接求取梯度, 于是,

$$\begin{aligned} \frac{d}{d\alpha_k} \left( \sum_{i=1}^N \left( \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right) \right) &= \frac{d}{d\alpha_k} \left( \|\mathbf{x}_k - \mathbf{D}\alpha_k\|_2^2 + \lambda \|\alpha_k\|_1 \right) \\ &= \frac{d}{d\alpha_k} ((\mathbf{x}_k - \mathbf{D}\alpha_k)^T (\mathbf{x}_k - \mathbf{D}\alpha_k)) + \lambda \text{sign}(\alpha_k) \\ &= \frac{d}{d\alpha_k} (-2\alpha_k^T \mathbf{D}^T \mathbf{x}_k + \alpha_k^T \mathbf{D}^T \mathbf{D}\alpha_k) + \lambda \text{sign}(\alpha_k) \\ &= -2\mathbf{D}^T \mathbf{x}_k + 2\mathbf{D}^T \mathbf{D}\alpha_k + \lambda \text{sign}(\alpha_k) = 0. \end{aligned} \quad (2.19)$$

由于符号函数的存在, 尽管我们无法写出(2.19)的解析解, 但求得其数值解仍然是简单的。换言之,  $\alpha_k$  可以快速求解。

### 快速求取 Lasso 数值解

尽管节 2.3.1 指出 Lasso 问题求数值解是快速的, 但究竟能快到什么程度, 还要依赖具体的算法。提出了一种求取 Lasso 数值解的快速算法。

(2.19) 存在一个最大的不可解要素就是,  $\text{sign}(\alpha_k)$  在 0 处不可导。然而, 0 值恰恰是一个需要重点讨论的值, 为了解决这个问题, 可以假设一个变量  $\theta_k$ , 用来代替  $\text{sign}(\alpha_k)$  参与求解, 于是, 这个“伪解析解”是

$$\alpha_k = (\mathbf{D}^T \mathbf{D})^{-1} (\mathbf{D}^T \mathbf{x}_k - \frac{\lambda}{2} \theta_k). \quad (2.20)$$

(2.20) 能代替解析解的情况有且仅有一种, 即是  $\theta_k = \text{sign}(\alpha_k)$  ( $\theta \in \{0, 1, -1\}$ )。代入到(2.19), 可以得到

$$\begin{aligned} \frac{1}{\lambda} \frac{d\text{Lasso}}{d\alpha_k} &= \text{sign} \left( (\mathbf{D}^T \mathbf{D})^{-1} (\mathbf{D}^T \mathbf{x}_k - \frac{\lambda}{2} \theta_k) \right) - \theta_k \\ &= \text{sign} \left( \mathbf{D}^T \mathbf{x}_k - \frac{\lambda}{2} \theta_k \right) - \theta_k = \text{sign} (\mathbf{y}_k - \lambda \theta_k) - \theta_k = 0. \end{aligned} \quad (2.21)$$

可见, 为了使最小条件实现, (2.21) 必须对解的本身有所要求。具体来说, 设向量  $\mathbf{y}_k = 2\mathbf{D}^T \mathbf{x}_k$ , 它恰好是  $\|\mathbf{x}_k - \mathbf{D}\alpha_k\|_2^2$  在假设  $\alpha_k$  为 0 时的反向偏导。于是可知, 在  $\mathbf{y}_k > \lambda$  的时候, 对应有  $\theta_k = 1$ ; 在  $\mathbf{y}_k < -\lambda$  的时候, 对应有  $\theta_k = -1$ , 这两种情况下, (2.20) 都正是我们想要的解。

于是, 一个问题出现了, 当  $\mathbf{y}_k \in [-\lambda, \lambda]$  的时候该怎么办呢? (2.21) 告诉我们, 在这种情况下,  $\frac{d\text{Lasso}}{d\alpha_k}$  不是  $> 0$  就是  $< 0$ , 至少代入(2.20)的做法是不可取的, 故而, 让我们回到(2.19)的结论。一旦  $\alpha_k > 0$ , 第一项和第三项的和必定为正, 而第二项由于  $\alpha_k > 0$  也取正, 故  $\frac{d\text{Lasso}}{d\alpha_k} > 0$ ; 反之, 若  $\alpha_k < 0$ , 第一项和第三项的和必定为负, 而第二项也取负, 同理, 有,  $\frac{d\text{Lasso}}{d\alpha_k} < 0$ 。故而可知, 在  $\mathbf{y}_k \in [-\lambda, \lambda]$  的时候, 尽管令  $\alpha_k = 0$  得到的并非 0 值, 但它的 + 侧为正值, - 侧为负值, 仍然符合在  $\alpha_k = 0$  取到最小值。因此,  $\alpha_k$  在许多情况下都可能收敛到 0, 这就实现了 Lasso 问题的稀疏性。

值得注意的是, 上述步骤中, 都是针对向量进行讨论的。这只是为了方便表达, 并不是在说,  $\alpha_k = 0$  全部都要取相同的值或者同号。事实上, 如果把  $\mathbf{D}$  拆解成列向量, 上述讨论就可以转换成针对第  $p$  列的元素讨论。实际操作中为了提高效率, 可以取  $\alpha_k$  的元素具有同性质(例如同号)的子集。

