北京航空航天大学学院路校区
最优化方法讲义(2022, 秋季)
Optimization Method

授课教师：刘红英

Email: liuhongying@buaa.edu.cn

August 27, 2022

Abstract

这些讲义的目的是给出连续优化中一些重要内容的基本介绍. 重点聚焦于那些在机器学习和数据分析中出现的方法，特别突出了凸性、鲁棒性和在这些领域的实现. 大量现有的ipython notebooks扩大了理论进展, 已经连接到文本并且可在如下网页下载:

https://ee227c.github.io/.

这些讲义的原始素材主要出自Berkeley于2018年春季教授的课程*EE227C: Convex Optimization and Approximation*(凸优化和近似，https://ee227c.github.io/). 此外，也参考了：

1. MPhilippe Rigollet 教授在MIT 2015 fall的18.657课程使用的讲义“mathematics of machine learning" (机器学习数学， https://ocw.mit.edu/courses/18-657-mathematics-of-machine-learning-fall-2015/pages/syllabus/),
2. Arkadi Nemirovski博士在GIT2018 Spring的 ISyE 6663 课程使用的幻灯片“Optimization III: Convex Analysis, Nonlinear Programming Theory, Nonlinear Programming Algorithms”，
3. Sébastien Bubeck, Convex Optimization: Algorithms and Complexity (凸优化：算法和计算复杂性, https://academic.microsoft.com/paper/2296319761), 和
4. Anthony Man-Cho So教授在香港中文大学2021 Fall 的ENGG 5501课程使用的讲义“Foundations of Optimization”(苏文藻，最优化基础, https://www1.se.cuhk.edu.hk/ manchoso/).

该讲义的特点是以最优化方法为线索，仅在需要相关概念和理论时才引入. 此外，也融入了相关方向最新的理论进展和应用.

目录

I 梯度法 **1**

1 凸性 **1**
1.1 凸集 1
1.2 凸函数 3
1.3 凸优化 6

2 梯度法 **8**
2.1 梯度下降法 9
2.2 Lipschitz函数 10
2.3 光滑函数 13

3 强凸性 **15**
3.1 提醒 15
3.2 强凸性 16
3.3 针对强凸函数的收敛速率 16
3.4 针对光滑强凸函数的收敛速率 18

4 梯度方法的若干应用 **19**

5 镜像下降法 **19**
5.1 Bregman投影 20
5.2 镜像下降算法 22
5.3 注记 24

6 条件梯度法 **27**
6.1 算法 27
6.2 条件梯度法的收敛分析 28
6.3 应用于核范数优化问题 29

II 加速梯度法 **31**

7 探索加速 **31**
7.1 二次函数 31
7.2 二次函数的梯度下降法 32
7.3 与多项式逼近的联系 33
7.4 Chebyshev多项式 35

8 Krylov子空间、特征值和共轭梯度法 **38**
8.1 Krylov子空间 38
8.2 求特征向量 39
8.3 应用Chebyshev多项式 40
8.4 共轭梯度法 40

9 Nesterov加速梯度下降法 **42**
9.1 收敛分析 43
9.2 强凸情况 44

10 下界与稳健性之间的权衡 **45**
10.1 下界 45
10.2 稳健性与加速之间的折中 49

III 随机优化 **51**

11 随机梯度法 **51**
11.1 风险极小化与经验风险极小化 51
11.2 外部随机优化问题的随机梯度法 53
11.3 随机梯度法 55
11.4 随机镜像下降法 56
11.5 在线学习与乘性权重更新 57

12 坐标下降法 **59**
12.1 随机坐标下降法 59
12.2 重要性采样 60
12.3 针对光滑坐标下降法的重要性采样 60
12.4 随机坐标下降法与随机梯度下降法 63
12.5 坐标下降的其它推广： 63

13 学习、稳定性、正则化 **63**
13.1 经验风险和推广误差 64
13.2 算法稳定性 64
13.3 经验风险极小化的稳定性 66
13.4 正则化 66
13.5 隐正则化 67

IV 对偶方法 **68**

14 对偶定理 **68**
14.1 等式约束优化的最优性条件 68
14.2 非线性约束 69
14.3 对偶问题 72
14.4 弱对偶性 72
14.5 强对偶性 73

15 利用对偶性的算法 **73**
15.1 对偶函数的性质 73
15.2 对偶梯度上升法 74
15.3 增广Lagrange函数法/乘子法 74

15.4 对偶分解法 75
15.5 ADMM-交替方向乘子法 77

16 Fenchel对偶与算法 **78**
16.1 得到经验风险最小化的对偶问题 79
16.2 随机对偶坐标上升法 81

17 反向传播与伴随 **82**
17.1 热身 82
17.2 通用表述 83
17.3 与链式法则的联系 85
17.4 举例说明 86

V 非凸优化 **88**

18 非凸问题 **88**
18.1 局部极小点 88
18.2 最速下降法的全局收敛性 90
18.3 鞍点 92

19 逃离鞍点 **93**
19.1 动力系统视角 94
19.2 二次情况 94
19.3 一般情况 95

20 交替极小化和期望极大化(EM) **96**

21 无导数优化、策略梯度和控制 **96**

22 非凸目标函数与凸松弛 **97**
22.1 难度 97
22.2 凸松弛 98

23 非凸约束与投影梯度下降法 **102**

VI 高阶和内点法 **106**

24 牛顿法 **106**
24.1 阻尼更新 108
24.2 拟牛顿法 109

25 二阶方法的实验 **110**

26 内点法入门 **110**
26.1 障碍法 110
26.2 线性规划 112

27 原始-对偶内点法 **115**
27.1 得到对偶问题 116
27.2 沿着中心路径的原始-对偶迭代 117
27.3 用牛顿步生成迭代 118

Part I
梯度法

极小化函数的方法中，**梯度下降法**(Gradient descent) 是应用最广泛的方法之一，适用于凸函数和非凸函数. 它的核心，是一种特定的**局部搜索**(local search)格式，多次迭代中都是在小区域内贪心地优化函数. 如果$f:\mathbb{R}\to\mathbb{R}$是二次连续可微的，那么Taylor定理表明

$$f(x+\delta)\approx f(x)+\delta f'(x)+\frac{1}{2}\delta^2 f''(x)\,.$$

该近似直接揭示出：对充分小的$\eta>0$，如果从x移到$x+\delta$，其中$\delta=-\eta\cdot f'(x)$，通常期望函数值能减小$\eta(f'(x))^2$. 利用Taylor定理的多元函数版本，可将该思想推广到多元函数. 这种使函数值减小的简单贪心方法就是著名的**梯度下降法**(gradient descent, GD).

梯度下降法收敛到一阶导数消失的点. 对于一大类**凸**函数, 这些点就是全局极小点. 此外，能够修正梯度下降法，使其适用于甚至不可微的凸函数. 后面，将证明梯度下降法也收敛到局部 (并且, 有时, 全局!)极小点.

关于这部分内容，在 第 1 节中介绍关于凸函数的预备知识以便推广梯度下降法. 关键之处是引入了次梯度的概念, 从而将梯度概念推广至非光滑凸函数. 在 第 2 节, 形式地引入(次)梯度下降法, 并证明当将梯度下降法应用到凸函数时得到的收敛速率. 在 第 3 节 引入**强凸**(strong convexity)假设, 这个更强的假设使得(次)梯度下降法享有更快的速率. **[Max's Note: finish]**

1 凸性

本节给出对凸集和凸函数而言是最重要的事实，后面会多次用到这些重要概念和结论. 当f是充分光滑的时，它们经常是Taylor定理的简单推论.

1.1 凸集

定义 1.1 (凸集)**.** 称集合$K\subseteq\mathbb{R}^n$是**凸的**(convex)，如果连接K 中任何两点的线段也包含在 K中. 正式地，对所有$x,y\in K$ 和所有标量$\gamma\in[0,1]$有$\gamma x+(1-\gamma)y\in K$.

定理 1.2 (分离定理)**.** 设$C,K\subseteq\mathbb{R}^n$均是凸集并且没有公共点，即$C\cap K=\emptyset$. 那么存在点$a\in\mathbb{R}^n$和数$b\in\mathbb{R}$使得

(i) 对所有$x\in C$, 有$\langle a,x\rangle\geqslant b$.

(ii) 对所有$x\in K$, 有$\langle a,x\rangle\leqslant b$.

如果C和K是闭集并且至少其中之一有界，那么可用严格不等式替换上面的不等式.

最关心的情况是当两个集合都是紧集(即有界闭集). 这里强调它的证明.

对于紧集证明定理 1.2. 在这种情况下，笛卡尔乘积$C \times K$也是紧的. 因此，距离函数$\|x-y\|$ 在$C \times K$上能取到最小值. 设p, q 是取到最小值的两个点. 过p和q 的中点且垂直于$q-p$ 的超平面是一个分离超平面. 即$a=q-p$，$b=(\langle a, q\rangle - \langle a, p\rangle)/2$. 用反证法，假设存在超平面上的点 r包含在两个集合中的一个，比如说 C. 那么由凸性，连接p和r 的线段也包含在 C中. 沿着这个线段可以找到点比p更接近q，这就与假设矛盾. ■

例子 1.3. 常用凸集

- **超平面(hyperplane)**$\{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}$ 和**仿射半空间(affine half space)**$\{x \in \mathbb{R}^n \mid \langle a, x\rangle \geqslant b\}$，其中$0 \neq a \in \mathbb{R}^n, b \in \mathbb{R}$ 是已知的. 称a 是超平面的**法向量(normal vector)**.

- 凸集的交集. 特别地，仿射子空间$\{x \in \mathbb{R}^n \mid Ax = b\}$和多面体(polyhedral set)$\{x \in \mathbb{R}^n \mid Ax \leqslant b\}$ 也是凸的，其中$0 \neq A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ 是已知的. 事实上，(由分离超平面定理知)每个闭凸集等于包含它的所有仿射半空间的交集.

- 凸集的仿射变换. 如果$K \subseteq \mathbb{R}^n$是凸的，对任何$A \in \mathbb{R}^{m \times n}$ 和$b \in \mathbb{R}^m$，$\{Ax+b \mid x \in K\}$也是凸的.

- 半正定矩阵锥，记作$S_+^n = \{A \in \mathbb{R}^{n \times n} \mid A \succeq 0\}$. 这里用$A \succeq 0$表示对所有$x \in \mathbb{R}^n$有$x^\top Ax \geqslant 0$ 成立. 可由凸集的定义直接验证S_+^n是凸的，但是也可以由已知推导出来. 的确，用$S_n = \{A \in \mathbb{R}^{n \times n} \mid A^\top = A\}$ 表示所有$n \times n$ 阶对称矩阵组成的集合，能够将S_+^n写作(无限个)半空间的交集：$S_+^n = \bigcap_{x \in \mathbb{R}^n \setminus \{0\}} \{A \in S_n \mid x^\top Ax \geqslant 0\}$.

- 如果$x_1 - x_0, \cdots, x_m - x_0$线性无关，称

$$\Delta(x_0, \cdots, x_m) = \left\{ \sum_{i=0}^m \theta_i x_i : \sum_{i=0}^m \theta_i = 1, \theta_i \geqslant 0, \forall i \right\}.$$

  是顶点为$x_0, x_1, \cdots, x_m$的$m-$维单纯形. 单纯形中每个点是顶点的凸组合，并且系数由该点唯一确定. $2-$维单纯形是由3个不共线的点确定的，是以这三个点为顶点的三角形；设$e_1, \cdots, e_n$是$\mathbb{R}^n$中的标准正交基. 由它们确定的$(n-1)-$维单纯形是**标准单纯形**(standard simplex)

$$\Delta_n = \{x \in \mathbb{R}^n : x \geqslant \underline{0}, \sum_{i=1}^n x_i = 1\}.$$

  将$e_0 = \underline{0}$加入$e_1, \cdots, e_n$，得到对应的n 维单纯形是

$$\Delta_n^+ = \{x \in \mathbb{R}^n : x \geqslant \underline{0}, \sum_{i=1}^n x_i \leqslant 1\}.$$

- 更多的凸集参见Boyd-Vandenberghe的《凸优化》[BV04].

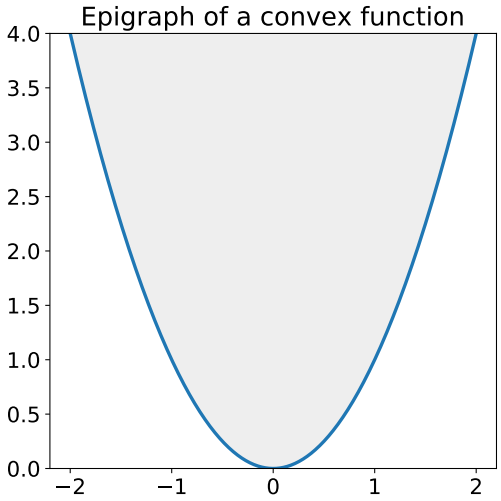


图 1.1: 函数的上图.

1.2 凸函数

定义 1.4 (凸函数)**.** 设$\Omega \subseteq \mathbb{R}^n$是凸集. 称函数$f\colon \Omega \to \mathbb{R}$ 是**凸的**(convex) 如果对所有$x, y \in \Omega$ 和所有标量$\gamma \in [0,1]$ 有$f((1-\gamma)x + \gamma y) \leqslant (1-\gamma)f(x) + \gamma f(y)$.

Jensen (1905)证明对于**连续**函数，能由中点条件—- 对于所有$x, y \in \Omega$,

$$f\left(\frac{x+y}{2}\right) \leqslant \frac{f(x)+f(y)}{2}$$

得到凸性. 在已知函数是连续的情况下，该结论有时能简化证明.

定义 1.5. 函数 $f\colon \Omega \to \mathbb{R}$的**上图**(epigraph)定义为

$$\operatorname{epi}(f) = \{(x,t) \in \Omega \times \mathbb{R} \mid f(x) \leqslant t\}.$$

事实 1.6. 函数是凸的当且仅当它的上图是凸的. 上图的几何直观见图*1.1*.

命题 1.7 (Jensen不等式)**.** 假设$f\colon \Omega \to \mathbb{R}$是凸函数，并且

$$x_1, \cdots, x_k \in \Omega,$$

权重$\gamma_i > 0$. 那么

$$f\left(\frac{\sum_{i=1}^k \gamma_i x_i}{\sum_{i=1}^k \gamma_i}\right) \leqslant \frac{\sum_{i=1}^k \gamma_i f(x_i)}{\sum_{i=1}^k \gamma_i}.$$

如下链接是一种图形”证明“ this link.

1.2.1 一阶刻画

将凸性和Taylor定理联系起来是有益的. 下面先回忆Taylor定理. 可微函数 $f\colon \Omega \to \mathbb{R}$在$x \in \Omega$处的**梯度(gradient)**定义为由偏导数

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_i}\right)_{i=1}^{n}$$

作分量组成的向量. 特别指出如下简单事实，它将梯度的线性型与一元函数在 0 处的导数值联系起来，这是由多元函数链式法则得到的结果.

事实 1.8. 假设$f\colon \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$连续可微. 设$x$ 是Ω 的内点, $0 \neq p \in \mathbb{R}^n$. 考虑

$$\phi(\gamma) = f(x + \gamma p). \tag{1.1}$$

那么，

$$\phi'(\gamma) = \nabla f(x + \gamma p)^\top p.$$

特别地，$\phi'(0) = \nabla f(x)^\top p$.

Taylor定理蕴含着如下命题.

命题 1.9. 假设$f\colon \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$在沿着连接两点$x$ 和y的线段上可微. 那么

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \int_0^1 (1 - \gamma)\phi''(\gamma)\,\mathrm{d}\gamma.$$

Proof. 针对函数$\phi(\gamma) = f(x + \gamma(y - x))$利用二阶Taylor 展式，并用事实 1.8替换其中的一阶项. ■

对于可微函数而言，凸性等价于性质：一阶 Taylor近似提供了函数的整体下界，即函数在一点的线性近似是函数的**偏低**估计.

命题 1.10 (梯度不等式)**.** 假设$f\colon \Omega \to \mathbb{R}$是可微的. 那么, f在Ω 上是凸的当且仅当对所有$x, y \in \Omega$有

$$f(y) \geqslant f(x) + \nabla f(x)^\top (y - x)\,. \tag{1.2}$$

Proof. 首先，假设f是凸的，那么由定义

$$\begin{aligned} f(y) &\geqslant \frac{f((1-\gamma)x + \gamma y) - (1-\gamma)f(x)}{\gamma} \\ &\geqslant f(x) + \frac{f(x + \gamma(y - x)) - f(x)}{\gamma} \\ &\to f(x) + \nabla f(x)^\top (y - x) \quad \text{当 } \gamma \to 0 + 0 \qquad \text{(由事实 1.8.)} \end{aligned}$$

另一方面，固定两个点 $x, y \in \Omega$ 和 $\gamma \in [0, 1]$. 令$z = (1 - \gamma)x + \gamma y$并两次应用 (1.2)得到

$$f(x) \geqslant f(z) + \nabla f(z)^\top (x - z) \quad \text{和} \quad f(y) \geqslant f(z) + \nabla f(z)^\top (y - z)$$

给这两个不等式两边分别乘以$(1 - \gamma)$和γ相加，得到

$$(1 - \gamma)f(x) + \gamma f(y) \geqslant f(z),$$

由此得到凸性. ■

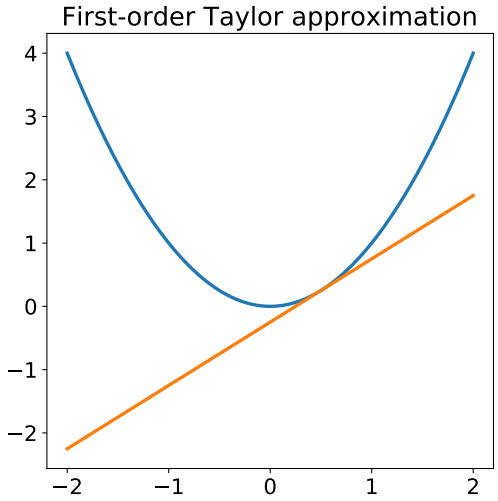


图 1.2: 函数$f(x) = x^2$在点0.5的Taylor近似.

推论 1.11. 设Ω 是凸集，f是定义在Ω上的凸函数，并且在包含Ω的开集上可微. 那么x_*是f在Ω上的全局极小点当且仅当

$$\langle \nabla f(x_*), x - x_* \rangle \geqslant 0 \quad \forall x \in \Omega.$$

特别地，如果$\nabla f(x_*) = 0$，即梯度在点 x_*处消失，那么 x_* 一定是 f的全局极小点.

当然，并不是所有凸函数都是可微的. 比如绝对值 $f(x) = |x|$, 它是凸的，但在 0 处不可微. 从而有必要扩展梯度的概念.

定义 1.12 (次梯度)**.** 设$\Omega \subseteq \mathbb{R}^n, x \in \Omega$，$f : \Omega \to \mathbb{R}$. 若向量$g \in \mathbb{R}^n$ 满足

$$f(y) \geqslant f(x) + g^\top (y - x) \quad \forall y \in \Omega,$$

则称该向量为f在x处的次梯度(subgradient). 记这种向量的集合为$\partial f(x)$，称为f在x处的**次微分**(sub-differential). 如果$\partial f(x)$非空，就称f 在x 处是次可微的.

次梯度的几何直观如图1.3，它本质上与梯度对应，但又不像梯度，对于凸函数，次梯度总是存在的. 下面定理表明甚至在不可微的情况下也是如此.

定理 若$f : \Omega \to \mathbb{R}$是凸的，则对于所有的$x$，$\partial f(x) \neq \emptyset$. 另外，若$f$ 在x处可微，那么$\partial f(x) = \{\nabla f(x)\}$.

证明略. 第一个结论需要凸集分离定理. 第二个结论由次梯度的定义和函数可微的性质可证明.

1.2.2 二阶刻画

将$f\colon \Omega \to \mathbb{R}$在点$x \in \Omega$处的Hessian阵定义为由二阶偏导数作元素组成的矩阵:

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{i,j \in [n]}.$$

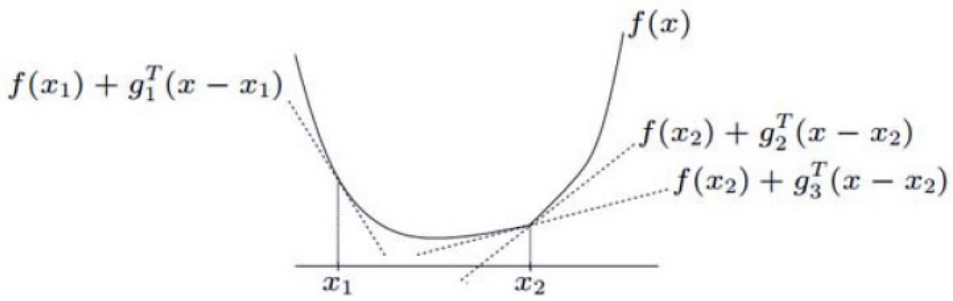


图 1.3: 函数f在x_1处是可微的，在x_2处是次可微的，g_1和g_2 均是次梯度.

倘若f的二阶偏导数在x的一个开邻域上是连续可微的，Schwarz 定理蕴含着点x处的Hessian阵是对称的.

类似于事实 1.8, 使用链式法则，能将Hessian阵定义的二次型与一元函数的导数联系起来.

事实 1.13. 假设$f\colon \Omega \to \mathbb{R}$在沿着从$x$到$y$的线段是二次连续可微的. 令$p = y - x$，考虑由式 (1.1)定义的一元函数. 那么

$$\phi''(\gamma) = p^\top \nabla^2 f(x + \gamma p) p.$$

命题 1.14. 如果f在它的定义域 Ω上是连续可微的，那么f是凸的当且仅当对所有$x \in \Omega$有$\nabla^2 f(x) \succeq 0$ 成立.

Proof. 假设f是凸的，并且目标是证明Hessian阵是半正定的. 针对某个任意的向量 u和标量α，设$y = x + \alpha u$. 命题 1.10表明

$$f(y) - f(x) - \nabla f(x)^\top (y - x) \geqslant 0$$

因此，由命题 1.9,

$$\begin{aligned} 0 &\leqslant \int_0^1 (1-\gamma)\phi''(\gamma)\,\mathrm{d}\gamma \\ &= (1-\bar{\gamma})\phi''(\bar{\gamma}) \quad \text{对某个}\bar{\gamma} \in (0,1) && \text{(由中值定理)} \\ &= (1-\bar{\gamma})(y-x)^\top \nabla^2 f(x + \bar{\gamma}(y-x))(y-x)\,. && \text{(由事实 1.13)} \end{aligned}$$

代入选择的 y, 这表明 $0 \leqslant u^\top \nabla^2 f(x + \alpha\gamma u)u$. 令 α趋于零就证明了$\nabla^2 f(x) \succeq 0$. (请注意这里的γ通常依赖于α 但总是不超过 1.)

现在，假设定义域Ω内任意点处的Hessian阵是半正定的，目标是证明函数 f 是凸的. 使用和上面同样的推导，能够证明Taylor定理中的二阶误差项必定是非负的. 因此一阶近似是整体下界，从而由 命题 1.10知函数 f是凸的. ■

1.3 凸优化

本课程的大部分内容是关于求解凸优化的不同方法. 凸优化即在凸集Ω上极小化凸函数 $f\colon \Omega \to \mathbb{R}$：

$$\min_{x \in \Omega} f(x). \tag{1.3}$$

首先，通过证明几乎所有优化问题可表述为一个凸问题来澄清一个误解“凸问题容易”. 为此，按如下方式重新表述给定的优化问题，即

$$\min_{x\in\Omega} f(x) \quad \Leftrightarrow \quad \min_{t\geqslant f(x), x\in\Omega} t \Leftrightarrow \min_{(x,t)\in \mathrm{epi}(f)} t,$$

其中$\mathrm{epi}(f)$是函数f的上图. 现在，已知集合D与向量c，观察到对于线性函数[1]，

$$\min_{x\in D} c^\top x = \min_{x\in \mathrm{conv}(D)} c^\top x,$$

其中凸包定义为

$$\mathrm{conv}(D) = \{y : \exists k \in \mathbb{Z}_+, x_1, \cdots, x_k \in D, \gamma_i \geqslant 0, \sum_{i=1}^k \gamma_i = 1, y = \sum_{i=1}^k \gamma_i x_i\}.$$

由于$D \subset \mathrm{conv}(D)$，因此上式的左边至少不小于右边. 为了表述简洁，下面用Δ_k表示$k-1$维标准单纯形. 为了证明另外一个方向，有

$$\begin{aligned}
\min_{x\in\mathrm{conv}(D)} c^\top x &= \min_k \min_{x_1,\cdots,x_k\in D} \min_{\gamma\in\Delta_k} c^\top \sum_{i=1}^k \gamma_i x_i \\
&= \min_k \min_{x_1,\cdots,x_k\in D} \min_{\gamma\in\Delta_k} \sum_{i=1}^k \gamma_i c^\top x_i \\
&\geqslant \min_k \min_{x_1,\cdots,x_k\in D} \min_{\gamma\in\Delta_k} \sum_{i=1}^k \gamma_i \min_{x\in D} c^\top x \\
&= \min_{x\in D} c^\top x,
\end{aligned}$$

因此得到

$$\min_{x\in\Omega} f(x) \Leftrightarrow \min_{(x,t)\in\mathrm{conv}(\mathrm{epi}(f))} t,$$

后者是一个凸问题.

1.3.1 为什么希望问题具有凸性？

下面将证明：根据凸性，能够从局部信息中推断出全局信息.

定理 1.15 (凸优化的性质及最优解的刻画)**.** 设f, Ω是凸的. 若x为f在Ω上的局部极小点，那么它也是全局极小点. 进一步，它是全局极小点当且仅当$0 \in \partial f(x)$.

Proof. $0 \in \partial f(x)$当且仅当对于所有的$y \in \Omega$有$f(x) - f(y) \leqslant 0$. 这等价于x 为全局极小点.

进一步，假定x为局部极小点，则对于所有的$y \in \Omega$，存在足够小的$\varepsilon > 0$，使得

$$f(x) \leqslant f\left((1-\varepsilon)x + \varepsilon y\right) \leqslant (1-\varepsilon) f(x) + \varepsilon f(y),$$

由此得到对于所有的$y \in \Omega$有$f(x) \leqslant f(y)$成立. ■

[1]将这里的线性函数换成凹函数，结论也成立. 本质是凸函数的最大值具有顶/极点可达性：设Ω是凸集，f在Ω上是凸函数. 如果f 在Ω能取到 最大值(上确界)，那么可在Ω 的极点处达到最大值.

考虑次梯度不仅让我们知道局部极小点是全局极小点，也告诉我们如果$g^\top(y-x) > 0$，则$f(x) < f(y)$. 这意味着$f(y)$ 不可能是极小值. 因此可将搜索限制在使得$g^\top(y-x) < 0$ 的y上. 在一维(单变量最优化)问题中，若$g > 0$，这对应着射线$\{y \in \mathbb{R} : y \leqslant x\}$；若$g < 0$，这对应着射线$\{y \in \mathbb{R} : y \geqslant x\}$. 这个概念也即梯度下降法的思想.

对于初学者，凸集也不必拥有紧凑描述. 当求解涉及凸集的计算问题时，需要关注如何表示正在处理的凸集. 不需要集合的显式表示，而是要求一个计算上称作**分离oracle**的抽象概念.

定义 1.16. 凸集 K的**分离oracle**(separation oracle)是一种装置，对任何 $x \notin K$的已知点，它返回一个将x与K 分离的超平面.

另一个计算上的抽象概念是**一阶oracle**，对任何已知点$x \in \Omega$，它返回$\nabla f(x)$. 类似地，**二阶oracle**返回$\nabla^2 f(x)$. 函数值oracle或者**零阶oracle** 仅返回$f(x)$. 一阶方法是使用一阶oracle 的算法. 相仿地，能够定义零阶方法，二阶方法.

1.3.2 什么是有效的?

经典的计算复杂性理论中，典型作法是用比特输入复杂性来量化一个算法消耗的资源(运行时间或者内存). 比如像 “使用长乘法，在$O(n^2)$时间内能得到两个n- 比特数的乘积."

这种计算方法在凸优化中复杂而低效，并且大部分课本回避了它. 相反，在凸优化领域，常规做法是用更抽象的资源，比如多久访问一次上面提到的某个oracle，来量化算法的成本. 统计oracle就能大致掌握预期一个方法工作的有多好.

在优化领域，“有效” 的定义并非完全一成不变. 典型地，目的是证明算法找到一个解 x 满足

$$f(x) \leqslant \min_{x \in \Omega} f(x) + \epsilon,$$

其中$\epsilon > 0$是某个正误差. 算法的代价与目标误差有关. 高度实用的算法通常和ϵ的多项式有关，诸如 $O(1/\epsilon)$或者甚至$O(1/\epsilon^2)$. 其它算法理论上达到 $O(\log(1/\epsilon))$步，但是实际计算成本高得惊人. 从技术上讲，如果想要将参数ϵ作为输入的一部分，它仅需要$O(\log(1/\epsilon))$ 比特来描述误差参数. 因此，与高于$1/\epsilon$的对数有关的算法，就它的输入大小而言不是多项式时间算法.

In this course, we will make an attempt to highlight both the theoretical performance and practical appeal of an algorithm. Moreover, we will discuss other performance criteria such as robustness to noise. How well an algorithm performs is rarely decided by a single criterion, and usually depends on the application at hand.

2 梯度法

本节学习至关重要的基石**梯度法**和一些分析其收敛行为的方法. 这里的目的是求解形如

$$\min_{x \in \Omega} \quad f(x)$$

的问题. 为了求解问题，将需要对**目标函数**(objective function) $f\colon \Omega \to \mathbb{R}$和约束集 Ω做一些假设. 在 $\Omega = \mathbb{R}^n$的情况下， 称作**无约束**(unconstrained)优化问题.

证明严格遵循Bubeck [Bub15]的课本中的对应章节.

2.1 梯度下降法

对于可微函数 f, 从一个初始点 x_1出发，基本梯度法定义成迭代更新公式

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \qquad t = 1, 2, \ldots$$

称其中的标量η_t是**步长**(step size), 有时也称作**学习率**(learning rate), 其可以随着 t变化. 有各种方式可以选取步长，这些选取方式对梯度下降的性能产生重要影响. 对于本讲看到的几种步长选取方式，定理确保了梯度下降法的收敛性. 但是这些步长对于实际应用并不必是最理想的.

2.1.1 到闭凸集的投影

在约束集 Ω不是整个 $\mathbb{R}^n$的情况下，梯度更新可能跃出定义域 Ω. 如何确保$x_{t+1} \in \Omega$? 一种自然的方法是将每个迭代投影到定义域 Ω上. 像将要看到的那样，这实际上并不会使分析变得更困难，因此一开始就将它包括进来.

定义 2.1 (投影)**.** 点 y在集合Ω上的**投影 (projection)**定义为

$$\Pi_\Omega(y) \in \arg\min_{x \in \Omega} \|y - x\|_2 .$$

例子 2.2. 在欧氏球B_2上的投影恰好是规范化: $\Pi_{B_2}(y) = \frac{y}{\|y\|}$.

下面定理给出了投影存在并且唯一的充分条件，同时给出了投影的刻画. 几何直观见图2.1

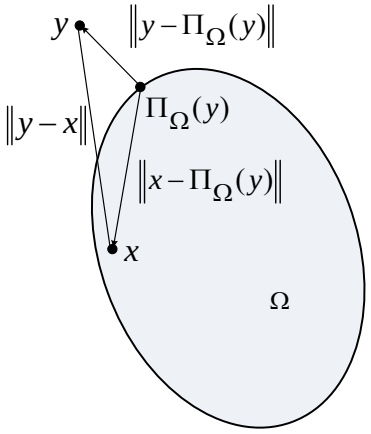


图 2.1: 投影的刻画与性质.

定理 2.3 (投影的存在唯一性及其刻画)**.** 设Ω是$\mathbb{R}^n$的非空闭凸子集. 那么$\forall y \in \mathbb{R}^n$，投影$\Pi_\Omega(y) \in \Omega$存在并且唯一. 进一步，$\pi$是$y$在$\Omega$上的投影当且仅当

$$\langle y - \pi, x - \pi \rangle \leqslant 0, \quad \forall x \in \Omega. \tag{2.1}$$

Proof. 存在性：因为$\Omega \neq \emptyset$，存在$\bar{x} \in \Omega$. 从而

$$\Pi_\Omega(y) \in \arg\min\{\|y - x\|_2 : x \in \Omega, \|y - x\|_2 \leqslant \|y - \bar{x}\|_2\}.$$

该问题的约束域是有界闭集，目标函数连续，从而由连续函数在有界闭集上能取到最小值知投影$\Pi_\Omega(y)$是存在的.

投影的刻画：任取$x \in \Omega$，并且对于$t \in (0, 1]$，定义$v = (1-t)\pi + tx$. 由于Ω是凸的，从而$v \in \Omega$. 由π的最优性有

$$\|y - \pi\|^2 \leqslant \|y - v\|^2 = \|y - \pi - t(x - \pi)\|^2.$$

将右边展开，得

$$\|y - \pi\|^2 \leqslant \|y - \pi\|^2 - 2t\langle y - \pi, x - \pi\rangle + t^2\|x - \pi\|^2.$$

这等价于

$$\langle y - \pi, x - \pi\rangle \leqslant \tfrac{t}{2}\|x - \pi\|^2.$$

因为对所有的$t \in (0, 1)$，上式都成立，令$t \to 0 + 0$得到式(2.1).

证明唯一性. 假设$\pi_1, \pi_2 \in \Omega$满足

$$\langle y - \pi_1, x - \pi_1\rangle \leqslant 0, \quad \forall x \in \Omega, \ \langle y - \pi_2, x - \pi_2\rangle \leqslant 0, \quad \forall x \in \Omega.$$

在第一个不等式中取$z = \pi_2$，第二个不等式中取$z = \pi_1$，得到

$$\langle y - \pi_1, \pi_2 - \pi_1\rangle \leqslant 0, \ \langle y - \pi_2, \pi_1 - \pi_2\rangle \leqslant 0.$$

将两个不等式相加得$\|\pi_1 - \pi_2\|^2 \leqslant 0$. 因此$\pi_1 = \pi_2$. ■

投影的一个重要性质是当 $x \in \Omega$时，对任何y(可能在Ω之外)，有

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2.$$

即x在包含y的凸集上的投影比x更接近y. 实际上，由毕达哥拉斯定理得到的一个更强的断言成立.

引理 2.4 (投影的压缩性)**.**

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2 - \|y - \Pi_\Omega(y)\|^2$$

因此，原来的步骤经修正后如 图 2.2所示. 并且保证$x_{t+1} \in \Omega$. 注意到关于问题计算的最难部分应该是计算投影. 然而，存在一些凸集，已经明确地知道如何计算投影 (比如例 2.2). 在后面的讲义中将会看到几个其它非平凡的例子.

2.2 Lipschitz函数

第一个假设是目标函数的梯度在定义域上不能太大，由其可得到收敛性分析. 这可由正常的Lipschitz连续性假设得到.

Starting from $x_1 \in \Omega$, repeat:

$$y_{t+1} = x_t - \eta g_t,\ g_t \in \partial f(x_t) \quad \text{(梯度步)}$$
$$x_{t+1} = \Pi_\Omega(y_{t+1}) \quad \text{(投影)}$$

图 2.2: 投影梯度下降法

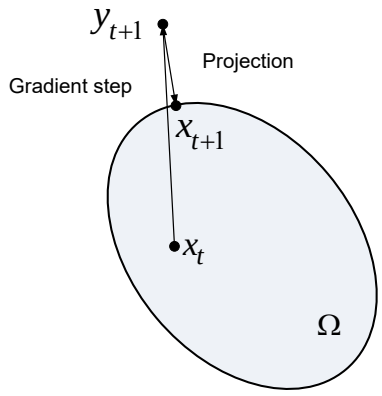


图 2.3: 投影梯度法.

定义 2.5 (L-Lipschitz)**.** 如果对任何$x, y \in \Omega$, 有

$$|f(x) - f(y)| \leqslant L\|x - y\|$$

成立，那么称函数 $f\colon \Omega \to \mathbb{R}$ 是$L-$**Lipschitz**(L-Lipschitz).

下面是$L-$Lipschitz 连续凸函数的次梯度的有界性. 该性质在梯度下降法的复杂性分析中很重要.

命题 2.6. 设$\mathcal{C} \subseteq \mathbb{R}^n$是开凸集，并且$f : \mathcal{C} \to \mathbb{R}$在集合$\mathcal{C}$上是凸的. 那么$f$ 在$\mathcal{C}$ 上是$L-$Lipschitz 连续的**当且仅当**对所有$x \in \mathcal{C}$和$g \in \partial f(x)$，有$\|g\| \leqslant L$.

证明. 假设对所有$g \in \partial f(x)$有$\|g\| \leqslant L$. 先由次梯度的定义，有

$$f(x) - f(y) \leqslant \langle g, x - y\rangle.$$

再用Cauchy-Schwartz不等式界定不等式的右边，得

$$f(x) - f(y) \leqslant \|g\|\|x - y\| \leqslant L\|x - y\|.$$

类似讨论可得

$$f(y) - f(x) \leqslant L\|x - y\|.$$

因此f是是$L-$Lipschitz连续的.

现在，假设f是$L-$Lipschitz连续的. 任取$x \in \mathcal{C}$和$g \in \partial f(x)$. 因为$\mathcal{C}$ 是开集，所以存在$\epsilon > 0$使得

$$y := x + \epsilon \frac{g}{\|g\|} \in \mathcal{C}.$$

因此$\langle y-x, g\rangle = \epsilon\|g\|, \|y-x\| = \epsilon$. 一方面，由次梯度的定义，得

$$f(y) - f(x) \geqslant \langle g, y-x\rangle = \epsilon\|g\|.$$

另一方面，由f是$L-$Lipschitz 连续的，有

$$L\epsilon = L\|y-x\| \geqslant f(y) - f(x).$$

综合上面两个不等式，得$\|g\| \leqslant L$.

综上，命题得证. □

现在能够证明梯度下降法的第一个收敛速率.

定理 2.7 (投影梯度法的复杂性)**.** 假设f在凸定义域 Ω上是L-Lipschitz的凸函数，并且在包含Ω的开集上可微. 设 R是初始点x_1与$x_* \in \arg\min_{x\in\Omega} f(x)$之间距离的上界. 设$x_1, \cdots, x_t$是投影梯度下降法计算产生的$t$步迭代序列，其中步长$\eta = \frac{R}{L\sqrt{t}}$. 那么

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x_*) \leqslant \frac{RL}{\sqrt{t}}.$$

这意味着优化过程中平均点处的函数值与最优值之差具有常数正比例于$\frac{1}{\sqrt{t}}$的上界.

在证明定理之前，回忆“最优化基本定理”(当$n = 2$时即余弦定理)，其表明内积可以写作范数的代数和:

$$u^\top v = \tfrac{1}{2}(\|u\|^2 + \|v\|^2 - \|u-v\|^2). \tag{2.2}$$

该性质源于更熟悉的恒等式 $\|u-v\|^2 = \|u\|^2 + \|v\|^2 - 2u^\top v$.

