

大学教材

优化选讲

董云达著

内 容 简 介

本书是作者在多年教学讲义的基础上编写而成的。它包含了优化的基本理论和方法。

首先，介绍了无约束优化方法：最陡下降法，共轭方向法，Newton-Raphson 方法，Davidon 变尺度法（也称拟 Newton 法）等基本的迭代方法。然后，又详细地介绍了约束优化的 Karush-Kuhn-Tucker 定理和相应的迭代方法：增广 Lagrange 乘子法，原始-对偶内点法，线性规划的原始对偶内点法，大规模半定规划的 Douglas-Rachford 方法等。书中的带*部分，可以跳过不讲。

本书的特色不求杂而全，只求取材精要，再进行透彻地论述。绝大部分是经典理论，也有少部分的最新结果。

作为大学教材，本书适合于数学专业的高年级本科生、研究生，也是广大教师和科研人员了解优化、研究优化的好帮手。

目录

第一章 绪论	1
1.1 优化问题的几个方面	3
1.2 Taylor 定理以及复合函数的求梯度法则	5
1.3 无约束优化问题的最优性条件	7
1.4 几个常用的假设	9
1.5 下降方向	11
1.6 精确线搜索	11
1.7 不精确线搜索	13
1.8 Kantorovich 不等式	17
1.9 Sherman-Morrison 公式	18
1.10 无约束优化方法的一般框架以及评价标准	19
第二章 最陡下降法	21
2.1 方法的导出	21
2.2 收敛性	23
2.3 收敛率	25
2.4 执行细节	27
2.5 克服锯齿现象的一种方法	27
2.6 在深度学习方面的一个应用	29
第三章 共轭梯度法	31
3.1 所研究的问题	31
3.2 共轭方向法的一般描述	31
3.3 线性共轭梯度法	33
3.4 收敛率	35
3.5 预处理	37
3.6 非对称情形	39

3.7	非线性情形	39
3.8	实用的 PR' 方法	45
3.9	一个密切相关的方法	47
第四章	Newton-Raphson 方法	49
4.1	方法的导出	49
4.2	收敛性和收敛率	51
4.3	一个实用形式	53
4.4	自协调函数	54
4.5	Nesterov-Nemirovski 方法	56
4.6	一个有趣的例子	60
第五章	Davidon 变尺度方法	63
5.1	割线方程与 Davidon 方法	63
5.2	对称秩二校正公式	64
5.3	DFP 方法	67
5.4	BFGS 方法	69
5.5	BFGS 方法的收敛性	73
5.6	BFGS 方法的超线性收敛性*	78
5.7	Perry 方法和 Perry-Shanno 方法	82
5.8	对称秩一校正公式	84
5.9	附 I: 一个数值例子	86
5.10	附 II: Fredholm 第二类型积分方程	87
第六章	Marquardt 方法	89
6.1	最小二乘问题	89
6.2	Gauß-Newton 方法	90
6.3	Levenberg-Marquardt 方法	91
6.4	收敛性分析	95
6.5	子问题的解法	96
6.6	另一重要形式	98
6.7	非线性方程组的 Broyden 方法	99
6.8	最小二乘求解器	100
第七章	约束优化的基本理论	103
7.1	一些基本概念	103
7.2	Fritz John 条件	104
7.3	约束限制与 KKT 定理	109

7.4	凸规划的最优性条件	112
7.5	二阶充分条件	115
7.6	KKT 系统的进一步讨论	117
7.7	Motzkin 定理的补充证明	119
第八章	增广 Lagrange 乘子法	123
8.1	Lagrange 乘子法	123
8.2	二次惩罚函数法	127
8.3	增广 Lagrange 乘子法	131
8.4	一般约束下的增广 Lagrange 乘子法	135
8.5	Debreu 引理的补充证明	136
第九章	原始对偶内点法	139
9.1	对数障碍法及其收敛性	139
9.2	与对数障碍法相关的外推技术	145
9.3	原始对偶内点法	146
第十章	线性规划: 原始对偶内点法	149
10.1	历史背景	149
10.2	中心路径	150
10.3	原始、对偶和原始对偶内点法	154
10.4	Mehrotra 预估校正算法	155
10.5	Salahi 预估校正算法	158
第十一章	二次规划	165
11.1	问题的提出	165
11.2	扰动的 KKT 条件和对偶理论	166
11.3	关于 H 的正半定性	167
11.4	凸二次规划的积极集法	168
11.5	凸二次规划的内点法	171
11.6	特殊非凸二次规划的最陡下降法	172
11.7	解线性规划的逐步二次规划方法	173
11.8	附: 机器学习中的支撑向量	174
第十二章	凸集与凸函数	179
12.1	问题的提出	179
12.2	凸集的定义、相对内部和凸锥	180
12.3	凸集分离定理、闭凸集外表示	182

12.4 凸函数的有效域、上图和闭性	185
12.5 连续可微函数的凸性判定	188
12.6 凸函数之间的运算	189
12.7 方向导数	192
12.8 次梯度和次微分	194
12.9 次微分的有效域的稠密性	200
12.10 凸优化的最优性条件	201
12.11 初识凸函数的一阶逼近和二阶逼近	203
12.12 Fenchel 共轭	204
12.13 附 I: 线性广义梯度	210
第十三章 Hilbert 空间简介	213
13.1 赋范线性空间和内积空间	213
13.2 Banach 空间和 Hilbert 空间	215
13.3 正交投影与正交分解	217
13.4 线性算子和线性泛函	218
13.5 有界线性泛函的表示定理	221
13.6 对偶	222
13.7 一致有界性原则与共鸣定理	223
13.8 弱收敛与弱*收敛	224
13.9 伴算子和自伴算子	227
13.10 正算子	230
13.11 Fredholm 第二型积分方程的离散化	233
第十四章 单调算子和邻近点方法	235
14.1 问题的提出	235
14.2 凸函数次微分的极大单调性	236
14.3 极大单调算子和 Minty 定理	238
14.4 邻近点方法	243
14.5 附: Yosida 逼近	247
第十五章 算子分裂方法	253
15.1 问题提出	253
15.2 基本的算子分裂方法	254
15.3 收敛性分析	255
15.4 算法 15.1 的若干特例	260
15.4.1 第一个特例	260

15.4.2 第二个特例	261
15.5 一元函数 $ x $ 的邻近映射	263
第十六章 半定规划	265
16.1 标准形式与对偶	265
16.2 原始对偶内点法	267
16.3 半正定锥上的投影	268
16.4 增广 Lagrange 乘子法	271
16.5 Douglas-Rachford 分裂方法	272
16.6 收敛性分析	273
16.7 收敛率	275
16.8 一个实用的算法	277
16.9 半定规划的对数行列式函数	280
16.10 其它	281
16.10.1 非凸二次规划	281
16.10.2 在超椭球体上极小化线性函数	282
16.10.3 凸二次半定规划简介	283
16.11 附 I: 半定锥的示性函数	283
附录 A 总结	287
附录 B 线性代数的一些基本知识	289
附录 C 强 Wolfe 条件的执行细节	293
附录 D 割线法	295
附录 E Dong 条件的 Matlab 程序	297
附录 F 降维法和 Nelder-Mead 直接法	299
附录 G Matlab 中的导数符号运算和 fmincon 简介	301
附录 H 算法 3.4 的 Matlab 程序	303
附录 I Hölder 不等式及其特例	305
附录 J Nyström 近似	307
附录 K 测试函数集	309

索 引

318

后 记

319

经计算

$$\varphi(\alpha) = \alpha^3 - 3.9996\alpha, \quad \hat{\varphi}(\alpha) = 3\alpha^2 + 3.6\alpha - 4.$$