综上, 经过多番迭代, 核心步骤只有两种, (1) 考察所有  $\alpha_{pk} = 0$  的元素, 并按照(2.20)更新所有能令  $|y_{pk}| > \lambda$  的  $\alpha_{pk}$ ; (2) 考察所有其他元素, 如果  $|y_{pk}| < \lambda$ , 就令所有对应的  $\alpha_{pk} = 0$ 。

### 拉格朗日对偶法求字典

接下来考虑另一情况, 即, 固定  $\{\alpha_i\}_{i=1}^N$ , 求解  $\mathbf{D}$ 。这种情况下, 限制条件(2.17-2)不可忽视, 这意味着我们必须运用对偶问题来求解该问题; 但另一方面, 所幸的是由于  $\{\alpha_i\}_{i=1}^N$  成为了常数, 稀疏项可以抹去。

考虑到  $l_1$ -norm 被去除了, 书写  $\{\alpha_i\}_{i=1}^N$  这样的向量形式是很麻烦的, 不妨设矩阵

$$\mathbf{A} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_N]. \quad (2.22)$$

于是, 可以用矩阵的 Frobenius norm 改写原问题, 等效形式如下:

问题 2.5 – 有限制条件的字典优化.

$$\min_{\mathbf{D}} \|\mathbf{X} - \mathbf{DA}\|_F^2, \quad (2.23-1)$$

$$\text{s.t. } \|D(:, k)\|_2 \leq 1, \forall k \in \{1, 2, \dots, K\}. \quad (2.23-2)$$

考虑(2.2)给出的解法, 可以先写出拉格朗日函数:

$$\mathcal{L}(\mathbf{D}, \boldsymbol{\mu}) = \|\mathbf{X} - \mathbf{DA}\|_F^2 + \sum_{j=1}^K \mu_j \sum_{i=1}^D (D_{ij}^2 - 1). \quad (2.24)$$

其中, *Frobenius norm* 可以被改写成矩阵的迹, 于是

$$\begin{aligned} \|\mathbf{X} - \mathbf{DA}\|_F^2 &= \text{Tr}((\mathbf{X} - \mathbf{DA})(\mathbf{X} - \mathbf{DA})^T) \\ &= \text{Tr}(\mathbf{XX}^T) + \text{Tr}(\mathbf{DAA}^T \mathbf{D}^T) - 2\text{Tr}(\mathbf{DAX}^T). \end{aligned} \quad (2.25)$$

设对角矩阵  $\boldsymbol{\Lambda} = \text{diag}(\{\mu_j\}_{j=1}^N)$ , 则剩余的部分同样也可以改写成矩阵的迹, 于是

$$\begin{aligned} \sum_{j=1}^K \mu_j \sum_{i=1}^D (D_{ij}^2 - 1) &= \sum_{j=1}^K \mu_j \sum_{i=1}^D (D_{ij}^2) - \sum_{j=1}^K \mu_j \\ &= \text{Tr}(\mathbf{D}\boldsymbol{\Lambda}\mathbf{D}^T - \boldsymbol{\Lambda}). \end{aligned} \quad (2.26)$$

(2.24)可以被改写成

$$\mathcal{L}(\mathbf{D}, \boldsymbol{\Lambda}) = \text{Tr}(\mathbf{XX}^T + \mathbf{DAA}^T \mathbf{D}^T - 2\mathbf{DAX}^T + \mathbf{D}\boldsymbol{\Lambda}\mathbf{D}^T - \boldsymbol{\Lambda}). \quad (2.27)$$

于是, 求对  $\mathbf{D}$  的梯度, 有

$$\begin{aligned} \frac{d}{d\mathbf{D}} (\mathcal{L}(\mathbf{D}, \boldsymbol{\Lambda})) &= \frac{d}{d\mathbf{D}} (\text{Tr}(\mathbf{DAA}^T \mathbf{D}^T - 2\mathbf{DAX}^T + \mathbf{D}\boldsymbol{\Lambda}\mathbf{D}^T)) \\ &= 2\mathbf{DAA}^T - 2\mathbf{XA}^T + 2\mathbf{D}\boldsymbol{\Lambda} = 0. \\ \mathbf{D} &= \mathbf{XA}^T (\mathbf{AA}^T + \boldsymbol{\Lambda})^{-1}. \end{aligned} \quad (2.28)$$