定理 2.7的证明. 首先确定函数值之差$f(x_s) - f(x_*)$的上界.

$$\begin{aligned}
f(x_s) - f(x_*) &\leqslant g_s^\top (x_s - x_*) && (\text{由} g_s \in \partial f(x_s)\text{和梯度不等式})\\
&= \tfrac{1}{\eta}(x_s - y_{s+1})^\top (x_s - x_*) && (\text{由更新规则})\\
&= \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x_*\|^2\right) && (\text{由 (2.2)})\\
&= \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 - \|y_{s+1} - x_*\|^2\right) + \tfrac{\eta}{2}\|g_s\|^2 && (\text{由更新规则})\\
&\leqslant \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 - \|y_{s+1} - x_*\|^2\right) + \tfrac{\eta L^2}{2} && (\text{Lipschitz条件})\\
&\leqslant \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2\right) + \tfrac{\eta L^2}{2} && (\text{引理 2.4})
\end{aligned}$$

现在，将这些差从 $s = 1$到$s = t$求和:

$$\begin{aligned}
\sum_{s=1}^{t}(f(x_s) - f(x_*)) &\leqslant \tfrac{1}{2\eta}\sum_{s=1}^{t}\left(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2\right) + \tfrac{\eta L^2 t}{2}\\
&= \tfrac{1}{2\eta}\left(\|x_1 - x_*\|^2 - \|x_{t+1} - x^*\|^2\right) + \tfrac{\eta L^2 t}{2} && (\text{裂项求和})\\
&\leqslant \tfrac{1}{2\eta}\|x_1 - x_*\|^2 + \tfrac{\eta L^2 t}{2} && (\text{由于}\|x_{t+1} - x_*\| \geqslant 0)\\
&\leqslant \tfrac{R^2}{2\eta} + \tfrac{\eta L^2 t}{2} && (\text{由于}\|x_1 - x_*\| \leqslant R)
\end{aligned}$$

然后，给最终得到的上述不等式两边同时除以t，再由Jensen不等式(命题 1.7)，得到

$$\begin{aligned} f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x_*) &\leqslant \frac{1}{t}\sum_{s=1}^{t} f(x_s) - f(x_*) && \text{(由凸函数的定义)} \\ &\leqslant \frac{R^2}{2\eta t} + \frac{\eta L^2}{2} && \text{(上面的不等式)} \\ &= \frac{RL}{\sqrt{t}} && \text{(对于}\eta = R/L\sqrt{t}.\text{)} \end{aligned}$$

■

2.3 光滑函数

将要遇到的下一个性质称作**光滑性**(smoothness). 光滑性的要点是允许控制Taylor近似中的二阶项. 这经常导致以相对强的假设作为代价，从而得到更强的收敛性保证.

定义 2.8 (光滑性)**.** 称连续可微函数 f是$\beta-$光滑的，如果梯度映射$\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n$是$\beta$-Lipschitz的，即，

$$\|\nabla f(x) - \nabla f(y)\| \leqslant \beta\|x - y\| \quad \forall x, y \in \Omega.$$

在分析针对光滑函数的梯度下降法之前，需要一些技术引理. 首次阅读时，跳过这些技术引理的证明不影响后续内容的理解.

引理 2.9. 设函数 f在$\mathbb{R}^n$上是β-光滑的. 那么，对每个$x, y \in \mathbb{R}^n$,

$$\left|f(y) - f(x) - \nabla f(x)^\top (y - x)\right| \leqslant \frac{\beta}{2}\|y - x\|^2.$$

Proof. 将$f(x) - f(y)$表示成积分，然后应用Cauchy-Schwarz不等式和 β-光滑性. 具体地，

$$\begin{aligned} |f(y) - f(x) - \nabla f(x)^\top (y - x)| &= \left|\int_0^1 \nabla f(x + t(y - x))^\top (y - x)dt - \nabla f(x)^\top (y - x)\right| \\ &\leqslant \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\| dt \\ &\leqslant \int_0^1 \beta t\|y - x\|^2 dt \\ &= \frac{\beta}{2}\|y - x\|^2 \end{aligned}$$

■

该引理的意义在于：选择

$$y = x - \frac{1}{\beta}\nabla f(x)$$

能得到

$$f(y) - f(x) \leqslant -\frac{1}{2\beta}\|\nabla f(x)\|^2. \tag{2.3}$$

这意味着**梯度更新能使函数值以正比例于梯度范数平方的数量减小**.

引理 2.10. 设f是β-光滑的凸函数，那么对每个$x, y \in \mathbb{R}^n$, 有

$$f(x) - f(y) \leqslant \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 .$$

Proof. 设 $z = y - \frac{1}{\beta}[\nabla f(y) - \nabla f(x)]$. 那么

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\leqslant \nabla f(x)^\top (x - z) + \nabla f(y)^\top (z - y) + \frac{\beta}{2}\|z - y\|^2 \\ &= \nabla f(x)^\top (x - y) + [\nabla f(x) - \nabla f(y)]^\top (y - z) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \\ &= \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \end{aligned}$$

这里的第一个不等式同时利用了梯度不等式和引理 2.9.

因此，由凸性和光滑性可以得到不等式. ■

将证明更新规则为

$$x_{t+1} = x_t - \eta \nabla f(x_t) \tag{2.4}$$

的梯度下降法在$\beta-$光滑性条件下，能达到更快的收敛速率. 利用上面两个引理可以证明如下结论.

定理 2.11 (梯度下降法的复杂性)**.** 设f是$\mathbb{R}^n$上的$\beta-$光滑凸函数. 那么$\eta = \frac{1}{\beta}$ 的梯度下降法 (2.4)满足

$$f(x_t) - f(x_*) \leqslant \frac{2\beta\|x_1 - x_*\|^2}{t-1}.$$

Proof. 由更新规则 (2.4)和 引理 2.9 有

$$f(x_{s+1}) - f(x_s) \leqslant -\frac{1}{2\beta}\|\nabla f(x_s)\|^2.$$

特别地，记$\delta_s = f(x_s) - f(x_*)$. 这说明

$$\delta_{s+1} \leqslant \delta_s - \frac{1}{2\beta}\|\nabla f(x_s)\|^2.$$

由凸性也有

$$\delta_s \leqslant \nabla f(x_s)^\top (x_s - x_*) \leqslant \|x_s - x_*\| \cdot \|\nabla f(x_s)\|.$$

将证明$\|x_s - x_*\|$关于s是递减的，这和上面的两个不等式蕴含着

$$\delta_{s+1} \leqslant \delta_s - \frac{1}{2\beta\|x_1 - x_*\|^2}\delta_s^2.$$

下来求解这个递推公式. 设 $w = \frac{1}{2\beta\|x_1 - x_*\|^2}$, 那么

$$w\delta_s^2 + \delta_{s+1} \leqslant \delta_s \iff w\frac{\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} \leqslant \frac{1}{\delta_{s+1}} \implies \frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geqslant w \implies \frac{1}{\delta_t} \geqslant w(t-1)$$

为了结束证明，还需要证$\|x_s - x_*\|$关于s是递减的. 使用引理 2.10, 得到

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geqslant \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 .$$

利用这个不等式和事实$\nabla f(x_*) = 0$来证明：

$$\begin{aligned}\|x_{s+1} - x_*\|^2 &= \|x_s - \tfrac{1}{\beta}\nabla f(x_s) - x_*\|^2 \\ &= \|x_s - x_*\|^2 - \tfrac{2}{\beta}\nabla f(x_s)^\top (x_s - x_*) + \tfrac{1}{\beta^2}\|\nabla f(x_s)\|^2 \\ &\leqslant \|x_s - x_*\|^2 - \tfrac{1}{\beta^2}\|\nabla f(x_s)\|^2 \\ &\leqslant \|x_s - x_*\|^2 .\end{aligned}$$

■

3 强凸性

本讲引入强凸概念，并结合光滑性发展出条件数的概念. 当由光滑性给出Taylor近似中二阶项的上界时，强凸性将提供一个下界. 当将这两个假设放在一起时，导致更快的形如$\exp(-\Omega(t))$的收敛速率，因此这两个假设相当强大. 口头上，对光滑和强凸函数，梯度下降法在每次迭代中以某严格小于 1的因子成倍地减少误差. 技术部分源于Bubeck的课本 [Bub15]的相应章节.

3.1 提醒

回忆对凸性和光滑性，各自(至少)有两个定义: 针对所有函数的一般定义和针对(二次-)可微函数的更紧凑的定义.

函数 f是凸的，如果对于每个输入，存在对函数整体有效的**线性**(linear) 下界:

$$f(y) \geqslant f(x) + g(x)^\top (y - x).$$

对于可微函数，由梯度扮演角色g. 函数f是β-光滑的，如果对于每个输入，存在关于函数整体有效的由(有限)参数β定义的**二次(quadratic)**上界:

$$f(y) \leqslant f(x) + g(x)^\top (y - x) + \tfrac{\beta}{2}\|x - y\|^2 .$$

用更诗意的表述，光滑凸函数“囿于一条抛物线和一条直线之间”. 由于β 随仿射变换是协变的，比如改变测量单位，常将 β- 光滑称作简单光滑.

对于二次可微函数，这些性质就Hessian阵(二阶偏导数为元素组成的矩阵)而言拥有简单条件. 一个C^2函数f，如果$\nabla^2 f(x) \succeq 0$，那么是凸的；并且如果$\nabla^2 f(x) \preceq \beta I$，那么是$\beta$-光滑的.

进一步定义L-Lipschitz的概念. 称函数f是L-Lipschitz的，如果它“stretches”自己输入的大小由L上控决定:

$$|f(x) - f(y)| \leqslant L\|x - y\| .$$

请注意，对可微函数而言，f的β-光滑性等价于梯度的β-Lipschitz 性.

3.2 强凸性

有了这三个概念，就能够证明梯度下降法(和它的投影、随机和次梯度版本)的两个误差衰减速率. 然而，总体来说这些速率比实践中在某种设置下所观察到的更慢.

注意到(源于凸性的)线性下界和(源于光滑性的)二次上界不具有对称性，通过将下界升级到二阶，将引入一种新的，更受限的函数类.

定义 3.1 (强凸性)**.** 函数$f\colon \Omega \to \mathbb{R}$ 是α**-强凸的(**α**-strongly convex)**，如果存在某$\alpha > 0$使得不等式

$$f(y) \geqslant f(x) + g(x)^\top (y-x) + \tfrac{\alpha}{2}\|x-y\|^2.$$

对所有$x, y \in \Omega$成立.

就像光滑性那样，经常将"α-强凸“简称为”强凸“. 强凸光滑函数是能够" 夹在两条抛物线之间的函数". 如果β-光滑性是一件好事，那么α- 凸性确保从这件好事得不到太多. 对于二次可微函数，如果 $\nabla^2 f(x) \succeq \alpha I$，那么它是 $\alpha-$强凸的. 再一次，请注意参数α在仿射变换下会发生改变.

针对α-强凸和β-光滑函数，可以定义称作**条件数**(condition number) 的量. 它与基无关，所以极为方便.

定义 3.2 (条件数)**.** α-强凸和β-光滑函数f 具有**条件数**(condition number)$\frac{\beta}{\alpha}$.

对于正定二次函数 f, 该条件数的定义与基于二次函数的矩阵条件数是一致的，而后者大家可能更熟悉.

回顾与展望. 下面的表总结了之前讲义中的结论和本讲中将要得到的结论[2]. 在两种情况下, 值 ϵ 是在由梯度下降法的输出计算得到的某个x'处的目标函数值与在最优点x_*处计算得到的目标函数值之差.

表 3.1: 将梯度下降法应用到各种函数类时，作为所采取步数t的函数得到的误差ϵ的上界.

	凸	强凸
Lipschitz	$\epsilon \leqslant O(1/\sqrt{t})$(定理 2.7)	$\epsilon \leqslant O(1/t)$ (定理 3.3)
光滑	$\epsilon \leqslant O(1/t)$(定理 2.11)	$\epsilon \leqslant e^{-\Omega(t)}$(定理 3.4)

对于就误差大小而言的指数速率就bit精度而言是线性的，因而这种速率称作**线性的(linear)**. 现在证明这些速率.

3.3 针对强凸函数的收敛速率

从强凸Lipschitz函数的收敛界开始，这种情况下得到收敛速率$O(1/t)$.

[2]这里Ω是计算复杂性符号，表示下界，大于等于的意思. 比如存在常数C使得$f(n) \geqslant Cg(n)$可记作$f(n) = \Omega(g(n))$.

定理 3.3 (针对强凸函数的梯度投影法的复杂性)**.** 假设 $f\colon \Omega \to \mathbb{R}$是$\alpha$-强凸和$L$-Lipschitz的. 设 x_* 是f 的最优点, 并设x_s 是使用投影下降法时在步s得到的更新点. 设最大迭代步数是t, 使用自适应步长$\eta_s = \frac{2}{\alpha(s+1)}$, 那么

$$f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x_*) \leqslant \frac{2L^2}{\alpha(t+1)}.$$

定理蕴含着投影梯度下降法对α-强凸函数的收敛速率与对β-光滑函数的类似, 具有误差界 $\epsilon \leqslant O(1/t)$. 为了证明 定理 3.3, 需要Jensen不等式(命题 1.7).

定理 3.3 的证明. 回忆投影梯度下降法的两步更新规则

$$\begin{aligned} y_{s+1} &= x_s - \eta_s g_s, \ \ g_s \in \partial f(x_s) \\ x_{s+1} &= \Pi_\Omega(y_{s+1}). \end{aligned}$$

首先，从探索函数值$f(x_s)$和$f(x_*)$之差开始.

$$\begin{aligned} & f(x_s) - f(x_*) \\ \leqslant\ & g_s^\top (x_s - x_*) - \tfrac{\alpha}{2}\|x_s - x_*\|^2 && \text{(由强凸性)} \\ =\ & \tfrac{1}{\eta_s}(x_s - y_{s+1})^\top (x_s - x_*) - \tfrac{\alpha}{2}\|x_s - x_*\|^2 && \text{(由更新规则)} \\ =\ & \tfrac{1}{2\eta_s}(\|x_s - x_*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x_*\|^2) - \tfrac{\alpha}{2}\|x_s - x_*\|^2 \\ & && \text{(由 "最优化基本定理")} \\ =\ & \tfrac{1}{2\eta_s}(\|x_s - x_*\|^2 - \|y_{s+1} - x_*\|^2) + \tfrac{\eta_s}{2}\|g_s\|^2 - \tfrac{\alpha}{2}\|x_s - x_*\|^2 && \text{(由更新规则)} \\ \leqslant\ & \tfrac{1}{2\eta_s}(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2) + \tfrac{\eta_s}{2}\|g_s\|^2 - \tfrac{\alpha}{2}\|x_s - x_*\|^2 && \text{(由 引理 2.4)} \\ \leqslant\ & \left(\tfrac{1}{2\eta_s} - \tfrac{\alpha}{2}\right)\|x_s - x_*\|^2 - \tfrac{1}{2\eta_s}\|x_{s+1} - x_*\|^2 + \tfrac{\eta_s L^2}{2} && \text{(由Lipschitz性)} \end{aligned}$$

给两边同时乘以s，并且代入步长$\eta_s = \frac{2}{\alpha(s+1)}$, 得到

$$s(f(x_s) - f(x_*)) \leqslant \frac{L^2}{\alpha} + \frac{\alpha}{4}\left[s(s-1)\|x_s - x_*\|^2 - s(s+1)\|x_{s+1} - x_*\|^2\right]$$

最后，能够找到定理 3.3中显示的由t步投影梯度下降法得到的函数值的上界：

$$\begin{aligned} f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) &\leqslant \sum_{s=1}^{t} \frac{2s}{t(t+1)} f(x_s) && \text{(由命题 1.7，即Jensen不等式)} \\ &\leqslant \frac{2}{t(t+1)} \sum_{s=1}^{t} \left[s f(x_*) + \frac{L^2}{\alpha} + \frac{\alpha}{4}\left[s(s-1)\|x_s - x_*\|^2 - s(s+1)\|x_{s+1} - x_*\|^2\right]\right] \\ &= \frac{2}{t(t+1)} \sum_{s=1}^{t} s f(x_*) + \frac{2L^2}{\alpha(t+1)} - \frac{\alpha}{2}\|x_{t+1} - x_*\|^2 && \text{(由裂项拆分求和)} \\ &\leqslant f(x_*) + \frac{2L^2}{\alpha(t+1)} \end{aligned}$$

由此断定，用投影梯度下降法求解具有强凸目标函数的优化问题时，收敛速率的阶为$\frac{1}{t+1}$，与具有纯粹Lipschitz性的凸函数的相比，这个更快. ■

3.4 针对光滑强凸函数的收敛速率

定理 3.4 (针对光滑强凸函数的梯度下降法的复杂性)**.** 假设$f\colon \mathbb{R}^n \to \mathbb{R}$是$\alpha$-强凸和$\beta$-光滑的. 设$x_*$是$f$的极小点, 并且设$x_t$是使用常数步长$\frac{1}{\beta}$的梯度下降法 (2.4)在步骤$t$的更新点. 那么[3]

$$\|x_{t+1} - x_*\|^2 \leqslant \exp\left(-t\tfrac{\alpha}{\beta}\right)\|x_1 - x_*\|^2.$$

为了证明定理 3.4, 需要使用如下引理. 请注意 (3.1)与 (2.3)的联系与区别.

引理 3.5. 假设 f如定理 3.4中所描述. 那么 $\forall x, y \in \mathbb{R}^n$ 和更新形式

$$x^+ = x - \tfrac{1}{\beta}\nabla f(x),$$

有

$$f(x^+) - f(y) \leqslant \nabla f(x)^\top (x - y) - \tfrac{1}{2\beta}\|\nabla f(x)\|^2 - \tfrac{\alpha}{2}\|x - y\|^2. \tag{3.1}$$

引理 3.5 的证明.

$$\begin{aligned}
&f(x^+) - f(x) + f(x) - f(y)\\
\leqslant&\nabla f(x)^\top (x^+ - y) + \tfrac{\beta}{2}\|x^+ - x\|^2 - \tfrac{\alpha}{2}\|x - y\|^2 && (\beta-\text{光滑和}L-\text{强凸})\\
=&\nabla f(x)^\top (x - y) - \tfrac{1}{2\beta}\|\nabla f(x)\|^2 - \tfrac{\alpha}{2}\|x - y\|^2 && (x^+\text{的定义})
\end{aligned}$$

■

现在用引理 3.5能够证明定理 3.4.

定理 3.4 的证明.

$$\begin{aligned}
\|x_{t+1} - x_*\|^2 &= \|x_t - \tfrac{1}{\beta}\nabla f(x_t) - x_*\|^2\\
&= \|x_t - x_*\|^2 - \tfrac{2}{\beta}\nabla f(x_t)^\top (x_t - x_*) + \tfrac{1}{\beta^2}\|\nabla f(x_t)\|^2\\
&\leqslant \left(1 - \tfrac{\alpha}{\beta}\right)\|x_t - x_*\|^2 && (\text{使用引理 3.5，其中} y = x_*, x = x_t)\\
&\leqslant \left(1 - \tfrac{\alpha}{\beta}\right)^t\|x_1 - x_*\|^2\\
&\leqslant \exp\left(-t\tfrac{\alpha}{\beta}\right)\|x_1 - x_*\|^2
\end{aligned}$$

■

针对约束情况使用投影梯度下降法也能证明相同的结论.

定理 3.6 (针对光滑强凸函数的梯度下降法)**.** 假设$f\colon \Omega \to \mathbb{R}$是$\alpha$-强凸和$\beta$-光滑的. 设$x_*$是$f$的极小点, 并且设$x_t$是使用常数步长$\frac{1}{\beta}$时投影梯度下降法在步骤$t$ 的更新点, 即使用更新规则

$$x_{t+1} = \Pi_\Omega\Big(x_t - \tfrac{1}{\beta}\nabla f(x_t)\Big),$$

其中Π_Ω 是投影算子. 那么

$$\|x_{t+1} - x_*\|^2 \leqslant \exp\left(-t\tfrac{\alpha}{\beta}\right)\|x_1 - x_*\|^2.$$

[3]请注意，这里是点列收敛.

像定理 3.4那样, 为了证明 定理 3.6， 需要使用类似于引理 3.5的如下引理.

引理 3.7. 假设f如定理 3.6 中所描述. 那么$\forall x, y \in \Omega$, 定义$x^+ = \Pi_\Omega(x - \frac{1}{\beta}\nabla f(x))$, 并且定义函数$g\colon \Omega \to \mathbb{R}$为$g(x) = \beta(x - x^+)$. 那么

$$f(x^+) - f(y) \leqslant g(x)^\top (x - y) - \frac{1}{2\beta}\|g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2.$$

引理 3.7 的证明. 对于引理 3.7 中所定义的x, x^+和y，由投影定理(定理 2.3)，得

$$\nabla f(x)^\top (x^+ - y) \leqslant g(x)^\top (x^+ - y).$$

因此， 与引理 3.5的证明类似，有

$$\begin{aligned} f(x^+) - f(x) + f(x) - f(y) &\leqslant \nabla f(x)^\top (x^+ - y) + \frac{1}{2\beta}\|g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\ &\leqslant g(x)^\top (x^+ - y) + \frac{1}{2\beta}\|g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\ &= g(x)^\top (x - y) - \frac{1}{2\beta}\|g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \end{aligned}$$ ■

再用适当的投影梯度下降更新代替标准梯度下降更新后， 用引理 3.7代替 引理 3.5， 那么定理 3.6 的证明与定理 3.4的证明完全相同.

4 梯度方法的若干应用

该讲是一系列代码的例子，可以在这里找到：

Lecture 4
(在你的浏览器中打开)

5 镜像下降法

Boosting. 对Boosting而言，以前要求φ是$L-$Lipschitz 的一元损失函数，优化问题为

$$\min_{\substack{\theta \in \mathbb{R}^M \\ \|\theta\|_1 \leqslant 1}} \widehat{R}_\varphi(f_\theta) := \frac{1}{m}\sum_{i=1}^m \varphi(-Y_i f_\theta(X_i)), \qquad \text{(Boosting)}$$

其中$f_\theta = \sum_{j=1}^M \theta_j f_j$，$f_j$为第$j$个弱分类器. 考虑用梯度下降法求解 (Boosting). 首先，$\varphi-$ 损失的梯度

$$\nabla \widehat{R}_\varphi(f_\theta) = \frac{1}{m}\sum_{i=1}^m \varphi'(-Y_i f_\theta(X_i))(-Y_i)F(X_i),$$

其中$F(x)$是列向量$[f_1(x), \cdots, f_M(x)]^\top$. 由$|Y_i| \leqslant 1$和$\varphi' \leqslant L$，再使用三角不等式以及$F(X_i)$为$M$维向量并且其分量的绝对值不大于 1，得到梯度$\ell_2$ 范数的上界：

$$\left\|\nabla \widehat{R}_\varphi(f_\theta)\right\|_2 \leqslant \frac{L}{m}\sum_{i=1}^m \|F(X_i)\| \leqslant L\sqrt{M}.$$

由于ℓ_1球的直径为 2，所以$R = 2$，并且与$\varphi-$ 风险有关的Lipschitz 常数是$L\sqrt{M}$，其中L为φ的Lipschitz常数. 误差上界$\frac{RL}{\sqrt{t}}$ 变成$2L\sqrt{\frac{M}{t}}$. 发现，为了把误差控制在$\frac{1}{m}$以内，则发现$t \sim m^2 M$，由于M很大，从而投影梯度下降法性能很差. 希望得到某种其它随着$\ln M$的增长而增长的速率，即希望有$t \sim m^2 \ln M$.

以Boosting为例，这里想在一个非欧空间(特别是ℓ_1空间)里执行梯度下降. ℓ_2 范数的对偶为其自身，然而ℓ_1范数的对偶为ℓ_∞ 或上确界范数. 如果遇到ℓ_1 约束，希望使用对偶范数. 但是如此做的原因并不直观，因为是在相同的空间$\mathbb{R}^n$ 中进行度量，可是当在其它空间里考虑最优化时，想用一个与所用度量不相同的程式. 镜像下降法能实现这个目标.

5.1 Bregman投影

定义 5.1. 如果$\|\cdot\|$为$\mathbb{R}^n$上的某种范数，则$\|\cdot\|_*$ 为其对偶范数.

例子 5.2. 设$p \geqslant 1$. 若ℓ_p范数$\|\cdot\|_p$的对偶范数为ℓ_q 范数$\|\cdot\|_q$，则$\frac{1}{p} + \frac{1}{q} = 1$；特别地当$p = 1$时，$q = +\infty$. 这就是Hölder 不等式所限定的情况.

如果x是原始空间中的向量，y属于对偶空间，通常可将$\mathbb{R}^n$ 中内积的上界提高为$x^\top y \leqslant \|x\|\|y\|_*$. 用同样的方式思考，将梯度看成原始空间上的线性算子，比如在$g_s^\top(x - x_*)$ 中，$x - x_*$ 属于原始空间，因此g_s 属于对偶空间，所以梯度属于对偶空间. 尽管所有向量都在$\mathbb{R}^n$ 中，但向量的转置表示这些向量来自不同的测度空间.

定义 5.3. 称凸集$\mathcal{D}$上的凸函数Φ关于$\|\cdot\|$而言是

(i) $L-$Lipschitz的，若$\|g\|_* \leqslant L, \forall g \in \partial\Phi(x)$,$\forall x \in \mathcal{D}$；

(ii) $\alpha-$强凸的，如果对所有的$x, y \in \mathcal{D}$和所有的$g \in \partial\Phi(x)$ 有

$$\Phi(y) \geqslant \Phi(x) + g^\top(y - x) + \tfrac{\alpha}{2}\|y - x\|^2,$$

其中$\alpha > 0$是参数.

例子 5.4. 如果Φ二次可微，其Hessian阵是$\nabla^2\Phi(x)$，$\|\cdot\|$为ℓ_2 范数，如果对所有x有$\mathrm{eig}(\nabla^2\Phi(x)) \geqslant \alpha$，那么$\Phi$关于$\ell_2$ 范数是强凸的.

定义 5.5 (Bregman散度,1967)**.** 已知凸集$\mathcal{D}$上的可微凸函数Φ 和$x, y \in \mathcal{D}$，称

$$D_\Phi(y, x) = \Phi(y) - [\Phi(x) + \nabla\Phi(x)^\top(y - x)]$$

是与Φ关联的Bregman散度.

该散度表示Φ与Φ在x处的线性近似之差. 由关于凸函数的梯度不等式知$D_\Phi(y, x) \geqslant 0$. 如果Φ是$\alpha-$ 强凸的，则$D_\Phi(y, x) \geqslant \frac{\alpha}{2}\|y - x\|^2$；并且如果二次近似是良好的，则该近似中的等式成立，并且该散度的行为像欧氏范数. 但是，请注意该量关于x 和y不是对称的.

命题 5.6 (三点性质)**.** 已知凸集$\mathcal{D}$上的凸函数Φ 和$x, y, z \in \mathcal{D}$，有

$$(\nabla\Phi(x) - \nabla\Phi(y))^\top(x - z) = D_\Phi(x, y) + D_\Phi(z, x) - D_\Phi(z, y).$$

证明. 请看上式的右边，有

$$\begin{aligned}
\text{右边} &= \Phi(x) - \Phi(y) - \nabla\Phi(y)^\top(x-y) + \Phi(z) - \Phi(x) - \nabla\Phi(x)^\top(z-x) \\
&\qquad - \left[\Phi(z) - \Phi(y) - \nabla\Phi(y)^\top(z-y)\right] \\
&= \nabla\Phi(y)^\top(y - x + z - y) - \nabla\Phi(x)^\top(z-x) \\
&= (\nabla\Phi(x) - \nabla\Phi(y))^\top(x-z).
\end{aligned}$$

□

前面分析了投影梯度下降算法的收敛性，并且证明了(定理 2.7)：考虑闭凸集Ω上的$L-$Lipschitz凸函数f，当优化步长$\eta_s = \frac{R}{L\sqrt{t}}$时，$t$次迭代后的精度为$f(\bar{x}) \leqslant f(x_*) + \frac{LR}{\sqrt{t}}$，其中$R = \|x_1 - x_*\|$.

尽管看上去好像投影梯度下降法的收敛速率与维数无关，但情况不完全如此. 回顾收敛速率的证明过程，发现：当目标函数f和约束集Ω具备优良的欧氏范数特性(也就是说，对于所有的$x \in \Omega$，$g \in \partial f(x)$，$\|x\|_2$和$\|g\|_2$ 与背景维数无关)时，收敛速率才与维数无关. 下面给出一个反例.

考虑欧氏球$B_{2,n}$上的可微凸函数f，其满足$\|\nabla f(x)\|_\infty \leqslant 1, \forall x \in B_{2,n}$. 这意味着$\|\nabla f(x)\|_2 \leqslant \sqrt{n}$，并且投影梯度下降算法以速度$\sqrt{\frac{n}{t}}$收敛到$f$ 在$B_{2,n}$上的极小值. 根据镜像下降法，得到的收敛速率是$\sqrt{\frac{\ln n}{t}}$.

为了对最优化问题得到更好的收敛速率，可使用**镜像下降算法**(Mirror Descent Algorithm, MDA). 其思想是将欧氏几何变成一个与待解决问题更相关的几何. 将采用所谓的**势函数**(potential function)$\Phi(x)$ 来定义新几何. 具体地，将利用基于Bregman 散度的Bregman 投影来定义这种几何.

镜像下降算法背后的几何直观如下：前面介绍的在任意Hilbert 空间$\mathcal{H}$ 上施行的投影梯度法将向量的范数与内积关联起来. 现在，假设对Banach 空间中开凸集$\mathcal{D}$ 上的最优化感兴趣. 换言之，所使用的范数(或者说距离的度量)不是由内积诱导出来的. 在此情况下，因为梯度$\nabla f(x)$ 是对偶空间的元素. 因此不能执行运算$x - \eta\nabla f(x)$. 所以梯度下降不合乎情理. (请注意投影梯度下降中的Hilbert 空间，也即$\mathcal{H}$ 的对偶空间与$\mathcal{H}$ 是等距的，从而不会遇到任何这样的问题.)

镜像下降算法的几何学洞察是为了在原始空间的集合$\mathcal{D}$上执行优化问题，将原始空间上的点$x \in \mathcal{D}$先映射到对偶空间$\mathcal{D}^*$，然后在对偶空间执行梯度更新，最后再将最优点映回原始空间. 请注意在每个更新步，原始空间集合$\mathcal{D}$ 中的新点有可能会跑到约束集$\Omega \subset \mathcal{D}$ 之外，在这种情况下需要将它投影到约束集Ω 上. 与镜像下降算法相关联的投影就是基于Bregman 散度概念定义的Bregman 投影.

定义 5.7 (Bregman投影)**.** 设Φ为$\mathcal{D} \subseteq \mathbb{R}^n$ 上的可微凸函数，并且闭凸集$\Omega \subset \mathrm{cl}(\mathcal{D})$. 定义$x \in \mathbb{R}^n$ 在Ω上关于Φ 的Bregman 投影为

$$\Pi_\Omega^\Phi(x) \in \operatorname*{argmin}_{z\in\Omega} D_\Phi(z, x).$$

在分析镜像下降算法的收敛速率时，需要Bregman投影的刻画. 所得结果与2.3.1 节中点在闭凸集上欧氏投影的刻画结果类似.

命题 5.8 (刻画Bregeman投影)**.** 已知$x \in \Omega \cap \mathcal{D}$. 那么

$$[\nabla\Phi(\Pi_\Omega^\Phi(x) - \nabla\Phi(x)]^\top[\Pi_\Omega^\Phi(x) - z] \leqslant 0, \forall z \in \Omega \cap \mathcal{D}$$

并且$D_\Phi(z, \Pi_\Omega^\Phi(x) \leqslant D_\Phi(z, x)$.

证明 记$\pi := \Pi_\Omega^\Phi(x)$，并定义$h(t) = D_\Phi(\pi + t(z - \pi), x)$. 由于$h(t)$在区间$[0, 1]$上的最小值在$t = 0$处取得(根据Bregman投影的定义)，有

$$h'(0) = \nabla_y D_\Phi(y, x)|_{y=\pi}(z - \pi) \geqslant 0.$$

根据Bregman散度的定义，

$$\nabla_y D_\Phi(y, x) = \nabla\Phi(y) - \nabla\Phi(x).$$

因此，

$$(\nabla\Phi(\pi) - \nabla\Phi(x))^\top(\pi - z) \leqslant 0.$$

再由三点性质，知道

$$(\nabla\Phi(\pi) - \nabla\Phi(x))^\top(\pi - z) = D_\Phi(\pi, x) + D_\Phi(z, \pi) - D_\Phi(z, x) \leqslant 0.$$

因为$D_\Phi(\pi, x) \geqslant 0$，所以$D_\Phi(z, \pi) \leqslant D_\Phi(z, x)$. □

5.2 镜像下降算法

如前所述，镜像下降算法的每一步中都应用Bregman散度将更新值投影到约束集上. 镜像下降算法的正式描述如下.

Algorithm 1 Mirror Descent algorithm

Require: $x_1 \in \arg\min_{\Omega\cap\mathcal{D}} \Phi(x), \zeta : \mathbb{R}^n \to \mathbb{R}^n$ such that $\zeta(x) = \nabla\Phi(x)$, constant $\eta > 0$.

1: **for** $s = 1$ to $t - 1$ **do**
2: $\quad\zeta(y_{s+1}) = \zeta(x_s) - \eta g_s$ for $g_s \in \partial f(x_s)$
3: $\quad x_{s+1} = \Pi_\Omega^\Phi(y_s + 1)$
4: **end for**

Ensure: Either $\bar{x} = \frac{1}{t}\sum_{s=1}^t x_s$ or $x^\circ \in \underset{x\in\{x_1,\cdots,x_t\}}{\operatorname{argmin}} f(x)$

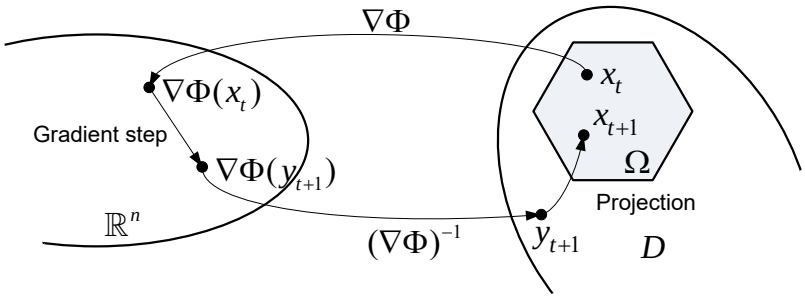


图 5.1: 镜像下降法.

在投影梯度下降法的散度证明中用到了欧氏范数如下的性质((2.2)式)：

$$2u^\top v = \|u\|^2 + \|v\|^2 - \|u-v\|^2,$$

同时选取$u = x_s - y_{s+1}$，$v = x_s - x_*$. 关于Bregman散度的三点性质(命题 5.6)表明Bregman 散度本质上表现为投影中欧氏范数的平方. 用类似的方式和Bregman 散度的三点性质证明镜像下降算法的复杂性.

定理 5.9 (镜像下降法的复杂性)**.** 设凸函数f关于$\|\cdot\|$是$L-$Lipschitz的，并且满足$x_* \in \operatorname{argmin}_\Omega f(x)$存在，$\Phi$在$\Omega \cap \mathcal{D}$ 上关于$\|\cdot\|$ 是$\alpha-$ 强凸的，并且

$$R^2 = \sup_{x\in\Omega\cap\mathcal{D}} \Phi(x) - \min_{x\in\Omega\cap\mathcal{D}} \Phi(x),$$

取$x_1 = \operatorname{argmin}_{x\in\Omega\cap\mathcal{D}} \Phi(x)$(假定它存在)，则在$\eta = \frac{R}{L}\sqrt{\frac{2\alpha}{t}}$时，由镜像下降算法得到

$$f(\bar{x}) - f(x_*) \leqslant RL\sqrt{\tfrac{2}{\alpha t}},\ \ f(x^\circ) - f(x_*) \leqslant RL\sqrt{\tfrac{2}{\alpha t}}.$$

证明 取$x^\sharp \in \Omega\cap\mathcal{D}$. 与投影梯度下降算法证明类似，有：

$$\begin{aligned}
f(x_s) - f(x^\sharp) &\overset{(i)}{\leqslant} g_s^\top (x_s - x^\sharp)\\
&\overset{(ii)}{=} \tfrac{1}{\eta}(\zeta(x_s) - \zeta(y_{s+1}))^\top (x_s - x^\sharp)\\
&\overset{(iii)}{=} \tfrac{1}{\eta}(\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}))^\top (x_s - x^\sharp)\\
&\overset{(iv)}{=} \tfrac{1}{\eta}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, y_{s+1})\right]\\
&\overset{(v)}{\leqslant} \tfrac{1}{\eta}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\right]\\
&\overset{(vi)}{\leqslant} \tfrac{\eta L^2}{2\alpha} + \tfrac{1}{\eta}\left[D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\right]
\end{aligned}$$

上式中的(i)由凸函数f在x处次梯度的定义得到. 上式中的(ii)和(iii)直接由镜像下降算法得到. 上式中的(iv)根据三点性质(命题 5.6)得到. 对于不等式(v)，由于$x_{s+1} = \Pi_\Omega^\Phi(y_{s+1})$，因此由本节Bregman 投影的刻画和性质(命题 5.8)，对于$x^\sharp \in \Omega\cap\mathcal{D}$，有$D_\Phi(x^\sharp, y_{s+1}) \geqslant D_\Phi(x^\sharp, x_{s+1})$. 下面证明不等式(vi).

$$\begin{aligned}
D_\Phi(x_s, y_{s+1}) &\overset{(a)}{=} \Phi(x_s) - \Phi(y_{s+1}) - \nabla\Phi(y_{s+1})^\top (x_s - y_{s+1})\\
&\overset{(b)}{\leqslant} [\nabla\Phi(x_s) - \nabla\Phi(y_{s+1})]^\top (x_s - y_{s+1}) - \tfrac{\alpha}{2}\|y_{s+1} - x_s\|^2\\
&\overset{(c)}{\leqslant} \eta\|g_s\|_*\|x_s - y_{s+1}\| - \tfrac{\alpha}{2}\|y_{s+1} - x_s\|^2\\
&\overset{(d)}{\leqslant} \tfrac{\eta^2 L^2}{2\alpha}.
\end{aligned}$$

由Bregman散度的定义得到等式(a). Φ是$\alpha-$强凸的事实蕴含着

$$\Phi(y_{s+1}) - \Phi(x_s) \geqslant \nabla\Phi(x_s)^\top (y_{s+1} - x_s) + \tfrac{\alpha}{2}\|y_{s+1} - x_s\|^2.$$

由此得到不等式(b). 根据镜像下降算法，$\nabla\Phi(x_s)-\nabla\Phi(y_{s+1})=\eta g_s$. 再利用Hölder不等式证明$g_s^\top(x_s-y_{s+1})\leqslant\|g_s\|_*\|x_s-y_{s+1}\|$，然后推导得到不等式(c). 对于$a,b>0$，不难证明二次项$ax-bx^2$的最大值为$\frac{a^2}{4b}$，再结合

$$x=\|y_{s+1}-x_s\|, a=\eta\|g_s\|_*\leqslant\eta L, b=\tfrac{\alpha}{2},$$

可推导得到不等式(d).

利用裂项求和得到

$$\tfrac{1}{t}\sum_{s=1}^{t}[f(x_s)-f(x^\sharp)]\leqslant\tfrac{\eta L^2}{2\alpha}+\tfrac{D_\Phi(x^\sharp,x_1)}{t\eta}. \tag{5.1}$$

根据Bregman散度定义得到

$$\begin{aligned}D_\Phi(x^\sharp,x_1)&=\Phi(x^\sharp)-\Phi(x_1)-\nabla\Phi(x_1)(x^\sharp-x_1)\\&\leqslant\Phi(x^\sharp)-\Phi(x_1)\\&\leqslant\sup_{x\in\Omega\cap\mathcal{D}}\Phi(x)-\min_{x\in\Omega\cap\mathcal{D}}\Phi(x)\\&=R^2.\end{aligned}$$

上面推导中的第一个不等式利用了$x_1\in\operatorname{argmin}_{\Omega\cap\mathcal{D}}\Phi(x)$ 蕴含着事实$\nabla\Phi(x_1)(x^\sharp-x_1)\geqslant 0$. 再将不等式代入(5.1)，并对得到的不等式的右边关于η 极小化，取$\eta=\frac{R}{L}\sqrt{\frac{2\alpha}{t}}$，得到

$$\tfrac{1}{t}\sum_{s=1}^{t}[f(x_s)-f(x^\sharp)]\leqslant RL\sqrt{\tfrac{2}{\alpha t}}.$$

取$x^\sharp$为x_*或者x°，均有$x^\sharp\in\Omega$，从而得到所需证明的结论. □

请注意选取适当的几何，投影梯度下降法就是镜像下降算法的实例.