它们的图像如下所示, 其中红色区间 $[0.8907, 1.9999]$ 为 α 的可接受范围。

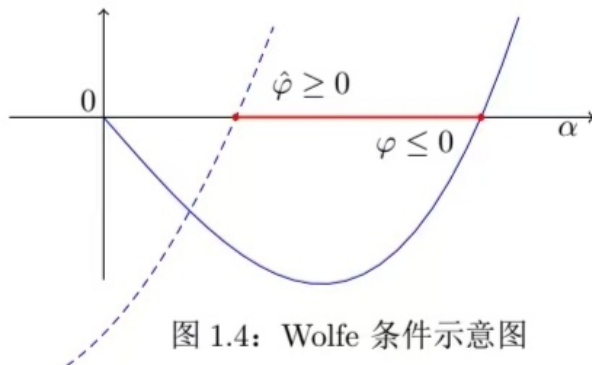


图 1.4: Wolfe 条件示意图

Wolfe 条件有一个加强的形式, 即下面的强 Wolfe 条件:

$$f(x + \alpha d) \leq f(x) + c_1 \alpha \nabla f(x)^T d, \quad (1.13)$$

$$|\nabla f(x + \alpha d)^T d| \leq -c_2 \nabla f(x)^T d, \quad (1.14)$$

其中, $0 < c_1 < c_2 < 1$ 。

从上面的讨论, 我们可以看出: 只要 d 是 f 在 x 处的一个下降方向, 那么无论 α 是由上述哪一个步长条件确定的, 总有 $f(x + \alpha d) < f(x)$ 。

接下来, 我们将考虑一个仅仅包含梯度信息的步长准则 — Dong 条件。假设 $f: R^n \rightarrow R$ 是连续可微凸的, $x, d \in R^n$, 则由 (1.6) 可得

$$(\nabla f(x + \alpha d) - \nabla f(x + td))^T (x + \alpha d - (x + td)) \geq 0.$$

当 $t \leq \alpha$ 时, 我们就有

$$(\nabla f(x + \alpha d) - \nabla f(x + td))^T d \geq 0.$$

结合

$$f(x + \alpha d) = f(x) + \int_0^\alpha d f(x + td) = f(x) + \int_0^\alpha \nabla f(x + td)^T d dt,$$

可以推导出

$$f(x + \alpha d) \leq f(x) + \alpha \nabla f(x + \alpha d)^T d, \quad \forall \alpha > 0. \quad (1.15)$$

于是, 我们就有了下面确定步长的技巧 — Dong 条件:

$$c_2 \nabla f(x)^T d \leq \nabla f(x + \alpha d)^T d \leq c_1 \nabla f(x)^T d, \quad (1.16)$$

其中 $0 < c_1 < c_2 < 1$ 。

接下来的问题是：如何选取步长 α 呢？Cauchy 建议使用下面的精确线搜索

$$\min \{f(x - \alpha \nabla f(x)) : \alpha \geq 0\}$$

来确定步长。于是我们就得到了下面的最陡下降法。

算法 2.1 最陡下降法 (steepest descent method)

0. 任取 $x^0 \in R^n, \varepsilon > 0$ 。令 $k := 0$ 。

1. 计算 $\nabla f(x^k)$ 。若 $\|\nabla f(x^k)\| \leq \varepsilon$ ，则算法停止。否则，计算 α_k

$$f(x^k - \alpha_k \nabla f(x^k)) = \min_{\alpha \geq 0} f(x^k - \alpha \nabla f(x^k)). \quad (2.2)$$

2. 计算下一个迭代点 $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$ 。令 $k := k + 1$ 。

注意：由 (2.2) 所确定的步长往往被称之为最优步长，通常取作

$$\alpha_k = \min \{\alpha : \nabla f(x^k - \alpha \nabla f(x^k))^T \nabla f(x^k) = 0, \alpha \geq 0\}. \quad (2.3)$$

注意：对于精确线搜索下的最陡下降法来说，即相邻的两个迭代点处得负梯度方向是垂直的： $\nabla f(x^{k+1})^T \nabla f(x^k) = 0$ 。于是，一旦靠近解点或者迭代点时必须通过狭长并且弯曲的“山谷”时，迭代过程会变得十分缓慢了。最著名的测试函数 – Rosenbrock 函数

$$f(x) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2, \quad x^0 = (-1.2, 1)^T$$

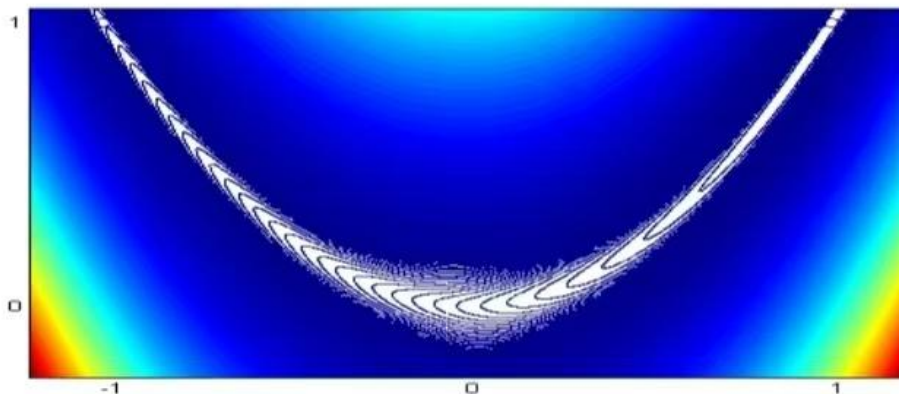


图2.1: Rosenbrock 函数等高线

可以较好说明这一点，因为从指定的初始点到解点 $x^* = (1, 1)^T$ ，必须经过一条香蕉状的“山谷”，所以人们也形象地称之为香蕉函数。沿着这条“山谷”，从 x^0 出发，经过 0 附近，最终逼近或到达解点，其函数值依次为 24.2, 1, 0。

算法 3.5 FR 方法

0. 选取 $x^0 \in R^n$, 计算 $d^0 = -g^0$ 。 $k := 0$ 。
1. 利用某种方式确定步长 α_k 。 计算 $x^{k+1} = x^k + \alpha_k d^k$ 。
2. 计算 $g^k = g(x^k)$ 以及

$$\beta_{k+1} = \|g^{k+1}\|^2 / \|g^k\|^2, \quad (3.16)$$

$$d^{k+1} = -g^{k+1} + \beta_{k+1} d^k. \quad (3.17)$$

令 $k := k + 1$ 。

在 1969 年, Polak 和 Ribière 没有采用 (3.16) 的取法而是取

$$\beta_{k+1} = \frac{(g^{k+1} - g^k)^T g^{k+1}}{\|g^k\|^2}. \quad (3.18)$$

接下来, 我们验证: 在精确线搜索下, 当 (3.15) 中的 f 为严格凸二次函数时, (3.16), (3.18) 与线性共轭梯度法中的共轭参数

$$\beta_{k+1} = \frac{(g^{k+1})^T A d^k}{(d^k)^T A d^k}$$

是等价的。实际上, 结合

$$A x^{k+1} - b = A x^k - b + \alpha_k A d^k \quad \Leftrightarrow \quad g^{k+1} = g^k + \alpha_k A d^k,$$

其中 $g^k = A x^k - b$, 我们有

$$\beta_{k+1} = \frac{(g^{k+1})^T A d^k}{(d^k)^T A d^k} = \frac{(g^{k+1})^T (g^{k+1} - g^k)}{(d^k)^T (g^{k+1} - g^k)}.$$

而在精确线搜索下呢, 最右端分母满足

$$(d^k)^T (g^{k+1} - g^k) = -(g^k)^T d^k = -(g^k)^T (-g^k + \beta_k d^{k-1}) = \|g^k\|^2.$$

这就证明了与 (3.18) 有关的等价性。与 (3.16) 有关的等价性, 再结合

$$\begin{aligned} (g^{k+1})^T g^k &= (g^{k+1})^T (-d^k + \beta_k d^{k-1}) \\ &= -(g^{k+1})^T d^k + \beta_k (g^{k+1})^T d^{k-1} \end{aligned}$$

和定理 3.2.1 (b) 即可。从这个意义上讲, 在共轭参数的选取方式上, (3.16), (3.18) 可以分别看作线性情形的一个推广。

在 1985 年, Al-Baali 利用下面的强 Wolfe 条件

$$f(x^k + \alpha_k d^k) \leq f(x^k) + c_1 \alpha_k (g^k)^T d^k, \quad (3.19)$$

$$|g(x^k + \alpha_k d^k)^T d^k| \leq -c_2 (g^k)^T d^k, \quad (3.20)$$

结合引理 4.4.2, 可以断定: $x^+ \in \text{int}X$. 接下来, 我们证明第二个结论. 暂时约定 $\delta = \delta(f, x)$. 于是, 我们就有

$$\delta = \sqrt{d_N^T \nabla^2 f(x) d_N}, \quad \nabla f(x)^T d_N = -\delta^2.$$