这是唯一解, 在这里略去它是最小值解的证明。于是, 代入  $\mathbf{D}$  到(2.27), 有

$$\begin{aligned} \min_{\mathbf{D}} \mathcal{L}(\mathbf{D}, \boldsymbol{\Lambda}) &= \text{Tr}(\mathbf{XX}^T + \mathbf{DAA}^T \mathbf{D}^T - 2\mathbf{DAX}^T + \mathbf{D}\boldsymbol{\Lambda}\mathbf{D}^T - \boldsymbol{\Lambda}) \\ &= \text{Tr}(\mathbf{XX}^T + \mathbf{D}(\mathbf{AA}^T + \boldsymbol{\Lambda})\mathbf{D}^T - 2\mathbf{DAX}^T - \boldsymbol{\Lambda}) \\ &= \text{Tr}(\mathbf{XX}^T + \mathbf{XA}^T (\mathbf{AA}^T + \boldsymbol{\Lambda})^{-1} (\mathbf{AA}^T + \boldsymbol{\Lambda}) (\mathbf{AA}^T + \boldsymbol{\Lambda})^{-1} \mathbf{AX}^T \\ &\quad - 2\mathbf{XA}^T (\mathbf{AA}^T + \boldsymbol{\Lambda})^{-1} \mathbf{AX}^T - \boldsymbol{\Lambda}) \\ &= \text{Tr}(\mathbf{XX}^T - \mathbf{XA}^T (\mathbf{AA}^T + \boldsymbol{\Lambda})^{-1} \mathbf{AX}^T - \boldsymbol{\Lambda}). \end{aligned} \quad (2.29)$$

接下来需要求取  $\frac{\partial \min_{\mathbf{D}} \mathcal{L}}{\partial \mu_i}$  的值, 由(2.29)知这是一个实函数对实变参量求导的过程, 所得也

是一个实变参量。由于求导的对象是实变参量, 可以将分母部分写到迹里面, 于是有:

$$\begin{aligned}\frac{\partial \min_{\mathbf{D}} \mathcal{L}}{\partial \mu_i} &= \text{Tr} \left( \frac{\partial \mathbf{X} \mathbf{X}^T}{\partial \mu_i} - \frac{\partial \mathbf{X} \mathbf{A}^T (\mathbf{A} \mathbf{A}^T + \boldsymbol{\Lambda})^{-1} \mathbf{A} \mathbf{X}^T}{\partial \mu_i} - \frac{\partial \boldsymbol{\Lambda}}{\partial \mu_i} \right) \\ &= -\text{Tr} \left( \frac{\partial \mathbf{X} \mathbf{A}^T (\mathbf{A} \mathbf{A}^T + \boldsymbol{\Lambda})^{-1} \mathbf{A} \mathbf{X}^T}{\partial \mu_i} \right) - 1.\end{aligned}\quad (2.30)$$

$\frac{\partial \boldsymbol{\Lambda}}{\partial \mu_i}$  得到的是一个对角线上只有一个元素为 1 的对角矩阵, 原理参见(2.31)。于是, 该问题从求取原函数迹的导数, 转化成了求取中间项的导数的迹。这是一个蔚为有趣的问题, 下面将在给出一个完整的推导过程。

■ 例 2.3 考虑矩阵  $\mathbf{P}$  有  $[N, M]$  的大小, 且存在对角矩阵  $\mathbf{X}$ , 任意矩阵  $\mathbf{A}$ , 大小均为  $[N, N]$ ;  $\mathbf{X} = \text{diag}(x_1, x_2, \dots, x_N)$ 。于是, 试求  $\text{Tr} \left( \frac{\partial \mathbf{P}^T (\mathbf{X} + \mathbf{A})^{-1} \mathbf{P}}{\partial x_i} \right)$  的值。

设存在  $\mathbf{X}$  的小量增量  $\Delta \mathbf{X}$ , 满足:

$$\Delta \mathbf{X} = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & \Delta x_i & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}. \quad (2.31)$$

可见, (2.31)对  $x_i$  求导, 得到的将是一个只有对角线上第  $i$  个元素为 1, 其他位置全部为 0 的矩阵。

于是, 整个问题可以考虑改写为导数的定义式:

$$\begin{aligned}\frac{\partial \mathbf{P}^T (\mathbf{X} + \mathbf{A})^{-1} \mathbf{P}}{\partial x_i} &= \frac{\mathbf{P}^T (\mathbf{X} + \mathbf{A} + \Delta \mathbf{X})^{-1} \mathbf{P} - \mathbf{P}^T (\mathbf{X} + \mathbf{A})^{-1} \mathbf{P}}{\Delta x_i} \\ &= \frac{\mathbf{P}^T ((\mathbf{X} + \mathbf{A} + \Delta \mathbf{X})^{-1} - (\mathbf{X} + \mathbf{A})^{-1}) \mathbf{P}}{\Delta x_i} \\ &= \frac{\mathbf{P}^T (\mathbf{X} + \mathbf{A})^{-1} ((\mathbf{X} + \mathbf{A}) - (\mathbf{X} + \mathbf{A} + \Delta \mathbf{X})) (\mathbf{X} + \mathbf{A} + \Delta \mathbf{X})^{-1} \mathbf{P}}{\Delta x_i} \\ &= -\frac{\mathbf{P}^T (\mathbf{X} + \mathbf{A})^{-1} \Delta \mathbf{X} (\mathbf{X} + \mathbf{A} + \Delta \mathbf{X})^{-1} \mathbf{P}}{\Delta x_i} \\ &= - \left( \mathbf{P}^T (\mathbf{X} + \mathbf{A})^{-1} \frac{\Delta \mathbf{X}}{\Delta x_i} (\mathbf{X} + \mathbf{A})^{-1} \mathbf{P} \right) \\ &= - \left( \mathbf{P}^T (\mathbf{X} + \mathbf{A})^{-1} \mathbf{e}_i \mathbf{e}_i^T (\mathbf{X} + \mathbf{A})^{-1} \mathbf{P} \right).\end{aligned}\quad (2.32)$$