5.3 注记

有时，镜像下降法也被称作为**镜像邻近**(Mirror Proximal)法. 可将x_{s+1} 写作

$$\begin{aligned}x_{s+1}&=\operatorname*{argmin}_{x\in\Omega}D_\Phi(x,y_{s+1})\\&=\operatorname*{argmin}_{x\in\Omega}\Phi(x)-(\nabla\Phi(y_{s+1}))^\top x\\&=\operatorname*{argmin}_{x\in\Omega}\Phi(x)-[\nabla\Phi(x_s)-\eta g_s]^\top x\\&=\operatorname*{argmin}_{x\in\Omega}\eta(g_s^\top x)+\Phi(x)-(\nabla\Phi(x_s))^\top x\\&=\operatorname*{argmin}_{x\in\Omega}\eta(g_s^\top x)+D_\Phi(x,x_s),\end{aligned}$$

这样，得到

$$x_{s+1}=\operatorname*{argmin}_{x\in\Omega}f(x_s)+g_s^\top(x-x_s)+\tfrac{1}{\eta}D_\Phi(x,x_s).$$

为了得到x_{s+1}，在右边的第一项看到了沿次梯度g_s确定的方向在靠近x_s 的线性近似. 如果函数是线性的，只需该线性近似项. 但是如果函数不是线性的，则该线性近似只在x_s 的一个小邻域内有效. 因此添加惩罚项$D_\Phi(x, x_s)$. 当选取$D_\Phi(x, x_s) = \frac{1}{2}\|x - x_s\|^2$时，使用范数的平方作为惩罚. 此时的镜像下降算法就是投影梯度下降算法.

但是，如果选取不同的散度$D_\Phi(x, x_s)$，就改变了几何，并且在与几何相关的不同方向上做不同的惩罚.

如此，使用镜像下降算法，可用另外的范数代替投影梯度下降算法中的2-范数，希望获得限制性更少的Lipschitz 常数. 另一方面，强凸参数是范数的下界. 因此，在提高收敛速率上存在着折中因子.

例子 5.10 (欧氏设置). $\Phi(x) = \|x\|_2^2, \|\cdot\| = \|\cdot\|_2, \mathcal{D} = \mathbb{R}^n, \nabla\Phi(x) = \zeta(x) = 2x$. 这里

$$\begin{aligned} D_\Phi(x, y) &= \|x\|_2^2 - \|y\|_2^2 - 2y^\top x + 2\|y\|_2^2 \\ &= \|x - y\|_2^2. \end{aligned}$$

从而更新过程与梯度下降类似. 因此，用这种势函数$\Phi(x)$的Bregman投影与平常的欧氏投影是一样的，且镜像下降算法与梯度投影下降算法完全一样，因为它们具有相同的更新和相同的投影算子. 请注意，由于

$$D_\Phi(x, y) \geqslant \tfrac{2}{2}\|x - y\|^2,$$

因此$\alpha = 2$. 这时上述定理中的步长$\eta = \frac{2R}{L\sqrt{t}}$，已验证对应的镜像下降法的迭代还原成投影梯度下降法，得到的误差上界为$\frac{LR}{\sqrt{t}}$. 这恰好是定理 2.7的结论.

例子 5.11 (ℓ_1设置). 考虑$\mathbb{R}_+^n \setminus \{0\}$. 定义$\Phi(x)$ 为负熵，因此

$$\Phi(x) = \sum_{i=1}^d x_i \ln(x_i), \qquad \zeta(x) = \nabla\Phi(x) = \{1 + \ln(x_i)\}_{i=1}^d.$$

请注意，本段分析为了表示向量的分量，从而将迭代指标作为上角标，并使用了$(\cdot)$. 对于上面的设置，由更新函数

$$\nabla\Phi(y^{(s+1)}) = \nabla\Phi(x^{(s)}) - \eta g^{(s)},$$

得到$\ln(y_i^{(s+1)}) = \ln(x_i^{(s)}) - \eta g_i^{(s)}$. 因此，对于所有的$i = 1, \cdots, n$，有

$$y_i^{(s+1)} = x_i^{(s)} \exp(-\eta g_i^{(s)}).$$

因此，

$$y^{(s+1)} = x^{(s)} \exp(-\eta g^{(s)}).$$

称该设置为指数梯度下降或者带有乘性权重的镜像下降.

该镜像映射的Bregman散度

$$\begin{aligned} D_\Phi(x, y) &= \Phi(x) - \Phi(y) - (\nabla\Phi(y))^T(x - y) \\ &= \sum_{i=1}^n x_i \ln(x_i) - \sum_{i=1}^n y_i \ln(y_i) - \sum_{i=1}^n (1 + \ln(y_i))(x_i - y_i) \\ &= \sum_{i=1}^n x_i \ln\left(\tfrac{x_i}{y_i}\right) - \sum_{i=1}^n (x_i - y_i). \end{aligned}$$

请注意，$\sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{y_i}\right) - \sum_{i=1}^{n}(x_i - y_i)$ 是x 与y 间的Kullback-Leibler 散度(KL-div).

证明：在单纯形

$$\triangle_n = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, x_i \geqslant 0 \right\}.$$

上关于该Bregman 散度的投影等同于简单的重正则化$y \mapsto y/\|y\|_1$. 要证明该结论，给出如下Lagrange 函数：

$$\mathcal{L} = \sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{y_i}\right) - \sum_{i=1}^{n}(x_i - y_i) + \lambda\left(\sum_{i=1}^{n} x_i - 1\right).$$

为了得到Bregman投影，对所有$i = 1, \cdots, n$，可写

$$\frac{\partial}{\partial x_i}\mathcal{L} = \ln\left(\frac{x_i}{y_i}\right) + \lambda = 0.$$

因此，令$\gamma = \exp(-\lambda)$, 则对所有i，有$x_i = \gamma y_i$. 知道$\sum_{i=1}^{n} x_i = 1$. 所以$\gamma = \frac{1}{\sum y_i}$. 如此，得到$\Pi_{\Delta_d}^{\Phi}(y) = \frac{y}{\|y\|_1}$. 具有这种更新和投影的镜像下降算法为：

$$\begin{aligned} y_{s+1} &= x_s \exp(-\eta g_s) \\ x_{s+1} &= \frac{y_{s+1}}{\|y_{s+1}\|_1}. \end{aligned}$$

为了分析收敛速率，要研究Δ_n上的ℓ_1范数. 因此，需要证明：对某α，Φ在Δ_n 上关于$\|\cdot\|_1$是$\alpha-$强凸函数. 演算如下：

$$\begin{aligned} D_\Phi(x, y) &= \sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{y_i}\right) - \sum_i (x_i - y_i) \\ &= \sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{y_i}\right) \\ &\geqslant \frac{1}{2}\|x - y\|_1^2. \end{aligned}$$

在上式中，利用了$x, y \in \Delta_n$的事实来证明$\sum_i (x_i - y_i) = 0$. 此外，还利用了Pinsker不等式得到上式结果. 所以，Φ 在Δ_n上是关于$\|\cdot\|_1$ 的$1-$ 强凸函数.

由于$\Phi(x) = \sum_{i=1}^{n} x_i \ln(x_i)$定义为负熵，则对$x \in \Delta_n$，$-\ln n \leqslant \Phi(x) \leqslant 0$. 因此，

$$R^2 = \max_{x \in \Delta_n} \Phi(x) - \min_{x \in \Delta_n} \Phi(x) = \ln n.$$

推论 5.12. 设Δ_n上的凸函数f满足

$$\|g\|_\infty \leqslant L, \quad \forall g \in \partial f(x), \quad \forall x \in \Delta_n.$$

则根据定理 5.9，由$\eta = \frac{1}{L}\sqrt{\frac{2\ln n}{t}}$ 的镜像下降得到

$$f(\bar{x}) - f(x_*) \leqslant L\sqrt{\frac{2\ln n}{t}}, \quad f(x^\circ) - f(x_*) \leqslant L\sqrt{\frac{2\ln n}{t}}.$$

Boosting: 对于弱分类器$f_1(x), \cdots, f_M(x)$ 以及$\theta \in \Delta_M$，定义

$$f_\theta = \sum_{j=1}^{M} \theta_j f_j, \qquad F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{pmatrix}$$

因此$f_\theta(x)$是加权的多数选票分类器. 注意，$\|F\|_\infty \leqslant 1$. 业已证明，在Boosting 中，有

$$g = \nabla \widehat{R}_\varphi(f_\theta) = \frac{1}{m} \sum_{i=1}^{m} \varphi'(-y_i f_\theta(x_i))(-y_i)F(x_i),$$

由于$\|F\|_\infty \leqslant 1$，$\|y\|_\infty \leqslant 1$，则$\|g\|_\infty \leqslant L$，此处L 为φ 的Lipschitz 常数(比如对于指数损失函数$\phi(x) = e^x$，该常数为e).

$$\widehat{R}_\varphi(f_{\theta_t^\circ}) - \min_{\theta \in \Delta_M} \widehat{R}_\varphi(f_\theta) \leqslant L\sqrt{\tfrac{2 \ln M}{t}},$$

为了误差小于$\frac{1}{m}$，需要的迭代次数$t \approx m^2 \ln M$.

函数f_j可能取到所有的顶点. 因此，如果希望把它们装入一个球内，该球的半径必须为$\sqrt{M}$. 这就是投影梯度下降速率为$\sqrt{\frac{M}{t}}$ 的原因. 但是通过审视这个梯度，可确定恰当的几何. 在这种情况下，该梯度由$\sup -$ 范数给出其界，而$\sup -$ 范数通常是投影梯度下降中最具限制性的范数. 这样，利用镜像下降将获得很大裨益.

其它势函数： 还有其它的势函数，它们关于ℓ_1 范数是强凸的. 在实践中，有

$$\Phi(x) = \tfrac{1}{p}\|x\|_p^p, \qquad p = 1 + \tfrac{1}{\ln n},$$

则Φ关于ℓ_1范数是$c\sqrt{\ln n}-$强凸的.

6 条件梯度法

在这讲中讨论条件梯度法，也称作Frank-Wolfe (FW)算法 [FW56]. 方法的动机是在有些场景下，投影梯度下降法中的投影步无法有效计算. 这时，条件梯度法提供了一种迷人的选择.

6.1 算法

条件梯度法使用一个灵活的思想避开投影步. 考虑从某个点 $x_0 \in \Omega$ 出发. 那么，对时间步从 $t = 1$ 到 T, 其中 T 是最终的时间步, 置

$$x_{t+1} = x_t + \eta_t(\bar{x}_t - x_t)$$

其中

$$\bar{x}_t = \arg\min_{x \in \Omega} f(x_t) + \langle \nabla f(x_t), x - x_t \rangle.$$

该表达式简化成:

$$\bar{x}_t = \arg\min_{x \in \Omega} \langle \nabla f(x_t), x \rangle.$$

请注意需要步长$\eta_t \in [0,1]$来保证 $x_{t+1} \in \Omega$. 特别地，当Ω是凸集，f是Ω上的凹函数时，由凹函数在区间上的极小值在端点处取得知步长$\eta_t = 1$.

因此，不是走一个梯度步然后投影到约束集上. 而是如图 6.1总结的那样，在约束集上优化一个线性函数.

Starting from $x_1 \in \Omega$, repeat:

$$y_t = \arg\min_{x\in\Omega} \langle \nabla f(x_t), x\rangle \qquad \text{(线性优化)}$$

$$\eta_t \in \arg\min_{\eta\in\mathbb{R}} f(x_t + \eta(\bar{x}_t - x_t)), \qquad \text{(线搜索)}$$

$$x_{t+1} = x_t + \eta_t(y_t - x_t) \qquad \text{(更新步)}$$

图 6.1: 条件梯度法

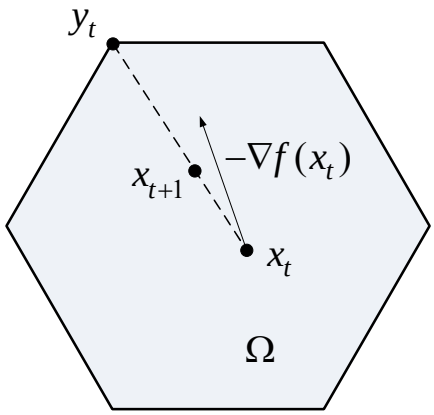


图 6.2: 条件梯度/Frank-Wolfe法.

6.2 条件梯度法的收敛分析

像将看到的那样，条件梯度法享有类似于前面已经看到的投影梯度法那样的收敛保证.

定理 6.1 (条件梯度法的复杂性)**.** 假设函数$f\colon \Omega \to \mathbb{R}$是凸的和$\beta$-光滑的，并且在点 $x_* \in \Omega$取到它的全局最小值. 那么，步长为 $\eta_t = \frac{2}{t+2}$ 的Frank-Wolfe法获得

$$f(x_t) - f(x_*) \leqslant \frac{2\beta D^2}{t+2}, \tag{6.1}$$

其中$D := \max_{x-y\in\Omega} \|x - y\|$是$\Omega$的直径.

Proof of 定理 6.1. 由光滑性知引理 2.9成立，从而有

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

将$y = x_{t+1}$和$x = x_t$代入上式，并结合条件梯度法的迭代规则中的更新步，得到:

$$f(x_{t+1}) \leqslant f(x_t) + \eta_t \langle \nabla f(x_t), \bar{x}_t - x_t \rangle + \frac{\eta_t^2 \beta}{2} \|\bar{x}_t - x_t\|^2$$

根据定理 6.1中D的定义，并且观测到$\|\bar{x}_t - x_t\|^2 \leqslant D^2$. 再由$\bar{x}_t$的最优性和$x_* \in \Omega$，可将上述不等式进一步放大，得到

$$f(x_{t+1}) \leqslant f(x_t) + \eta_t \langle \nabla f(x_t), x_* - x_t \rangle + \frac{\eta_t^2 \beta D^2}{2}.$$

由f的凸性，也有

$$\nabla f(x_t)^\top (x_* - x_t) \leqslant f(x_*) - f(x_t).$$

这样，

$$f(x_{t+1}) - f(x_*) \leqslant (1 - \eta_t)[f(x_t) - f(x_*)] + \frac{\eta_t^2 \beta D^2}{2}. \tag{6.2}$$

下面基于 (6.2)，利用归纳法证明 (6.1).

基本情况 $t = 0$. **当**$t = 0$时，有$\eta_0 = \frac{2}{0+2} = 1$. 因此由 (6.2)有

$$f(x_1) - f(x_*) \leqslant (1 - 1)[f(x_0) - f(x_*)] + \frac{\beta D^2}{2} \leqslant \frac{2\beta D^2}{3}.$$

从而，归纳假设对于基本情况成立.

归纳步. **按**归纳法，假设不等 (6.1)对所有不超过t 的正整数成立，下面来证明断言对$t + 1$也成立. 由 (6.2),

$$\begin{aligned}
f(x_{t+1}) - f(x_*) &\leqslant \left(1 - \tfrac{2}{t+2}\right)[f(x_t) - f(x_*)] + \tfrac{4}{2(t+2)^2}\beta D^2 \\
&\leqslant \left(1 - \tfrac{2}{t+2}\right)\tfrac{2\beta D^2}{t+2} + \tfrac{2}{(t+2)^2}\beta D^2 \\
&= 2\beta D^2 \cdot \tfrac{t+1}{(t+2)^2} \\
&= 2\beta D^2 \cdot \tfrac{t+1}{t+2} \cdot \tfrac{1}{t+2} \\
&\leqslant 2\beta D^2 \cdot \tfrac{t+2}{t+3} \cdot \tfrac{1}{t+2} \\
&= 2\beta D^2 \tfrac{1}{t+3}
\end{aligned}$$

这样，不等式对$t + 1$的情况也成立.

■

6.3 应用于核范数优化问题

下面例子的代码见这个链接here.

6.3.1 核范数投影

矩阵A的**核范数**(nuclear norm)(有时也称作**Schatten 1-范数**或者**迹范数**)，记作$\|A\|_*$, 定义为它的奇异值之和：

$$\|A\|_* = \sum_i \sigma_i(A)\,.$$

可用A的奇异值分解来计算核范数. 用

$$B_*^{m\times n} = \{A \in \mathbb{R}^{m\times n} \mid \|A\|_* \leqslant 1\}$$

表示核范数单位球. 如何将一个矩阵投影到B_*上? 形式上，欲求解

$$\min_{X\in B_*} \|A - X\|_F^2.$$

由于Frobenius范数具有旋转不变性，将奇异值投影到单纯形上就可以得到解. 该算子对应于将所有奇异值平移相同的参数θ并在0处截断使得平移截断值之和等于 1. 在[DSSSC08]中能找到这个算法.

6.3.2 低秩矩阵补全

假设有一个部分可观测矩阵Y, 将它的一些缺失值填充成 0 ，欲将其投影到核范数球上，以找到它的补全形式. 正式描述为

$$\min_{X\in B_*} \tfrac{1}{2}\|Y - P_O(X)\|_F^2,$$

其中P_O是投影到由O指定的X的坐标子集的线性投影算子. 在该例中，$P_O(X)$ 将产生一个矩阵，其将Y中对应观测值的元素保持作X的对应元素，其它元素补 0 . 有$P_O(X) = X \odot O$，这里$\odot$表示Hardmard积(两个矩阵的逐元素乘积得到的矩阵), 其中O 是0-1 矩阵. 计算这个函数的梯度，得到

$$\nabla f(X) = Y - X \odot O\,.$$

可以使用投影梯度下降法求解这个问题，但是使用Frank-Wolfe算法会更有效，后者需要求解线性最优化oracle

$$\tilde{X}_t \in \operatorname*{argmin}_{X\in B_*} \langle \nabla f(X_t), X\rangle.$$

为了化简这个问题，需要一个简单事实，其源于奇异值分解.

事实 6.2. 核范数单位球是秩-1 矩阵的凸包：

$$\operatorname{conv}\{uv^\top : \|u\| = \|v\| = 1, u \in \mathbb{R}^m, v \in \mathbb{R}^n\} = \{X \in \mathbb{R}^{m\times n} \mid \|X\|_* \leqslant 1\}\,.$$

由该事实得到：$\langle \nabla f(X_t), X\rangle$在由单位向量$u$和$v$确定的秩-1 矩阵$uv^\top$处取到最小值. 等价地，可以在所有单位向量$u$和$v$上极大化 $-\langle \nabla f(X_t), uv^\top\rangle$. 置 $Z = -\nabla f(X_t)$并且注意到

$$\langle Z, uv^\top\rangle = \operatorname{tr}(Z^\top uv^\top) = \operatorname{tr}(u^\top Zv) = u^\top Zv\,.$$

另一种理解它的方式：注意到核范数的对偶范数是算子范数

$$\|Z\| = \max_{\|X\|_*\leqslant 1} \langle Z, X\rangle\,.$$

两种理解均表明：为了在核范数单位球上运行Frank-Wolfe法，仅需要计算矩阵最大左奇异值的方法. 能达到此目的的方式之一就是图 6.3描述的经典幂法.

- Pick a random unit vector x_1 and let $y_1 = A^\top x / \|A^\top x\|$.
- From $t = 1$ to $T - 1$:
  - Put $x_{t+1} = \frac{Ay_t}{\|Ay_t\|}$
  - Put $y_{t+1} = \frac{A^\top x_{t+1}}{\|A^\top x_{t+1}\|}$
- Return x_T and y_T as approximate top left and right singular vectors.

图 6.3: 幂法

Part II
加速梯度法

现在研发一套技术，使用其能得到比经典梯度法更快的收敛速率. 对于二次函数这个特例，这些思想简单、自然，并且能得到在实际中具有重要意义的算法. 可将**加速**(acceleration)的主题推广到任何光滑函数和凸函数上，尽管所得到的方法在实践中并不必是优越的.

以一个注意事项结束，将看到如何在加速和**稳健性**(robustness) 之间折中. 加速梯度法天生缺乏对于噪声的稳健性，而这正是基本梯度法所享有的.

7 探索加速

本讲力图寻找比前面讲过的方法收敛的更快的方法. 为了得到这种加速方法，先考虑优化二次函数这种特殊情况. 这里基本是按照Lax的优秀课本 [Lax07] 进行阐释的.

7.1 二次函数

定义 7.1 (二次函数)**.** 二次函数$f\colon \mathbb{R}^n \to \mathbb{R}$ 形如:

$$f(x) = \tfrac{1}{2}x^T Ax - b^T x + c, \tag{QF}$$

其中 $A \in S^n, b \in \mathbb{R}^n, c \in \mathbb{R}$.

请注意，像所期待的那样，将$n = 1$代入上面的定义能恢复出熟悉的一元二次函数

$$f(x) = ax^2 + bx + c$$

其中$a, b, c \in \mathbb{R}$. 该定义的精妙之处：限制 A 是对称的. 事实上，由于对任何 $A \in \mathbb{R}^{n\times n}$ 存在对称矩阵 $\tilde{A} = \frac{1}{2}\left(A + A^T\right)$ 满足:

$$x^T Ax = x^T \tilde{A}x \; \forall x \in \mathbb{R}^n,$$

从而可以允许 $A \in \mathbb{R}^{n\times n}$，并且这也能定义相同的函数类. 限制 $A \in S^n$确保每个二次函数的表示是**唯一的**(unique).

一般二次函数(QF)的梯度和Hessian阵形如:

$$\nabla f(x) = Ax - b, \quad \nabla^2 f(x) = A.$$

倘若A是非奇异的，二次函数有唯一临界点

$$x_* = A^{-1}b.$$

当$A \succ 0$, 二次函数是**严格凸的**(strictly convex) 并且这个临界点是唯一全局极小点.

7.2 二次函数的梯度下降法

本节考虑其中A是正定的二次函数$f(x)$，特别地：存在$0 < \alpha \leqslant \beta$ 使得

$$\alpha I \preceq A \preceq \beta I.$$

这蕴含着f是α-强凸和β-光滑的.

由定理 3.6 知道：在这些条件下，恰当步长的梯度下降法以速率$\exp\left(-t\frac{\alpha}{\beta}\right)$ 线性收敛. 显然$\frac{\alpha}{\beta}$ 的大小能极大地影响收敛保证. 事实上，对于二次函数(QF) 的情况，这与矩阵A 的**条件数**有关.

定义 7.2 (条件数)**.** 设A是实矩阵. 它关于(欧氏范数的)**条件数**(condition number)是

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)},$$

即最大奇异值和最小奇异值之比.

特别地，有 $\kappa(A) \leqslant \frac{\beta}{\alpha}$; 此后, 将假设$A$对称正定，并且$\alpha, \beta$ 对应于A 的最小和最大特征值，因此$\kappa(A) = \frac{\beta}{\alpha}$. 由定理 3.4 知步长为$\frac{1}{\beta}$的梯度下降法以

$$\|x_{t+1} - x_*\|^2 \leqslant \exp\left(-t\frac{1}{\kappa}\right)\|x_1 - x_*\|^2$$

的方式收敛. 在许多情况下，函数 f 是病态的，并且 κ 易于取很大的值. 在这些情况下，为了让误差变得很小，需要$t > \kappa$，所以收敛可能会非常慢. 能够做得比这更好吗？

为了回答这个问题，分析专门针对二次函数的梯度下降法，并像以前针对任何强凸光滑函数那样推导收敛界将非常有意义. 该练习将表明在哪里损失了性能，并且使人想到能达到更好保证的方法.

定理 7.3. 假设 $f\colon \mathbb{R}^n \to \mathbb{R}$是条件数为 κ的二次函数. 设 x^* 是f 的最小点, 并且设 x_t 是在步t使用步长$\frac{1}{\beta}$的梯度下降法得到的更新点，即使用更新规则

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

那么:

$$\|x_{t+1} - x_*\|^2 \leqslant \exp\left(-\tfrac{t}{\kappa}\right)\|x_1 - x_*\|^2.$$

Proof. 考虑二次函数(QF)，那么步长为 η_t 的梯度下降法更新形如：

$$x_{t+1} = x_t - \eta_t \nabla f(x_t) = x_t - \eta_t (Ax_t - b)$$

给该式的两边减去 x_*，并利用性质 $Ax_* - b = \nabla f(x_*) = 0$:

$$\begin{aligned} x_{t+1} - x_* &= [x_t - \eta_t (Ax_t - b]) - [x_* - \eta_t (Ax_* - b)] \\ &= (I - \eta_t A)(x_t - x_*) \\ &= \prod_{s=1}^{t} (I - \eta_s A)(x_1 - x_*) . \end{aligned}$$

这样，

$$\|x_{t+1} - x_*\|_2 \leqslant \left\| \prod_{s=1}^{t} (I - \eta_s A) \right\|_2 \|x_1 - x_*\|_2 \leqslant \left[\prod_{s=1}^{t} \|I - \eta_s A\|_2 \right] \|x_1 - x_*\|_2 .$$

对所有s，置 $\eta_s = \frac{1}{\beta}$. 注意到$\frac{\alpha}{\beta} I \preceq \frac{1}{\beta} A \preceq I$, 因此:

$$\left\| I - \tfrac{1}{\beta} A \right\|_2 = 1 - \tfrac{\alpha}{\beta} = 1 - \tfrac{1}{\kappa}.$$

得到

$$\|x_{t+1} - x_*\|_2 \leqslant \left(1 - \tfrac{1}{\kappa}\right)^t \|x_1 - x_*\|_2 \leqslant \exp\left(-\tfrac{t}{\kappa}\right) \|x_1 - x_*\|_2 . \qquad \blacksquare$$

7.3 与多项式逼近的联系

上一节证明了收敛速率的上界. 本节想要提高这个上界. 为了搞清楚如何提高，思考在上述讨论中是否有疏忽的地方？一个显然的可能之处是选取的步长, 那里选$\eta_s = \frac{1}{\beta}$ 相当随意. 事实上，由选取的 η_s 序列，能够选择**任何**形如

$$p_t(a) = \Pi_{s=1}^{t} (1 - \eta_s a)$$

的t-次多项式. 注意到

$$\|x_{t+1} - x_*\| \leqslant \|p_t(A)\|_2 \|x_1 - x_*\|_2$$

其中

$$p_t(A) := \prod_{k=1}^{t} (I - \eta_k A) , \|p_t(A)\|_2 = \max_{a \in \lambda(A)} |p_t(a)| .$$

通常，不知道特征值集合 $\lambda(A)$, 但是知道所有特征值属于区间 $[\alpha, \beta]$. 放松上界，得到

$$\|p_t(A)\|_2 \leqslant \max_{a \in [\alpha, \beta]} |p_t(a)| .$$

观察到：现在想找多项式 $p_t(a)$，其在$[\alpha, \beta]$ 上的模很小，同时满足额外的规范约束 $p_t(0) = 1$.

7.3.1 一个简单的多项式解

一个简单的解是常数步长 $\eta_s = \frac{2}{\alpha+\beta}$. 注意到

$$\max_{a\in[\alpha,\beta]} \left|1 - \frac{2}{\alpha+\beta}a\right| = \frac{\beta-\alpha}{\alpha+\beta} \leqslant \frac{\beta-\alpha}{\beta} = 1 - \frac{1}{\kappa},$$

这复原了前面证明的同样的收敛速率. 图 7.1显示了针对 $t = 3$次和$t = 6$次, $\alpha = 1$，$\beta = 10$ 所得到的多项式 $p_t(a)$. 请注意将次数 3 加倍仅使得多项式在$[\alpha, \beta]$上绝对值的最大值减半，这解释了为什么收敛很慢.

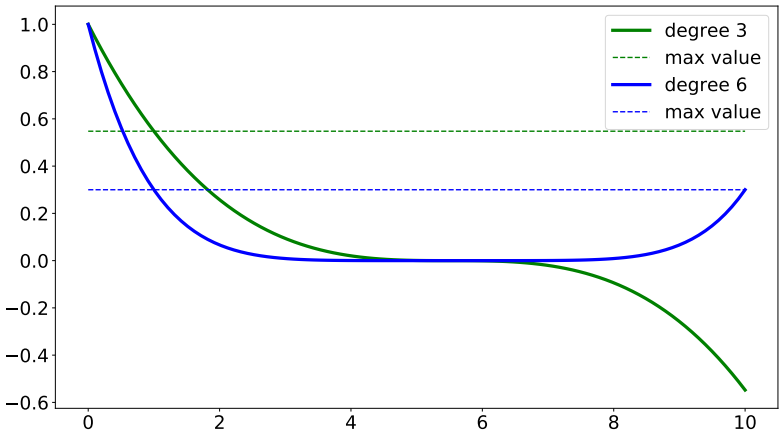


图 7.1: 朴素多项式

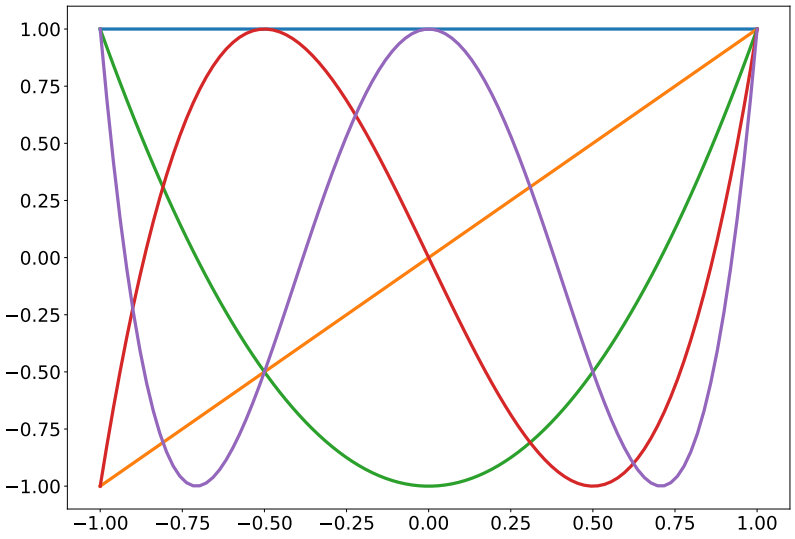


图 7.2: 前5个Chebyshev多项式：$T_0, T_1, \cdots, T_4$，这里定义中的$a = 1$.

7.4 Chebyshev多项式

幸运的是，当使用Chebyshev多项式来加速梯度下降法时，所得结果比这个更好. 这里使用由递归关系确定(第一种定义)的Chebyshev多项式:

$$T_0(a) = 1, \quad T_1(a) = a$$
$$T_{t+1}(a) = 2aT_t(a) - T_{t-1}(a), \text{ 当} t \geqslant 1 .$$

图7.2画出了前面几个Chebyshev多项式的图形. 请注意这里Chebyshev多项式的支集是$[-1, 1]$.

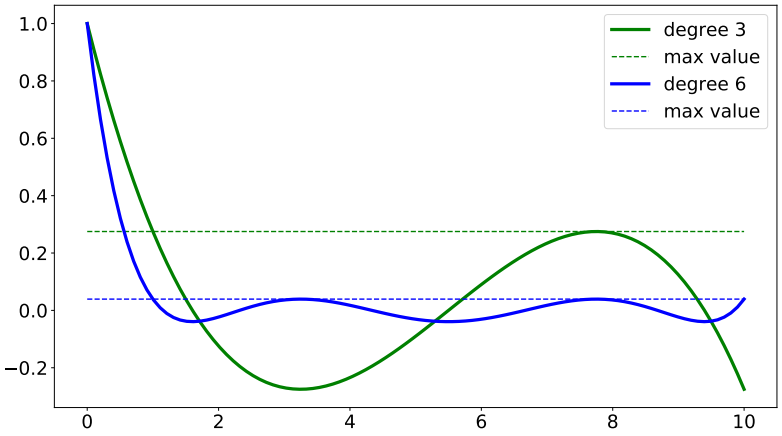


图 7.3: 重伸缩后的 3 次和 6 次Chebyshev多项式，其中$\alpha = 1, \beta = 10$.

为什么是Chebyshev多项式？经过适当伸缩，它们在感兴趣的区间$[\alpha, \beta]$ 使得绝对值最小，同时满足规范约束，即在原点处值为 1 .

回忆正在考虑的矩阵的特征值属于区间$[\alpha, \beta]$. 需要重新伸缩Chebyshev多项式使得它们以这个区间作为支集，并且在原点处的值仍然保持为 1. 下面的多项式可以做到：

$$P_t(a) = \frac{T_t\left(\frac{\beta+\alpha-2a}{\beta-\alpha}\right)}{T_t\left(\frac{\beta+\alpha}{\beta-\alpha}\right)}.$$

从图7.3看到，次数加倍对多项式在区间$[\alpha, \beta]$上大小的影响更显著. 将图7.4中这个漂亮的Chebyshev多项式与早前看到的朴素多项式进行比较. Chebyshev多项式表现得更好: 3-次多项式绝对值的最大值大约为0.3 (用朴素多项式需要 6 次), 6- 次多项式的在 0.1 以下.

7.4.1 加速梯度下降法

由Chebyshev多项式可得梯度下降法的加速版本. 在描述迭代过程之前，先看下由Chebyshev多项式产生了怎样的误差界.

为此，就是多项式在区间$[\alpha, \beta]$上有多大？请注意是在 α处取到最大值. 将这代入重伸缩的切比雪夫多项式的定义，对任何$a \in [\alpha, \beta]$得到上界：

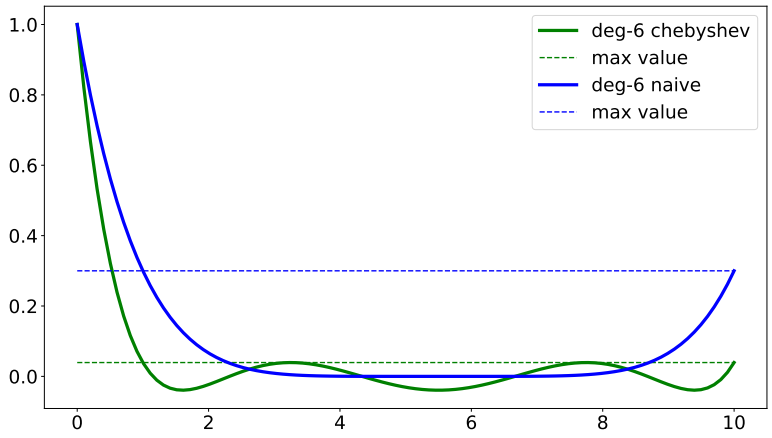


图 7.4: 重伸缩的Chebyshev多项式与朴素多项式

$$|P_t(a)| \leqslant |P_t(\alpha)| = \frac{|T_t(1)|}{\left|T_t\left(\frac{\beta+\alpha}{\beta-\alpha}\right)\right|} \leqslant \left|T_t\left(\frac{\beta+\alpha}{\beta-\alpha}\right)^{-1}\right|.$$

回忆条件数$\kappa = \beta/\alpha$，从而有

$$\frac{\beta+\alpha}{\beta-\alpha} = \frac{\kappa+1}{\kappa-1}.$$

通常 κ很大，如此上式形如$1+\epsilon$, 其中$\epsilon \approx \frac{2}{\kappa}$. 因此，有

$$|P_t(a)| \quad \leqslant |T_t(1+\epsilon)^{-1}|.$$

为了确定$|P_t|$的上界，需要确定 $|T_t(1+\epsilon)|$ 的下界.

事实: 对$a > 1$, $T_t(a) = \cosh\left(t \cdot \operatorname{arccosh}(a)\right)$，其中

$$\cosh(a) = \frac{e^a+e^{-a}}{2}, \quad \operatorname{arccosh}(a) = \ln\left(a + \sqrt{a^2-1}\right).$$

现在，设 $\phi = \operatorname{arccosh}(1+\epsilon)$:

$$e^{\phi} = 1 + \epsilon + \sqrt{2\epsilon+\epsilon^2} \geqslant 1 + \sqrt{\epsilon}.$$

因此，能给出$|T_t(1+\epsilon)|$的下界:

$$\begin{aligned} |T_t(1+\epsilon)| &= \cosh\left(t \cdot \operatorname{arccosh}(1+\epsilon)\right) \\ &= \cosh(t\phi) \\ &= \frac{(e^{\phi})^t + (e^{-\phi})^t}{2} \\ &\geqslant \frac{(1+\sqrt{\epsilon})^t}{2}. \end{aligned}$$

那么，对等的是为算法的误差确定上界，所以有：

$$|P_t(a)| \leqslant |T_t k(1+\epsilon)^{-1}| \leqslant 2(1+\sqrt{\epsilon})^{-t}.$$

如此表明Chebyshev多项式获得的误差界:

$$\begin{aligned}\|x_{t+1}-x_*\| &\leqslant 2(1+\sqrt{\epsilon})^{-t}\|x_1-x_*\| \\ &\approx 2\left(1+\sqrt{\tfrac{2}{\kappa}}\right)^{-t}\|x_1-x_*\| \\ &\leqslant 2\exp\left(-t\sqrt{\tfrac{2}{\kappa}}\right)\|x_1-x_*\|.\end{aligned}$$

这意味着对于大的κ，在误差指数地减小之前，需要的次数获得二次节省. 图7.5显示了不同的收敛速率. 能明显地看到，随着ϵ的增大，二者的差别显著增加.

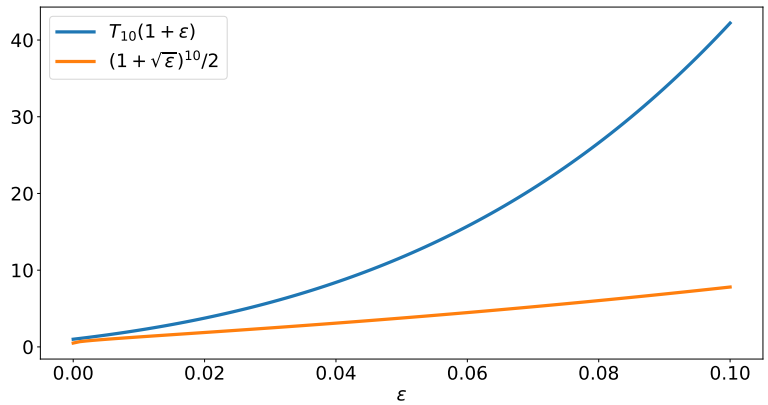


图 7.5: 朴素多项式和Chebyshev多项式的收敛

7.4.2 Chebyshev多项式的递归关系

由Chebyshev多项式的递归定义，直接可得递归算法. 为此，先将递归定义转换成重伸缩后的切比雪夫多项式，有:

$$P_{t+1}(a) = (\gamma_t - \eta_t a)P_t(a) + \mu_t P_{t-1}(a).$$

其中的系数η_t, γ_t, μ_t能由递归定义算出来. 由于 $P_t(0) = 1$, 从而必有 $\gamma_t + \mu_t = 1$. 由此得到迭代的简单更新规则:

$$\begin{aligned}x_{t+1} &= (\gamma_t - \eta_t A)x_t + \mu_t x_{t-1} + \eta_t b \\ &= (-\eta_t A + (1-\mu_t))x_t + \mu_t x_{t-1} + \eta_t b \\ &= x_t - \eta_t(Ax_t - b) + \mu_t(x_{t-1} - x_t).\end{aligned}$$

看到除过多出来的项 $\mu_t(x_{t-1} - x_t)$外，上面的更新规则实际上与未加修正项的梯度下降法非常相似. 可将这一项解释成**动量**(momentum) 项，沿着以前前进的方向推进算法. 在下一讲，将深挖动量项，并看如何将针对二次函数的结论推广到一般凸函数.

8 Krylov子空间、特征值和共轭梯度法

这一讲将研发求解特征值问题$Ax = \lambda x$和线性方程组$Ax = b$的统一方法. 特别地, 将对表8.1进行解释.

表 8.1: 求解特征值问题和线性方程组的统一方法

方法	$Ax = b$	$Ax = \lambda x$
基本	梯度下降法	幂法
加速	Chebyshev迭代	Chebyshev迭代
加速无步长	共轭梯度法	Lanczos

上次看到的是基本梯度下降法和求二次函数极小点的Chebyshev迭代. Chebyshev迭代要求**精心选取步长**. 本节, 将观看如何得到一个"不受步长限制"(step-size free)的加速方法, 称作**共轭梯度法**(conjugate gradient). 将所有这一切联系起来的是Krylov子空间的概念和与之关联的低次多项式. 这里的阐释源于 Trefethen-Bau [TD97]中的第VI章.

8.1 Krylov子空间

讨论的方法均具有性质：迭代产生的点列包含在称作**Krylov** 的子空间中.