再利用, 便可以知道

$$\begin{aligned} f(x + \frac{1}{1+\delta} d_N) &\leq f(x) + \frac{1}{1+\delta} \nabla f(x)^T d_N - \frac{1}{1+\delta} \sqrt{d_N^T \nabla^2 f(x) d_N} \\ &\quad - \ln \left(1 - \frac{1}{1+\delta} \sqrt{d_N^T \nabla^2 f(x) d_N} \right) \\ &= f(x) - \frac{\delta^2}{1+\delta} - \frac{\delta}{1+\delta} - \ln \left(1 - \frac{\delta}{1+\delta} \right) \\ &= f(x) - \delta + \ln(1 - \delta). \end{aligned}$$

从而, 结论成立. □

定理 4.5.2 如果 $x \in \text{int}X$, 那么下面的不等式成立

$$\delta(f, x^+) \leq 2\delta^2(f, x).$$

证 对于所有的 $h \in R^n$, 定义 $\psi(t) = \nabla f(x + td_N)^T h$. 当 $0 \leq t \leq 1/(1+\delta)$ 时, 它是二次连续可微的, 并且

$$\psi'(t) = d_N^T \nabla^2 f(x + td_N) h, \quad \psi''(t) = \nabla^3 f(x + td_N)[h, d_N, d_N].$$

则

$$\begin{aligned} |\psi''(t)| &= |\nabla^3 f(x + td_N)[h, d_N, d_N]| \\ &\leq 2\sqrt{h^T \nabla^2 f(x + td_N) h} d_N^T \nabla^2 f(x + td_N) d_N. \end{aligned}$$

结合引理 4.4.2 中的不等式 (iii), 进一步可以知道

$$\begin{aligned} |\psi''(t)| &\leq 2(1 - t\delta)^{-3} \sqrt{h^T \nabla^2 f(x) h} d_N^T \nabla^2 f(x) d_N \\ &= 2\sqrt{h^T \nabla^2 f(x) h} \delta^2 (1 - t\delta)^{-3}. \end{aligned}$$

当然, 这也意味着

$$\psi''(t) \leq 2\sqrt{h^T \nabla^2 f(x) h} \delta^2 (1 - t\delta)^{-3}.$$

积分两次, $\psi(\frac{1}{1+\delta})$ 不会超过

$$\psi(0) + \frac{1}{1+\delta} \psi'(0) + \sqrt{h^T \nabla^2 f(x) h} \int_0^{\frac{1}{1+\delta}} \int_0^t 2\delta^2 (1 - \tau\delta)^{-3} d\tau dt,$$

将 (5.22) 写成

$$\tilde{M} = M(I - \frac{ss^T M}{s^T M s} + \frac{M^{-1}yy^T}{s^T y}).$$

两边取行列式，并且利用引理 5.5.1

$$\det(I + uv^T + wz^T) = (1 + v^T u)(1 + z^T w) - (u^T z)(v^T w)$$

(该公式的推导要求 $1 + v^T u$ 和 $1 + z^T w$ 不同时为 0，这儿的假设条件正好使之满足)，则有

$$\det(\tilde{M}) = \det(M) \frac{s^T y}{s^T M s} = \det(M) \frac{s^T y}{\|s\|^2} \frac{1}{q}.$$

两边取对数

$$\ln \det(\tilde{M}) = \ln \det(M) - \ln q + \ln \frac{s^T y}{\|s\|^2}. \quad (5.24)$$

将 (5.23) 减去 (5.24)，并且利用 $\varphi(\cdot) = \text{trace}(\cdot) - \ln \det(\cdot)$ ，整理可以得到

$$\varphi(\tilde{M}) = \varphi(M) + 1 - \frac{q}{\cos^2 \theta} + \ln \frac{q}{\cos^2 \theta} + \ln \cos^2 \theta + \frac{\|y\|^2}{s^T y} - \ln \frac{s^T y}{\|s\|^2} - 1.$$

既然对于所有的 $t > 0$ 总有 $1 - t + \ln t \leq 0$ 以及根据假设条件 (5.19)

$$\frac{s^T y}{\|s\|^2} \geq \mu, \quad \frac{\|y\|^2}{s^T y} \leq \bar{\mu}, \quad (5.25)$$

那么我们就有: $c := \bar{\mu} - \ln \mu - 1 > 0$ 以及

$$\varphi(\tilde{M}) \leq \varphi(M) + \ln \cos^2 \theta + c.$$

重新写出上、下标，则有

$$\varphi(M_{k+1}) \leq \varphi(M_k) + \ln \cos^2 \theta_k + c \leq \varphi(M_0) + \sum_{j=0}^k \ln \cos^2 \theta_j + (k+1)c.$$

由于当 M 正定时，

$$\varphi(M) = \text{trace}(M) - \ln \det(M) = \sum (\lambda_i - \ln \lambda_i) > 0,$$

其中 λ_i 为 M 的特征值，所以，我们可以推导出

$$k \min_{j=0, \dots, k-1} \{-\ln \cos^2 \theta_j\} \leq \sum_{j=0}^{k-1} (-\ln \cos^2 \theta_j) < \varphi(M_0) + kc.$$

这个不等式表明，当 $k > \varphi(M_0)$ 时，我们有

$$\min_{j=0, \dots, k-1} \{-\ln \cos^2 \theta_j\} < \varphi(M_0)/k + c < 1 + c.$$

Levenberg-Marquardt 方法的好处在于：一是使子问题 (6.6) 仅有一个极小点；二是该子问题比 (6.4) 更容易求解；三是有较好的收敛性。

有趣的是， $d^k = d^k(\mu_k)$ 既是下降方向，同时又包含了步长信息。显然，就这一点而言，它是与最速下降法中的搜索方向截然不同。对于参数 $\bar{\mu}_k$ 的选取，也值得细心对待。当 ρ_k 接近于 1 时，说明线性逼近效果较好，可以在下一步使该参数尽可能地变小，从而 (6.7) 中的系数矩阵就可以包含问题数据本身的更多信息了。

记 $d(\mu) = -(J^T J + \mu I)^{-1} J^T r$ 。下面讨论，当 μ 变化时，其长度和方向的基本性质。

引理 6.3.1 当 $\mu \uparrow +\infty$ 时， $\|d(\mu)\| \downarrow 0$ 。

证 当 $\mu > 0$ 时， $J^T J + \mu I$ 是正定的，从而存在正交阵 U 使得

$$(J^T J + \mu I)^{-1} = U^T \text{diag}(1/(\lambda_1 + \mu), \dots, 1/(\lambda_n + \mu)) U,$$

其中 $\lambda_i, i = 1, \dots, n$ ，为矩阵 $J^T J$ 的所有特征值。令 $U J^T r$ 的第 i 个分量为 v_i ，则有

$$\|d(\mu)\|^2 = \sum_{i=1}^n v_i^2 / (\lambda_i + \mu)^2.$$

证完。 □

类似地，我们可以证明下面的结论成立。

引理 6.3.2 负梯度方向 $-\nabla f = -J^T r$ 与 $d(\mu)$ 的夹角 $\theta(\mu)$ 关于 $\mu > 0$ 是递减的。

证 记号 $\sum_{i=1}^n$ 用 \sum 来代替。当 $\mu > 0$ 时，对称矩阵 $J^T J + \mu I$ 是正定的，从而存在正交阵 U 使得

$$(J^T J + \mu I)^{-1} = U^T \text{diag}(1/(\lambda_1 + \mu), \dots, 1/(\lambda_n + \mu)) U,$$