其中,  $\mathbf{e}_i = [0 \ 0 \ \cdots \ 1 \ \cdots \ 0]^T$ , 表示第  $i$  个元素为 1, 其他元素为 0 的单位列向量。由(2.31), 显然有:

$$\frac{\Delta \mathbf{X}}{\Delta x_i} = \mathbf{e}_i \mathbf{e}_i^T. \quad (2.33)$$

故(2.32)成立。

显然, 由于  $\mathbf{e}_i$  是个列向量,  $\mathbf{P}^T(\mathbf{X} + \mathbf{A})^{-1}\mathbf{e}_i$  也是个列向量, 记为  $\mathbf{p}$ , 显然由于对称关系,  $\mathbf{p}^T = \mathbf{e}_i^T(\mathbf{X} + \mathbf{A})^{-1}\mathbf{P}$ 。于是, 用  $\mathbf{p}$  改写(2.32), 并代入原问题, 有

$$\begin{aligned}\text{Tr}\left(\frac{\partial \mathbf{P}^T(\mathbf{X} + \mathbf{A})^{-1}\mathbf{P}}{\partial x_i}\right) &= -\text{Tr}(\mathbf{p}\mathbf{p}^T) = -\mathbf{p}^T\mathbf{p} = -\|\mathbf{p}\|_2^2 \\ &= -\|\mathbf{P}^T(\mathbf{X} + \mathbf{A})^{-1}\mathbf{e}_i\|_2^2.\end{aligned}\quad (2.34)$$

即为所求。 ■

用(2.30)的参数形式改写例 2.3 的结论, 代入  $\mathbf{P} = \mathbf{AX}^T$ ,  $\mathbf{X} = \mathbf{\Lambda}$ ,  $\mathbf{A} = \mathbf{AA}^T$ , 于是,

$$\frac{\partial \min_{\mathbf{D}} \mathcal{L}}{\partial \mu_i} = \|\mathbf{XA}^T(\mathbf{AA}^T + \mathbf{\Lambda})^{-1}\mathbf{e}_i\|_2^2 - 1 = 0. \quad (2.35)$$

这是一个最大值点方程, 二阶求导的过程在此省去。虽然我们不能写出它的解析解, 但显然这是一个很容易求解的凸问题。故而, 拉格朗日对偶问题有解, 且解(2.35)的结果  $\mathbf{\Lambda}$  代入到(2.28), 即可求得最优的字典值  $\mathbf{D}^*$ 。通过交替解 Lasso 问题和求取最优化字典的过程, 即可完成整个稀疏编码的训练。

有人可能会问, 这里又没有证明 KKT 条件, 单凭一个对偶问题有解, 怎么就能武断地下(2.16-1)的不等式取等这样严格的限制条件成立的结论呢? 对于这个问题, 我们不妨回到(2.28), 看看我们之前解出的最优解是什么。多么巧妙啊, 将(2.28)代入到(2.35), 竟然能得到:

$$\|\mathbf{De}_i\|_2^2 = \|D(:, i)\|_2 = 1. \quad (2.36)$$

它反映了一个事实, 那就是限制条件(2.23-2)被以等号的形式严格满足了! 尽管我们并无心去考察拉格朗日对偶问题是否能代替原问题, 甚至我们原本的限制条件也只有一个松弛的不等号, 但该解法给出的结果却准确无误地表明, 我们所求的限制条件, 事实上确实以完全严格的形式满足了。

## 2.4 对偶条件

### 2.4.1 对偶问题的凹性质

在讨论对偶问题的相关条件之前, 我们首先要厘清的问题是, 为什么要解对偶问题? 在前文我们已经知道, 对于一个任意的优化问题, 它可能是非凸的, 还有诸多限制条件, 这都成为我们解这些问题的阻碍。因此我们考虑取一个对偶问题。那么这个对偶问题相比于原问题有什么优势呢? (1) 首先, 它成功地将限制条件写进了目标函数里面, 让我们控制可行域变得非常容易; (2) 对偶问题一定是一个凸问题, 换言之, 它的目标函数和所有的限制条件都一定是凸函数。对于(1)我们可以从直观上感受到; 对于(2), 对偶问题的限制条件都是线性的, 故而下面叙述对目标函数凹性质的证明。

**引理 2.4.1 – 对偶问题的等价简化.** 问题 2.2, 亦即  $\max_{\lambda, \mu} \min_x \mathcal{L}(x, \lambda, \mu)$ , 可以简化、等价成不含

$\mu$  的形式, 完整的表述为:

$$\max_{\lambda} \min_x \mathcal{L}(x, \lambda), \quad (2.37-1)$$

$$s.t. \mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x), \quad (2.37-2)$$

$$\forall i, \lambda_i \geq 0. \quad (2.37-3)$$

首先, 考虑将问题 2.1 中的  $h_i(x) = 0$  改写成  $h_i(x) \leq 0$  和  $h_i(x) \geq 0$  的形式, 于是有