定义 8.1 (Krylov子空间)**.** 对于矩阵$A \in \mathbb{R}^{n\times n}$和向量$b \in \mathbb{R}^n$, $t-$阶**Krylov序列**是由$b, Ab, A^2b, \cdots, A^tb$生成的. 定义**Krylov 子空间**为

$$K_t(A, b) = \text{span}\{b, Ab, A^2b, \cdots, A^tb\} \subseteq \mathbb{R}^n .$$

Krylov子空间与多项式逼近问题有着自然联系. 为了看清这一点, 回忆 t-次矩阵多项式展开: $p(A) = \sum_{i=0}^{t} \alpha_i A^i$.

事实 8.2 (与多项式的联系)**.** Krylov子空间满足

$$K_t(A, b) = \{p(A)b : \deg(p) \leqslant t\} .$$

Proof. 请注意

$$v \in K_t(A, b) \Longleftrightarrow \exists \alpha_i : \ v = \alpha_0 b + \alpha_1 Ab + \cdots + \alpha_t A^t b.$$

■

从现在开始, 假设对称矩阵 $A \in \mathbb{R}^{n\times n}$拥有单位正交特征向量$u_1, \cdots, u_n$和按模有序的特征值$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$. 这意味着

$$\langle u_i, u_j\rangle = 0, \quad \text{如果} i \neq j; \quad \langle u_i, u_i\rangle = 1.$$

使用 $A = \sum_j \lambda_j u_j u_j^\top$, 得到

$$p(A)u_j = p(\lambda_j)u_j .$$

现在假设用A的特征基将 b写作

$$b = \alpha_1 u_1 + \cdots + \alpha_n u_n, \tag{8.1}$$

其中$\alpha_j = \langle u_j, b\rangle$. 得到

$$p(A)b = \alpha_1 p(\lambda_1)u_1 + \alpha_2 p(\lambda_2)u_2 + \cdots + \alpha_n p(\lambda_n)u_n\,. \tag{8.2}$$

8.2 求特征向量

已知这些思想, 一种自然的求特征向量的方法是求多项式 p 使得

$$p(A)b \approx \alpha_1 u_1\,.$$

理想地, 式 (8.2)中将有 $p(\lambda_1) = 1$并且对 $i > 1$ 有$p(\lambda_i) = 0$, 但是这一般是不可能的, 除非让多项式的次数和A的互不相同特征值的个数一样大. 谨记最终的次数决定了迭代算法需要的步数. 因此想使它尽可能地小. 由此将产生最大特征值的良好近似. 这就是为什么选择满足$p(\lambda_1) = 1$, 并且$\max_{i>1} |p(\lambda_i)|$ 尽可能地小这样的近似解了. 实践中, 并不能提前知道 λ_1的值. 由这里的讨论知道，实际上关心的是比值, 以便不管λ_1如何, 第二大特征值将被p映到更小的值.

考虑如下的简单多项式 $p(\lambda) = \lambda^t$, 其满足

$$\frac{p(\lambda_2)}{p(\lambda_1)} = \left(\frac{\lambda_2}{\lambda_1}\right)^t$$

在$\lambda_1 = (1+\epsilon)\lambda_2$的情况下, 需要$t = O(1/\epsilon)$ 以便使得比值很小.

下一个引理将小比值演变成关于最大特征值的近似结果. 为了陈述这个引理, 用$\tan\angle(a,b)$表示向量a与b之间夹角的正切.

引理 8.3. $\tan\angle(p(A)b, u_1) \leqslant \max_{j>1} \frac{|p(\lambda_j)|}{|p(\lambda_1)|} \tan\angle(b, u_1)$.

Proof. 定义$\theta = \angle(b, u_1)$. 由 (8.1)和向量夹角的定义, 得到

$$\begin{aligned}
\cos^2\theta &= \tfrac{1}{\|b\|_2^2}|\alpha_1|^2,\\
\sin^2\theta &= \tfrac{1}{\|b\|_2^2}\sum_{j>1}\alpha_j^2,\\
\tan^2\theta &= \sum_{j>1}\tfrac{|\alpha_j^2|}{|\alpha_1|^2}.
\end{aligned}$$

类似地，由 (8.2)，有

$$\tan^2\angle(p(A)b, u_1) = \sum_{j>1}\frac{|p(\lambda_j)\alpha_j|^2}{|p(\lambda_1)\alpha_1|^2} \leqslant \max_{j>1}\frac{|p(\lambda_j)|^2}{|p(\lambda_1)|^2}\sum_{j>1}\frac{|\alpha_j|^2}{|\alpha_1|^2}$$

注意到上面不等式最后的和项$\sum_{j>1}\frac{|\alpha_j|^2}{|\alpha_1|^2} = (\tan\theta)^2$, 由此得到想要的结论. ■

对$p(\lambda) = \lambda^t$和$\lambda_1 = (1+\epsilon)\lambda_2$应用引理, 得到

$$\tan\angle(p(A)b, u_1) \leqslant (1+\epsilon)^{-t} \tan\angle(b, u_1)\,.$$

如果λ_1与λ_2之间存在大间隙, 这将使得收敛会很快, 但是如果$\lambda_1 \approx \lambda_2$, 它将会收敛地很慢. 再进一步, 通过重复地乘以A, 也可以看出表达式 $p(A)b = A^t b$ 能够构造迭代. 为了数值稳定性, 在每次矩阵-向量乘之后规范化是明智的. 这保持了迭代是同方向的, 因此不会改变收敛性分析. 随之产生的算法就是著名的幂法, 它的递归定义如下:

$$x_0 = \frac{b}{\|b\|}, \quad x_t = \frac{Ax_{t-1}}{\|Ax_{t-1}\|}.$$

该方法追溯到一百多年前由Müntz于1913写的一篇论文, 但是在今天仍能继续发现幂法的新应用.

8.3 应用Chebyshev多项式

和针对二次函数研发所期望的那样, 对上面提出的多项式近似问题, 可以使用切比雪夫多项式得到更好的解. 思想一摸一样, 微小的区别是规范化Chebyshev多项式时略有不同. 此时此刻, 想要确保 $p(\lambda_1) = 1$以便用正确的伸缩挑出最大特征值.

引理 8.4. 恰当地重伸缩 *t*次*Chebyshev*多项式达到

$$\min_{p(\lambda_1)=1} \max_{\lambda\in[\lambda_2,\lambda_n]} |p(\lambda)| \leqslant \frac{2}{(1+\max\{\sqrt{\epsilon},\epsilon\})^t}$$

其中 $\epsilon = \frac{\lambda_1}{\lambda_2} - 1$量化了第一大和第二大特征值之间的间隙.

请注意, 当 ϵ很小时, 这个界要比以前的那个更好. 对二次函数极小化的情况, 相应的"ϵ-值" 是条件数的倒数. 针对特征值, 这变成了第一大和第二大特征值之间的间隙. 表8.2是这些结论的总结.

表 8.2: 应用Chebyshev多项式得到的误差

问题	$Ax = b$	$Ax = \lambda x$
ϵ	$\frac{1}{\kappa} = \frac{\alpha}{\beta}$	$\frac{\lambda_1}{\lambda_2} - 1$

像前面看到的那样, 用Chebyshev多项式满足的递归关系能推导出获得如上所示上界的迭代法. 该方法的主要缺陷是它需要第一大和第二大特征值的位置. 不描述这个算法, 而是继续讨论一个不需要任何这样的信息也能工作的算法.

8.4 共轭梯度法

这时, 再转回针对对称正定矩阵 $A \in \mathbb{R}^{n\times n}$ 的线性方程组 $Ax = b$. 将要学习的方法被称作**共轭梯度法**(conjugate gradient), 是求解线性方程组的重要算法. 与之相对的是求特征值的Lanczos 方法. 尽管这些方法背后的思想是相似的, 线性方程的形况稍微直观些.

定义 8.5 (共轭梯度法)**.** 想要求解 $Ax = b$, 其中$A \succ 0$ 是对称的. 共轭梯度法维持三个点列:

$$\begin{aligned} x_0 &= 0 && (\text{“候选解”}) \\ r_0 &= b && (\text{“余量”}) \\ p_0 &= r_0 = -\nabla f(x_0) && (\text{“搜索方向”}) \end{aligned}$$

对 $t \geqslant 1$:

$$\begin{aligned} p_{t-1}^{\mathrm{mv}} &= Ap_{t-1} && (\text{“ 矩阵-向量乘”}) \\ \eta_t &= \frac{\|r_{t-1}\|^2}{\langle p_{t-1}, p_{t-1}^{\mathrm{mv}} \rangle} && (\text{“ 步长”}) \\ x_t &= x_{t-1} + \eta_t p_{t-1} \\ r_t &= r_{t-1} - \eta_t * p_{t-1}^{\mathrm{mv}} \\ p_t &= r_t + \frac{\|r_t\|^2}{\|r_{t-1}\|^2} p_{t-1} \end{aligned}$$

引理 8.6. 对于共轭梯度法如下三个方程总成立*:*

(i) $\mathrm{span}\{r_0, r_1, \cdots, r_t\} = K_t(A, b)$

(ii) 对 $j \leqslant t$ 有 $\langle r_{t+1}, r_j \rangle = 0$, 并且特别地 $r_{t+1} \perp K_t(A, b)$.

(iii) 搜索方向是共轭的*:* 对$i \neq j, p_i^\top A p_j = 0$.

Proof. 用归纳法可以证明(参见Trefethen and Bau). 证明这些条件初始时成立, 并且应用更新规则后仍然成立即可. ■

引理 8.7. 设 $\langle u, v \rangle_A = u^\top A v$, $\|u\|_A = \sqrt{u^\top A u}, e_{t+1} = x_{t+1} - x_*$. 那么 $e_{t+1} = \min_{x \in K_t(A,b)} \|x - x_*\|_A$, 即

$$x_{t+1} \in \arg \min_{x \in K_t(A,b)} \|x_* - x\|_A.$$

Proof. 由迭代格式, 有 $x_{t+1} \in K_t(A, b)$. 设 $x \in K_t(A, b)$. 定义 $\delta = x - x_{t+1}$. 那么

$$e := x - x_* = \delta + e_{t+1}.$$

用A范数计算误差:

$$\begin{aligned} \|e\|_A^2 = \|x - x_*\|_A^2 &= (\delta + e_{t+1})^\top A (\delta + e_{t+1}) \\ &= e_{t+1}^\top A e_{t+1} + \delta^\top A \delta + 2 e_{t+1}^\top A \delta \end{aligned}$$

由定义 $A \top e_{t+1} = r_{t+1}$. 请注意 $\delta \in K_t(A, b)$. 由 引理 8.6(ii), 有 $r_{t+1} \perp K_t(A, b)$. 因此, $e_{t+1}^\top A \delta = 0$. 将此带入上式，得

$$\|e\|_A^2 = \|x - x_*\|_A^2 = e_{t+1}^\top A e_{t+1} + \delta^\top A \delta \geqslant \|e_{t+1}\|_A^2 .$$

最后一步利用了 $A \succ 0$. ■

引理 8.6表明，共轭梯度法具有二次终止性，即对于二次函数，有限步可以得到精确解. 此外，这表明共轭梯度法本质上求解多项式逼近问题:

$$\min_{p:\ \deg(p)\leqslant t, p(0)=1} \|p(A)e_0\|_A .$$

此外，不难证明

$$\min_{p:\ \deg(p)\leqslant t, p(0)=1} \frac{\|p(A)e_0\|_A}{\|e_0\|_A} \leqslant \min_{p:\ \deg(p)\leqslant t, p(0)=1} \max_{a\in\Lambda(A)} |p(a)| .$$

换句话说，共轭梯度法得到的误差并不比关于右端多项式近似的误差糟糕，而切比雪夫多项式近似求解该问题. 由此得到就$\|\cdot\|_A$- 范数而言，共轭梯度法至少和Chebyshev迭代一样快.

9 Nesterov加速梯度下降法

前面，针对极小化二次函数(QF), 其中A 是对称正定矩阵，看到如何来加速梯度下降法. 特别地，与标准梯度下降法相比，加速版获取了与矩阵A 的条件数无关的二次提高. 产生的更新规则形如

$$x_{t+1} = x_t - \eta_t \nabla f(x_t) + \mu_t (x_t - x_{t-1}) ,$$

可将其中最后一项解释为“动量". 有时也将以这种简单形式呈现的更新规则称作Polyak **重球法**(heavy ball method).

为了针对一般光滑凸函数得到如已经看到的针对二次函数相同的加速收敛速率，必须更加地努力. 下面研究Nesterov著名的**加速梯度法**(accelerated gradient method) [Nes83, Nes04]

具体而言，将看到Nesterov的方法对于$\beta-$光滑函数得到的收敛速率为 $\mathcal{O}\left(\frac{\beta}{t^2}\right)$. 对于$\alpha$-强凸且$\beta-$光滑的函数，得到的收敛速率是 $\exp\left(-\Omega\left(\sqrt{\frac{\beta}{\alpha}}t\right)\right)$[4].

更新规则比朴素动量规则稍微复杂些，具体如下:

$$\begin{aligned} x_0 &= y_0 = z_0, \\ x_{t+1} &= \tau z_t + (1-\tau) y_t, t \geqslant 0, \\ y_t &= x_t - \tfrac{1}{\beta}\nabla f(x_t), t \geqslant 1, \\ z_t &= z_{t-1} - \eta \nabla f(x_t), t \geqslant 1. \end{aligned} \tag{9.1}$$

这里，参数 β是正在极小化的函数的光滑性常数. 下面将选取步长η 和参数 τ以便给出收敛保证.

[4]这里Ω是计算复杂性符号，表示下界，大于等于的意思. 比如存在常数C使得$f(n) \geqslant Cg(n)$可记作$f(n) = \Omega(g(n))$.

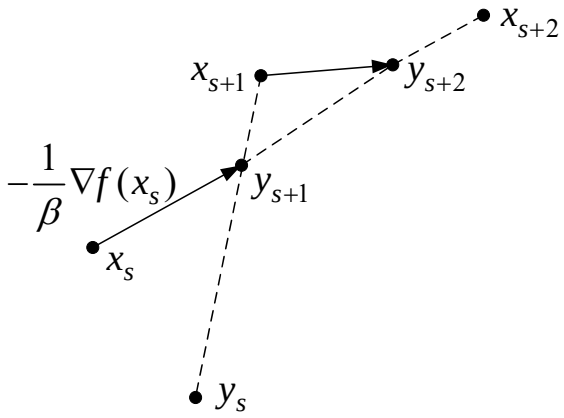


图 9.1: Nesterov加速梯度法

9.1 收敛分析

首先证明，针对一种简单的步长设置，算法能将它的初始误差从某个值 d减小到 $d/2$. 那么可以重复地重新开始算法以继续减小误差. 这稍微偏离了Nesterov的方法，因为其不需要重新开始，只是要求更细致的步长策略，但那样会让分析变复杂.

引理 9.1. 假设 $f\colon \mathbb{R}^n \to \mathbb{R}$ 是β-光滑的凸函数，它在点 $x_* \in \mathbb{R}^n$取到它的最小值. 假设初始点满足 $\|x_0 - x_*\| \leqslant R$ 并且$f(x_0) - f(x_*) \leqslant d$. 置 $\eta = \frac{R}{\sqrt{d\beta}}$, 并选取 τ 使得 $\frac{1-\tau}{\tau} = \eta\beta$. 那么在 $T = 4R\sqrt{\beta/d}$ 步之后, 迭代平均值 $\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$ 满足

$$f(\bar{x}) - f(x_*) \leqslant d/2\,.$$

Proof. 当在第 2.3 节引入光滑性的时候, 看到 引理 2.9 蕴含着 (2.3)式，即

$$f(y_t) - f(x_t) \leqslant -\tfrac{1}{2\beta}\|\nabla f(x_t)\|^2\,. \tag{9.2}$$

由式 (2.2)的 “最优化基本定理” , 对所有$u \in \mathbb{R}^n$ 有：

$$\langle \eta\nabla f(x_{t+1}), z_t - u\rangle = \tfrac{\eta^2}{2}\|\nabla f(x_{t+1})\|^2 + \tfrac{1}{2}\|z_t - u\|^2 - \tfrac{1}{2}\|z_{t+1} - u\|^2\,. \tag{9.3}$$

将(9.2)式代入上式，得到

$$\eta\langle \nabla f(x_{t+1}, z_t - u\rangle \leqslant \eta^2\beta(f(x_{t+1}) - f(y_{t+1})) + \tfrac{1}{2}\|z_t - u\|^2 - \tfrac{1}{2}\|z_{t+1} - u\|^2 \tag{9.4}$$

为了能够变成这些项的裂项求和，计算如下差值

$$\begin{aligned}
&\eta\langle \nabla f(x_{t+1}), x_{t+1} - u\rangle - \eta\langle \nabla f(x_{t+1}), z_t - u\rangle \\
=&\eta\langle \nabla f(x_{t+1}), x_{t+1} - z_t\rangle \\
=&\tfrac{1-\tau}{\tau}\eta\langle \nabla f(x_{t+1}), y_t - x_{t+1}\rangle \quad \text{由 (9.1)式} \\
\leqslant&\tfrac{1-\tau}{\tau}\eta(f(y_t) - f(x_{t+1})) \quad \text{(由梯度不等式).}
\end{aligned} \tag{9.5}$$

结合(9.4)和(9.5), 并置 $\frac{1-\tau}{\tau} = \eta\beta$，那么针对所有$u \in \mathbb{R}^n$ 产生：

$$\eta\langle\nabla f(x_{t+1}), x_{t+1}-u\rangle \leqslant \eta^2\beta[f(y_t)-f(y_{t+1})]+\tfrac{1}{2}\|z_t-u\|^2-\tfrac{1}{2}\|z_{t+1}-u\|^2.$$

像定理 2.7中的分析所作的那样，对 $u=x_*$ 应用这个不等式，将其从 $t=0$ 到 $T-1$ 求和，并利用裂项求和，得

$$\eta T[f(\bar{x})-f(x_*)] \leqslant \sum_{t=0}^{T-1}\eta\langle\nabla f(x_{t+1}), x_{t+1}-x_*\rangle \leqslant \eta^2\beta d+R^2,$$

重新整理，由于 $\eta=R/\sqrt{d\beta}$和 $T\geqslant 4R\sqrt{\beta/d}$，得

$$f(\bar{x})-f(x_*) \leqslant \tfrac{\eta\beta d}{T}+\tfrac{R^2}{\eta T}=\tfrac{2\sqrt{\beta d}}{T}R \leqslant \tfrac{d}{2}.$$

■

这个引理出现在Allen-Zhu和Orecchia [AZO17]的工作中，他们将Nesterov的方法解释成分析梯度下降法的两种方式的耦合. 一个是 (9.2) 中的不等式，通常用在针对光滑函数的梯度下降法的分析中. 另一个是 (9.3)式，一般被用在针对非光滑函数的收敛性分析中. 前者出现在定理 2.11中，后者出现在定理 2.7 中.

定理 9.2. 在引理 9.1的假设下, 通过重复地重新开始算法，能够找到点x 使得最多在$O(R\sqrt{\beta/\epsilon})$次梯度迭代后，有

$$f(x)-f(x_*)\leqslant\epsilon.$$

Proof. 由 引理 9.1, 存在某常数C，用$CR\sqrt{\beta/d}$次梯度更新，能将误差从d降到$d/2$. 用上一轮运行的输出作为每次运行的初始值，因此能够将初始误差d逐次降到$d/2$，降到$d/4$，诸如此类，直到运行$O(\log(d/\epsilon))$次算法后，得到误差 ϵ. 所作的梯度步总数是

$$CR\sqrt{\beta/d}+CR\sqrt{2\beta/d}+\cdots+CR\sqrt{\beta/\epsilon}=O\left(R\sqrt{\beta/\epsilon}\right).$$

请注意，在差一个常数因子的意义上，算法最后一次运行控制了总步数. ■

9.2 强凸情况

可以证明引理 9.1的一个变形，其可应用于函数也是 α- 强凸的场景，并最终得到线性收敛速率. 所用思想是常用技巧：将针对光滑函数的收敛速率转换成在定义域内使用强凸性得到的收敛速率.

引理 9.3. 在引理 9.1的假设和函数 f 是 α-强凸的额外假设下，能够在$t=O\left(\sqrt{\frac{\beta}{\alpha}}\right)$次梯度下降法更新后找到点 x 使得

$$\|x-x_*\|^2\leqslant\tfrac{1}{2}\|x_0-x_*\|^2.$$

Proof. 请注意 $\|x_0-x_*\|^2\leqslant R^2$, 对误差参数 $\epsilon=\frac{\alpha}{4}\|x_0-x_*\|^2$ 应用定理 9.2，仅需要进行 $O\left(\sqrt{\beta/\alpha}\right)$步迭代，就能找到点 x 使得

$$f(x)-f(x_*)\leqslant\tfrac{\alpha}{4}\|x_0-x_*\|^2.$$

由强凸性的定义得到

$$\tfrac{\alpha}{2}\|x-x_*\|^2\leqslant f(x)-f(x_*).$$

综合这两个不等式给出需要证明的命题. ■

从引理看出，针对强凸函数，每步实际上将迭代点到最优点的距离减小常数因子倍. 因此重复应用引理能得到线性收敛速率.

表 9.1比较了当将Nesterov方法和普通梯度下降法应用于不同函数时，得到的关于误差$\epsilon(t)$的界，这些界均是总迭代步数t的函数.

表 9.1: 关于误差 ϵ的上界，其是不同方法所需迭代次数 t的函数.

f的性质	普通梯度下降法	Nesterov的加速梯度下降法
β-光滑, 凸	$O(\beta/t)$(定理 2.11)	$O\left(\beta/t^2\right)$(定理 9.2)
β-光滑, α-强凸	$\exp\left(-\Omega(t\alpha/\beta)\right)$(定理 3.4)	$\exp\left(-\Omega(t\sqrt{\alpha/\beta})\right)$

10 下界与稳健性之间的权衡

本讲的第一部分，研究以前得到的各种方法的收敛速率是否是紧的. 针对几类最优化问题(光滑、强凸等), 表明答案的确是肯定的. 这种分析最精彩的部分是证明Nesterov 的加速梯度法所达到的速率 $O(1/t^2)$对于光滑凸函数(在弱技术意义下)是最优的.

本讲的第二部分，跳出收敛速率的研究，关注对比算法的其它方面. 说明加速梯度法提高的速率是以对噪声的稳健性为代价的. 特别地，如果限制只能使用近似梯度，标准梯度法基本不会变慢，而加速梯度法的误差关于迭代次数是线性累积的.

10.1 下界

在开始讨论下界之前，先简要回顾下截至目前所得到的上界是有益的. 针对凸函数 f, 表 10.1总结了假设和前面已经证明的速率. 下面说明表 10.1中列出的复杂性都是最优的.

表 10.1: 讲义 2-8中的上界

f的性质	算法	速率
凸、Lipschitz	梯度下降法	$RL/\sqrt{t}$ (定理 2.7)
强凸、Lipschitz	梯度下降法	$L^2/(\alpha t)$(定理 3.3)
凸、光滑	加速梯度下降法	$\beta R^2/t^2$(定理 9.2)

使用梯度下降法的某种变形，可以得到表 10.1中的每一种速率. 可将这些算法看作将过往的点和次梯度$(x_1, g_1, \cdots, x_t, g_t)$映射成新点 x_{t+1}的程序(preceduce). 为了证明下界，限制到与这些程序相似的算法类上考虑. 形式上，定义黑盒方法如下.

定义 10.1 (黑盒程序)**. 黑盒程序**产生点列 $\{x_t\}$使得

$$x_{t+1} \in x_1 + \text{span}\{g_1, \cdots, g_t\},$$

其中$g_s \in \partial f(x_s), s = 1, \cdots, t.$

自始至终，将进一步假设$x_1 = 0$. 像所预期的，梯度下降法是一个黑盒程序. 的确，把迭代展开，可将x_{t+1}表示为

$$\begin{aligned} x_{t+1} &= x_t - \eta g_t \\ &= x_{t-1} - \eta g_{t-1} - \eta g_t \\ &= x_1 - \sum_{i=1}^{t} \eta g_i. \end{aligned}$$

现在证明针对任何黑盒程序的收敛速率的下界. 第一个定理是关于非光滑函数在约束集上的极小化的情况. 定理来自[Nes83], 但是陈述遵循[Nes04].

定理 10.2 (带约束的非光滑f)**.** 设 $t \leqslant n$, $L, R > 0$. 存在 L-Lipschitz的凸函数 f 使得黑盒程序满足

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2(R)} f(x) \geqslant \frac{RL}{2(1+\sqrt{t})}. \tag{10.1}$$

进一步，存在L-Lipschitz的α-强凸函数 f使得

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2(\frac{L}{2\alpha})} f(x) \geqslant \frac{L^2}{8\alpha t}. \tag{10.2}$$

证明策略就是构造一个凸函数f， 以便对任何黑盒程序， $\operatorname{span}\{g_1, \cdots, g_t\} \subset \operatorname{span}\{e_1, \cdots, e_t\}$, 其中 e_t 是第 t个标准基向量. 对于 $t < n$, 在t步后，至少有 $n - t$ 个坐标精确地是 0, 由每个坐标的误差下界恰好是零得到定理.

Proof. 对于某$\gamma, \alpha \in \mathbb{R}_{++}$，考虑函数[5]

$$f(x) = \gamma \max_{1 \leqslant i \leqslant t} x[i] + \tfrac{\alpha}{2}\|x\|^2.$$

这里$\alpha > 0, \gamma > 0$是待定参数. 函数 f显然是α-强凸的. 由次微分计算法则，得

$$\partial f(x) = \alpha x + \gamma \operatorname{conv}\left\{e_i \colon i \in \operatorname*{argmax}_{1 \leqslant j \leqslant t} x[j]\right\}.$$

进一步, 如果 $\|x\| \leqslant R$ 和$g \in \partial f(x)$, 那么 $\|g\| \leqslant \alpha R + \gamma$, 因此 f 在在$B_2(R)$上是$(\alpha R + \gamma)$-Lipschitz的.

考虑 $x_* \in \mathbb{R}^n$ 使得

$$x_*[i] = \begin{cases} -\frac{\gamma}{\alpha t} & \text{如果} 1 \leqslant i \leqslant t \\ 0 & \text{否则.} \end{cases}$$

由于 $0 \in \partial f(x_*)$, 所以x_* 是f的最小点，其目标值

$$f(x_*) = \frac{-\gamma^2}{\alpha t} + \frac{\alpha}{2}\frac{\gamma^2}{\alpha^2 t} = -\frac{\gamma^2}{2\alpha t}.$$

下面构造例子所需要的参数.

[5]为了避免与迭代指标混淆，本节用$x[i]$表示x的第i个分量.

假设梯度oracle返回$g_i = \alpha x + \gamma e_i$, 其中 i 是满足$x[i] = \max_{1\leqslant j\leqslant t} x[j]$的第一个坐标. 用归纳法可以证明

$$x_s \in \mathrm{span}\{e_1, \cdots, e_{s-1}\}, \quad s \geqslant 2.$$

结果，对于$s \leqslant t$, $f(x_s) \geqslant 0$. 因此 $f(x_s) - f(x_*) \geqslant \frac{\gamma^2}{2\alpha t}$.

通过恰当地选取α和γ，即可完成证明. 具体地，在凸Lipschitz情况下，α和γ 都是自由参数. 置

$$\alpha = \frac{L}{R}\frac{1}{1+\sqrt{t}} \quad , \quad \gamma = L\frac{\sqrt{t}}{1+\sqrt{t}}.$$

那么，f是L-Lipschitz的，并且

$$\|x_1 - x_*\| = \|x_*\| = \sqrt{t\left(\frac{-\gamma}{\alpha t}\right)^2} = \frac{\gamma}{\alpha\sqrt{t}} =: R$$

因此

$$f(x_s) - \min_{x\in B_2(R)} f(x) = f(x_s) - f(x_*) \geqslant \frac{\gamma^2}{2\alpha t} = \frac{RL}{2(1+\sqrt{t})}.$$

在强凸情况下，γ是自由参数. 置$\gamma = \frac{L}{2}$，并取 $R = \frac{L}{2\alpha}$. 那么，f是L-Lipschitz的，并且有

$$\|x_1 - x_*\| = \|x_*\| = \frac{\gamma}{\alpha\sqrt{t}} = \frac{L}{2\alpha\sqrt{t}} = \frac{R}{\sqrt{t}} \leqslant R,$$

因此

$$f(x_s) - \min_{x\in B_2(L/2\alpha)} f(x) = f(x_s) - f(x_*) \geqslant \frac{LR}{4t} = \frac{L^2}{8\alpha t}.$$

■

接下来，研究光滑凸的情况，并证明加速梯度下降法达到的速率$O(1/t^2)$是最优的. 与以前的定理类似，证明策略是展示一个病态凸函数. 在这种情况下，选取Nesterov称作的所谓“世界上最差的函数” [Nes04].

定理 10.3 (光滑-f)**.** 设 $t \leqslant \frac{n-1}{2}$, $\beta > 0$. 存在 β-光滑的二次凸函数f 使得黑盒方法满足

$$\min_{1\leqslant s\leqslant t} f(x_s) - f(x_*) \geqslant \frac{3\beta\|x_1 - x_*\|_2^2}{32(t+1)^2}. \tag{10.3}$$

Proof. 不失一般性，设 $n = 2t + 1$. 设 $L \in \mathbb{R}^{n\times n}$是三对角矩阵

$$L = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}.$$

矩阵 L是圈的Laplace矩阵.[6] 请注意

$$x^\top Lx = x[1]^2 + x[n]^2 + \sum_{i=1}^{n-1}(x[i] - x[i+1])^2,$$

并且由这个表达式，易于验证 $0 \preceq L \preceq 4I$. 定义如下的β-光滑函数

$$f(x) = \frac{\beta}{8}x^\top Lx - \frac{\beta}{4}\langle x, e_1\rangle.$$

它的最小点x_*满足$Lx_* = e_1$, 求解该方程得到

$$x_*[i] = 1 - \tfrac{i}{n+1},$$

对应的目标函数值

$$\begin{aligned} f(x_*) &= \frac{\beta}{8}{x_*}^\top Lx_* - \frac{\beta}{4}\langle x_*, e_1\rangle \\ &= -\frac{\beta}{8}\langle x_*, e_1\rangle = -\frac{\beta}{8}(1 - \tfrac{1}{n+1}). \end{aligned}$$

如果$x_1 = 0$, 类似于定理 10.2的证明, 由f的构造可以证明

$$x_s \in \operatorname{span}\{e_1, \cdots, e_{s-1}\},$$

那么对任何黑盒程序，当$i \geqslant s$时有$x_s[i] = 0$. 设

$$x_*^s = \operatorname*{argmin}_{x:\, i\geqslant s, x[i]=0} f(x).$$

考虑由L 的前$s-1$ 行和前$s-1$列确定的$(s-1)\times(s-1)$的Laplace矩阵，注意到x_*^s的前$s-1$个分量是这个子Laplace矩阵定义的方程组的解，因此

$$x_*^s[i] = \begin{cases} 1 - \frac{i}{s} & \text{如果} i < s \\ 0 & \text{否则,} \end{cases}$$

对应的目标值$f(x_*^s) = -\frac{\beta}{8}(1 - \frac{1}{s})$. 因此，对任何$s \leqslant t$,

$$\begin{aligned} f(x_s) - f(x_*) &\geqslant f(x_*^t) - f(x_*) \\ &\geqslant \frac{\beta}{8}\left(\frac{1}{t} - \frac{1}{n+1}\right) \\ &\geqslant \frac{\beta}{8}\left(\frac{1}{t+1} - \frac{1}{2(t+1)}\right) \\ &= \frac{\beta}{8}\frac{1}{2(t+1)}. \end{aligned}$$

[6] https://en.wikipedia.org/wiki/Laplacian_matrix

最后，确定最优解与初始点之间的距离. 回忆 $x_1 = 0$,

$$\begin{aligned}\|x_1 - x_*\|^2 = \|x_*\|^2 &= \sum_{i=1}^{n} \left(1 - \tfrac{i}{n+1}\right)^2 \\ &= n - \tfrac{2}{n+1}\sum_{i=1}^{n} i + \tfrac{1}{(n+1)^2}\sum_{i=1}^{n} i^2 \\ &\leqslant n - \tfrac{2}{n+1}\sum_{i=1}^{n} i + \tfrac{1}{(n+1)^2}\int_1^{n+1} x^2\,dx \\ &\leqslant n - \tfrac{2}{n+1}\tfrac{n(n+1)}{2} + \tfrac{1}{(n+1)^2}\tfrac{(n+1)^3}{3} \\ &= \tfrac{(n+1)}{3} \\ &= \tfrac{2(t+1)}{3}.\end{aligned}$$

综合上面两个演算，对任何$s \leqslant t$,

$$f(x_s) - f(x_*) \geqslant \tfrac{\beta}{8}\tfrac{1}{2(t+1)} \geqslant \tfrac{3\beta\|x_1 - x_*\|^2}{32(t+1)^2}.$$

■

10.2 稳健性与加速之间的折中

该课程的第一部分基本上完全集中在最优化算法的收敛速率上. 从这个方面看，收敛速率越快，算法越好. 止步于收敛速率的最优化算法理论是不完整的. 经常存在其它重要的算法设计目标，比如对噪声或者数值误差的鲁棒性, 重视收敛速率而忽略了它, 当这些目标变成首要的时，过分强调速率会导致从业者选取错误的算法. 这节处理这种情况.

狭义上，上面几节技术上的价值是：Nesterov加速梯度下降法是一个“最优”算法，具备匹配它的收敛速率的上界和下界. 盲目地一味追求收敛速率表明人们总应该使用Nesterov方法. 在为Nesterov方法加冕前，考虑有噪声时它的表现具有重要意义.

图 10.1 比较了普通梯度下降法和Nesterov加速梯度下降法对于定理 10.3中证明的函数f. 在无噪声情况下，加速方法(NAG) 获得预期在梯度下降法(GD)之上的提速. 然而，如果给梯度增加少量球面噪声，不仅提速消失了，而且梯度下降法开始胜过加速方法，后者在若干次迭代后开始发散.

上面的例子在任何意义上不是邪恶地病态. 相反地，它说明了一种普遍现象. Devolder, Glineur 和Nesterov [DGN14]的工作表明：在如下精确意义下，加速和稳健性之间存在基本的权衡.

首先，定义非精确梯度oracle的概念. 回忆对于一个β-光滑的凸函数 f 和任何 $x, y \in \Omega$, 有

$$0 \leqslant f(x) - [f(y) + \langle \nabla f(y), x - y\rangle] \leqslant \tfrac{\beta}{2}\|x - y\|^2. \tag{10.4}$$

对任何 $y \in \Omega$, 精确一阶oracle返回一对 $(f(y), g(y)) = (f(y), \nabla f(y))$ 其保证对每个 $x \in \Omega$， (10.4)式精确成立. 一个非精确oracle，返回一对使得 (10.4)对某个松弛的 δ 成立.

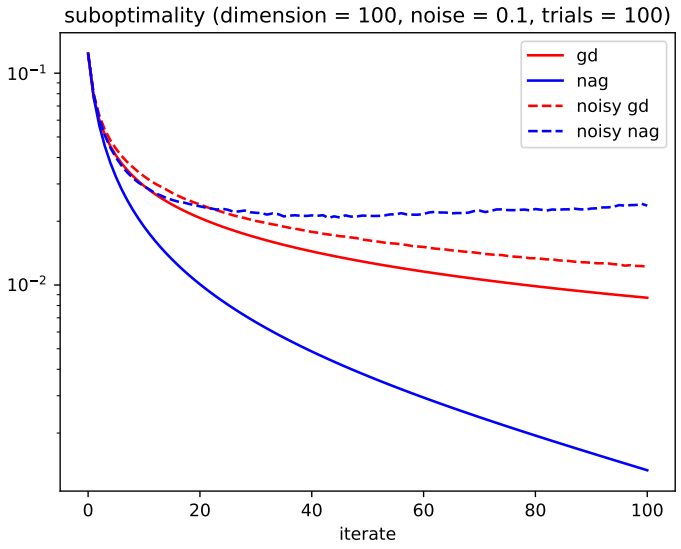


图 10.1: 将梯度下降法(GD)和Nesterov加速梯度下降法(NAG)应用到$n = 100$的世界上最差的函数时，最优性间隙随迭代的变化. 请注意，加速极大地得益于精确梯度. 然而，当给梯度增加均匀半径$\delta = 0.1$的球面噪声后，随机梯度下降法(noisy GD) 仍然表现稳健，而随机加速梯度下降法(noisy NAG)会累积误差. 随机方法的数据是100 次试验结果的平均.

定义 10.4 (非精确oracle)**.** 设 $\delta > 0$. 对任何 $y \in \Omega$, δ-非精确oracle返回一对 $(f_\delta(y), g_\delta(y))$ 使得对每个 $x \in \Omega$,有

$$0 \leqslant f(x) - [f_\delta(y) + \langle g_\delta(y), x - y\rangle] \leqslant \frac{\beta}{2}\|x - y\|^2 + \delta\,.$$

考虑返回δ-非精确oracle的梯度下降法. Devolder等 [DGN14]证明，在t步之后，

$$f(x_t) - f(x_*) \leqslant \frac{\beta R^2}{2t} + \delta.$$

将该速率与表 10.1中的相比，普通梯度下降法不受非精确oracle的影响，没有误差积累. 另一方面，如果运行δ-非精确oracle的加速梯度法，那么在 t 步之后，

$$f(x_t) - f(x_*) \leqslant \frac{4\beta R^2}{(t+1)^2} + \frac{1}{3}(t+3)\delta.$$

换句话说，加速梯度法关于迭代次数线性地累积误差! 此外，这个松弛不是分析的产物. 像如下定理所精确描述的，任何黑盒法如果在非精确情况下被加速，它必定会累积误差.

定理 10.5 ([DGN14]，定理6)**.** 考虑收敛速率为$O\left(\frac{\beta R^2}{t^p}\right)$的黑盒法使用非精确oracle的情况. 用$\delta$-非精确oracle, 假设算法获得速率

$$f(x_t) - f(x_*) \leqslant O\left(\frac{\beta R^2}{t^p}\right) + O\left(t^q \delta\right), \tag{10.5}$$

那么$q \geqslant p - 1$.

特别地，对任何$p > 1$的加速方法，结果有 $q > 0$，因此方法关于迭代次数累积的误差至少为 $O(t^{p-1}\delta)$.

Part III
随机优化

这部分介绍随机优化，以此来总结优化技术的研究工作. 当能给出梯度噪声的上界时，将要研究的工具对于最小化$\varphi-$ 风险很有效.

11 随机梯度法

考虑随机函数$x \mapsto \ell(x, Z)$，其中x为优化参数，Z 为随机变量. 设P_Z 是Z 的分布. 假设$x \mapsto \ell(x, Z)$ 关于P_Z是几乎处处凸的. 特别地，$f(x) := \mathbb{E}_{Z\sim P_Z}[\ell(x, Z)]$ 也将是凸的. 随机凸优化的目的是

$$\min_{x\in\Omega} f(x) := \mathop{\mathbb{E}}_{Z\sim P_Z}[\ell(x, Z)]. \tag{SO}$$

本课程感兴趣的问题中Ω是确定的凸集. 然而，随机凸优化可定义得更为宽泛. 约束本身也可以是随机的. 接下来，将针对Ω 是确定的情况进行介绍. 处理过的一些优化问题也可用该新框架来阐述. 当分布P_Z的支集是有限的时，一般可将函数写作

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{11.1}$$

其中函数$f_1, \cdots, f_m$关于x是凸的. 为了求解 (11.1)，分析随机梯度法和它的典型应用.

11.1 风险极小化与经验风险极小化

有两个对象空间$\mathcal{X}$和$\mathcal{Y}$, 其中将$X \in \mathcal{X}$看作**实例**(instance)或者**样例**(example) 空间, 将$Y \in \mathcal{Y}$ 看作**标签**(label) 或者**类别**(class)集合. 目的是**学习**(learn)函数$h\colon \mathcal{X} \to \mathcal{Y}$，当已知$X \in \mathcal{X}$时，由它输出对象$Y \in \mathcal{Y}$. 假设$Z = (X, Y)$ 在空间 $\mathcal{X} \times \mathcal{Y}$ 上服从联合分布 P_Z. 已知从P_Z抽取的m个独立同分布样例 $S = ((X_1, Y_1), \ldots, (X_m, Y_m))$. 现在定义非负实值损失函数 $\ell(y', y)$ 来度量预测值 y'和真实结果y之间的差异.