其中 $\lambda_i, i = 1, \dots, n$ 为矩阵 $J^T J$ 的所有特征值。令 $U J^T r$ 的第 i 个分量为 v_i ，则负梯度方向 $-J^T r$ 与 $d(\mu) = -(J^T J + \mu I)^{-1} J^T r$ 的夹角 $\theta(\mu)$ 满足

$$\cos \theta(\mu) = \frac{(J^T r)^T (J^T J + \mu I)^{-1} J^T r}{\|J^T r\| \|(J^T J + \mu I)^{-1} J^T r\|} = \frac{\sum v_i^2 / (\lambda_i + \mu)}{\|v\| (\sum v_i^2 / (\lambda_i + \mu)^2)^{1/2}}.$$

于是，我们可以算出 $\frac{d}{d\mu} \cos \theta(\mu)$ ，它是一个比式，分子为

$$\left(\sum \frac{v_i^2}{\lambda_i + \mu} \right) \left(\sum \frac{v_i^2}{(\lambda_i + \mu)^3} \right) - \left(\sum \frac{v_i^2}{(\lambda_i + \mu)^2} \right)^2$$

定理 7.7.2 (Tucker 定理) 设 $D \in R^{m \times n}$, 则不等式组

$$Dx \geq 0, D^T y = 0, y \geq 0$$

总存在一组解 (\bar{x}, \bar{y}) 使得 $D\bar{x} + \bar{y} > 0$ 。

证 记 $D^T = (d^1, d^2, \dots, d^m)$ 。构造相应的集合

$$X_1 = \left\{ \sum_{i \neq 1} -y_i d^i : y_i \geq 0, i \neq 1 \right\}. \quad (7.36)$$

分两种情形进行讨论。

情形一 $d^1 \in X_1$ 。此时, 必然存在 $y_i \geq 0, i = 2, \dots, m$ 使得

$$d^1 = \sum_{i \neq 1} -y_i d^i.$$

令 $x^1 = 0, y^1 = (1, y_2, \dots, y_m)^T$, 则 (x^1, y^1) 是不等式组的解, 并且 $Dx^1 + y^1$ 的第一个分量为正。

情形二 $d^1 \notin X_1$ 。考虑到 X_1 是一个非空闭凸锥, 用点到闭凸锥分离定理, 可以知道: 存在一个 $p \in R^n$ 使得

$$p^T d^1 < 0 \leq p^T x, \quad \forall x \in X_1.$$

既然 $-d^i, i = 2, \dots, m$ 都属于 X_1 , 那么

$$(d^i)^T (-p) = p^T (-d^i) \geq 0, \quad i = 2, \dots, m.$$

另外, 我们也有 $(d^1)^T (-p) > 0$. 令 $x^1 = -p, y^1 = 0$, 则 (x^1, y^1) 是不等式组的解, 并且 $Dx^1 + y^1$ 的第一个分量为正。

类似(7.36), 我们构造相应的集合

$$X_2 = \left\{ \sum_{i \neq 2} -y_i d^i : y_i \geq 0, i \neq 2 \right\}.$$

然后, 分 $d^2 \in X_2$ 和 $d^2 \notin X_2$ 两种情形进行讨论, 可以得到 (x^2, y^2) , 它是不等式组的解, 并且 $Dx^2 + y^2$ 的第二个分量为正。

依此类推, 我们断言: 对于每个 i , 都存在相应的 (x^i, y^i) 使得

$$Dx^i \geq 0, D^T y^i = 0, y^i \geq 0, Dx^i + y^i \text{ 的第 } i \text{ 个分量为正}.$$

令 $\bar{x} = \sum_{i=1}^m x^i, \bar{y} = \sum_{i=1}^m y^i$, 则 (\bar{x}, \bar{y}) 是不等式组的解, 并且

$$D\bar{x} + \bar{y} = \sum_{i=1}^m (Dx^i + y^i).$$

由于 $Dx^i + y^i$ 的每个分量都是非负的以及它的第 i 个分量为正, 从而 $D\bar{x} + \bar{y}$ 的每个分量都是正的。□

其中 $\mu > 0, \rho > 0$ 。由于一元函数 $(\max\{0, t\})^2$ 在 $t = 0$ 处的左导数和右导数相等, 从而它是处处可导的, 并且

$$\frac{d}{dt}(\max\{0, t\})^2 = 2\max\{0, t\}, \quad (8.11)$$

所以, 这个惩罚函数 $p(x, \mu, \rho)$ 关于 x 的梯度为

$$\nabla f(x) - \mu\rho \sum_{i=1}^l \max\{0, -g_i(x)\} \nabla g_i(x) + \mu \sum_{j=1}^m h_j(x) \nabla h_j(x).$$

下面, 我们给出一般约束下的二次惩罚函数法。

算法 8.3 二次惩罚函数法

0. 选取初始惩罚因子 $\mu_0 > 0, \rho_0 > 0$ 。 $k := 0$ 。
1. 计算子问题 $\min p(x, \mu_k, \rho_k)$ 的全局极小点, 然后, 作为 x^k 。
2. 选取 $\mu_{k+1} > \mu_k$ 。 计算

$$c_k := \sum_{i=1}^l (\max\{0, -g_i(x^k)\})^2, \quad \hat{c}_k := \sum_{j=1}^m h_j^2(x^k).$$

若 $c_k > 2\hat{c}_k$, 则 $\rho_{k+1} > \rho_k$ 。 否则, $\rho_{k+1} = \rho_k$ 。 令 $k := k + 1$ 。

对于二次惩罚函数法, 我们刚才给出了自适应地校正 ρ_k 的策略: 在迭代的第 k 步, 首先, 计算出第 k 个迭代点违反不等式的程度 c_k 以及违反等式约束的程度 \hat{c}_k 。如果前者较大, 那么我们应该 (在下次迭代中) 加大 ρ 的值以便加大相应的惩罚因子 $\rho\mu$ 。否则, 保持不变。从某种意义上讲, 它在两者之间起了协调、平衡的作用。

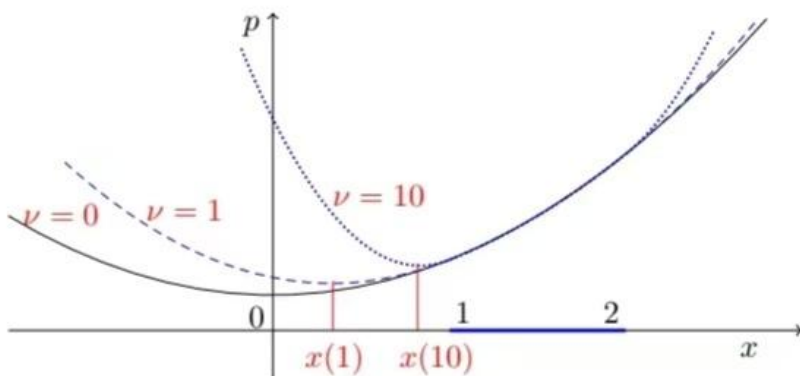


图 8.1: 二次惩罚函数法收敛行为示意图

例子 8.2.2 考察下面的一维极小化问题

$$\min f(x) = x^2 + 1, \quad \text{s.t. } x - 1 \geq 0, 2 - x \geq 0.$$

其中 z 称为 Lagrange 乘子或对偶变量。一阶最优性条件为

$$F(x, \lambda, z) := \begin{pmatrix} \nabla f(x) - \sum_{i=1}^l z_i \nabla g_i(x) \\ -\mu e + \Lambda Z e \\ g(x) - \lambda \end{pmatrix} = 0, \quad \lambda > 0, \quad z > 0, \quad (9.7)$$

其中 $e = (1, \dots, 1)^T$ 以及

$$\Lambda := \text{diag}(\lambda_1, \dots, \lambda_l), \quad Z := \text{diag}(z_1, \dots, z_l), \quad g = (g_1, \dots, g_l)^T.$$

注意：在上面这个最优性条件中，除了 $F(x, \lambda, z) = 0$ ，还包括 $\lambda > 0$ 以及 $z > 0$ 。要求 $\lambda > 0$ 是显然的，只有这样，相应的对数内点函数才有意义。再根据 $-\mu e + \Lambda Z e = 0$ ，就必须做出 $z > 0$ 这样的要求了。