**问题 2.6 – 普通优化问题的等价改写.**

$$\min_x f_0(x), \quad (2.38-1)$$

$$s.t. f_i(x) \leq 0, i = 1, \dots, m, \quad (2.38-2)$$

$$h_i(x), -h_i(x) \leq 0, i = 1, \dots, p. \quad (2.38-3)$$

进而, 可以改写拉格朗日函数

$$\begin{aligned} \mathcal{L}(x, \lambda, \mu) &= f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_{i+} h_i(x) + \sum_{i=1}^p (-\mu_{i-}) (-h_i(x)), \\ &= f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p (\mu_{i+} + \mu_{i-}) h_i(x), \end{aligned} \quad (2.39-1)$$

$$s.t. \lambda_i \geq 0, \mu_{i+} \geq 0, -\mu_{i-} \geq 0. \quad (2.39-2)$$

显然, 如果我们用  $\mu_i$  代替  $(\mu_{i+} + \mu_{i-})$ , 我们就可以得到(2.11-1)形式的表述; 如果我们将  $\mu_{i+}, -\mu_{i-}$  并入  $\lambda_i$  中, 我们就能得到(2.37-2)。故而, 在下面的讨论中, 只考虑有  $\lambda_i$  的情况, 因为  $\mu_i$  可以被  $\lambda_i$  完全等价地表示出来。

**定理 2.4.1 – 对偶问题的凹性质.** 对偶问题的目标函数  $\mathcal{L}(\lambda) = \min_x \mathcal{L}(x, \lambda)$  一定是一个凹函数。

设有一个固定的  $x_0$ , 那么对它而言  $\lambda_i f_i(x_0)$  对任何一个  $\lambda_i$  都是一个线性变化的函数,  $f_0(x_0)$  是一个定值。故而,  $\mathcal{L}(x_0, \lambda)$  是一组线性函数的和, 亦即一个多项式次数不超过 1 的函数, 这样的函数称为**仿射函数**。

**引理 2.4.2 – 仿射函数的凹性质.** 具有相同参变量的两个凹函数  $f_{c1}(\mathbf{x}), f_{c2}(\mathbf{x})$  的最小值函数  $f_m(\mathbf{x}) = \min(f_{c1}(\mathbf{x}), f_{c2}(\mathbf{x}))$  仍然是一个凹函数。因此, 任意数目的仿射函数共同的最小值函数, 是一个凹函数。

由于  $f_{c1}$  和  $f_{c2}$  都是凹函数,  $f_{c1}(\alpha\mathbf{x} + \beta\mathbf{y}) \geq \alpha f_{c1}(\mathbf{x}) + \beta f_{c1}(\mathbf{y})$ , 对  $f_{c2}$  亦然。于是

$$\begin{aligned} \alpha f_m(\mathbf{x}) + \beta f_m(\mathbf{y}) &= \alpha \min(f_{c1}(\mathbf{x}), f_{c2}(\mathbf{x})) + \beta \min(f_{c1}(\mathbf{y}), f_{c2}(\mathbf{y})) \\ &\leq \alpha f_{c1}(\mathbf{x}) + \beta f_{c1}(\mathbf{y}) \leq f_{c1}(\alpha\mathbf{x} + \beta\mathbf{y}). \end{aligned} \quad (2.40)$$

$$\begin{aligned}\alpha f_m(\mathbf{x}) + \beta f_m(\mathbf{y}) &= \alpha \min(f_{c1}(\mathbf{x}), f_{c2}(\mathbf{x})) + \beta \min(f_{c1}(\mathbf{y}), f_{c2}(\mathbf{y})) \\ &\leq \alpha f_{c2}(\mathbf{x}) + \beta f_{c2}(\mathbf{y}) \leq f_{c2}(\alpha\mathbf{x} + \beta\mathbf{y}).\end{aligned}\quad (2.41)$$

因为  $f_m(\alpha\mathbf{x} + \beta\mathbf{y})$  必然取  $f_{c1}(\alpha\mathbf{x} + \beta\mathbf{y})$  或  $f_{c2}(\alpha\mathbf{x} + \beta\mathbf{y})$  其中一个值。故而, 无论取哪个值, 总有  $\alpha f_m(\mathbf{x}) + \beta f_m(\mathbf{y}) \leq f_m(\alpha\mathbf{x} + \beta\mathbf{y})$  成立。亦即,  $f_m$  也是一个凹函数。

因为单个仿射函数是线性函数, 线性函数既是凹函数又是凸函数。因此, 由数学归纳法可知, 任意个仿射函数的最小值函数总是一个凹函数。

在已知该结论的情况下, 容易得到, 设有无数组  $\{x_0, x_1, x_2, \dots, x_i, \dots\}$  共同构成了  $x$  的定义域, 于是,  $\mathcal{L}(\boldsymbol{\lambda})$  可以看成这无数个元素对应的各自的仿射函数  $\mathcal{L}(x_i, \boldsymbol{\lambda})$  共同构成的最小值函数, 亦即:

$$\mathcal{L}(\boldsymbol{\lambda}) = \min_x \mathcal{L}(x, \boldsymbol{\lambda}) = \min \{\mathcal{L}(x_i, \boldsymbol{\lambda}), \mathcal{L}(x_2, \boldsymbol{\lambda}), \dots, \mathcal{L}(x_i, \boldsymbol{\lambda}), \dots\}. \quad (2.42)$$