定义 11.1. 函数$h\colon \mathcal{X} \to \mathcal{Y}$的**风险**(risk) 定义为

$$R(h) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}} \ell(h(X), Y)\,.$$

学习算法的终极目标是在函数类$\mathcal{H}$中找到极小化$R(h)$ 的h^*:

$$h^* \in \arg\min_{h\in\mathcal{H}} R(h).$$

当将函数类参数化后，相应的$\ell(h(X), Y)$变成$\ell(x, Z)$，其中x是确定函数h的参数. 由此得到随机优化 (SO). 下面是一些典型例子.

Boosting. 典型的集成学习方法Boosting，其对应于函数类

$$\mathcal{H} = \left\{ f_\theta = \sum_{j=1}^{M} \theta_j h_j(\cdot) : \theta \in \mathbb{R}^M, \|\theta\|_1 \leqslant 1, \text{分类器 } h_j : \mathcal{X} \mapsto [-1, 1], j \in \{1, 2, \cdots, M\} \right\},$$

和指数损失函数$\varphi(z) = e^{-z}$. 从而Boosting的目标是极小化风险：

$$\min_{\theta\in\Delta_M} \mathbb{E}[e^{-Yf_\theta(X)}],$$

其中$f_\theta = \sum_{j=1}^{M} \theta_j h_j(\cdot), \theta \in \Delta_M$. 这里$\Delta_M$是$\mathbb{R}^M$中的$M$维单纯形. 定义$Z = (X, Y)$ 以及随机函数$\ell(\theta, Z) = \varphi(-Yf_\theta(X))$，则其关于$P_Z$在几乎处处的含义下关于$\theta$是凸的.

线性回归. 该问题的目标是极小化ℓ_2- 风险：

$$\min_{\theta\in\mathbb{R}^d} \mathbb{E}[(Y - f_\theta(X))^2],$$

其中$f_\theta(X) = \theta^\top X$. 定义$Z = (X, Y)$以及随机函数$\ell(\theta, Z) = (Y - f_\theta(X))^2$，则其关于$P_Z$ 在几乎处处的含义下关于θ是凸的.

最大似然. 考虑独立同分布样本$Z_1, \cdots, Z_m$，其概率密度为p_θ，$\theta \in \Theta$. 比如$Z \sim \mathcal{N}(\theta, 1)$. 与该组样本相关的似然函数，对应于映射$\theta \mapsto \prod_{i=1}^m p_\theta(Z_i)$. 令$p^*(Z)$ 表示Z的真实密度(它不必具有p_θ 的形式，其中$\theta \in \Theta$)，那么

$$\frac{1}{m}\mathbb{E}\left[\ln\prod_{i=1}^m p_\theta(Z_i)\right] = -\int \ln\left(\frac{p^*(z)}{p_\theta(z)}\right)p^*(z)dz + C = -\mathrm{KL}(p^*, p_\theta) + C,$$

其中C是与θ无关的常数. 因此极大化对数似然的期望等效于极小化Kullback-Leibler散度：

$$\max_\theta \mathbb{E}\left[\ln\prod_{i=1}^m p_\theta(Z_i)\right] \Longleftrightarrow \min_\theta \mathrm{KL}(p^*, p_\theta).$$

外部随机化(External randomization)与经验风险极小化. 通常，因为联合分布未知，所以无法计算风险$R(h)$. 因此，取而代之的是极小化所谓的**经验风险**(empirical risk). 经验风险定义为损失函数在训练集上的平均值:

$$R_S(h) = \frac{1}{m}\sum_{i=1}^m \ell(h(x_i), y_i).$$

经验风险最小点(empirical risk minimizer, ERM)是任何 $h^* \in \arg\min_{h\in\mathcal{H}} R_S(h)$. 易见经验风险极小化问题形如 (11.1).

假设要最小化 (11.1) 如所见，这在经验风险最小化里应用很多. 在这种情况下，视该问题为确定问题，但是引入如下人造随机性. 令I 为随机变量，均匀分布在$[m] := \{1, \cdots, m\}$ 上，此即为$Z = I$，$\ell(x, I) = f_I(x)$的随机凸优化，表示为$f(x) = \mathbb{E}_I[f_I(x)]$.

如果每个样本仅被使用一次，可将随机梯度法看作在直接极小化风险. 在训练集上多轮使用的情况，最好看作是在极小化经验风险，这与极小化风险给出的解不同. 后面将研究将风险与经验风险关联起来的工具.

重要注记 假设已知独立随机变量的情况和生成人工随机性的情况之间有非常关键的区别. 以Boosting为例来阐明这种区别. 已知$(X_1, Y_1), \cdots, (X_m, Y_m)$是独立并且源于某未知分布. 在第一个例子中，目标是基于这m 个观测数据极小化$\mathbb{E}[\varphi(-Yf_\theta(X))]$，并且将看到随机梯度法允许每次迭代仅取一对$(X_i, Y_i)$来极小

化$\mathbb{E}[\varphi(-Yf_\theta(X))]$. 特别地，对于每对数据数据，最多只能利用一次. 称作关于数据执行了**一轮**(one pass).

也能够利用前面讲义中的经验风险极小点的统计分析并尝试最小化 经验$\varphi-$ 风险：

$$\widehat{R}_\varphi(f_\theta) = \frac{1}{m}\sum_{i=1}^{m}\varphi(-Y_i f_\theta(X_i)).$$

具体地，产生k 个在$\{1,\cdots,m\}$ 上均匀分布的独立随机变量$I_1,\cdots,I_k$，并在每次中迭代使用一个随机变量I_j 来执行随机梯度下降. 这里的区别在于：无论观测数量m 有多大，k 可以任意大(即在数据上执行多轮). 然而，最小化

$$\mathbb{E}_I[\varphi(-Y_I f_\theta(X_I))|X_1,Y_1,\cdots,X_m,Y_m]$$

的效果并不比经验风险最小点(其统计性能受限于观测数量m) 的好.

11.2 外部随机优化问题的随机梯度法

参照Robbins-Monro [RM51], 定义求解外部随机优化问题 (11.1)的随机梯度法如下.

定义 11.2. 随机梯度算法(Stochastic gradient descent algorithm)从点$x_0 \in \Omega$ 开始，接着根据更新规则产生

$$x_{t+1} = x_t - \eta_t \nabla f_{i_t}(x_t)$$

其中 $i_t \in \{1,\cdots,m\}$在每一步随机选取，或者通过$\{1,\cdots,m\}$ 的随机置换来循环.

上面两种选择 i_t的方法都能得到事实

$$\mathbb{E}[\nabla f_{i_t}(x)|x] = \nabla f(x).$$

下面对一个简单问题应用随机梯度法，能产生最优解.

例子 11.3. 设 $p_1,\cdots,p_m \in \mathbb{R}^n$, 并定义 $f\colon \mathbb{R}^n \to \mathbb{R}_+$:

$$\forall x \in \mathbb{R}^n,\ f(x) = \frac{1}{2m}\sum_{i=1}^{m}\|x-p_i\|_2^2.$$

请注意这里 $f_i(x) = \frac{1}{2}\|x-p_i\|_2^2, \nabla f_i(x) = x - p_i$. 此外，易见该问题的最优解

$$x_* = \operatorname*{argmin}_{x\in\mathbb{R}^n} f(x) = \frac{1}{m}\sum_{i=1}^{m} p_i.$$

现在，用步长$\eta_t = \frac{1}{t}$，并按循环顺序运行随机梯度法，即$i_t = t$ 和 $x_0 = 0$:

$$\begin{aligned}
x_0 &= 0\\
x_1 &= 0 - \tfrac{1}{1}(0-p_1) = p_1\\
x_2 &= p_1 - \tfrac{1}{2}(p_1-p_2) = \tfrac{p_1+p_2}{2}\\
&\vdots\\
x_m &= \frac{1}{m}\sum_{i=1}^{m} p_i = x_*
\end{aligned}$$

1958年的纽约时报(New York Times)写道 感知器(Perceptron)[Ros58] 是:

> *the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.*

那么，来看下.

定义 11.4 (感知器)**.** 已知带标签的点 $((x_1, y_1), \cdots, (x_m, y_m)) \in (\mathbb{R}^n \times \{-1, 1\})^m$，和初始点 $w_0 \in \mathbb{R}^n$，感知器是如下算法. 对于随机均匀选取的 $i_t \in \{1, \cdots, m\}$,

$$w_{t+1} = w_t(1-\gamma) + \eta \begin{cases} y_{i_t} x_{i_t} & \text{如果} y_{i_t} \langle w_t, x_{i_t} \rangle < 1 \\ 0 & \text{否则} \end{cases}$$

其中γ和η是正参数.

逆向工程该算法，可以看出感知器等价于对正则化的支撑向量机(Support Vector Machine, SVM) 的经验风险极小化执行SGM.

例子 11.5 (SVM)**.** 已知带标签的点 $((x_1, y_1), \cdots, (x_m, y_m)) \in (\mathbb{R}^n \times \{-1, 1\})^m$, SVM的目标函数是:

$$f(w) = \frac{1}{m} \sum_{i=1}^{m} \max\{1 - y_i \langle w, x_i \rangle,\ 0\} + \frac{\lambda}{2} \|w\|_2^2$$

称损失函数 $\varphi_i(z_i) = \max\{1 - y_i z_i,\ 0\}$是合页损失(Hinge Loss)，这里$z_i = \langle w, x_i \rangle$. 称额外的项 $\lambda \|w\|_2^2$ 是正则化(regularization) 项，$\lambda > 0$ 是权衡参数.

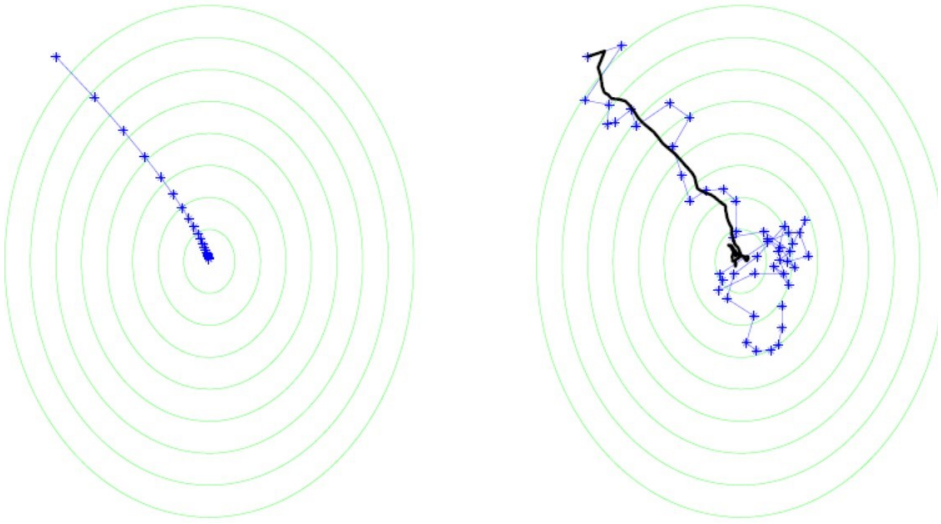


图 11.1: 求解外部随机化问题的梯度下降法和随机梯度法

11.3 随机梯度法

如果Z的分布已知，那么函数$f: x \mapsto \mathbb{E}[\ell(x,Z)]$是已知的，可应用梯度下降法、投影梯度下降法或其它优化方法进行求解，就如同求解以前的确定性问题一样. 然而，若**真实分布P_Z未知**，并且仅能给出样本$Z_1, \cdots, Z_t$和随机函数$\ell(x,Z)$. 在下面的叙述中，用$\partial\ell(x,Z)$表示函数$y \mapsto \ell(y,Z)$在x点处次梯度的集合，即次微分(是闭凸集). 定义$f_s(x) := \ell(x,Z_s)$，那么$\partial f_s(x) = \partial\ell(x,Z_s)$. 求解一般随机优化问题 (SO)的随机梯度算法的正式描述见算法 2.

Algorithm 2 Stochastic Gradient Descent algorithm

Require: $x_1 \in \Omega$, positive sequence $\{\eta_s\}_{s\geqslant 1}$, independent random variables $Z_1, \cdots, Z_t$ with distribution P_Z.

1: **for** $s = 1$ to t **do**
2: $\quad y_{s+1} = x_s - \eta_s \widehat{g}_s, \ \widehat{g}_s \in \partial f_s(x_s)$
3: $\quad x_{s+1} = \pi_C(y_{s+1})$
4: **end for**

Ensure: $\bar{x}_t = \frac{1}{t}\sum_{s=1}^{t} x_s$

因为$Z_s \sim P_Z$，所以有$\mathbb{E}[\widehat{g}_s|x_s] \in \partial f(x_s)$. 从而这里的随机梯度法是(一阶)随机oracle：输入$x \in \mathbb{R}^n$，输出一个随机变量$\widehat{g}(x)$满足**无偏假设**

$$\mathbb{E}[\widehat{g}(x)|x] \in \partial f(x). \tag{GUE}$$

其中当查询点有可能是随机变量(由以前查询点的orale得到的). 无偏假设(GUE)本身不足以得到收敛速率. 还需要对$\widehat{g}(x)$的波动性作假设. 在非光滑情况下，一个基本假设是：存在$L > 0$使得

$$\mathbb{E}[\|\widehat{g}(x)\|_*^2|x] \leqslant L^2, \forall x \in \mathbb{R}^n, \forall \widehat{g}(x) \in \partial\ell(x,Z). \tag{VB}$$

需要注意的是，与有偏oracle相关的方差有界假设与此相当不同. 除此之外，该算法与确定梯度下降法的区别在于：后者返回$\bar{x}$或者

$$x_t^\circ = \arg\min_{x\in\{x_1,\cdots,x_t\}} f(x).$$

在随机框架下，函数$f(x) = \mathbb{E}[\ell(x,Z)]$典型地是未知的，$x_t^\circ$也是不可计算的.

定理 11.6. 设Ω是$\mathbb{R}^n$的闭凸子集，并且$\mathrm{diam}(\Omega) \leqslant R$. 假定凸函数$f(x) = \mathbb{E}[\ell(x,Z)]$在$x_* \in \Omega$处取到它在$\Omega$上的最小值，还假定$\ell(x,Z)$关于$P_Z$在几乎处处的含义下关于$x$是凸的，并且方差有界假设(VB)成立. 那么，如果$\eta_s \equiv \eta = \frac{R}{L\sqrt{t}}$，有

$$\mathbb{E}[f(\bar{x}_t)] - f(x_*) \leqslant \frac{LR}{\sqrt{t}}.$$

Proof. 由$\widehat{g}_s \in \partial f_s(x_s)$，有

$$f_s(x_s) - f_s(x_*) \leqslant \widehat{g}_s^\top (x_s - x_*).$$

假设x_s已知，上式两边关于Z_s取条件期望，得

$$\begin{aligned} f(x_s) - f(x_*) &\leqslant \mathbb{E}[\widehat{g}_s^\top (x_s - x_*)|x_s] \\ &= \frac{1}{\eta}\mathbb{E}[(x_s - y_{s+1})^\top (x_s - x_*)|x_s] \\ &= \frac{1}{2\eta}\mathbb{E}[\|x_s - y_{s+1}\|^2 + \|x_s - x_*\|^2 - \|y_{s+1} - x_*\|^2|x_s] \\ &\leqslant \frac{1}{2\eta}(\eta^2\mathbb{E}[\|\widehat{g}_s\|^2|x_s] + \mathbb{E}[\|x_s - x_*\|^2|x_s] - \mathbb{E}[\|x_{s+1} - x_*\|^2|x_s]). \end{aligned}$$

将s从1到t得到的不等式求和，并关于$Z_1, \cdots, Z_{t-1}$取期望，得到

$$\mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t}[f(x_s) - f(x_*)]\right] \leqslant \frac{\eta L^2}{2} + \frac{R^2}{2\eta t}.$$

根据Jensen不等式(命题 1.7)，并选取$\eta = \frac{R}{L\sqrt{t}}$，得到

$$\mathbb{E}[f(\bar{x}_t)] - f(x_*) \leqslant \frac{LR}{\sqrt{t}}.$$

■

11.4 随机镜像下降法

可将镜像下降法扩展到如下随机镜像下降算法.

Algorithm 3 Stochastic Mirror Descent algorithm

Require: $x_1 \in \operatorname{argmin}_{\Omega\cap\mathcal{D}} \Phi(x)$, $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ such that $\zeta(x) = \nabla\Phi(x)$, independent random variables $Z_1, \cdots, Z_t$ with distribution P_Z.
  for $s = 1, \cdots, t$ **do**
    $\zeta(y_{s+1}) = \zeta(x_s) - \eta_s\widehat{g}_s$ for $\widehat{g}_s \in \partial\ell(x_s, Z_s)$
    $x_{s+1} = \Pi_\Omega^\Phi(y_{s+1})$
  end for
Ensure: $\bar{x}_t = \frac{1}{t}\sum_{s=1}^{t} x_s$

定理. 假设Φ是$\Omega\cap\mathcal{D}$上关于$\|\cdot\|$的$\alpha-$强凸函数，以及

$$R^2 = \sup_{x\in\Omega\cap\mathcal{D}} \Phi(x) - \min_{x\in\Omega\cap\mathcal{D}} \Phi(x)$$

取$x_1 = \operatorname{argmin}_{x\in\Omega\cap\mathcal{D}} \Phi(x)$(假设它存在). 并假设方差有界假设(VB)成立. 则当$\eta = \frac{R}{L}\sqrt{\frac{2\alpha}{t}}$时，由随机镜像下降算法得到的$\bar{x}_t$满足

$$\mathbb{E}[f(\bar{x}_t)] - f(x_*) \leqslant RL\sqrt{\frac{2}{\alpha t}}.$$

证明. 本质上是在重复镜像下降算法的证明. 取$x^\sharp \in \Omega \cap \mathcal{D}$，并且记$g_s = \mathbb{E}[\widehat{g}|x_s]$，那么$g_s \in \partial f(x_s)$，从而有

$$\begin{aligned}
f(x_s) - f(x^\sharp) &\leqslant g_s^\top (x_s - x^\sharp) \\
&= \mathbb{E}[\widehat{g}_s^\top (x_s - x_*)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(\zeta(x_s) - \zeta(y_{s+1}))^\top (x_s - x^\sharp)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}))^\top (x_s - x^\sharp)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, y_{s+1})|x_s\right] \\
&\leqslant \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})|x_s\right] \\
&\leqslant \frac{\eta}{2\alpha}\mathbb{E}[\|\widehat{g}_s\|_*^2|x_s] + \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})|x_s\right]
\end{aligned}$$

其中最后一个不等式是由以下推导得到的：

$$\begin{aligned}
D_\Phi(x_s, y_{s+1}) &= \Phi(x_s) - \Phi(y_{s+1}) - \nabla\Phi(y_{s+1})^\top (x_s - y_{s+1}) \\
&\leqslant \left[\nabla\Phi(x_s) - \nabla\Phi(y_{s+1})\right]^\top (x_s - y_{s+1}) - \tfrac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\leqslant \eta\|\widehat{g}_s\|_*\|x_s - y_{s+1}\| - \tfrac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\leqslant \tfrac{\eta^2\|\widehat{g}_s\|_*^2}{2\alpha}.
\end{aligned}$$

将s从1到t得到的不等式求和，并关于$Z_1, \cdots, Z_{t-1}$取期望，得到

$$\mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t}\left[f(x_s) - f(x^\sharp)\right]\right] \leqslant \frac{\eta L^2}{2\alpha} + \frac{D_\Phi(x^\sharp, x_1)}{t\eta}. \tag{11.2}$$

得到如前一讲的结论. □

11.5 在线学习与乘性权重更新

学习设置的一个有趣变形是所谓的**在线学习**(online learning). 这里没有整个训练集，而且需要依次作一系列决策.

听取专家的建议. **想**象可以利用n个专家的预测. 从关于专家的初始分布开始，已知权重 $w_1 \in \Delta_n = \{w \in \mathbb{R}^n \colon \sum_i w_i = 1, w_i \geqslant 0\}$.

在每一步 $s = 1, \ldots, T$:

- 根据 w_s随机地选取专家.
- 自然地分配损失函数 $f_s \in [-1, 1]^n$, 为专家i 指定损失 $f_s[i]$，即专家i在时刻s 的预测引起的损失.
- 遭受的期望损失$\mathbb{E}_{i\sim w_s} f_s[i] = \langle w_s, f_s\rangle$.
- 将分布从w_s更新为w_{s+1}.

一天结束后，度量相对于事后关于专家最好的固定分布而言，执行地有多好. 将此称作**遗憾**(regret):

$$R_t = \sum_{s=1}^{t} \langle w_s, f_s \rangle - \min_{w \in \Delta_n} \sum_{s=1}^{t} \langle w, f_s \rangle$$

这是一个相对基准. 小的遗憾并不意味着损失必定很小. 仅表明，即使有后见之明，并在所有步使用该相同策略，也不可能做的更好.

最重要的在线算法也许是**乘性权重更新**(multiplicative weights update). 从均匀分布 w_1开始， 继而对于$s > 1$ 根据如下简单规则更新，

$$v_s[i] = w_{s-1}[i] e^{-\eta f_s[i]} \qquad \text{(指数权重更新)}$$

$$w_s = v_s / (\textstyle\sum_i v_s[i]) \qquad \text{(归一化)}$$

问题是**如何确定乘性权重更所得遗憾的上界?** 能进行直接分析，但是这里选择把乘性权重与梯度下降法关联起来，从而使用已经知道的收敛性结论.

可将乘性权重解释成用镜像下降法求解上述在线凸优化，即 乘性权重更新是镜像下降法的实例. 具体地，取镜像映射$\Phi(w) = \sum_{i=1}^{n} w_i \ln w_i$是负熵函数，选$\|\cdot\|_1$. 有

$$\nabla \Phi(w) = 1 + \ln w,$$

其中对数是逐分量的. 镜像下降法的更新规则是

$$\nabla \Phi(v_{s+1}) = \nabla \Phi(w_s) - \eta_s f_s \,,$$

这蕴含着

$$v_{s+1} = w_s e^{-\eta_s f_s},$$

由此复原了乘性权重更新.

现在计算投影步. 与Φ对应的Bregman散度是

$$\begin{aligned} D_\Phi(x, y) &= \Phi(x) - \Phi(y) - \nabla \Phi(y)^T (x - y) \\ &= \sum_i x_i \ln(x_i / y_i) - \sum_i x_i + \sum_i y_i \,, \end{aligned}$$

发现这就是单纯形上的相对熵或者Kullback-Leibler散度. 因此选取概率单纯形

$$\Omega = \{w \in \mathbb{R}^n \mid \sum_i w_i = 1, w_i \geqslant 0\}$$

作为定义域Ω. 投影

$$\Pi_\Omega^\Phi(y) = \operatorname*{argmin}_{x \in \Omega} D_\Phi(x, y)$$

恰好对应于乘性权重算法中更新规则的归一化步.

收敛速率. 为了从上面的定理得到具体的收敛速率，仍然需要确定设置中涉及到的强凸性常数 α. 这里，选取了ℓ_1- 范数. 由Pinsker不等式得到Φ 关于ℓ_1- 范数是1-强凸的. 此外，就ℓ_∞-范数而言，由于损失函数的值域包含在 $[-1, 1]$，从而所有梯度的ℓ_∞范数以 1 为界. 最终，初始均匀分布与任一其它分布的相对熵至多是 $\ln n$. 将这些事实放在一起，并平衡步长 η的值，能得到平均遗憾界

$$O\left(\sqrt{\tfrac{\ln n}{t}}\right).$$

特别地，这表明随着时间趋于无穷，乘性更新规则的平均遗憾会趋于 0 .

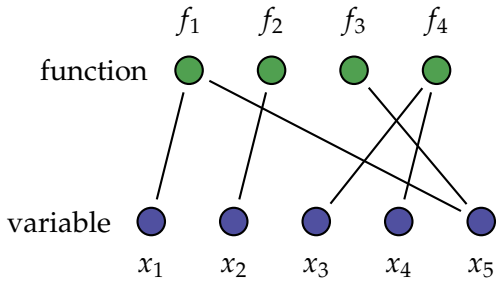


图 12.1: 由函数$f : \mathbb{R}^n \to \mathbb{R}^m$的群稀疏结构诱导的函数分量$f_i$和变量$x_j$间的二分图例子. f_i和x_j之间的边传达了第i个分量函数与输入的第j个坐标有关.

12 坐标下降法

有许多函数，易于计算沿着标准基向量$e_i, i \in [n]$的方向导数. 比如

$$f(x) = \|x\|_2^2 \quad \text{或者} \quad f(x) = \|x\|_1$$

对于一般正则化子也是如此，其一般形如

$$R(x) = \sum_{i=1}^{n} R_i(x_i) \, .$$

更一般的，许多目标和正则化子呈现出“群稀疏性”；即

$$R(x) = \sum_{j=1}^{m} R_j(x_{S_j})$$

其中对$j \in [m]$，S_j是$[n]$的子集，针对$f(x)$也是类似的.

12.1 随机坐标下降法

设函数f在$\mathbb{R}^n$上是可微的凸$L-$Lipschitz函数. 记f在方向e_i 上的偏导数为$\nabla_i f$. 梯度下降算法的一个缺点是：在每一步中需要计算梯度的所有分量$\nabla_i f$，以便更新每一个分量. 随机坐标下降算法的思路是：在每一步中均匀地选取一个方向e_j，并选择e_j为该步的下降方向. 准确地说，如果I 是$[n]$ 上的均匀分布. 那么

$$\mathbb{E}[n \nabla_I f(x) e_I | x] = \nabla f(x). \tag{12.1}$$

因此，只有一个非零坐标的向量$n \nabla_I f(x) e_I$ 是梯度$\nabla f(x)$的无偏估计. 可利用该估计执行随机梯度下降算法.

因为(12.1)，所以SCD是一阶随机oracle 方法. 进一步，计算易得

$$\mathbb{E}[\|n \nabla_I f(x) e_I\|_2^2] = n \|\nabla f(x)\|_2^2,$$

Algorithm 4 Stochastic Coordinate Descent algorithm

Require: $x_1 \in \Omega$, positive sequence $\{\eta_s\}_{s\geqslant 1}$, independent random variables $I_1, \cdots, I_t$ uniform over $[n]$.
  for $s = 1$ to t **do**
    $y_{s+1} = x_s - \eta_s d \nabla_{I_s} f(x_s) e_{I_s}$
    $x_{s+1} = \pi_\Omega(y_{s+1})$
  end for
Ensure: $\bar{x}_t = \frac{1}{t}\sum_{s=1}^{t} x_s$

再由f是$L-$Lipschtize的，得方差有界条件(VB)中的上界参数是nL^2. 这样，在随机梯度下降法的复杂性结论(定理 11.6)，置$\eta = \frac{R}{L\sqrt{nt}}$，直接得到

$$\mathbb{E}[f(\bar{x}_t)] - f(x_*) \leqslant RL\sqrt{\frac{n}{t}}.$$

在这里取了一个折衷：更新过程易于执行，但需要执行更多的步骤以达到与梯度下降算法同样的精度.

12.2 重要性采样

在上面，决定使用均匀分布来采样每个坐标. 但是假设有更细粒度的信息. 特别地，如果知道能够上控$\sup_{x\in\Omega} \|\nabla_i f(x)\|_2 \leqslant L_i$，使用哪种采样? 一种方法是以某种方式将$L_i$纳入采样考量. 这激发了$\nabla f(x)$的“重要性采样”估计量, 形如

$$\widehat{g}_s = \frac{1}{p_{i_s}} \cdot \nabla_{i_s} f(x_s) e_{i_s},$$

其中$\mathbb{P}(i_s = i) = p_i, i = 1, \cdots, n$. 请注意那么 $\mathbb{E}[\widehat{g}_s | x_s] = \nabla f(x_s)$, 但是

$$\mathbb{E}[\|\widehat{g}_s\|_2^2] = \sum_{i=1}^{n} (\nabla_i f(x_s))^2 / p_i \leqslant \sum_{i=1}^{n} L_i^2 / p_i.$$

这种情况下，在随机梯度下降法的复杂性结论(定理 11.6)得到收敛速率

$$\mathbb{E}\left[f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right)\right] - \min_{x\in\Omega} f(x) \leqslant \frac{R}{\sqrt{t}} \cdot \sqrt{\sum_{i=1}^{n} \frac{L_i^2}{p_i}}.$$

在很多情况下，如果L_i的值是异构的，由此能够优化p_i的值.

12.3 针对光滑坐标下降法的重要性采样

本节考虑使用梯度**有偏**(biased)估计量的坐标下降法. 假设对于$x \in \mathbb{R}^n$ 和$\alpha \in \mathbb{R}$, 存在$\beta_i > 0$使得不等式

$$|\nabla_i f(x) - \nabla_i f(x + \alpha e_i)| \leqslant \beta_i |\alpha|$$

成立，其中β_i可能是异构的. 请注意，如果f是二次连续可微的，那么上面的定义等价于$\nabla^2_{ii} f(x) \leqslant \beta_i$, 或者 $\mathrm{diag}(\nabla^2 f(x)) \leqslant \mathrm{diag}(\boldsymbol{\beta} I)$，这里用$\mathrm{diag}(\cdot)$ 表示一个矩阵的对角线元素所得向量. 已知参数$\gamma > 0$，定义

$$p_i^\gamma = \frac{\beta_i^\gamma}{\sum_{j=1}^n \beta_j^\gamma}.$$

对于分布p^γ，考虑使用称作RCD(γ) 规则的梯度下降法

$$x_{t+1} = x_t - \frac{1}{\beta_{i_t}} \cdot \nabla_{i_t} f(x_t) \cdot e_{i_t}, \text{其中} i_t \sim p^\gamma$$

注意到随着$\gamma \to \infty$, 较大的β_i值对应的坐标被选择的更频繁些. 需要注意的是, 因为

$$\mathbb{E}\left[\frac{1}{\beta_{i_t}} \nabla_{i_t} f(x_t) e_{i_t}\right] = \frac{1}{\sum_{j=1}^n \beta_j^\gamma} \cdot \sum_{i=1}^n \beta_i^{\gamma-1} \nabla_i f(x_t) e_i = \frac{1}{\sum_{j=1}^n \beta_j^\gamma} \cdot \nabla f(x_t) \circ (\beta_i^{\gamma-1})_{i\in[n]}$$

所以这一般不等价于SGD. 当$\gamma = 1$时，这仅是伸缩版的$\nabla f(x_t)$. 仍然可以证明如下定理：

定理 12.1 (定理 2.11的随机版)**.** 已知$\gamma > 0, \beta \in \mathbb{R}^n_{++}$. 定义加权范数

$$\|x\|^2_{[\gamma]} := \sum_{i=1}^n x_i^2 \beta_i^\gamma \quad \text{和} \|x\|^{*2}_{[\gamma]} := \sum_{i=1}^n x_i^2 \beta_i^{-\gamma}.$$

注意这一对范数互为对偶. 然后，规则RCD(γ)产生的迭代满足

$$\mathbb{E}[f(x_t) - \min_{x\in\mathbb{R}^n} f(x)] \leqslant \frac{2R^2_{1-\gamma} \cdot \sum_{i=1}^n \beta_i^\gamma}{t-1},$$

其中 $R^2_{1-\gamma} = \sup_{x\in\mathbb{R}^n : f(x)\leqslant f(x_1)} \|x - x_*\|^2_{[1-\gamma]}$.

Proof. 由引理 2.9后面的说明，知道针对一般的β_φ-光滑凸函数 φ, 有

$$\varphi\left(u - \frac{1}{\beta_\varphi} \nabla\varphi(u)\right) - \varphi(u) \leqslant -\frac{1}{2\beta_\varphi} \|\nabla\varphi\|^2.$$

考虑函数 $\varphi_i(u; x) = f(x + ue_i)$, 看到$\varphi_i'(u; x) = \nabla_i f(x + ue_i)$, 并且$\varphi_i$ 是 β_i光滑的. 因此有

$$f\left(x - \frac{1}{\beta_i} \nabla_i f(x) e_i\right) - f(x) = \varphi_i\left(0 - \frac{1}{\beta_i} \varphi_i'(0; x); x\right) - \varphi(0; x) \leqslant -\frac{\varphi_i'(0;x)^2}{2\beta_i} = -\frac{\nabla_i f(x)^2}{2\beta_i}.$$

从而，如果 $i_t \sim p^\gamma$, 有

$$\begin{aligned}
\mathbb{E}\left[f\left(x - \frac{1}{\beta_{i_t}} \nabla_{i_t} f(x) e_{i_t}\right) - f(x)\right] &\leqslant \sum_{i=1}^n p_i^\gamma \cdot \left(-\frac{\nabla_i f(x)^2}{2\beta_i}\right) \\
&= -\frac{1}{2\sum_{i=1}^n \beta_i^\gamma} \sum_{i=1}^n \beta_i^{\gamma-1} \nabla_i f(x)^2 \\
&= -\frac{\|\nabla f(x)\|^{*2}_{[1-\gamma]}}{2\sum_{i=1}^n \beta_i^\gamma}
\end{aligned}$$

因此，如果定义 $\delta_t = \mathbb{E}[f(x_t) - f(x_*)]$, 有

$$\delta_{t+1} - \delta_t \leqslant -\frac{\|\nabla f(x_t)\|^{*2}_{[1-\gamma]}}{2\sum_{i=1}^n \beta_i^\gamma}. \tag{12.2}$$

此外，由上面，以概率1也有$f(x_{t+1}) \leqslant f(x_t)$. 现在继续用光滑梯度下降法的常规证明. 请注意

$$\begin{aligned} \delta_t &\leqslant \nabla f(x_t)^\top (x_t - x_*) \\ &\leqslant \|\nabla f(x_t)\|^*_{[1-\gamma]} \|x_t - x_*\|_{[1-\gamma]} \\ &\leqslant R_{1-\gamma} \|\nabla f(x_t)\|^*_{[1-\gamma]} . \end{aligned}$$

把这些事实放在一起蕴含着

$$\delta_{t+1} - \delta_t \leqslant -\frac{\delta_t^2}{2R^2_{1-\gamma}\sum_{i=1}^n \beta_i^\gamma}$$

回忆这就是在非随机-情况下，即定理 2.11用来证明收敛性的递归关系. ■

定理 12.2. 如果额外地，f关于范数$\|\cdot\|_{[1-\gamma]}$是α- 强凸的，那么得到

$$\mathbb{E}[f(x_{t+1}) - \min_{x\in\mathbb{R}^n} f(x)] \leqslant \left(1 - \frac{\alpha}{\sum_{i=1}^n \beta_i^\gamma}\right)^t (f(x_1) - f(x_*)) . \tag{12.3}$$

Proof. 需要如下引理：

引理 12.3. 设f关于范数$\|\cdot\|$是α-强凸的. 那么$f(x) - f(x_*) \leqslant \frac{1}{2\alpha}\|\nabla f(x)\|^2_* .$

Proof. 从$\alpha-$强凸的定义开始，对任意的x, y，得

$$\begin{aligned} f(x) - f(y) &\leqslant \nabla f(x)^\top (x - y) - \tfrac{\alpha}{2}\|x - y\|_2^2 \\ &\leqslant \|\nabla f(x)\|_* \|x - y\| - \tfrac{\alpha}{2}\|x - y\|_2^2 \\ &\leqslant \max_t \|\nabla f(x)\|_* t - \tfrac{\alpha}{2} t^2 \\ &= \tfrac{1}{2\alpha}\|\nabla f(x)\|^2_* . \end{aligned}$$

■

引理12.3表明

$$\|\nabla f(x_t)\|^{*2}_{[1-\gamma]} \geqslant 2\alpha\delta_t .$$

另一方面，将这个不等式与定理12.1证明了的不等式 (12.2)综合起来， 得到

$$\begin{aligned} \delta_{t+1} - \delta_t &\leqslant -\frac{\alpha\delta_t}{\sum_{i=1}^n \beta_i^\gamma} \\ \delta_{t+1} &\leqslant \delta_t \left(1 - \frac{\alpha}{\sum_{i=1}^n \beta_i^\gamma}\right) . \end{aligned}$$

递归地应用上面的不等式，并且回想起 $\delta_t = \mathbb{E}[f(x_t) - f(x_*)]$即给出结论.

■

12.4 随机坐标下降法与随机梯度下降法

出人意料的是，尽管RCD(γ)是随机的，但它是下降方法. 这对于一般的SGD是不成的. 但是，何时RCD(γ)实际上会表现地更好？如果 $\gamma = 1$, 节省量(the savings)与比值$\sum_{i=1} \beta_i/\beta \cdot (T_{\text{coord}}/T_{\text{grad}})$成正比. 当 f二次可微时，这个比值为

$$\frac{\operatorname{tr}(\max_x \nabla^2 f(x))}{\|\max_x \nabla^2 f(x)\|_{\text{op}}}(T_{\text{coord}}/T_{\text{grad}})$$

12.5 坐标下降的其它推广：

1. 非随机，循环SGD
2. 有放回采样
3. 强凸 + 光滑！？
4. 强凸 (广义SGD)
5. 加速？参见 [TVW+17]

13 学习、稳定性、正则化

在本讲中重新审视机器学习，特别是经验风险最小化. 定义$\mathcal{X} \times \mathcal{Y}$ 上分布为D的随机变量(数据)，其中 $\mathcal{X} \subseteq \mathbb{R}^d$，$\mathcal{Y}$是某离散集合，表示类别标签. 比如，在二分类任务中，Y 有两个标签，这时$\mathcal{Y} = \{-1, 1\}$.

- 由参数集合 $w \in \Omega \subseteq \mathbb{R}^n$来指定“模型”.
- "损失函数"记作 ℓ: $\Omega \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, 请注意$\ell(w, Z)$给出的是模型$w$关于实例$Z = (X, Y)$的损失.
- 模型的**风险**(risk)定义为$R(w) = \mathbb{E}_{Z\sim D}[\ell(w, Z)]$.

目的是找到极小化 $R(w)$的模型.

达成该目标的一种方法是对总体目标直接使用随机梯度法：

$$w_{t+1} = w_t - \eta \nabla \ell(w_t, Z_t), \quad Z_t \sim D$$

当已知数据集有限时，在数据上做多轮迭代是更有效的. 在这种情况下，随机梯度法不再直接极小化风险.

13.1 经验风险和推广误差

考虑有限样本. 假设

$$S = ((X_1, Y_1), \cdots, X_m, Y_m) \in (X \times Y)^m,$$

其中$Z_i = (Z_i, Y_i)$代表第i个带标签的样例. **经验风险**(empirical risk)定义为

$$R_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, Z_i)\,.$$

经验风险极小化(empirical risk minimization，ERM)通常是极小化未知的总体风险的某种替代. 但是这种替代有多好？理想地，想要通过经验风险极小化找到的点 w 满足 $R_S(w) \approx R(w)$. 然而，情况并不是这样的，因为风险 $R(w)$捕捉了未看到的样本上的损失，而经验风险$R_S(w)$捕捉了看到的样本上的损失. 通常期望对看到的样本比对未看到的样本做的更好. 将看到的样本与未看到的样本上的这种性能间隙称作**推广误差**(generalization error).

定义 13.1 (推广误差)**.** 模型w的**推广误差**(generalization error)定义为

$$\epsilon_{\text{gen}}(w) = R(w) - R_s(w)\,.$$

请注意如下尽管是同义反复，然而却很重要的等式:

$$R(w) = R_S(w) + \epsilon_{\text{gen}}(w). \tag{13.1}$$

特别地，这表明：如果通过优化设法使经验风险$R_S(w)$很小，那么剩下唯一需要操心的就是推广误差.

因此，如何能够上控推广误差? 下面将建立基本关系：推广误差等价于称作**算法稳定性**(algorithmic stability)的稳健性质. 直观上，算法稳定性度量了算法对单个训练样本改变的敏感程度.