已知 (x^k, λ^k, z^k) 以及 $\lambda^k > 0, z^k > 0$ 。找 $\Delta x^k, \Delta \lambda^k, \Delta z^k$ ，使得

$$F(x^k, \lambda^k, z^k) + F'(x^k, \lambda^k, z^k) \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \\ \Delta z^k \end{pmatrix} = 0, \quad (9.8)$$

其中

$$F'(x^k, \lambda^k, z^k) = \begin{pmatrix} M_k & 0 & -J_k^T \\ 0 & Z_k & \Lambda_k \\ J_k & -I & 0 \end{pmatrix}, \quad M_k := \nabla^2 f(x^k) - \sum_{i=1}^l z_i^k \nabla^2 g_i(x^k)$$

为 F 在 (x^k, λ^k, z^k) 的 Jacobi 矩阵而 J_k 是 g 在 x^k 处的 Jacobi 矩阵。线性方程组 (9.8) 称为原始对偶系统。

接下来，我们考虑 (9.8) 的一个等价形式

$$\begin{aligned} M_k \Delta x^k - J_k^T \Delta z^k &= -\nabla f(x^k) + \sum z_i^k \nabla g_i(x^k), \\ Z_k \Delta \lambda^k + \Lambda_k \Delta z^k &= \mu e - \Lambda_k Z_k e, \end{aligned} \quad (9.9)$$

$$J_k \Delta x^k - \Delta \lambda^k = \lambda^k - g(x^k), \quad (9.10)$$

其中 \sum 表示 $\sum_{i=1}^l$ 。由于 Z_k 是可逆的，所以可以从 (9.9) 中解得 $\Delta \lambda^k$ 。然后，再将其代入 (9.10)。于是，我们就有

$$\begin{aligned} \begin{pmatrix} -M_k & J_k^T \\ J_k & Z_k^{-1} \Lambda_k \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \Delta z^k \end{pmatrix} &= \begin{pmatrix} \nabla f(x^k) - \sum z_i^k \nabla g_i(x^k) \\ \lambda^k - g(x^k) - \mu Z_k^{-1} e - \Lambda_k e \end{pmatrix}, \\ \Delta \lambda^k &= -Z_k^{-1} \Lambda_k \Delta z^k + \mu Z_k^{-1} e - \Lambda_k e. \end{aligned}$$

由于这个线性方程组的系数矩阵是对称的，所以一个常用的直接解法是 LDL^T 分解。具体地讲，先给出该系数矩阵的 LDL^T 分解，其中 L 是一个单位下三角矩阵

证 根据 (10.18), 我们有

$$\sqrt{\frac{z_i}{x_i}} \Delta x_i + \sqrt{\frac{x_i}{z_i}} \Delta z_i = \frac{\mu_{\min}}{\sqrt{x_i z_i}} - \sqrt{x_i z_i} - \frac{\Delta x_i^a \Delta z_i^a}{\sqrt{x_i z_i}}.$$

两边平方后, 对 $i = 1, \dots, n$ 求和, 右端为

$$\begin{aligned} & \sum_{i=1}^n \left(\frac{\mu_{\min}^2}{x_i z_i} + x_i z_i + \frac{(\Delta x_i^a \Delta z_i^a)^2}{x_i z_i} - 2\mu_{\min} - 2\mu_{\min} \frac{\Delta x_i^a \Delta z_i^a}{x_i z_i} + 2\Delta x_i^a \Delta z_i^a \right) \\ & \leq \frac{n\mu_{\min}^2}{\gamma t} + nt + \frac{1}{16}nt + \frac{1}{16\gamma}n^2t - 2n\mu_{\min} + \frac{n\mu_{\min}}{2\gamma}, \end{aligned}$$

其中的不等式用到了 $x_i z_i \geq \gamma t$, $(\Delta x^a)^T \Delta z^a = 0$ 以及引理 10.5.3。而两边平方后的左端为

$$\begin{aligned} & \sum_{i=1}^n \left(\frac{z_i}{x_i} (\Delta x_i)^2 + \frac{x_i}{z_i} (\Delta z_i)^2 \right) + 2\Delta x^T \Delta z \\ & = \sum_{i=1}^n \left(\frac{z_i}{x_i} (\Delta x_i)^2 + \frac{x_i}{z_i} (\Delta z_i)^2 \right) \\ & \geq \sum_{i=1}^n 2|\Delta x_i \Delta z_i|. \end{aligned}$$

从而

$$\begin{aligned} \sum_{i=1}^n |\Delta x_i \Delta z_i| & \leq \frac{1}{2} \left(\frac{n\mu_{\min}^2}{\gamma t} + nt + \frac{1}{16}nt + \frac{1}{16\gamma}n^2t + \left(\frac{1}{2\gamma} - 2 \right) n\mu_{\min} \right) \\ & \leq n^2 t \gamma^{-1} \left(\frac{1}{32} + \frac{25\gamma}{32n} - \frac{\gamma^2}{2n} \right) \\ & < 0.1 n^2 t \gamma^{-1}, \end{aligned}$$

其中最后的不等式用到了引理 10.5.1 中 $\mu_{\min} \leq \gamma t$ 的结论。 \square

引入记号

$$(x(\alpha), \lambda(\alpha), z(\alpha)) := (x, \lambda, z) + \alpha(\Delta x, \Delta \lambda, \Delta z), \quad t(\alpha) := x(\alpha)^T z(\alpha)/n.$$

假设 10.5.1 正数 $\gamma < 0.01$ 以及正数 α 满足

$$\alpha < (2.5 - 5\gamma)\gamma^2/n^2.$$

引理 10.5.5 对于 Salah 预估校正算法来说, 若假设 10.5.1 成立, 则

$$(x(\alpha), \lambda(\alpha), z(\alpha)) \in \mathcal{N}_{\infty}^-(\gamma).$$

证 先梳理一下要证什么。(i) $(x(\alpha), \lambda(\alpha), z(\alpha))$ 仍然满足原始对偶线性约束条件。(ii) $x(\alpha) > 0, z(\alpha) > 0$ 。(iii) $x_i(\alpha)z_i(\alpha) - \gamma \frac{x(\alpha)^T z(\alpha)}{n} \geq 0, i = 1, \dots, n$ 。

11.7 解线性规划的逐步二次规划方法

对于标准的线性规划，它有多种数值方法。这儿，我们将要讨论的方法是，首先利用对数障碍函数法，将标准的线性规划转化为一个等式约束的极小化问题。并且写出相应的 Lagrange 函数。接下来，利用 Newton-Raphson 方法，解 Lagrange 函数所确定的驻点方程。最后，我们将说明，这样一个处理过程为什么等价于一个逐步二次规划方法。

考虑标准的线性规划

$$\min c^T x, \quad \text{s.t. } Ax = b, x \geq 0,$$

其中 $c \in R^n$, $b \in R^m$, $A \in R^{m \times n}$ 是行线性无关的。一个密切相关的问题是

$$\min f(x) := c^T x - \mu \sum_{i=1}^n \ln x_i, \quad \text{s.t. } Ax = b, \quad (11.12)$$

其中 $\mu > 0$ 是一个障碍因子。由线性约束下优化问题的 KKT 定理，存在 $\lambda \in R^m$, x 为 Lagrange 函数 $f(x) - \lambda^T(Ax - b)$ 的驻点，即

$$F(x, \lambda) = \begin{pmatrix} \nabla f(x) - A^T \lambda \\ -Ax + b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

在 (x^k, λ^k) 处，相应的 Newton-Raphson 方向满足

$$J(x^k, \lambda^k) \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix} = -F(x^k, \lambda^k), \quad (11.13)$$

其中 J 为 F 在该点处 Jacobi 矩阵

$$J(x^k, \lambda^k) = \begin{pmatrix} \nabla^2 f(x^k) & -A^T \\ -A & 0 \end{pmatrix} = \begin{pmatrix} \mu X_k^{-2} & -A^T \\ -A & 0 \end{pmatrix}.$$

由 μX_k^{-2} 的正定性，该矩阵可逆。

引入辅助函数 $\phi(x, \lambda) := \|F(x, \lambda)\|^2$ ，则 $\nabla \phi(x, \lambda) = 2J(x, \lambda)^T F(x, \lambda)$ 。结合 (11.13)

$$\nabla \phi(x^k, \lambda^k)^T \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix} = -2\phi(x^k, \lambda^k).$$