由引理 2.4.2 可以直接导出,  $\mathcal{L}(\boldsymbol{\lambda})$  是一个凹函数。且这个凹函数有一个最大值点, 亦即拉格朗日对偶问题的解。

#### 2.4.2 弱对偶与强对偶

由(2.16-1), 我们知道

$$d^* \leq p^*. \quad (2.43)$$

我们把(2.43)这样的条件称为弱对偶; 当且仅当(2.43)严格取等的时候, 我们把这样的条件称为强对偶。弱对偶是所有拉格朗日对偶问题都必定满足的条件, 而强对偶却只能满足于部分特殊的对偶问题。图 2.1 已经展示了两种对偶条件下求解的区别, 找到一种证明强对偶是否成立的方法, 能给我们对对偶问题的应用带来巨大的便利, 这将在稍后说明原因。

了解这个性质, 一个关键点是, 要了解是什么造成了  $d^* \leq p^*$  这样的情况有可能发生。鉴于  $p^*$  表示的是可行域内的最小值,  $d^*$  比它更小的情况下, 一定表示对偶问题的最优解不在原问题的可行域内, 换言之, 这样的解对我们而言是没有太大意义的, 如果有意义, 也仅仅是提供了一个真正的最小值的下界。

现在回头看(2.36), 它并没有直接表明, 拉格朗日对偶问题的解是可行域内的最优解, 但是它蕴涵了一个明显的暗示, 接下来我们将讨论为什么它能导出拉格朗日问题的解就是我们想要的最优解。

**引理 2.4.3 – 最优解的等价性.** 一个原问题的对偶问题解  $\{x | \arg_x d^*\}$  在原问题的可行域内, 是该解与原问题的最优解等价的充要条件。

证明引理 2.4.3 的过程很简单。因为(2.43)必定成立, 如果这个解本身就在原问题的可行域内, 可行域内就再也找不出一个比它对应的  $d^*$  更小的  $p^*$ 。由此, 可以推出: 对偶问题的解在可行域中是对偶问题的解等价于原问题的最优解的充分条件; 而原问题的最优解要等价于对偶问题的解, 则一定需要对偶问题的解在可行域中, 故而二者构成了充要条件。换言之, 在可行域内对偶问题解与原问题的最优解是完全等价的。

同时, 在这一情况下, 我们也可以导出  $d^* = p^*$ , 这就是强对偶条件。换言之,

**定理 2.4.2 – 对偶问题的解在可行域中的必要条件.** 一个原问题的对偶问题解  $\{x | \arg_x d^*\}$ , 如果能被证明在原问题的可行域内, 则满足必要条件  $d^* = p^*$ 。

可见, 强对偶条件是对偶问题的解在可行域中的必要条件。那它是否是解在可行域中的充分条件呢? 遗憾的是, 在绝大多数情况下, 这一推论并不成立。下面将说明导致充分性无法满足的原因。

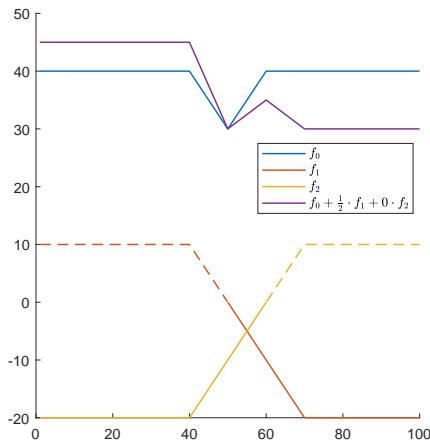
**引理 2.4.4** — 强对偶不是对偶问题在可行域中的充分条件. 强对偶条件不能推出对偶问题的解在可行域中, 亦即该条件下对偶问题的解不一定等价于原问题的解。

下面将用反证法证明: 绝大多数情况下, 可以找到符合问题 2.1 的反例, 使得不在原问题可行域内的对偶问题解也能导出强对偶条件。

考察这一问题, 重要的是找到什么样的反例能提供对强对偶条件充分性的反证。如果考虑存在加在限制条件上的一组拉格朗日乘子  $\{\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*\}$ , 使得  $\mathcal{L}(x, \lambda_1^*, \lambda_2^*, \dots, \lambda_n^*)$  对  $x$  取到的最小值  $d^*$  满足  $d^* = p^*$ , 则使得上述条件成立的  $x$  都是拉格朗日对偶函数的最优解(因为  $d^* \leq p^*$  必须成立, 可知  $d^*$  最大值就是  $p^*$ )。如果这样的  $x$  能找到不在可行域内的解, 说明拉格朗日对偶问题的最优解虽然满足强对偶条件, 却可能不在可行域内, 继而由必要性可知, 这样的解不能等效为原问题的最优解。