13.2 算法稳定性

为了引入稳定性的思想，选取两个独立样本 $S = (Z_1, \cdots, Z_m)$和 $S' = (Z'_1, \cdots, Z'_m)$，每个都是从$D$中独立同分布抽取的. 这里，将第二个样本$S'$ 称作**幽灵样本**(ghost sample)，其主要目的是为分析服务.

用单个点将两个样本关联起来，引入混合样本 $S^{(i)}$:

$$S^{(i)} = (Z_1, \cdots, Z_{i-1}, Z'_i, Z_{i+1}, \cdots, Z_m)$$

注意，这里第i个样本取自 S', 而所有其它的取自 S. 用这个得力记号，可以引入平均稳定性的概念.

定义 13.2 (平均稳定性)**.** 算法$A : (\mathcal{X} \times \mathcal{Y})^m \to \Omega$的**平均稳定性**(average stability):

$$\Delta(A) = \mathop{\mathbb{E}}_{S,S'} \left[\frac{1}{m} \sum_{i=1}^{m} \Big(\ell(A(S), Z'_i) - \ell(A(S^{(i)}), Z'_i) \Big) \right].$$

这里将算法量化成映射A，其中$A(S)$是映射的像，表示由算法A得到的模型参数. 可将该定义解释为算法在一个看不到的样本与一个看到的样本上性能的比对. 这是为什么平均稳定性实际上等于推广误差的直观解释. 称算法A是稳定的，如果它的输入S发生小的变化将导致由它输出的假设出现小的改变. 该定义中，输入量小的改变体现为替换其中一个样本. 输出量的改变是依次改变其中一个样本带来输出量差异的平均值关于(S, S')的期望.

定理 13.3. 对于算法A，有 $\mathbb{E}[\epsilon_{\text{gen}}(A)] = \Delta(A)$.

Proof. 请注意，由定义有

$$\begin{aligned}
\mathbb{E}[\epsilon_{\text{gen}}(A)] &= \mathbb{E}\left[R(A(S)) - R_S(A(S))\right], \\
\mathbb{E}[R(A(S))] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(A(S), Z_i')\right], \\
\mathbb{E}\left[R_S(A(S))\right]] &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(A(S), Z_i)\right].
\end{aligned}$$

同时，由于Z_i和Z_i'是同分布，并且与其它样例独立，所以有

$$\mathbb{E}[\ell(A(S), Z_i)] = \mathbb{E}[\ell(A(S^{(i)}), Z_i')].$$

对上面经验风险中的每一项应用该等式，并且与$\Delta(A)$的定义进行比对，得到

$$\mathbb{E}[R(A(S)) - R_S(A(S))] = \Delta(A).$$

■

13.2.1 一致稳定性

尽管平均稳定性给出了推广误差的精确刻画，但它需要关于S和S'求期望，从而很难. 现在给出一致稳定性的概念，它用上确界代替平均，是一个更强但是很有用的概念 [BE02].

定义 13.4 (一致稳定性)**.** 算法 A的**一致稳定性**定义为

$$\Delta_{\text{sup}}(A) = \sup_{S,S' \in (\mathcal{X}\times\mathcal{Y})^m} \sup_{i\in[m]} |\ell(A(S), Z_i') - \ell(A(S^{(i)}, Z_i')|.$$

由于一致稳定性是平均稳定性的上界，从而一致稳定性也给出了(期望意义上)推广误差的上界.

推论 13.5. $\mathbb{E}[\epsilon_{\text{gen}}(A)] \leqslant \Delta_{\text{sup}}(A)$.

结果发现该推论惊人地有用，因为许多算法是一致稳定的. 比如，像将要证明的那样，强凸损失函数对于稳定性是充分的，因此强凸损失函数对于推广性也是充分的.

13.3 经验风险极小化的稳定性

下一个定理归功于 [SSSSS10]，它表明强凸损失函数的经验风险极小点是一致稳定的. 有趣的是没有显式提到函数类的复杂性. 下面给出描述和证明.

定理 13.6. 假设$\ell(\cdot, z)$在定义域 Ω上是α-强凸和L-Lipschitz 的. 设

$$\widehat{w}_S = \arg\min_{w\in\Omega} R_S(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, Z_i)$$

表示经验风险最小点(empirical risk minimizer, ERM). 那么 ERM 满足

$$\Delta_{\sup}(\mathrm{ERM}) \leqslant \frac{4L^2}{\alpha m}.$$

Proof. 用$\widehat{w}_S$表示与样本 S 关联的经验风险最小点. 固定大小为m的任意样本 S, S'和指标 $i \in [m]$. 需要证明

$$|(\ell(\widehat{w}_{S^{(i)}}, Z_i') - \ell(\widehat{w}_S, Z_i'))| \leqslant \frac{4L^2}{\alpha m}.$$

一方面，由强凸性知道

$$R_S(\widehat{w}_{S^{(i)}}) - R_S(\widehat{w}_S) \geqslant \frac{\alpha}{2}\|\widehat{w}_S - \widehat{w}_{S^{(i)}}\|^2. \tag{13.2}$$

另一方面，

$$\begin{aligned}
&R_S(\widehat{w}_{S^{(i)}}) - R_S(\widehat{w}_S)\\
=&\tfrac{1}{m}[\ell(\widehat{w}_{S^{(i)}}, Z_i) - \ell(\widehat{w}_S, Z_i)] + \tfrac{1}{m}\sum_{j\neq i}[\ell(\widehat{w}_{S^{(i)}}, Z_j) - \ell(\widehat{w}_S, Z_j)]\\
=&\tfrac{1}{m}[\ell(\widehat{w}_{S^{(i)}}, Z_i) - \ell(\widehat{w}_S, Z_i)] + \tfrac{1}{m}[\ell(\widehat{w}_S, Z_i') - \ell(\widehat{w}_{S^{(i)}}, Z_i')] + \left[R_{S^{(i)}}(\widehat{w}_{S^{(i)}}) - R_{S^{(i)}}(\widehat{w}_S)\right]\\
\leqslant&\tfrac{1}{m}|\ell(\widehat{w}_{S^{(i)}}, Z_i) - \ell(\widehat{w}_S, Z_i)| + \tfrac{1}{m}|(\ell(\widehat{w}_S, Z_i') - \ell(\widehat{w}_{S^{(i)}}, Z_i'))|\\
\leqslant&\tfrac{2L}{m}\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\|.
\end{aligned} \tag{13.3}$$

这里，利用了 $R_{S^{(i)}}(\widehat{w}_{S^{(i)}}) - R_{S^{(i)}}(\widehat{w}_S)) \leqslant 0$ 和损失函数ℓ是$L-$Lipschitz的事实.

将 (13.2)和 (13.3)结合起来，得到$\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\| \leqslant \frac{4L}{\alpha m}$. 再次由损失函数的Lipschitz性，有

$$|(\ell(\widehat{w}_{S^{(i)}}, Z_i') - \ell(\widehat{w}_S, Z_i'))| \leqslant L\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\| \leqslant \tfrac{4L^2}{\alpha m}.$$

因此， $\Delta_{\sup}(\mathrm{ERM}) \leqslant \frac{4L^2}{\alpha m}$.

■

13.4 正则化

并不是所有的ERM问题都是强凸的. 然而如果问题是凸的，可以考虑正则化的目标

$$R_S(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, Z_i) + \frac{\alpha}{2}\|w\|^2$$

正则化损失$R_S(w)$是$\alpha-$强凸的. 根据所在领域，称最后一项是ℓ_2- 正则化，权重衰减或者Tikhonov正则化. 因此，有如下链式蕴含关系:

$$\text{regularization} \Rightarrow \text{strong convexity} \Rightarrow \text{uniform stability} \Rightarrow \text{generalization}$$

也能够证明：求解正则化目标也求解了未正则化的目标. 假设 $\Omega \subseteq \mathcal{B}_2(R)$, 通过置$\alpha \approx \frac{L^2}{R^2 m}$ 能够证明正则化的风险最小点也极小化未正则化的风险，其误差是$\mathcal{O}(\frac{LR}{\sqrt{m}})$. 此外，有前一个定理，推广误差也将是$\mathcal{O}(\frac{LR}{\sqrt{m}})$. 细节请参见 [SSSSS10]中的定理3.

13.5 隐正则化

在隐式正则化中，算法本身在正则化目标，而不是显式增加正则化项. 如下定理描述了随机梯度法(SGM)的正则化效果.

定理 13.7. 假设 $\ell(\cdot, Z)$是凸的，β-光滑和L-Lipschitz 的. 如果运行t步 *SGM* ，那么算法具有一致稳定性

$$\Delta_{\sup}(\mathrm{SGM}_t) \leqslant \frac{2L^2}{m} \sum_{s=1}^{t} \eta_s$$

注意对于 $\eta_s \approx \frac{1}{m}$那么 $\Delta_{\sup}(\mathrm{SGM}_t) = \mathcal{O}(\frac{\log(t)}{m})$, 并且对于$\eta_s \approx \frac{1}{\sqrt{m}}$ 和$t = \mathcal{O}(m)$那么$\Delta_{\sup}(\mathrm{SGM}_t) = \mathcal{O}(\frac{1}{\sqrt{m}})$. 证明参见[HRS15].

Part IV
对偶方法

14 对偶定理

这些笔记是基于Benjamin Recht和Ashia Wilson之前的讲义整理的.

14.1 等式约束优化的最优性条件

设$\Omega \subseteq \mathbb{R}^n$是闭凸集，函数$f$在包含$\Omega$的开集上是光滑和凸的. 那么推论 1.11表明，

$$x_* \in \arg\min_{x\in\Omega} f(x) \tag{14.1}$$

当且仅当

$$\langle \nabla f(x_*), x - x_* \rangle \geqslant 0 \quad \forall x \in \Omega\,. \tag{14.2}$$

下来专门研究(14.1)中的Ω是仿射集的特殊情况. 设A 是$m \times n$矩阵，秩为m，并且存在$b \in \mathbb{R}^m$使得$\Omega = \{x \,:\, Ax = b\}$. 请注意总是假设 $\operatorname{rank} A = m$，否则，会有冗余约束. 也能够将$\Omega$ 参数化成$\Omega = \{x_0 + u : Au = 0\}$，这对任何$x_0 \in \Omega$ 都是成立的. 那么应用 (14.2), 有

$$\langle \nabla f(x_*), x - x_* \rangle \geqslant 0 \ \ \forall x \in \Omega \ \ \text{当且仅当} \ \ \langle \nabla f(x_*), u \rangle \geqslant 0 \ \ \forall u \in \operatorname{null} A\,.$$

但是由于$\operatorname{null} A$是子空间，这成立当且仅当对所有$u \in \operatorname{null} A$有$\langle \nabla f(x_*), u \rangle = 0$. 特别地，这意味着$\nabla f(x_*)$必须属于$(\operatorname{null} A)^\perp$. 记$A^\top$的像空间

$$\operatorname{Im} A^\top := \{A^\top \lambda : \lambda \in \mathbb{R}^m\}.$$

由于有$\mathbb{R}^n = \operatorname{null} A \oplus \operatorname{Im} A^T$, 这意味着存在$\lambda \in \mathbb{R}^m$使得$\nabla f(x_*) = A^T\lambda$.

总而言之，这意味着x_*是f在Ω上的极小点当且仅当存在$\exists \lambda^* \in \mathbb{R}^m$使得

$$\begin{aligned} \nabla f(x_*) + A^T\lambda^* &= 0 \\ Ax_* &= b. \end{aligned}$$

这些最优性条件就是著名的**Karush-Kuhn-Tucker条件**或者**KKT条件**应用于线性等式约束优化问题

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & Ax = b \end{aligned} \tag{14.3}$$

所得到的. 作为例子，考虑等式约束二次优化问题

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}x^TQx - c^Tx \\ \text{subject to} \quad & Ax = b, \end{aligned}$$

其中对称的$Q \in \mathbb{R}^{n\times n}$和$c \in \mathbb{R}^n$是已知的. 它的KKT条件用矩阵形式表示成

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} c \\ b \end{bmatrix}.$$

14.2 非线性约束

下面考虑包含非线性约束的情况. 假设想要在闭凸集Ω和仿射集$\mathcal{A} = \{x : Ax = b\}$的交集上极小化连续可微函数 f：

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \Omega, Ax = b \end{array} \tag{14.4}$$

其中 A再一次是满秩的$m \times n$矩阵. 为了刻画最优性，引入凸集在一点的切锥和法锥的概念.

设Ω是闭凸集. 定义Ω在x处的**切锥**(tangent cone)[7]为

$$\mathcal{T}_\Omega(x) = \text{cone}\{z - x : z \in \Omega\}$$

切锥是由所有满足如下条件的方向组成的集合：满足从x出发，沿着该方向移动还能保持在Ω内. 定义Ω在x处的**法锥**(normal cone)为集合

$$\mathcal{N}_\Omega(x) = \mathcal{T}_\Omega(x)^\circ := \{u \ : \ \langle u, v\rangle \leqslant 0, \ \ \forall v \in \mathcal{T}_\Omega(x)\}.$$

切锥和法锥的几何直观见图 14.2.

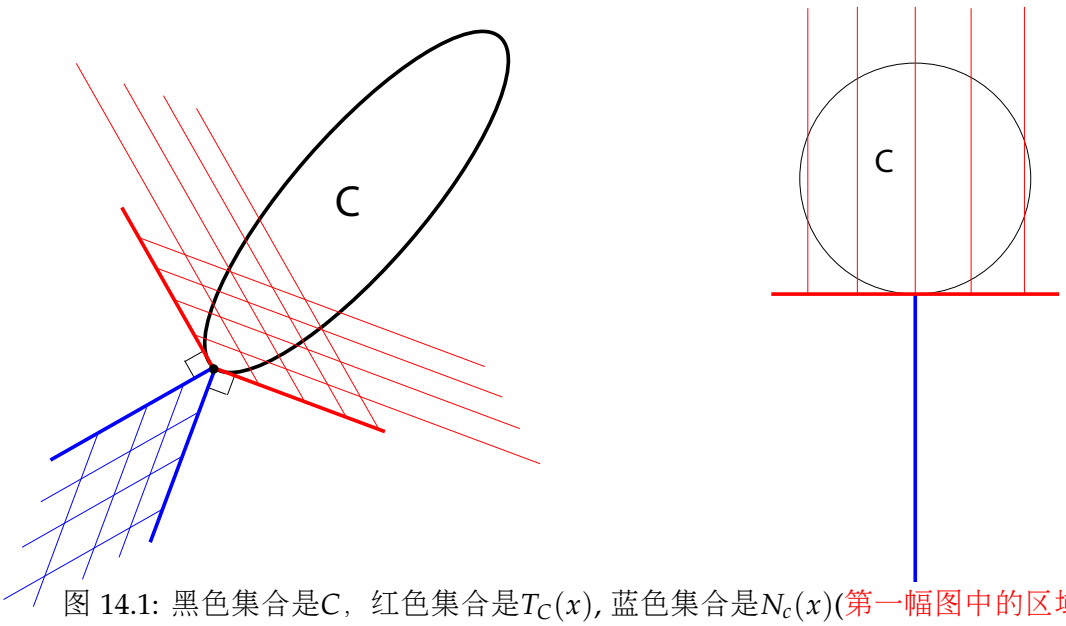


图 14.1: 黑色集合是C，红色集合是$T_C(x)$, 蓝色集合是$N_c(x)$(第一幅图中的区域Ω在所考虑点，它的边界应该是由两端光滑曲线拼接而成的才对. 这时候，那两组平行线分别与那一点两个光滑曲线的切线平行).

考虑这里定义的仿射集$\mathcal{A}$，由切锥和法锥的定义，有

$$\mathcal{N}_\mathcal{A}(x) = \text{null}\,\mathcal{A}, \quad \mathcal{T}_\mathcal{A}(x) = \text{Im}\,A^\top.$$

[7]切锥可看成光滑曲线在一点的切线的推广. 考虑由光滑曲线围成的区域Ω，考虑过Ω的边界某一点的切线，即这里定义的切锥. 集合Ω可以是无界区域，或者其边界也可以是由分段光滑曲线拼接而成的. 这样，在拼接点，切线就扩展成切锥了.

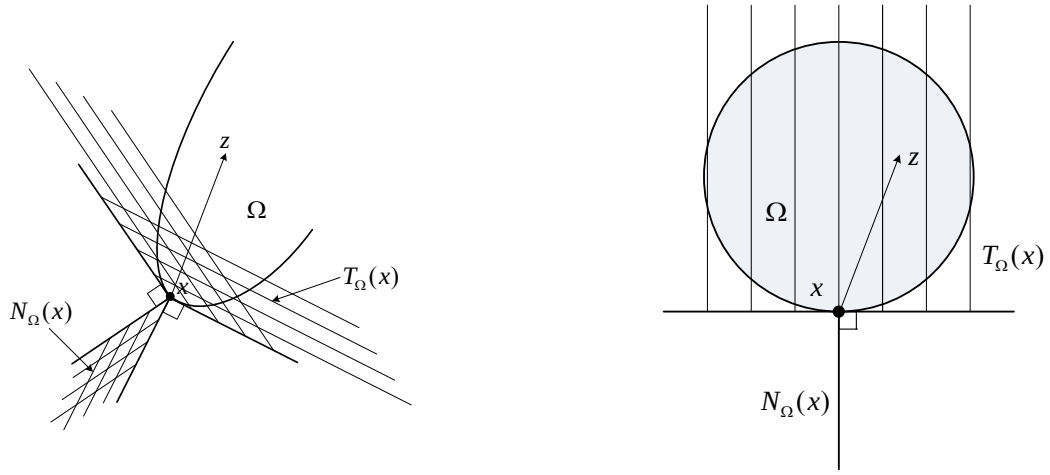


图 14.2: 凸集的切锥与法锥

.

当$n = 2$与3时，$\mathcal{A}$的切锥分别还原成直线的法向量和平面的法向量. 请注意，当等式约束 $Ax = b$不存在时，约束问题(14.1) 的最优性条件(14.2)完全等价于断言

$$-\nabla f(x_*) \in \mathcal{N}_\Omega(x_*)\,. \tag{14.5}$$

本节将推广 (14.2) 以证明最优化问题 (14.4)的最优性条件.

命题 14.1 (几何最优性条件)**.** 向量x_* 是 (14.4)的最优解当且仅当存在$\lambda^* \in \mathbb{R}^m$ 使得

$$\begin{aligned} -\nabla f(x_*) + A^\top \lambda^* &\in \mathcal{N}_\Omega(x_*), \\ x_* &\in \Omega \cap \mathcal{A}. \end{aligned}$$

这里分析的核心将依赖于凸分析. 与一般凸优化(14.1)的最优性条件 (14.5)相比，为了证明命题 14.1, 理解集合$\Omega \cap \mathcal{A}$在点x_*的法锥是充分的. 为了得到合理的刻画，从证明一个一般事实开始.

命题 14.2. 设$\Omega \subseteq \mathbb{R}^n$是闭凸集. 设$\mathcal{A}$表示仿射集$\{x : Ax = b\}$，其中$A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. 假设集合$\operatorname{ri}\Omega \cap \mathcal{A}$非空[8]. 那么对于任何$x \in \Omega \cap \mathcal{A}$,

$$\mathcal{N}_{\Omega\cap\mathcal{A}}(x) = \mathcal{N}_\Omega(x) + \mathcal{N}_\mathcal{A}(x)\,.$$

Proof. 断言"$\supseteq$"是直接的. 为了说明它， 设$u \in \mathcal{N}_\Omega(x)$, 对于$\lambda \in \mathbb{R}^m$. 假设$z \in \Omega \cap \mathcal{A}$，那么

$$\langle z - x, u + A^\top\lambda\rangle = \langle z - x, u\rangle \leqslant 0,$$

等式是因为$z - x \in \operatorname{null} A$，不等式是因为$u \in N_\Omega(x)$. 这蕴含着 $u + A^\top\lambda \in \mathcal{N}_{\Omega\cap\mathcal{A}}(x)$.

对于反包含关系，设$v \in \mathcal{N}_{\Omega\cap\mathcal{A}}(x)$. 那么有

$$v^\top(z - x) \leqslant 0 \;\forall\, z \in \Omega \cap \mathcal{A}.$$

[8]此即经典的Slater条件.

现在定义集合

$$C_1 = \left\{(y,\mu) \in \mathbb{R}^{n+1} \,:\, y = z - x,\ z \in \Omega,\ v^\top y \geqslant \mu\right\},$$
$$C_2 = \left\{(y,\mu) \in \mathbb{R}^{n+1} \,:\, y \in \operatorname{null} A,\ \mu = 0\right\}.$$

请注意$\operatorname{ri} C_1 \cap C_2 = \emptyset$，因为否则会存在$(\widehat{y},\widehat{\mu})$使得

$$v^\top \widehat{y} > \widehat{\mu} = 0$$

和 $\widehat{y} \in \mathcal{T}_{\Omega\cap\mathcal{A}}(x)$. 这将与假设 $v \in \mathcal{N}_{\Omega\cap\mathcal{A}}(x)$矛盾. 由于它们的交集是空的，可以将$\operatorname{ri} C_1$与 C_2恰当地分开，即存在(w,γ)使得

$$\sup_{(y,\mu)\in C_1} \{w^\top y + \gamma\mu\} < \inf_{(y,\mu)\in C_2} \{w^\top y + \gamma\mu\} = 0 \tag{14.6}$$

进一步，由于C_2 是子空间，同时有

$$w^\top u = 0\ \forall u\ \in \operatorname{null} A. \tag{14.7}$$

特别地，这意味着存在某$\lambda \in \mathbb{R}^m$使得$w = A^T\lambda$ 成立. 现在，γ必须是非负的，因为否则，由C_1的构造，(通过令μ趋于负无穷可以看到)有

$$\sup_{(y,\mu)\in C_1} \{w^\top y + \gamma\mu\} = \infty.$$

如果$\gamma = 0$, 那么由(14.6)和(14.7)，有

$$\sup_{z\in\Omega} w^\top(z - x) < \inf_{y\in \operatorname{Null} A} w^\top y = 0.$$

这意味着$\Omega - \{x\}$的相对内部与A的核没有交集，这与假设Slater条件满足矛盾. 这样，可以断言γ是严格正的. 由齐次性，不妨设$\gamma = 1$.

为了完成讨论，注意到由(14.6)和(14.7)，现在有

$$(w + v)^\top(z - x) \leqslant 0\ \forall z\ \in \Omega.$$

这意味着$v + w \in \mathcal{N}_\Omega(x)$，并且已经证明 $w = A^T\lambda$. 这样，

$$v = (v + w) - w \in \mathcal{N}_\Omega(x) + \mathcal{N}_\mathcal{A}(x).$$

■

现在针对这里关心的问题翻译命题的结果. 利用 (14.5) 和命题 14.2, 知x_*是问题 (14.4) 的最优解当且仅当$x_* \in \Omega, Ax_* = b$，并且存在$\lambda^* \in \mathbb{R}^m$使得

$$-\nabla f(x_*) + A^\top\lambda^* \in \mathcal{N}_\Omega(x_*).$$

不能直接用这个化简结果，因为它不能提供求解约束优化问题的数值方法. 然而，它是深入研究对偶性的基础.

14.3 对偶问题

对偶性将任何一个约束优化问题 (14.4)与一个凹极大化联系在一起，这个关联问题的解为原始问题的最优值提供一个下界. 特别地，在温和的假设下，将说明通过求解对偶问题能间接地求解原始问题.

继续关注有等式约束的原始问题 (14.4). 这里，假设Ω是闭凸集, f是可微的，A满秩. Lagrange对偶性的关键是问题 (14.4)等价于

$$\min_{x\in\Omega}\max_{\lambda\in\mathbb{R}^n} f(x)+\lambda^T(Ax-b).$$

为了理解这个事实，仅需注意到如果$Ax\neq b$, 那么关于λ的最大值是正无穷. 另一方面，如果$Ax=b$是可行的，那么关于λ的最大值等于$f(x)$.

与 (14.4)关联的**对偶问题**(dual problem) 是

$$\max_{\lambda\in\mathbb{R}^n}\min_{x\in\Omega} f(x)+\lambda^T(Ax-b),$$

对偶函数

$$g(\lambda):=\min_{x\in\Omega} f(x)+\lambda^T(Ax-b)$$

是线性函数的逐点下确界，从而对偶函数总是凹函数. 因此不管f 是哪种形式，Ω 取什么，对偶问题都是**凹极大化问题**.

14.4 弱对偶性

现在证明对偶问题总为原始问题提供了下界.

命题 14.3 (弱对偶性)**.** 对于所有$\lambda\in\mathbb{R}^m$和 (14.4) 的所有可行解x，有

$$g(\lambda)\leqslant f(x).$$

Proof. 由g的定义和$x\in\Omega$有

$$g(\lambda)\leqslant f(x)+\lambda^\top(Ax-b)=f(x)$$

其中的等式是因为$Ax=b$. ■

基于弱对偶性，可将对偶问题理解成为 (14.4) 确定下界的系统方法.

命题 14.4. 针对任何函数 $\varphi(x,z)$，有

$$\inf_x\sup_z \varphi(x,z)\geqslant\sup_z\inf_x\varphi(x,z).$$

Proof. 证明本质上是同义反复的. 注意总是有

$$\varphi(x,z)\geqslant\inf_x\varphi(x,z),\ \forall x,\forall z$$

关于第二个变元取最大验证了

$$\sup_z\varphi(x,z)\geqslant\sup_z\inf_x\varphi(x,z)\quad\forall x\,.$$

现在，关于x极小化左边证明了

$$\inf_x\sup_z \varphi(x,z)\geqslant\sup_z\inf_x\varphi(x,z).$$

这恰好就是需要的论断. ■

14.5 强对偶性

针对凸优化问题，能证明一个相当深刻的结论. 即，原始和对偶问题取得相同的最优值. 此外，如果知道对偶最优解，能从一个更简单的优化问题提取出原始最优解.

定理 14.5 (强凸性)**.**

(i) 如果 $\exists \hat{x} \in \operatorname{ri}\Omega$ 也满足等式约束(Slater条件)，同时原始问题 (14.4)有一个最优解，那么对偶问题也有最优解，并且原始-对偶最优值相等.

(ii) 事实 x_*是原始最优的同时λ^*是对偶最优解的充要条件是$Ax_* = b$, $x_* \in \Omega$和

$$x_* \in \arg\min_{x\in\Omega} \ \mathcal{L}(x, \lambda^*) := f(x) + {\lambda^*}^T(Ax - b)$$

Proof. 现在由命题 14.1, x_*是 (14.4)最优解当且仅当存在$\lambda^* \in \mathbb{R}^m$ 使得

$$\langle \nabla f(x_*) + A^T\lambda^*, x - x_* \rangle \geqslant 0 \ \ \forall x \in \Omega$$

和$Ax_* = b$. 请注意该条件蕴含着x_*在Ω上极小化$\mathcal{L}(x, \lambda^*)$.

由前面的讨论，现在得到

$$\begin{aligned} g(\lambda^*) &= \inf_{x\in\Omega} \mathcal{L}(x, \lambda^*) \\ &= \mathcal{L}(x_*, \lambda^*) \\ &= f(x_*) + {\lambda^*}^T(Ax_* - b) = f(x_*). \end{aligned}$$

由此完成证明. ■

15 利用对偶性的算法

前一讲的Lagrange对偶理论可用于设计求解对偶问题的最优化算法，其针对对偶函数执行优化. 通常，转移到对偶能简化计算，或者能执行并行计算.

15.1 对偶函数的性质

回忆**原始问题**(primal problem) (14.4). 通过考虑**Lagrange函数**

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

能得到相应的对偶问题，称其中的λ_i是**Lagrange乘子**(Lagrange multipliers). 定义**对偶函数**(dual function)

$$g(\lambda) := \inf_{x\in\Omega} L(x, \lambda).$$

对偶问题(dual problem)是

$$\sup_{\lambda\in\mathbb{R}^m} g(\lambda).$$

定义 15.1 (凹函数)**.** 函数f是凹的 $\iff$ $-f$是凸的.

事实 15.2. (即使f和Ω不是凸的)对偶函数总是凹的.

Proof. 对任何 $x \in \Omega$, $L(x,\lambda)$关于λ是线性函数，因此$g(\lambda)$ 是一族线性函数的下确界，因此是凹的. ■

由于对偶函数g是凹函数，因此对偶问题的任何局部极大点都是全局极大点. 这使得求解对偶问题是一个很吸引人的想法. 然而，求解对偶问题的主要困难是，由于仅在求解了子问题后才能得到对偶函数在一点的值，因此对偶函数通常没有解析表达式. 下面研究对偶函数的可微性和次可微性. 这些性质在极大化对偶函数时很有用.

命题 15.3 (P.276 Dinskin定理)**.** 如果λ使得$g(\lambda)$有限，那么$-g$在λ处是次可微的，并且

$$b - Ax(\lambda) \in \partial(-g(\lambda)),$$

其中$x(\lambda) \in \arg\inf_{x\in\Omega} L(x,\lambda)$.

15.2 对偶梯度上升法

对偶函数$g(\lambda)$的凹性确保$-g(\lambda)$的次梯度是存在的，因此能用次梯度法来优化$g(\lambda)$. **对偶梯度上升**(dual gradient ascent) 算法如下:

从初始点λ_0开始. 对所有$t \geqslant 0$:

$$x_t = \arg\inf_{x\in\Omega} L(x,\lambda_t)$$

$$\lambda_{t+1} = \lambda_t + \eta(Ax_t - b)$$

由次梯度法(定理 2.7)，该算法的收敛速率是$O(1/\sqrt{t})$.

15.3 增广Lagrange函数法/乘子法

鉴于对偶梯度上升法通过在(次)梯度方向走一步来更新 λ_{t+1}，从使用临近算子作为更新规则迭代地优化λ 出发，可以激发众所周知的**对偶临近点法**(dual proximal point method):

$$\lambda_{t+1} = \mathrm{prox}_{\eta_t g}(\lambda_t) = \arg\sup_{\lambda} \underbrace{\underbrace{\inf_{x\in\Omega} f(x) + \lambda^T(Ax-b)}_{g(\lambda)} - \underbrace{\frac{1}{2\eta_t}\|\lambda-\lambda_t\|^2}_{\text{proximal term}}}_{h(\lambda)}$$

请注意该表达式包含了邻近项，这使得$h(\lambda)$变成强凸的.

然而，该更新并不总是直接有用的，由于它要求关于λ 优化$h(\lambda)$，这不一定有闭合解. 相反地，注意到如果能交换inf 和 sup (比如强对偶性，当Ω是凸集时应

用Sion定理)那么可以重写

$$\begin{aligned}&\sup_{\lambda}\inf_{x\in\Omega} f(x)+\lambda^T(Ax-b)-\tfrac{1}{2\eta_t}\|\lambda-\lambda_t\|^2\\ =&\inf_{x\in\Omega}\sup_{\lambda} f(x)+\lambda^T(Ax-b)-\tfrac{1}{2\eta_t}\|\lambda_t-\lambda\|^2\\ =&\inf_{x\in\Omega} f(x)+\lambda_t^T(Ax-b)+\tfrac{\eta_t}{2}\|Ax-b\|^2\end{aligned}$$

其中内部sup的最优解具有闭合式$\lambda=\lambda_t+\eta_t(Ax-b)$. 为了单独考虑剩下的关于$x$的优化问题，做如下定义.

定义 15.4 (增广Lagrange函数)**. 增广Lagrange函数(augmented Lagrangian)**是

$$L_\eta(x,\lambda)=f(x)+\lambda^T(Ax-b)+\tfrac{\eta}{2}\|Ax-b\|^2$$

增广Lagrange函数法(Augmented Lagrangian Method, ALM)，亦称乘子法(Method of Multipliers, MM)是按如下方式定义迭代:

$$\begin{aligned}x_t&=\arg\inf_{x\in\Omega} L_{\eta_t}(x,\lambda_t)\\ \lambda_{t+1}&=\lambda_t+\eta_t(Ax_t-b)\end{aligned}$$

尽管该迭代看起来与对偶梯度上升法相似，但也存在显著不同：

- 乘子法能够加速收敛(比如针对非光滑函数), 但是由于增广Lagrange函数中额外的项，可能会使得计算x_t更加困难.
- $L(x,\lambda)$关于x是凸的，但是$L_\eta(x,\lambda)$ 关于x 是强凸的 (如果A是满秩的)
- 收敛速率是 $O(1/t)$. 更精确地，针对大小为η 的常数步长, 能证明乘子法满足

$$g(\lambda_t)-g^*\geqslant-\tfrac{\|\lambda^*\|^2}{2\eta t}.$$

15.4 对偶分解法

对偶分解的主要优点是由它易于得到可并行化的更新规则. 假设能将原始问题剖分成大小为$(n_i)_{i=1}^N$的块，即

$$x^T=((x^{(1)})^T,\cdots,(x^{(N)})^T)\qquad x^{(i)}\in\mathbb{R}^{n_i},\sum_{i=1}^N n_i=n$$

$$A=[A_1|\cdots|A_N]\qquad Ax=\sum_{i=1}^N A_ix^{(i)}$$

$$f(x)=\sum_{i=1}^N f_i(x^{(i)})$$

那么，Lagrange函数

$$L(x,\lambda)=\sum_{i=1}^N(\underbrace{f_i(x^{(i)})+\lambda^TA_ix^{(i)}-\tfrac{1}{N}\lambda^Tb}_{L_i(x^{(i)},\lambda)})=\sum_{i=1}^N L_i(x^{(i)},\lambda)$$

关于x也是分离的. 和式中的每一项由一个非耦合剖分$(x^{(i)}, A_i, f_i)$ 组成，因此能够并行地极小化和式中的每一项. 由此得到**对偶分解算法(dual decomposition algorithm)**:

- 在员工节点并行: $x_t^{(i)} = \arg\inf_{x^{(i)}} L_i(x^{(i)}, \lambda_t)$
- 在主节点: $\lambda_{t+1} = \lambda_t + \eta(Ax_t - b)$

例子 15.5 (共识优化)**.** 共识优化(Consensus optimization)是分布式计算中产生的应用，可利用对偶分解算法来求解. 已知$G = (V, E)$,

$$\min_x \sum_{v \in V} f_v(x_v) : x_v = x_u \ \forall (u, v) \in E.$$

该问题关于$v \in V$是可分离的，因此可以应用对偶分解法.

例子 15.6 (网络效用最大化)**.** 假设拥有的网络共有k条链路，第ℓ条链路的容量是c_ℓ. 感兴趣的是为N 条穿过这些链路的固定路由分配不同的流，使得在满足资源约束不超限的前提下，极大化总的效用. 设$x_i \in \mathbb{R}$代表分给流i的数量，$U_i : \mathbb{R} \to \mathbb{R}$是凸效用函数，其返回流$i$数量为$x_i$时所得到的效用大小. 最优化问题是

$$\max_{x \geqslant 0} \sum_{i=1}^{N} U_i(x_i) : Rx \leqslant c$$

其中R是$k \times N$的路由矩阵，如果流i 的路由通过链路ℓ，它的第(ℓ, i)个元素是 1 ，否则是 0 .

取相反数，可将原始问题写成标准形，即极小化问题:

$$\min_{x \geqslant 0} -\sum_i U_i(x_i) : Rx \leqslant c.$$

那么对偶问题是

$$\max_{\lambda \geqslant 0} \min_{x \geqslant 0} \sum_i -U_i(x_i) + \lambda^T(Rx - c)$$

其中原始不等式约束 $Rx \leqslant c$产生了$\lambda \geqslant 0$约束. 可将第二项写作$\lambda^T \left(\sum_i [R_i x_i - \frac{1}{N} c \right])$, 求对偶函数时，即已知$\lambda$，求$L(x, \lambda)$关于$x$的极小值时，对应的问题关于$i$ 是可分离的，因此可以用对偶分解法. 由此得到一个并行算法，其中每个员工节点计算

$$\arg\max_{x_i \geqslant 0} U_i(x_i) - \lambda^T R_i x_i$$

主节点计算

$$\lambda_{t+1} = [\lambda_t + \eta(Rx - c)]_+$$

取正部是因为$\lambda \geqslant 0$约束.

旁白: 在资源分配问题中，最优点处的对偶变量λ的值具有经济学解释，即资源的“价格”. 在该例中，可将λ_ℓ解释成链路ℓ上每单位流的运输价格.

15.5 ADMM-交替方向乘子法

尽管能利用对偶分解来并行化对偶次梯度上升法，但是增广拉格朗日函数法却不行，这是因为$\|Ax-b\|^2$项引入了决策变量之间的耦合，这使得增广Lagrange函数关于x不再是可分离的.

交替方向乘子法(Alternating Direction Method of Multipliers, ADMM) 的目的是两全其美：既希望享有对偶分解法提供的并行化，也想获得增广Lagrange函数法快的收敛速率. 将看到，类似于对偶分解法，ADMM将决策变量剖分成两块. 也类似于乘子法，ADMM 使用增广Lagrange函数$L_\eta(x,z,\lambda_t)$. 可将ADMM看成松弛版的乘子法.

考虑形如

$$\begin{aligned}\underset{x,z}{\text{minimize}} \quad & f(x)+g(z)\\ \text{subject to} \quad & Ax+Bz=c\end{aligned}$$

的问题. 换句话说，可将目标和约束拆分成x和z两块. 在单个优化步内，乘子法关于两块变量来联合优化增广Lagrange函数:

$$\begin{aligned}(x_{t+1},z_{t+1}) &= \inf_{x,z} L_\eta(x,z,\lambda_t)\\ \lambda_{t+1} &= \lambda_t+\eta(Ax_{t+1}+Bz_{t+1}-c)\end{aligned}$$

与此相反，ADMM在关于x和z优化增广Lagrange函数之间交替地("ADMM" 中的A)进行:

$$\begin{aligned}x_{t+1} &= \inf_x L_\eta(x,z_t,\lambda_t)\\ z_{t+1} &= \inf_z L_\eta(x_{t+1},z,\lambda_t)\\ \lambda_{t+1} &= \lambda_t+\eta(Ax_{t+1}+Bz_{t+1}-c)\end{aligned}$$

这与乘子法不同，后者不能并行化，因为在z_{t+1}之前必须计算出x_{t+1}. 还有，收敛保证更弱：得不到收敛速率，仅能得到渐近收敛保证.

定理 15.7. 假设f和g的上图均是非空闭凸集，并且*Laglange*函数L 具有鞍点x_*, z_*, λ^*, 即

$$\forall x,z,\lambda : L(x_*,z_*,\lambda)\leqslant L(x_*,z_*,\lambda^*)\leqslant L(x,z,\lambda^*).$$

那么，随着$t\to\infty$, ADMM 满足$f(x_t)+g(z_t)\to p^*, \lambda_t\to\lambda^*$.

按语: 鞍点很有用，因为inf和sup可以交换. 为了说明，请注意鞍点条件

$$L(x_*,\lambda)\leqslant L(x_*,\lambda^*)\leqslant L(x,\lambda^*)$$

蕴含着

$$\begin{aligned}\inf_x\sup_\lambda L(x,\lambda) &\leqslant \sup_\lambda L(x_*,\lambda)\\ &\leqslant L(x_*,\lambda^*)\\ &= \inf_x L(x,\lambda^*)\\ &\leqslant \sup_\lambda\inf_x L(x,\lambda).\end{aligned}$$

16 Fenchel对偶与算法

本节介绍Fenchel共轭. 首先，回忆对于一元可微实值凸函数$f(x)$, 定义

$$f^*(p) := \sup_x px - f(x)$$

是f的Legendre变换. 几何直观见图**??**. 将Legendre变换扩展到可能是非凸和/不可微函数，就是Fenchel共轭.

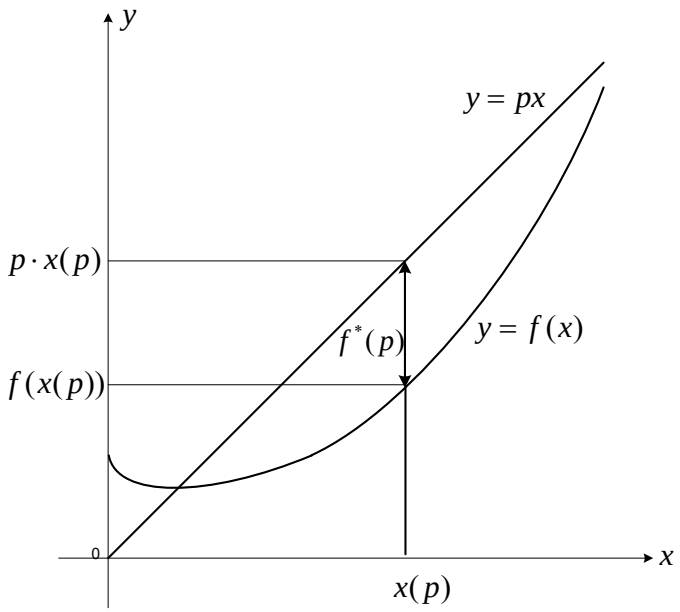


图 16.1: Legendre变换. 函数$f^*(p)$是为了让线性函数px在x 与f 相切，而需要垂直平移f 上图的大小.