这个等式表明，如果 (x^k, λ^k) 不是最优解，那么 $\begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix}$ 是 ϕ 在 (x^k, λ^k) 处的一个下降方向。由此，我们给出下面的迭代方法。

相加可以得到

$$\begin{aligned} f(y) + g(Qy - q) &\geq f(x) + g(Qx - q) + \langle s + Q^T \hat{s}, y - x \rangle \\ &= f(x) + g(Qx - q). \end{aligned}$$

这表明: x 为问题 (12.17) 的一个解点。由于这是一个凸优化, 从而该点也是一个全局极小点。 \square

特别地, 若 $g = \delta_C$, 其中 C 为一个非空闭凸集合, 则 (12.17) 变为

$$\min f(x), \quad \text{s.t. } Qx - q \in C. \quad (12.18)$$

接下来, 我们考虑凸优化问题的另一基本形式

$$\min f(x), \quad \text{s.t. } g_i(x) \leq 0, \quad i = 1, \dots, l, \quad (12.19)$$

其中 $f: E \rightarrow R$ 是一个闭的真凸函数, 每一个 $g_i: E \rightarrow R$ 都是连续可微凸函数。

定理 12.10.2 考虑问题 (12.19)。记 $g := \max\{g_1, \dots, g_l\}$ 。假设 $\text{ri dom } f \cap \text{ri dom } g \neq \emptyset$ 或者在每一个 g_i 都是分片线性的条件下假设 $\text{ri dom } f \cap \text{dom } g \neq \emptyset$ 。若 x^* 为该问题的一个解点, 则存在一组实数 $\lambda_0^*, \lambda_1^*, \dots, \lambda_l^*$ 使得

$$\lambda_0^* \geq 0, \lambda_1^* \geq 0, \dots, \lambda_l^* \geq 0, \quad \lambda_0^* + \lambda_1^* + \dots + \lambda_l^* = 1, \quad (12.20)$$

$$0 \in \lambda_0^* \partial f(x^*) + \sum_{i=1}^l \lambda_i^* \nabla g_i(x^*), \quad (12.21)$$

$$\lambda_i^* \geq 0, \quad g_i(x^*) \leq 0, \quad \lambda_i^* g_i(x^*) = 0, \quad i = 1, \dots, l. \quad (12.22)$$

证 考察下面的辅助函数

$$\varphi(x) = \max\{f(x) - f(x^*), g_1(x), \dots, g_l(x)\}.$$

显然, $\varphi(x)$ 在其有效域上总是非负的。同时, 定理 12.6.4 表明: φ 是一个闭的真凸函数。所以, x^* 是凸函数 $\varphi(x)$ 的一个极小点。相应的最优性条件为

$$0 \in \partial \varphi(x^*).$$

结合定理 12.8.7 可以知道

$$0 \in \text{co}\{\partial f(x^*), \nabla g_i(x^*): i \in I(x^*)\},$$

其中 $I(x^*) := \{i: g_i(x^*) = 0\}$ 。从而, 存在一组实数 $\lambda_0^*, \lambda_i^*, i \in I(x^*)$, 使得

$$\lambda_0^* \geq 0, \lambda_i^* \geq 0, \quad \lambda_0^* + \sum_i \lambda_i^* = 1, \quad i \in I(x^*),$$

$$0 \in \lambda_0^* \partial f(x^*) + \sum_i \lambda_i^* \nabla g_i(x^*), \quad i \in I(x^*).$$

引入 λ_i^* , 其中 $i \in \{1, \dots, l\}$ 但不属于 $I(x^*)$ 。令这些引入的 $\lambda_i^* = 0$ 。证完。 \square

13.7 一致有界性原则与共鸣定理

在 1876 年, P. du Bois Reymond 构造了一个周期为 2π 的连续函数, 使得其 Fourier 级数在给定的点处发散。在总结前人五十年来大量工作的基础上, Banach 和 Steinhaus 通过借鉴 Osgood 定理的证明方法, 抽象提取出了共鸣(resonance)定理。它刻画了有界线性算子族的基本性质, 从而成为泛函分析中最为重要的研究成果之一。

引理 13.7.1 设 T 是从 Banach 空间 \mathcal{B} 映射到赋范线性空间 \mathcal{Y} 上的有界线性算子。则对于任意的 $x \in \mathcal{B}$ 和 $r > 0$, 有

$$\sup\{\|Tx'\|: x' \in \mathcal{B}(x, r)\} \geq r\|T\|,$$

其中 $\mathcal{B}(x, r) = \{x' \in \mathcal{B}: \|x' - x\| \leq r\}$ 。

证 对于任意的 $\xi \in \mathcal{B}$, 我们有

$$\begin{aligned} 2 \max\{\|T(x + \xi)\|, \|T(x - \xi)\|\} &\geq \|T(x + \xi)\| + \|T(x - \xi)\| \\ &\geq \|T(x + \xi - (x - \xi))\| \\ &= 2\|T\xi\|. \end{aligned} \quad (13.13)$$

由线性算子范数的定义得 $\sup\{\|T(r^{-1}\xi)\|: \|r^{-1}\xi\| \leq 1\} \geq \|T\|$ (以二维空间为例, 前者在单位圆面上取而后者在单位圆周上取上确界。当然, 由于 T 是有界线性的, 所以实际上两者的上确界相等)。于是, 我们就有

$$\forall \xi \in \mathcal{B}(0, r), \sup\|T\xi\| = r \sup\|T(r^{-1}\xi)\| \geq r\|T\|.$$

另一方面, 因为当 $\xi \in \mathcal{B}(0, r)$ 时, $x \pm \xi \in \mathcal{B}(x, r)$, 所以

$$\sup\{\|Tx'\|: x' \in \mathcal{B}(x, r)\} \geq \max\{\|T(x + \xi)\|, \|T(x - \xi)\|\}.$$

在 (13.13) 式两边关于 $\xi \in \mathcal{B}(0, r)$ 取上确界即可。 \square

定理 13.7.1 (一致有界性原则) 设 \mathcal{F} 是一族从 Banach 空间 \mathcal{B} 映射到赋范线性空间 \mathcal{Y} 上的有界线性算子。若任给的 $T \in \mathcal{F}$ 是逐点有界的, 即对于所有的 $x \in \mathcal{B}$, 都存在相应的 $\beta_x > 0$ 使得 $\|Tx\| \leq \beta_x$, 则必然存在不再依赖于 x 的 $\beta > 0$ 使得

$$\|T\| \leq \beta, \quad \forall T \in \mathcal{F}.$$

讨论 取 \mathcal{B} 和 \mathcal{Y} 为一维实空间, $T_n(x) = nx$, $n = 1, 2, \dots$ 是一族线性算子。试问: 满足上述定理的题设条件吗? 举例说明: 如果 \mathcal{B} 换成不完备的赋范线性空间, 那么一致有界性原则不成立。

一致有界性原则的逆否命题为下面的共鸣定理。

特别地, 若 $\{\lambda_k\}$ 有正的下界, 则

$$|A(x^k) - 0|^2 \leq o(1/k), \quad k = 1, 2, \dots$$

证 根据迭代公式 (14.13), 我们有

$$\lambda_k^{-1}(x^k - x^{k+1}) \in A(x^{k+1}). \quad (14.21)$$

再根据假设 $0 \in A(x^*)$ 以及 A 的单调性, 我们进一步有

$$\begin{aligned} \langle x^{k+1} - x^*, \lambda_k^{-1}(x^k - x^{k+1}) - 0 \rangle &\geq 0 \quad \xrightarrow{\lambda_k \geq 0} \\ \langle x^{k+1} - x^*, x^k - x^{k+1} \rangle &\geq 0. \end{aligned}$$

因此

$$\begin{aligned} \|x^k - x^*\|^2 &= \|x^k - x^{k+1} + x^{k+1} - x^*\|^2 \\ &= \|x^k - x^{k+1}\|^2 + 2\langle x^k - x^{k+1}, x^{k+1} - x^* \rangle + \|x^{k+1} - x^*\|^2 \\ &\geq \|x^k - x^{k+1}\|^2 + \|x^{k+1} - x^*\|^2. \end{aligned} \quad (14.22)$$