事实证明, 设计合理的  $f_0$  和  $f_i$ , 尤其是在限制条件不止 1 个的情况下, 极容易找到符合上述条件的反例。只讨论输入只有一个的实变量  $x$ , 可行域只由  $f_1(x)$  和  $f_2(x)$  决定的情况。其中,  $f_1$  单减, 用来决定左边界,  $f_2$  单增, 用来决定右边界。在这样一个简单的模型下, 只存在  $\lambda_1$  和  $\lambda_2$  两个拉格朗日乘子。

记原问题  $f_0(x)$  在可行域内的最优解为  $x_p$ , 在可行域外的最优解为  $x_q$ , 对偶问题的最优解在  $x_d$ , 于是:



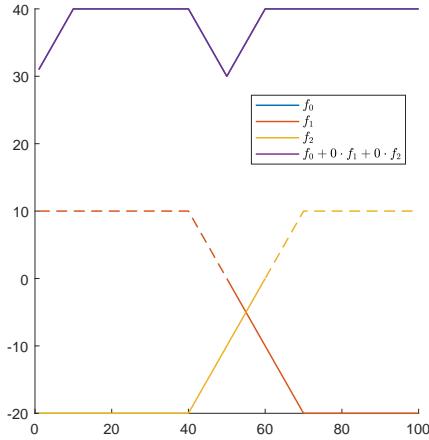
(a)  $\mathcal{L}(x_p) < \mathcal{L}(x_q)$  时, 反例在  $\lambda_1 = 0.5, \lambda_2 = 0$

(b) 反例的动态演示

图 2.3: 强对偶条件下对偶解推出原问题最优解的的反例 1。

$\mathcal{L}(x_p) < \mathcal{L}(x_q)$  时, 设  $f_0$  是一个仅仅在  $(40, 60)$  之间  $< 40$  的函数, 单减、单增的部分斜率各自为  $\mp 1$ 。 $f_1$  和  $f_2$  在单减/单增的部分斜率分别为  $\mp 1$ , 且各自定义了可行域的下界和上界。于是, 可行域在  $[50, 60]$ 。且可行域内的最小值在  $x = 50, p^* = 30$ , 即可行域的边缘上, 如图 2.3。虽然  $f_0$  简单到可以被看作是一个不需要可行域也能解出有效最优解的目标函数。然而, 令  $\lambda_1 = 0.5, \lambda_2 = 0$ , 却可以取到  $\{x|x = 50, x \in [70, +\infty)\}$  均满足  $d^* = 30$ , 于是, 这成

为一个反例, 因为满足  $\lambda_1 = 0.5, \lambda_2 = 0$  解到的最优解  $[70, +\infty)$  有无数个, 且全部都不在可行域内<sup>2</sup>。



(a)  $\mathcal{L}(x_p) = \mathcal{L}(x_q)$  时, 反例在  $\lambda_1 = 0, \lambda_2 = 0$

(b) 反例的动态演示

图 2.4: 强对偶条件下对偶解推出原问题最优解的的反例 2。

$\mathcal{L}(x_p) = \mathcal{L}(x_q)$  时, 一定可以成为充分条件的反例。因为任何满足该条件的目标函数在  $\lambda$  全部为 0 时, 均可得到  $d^* = \mathcal{L}(x_q), p^* = \mathcal{L}(x_p)$ , 此时显然  $x_p, x_q$  都是**对偶问题**的最优解, 但是  $x_q$  却不在原问题的可行域内。图 2.4 图示了这种反例, 只需要微调目标函数  $f_0$ , 在  $x = 1$  设立一个最小值恰好为 30 即可。此时  $\lambda_1 = 0, \lambda_2 = 0$  构成一个反例。对偶解在  $\{1, 50\}$ , 而显然  $x = 1$  不在可行域内。

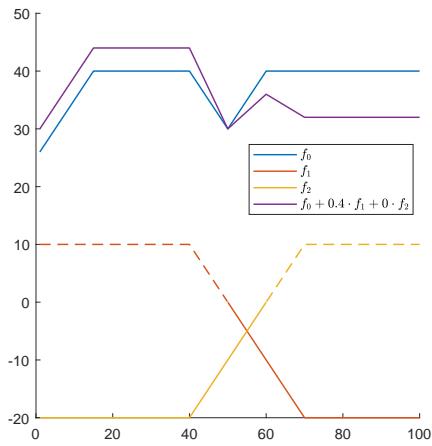
$\mathcal{L}(x_p) > \mathcal{L}(x_q)$  时, 这往往是需要求解**对偶问题**的场景。然而, 这一条件下同样可以设计出反例, 而且很容易设计。仍然只考虑  $\lambda_2 = 0$  的场合, 这样的反例往往满足(但不一定必须满足)以下条件:

- (1) 可行域内的最小值解  $x_p$  在可行域的边缘(亦即至少存在一个限制条件, 令  $f_i(x_p) = 0$ );
- (2) 在  $f_i(x_p) = 0$  对应的  $\lambda_i$  增大的过程中, 对  $\forall x \in \mathcal{D}$ , 满足  $0 \leq \lambda_i \leq \frac{f_0(x_p) - f_0(x)}{f_i(x)}$ , (可定义  $\bar{\lambda}_i = \min_x \frac{f_0(x_p) - f_0(x)}{f_i(x)}$ ) 亦即可行域内的最小值在  $\lambda$  足够小的范围内保持不变;
- (3) 可行域外的最小值  $x_d$  对应的同一个条件  $f_i(x_d)$  是个很大的值, 容易取到  $\lambda_i^* = \frac{f_0(x_p) - f_0(x_d)}{f_i(x_d)} \leq \bar{\lambda}_i$ 。

图 2.5 图示了在这一目标下设计的反例,  $\lambda_1 = 0.4, \lambda_2 = 0$  时,  $\mathcal{L}$  第一次对  $\lambda_1$  达到最大值  $d^* = 30$ , 然而这一情况下  $\{1, 50\}$  均是**对偶问题**的解, 且  $x = 1$  不在可行域内, 构成反例。

综上, 在各种情况下, 都存在反例, 使得由**强对偶条件**不能反推出**对偶问题**的解与原问题的解等价。但是观察这些反例, 可以发现, 造成反例的原因都是由于**对偶问题**解出了不止一个最优解。虽然存在不在可行域内的对偶解, 但在可行域的最优解仍然位列这些对偶解之中。由此, 我们可以得到一个更弱的推论。

<sup>2</sup>不在可行域内的原因: 可以发现, 如果只看  $f_1$  条件, 其实这些解是在  $f_1$  决定的区域内的; 然而由于  $f_2$  限定了上界, 这些解就不在可行域中。

(a)  $\mathcal{L}(x_p) > \mathcal{L}(x_q)$  时, 反例在  $\lambda_1 = 0, \lambda_2 = 0$ 

(b) 反例的动态演示

图 2.5: 强对偶条件下对偶解推出原问题最优解的反例 3。

**引理 2.4.5 – 对偶问题的解与原问题解的关系.** 在强对偶条件 ( $d^* = p^*$ ) 下, 设原问题有唯一的最优解  $x_p$ , 其对应的对偶问题最优解存在, 构成集合  $\mathbb{X}_d$ , 则  $x_p \in \mathbb{X}_d$ 。

证明引理 2.4.5 的过程较为简单。设  $\forall x \in \mathbb{X}_d$  满足  $\mathcal{L}(x, \boldsymbol{\lambda}_x) = d^*$ 。考虑可行域  $\mathcal{D}$  内, 由于  $\forall i, \lambda_i \geq 0, f_i(x) \leq 0, \lambda_i$  的任意增量  $\Delta \lambda_i$  都会导致  $\mathcal{L}$  减小。继而, 因为  $\mathcal{L}(x_p, 0) = p^*$ , 且可行域  $\mathcal{D}$  内  $\mathcal{L}$  对任意  $\boldsymbol{\lambda}$  都是单减的, 则有

- (1) 若在该组  $\boldsymbol{\lambda}_x$  下,  $\mathcal{L}(x_p, \boldsymbol{\lambda}_x) < p^*$ : 则存在  $x_p$  令  $\mathcal{L}(x_p, \boldsymbol{\lambda}_x) < \mathcal{L}_{x \in \mathbb{X}_d}(x, \boldsymbol{\lambda}_x)$ , 此时  $x_p$  取到了更小的拉格朗日函数值, 与  $\mathbb{X}_d$  在该组  $\boldsymbol{\lambda}_x$  目标函数值最小的条件矛盾, 故不存在这样的情况;
- (2) 若在该组  $\boldsymbol{\lambda}_x$  下,  $\mathcal{L}(x_p, \boldsymbol{\lambda}_x) > p^*$ : 显然这不可能成立, 因为在可行域内,  $\mathcal{L}$  在  $x_p$  固定时, 一定是单减的, 故不存在这样的情况;
- (3) 若在该组  $\boldsymbol{\lambda}_x$  下,  $\mathcal{L}(x_p, \boldsymbol{\lambda}_x) = p^*$ : 显然  $\mathcal{L}(x_p, \boldsymbol{\lambda}_x) = d^*$ , 于是  $x_p \in \mathbb{X}_d$ 。

综上,  $x_p \in \mathbb{X}_d$  一定成立。

由上述推论进一步可以得到的结论是,

**定理 2.4.3 – 限定条件下, 强对偶是对偶问题在可行域中的充分条件.** 对具有唯一最小值的原问题, 若强对偶条件成立, 且只能取到唯一的一组  $\boldsymbol{\lambda}$  和唯一一个对应的对偶问题最优解, 则该最优解一定是原问题的最优解。

结合定理 2.4.2 和定理 2.4.3, 可以确信的是考察强对偶条件对我们能否用对偶问题代替原问题具有重要的指导意义。事实上, 下面将介绍的, 是许多围绕这一性质的拓展工作, 它们共同的特点是, 通过证明在某些限定条件下强对偶成立, 从而间接证明对偶问题的解等价于原问题的解。

### 2.4.3 Slater 条件