定义 16.1 (Fenchel共轭)**.** 函数$f : \mathbb{R}^n \to \mathbb{R}$的**Fenchel共轭**(Fenchel Conjugate)是

$$f^*(p) = \sup_x \langle p, x\rangle - f(x)$$

现在给出关于Fenchel共轭有用的事实. 其证明留作练习.

事实 16.2. f^* 关于p是凸函数.

的确，f^*是一族关于p的仿射函数族$f_x := \langle x, p\rangle - f(x), x \in \mathbb{R}^n$ 的逐点上确界，因此是凸的. 这样，也将f 的Fenchel 共轭称作它的凸共轭.

事实 16.3. 如果f是凸的，那么$f^*(f^*(x)) = f$.

换句话说，对于凸函数，Fenchel共轭是自己的反函数. 现在，也能将函数的次微分与它的Fenchel 共轭关联起来. 直观上，观察$0 \in \partial f^*(p) \iff 0 \in p - \partial f(x) \iff p \in \partial f(x)$. 这被总结成如下更一般的事实.

事实 16.4 (刻画次微分)**.** f^*在p处的次微分是$\partial f^*(p) = \{x : p \in \partial f(x)\}$.

实际上，$\partial f^*(0)$是f的最小点集合. 在如下定理中，引入Fenchel 对偶性.

定理 16.5 (Fenchel对偶性)**.** 假设f是真凸的，g是真凹的. 那么

$$\min_x f(x) - g(x) = \max_p g^*(p) - f^*(p)$$

其中g^*是g的凹共轭，定义为$\inf_x \langle p, x\rangle - g(x)$.

图16.2给出了一维情况时Fenchel对偶性的几何直观.

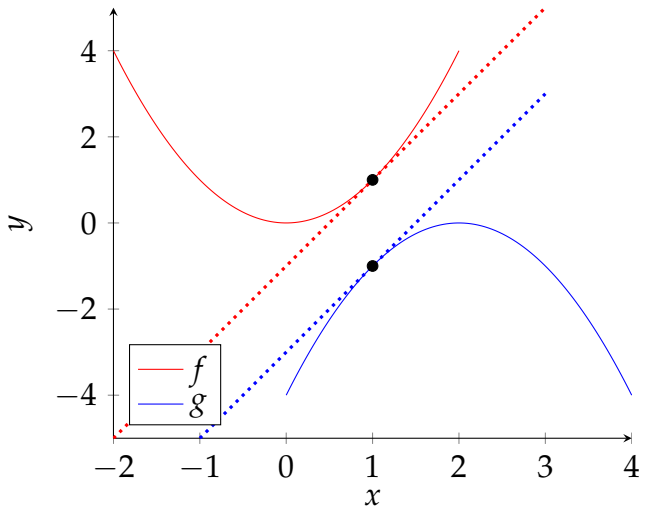


图 16.2: 一维中的Fenchel对偶性

在极小化问题中，想要找x使得f和g在x处的垂直距离尽可能地小. 在(对偶)极大化问题中，画f 和g图形的切线，使得二者的切线有相同的斜率p, 想找p 使得切线之间的垂直距离尽可能地大. 上面的对偶定理表明：强对偶性成立，即两个问题有相同的最优值.

能从已经研究过的Lagrange对偶得到Fenchel对偶. 为此，需要为定理 16.5 中的极小化问题引入约束. 约束版的自然重新表述如下：

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & f(x) - g(z) \\ \text{subject to} \quad & x = z. \end{aligned} \tag{16.1}$$

16.1 得到经验风险最小化的对偶问题

在经验风险最小化中，经常想要极小化正则化的经验风险，即

$$P(w) = \frac{1}{m}\sum_{i=1}^{m} \phi_i(\langle w, x_i\rangle) + \lambda g(w). \tag{16.2}$$

这里将$w \in \mathbb{R}^d$ 看作模型参数，想要关于它优化(在这种情况下它对应着挑选一个超平面), x_i 看作训练集中第i 个样例的特征. $\phi_i(\langle \cdot, x_i\rangle)$ 是针对第i 个训练样例的损失函数，也可能与它的标签有关. $g(w)$ 是正则化子，典型选取形如 $g(w) = \frac{1}{2}\|w\|_2^2$，$\lambda > 0$是超参数，用于控制模型复杂度.

可将原始问题$\min_{w\in\mathbb{R}^n} P(w)$等价地表示为

$$\begin{array}{ll}\underset{w,z}{\text{minimize}} & \frac{1}{m}\sum_{i=1}^m \phi_i(z_i) + \lambda g(w) \\ \text{subject to} & \frac{1}{m}X^\top w = \frac{1}{m}z,\end{array} \tag{16.3}$$

其中$X = [x_1, \cdots, x_m] \in \mathbb{R}^{d\times m}$. 由Lagrange对偶性，其对偶函数

$$\begin{aligned} D(\alpha) &= \min_{w,z} \frac{1}{m}\sum_{i=1}^m \phi_i(z_i) + \lambda g(w) + \frac{1}{m}\alpha^\top(z - X^\top w) \\ &= \min_{w,z} \frac{1}{m}\sum_{i=1}^m [\phi_i(z_i) + \alpha_i z_i] + \lambda g(w) - \frac{1}{m}\alpha^\top X^\top w \\ &= -\left[\max_{w,z} -\sum_{i=1}^m \frac{1}{m}[\phi_i(z_i) + \alpha_i z_i] + \frac{1}{m}(X\alpha)^\top w - \lambda g(w)\right] \\ &= -\frac{1}{m}\sum_{i=1}^m \max_{z_i}[-\phi_i(z_i) - \alpha_i z_i] - \lambda \max_w [(\tfrac{X\alpha}{m\lambda})^\top w - g(w)] \\ &= -\frac{1}{m}\sum_{i=1}^n \phi_i^*(-\alpha_i) - \lambda g^*(\tfrac{X\alpha}{m\lambda}) \end{aligned}$$

其中 ϕ_i^*和g^*分别是ϕ_i和g的Fenchel共轭. 由弱对偶性, $D(\alpha) \leqslant P(w)$. 对偶问题是

$$\max_{\alpha\in\mathbb{R}^m} D(\alpha) := -\frac{1}{m}\sum_{i=1}^m \phi_i^*(-\alpha_i) - \lambda g^*(\tfrac{X\alpha}{m\lambda}). \tag{16.4}$$

对于$g(w) = \frac{1}{2}\|w\|_2^2$, $g^*(p) = \frac{1}{2}\|p\|_2^2$. 因此$g$ 是自己的凸共轭. 在这种情况下对偶变成:

$$\max_{\alpha\in\mathbb{R}^m} D(\alpha) := -\sum_{i=1}^m \phi_i^*(-\alpha_i) - \tfrac{\lambda}{2m}\left\|\frac{1}{\lambda}\sum_{i=1}^m \alpha_i x_i\right\|_2^2. \tag{16.5}$$

在求解正则化项的共轭函数时，得到映射

$$w(\alpha) = \frac{1}{m\lambda}\sum_{i=1}^m \alpha_i x_i,$$

它把原始和对偶变量联系了起来. 特别地，这表明最优超平面的法向量位于数据生成的子空间内. 这里有一些可以使用该框架的模型案例. 以下记$z_i = \langle w, x_i\rangle$.

例子 16.6 (线性SVM)**.** 已知$y_i \in \{-1, 1\}$. 使用合页损失作为ϕ_i. 这对应于

$$\phi_i(z_i) = \max\{0, 1 - y_i z_i\}, \quad \phi_i^*(-\alpha_i) = -\alpha_i y_i, \alpha_i y_i \in [0, 1].$$

例子 16.7 (最小二乘线性回归)**.** 已知$y_i \in \mathbb{R}$. 使用平方损失作为ϕ_i. 这对应于

$$\phi_i(z_i) = (z_i - y_i)^2, \quad \phi_i^*(-\alpha_i) = -\alpha_i y_i + \alpha_i^2/4.$$

最后，用一个事实结尾，其将ϕ_i的光滑性与ϕ_i^*的强凸性关联起来.

事实 16.8. 如果ϕ_i是$\frac{1}{\gamma}$-光滑的，那么ϕ_i^*是γ- 强凸的.

16.2 随机对偶坐标上升法

本节讨论一个针对经验风险极小化 (16.3)的特定算法，随机对偶坐标上升法(stochastic dual coordinate ascent, SDCA)，其本质是求解对偶问题 (16.4)的随机坐标上升法. 它的主要思想是随机地挑选一个指标$i \in [m]$，然后在保持其它坐标固定的同时，关于坐标i 求解对偶问题.

更精确地，算法执行如下步骤:

1. Start from $w^0 := w(\alpha^0)$
2. For $t = 1, \ldots, T$:
   (a) Randomly pick $i \in [m]$
   (b) Find $\Delta\alpha_i$ which maximizes

$$-\phi_i^*\left(-(\alpha_i^{t-1} + \Delta\alpha_i)\right) - \frac{\lambda}{2m}\left\|w^{t-1} + \frac{1}{\lambda}\Delta\alpha_i x_i\right\|^2 \tag{16.6}$$

3. Update the dual and primal solution
   (a) $\alpha^t = \alpha^{t-1} + \Delta\alpha_i e_i$
   (b) $w^t = w^{t-1} + \frac{1}{m\lambda}\Delta\alpha_i x_i$

针对某些损失函数，子问题 (16.6)的极大点$\Delta\alpha_i$具有解析解. 比如，对于合页损失，可以给出显式解:

$$\Delta\alpha_i = y_i \max\left\{0, \min\left\{1, \frac{1 - x_i^T w^{t-1} y_i}{\|x_i\|^2/\lambda m} + \alpha_i^{t-1} y_i\right\}\right\} - \alpha_i^{t-1},$$

对于平方损失，它的显式解为:

$$\Delta\alpha_i = \frac{y_i - x_i^T w^{t-1} - 0.5\alpha_i^{t-1}}{0.5 + \|x\|^2/\lambda m}.$$

请注意这些更新要求对原始解和对偶解都执行更新.

现在，陈述[SSZ13]中给出的引理, 其蕴含着SDCA的线性收敛性. 下面假设对所有x 有$\|x_i\| \leqslant 1$, $\phi_i(z_i) \geqslant 0$, 并且 $\phi_i(0) \leqslant 1$.

引理 16.9. 假设ϕ_i^*是γ-强凸的，其中$\gamma > 0$. 那么:

$$\mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})] \geqslant \frac{s}{m}\mathbb{E}[P(w^{t-1}) - D(\alpha^{t-1})],$$

其中 $s = \frac{\lambda m\gamma}{1+\lambda m\gamma} \in (0, 1)$.

略去该结论的证明， 然而给出使用这个引理证明SDCA的线性收敛性的简短讨论. 记

$$\epsilon_D^t := D(\alpha^*) - D(\alpha^t).$$

因为对偶解为原始问题的最优值提供了下界，从而有:

$$\epsilon_D^t \leqslant P(w^t) - D(\alpha^t).$$

进一步，可以写：

$$D(\alpha^t) - D(\alpha^{t-1}) = \epsilon_D^{t-1} - \epsilon_D^t.$$

给该等式两边取期望，并应用引理 16.9, 得到：

$$\begin{aligned}\mathbb{E}[\epsilon_D^{t-1} - \epsilon_D^t] &= \mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})] \\ &\geqslant \tfrac{s}{m}\mathbb{E}[P(w^{t-1}) - D(\alpha^{t-1})] \\ &\geqslant \tfrac{s}{m}\mathbb{E}[\epsilon_D^{t-1}].\end{aligned}$$

重新整理，并递归地应用前面的讨论产生：

$$\mathbb{E}[\epsilon_D^t] \leqslant \left(1 - \tfrac{s}{m}\right)\mathbb{E}[\epsilon_D^{t-1}] \leqslant \left(1 - \tfrac{s}{m}\right)^t \epsilon_D^0.$$

由这个不等式，能得到：为了获得ϵ对偶误差，需要$O(m + \frac{1}{\lambda\gamma}\log(1/\epsilon))$ 步迭代.

利用引理 16.9, 也能够上控原始误差. 再次使用对偶解是原始解的下估计这个事实，以如下方式提供界：

$$\begin{aligned}\mathbb{E}[P(w^t) - P(w^*)] &\leqslant \mathbb{E}[P(w^t) - D(\alpha^t)] \\ &\leqslant \tfrac{m}{s}\mathbb{E}[D(\alpha^{t+1}) - D(\alpha^t)] \\ &\leqslant \tfrac{m}{s}\mathbb{E}[\epsilon_D^t],\end{aligned}$$

其中最后一个不等式忽略了负项 $-\mathbb{E}[\epsilon_D^{t-1}]$.

17 反向传播与伴随

从现在开始，放弃凸性所提供的奢华，进入非凸函数领域. 截至目前所看到的问题中，得到梯度的闭式表达式是相当直接的. 但是，做到这一点对一般的非凸函数可能是一个极具挑战性的任务. 本次课，准备聚焦于能表示成多个函数的复合函数. 接下来将引入**反向传播(backpropagation, BP)** - 一种流行的技术，其利用函数的复合本质逐步计算梯度.

下面的阐述基于Tim Viera博大而富有洞见的关于BP的笔记. 这些笔记针对感兴趣的读者也提供了优秀的演示代码.

17.1 热身

关于BP的共识是**"它恰好是链式法则"**("it's just chain rule"). 这种观点并非特别有益. 然而，将会看到远不止于此. 作为热身的例子，考虑如下与$g : \mathbb{R}^n \to \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$有关的优化问题:

$$\min_x f(g(x)).$$

使用在ADMM背景下看到的类似技巧，能将这个问题重新写作

$$\begin{aligned}\min_{x,z} \quad & f(z) \\ \text{s.t.} \quad & z = g(x)\end{aligned}$$

请注意已经将关于x的原始无约束优化问题转化成关于x和z的约束优化问题，后者的拉格朗日函数：

$$\mathcal{L}(x, z, \lambda) = f(z) + \lambda(g(x) - z).$$

置$\nabla\mathcal{L} = 0$, 得到如下最优性条件：

$$0 = \nabla_x \mathcal{L} = \lambda g'(x) \Leftrightarrow 0 = \lambda g'(x), \tag{17.1a}$$
$$0 = \nabla_z \mathcal{L} = f'(z) - \lambda \Leftrightarrow \lambda = f'(z), \tag{17.1b}$$
$$0 = \nabla_\lambda \mathcal{L} = g(x) - z \Leftrightarrow z = g(x). \tag{17.1c}$$

这蕴含着

$$0 = f'(g(x))g'(x) = \nabla_x f(g(x)). \qquad \text{(由链式法则)}$$

因此，求解拉格朗日方程给出了计算梯度的逐步法. 像将要看到的那样，在相当普遍的情况下，这也是成立的. 注意到“当求解 (17.1)中的方程组时，并**没有**使用链式法则”是非常重要的. 链式法则仅出现在了正确性的证明中.

17.2 通用表述

任何复合函数都可以用它的计算图来描述. 只要计算图中的基本函数是可微的，就能执行和上面一样的程式. 在进一步采取行动之前，先引入一些记号:

- 有向非循环计算图：$\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- 节点数目： $|V| = n$
- 第 i个节点的父节点集：$\alpha(i) = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}\}$
- 第i个节点的子节点集：$\beta(i) = \{j \in \mathcal{V} : (i, j) \in \mathcal{E}\}$
- 在第i个节点处的计算：$f_i(z_{\alpha(i)})$，其中$f_i : \mathbb{R}^{|\alpha(i)|} \to \mathbb{R}^{|\beta(i)|}$. 请注意$f_i$是向量值函数，是映射.
- 节点：
  - 输入节点- $z_1, \ldots, z_d$
  - 中间节点 - $z_{d+1}, \ldots, z_{n-1}$
  - 输出节点 - z_n

那么，一般表述是

$$\begin{array}{ll} \min & z_n \\ \text{s.t.} & z_i = f_i(z_{\alpha(i)}), i = 1, \cdots, n. \end{array}$$

它的拉格朗日函数是

$$\mathcal{L}(z, \lambda) = z_n + \sum_{i=1}^{n} \lambda_i (f_i(z_{\alpha(i)}) - z_i).$$

像那个热身的例子中，置 $\nabla\mathcal{L} = 0$. 可将这看作一个由两步组成的算法：

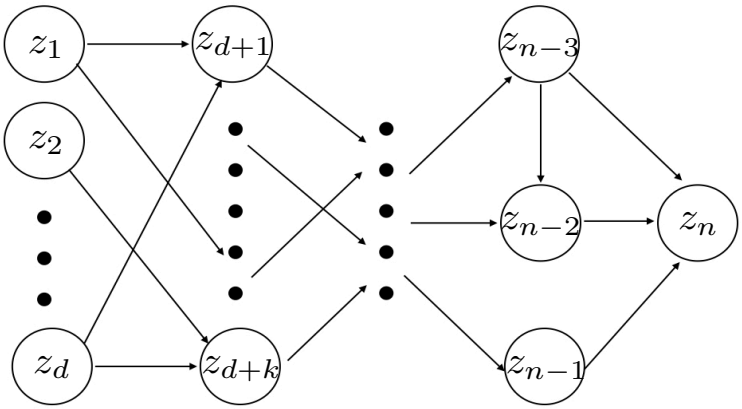


图 17.1: 计算图

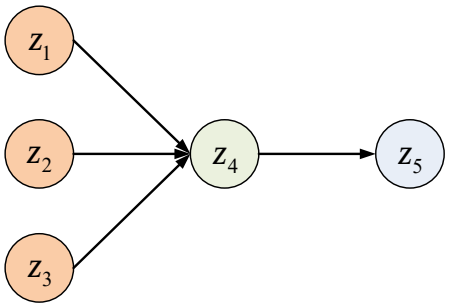


图 17.2: 热身例子的计算图，其中$d = 3, L = 3, n = 5$

BP算法

- **步 1**: 置$\nabla_\lambda \mathcal{L} = 0$, 即,

$$\nabla_{\lambda_i} \mathcal{L} = z_i - f_i(z_{\alpha(i)}) = 0 \Leftrightarrow z_i = f_i(z_{\alpha(i)}) \tag{FP}$$

  观察: 由 (FP)，节点(z_i)处的值是用父节点集的值来计算得到的，因此称 (FP)是**向前传递((forward pass)或者向前传播(forward propagation)**，因为是已知输入x，计算输出$f_n(x)$的过程.

- **步 2**: 置$\nabla_{z_j} \mathcal{L} = 0$,
  - 对于$j = n$,

$$0 = \nabla_{z_n} \mathcal{L} = 1 - \lambda_n \Leftrightarrow \lambda_n = 1$$

  - 对于$j < n$,

$$\begin{aligned} 0 &= \nabla_{z_j} \mathcal{L} \\ &= \nabla_{z_j} \Big(z_n + \sum_i \lambda_i (f_i(z_{\alpha(i)}) - z_i) \Big) \\ &= \sum_i \lambda_i (\nabla_{z_j} f_i(z_{\alpha(i)}) - \nabla_{z_j} [z_i]) \\ &= \sum_i \lambda_i \nabla_{z_j} f_i(z_{\alpha(i)}) - \lambda_j \\ &= \sum_{i \in \beta(j)} \lambda_i \frac{\partial f_i(z_{\alpha(i)})}{\partial z_j} - \lambda_j \\ \Leftrightarrow \quad \lambda_j &= \sum_{i \in \beta(j)} \lambda_i \frac{\partial f_i(z_{\alpha(i)})}{\partial z_j} \end{aligned} \tag{BP}$$

  观察: 因为计算 λ_j时，使用的是计算图中子节点处的梯度和λ值，称 (BP)是**向后传递(backward pass)或者向后传播(back propagation)**.

17.3 与链式法则的联系

在本节，将证明一个定理，其解释了为什么由BP能逐步地计算梯度.

定理 17.1. 针对所有 $1 \leqslant j \leqslant n$，有

$$\lambda_j = \frac{\partial f(x)}{\partial z_j},$$

即函数f关于图中第j个节点在x处的偏导数.

Proof. 为了简单，假设计算图有L层，并且仅在连续的两层之间存在边，即，$f = f_L \circ \cdots \circ f_1$. 证明是从输出层开始的 关于层的归纳法.

Base case: $\lambda_n = 1 = \frac{\partial f_n(x)}{\partial z_n} = \frac{\partial z_n}{\partial z_n}$.

Induction: 固定第 p层，并假设论断对后续层$\ell > p$中的节点成立. 那么，针对第p层中的节点z_j，

$$\begin{aligned}\lambda_j &= \sum_{i\in\beta(j)} \lambda_i \frac{\partial f_i(z_{\alpha(i)})}{\partial z_j} \qquad \text{(BP)} \\ &= \sum_{i\in\beta(j)} \frac{\partial f(x)}{\partial z_i}\frac{\partial z_i}{\partial z_j} \qquad \text{(因为}z_{\beta(j)}\text{属于层}p+1\text{由归纳假设和 (BP))} \\ &= \frac{\partial f(x)}{\partial z_j} \qquad \text{(由多元链式法则).}\end{aligned}$$

■

请注意，由逆向计算图上的偏序归纳法进行针对任意计算图的证明.

按语

1. 假设基本的节点运算开销是常数时间，向前和向后传递的开销均是 $O(|\mathcal{V}| + |\mathcal{E}|) \Rightarrow$ 线性时间！
2. 注意到算法本身并没有使用链式法则，仅正确性证明使用了链式法则.
3. 算法等价于控制论中六十年代引入的"**伴随法(method of adjoints)**". 由Baur和Strassen于1983年为了计算偏导数而再次发现[BS83]. 近年来，自从二十世纪九十年代被深度学习群体采用而备受关注.
4. 算法也被称作**自动微分**(automatic differentiation), 注意不要与符号微分和数值微分混淆.

17.4 举例说明

例 1 [两层全连接神经网络] 假设有标签为$y \in \mathbb{R}^m$ 的批数据$X \in \mathbb{R}^{m\times d}$. 考虑权重为$W_1 \in \mathbb{R}^{d\times n}, W_2 \in \mathbb{R}^{n\times 1}$的两层全链接神经网络：

$$f(W_1, W_2) = \|\sigma(XW_1)W_2 - y\|^2$$

为了计算梯度，仅需关于基本运算：

- 范数的平方
- 减法/加法
- 逐分量非线性激活函数σ
- 矩阵乘法

执行向前/向后传递. 观察到前三个运算的偏导数是易于计算的. 因此，聚焦于矩阵乘法即可.

例 2 [针对矩阵填充的BP] 该背景下BP算法的两步是：

向前传递：

- 输入：$A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times d}$
- 输出： $C = AB \in \mathbb{R}^{m\times d}$

向后传递：

- 输入：偏导数$\Lambda \in \mathbb{R}^{m\times d}$ (还有从向前传递得到的A, B, C)
- 输出：
  - $\Lambda_1 \in \mathbb{R}^{m\times n}$ (左输入的偏导数)
  - $\Lambda_2 \in \mathbb{R}^{n\times d}$ (右输入的偏导数)

断言 17.2. $\Lambda_1 = \Lambda B^T, \Lambda_2 = A^T\Lambda$

Proof.

$$f = \sum_{i,j} \lambda_{ij} C_{ij} = \sum_{i,j} (AB)_{ij} = \sum_{i,j} \lambda_{ij} \Sigma_k a_{ik} b_{kj}.$$

如此，由拉格朗日更新规则，

$$(\Lambda_1)_{pq} = \frac{\partial f}{\partial a_{pq}} = \sum_{i,j,k} \lambda_{ij} \frac{\partial a_{ik}}{\partial a_{pq}} b_{kj} = \sum_j \lambda_{pj} b_{qj} = (\Lambda B^T)_{pq}.$$

使用针对关于B的偏导数相同的方法，得到

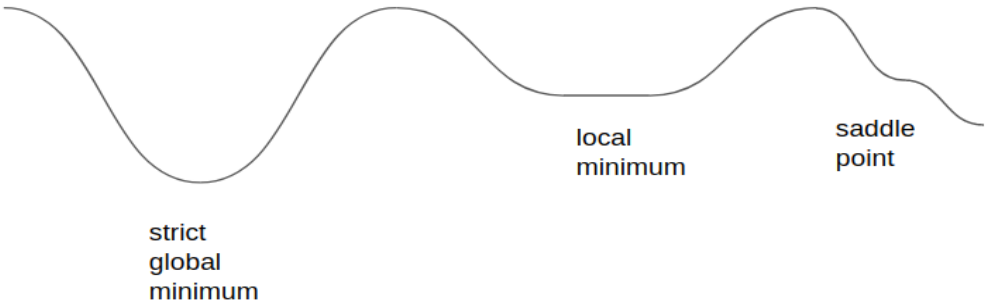
$$(\Lambda_2)_{pq} = (A^T\Lambda)_{pq}$$

■

Part V
非凸优化

18 非凸问题

本讲给出有关非凸问题如何不同于凸问题的重要信息. 对于非凸问题，算法易于陷入数目可能巨大的局部极小点和鞍点. 从而非凸问题的中心课题是很难找到全局极小点.



18.1 局部极小点

先着手讨论局部极小点、及局部极小点的充分和必要条件.

定义 18.1 (局部极小点)**.** 称点x_*是无约束**局部极小点(local minimum)**，如果存在$\delta > 0$使得对于所有满足$\|x - x_*\| < \delta$的x有$f(x_*) \leqslant f(x)$成立.

定义 18.2 (全局极小点)**.** 称点x_*是无约束**全局极小点**，如果对于所有x有$f(x_*) \leqslant f(x)$成立.

以上两个定义，如果这些不等式对于$x \neq x_*$是严格的，分别称作“严格局部极小点”和“严格全局极小点”.

命题 18.3 (局部极小点的必要条件)**.** 设x_*是$f\colon \mathbb{R}^n \to \mathbb{R}$的无约束局部极小点，并且假设$f$在包含$x_*$的开集上是连续可微的($f \in C^1$). 那么

(i) $\nabla f(x_*) = 0$,

(ii) 此外，如果f在包含x_*的开集上是二次连续可微的，那么$\nabla^2 f(x_*) \succeq 0$.

Proof. 固定任意方向$d \in \mathbb{R}^n$. 考虑$\phi(\alpha) := f(x_* + \alpha d)$. 那么

$$\begin{aligned} 0 &\leqslant \lim_{\alpha \to 0} \tfrac{f(x_* + \alpha d) - f(x_*)}{\alpha} \\ &= \tfrac{d\phi(0)}{d\alpha} \\ &= d^\top \nabla f(x_*) \end{aligned} \tag{18.1}$$

不等式(18.1)源于：x_*是局部极小点，所以对充分小的α 有$0 \leqslant f(x_* + \alpha d) - f(x_*)$. 由于$d$是任意的，这蕴含着$\nabla f(x_*) = 0$.

下面使用$\phi(\alpha)$在 0 处的二阶Taylor展式，有

$$\begin{aligned} f(x_* + \alpha d) - f(x_*) &= \alpha \nabla f(x_*)^\top d + \tfrac{\alpha^2}{2} d^\top \nabla^2 f(x_*) d + O(\alpha^2) \\ &= \tfrac{\alpha^2}{2} d^\top \nabla^2 f(x_*) d + O(\alpha^2) \end{aligned}$$

由x_*的最优性，

$$\begin{aligned} 0 &\leqslant \lim_{\alpha \to 0} \tfrac{f(x_* + \alpha d) - f(x_*)}{\alpha^2} \\ &= \lim_{\alpha \to 0} \tfrac{1}{2} d^\top \nabla^2 f(x_*) d + \tfrac{O(\alpha^2)}{\alpha^2} \\ &= \tfrac{1}{2} d^\top \nabla^2 f(x_*) d \end{aligned}$$

因为d是任意的，这蕴含着$\nabla^2 f(x_*)$是半正定的. ■

定义 18.4 (驻点)**.** 称点$x \in \mathbb{R}^n$是$f \colon \mathbb{R}^n \to \mathbb{R}$的**驻点(stationary point)**，如果 $\nabla f(x) = 0$.

称命题 18.3 (i)是最优性的一阶必要条件；命题 18.3 (i)和(ii)是最优性的二阶必要条件. 请注意$\nabla f(x_*) = 0$ 独自并不蕴含着x_*是局部极小点. 甚至必要条件$\nabla f(x_*) = 0$和$\nabla^2 f(x_*) \succeq 0$也不蕴含$x_*$是局部极小点. 这是因为有可能$\nabla^2 f(x_*) = 0$, 但是三阶导数不是 0 . 比如对一维的情况，$x_* = 0$ 针对$f(x) = x^3$满足这些条件，但它不是局部极小点. 现在，将考虑局部极小点的实用充分条件.

命题 18.5 (严格极小点的充分条件)**.** 设$f \colon \mathbb{R}^n \to \mathbb{R}$在开集$S$上是二次连续可微的. 假设$x_* \in S$ 使得$\nabla f(x_*) = 0$并且$\nabla^2 f(x_*) \succ 0$ (正定). 那么， x_* 是f的严格无约束局部极小点.

Proof. 固定$0 \neq d \in \mathbb{R}^n$. 请注意$d^\top \nabla^2 f(x_*) d \geqslant \lambda_{\min} \|d\|^2$, 其中$\lambda_{\min}$是$\nabla^2 f(x_*)$ 的最小特征值. 那么

$$\begin{aligned} f(x_* + d) - f(x_*) &= \nabla f(x_*)^\top d + \tfrac{1}{2} d^\top \nabla^2 f(x_*) d + o(\|d\|^2) && (18.2) \\ &\geqslant \tfrac{\lambda_{\min}}{2} \|d\|^2 + o(\|d\|^2) \\ &= \left(\tfrac{\lambda_{\min}}{2} + \tfrac{o(\|d\|^2)}{\|d\|^2} \right) \|d\|^2 \\ &> 0 && (18.3) \end{aligned}$$

等式(18.2)是由二阶Taylor展式推出来的. 不等式(18.3)是因为对充分小的$\|d\|$和$\lambda_{\min} > 0$. 因此，x_*必须是严格局部极小点. ■

例子 18.6. 考虑函数

$$f(x, y) = \frac{1}{2} x^2 + \frac{1}{4} y^4 - \frac{1}{2} y^2,$$

对应的梯度

$$\nabla f(x, y) = \begin{bmatrix} x \\ y^3 - y \end{bmatrix},$$

Hessian阵

$$\nabla^2 f(x,y) = \begin{bmatrix} 1 & 0 \\ 0 & 3y^2-1 \end{bmatrix}.$$

易见有三个驻点$(0,0),(0,-1),(0,1)$，其中$(0,-1)$ 和$(0,1)$ 满足二阶充分条件，是严格局部极小点.

18.2 最速下降法的全局收敛性

针对非凸问题，必须接受梯度下降法并不总能找到全局极小点，甚至不必是局部极小点的事实. 然而，可以保证收敛到驻点.

定义 18.7 (下降方向)**.** 已知$x \in \mathbb{R}^n$. 如果$d \in \mathbb{R}^n$使得$d^\top \nabla f(x) < 0$，称d是f 在x 处的下降方向.

假设 $x' = x + \eta d, \eta > 0$. 由一阶Taylor展式，得到

$$\begin{aligned} f(x') &= f(x) + \nabla f(x)^\top (x' - x) + o(\|x' - x\|) \\ &= f(x) + \eta d^\top \nabla f(x) + o(\eta \|d\|) \\ &= f(x) + \eta d^\top \nabla f(x) + o(\eta) \end{aligned} \tag{18.4}$$

在等式(18.4)中，因为η的大小是能控制的, 并且$d^\nabla f(x)$对于η而言是常数. 因此，充分小的步长$\eta > 0$ 能使得$f(x') < f(x)$. 接下来讨论用灵活的方法挑选一个步长.

18.2.1 线搜索与Armijo法则

已知下降方向d (比如 $d = -\nabla f(x)$), 设步长

$$\eta_* \in \operatorname*{argmin}_{\eta \geqslant 0} \phi(\eta) := f(x + \eta d).$$

因为沿着方向d搜索最好的步长，称使用这种格式的方法为**精确线搜索(Exact Line Search)**. 仅当f是严格凸二次函数时，精确步长才有解析表达式. 一般的都需要数值方法来求解. 精确线搜索的计算开销与单变量方程求根相似，成本很昂贵.

在这种情况的许多情形中，没必要精确地找到全局极小点. 只要ϕ有“本质”减少就足够了，乘对应的法是非精确线搜索(Inexact Line Search). Armijo 法则给出了“本质”减少之定义(及获取)的标准方式： 设$\phi(\eta)$在$\eta \geqslant 0$上连续可微，并且满足$\phi'(0) < 0$. 设$\rho \in (0,1)$，$\gamma \in (0,1)$是参数(通常选$\rho = 1/100$，$\gamma = 1/2$或者$\gamma = 0.1$).

称步长$\eta > 0$是**合适的**，如果**Armijo条件**

$$\phi(\eta) \leqslant \phi(0) + \rho\eta\phi'(0) \tag{18.5}$$

成立；称η是**几乎最大的**，如果$\frac{1}{\gamma}$倍的步长不再合适：

$$\phi\left(\frac{\eta}{\gamma}\right) > \phi(0) + \rho\frac{\eta}{\gamma}\phi'(0). \tag{18.6}$$

称步长$\eta > 0$通过Armijo法则的测试("本质"减少ϕ)，如果它既是合适的又是几乎最大的. 图 18.1给出了 (18.5)式右端线性函数$\rho-$线的图示. 这里的Armijo条件 (18.5)要

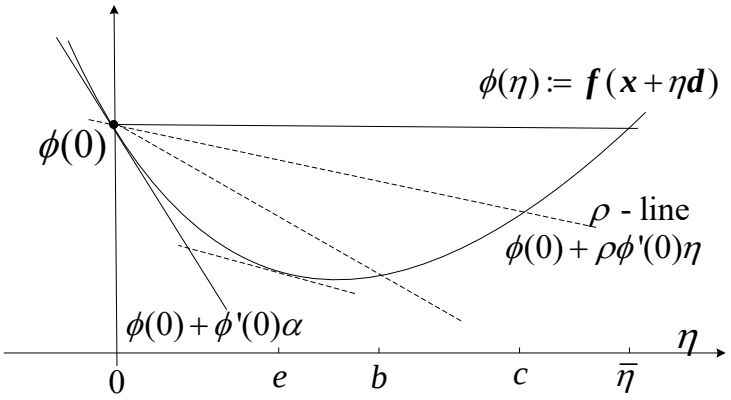


图 18.1: 非精确线搜索示意图

求所选步长使得函数ϕ的图形要在$\rho-$线图形的下方. 步长几乎是最大的条件 (18.6)表明将它扩大$\frac{1}{\gamma}$后，不再满足上述几何事实.

重要**事实**是假设在射线$\eta > 0$上ϕ有下界. 那么通过Armijo法则测试的步长肯定存在，并且能有效地找出来. 称满足 (18.5)和 (18.6)的η是Armijo- 可接受步长.

找Armijo-可接受步长的算法：

开始：选择$\bar{\eta} > 0$，并检查其是否满足 (18.5). 如果满足，转分支 A，否则转分支 B.

分支 A：$\bar{\eta}$满足 (18.5). 依次测试$\gamma\bar{\eta}, \gamma^{-2}\bar{\eta}, \gamma^{-3}\bar{\eta}, \cdots$，直到当前值首次不满足 (18.5)终止，那么前一个值通过Armijo法则的测试.

分支 B：$\bar{\eta}$不满足 (18.5). 依次测试$\gamma^1\bar{\eta}, \gamma^2\bar{\eta}, \gamma^3\bar{\eta}, \cdots$，直到当前值满足 (18.5)时终止，那么这个值通过Armijo法则的测试.

算法验证：显然，如果算法终止，那么结果确实通过Armijo测试，因此需要验证算法能有限步终止.

分支 A 明显是有限的：这里沿着序列$\eta_i = \gamma^{-i}\bar{\eta} \to \infty$检查不等式 (18.5). 当不等式首次满足时终止计算. 由于$\phi'(0) < 0$并且ϕ有下界，那么上述情况一定会发生.

分支 B 明显是有限的：这里沿着序列$\eta_i = \gamma^i\bar{\eta} \to 0+0$ 检查不等式 (18.5)，并且当不等式首次满足时终止计算. 由于$\gamma \in (0,1)$并且$\phi'(0) < 0$，故

$$\phi(\eta) = \phi(0) + \eta[\phi'(0) + \underbrace{R(\eta)}_{\to 0, \eta \to 0+0}]$$

从而不等式 (18.5)对于所有足够小的正值η都是满足的. 因为当i充分大时，η_i一定会变得“足够小”. 因此，分支 B 也是有限的.

将分支B称作**回溯Armijo线搜索**. 具体地，已知 $\gamma, \rho \in (0,1), \bar{\eta} > 0$. 置 $\eta = \gamma^m\bar{\eta}$，其中$m$是使得

$$f(x) - f(x+\gamma^m\bar{\eta}d) \geqslant -\rho\gamma^m\bar{\eta}\nabla f(x)^\top d$$

成立的最小正整数. 将$\bar{\eta}$看作初始学习率. 如果$\bar{\eta}$ 引起充分减小量那么停止，否则仍然乘以γ直到由它引起充分减小量. 这些参数的典型选择是

$$\gamma = \tfrac{1}{2}, \quad \rho = \tfrac{1}{100}, \quad \bar{\eta} = 1\,.$$

命题 18.8 (最速下降法收敛到驻点)**.** 假设f是连续可微的(C^1), 并且设 $\{x_t\}$是由

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

产生的序列，其中η_t满足Armijo法则. 那么，$\{x_t\}$的每个极限点都是驻点.

Proof. 设$\bar{x}$是某极限点. 由连续性知$\{f(x_t)\}$收敛于$f(\bar{x})$，因此

$$f(x_t) - f(x_{t+1}) \to 0. \tag{18.7}$$

由Armijo法则的定义有

$$f(x_t) - f(x_{t+1}) \geqslant \rho \eta_t \|\nabla f(x_t)\|^2 \tag{18.8}$$

和

$$f(x_t) - f\Big(x_t - \tfrac{\eta_t}{\gamma}\nabla f(x_t)\Big) < \tfrac{\rho\eta_t}{\gamma}\|\nabla f(x_t)\|^2. \tag{18.9}$$

用反证法. 假设$\bar{x}$不是f的驻点. 那么

$$\liminf_{t\to\infty} \|\nabla f(x_t)\|^2 > 0.$$

一方面， (18.7)和不等式 (18.8)蕴含着$\eta_t \to 0$. 另一方面，现在设$\tilde{\eta}_t = \frac{\eta_t}{\gamma}$，由 (18.9)可继续进行如下推导

$$\begin{aligned}
&\tfrac{f(x_t) - f(x_t - \tilde{\eta}_t \nabla f(x_t))}{\tilde{\eta}_t} < \rho \|\nabla f(x_t)\|^2 \\
\Rightarrow &\nabla f(x_t - \theta_t \nabla f(x_t))^T \nabla f(x_t) < \rho \|\nabla f(x_t)\|^2, \text{其中} \theta_t \in (0, \tilde{\eta}_t) \qquad (18.10)\\
\Rightarrow &\|\nabla f(x_t)\|^2 \leqslant \rho \|\nabla f(x_t)\|^2 \qquad (18.11)
\end{aligned}$$

不等式(18.10)是使用中值定理(Mean Value Theorem, MVT) 得到的. 取极限，由于$\eta_t \to 0 \Rightarrow \tilde{\eta}_t \to 0$ 得到的不等式(18.11). 这与$0 < \rho < 1$相矛盾. 因此，极限点$\bar{x}$是f的稳定点. ■

因此，如果确定的步长满足Armijo法则，就能保证梯度下降法收敛到驻点.