因此

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \|x^k - x^{k+1}\|^2. \quad (14.23)$$

可以推出

$$|x^{k+1} - A^{-1}(0)|^2 \leq |x^k - A^{-1}(0)|^2 - \lambda_k^2 \|\lambda_k^{-1}(x^k - x^{k+1})\|^2.$$

这表明: 序列 $\{|x^k - A^{-1}(0)|\}$ 单调递减有下界, 从而它的极限 l 存在. 因此, 我们有

$$\begin{aligned} &|x^{k+1} - A^{-1}(0)|^2 - l^2 \\ &\leq |x^k - A^{-1}(0)|^2 - l^2 - \lambda_k^2 \|\lambda_k^{-1}(x^k - x^{k+1})\|^2. \end{aligned} \quad (14.24)$$

引入记号

$$\alpha_k := \sqrt{|x^k - A^{-1}(0)|^2 - l^2}, \quad \beta_k := \lambda_k^2, \quad \gamma_k := \|\lambda_k^{-1}(x^k - x^{k+1})\|^2.$$

那么根据定理 14.4.1 和 Dong 引理, 我们有: 当假设 (14.20) 成立时, 我们有

$$\begin{aligned} \|\lambda_k^{-1}(x^k - x^{k+1})\|^2 \sum_{i=0}^k \lambda_i^2 &\leq 2\sqrt{|x^0 - A^{-1}(0)|^2 - l^2} \varepsilon_k, \\ \lim_{k \rightarrow +\infty} \varepsilon_k &= \lim_{k \rightarrow +\infty} \sqrt{|x^k - A^{-1}(0)|^2 - l^2} = 0. \end{aligned}$$

再结合 (14.21) 即可。

算法 15.3 算法 15.1 的一个特例

0. 选取 $x^0 \in \text{dom } B$ 和 $u^0 \in \mathcal{G}$. 令 $k := 0$.

1. 选取 $\alpha_k > 0$. 计算

$$x^k(\alpha_k) = (I + \alpha_k B)^{-1}(x^k - \alpha_k Q^* u^k),$$

记 $\bar{x}^k = x^k(\alpha_k)$. 选取 β_k 满足 (15.12) 并且找到 \bar{u}^k 使得

$$(\beta_k I + D^{-1})(\bar{u}^k) \ni \beta_k u^k + Q\bar{x}^k - q. \quad (15.29)$$

若 $\bar{x}^k = x^k$, $\bar{u}^k = u^k$, 停止. 否则, 进入步 2.

2. 计算

$$d_x^k = \alpha_k^{-1}(x^k - \bar{x}^k) - Q^*(u^k - \bar{u}^k), \quad (15.30)$$

$$d_u^k = \beta_k(u^k - \bar{u}^k), \quad (15.31)$$

$$\gamma_k = \frac{\langle x^k - \bar{x}^k, d_x^k \rangle + \langle u^k - \bar{u}^k, d_u^k \rangle}{\|d_x^k\|^2 + \|d_u^k\|^2}. \quad (15.32)$$

选取 $\theta_k \in (0, 2]$, 并且计算

$$x^{k+1} = P_{\mathcal{C}}[x^k - \theta_k \gamma_k d_x^k], \quad u^{k+1} = u^k - \theta_k \gamma_k d_u^k, \quad (15.33)$$

其中 $\mathcal{C} = \text{dom } B$. 令 $k := k + 1$, 并且转入步 1.

下面, 我们讨论算法 15.3 在深度学习中的应用。

将 (15.2) 中的最小二乘项变一下, 则

$$\min_x \|(e, XH)x - p\|_1/n + \lambda(|x_2| + \dots + |x_r|) + \lambda \left\| \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & H \end{pmatrix} x \right\|_1. \quad (15.34)$$

显然, (15.34) 一个等价形式为

$$\min_{x \in \mathcal{R}^{r+1}} |x_2| + \dots + |x_r| + \|Qx - q\|_1, \quad (15.35)$$

其中

$$Q = \begin{pmatrix} \frac{1}{n\lambda}e & \frac{1}{n\lambda}XH \\ 0 & \mathbf{0}^T \\ \mathbf{0} & H \end{pmatrix}, \quad q = \begin{pmatrix} \frac{1}{n\lambda}p \\ 0 \\ \mathbf{0} \end{pmatrix}.$$

注意: 矩阵 Q 的行数和列数分别为 $n + d + 1$ 和 $r + 1$ 。

记

$$f(x) = |x_2| + \dots + |x_r|, \quad g(\cdot) = \|\cdot\|_1.$$

算法 16.3 算法 15.1 应用于半定规划

0. 选取 $X^0 \in S^n$ 和 $u^0 \in R^m$. 令 $k := 0$.

1. 选取 $\alpha_k > 0$. 计算

$$\bar{X}^k = [X^k - \alpha_k(C + \mathcal{A}^T u^k)]_+. \quad (16.25)$$

选取 β_k 满足 (16.31) 并且找到 \bar{u}^k 使得

$$\bar{u}^k = u^k + \beta_k^{-1}(\mathcal{A}\bar{X}^k - b). \quad (16.26)$$

若 $\bar{X}^k = X^k$, $\bar{u}^k = u^k$, 停止. 否则, 进入步 2.

2. 计算

$$d_x^k = \alpha_k^{-1}(X^k - \bar{X}^k) - \mathcal{A}^T(u^k - \bar{u}^k), \quad (16.27)$$

$$d_u^k = \beta_k(u^k - \bar{u}^k), \quad (16.28)$$

$$\gamma_k = \frac{\langle X^k - \bar{X}^k, d_x^k \rangle + \langle u^k - \bar{u}^k, d_u^k \rangle}{\|d_x^k\|_F^2 + \|d_u^k\|^2}. \quad (16.29)$$

选取 $\theta_k \in (0, 2]$, 并且计算

$$X^{k+1} = X^k - \theta_k \gamma_k d_x^k, \quad u^{k+1} = u^k - \theta_k \gamma_k d_u^k, \quad (16.30)$$

令 $k := k + 1$, 并且转入步 1.

至于 β_k , 我们建议

$$\beta_k \geq \frac{2\varepsilon - \varepsilon^2 + \alpha_k^2 \|AA^T\|}{2(2 - \varepsilon)\alpha_k}, \quad 0 < \varepsilon < 2. \quad (16.31)$$

它对应于 (15.14).

注意两点: (16.26) 由 Moreau 恒等式得到; 调用 ARPACK 库中隐式重启的 *Arnoldi* 算法来计算

$$M := X^k - \alpha_k(C + \mathcal{A}^T u^k)$$

到正半定锥上的近似 (非精确!) 投影. 具体地讲, 假设理论上 M 的谱分解为

$$M = \lambda_1 z^1 (z^1)^T + \cdots + \lambda_n z^n (z^n)^T$$

则 M 到正半定锥上的投影为

$$M_+ = \max\{0, \lambda_1\} z^1 (z^1)^T + \cdots + \max\{0, \lambda_n\} z^n (z^n)^T$$

而近似投影为: 若 M 的最大特征值 $\lambda_1 \leq 0$, 则 M_+ 为零矩阵. 否则, 我们有

$$M_+ \stackrel{\epsilon}{\approx} \max\{0, \lambda_1\} z^1 (z^1)^T + \cdots + \max\{0, \lambda_r\} z^r (z^r)^T,$$

索引

- LDL^T 分解, 147, 157
 ℓ_2 空间, 215
 $L_2[a, b]$ 空间, 216
Banach 空间, 215
Cauchy 列, 215
Fenchel 对偶定理, 206
Fenchel 共轭函数, 204
Fitzpatrick 函数, 240
Hilbert 空间, 215
Schwarz 不等式, 214
0.618 法, 12
- Abadie 约束限制, 111, 112
Aitken, 97
Armijo 条件, 13, 14, 17, 125, 174
Arnoldi 算法, 278
Arrow-Hurwitz-Uzawa 约束限制, 111
- back propagation, 30
Bregman 距离, 87
Broyden 方法, 99
- Chebyshev 逼近, 4
Cholesky 分解, 31
constraint qualifications, 109
- Davidon 变尺度方法, 63
DFP 方法, 67
Dong 条件, 16, 41–43
Dong 引理, 244, 246
Farkas 引理, 105
- Fermat 定理, 7
Forsythe, 31
Fréchet 可微, 5
Fredholm 积分方程, 87, 233, 239
- Gauß 消去法, 31
Gordan 定理, 105
Graham 扫描法, 176
Guiguard 约束限制, 111, 112
- Hesse 矩阵, 5
Hestenes, 31
Householder 变换, 96
- Jacobi 矩阵, 6, 90, 106, 124, 147, 173
Jensen 不等式, 186
- Kantorovich 不等式, 17
Kiefer, 12
KKT 定理, 109, 113, 175
Korpelevich, 27
Kuhn-Tucker 约束限制, 111
Kullback-Leibler 散度, 188
- Lagrange 乘子向量, 123, 124, 131, 134, 135, 166, 266, 271
Lanczos, 31
Levenberg-Marquardt 方法, 89
Lipschitz 连续, 9
- Marquardt 方法, 89
Mehrotra 预估校正算法, 149, 155

- memoryless, 83
- MF 约束限制, 109–111
- Minty 定理, 235
- Moreau, 195
- Moreau envelope, 250
- Moreau-Yosida 逼近, 250
- Motzkin 定理, 104

- Passty, 200
- Perry-Shanno 方法, 82
- Perry 方法, 82
- Poisson 问题, 37
- Portfolio, 165
- proper separation, 182

- Raphson, 49
- ReLU 函数, 180, 205
- Riesz 表示定理, 221, 222
- Rosenbrock 函数, 22
- Rosser, 31

- Salahi 预估校正算法, 149, 158
- sequential quadratic programming, 174
- Sherman-Morrison 公式, 18, 70, 74
- Simpson, 49
- Slater 条件, 111–114
- Stiefel, 31

- Taylor 定理, 5
- Tucker 定理, 120

- Wolfe 条件, 13–16, 69, 74, 75, 77, 78

- Yosida 逼近, 235

- 鞍点, 8
- 伴算子, 227
- 伴随, 254, 261
- 半定规划, 3, 265
- 半光滑, 52, 122
- 半正定锥, 268, 269, 273, 279, 283

- 闭, 181
- 闭凸集外表示, 182
- 闭凸锥, 119–121
- 变分不等式, 260

- 不等式约束函数, 103
- 不精确线搜索, 13
- 不完全 Cholesky 分解, 38
- 步长, 11
- 插值法, 11, 12

- 常微分方程, 106
- 超椭球体, 3, 282
- 超线性收敛, 20, 67, 78, 82
- 稠密, 200

- 次梯度, 194
- 次微分, 194

- 单调, 9
- 单调包含, 243, 256
- 单调的, 238
- 单位下三角矩阵, 38
- 单形, 180
- 单形法, 149
- 等式约束函数, 103
- 低秩解, 279

- 对称秩二校正公式, 64, 98
- 对称秩一校正公式, 84, 98
- 对偶(空间), 222
- 对偶变量, 148, 150
- 对偶松弛变量, 150
- 对数障碍函数, 139
- 多项式时间, 149, 163
- 二次插值法, 13

- 二次惩罚函数法, 127
- 二次收敛, 20, 51, 53, 61, 62
- 二次有限终止性, 19, 47, 69, 77, 84
- 二阶充分条件, 115, 133

- 反向传播算法, 30
- 范数, 214
- 方向导数, 192
- 仿射超平面, 180
- 仿射流形, 180
- 非光滑优化, 4
- 非扩张算子, 239
- 非线性规划, 4
- 分离定理, 182, 183

- 赋范线性空间, 213
- 复合函数, 6, 107
- 负梯度方向, 11, 22, 27, 33, 35, 50, 74, 93, 94
- 割线, 63, 100
- 割线法, 11

- 公式法, 11
- 共鸣定理, 223, 224
- 共轭方向, 67
- 共轭梯度法, 31
- 共轭性, 67
- 光滑优化, 4

- 互补, 260
- 互补条件, 108, 118, 150
- 互补问题, 108, 122
- 黄金分割法, 11, 12
- 机器学习, 29, 174, 179
- 积极集法, 168, 169
- 积极约束函数, 103

- 极大单调, 238, 240, 254, 256, 261
- 极大单调算子, 235
- 极锥, 111

- 降维法, 172, 299
- 解析中心, 153

- 精确线搜索, 11, 22–24, 40, 45, 67–70, 78, 80, 82
- 局部极小点, 7
- 矩阵多项式, 35
- 锯齿现象, 27
- 卷积下确界, 191

- 可逆, 2
- 可微凸规划, 103, 111–114, 132, 133, 152, 153
- 可行方向, 104, 109, 115

- 连通, 140
- 连续可微, 6
- 连续可微曲线, 106, 107, 144, 145
- 临界点, 7
- 邻近点方法, 235, 243

- 模, 1
- 目标函数, 3, 103
- 内点法, 139, 149
- 内积, 1
- 内积的连续性定理, 216
- 内积空间, 214
- 拟 Newton 法, 63
- 逆二次插值法, 13

- 抛物线法, 13, 14

- 谱分解, 17

- 强 Wolfe 条件, 16, 40, 45

- 强单调, 9, 70, 254
- 强对偶, 284
- 强凸函数, 8
- 切方向, 111
- 切锥, 111
- 曲率条件, 14, 70

- 全局极小点, 7
- 弱*收敛, 225
- 弱闭, 227
- 弱对偶, 167, 269
- 弱聚点, 247, 260
- 弱收敛, 224, 255
- 三次插值法, 13

- 上图, 185
- 深度学习, 29, 30

- 示性函数, 104
- 收敛, 19
- 收敛率, 19

- 双共轭, 208
- 水平集, 187
- 搜索方向, 32–35, 69, 74, 82, 93, 148, 156
- 算子的偏序, 231
- 算子分裂方法, 253, 255
- 损失函数, 29, 180

- 特征算子, 256
- 特征向量, 2
- 特征值, 2, 17, 25, 30, 36
- 梯度, 5
- 条件数, 2, 35, 129, 137

- 停止准则, 19
- 投影, 106, 268, 269, 273, 275, 279, 280
- 凸包, 198
- 凸规划, 4
- 凸函数, 4, 179
- 凸函数的二阶逼近, 203, 211
- 凸集, 4, 179, 217
- 凸优化, 4
- 外点法, 131
- 外推, 145, 146
- 完备, 215

- 无穷模邻域, 158
- 五点差分, 37

- 下半连续的, 187
- 下降方向, 11, 14, 16, 21, 41, 59, 60, 72, 74, 93, 94, 104, 110, 115, 124
- 线性泛函, 218
- 线性复合, 253
- 线性广义梯度, 210
- 线性规划, 3, 4, 149
- 线性空间, 213
- 线性收敛, 19, 20, 47, 52, 62
- 线性算子, 218
- 线性无关, 2, 109, 110, 133, 142, 143, 173, 176, 290
- 线性无关约束限制, 109–111
- 线性约束限制, 109, 112
- 相对内点, 181
- 向前差分法, 27

- 斜伴随, 254, 255, 259
- 斜对称, 254

- 严格互补条件, 135, 143, 144
- 严格凸二次函数, 8, 19, 25, 40, 52, 56, 67, 69, 70, 77, 84

- 一致有界性原则, 223
- 遗传性, 67
- 隐函数定理, 145

有限内存, 72

有效域, 185, 194

预处理, 37

原始对偶内点法, 149, 267

约束函数, 3

约束限制, 109, 111, 112, 117, 119

增广 Lagrange 乘子法, 123

正交分解定理, 221, 222, 226

正交投影, 218

正算子, 231

支撑向量, 174, 175

执行集, 168

中心路径, 150

逐步二次规划, 174

逐步二次规划方法, 126, 173

驻点, 7

锥, 119, 120, 182

自伴随, 254

自伴有界线性算子, 230

自对称, 254

自协调函数, 54

最陡下降法, 21, 172

最小二乘求解器, 100

最小二乘问题, 89

最优步长, 22