18.3 鞍点

现在知道梯度下降将收敛到驻点，那么驻点多大程度上不是局部极小点呢？

定义 18.9 (鞍点)**.** 不是局部最优(既不是局部极小点，也不是局部极大点)的驻点是鞍点.

假设f二次连续可微. 如果$\nabla f(x) = 0$， 并且$\nabla^2 f(x)$ 既有正特征值，也有负特征值，那么x就是f 的鞍点.

例子 18.10. 考虑三个一元函数：

$$f_1(x) = x^2,\ f_2(x) = x^3,\ f_3(x) = -x^4.$$

易见$x_* = 0$是它们的驻点，并且分别是f_1的极小点和f_3的极大点，是f_2的鞍点.

18.3.1 鞍点是如何出现的？

在大多数非凸问题中，存在多个局部极小点. 在具有自然对称性的问题中易于看到这一点，比如图**??**的两层全连神经网络对应的训练问题.

请注意隐层单元的任何置换都将保持相同的函数值，因此至少有$h!$个局部极小点. 在非凸问题中，两个不同的局部极小点的凸组合通常不再是局部极小点. 在$\nabla f(x)$可微的情况下，由中值定理知道：在任何两个局部极小点之间必定存在另一个驻点. 因此，在任何两个不同的局部极小点之间，通常至少存在一个鞍点. 所以，大量的局部极小点往往会导致大量的鞍点.

然而，当前的工作表明鞍点通常不是问题.

(i) 从随机初始化出发的梯度下降法不会收敛到严格鞍点. [GHJY15]

(ii) 用加性噪声可以避免鞍点. [LPP+17]

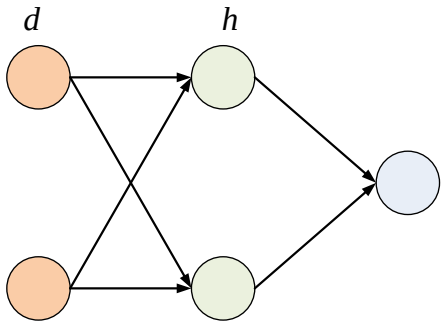


图 18.2: 两层全连接神经网络

19 逃离鞍点

本讲形式化表示并证明针对非凸优化的直观断言：**梯度下降法几乎从来不会收敛到(严格)鞍点.** 该结论的证明见 [LSJR16]. 先给一些定义.

定义 19.1 (驻点)**.** 称x^*是f的驻点，如果梯度在x^*消失，即 $\nabla f(x^*) = 0$.

可以进一步将驻点分成不同类别. 鞍点是一类重要的驻点.

定义 19.2 (鞍点)**.** 点x^*是**鞍点(saddle point)**，如果对所有$\epsilon > 0$, 存在$x, y \in B(x^*; \epsilon)$使得 $f(x) < f(x^*) < f(y)$.

定义 19.3 (严格鞍点)**.** 针对二次连续可微函数 f, 称驻点x^*是**严格鞍点(strict saddle point)**，如果该点处的Hessian阵不是半正定的，即$\lambda_{\min}(\nabla^2 f(x^*)) < 0$, 其中 $\lambda_{\min}$表示最小特征值.

19.1 动力系统视角

将梯度下降法的轨道看作动力系统是有益的. 为此，将每个梯度更新看作一个算子. 针对固定步长 η, 设

$$g(x) = x - \eta \nabla f(x) \tag{GM}$$

因此以前讨论的针对梯度下降法迭代的记号转变为

$$x_t = g^t(x_0) = g(g(\cdots g(x_0))),$$

即对初始点x_0作用t次算子g. 称g是**梯度映射**(gradient map). 请注意x^***是驻点当且仅当它是梯度映射的不动点，即** $g(x^*) = x^*$. 还请注意

$$J_g(x) = I - \eta \nabla^2 f(x) \ (g\text{的雅可比矩阵}), \tag{19.1}$$

该事实后面将变得很重要. 现在正式给出x^*的“吸引子”集合的概念.

定义 19.4. 点x^*的**全局稳定集/吸引子**(global stable set/attractors) 定义为

$$W^S(x^*) = \left\{ x \in \mathbb{R}^n : \lim_{t\to\infty} g^t(x) = x^* \right\}.$$

换句话说，这是用g作用多次，最终会收敛到x^*的点集.

用这个不寻常的定义，可以正式陈述主要断言.

定理 19.5. 假设$f \in C^2$是β-光滑的. 也假设步长 $\eta < 1/\beta$. 那么，对于所有严格鞍点x^*, 它的吸引子 $W^S(x^*)$的Lebesgue测度为 0.

按语 19.6. 事实上，用额外的技术能证明 $\bigcup_{\text{strict saddle points } x^*} W^S(x^*)$ 的*Lebesgue*测度也是 *0* . 这恰好是另一种呈现梯度下降法几乎处收敛到局部极小点的方式.

按语 19.7. 由定义，定理 *19.5*中的结论关于*Lebesgue*测度连续的任何概率测度也成立*(*比如任何连续概率分布*)*，即

$$\mathbb{P}\left(\lim_{t\to\infty} x_t = x^* \right) = 0.$$

然而，上面的定理仅是一种渐近方式的陈述. 非渐近地，甚至用相当自然的随机初始化策略和非病态函数，鞍点会使梯度下降法的收敛速度显著变慢. 最近的结论[DJL+17]表明梯度下降法需要花费指数时间来逃离鞍点(尽管上面的定理说他们最终能逃离). 该讲不证明这个结论.

19.2 二次情况

在证明定理 19.5之前，先看两个例子，这样会使得证明更直观.

例子 19.8. 设$f(x) = \frac{1}{2}x^T H x$, 其中$H$是$n \times n$对称的非半正定矩阵. 为了方便，假设 0 不是$H$的特征值. 因此0是该问题唯一的驻点和唯一的严格鞍点.

计算得

$$g(x) = x - \eta H x = (I - \eta H)x, \ g^t(x) = (I - \eta H)^t x.$$

意识到

$$\lambda_i(I-\eta H)=1-\eta\lambda_i,$$

其中$\lambda_1,\cdots,\lambda_n$表示$H$的$n$个特征值. 因此，设$x$是$H$的与$\lambda_i$对应的特征向量. 为了

$$\lim_t g^t(x)=\lim_t(1-\eta\lambda_i)^t x=0=:x_*$$

恰好需要

$$\lim_t(1-\eta\lambda_i)^t=0,$$

即$|1-\eta\lambda_i|<1$. 这蕴含着

$$W^S(0)=\operatorname{span}\left\{u:Hu=\lambda u,0<\lambda<\tfrac{\eta}{2}\right\}$$

即小于$\frac{\eta}{2}$的正特征值的特征向量组成的集合. 因为η 能任意大， 刚好考虑正特征值的特征向量这个更大的集合. 由关于H的假设，这个集合的维数小于n, 因此测度是 0 .

例子 19.9 (例 18.6续)**.** 对于函数

$$f(x,y)=\tfrac{1}{2}x^2+\tfrac{1}{4}y^4-\tfrac{1}{2}y^2,$$

对应的梯度映射

$$g(x,y)=\begin{bmatrix}(1-\eta)x\\(1+\eta)y-\eta y^3\end{bmatrix}.$$

易见$(0,0)$是唯一的严格驻点. 注意这里f不是二次函数，从而分析时，需要用$\nabla^2 f(x_*)$代替上个例子中的H. 这里$\dim W^S(0)=1$，和上一个例子类似，$W^S(0)$ 是低维子空间.

19.3 一般情况

在本讲结束之前，给出主要定理的证明.

定理 19.5的证明. 首先定义x^*的**局部稳定集/吸引子**(local stable set/attractors) 为

$$W^S_\epsilon(x^*)=\{x\in B(x^*;\epsilon):g^t(x)\in B(x^*;\epsilon)\ \forall t\}.$$

直观上，这描述了$B(x^*;\epsilon)$的一个子集，其中的元素在任意多次梯度映射的作用下，仍然停留在$B(x^*;\epsilon)$内. 局部稳定集取代了具有正测度的$B(x^*;\epsilon)$，从而建立了对梯度下降法的收敛性至关重要的局部概念.

现在陈述一个简化版的不加证明的**稳定流形定理**：针对微分同胚 $g:\mathbb{R}^n\to\mathbb{R}^n$, 如果 x^*是g的不动点，那么对于所有充分小的ϵ，$W^S_\epsilon(x^*)$是个子流形，其维数等于$J_g(x^*)$的不超过 1 的特征值的几何重数之和. 微分同胚，粗略地讲，是一个可微同构. 事实上，因为对于g假设了可微性，将聚焦于同构.

设x^*是严格鞍点. 一旦证得g是可微映射(使用假设$\eta<1/\beta$)这个事实，由于x^*是g的不动点，就能应用稳定流形定理. 因为$\nabla^2 f(x^*)$至少有一个负特征值，因此由式 (19.1)知，梯度映射在x_*的Jacobi矩阵$J_g(x^*)$ 必有大于 1 的特征值，因此$W^S_\epsilon(x^*)$ 的维数小于 n，从而$W^S_\epsilon(x^*)$ 的测度是 0.

如果$g^t(x)$收敛到x^*, 必存在足够大的T使得

$$g^T(x) \in W_\epsilon^S(x^*).$$

因此

$$W^S(x^*) \subseteq \bigcup_{t\geqslant 0} g^{-t}(W_\epsilon^S(x^*)).$$

对于每个t, g^t是同构的复合，从而也是同构；g^{-t}也是同构. 因此$g^{-t}(W_\epsilon^S(x^*))$的维数和$W_\epsilon^S(x^*)$ 的相同，从而$g^{-t}(W_\epsilon^S(x^*))$的测度也是 0. 所以可数个这种集合的并集合的测度也是 0. 这样它的子集$W^S(x^*)$ 的测度最终也是 0 ，从而得到想要的结论.

由于假设g是光滑的，最后证明g是双射即得到同构. 首先它是单射. 假设$g(x) = g(y)$, 那么由g的定义和f的光滑性，

$$\|x - y\| = \eta\|\nabla f(x) - \nabla f(y)\| \leqslant \eta\beta\|x - y\|.$$

因为$\eta\beta < 1$, 必有$\|x - y\| = 0$. 为了证明g是满的，$\forall y$，构造反函数

$$h(y) = \operatorname*{argmin}_x \tfrac{1}{2}\|x - y\|^2 - \eta f(x) \tag{19.2}$$

亦称临近更新. 对于 $\eta < 1/\beta$, 因为$f \in C^2$和f是$\beta-$光滑的知，问题 (19.2)中的目标函数关于x是$(1 - \eta\beta)-$强凸的. 因此驻点条件是该优化问题最优解的充分和必要条件. 从而有

$$y = h(y) - \eta\nabla f(h(y)) = g(h(y)).$$

证毕. ■

20 交替极小化和期望极大化(EM)

本讲是一系列代码示例，参见这里：

Lecture 19
(在你的浏览器中打开)

21 无导数优化、策略梯度和控制

本讲是一系列代码示例，参见这里：

Lecture 20
(在你的浏览器中打开)

22 非凸目标函数与凸松驰

凸极小化指在凸集上极小化凸函数. 这讲开始研究非凸优化. 分析一般的“非凸性”影响很困难，因为它可以指任何非凸问题，这是非常广泛的一类问题. 所以取而代之，将关注求解带稀疏约束的最小二乘：

$$\min_{\|x\|_0 \leqslant s} \|Ax - y\|_2^2, \tag{22.1}$$

其中$y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times d}$ 是已知的, $x \in \mathbb{R}^d$. 将证明，尽管求解该问题的一般形式很困难，但是针对有限的一类问题，其存在有效凸松驰.

可用带稀疏约束的最小二乘问题 (22.1)求解压缩感知和稀疏线性回归问题，它们在各个领域都很重要. 在压缩感知中，A 是测量模型，y 是某个稀疏信号x 的测量. 压缩感知被用来减少所需测量的数目，比如说，一个MRI，因为保留包含关于x 的稀疏约束，能够从更少的测量y恢复出信号x.

在稀疏线性回归中，A是数据矩阵，y是某个结果变量. 稀疏线性回归的目的是在稀疏的特征集合上恢复权重x 以解释结果变量. 在遗传学上，A 可以是病人的基因，y 是他们是否患有某种特定疾病. 那么目标就是在预测是否有疾病的稀疏基因集合上恢复权重x.

当线性方程组中不存在噪声时，问题可简化成ℓ_0极小化问题：

$$\begin{array}{ll} \text{minimize} & \|x\|_0 \\ \text{subject to} & Ax = y \end{array} \tag{22.2}$$

22.1 难度

甚至这个简化版的问题也是NP-难的，由于将证明精确3-覆盖是ℓ_0-极小化问题 (22.2) 的特例，而精确3-覆盖是NP- 完全的. 这里的证明源自[FR13].

定义 22.1. 3-集合精确覆盖(exact cover by 3-sets)问题：已知$[m]$的3- 元素子集$\{T_i\}$，记$d = |\{T_i\}|$. 那么是否存在集合$z \subseteq [d]$ 满足 $\cup_{j\in z} T_j = [m]$ 和 $T_i \cap T_j = \emptyset, i, j \in z, i \neq j$? 称满足该条件的集合$z$ 是$[m]$ 的**精确覆盖(exact cover)**.

定义 22.2. 向量$x \in \mathbb{R}^d$的**支集**定义为 $\operatorname{supp}(x) = \{i \in [d] \mid x_i \neq 0\}$.

定理 22.3. 针对一般的(A, y)的ℓ_0-极小化问题 (22.2) 是*NP*-难的.

Proof. 已知$[m]$的3-元素子集$\{T_1, \cdots, T_d\}$. 定义$m \times d$的矩阵A为

$$A_{ij} = \begin{cases} 1, & \text{如果} i \in T_j \\ 0, & \text{否则,} \end{cases}$$

y是全 1 向量. 请注意，由构造，A的每列有 3 个非零元素，所以有$\|Ax\|_0 \leqslant 3\|x\|_0$. 如果$x$满足$Ax = y$, 这样有$\|x\|_0 \geqslant \frac{\|y\|_0}{3} = \frac{m}{3}$. 现在考虑关于$(A, y)$ 的ℓ_0- 极小化问题，并设对应的最优解是$\hat{x}$. 存在两种情况：

1. 如果$\|\hat{x}\|_0 = \frac{m}{3}$, 那么$z = \operatorname{supp}(\hat{x})$ 是精确3-覆盖.
2. 如果$\|\hat{x}\|_0 > \frac{m}{3}$, 那么不存在精确 3- 覆盖. 否则，若存在，它将满足$\|\hat{x}\|_0 = \frac{m}{3}$，并因此与$\hat{x}$ 的最优性相矛盾.

这样，由于通过ℓ_0-极小化问题 (22.2)能求解精确 3-覆盖，所以ℓ_0- 极小化问题 (22.2) 必是NP- 难的. ■

22.2 凸松弛

尽管ℓ_0-极小化问题 (22.2)一般来说是NP- 难的，将证明对有限类的A, 能将ℓ_0-极小化问题松弛成ℓ_1- 极小化. 首先，定义支集为S 的近似稀疏向量集合是由它的ℓ_1质量被S控制的那些向量组成. 正式地，

定义 22.4. 支集为$S \subset [d]$的近似稀疏向量是

$$C(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{\bar{S}}\|_1 \leqslant \|\Delta_S\|_1\},$$

其中 $\bar{S} = [d]/S$，并且 Δ_S是Δ在S上的限制，即

$$(\Delta_S)_i = \begin{cases} \Delta_i & \text{如果} i \in S \\ 0 & \text{否则.} \end{cases}$$

回忆A的零空间是集合$\text{null}A = \{\Delta \in \mathbb{R}^d \mid A\Delta = 0\}$. 零空间在估计问题中是“坏”向量的集合. 考虑解$Ax = y$. 如果$\Delta \in$ null A, 由于

$$A(x + \Delta) = Ax + A\Delta = Ax = b,$$

从而$x + \Delta$也是一个解. 这样，专注于零空间与所关心的稀疏向量集上仅包含零的那些矩阵.

定义 22.5. 矩阵A关于支集S满足受限零空间性质(restricted nullspace property, RNP)，如果$C(S) \cup \text{null}A = \{0\}$.

举个满足RNP的矩阵A和集合S的例子. 用这些定义，现在能陈述主要定理.

定理 22.6. 已知 $A \in \mathbb{R}^{m\times d}$和$y \in \mathbb{R}^m$，考虑 ℓ_0- 极小化问题 (22.2)的解x^*. 假设x^*的支集是S并且矩阵A关于S 满足受限零空间性质. 已知ℓ_1-极小化问题的解

$$\begin{aligned} \widehat{x} = \arg \quad & \text{minimize } \|x\|_1 \\ & \text{subject to } Ax = y \end{aligned} \tag{22.3}$$

那么有$\widehat{x} = x^*$.

Proof. 注意到，由定义x^*和$\widehat{x}$都满足约束条件$Ax = y$. 设$\Delta = \widehat{x} - x^*$是差向量，有

$$A\Delta = A\widehat{x} - Ax^* = 0,$$

这蕴含着$\Delta \in \text{null}A$.

现在的目的是证明$\Delta \in C(S)$，那么由受限零空间性质将有$\Delta = 0$. 首先，由于$\widehat{x}$是ℓ_1优化的最优解，从而有

$$\|\widehat{x}\|_1 \leqslant \|x^*\|_1.$$

那么有

$$\begin{aligned} \|x_S^*\|_1 = \|x^*\|_1 &\geqslant \|\widehat{x}\|_1 \\ &= \|x^* + \Delta\|_1 \\ &= \|x_S^* + \Delta_S\|_1 + \|x_{\bar{S}}^* + \Delta_{\bar{S}}\|_1 && \text{通过拆分}\ell_1 \text{ 范数,} \\ &= \|x_S^* + \Delta_S\|_1 + \|\Delta_{\bar{S}}\|_1 && \text{由}\|x^*\|_1\text{支集的假设,} \\ &\geqslant \|x_S^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{\bar{S}}\|_1. \end{aligned}$$

因此$\|\Delta_S\|_1 \geqslant \|\Delta_{\bar{S}}\|_1$, 这蕴含着 $\Delta \in C(S)$. ■

到此还算顺利. 已经看到ℓ_1-松弛对某些矩阵是行得通的. 一个自然的问题是哪些矩阵满足受限零空间性质. 为了得到关于此问题的切入点，将研究矩阵的另一个好性质，所谓的**受限等距性质(restricted isometry property, RIP)**. 后面，将看到特定矩阵全体以很高的概率满足RIP.

定义 22.7. 矩阵A享有(s,δ)-RIP，如果对所有s-稀疏向量x ($\|x\|_0 \leqslant s$), 有

$$(1-\delta)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1+\delta)\|x\|_2^2.$$

直观上，A像等距算子(真正的等距算子满足$\delta = 0$)那样作用. RIP 很有用，由于它蕴含着：不能将两个s- 稀疏向量之差映射到 0 ，并且也蕴含着RNP. 通过考虑 A的奇异值，得到如下引理.

引理 22.8. 如果A具有(s,δ)-RIP, 那么对于所有基数为s的子集S，有

$$\|A_S^\top A_S - I_S\|_2 \leqslant \delta.$$

其中

$$(A_S)_{ij} = \begin{cases} A_{ij} & \text{如果 } i,j \in S \\ 0 & \text{否则.} \end{cases}$$

现在证明RIP蕴含着受限零空间性质.

定理 22.9. 如果矩阵A享有$(2s,\delta)$-RIP, 那么对于所有基数满足$|S| \leqslant s$ 的子集S，矩阵A均享有RNP.

Proof. 设$x \in \text{null}A$是任意的非零向量. 那么必须针对任意满足$|S| \leqslant s$ 的集合S证明

$$x \notin C(S).$$

特别地，设S_0是向量x的前s大元素的指标集. 证明

$$\|x_{S_0}\|_1 < \|x_{\overline{S}_0}\|_1$$

是充分的，因为如果上式成立的话，对任何别的满足$|S| \leqslant s$的子集S上式也成立.

将$\overline{S}_0$进行剖分：

$$\overline{S}_0 = \bigcup_{j=1}^{\lceil \frac{d}{s} \rceil - 1} S_j$$

其中

- S_1 是$\overline{S}_0$中与前s大元素对应的指标子集
- S_2 是$\overline{S}_0 \setminus S_1$中与前$s$大元素对应的指标子集
- S_3是$\overline{S}_0 \setminus S_1 \setminus S_2$ 中与前s 大元素对应的指标子集
- 以此类推$\cdots$

因此有$x = x_{S_0} + \sum_j x_{S_j}$. 已经将$x$分解成大小为$s$的块. 也称这是剥壳(shelling). 由RIP, 有

$$\|x_{S_0}\|_2^2 \leqslant \tfrac{1}{1-\delta}\|Ax_{S_0}\|_2^2.$$

由假设$x \in \mathrm{null}A$，有

$$A(x_{S_0} + \sum_{j\geqslant 1} x_{S_j}) = 0 \implies Ax_{S_0} = -\sum_{j\geqslant 1} Ax_{S_j}.$$

因此

$$\begin{aligned}
\|x_{S_0}\|_2^2 &\leqslant \tfrac{1}{1-\delta}\|Ax_{S_0}\|_2^2 \\
&= \tfrac{1}{1-\delta}\langle Ax_{S_0}, Ax_{S_0}\rangle \\
&= \tfrac{1}{1-\delta}\sum_{j\geqslant 1}\langle Ax_{S_0}, -Ax_{S_j}\rangle \\
&= \tfrac{1}{1-\delta}\sum_{j\geqslant 1}[\langle Ax_{S_0}, -Ax_{S_j}\rangle + \langle x_{S_0}, x_{S_j}\rangle] && \text{由于}\langle x_{S_0}, x_{S_j}\rangle = 0 \\
&= \tfrac{1}{1-\delta}\sum_{j\geqslant 1}\langle x_{S_0}, (I - A^\top A)x_{S_j}\rangle \\
&\leqslant \tfrac{\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_0}\|_2\|x_{S_j}\|_2 && \text{由引理 22.8.}
\end{aligned}$$

如此，有

$$\|x_{S_0}\|_2 \leqslant \tfrac{\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_j}\|_2. \tag{22.4}$$

由构造方式知，针对每个$j \geqslant 1$, 有

$$\|x_{S_j}\|_\infty \leqslant \tfrac{1}{s}\|x_{S_{j-1}}\|_1$$

并且因此

$$\|x_{S_j}\|_2 \leqslant \tfrac{1}{\sqrt{s}}\|x_{S_{j-1}}\|_1.$$

代入 (22.4), 得到

$$\begin{aligned}
\|x_{S_0}\|_1 &\leqslant \sqrt{s}\|x_{S_0}\|_2 \\
&\leqslant \tfrac{\sqrt{s}\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_j}\|_2 \\
&\leqslant \tfrac{\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_{j-1}}\|_1 \\
&= \tfrac{\delta}{1-\delta}(\|x_{s_0}\|_1 + \sum_{j>1}\|x_{S_{j-1}}\|_1)
\end{aligned}$$

这等价于

$$\|x_{s_0}\|_1 \leqslant \tfrac{\delta}{1-\delta}(\|x_{s_0}\|_1 + \|x_{\bar{s}_0}\|_1).$$

简单的代数演算表明：只要$\delta < \frac{1}{3}$，就有$\|x_{s_0}\|_1 < \|x_{\bar{s}_0}\|_1$. ■

现在证明了如果矩阵享有RIP，那么ℓ_1-松弛可以工作，接下来看一些天然存在的，并且具有这种性质的矩阵的例子.

定理 22.10. 设$A \in \mathbb{R}^{m\times d}$定义为$a_{ij} \sim N(0,1)$，并且各元素是独立同分布的. 那么对于不小于$\mathcal{O}\left(\frac{1}{\delta^2} s \log \frac{d}{s}\right)$ 的m，矩阵$\frac{1}{\sqrt{n}} A$ 具有(s,δ)-*RIP.*

对于次高斯分布而言，同样的结论也成立. 对于更加结构化的矩阵也有类似的结论，诸如子采样Fourier矩阵.

定理 22.11. 设$A \in \mathbb{R}^{m\times d}$是子采样Fourier矩阵. 那么针对不小于$\mathcal{O}\left(\frac{1}{\delta^2} s \log^2 s \log d\right)$的$m$，矩阵$A$ 具有(s,δ)-RIP.

该结论源于[HR15]使用[RV07, Bou14, CT06] 的工作. $\mathcal{O}\left(\frac{1}{\delta^2} s \log d\right)$是一个公开猜想.

有许多关于凸松弛的工作. 仅针对稀疏性，已经研究了许多变形，比如

- **基追踪消噪(Basic pursuit denoising, BPDN)**：

$$\min \|x\|_1 \ \text{ subject to } \ \|Ax - y\|_2 \leqslant \epsilon \tag{BPDN}$$

  其中$\epsilon > 0$是参数.

- **约束型LASSO**：

$$\min \|Ax - y\|_2^2 \ \text{ subject to } \ \|x\|_1 \leqslant t \tag{LASSO}$$

  其中$t > 0$是参数.

- **拉格朗日型/惩罚型LASSO**：

$$\min \|Ax - y\|_2^2 + \lambda \|x\|_1 \tag{22.5}$$

  其中$\lambda > 0$是参数.

也有针对其它非凸目标的凸松弛. 已知线性算子A和观测矩阵Y，比如

$$\min \ \operatorname{rank}(X) \quad \text{subject to} \quad A(X) = Y$$

很困难， 一个简单问题是求解核范数极小化：

$$\min \ \|X\|_* \quad \text{subject to} \quad A(X) = Y,$$

其中$\|X\|_* = \sum_i \sigma_i(X)$是矩阵$X$的奇异值之和. 这个非凸模型和凸松弛经常出现在针对图像的低秩估计或者矩阵补全问题中(参见5.3 节).

23 非凸约束与投影梯度下降法

上一讲提到在凸集上极小化非凸函数. 该类问题的典型实例是 (22.2)，称作l_0- 极小化，即在仿射约束$Ax = y$ 下极小化$\|x\|_0$. 当测量存在噪声时， 得到另一种类型的非凸优化，即在非凸集上极小化凸函数. 这类非凸优化问题的典型实例是 (22.1).

求解非凸优化问题 (22.2)的一种选项是将ℓ_0-目标松弛成凸的ℓ_1- 目标，得到 ℓ_1-极小化 (22.3). 请注意在关于(A, y)合适的假设下，观察到ℓ_1-极小化仍然给出正确答案(诸如RIP假设和RNP假设). 在当前设置下，如果将此思想用来求解 (22.1)， 有

非凸约束$\to$凸松弛$\to$ PGD (投影梯度下降法),

或者更直接的，应用PGD直接求解非凸约束优化问题.

讨论一些从jupyter notebook中拿来的内容. 尽管能有效求解凸松弛(比如，使用内点法), 但扩展到大规模实例仍存在问题. 因此考虑诸如投影梯度下降法的一阶方法来加速计算是有意义的. 发现直接投影到非凸集也是可以工作的. 并且当运行PGD时，很自然的问题是：是否需要首先进行凸松弛，还是仅直接在非凸集上运行PGD.

考虑形如$y = Ax$的具有s-稀疏x的测试问题，其中A 是$m \times d$矩阵，并且A的元素是独立同分布的高斯分布的样本，因此如果取样本容量m充分大的话，矩阵满足RIP. 可以检查为了ℓ_1-松弛能工作需要的样本容量的大小. 结果是独立同分布高斯矩阵需要$O(s \log d/s)$行以满足RIP，因此这对应于$m = O(s \log d/s)$个样本. 使用内点法的运行时间稍微有些慢：

- $d = 100, m = 50$: 10 毫秒,
- $d = 1000, m = 500$: 5-6 秒,
- $d = 4000, m = 2000$: 112 秒.

因此，为了求解规模非常大的实例，也应该考虑一阶方法. 可以直接对非凸稀疏向量集运行PGD，也称作**迭代硬阈值(Iterative Hard Thresholding, IHT)**，由于投影步(找到最近s-稀疏向量)对应于硬阈值向量(保持前 s 大元素不变，其余元素置 0). 对于上面的第三个实例，它仅需0.0357秒，比内点法快1000倍.

现在，讨论IHT，也被称作针对稀疏向量的投影梯度下降法(projected gradient descent, PGD)，它是针对如下问题的投影梯度下降法. 已知设置：

$$y = Ax + e$$

其中$y \in \mathbb{R}^m, x \in \mathbb{R}^d, A \in \mathbb{R}^{m \times d}$, e 是观测噪声，目标是：已知y 和 A，当$m \ll d$并且x近似地是s-稀疏的，估计x.

IHT 算法使用迭代

$$x^{i+1} = \Pi_s(x^i + A^\top(y - Ax^i))$$

其中Π_s是保持一个向量前s大元素的硬阈值算子. 因为这里需要使用脚标修饰迭代点，所以均使用上标作为迭代指标. 以下同理.

考虑目标函数$f(x) = \frac{1}{2}\|Ax - y\|_2^2$. 它的梯度 $\nabla f(x) = A^\top(Ax - y)$. 如果可以直接在非凸集上优化，那么不需要凸松弛. 想要证明IHT算法输出一个解. 算法定义如下：

请注意Π_s是在s-稀疏向量集合上的投影. 在本讲的剩余部分，研究这个投影如何工作和该方法到底有多快.

Algorithm 5 Iterative Hard Thresholding (ITH) Algorithm

Require: Parameters y, A, t, s.
1: **for** $i = 1$ to t **do**
2: $\quad \tilde{x}^{i+1} \leftarrow x^i - A^\top(Ax^i - y)$.
3: $\quad x^{i+1} \leftarrow \Pi_s(\tilde{x}^{i+1})$.
4: **end for**
Ensure: $\widehat{x} \leftarrow x^{t+1}$.

研究矩阵的一种优良性质：**受限等距性(Restricted Isometry Property, RIP)**. 因为该性质蕴含着不能将两个$s-$ 稀疏向量之差映射成 0 ，并且RIP 也蕴含着受限零空间性质，所以该性质很有用. 这里，受限零空间性质允许在非凸集上优化.

定义 23.1. 称矩阵A满足(s,δ)**-RIP**，如果对所有的s-稀疏向量，有：

$$(1-\delta)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1+\delta)\|x\|_2^2.$$

后果：对于大小为s的支集S，有

$$\|I_s - A_S^\top A_S\|_2 \leqslant \delta,$$

其中A_S是A在S上的限制(详细定义参见引理 22.8).

定理 23.2. 考虑如下设置：

$$y = Ax_* + e,$$

其中x_* 是支集为S的s-稀疏向量，A 享有$(3s, \frac{1}{4})$-RIP， e 是任意噪声. 那么对IHT，以下结论成立：

$$\|x^{i+1} - x_*\|_2 \leqslant \tfrac{1}{2}\|x^i - x_*\|_2 + 2\max_{|S|\leqslant 3s}\|A_S^\top e\|_2.$$

Proof. 设 S_i是(x^i)的支集，令$S' = S_{i+1} \cup S_i \cup S$ (因此 $|S'| \leqslant 3s$).
那么

$$\begin{aligned}
\|x^{i+1} - x_*\|_2 &\leqslant \|x^{i+1} - \tilde{x}^{i+1}_{S'}\|_2 + \|\tilde{x}^{i+1}_{S'} - x_*\|_2 && \text{(由三角不等式)}\\
&\leqslant 2\|\tilde{x}^{i+1}_{S'} - x_*\|_2 \\
&= 2\|x^i_{S'} - A^\top_{S'}(Ax^i - y) - x_*\|_2 && (\tilde{x}^{i+1}_{S'}\text{的定义})\\
&= 2\|x^i - A^\top_{S'}(A_{S'}x^i - A_{S'}x_* - e) - x_*\|_2 && \text{(代入}y\text{，并由}S'\text{的定义)}\\
&= 2\|x^i - x_* - A^\top_{S'}A_{S'}(x^i - x_*) + A^\top_{S'}e\|_2 \\
&\leqslant 2\|[I - A^\top_{S'}A_{S'}](x^i - x_*)\|_2 + 2\|A^\top_{S'}e\|_2 \\
&\leqslant 2\delta\|(x^i - x_*)\|_2 + 2\max_{|S|\leqslant 3s}\|A_S^\top e\|_2. && \text{(RIP)}
\end{aligned}$$

因为x^{i+1}是$\tilde{x}^{i+1}$的s-稀疏投影，这也蕴含着x^{i+1} 是$\tilde{x}^{i+1}_{S'}$的最好s- 稀疏逼近，因此有：

$$\|x^{i+1} - \tilde{x}^{i+1}_{S'}\| \leqslant \|x_* - \tilde{x}^{i+1}_{S'}\|_2.$$

由此得到上面第二个不等式. 将$\delta = \frac{1}{4}$带入上面的不等式，即得待证结论. ■

在上面的定理中，已经证明每次迭代使得误差以因子1/2减小(直到噪声阈值). 因此，像如下推论中那样，得到线性收敛速率是相当直接的：

推论 23.3. 设IHT的输出是$\widehat{x}$. 那么有

$$\|\widehat{x} - x_*\|_2 \leqslant \frac{1}{2^t}\|x_*\|_2 + \sqrt{5}\|e\|_2$$

成立. 因此$t = \log \frac{\|x_*\|_2}{\epsilon}$次迭代后有

$$\|\widehat{x} - x_*\|_2 \leqslant \epsilon + \sqrt{5}\|e\|_2.$$

由此得到PGD具有线性速率，然而分析看起来有些不同(没有凸性/ 光滑性)，并且也不需要步长.

定义 23.4. 称f是$L-$光滑的，如果存在常数L使得对于所有$x, \Delta \in \mathbb{R}^d$，有

$$f(x+\Delta) \leqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{L}{2}\|\Delta\|_2^2$$

成立.

现在考虑函数 $f = \frac{1}{2}\|Ax - y\|_2^2$, 它的梯度 $\nabla f(x) = A^\top(Ax - y)$. 该函数的光滑性意味着什么? 针对该函数利用上述定义，得

$$\frac{1}{2}\|A(x+\Delta) - y\|_2^2 \leqslant \frac{1}{2}\|Ax - y\|_2^2 + \Delta^\top A^\top (Ax - y) + \frac{L}{2}\|\Delta\|_2^2$$

将上式展开，整理后得

$$\frac{1}{2}\Delta^\top A^\top A\Delta \leqslant \frac{L}{2}\|\Delta\|_2^2,$$

因此$L-$光滑性等价于 $\|A\Delta\|_2^2 \leqslant L\|\Delta\|_2^2$. 对于强凸性有类似结论.

定义 23.5. 称f是$\ell-$强凸的，如果对所有$x, \Delta \in \mathbb{R}^d$，有

$$f(x+\Delta) \geqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{\ell}{2}\|\Delta\|_2^2$$

成立.

易于得到强凸性等价于$\|A\Delta\|_2^2 \geqslant \ell\|\Delta\|_2^2$.

请注意上面的不等式将光滑性和强凸性与矩阵的RIP联系起来了，只是将条件“对于所有Δ” (针对光滑性和强凸性)替换成“对于所有s- 稀疏的Δ” (针对RIP).

定义 23.6. 称f是**受限$\ell-$强凸(Restricted Strongly Convex, RSC)** 的，如果对于所有s- 稀疏的Δ，有

$$f(x+\Delta) \geqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{\ell}{2}\|\Delta\|_2^2$$

成立.

类似地，可以定义**受限$L-$光滑(Restricted $L-$ smooth, RM)**. 可以说

$$\text{RIP} = \text{RSC} + \text{RS}.$$

因此，如果矩阵A满足(s,δ)-RIP，取$L=1+\delta, \ell=1-\delta$，上面的讨论等同于

$$L/\ell = \frac{1+\delta}{1-\delta} \approx 1,$$

即函数f具有非常好的条件数L/ℓ (因此常数步长$\approx 1/L$). 有大量的工作在削弱该假设，并进一步限制集合.

凸松弛牵扯最优时，主要与条件数有关. PGD对于任何条件数都能工作，但是具有较差的统计速率. 能将凸松弛与非凸PGD匹配起来吗? 答案是肯定的!

已知稀疏性条件，在$O(d)$时间内做硬阈值是可能的. 已知低秩条件，针对$d_1 \times d_2$ 矩阵，在$O(d_1 d_2 \min\{d_1, d_2\})$时间内能计算SVD 并求得最大奇异值.

Part VI
高阶和内点法

24 牛顿法

到目前为止，仅考虑了优化函数的一阶方法. 现在，将利用二阶信息以期获得更快的收敛速率.

一如既往，目标是极小化函数$f\colon \mathbb{R}^n \to \mathbb{R}$. 牛顿法的基本思想是置梯度的一阶Taylor展式为零：

$$F(x) := \nabla f(x) = 0.$$

由这得到一种更新步，它将(在某些条件下)导致比梯度下降法显著地更快的收敛速率.

为了说明这一点，考虑单变量函数$\varphi\colon \mathbb{R} \to \mathbb{R}$. 目的是求解非线性方程

$$\varphi(x) = 0.$$

由Taylor定理，得到$\varphi(x)$的一阶展开为

$$\varphi(x) = \varphi(x_0) + \varphi'(x_0) \cdot (x - x_0) + o(|x - x_0|)$$

记$\delta = x - x_0$，等价地有

$$\varphi(x_0 + \delta) = \varphi(x_0) + \varphi'(x_0) \cdot \delta + o(|\delta|)$$

忽略$o(|\delta|)$项，求解关于δ的线性方程：

$$\varphi(x_0) + \varphi'(x_0)\delta = 0,$$

得到

$$\delta = -\frac{\varphi(x_0)}{\varphi'(x_0)},$$

由此得到迭代

$$x_{t+1} = x_t - \frac{\varphi(x_t)}{\varphi'(x_t)}.$$

对多元函数向量值函数$F : \mathbb{R}^n \to \mathbb{R}^n$可做类似讨论. 目标是求解方程组

$$F(x) = 0.$$

再一次，由Taylor定理，有

$$F(x + \Delta) = F(x) + J_F(x)\Delta + o(\|\Delta\|)$$

其中J_F是F的雅可比矩阵. 如果$J_F(x)$可逆，这时

$$\Delta = -J_F^{-1}(x)F(x),$$

迭代为

$$x_{t+1} = x_t - J_F^{-1}(x_t)F(x_t).$$

已知$f\colon \mathbb{R}^n \to \mathbb{R}$, 最优化中的牛顿法对 $F(x) = \nabla f(x) = 0$. 应用此种更新. 如果$\nabla^2 f(x_t)$正定，这时更新规则是

$$x_{t+1} = x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t).$$

牛顿步极小化f在x_t的二阶Taylor近似

$$f(x) \approx f(x_t) + \nabla f(x_t)^\top (x - x_t) + \tfrac{1}{2}(x - x_t)^\top \nabla^2 f(x_t)(x - x_t).$$

现在，将证明当初始点在局部极小点的充分小邻域内时，牛顿法收敛到该局部极小点.

定理 24.1. 已知$f : \mathbb{R}^n \to \mathbb{R}$，假设 $\nabla^2 f(x)$是β-Lipschitz连续的：

$$\|\nabla^2 f(x) - \nabla^2 f(x')\| \leqslant \beta \|x - x'\|.$$

设x_*是f的满足二阶充分条件的局部极小点，即存在$\alpha > 0$ 使得$\nabla f(x_*) = 0, \nabla^2 f(x_*) \succeq \alpha I$. 那么当初始点$x_0$满足

$$\|x_0 - x_*\| \leqslant \tfrac{\alpha}{2\beta}$$

那么$\forall t \geqslant 0$，牛顿法是适定的，而且满足

$$\|x_{t+1} - x_*\| \leqslant \tfrac{\beta}{\alpha} \|x_t - x_*\|^2.$$

Proof. 由已知$\nabla f(x_*) = 0$, 有

$$\begin{aligned} x_{t+1} - x_* &= x_t - x_* - \nabla^2 f(x_t)^{-1} \nabla f(x_t) \\ &= \nabla^2 f(x_t)^{-1} [\nabla^2 f(x_t)(x_t - x_*) - [\nabla f(x_t) - \nabla f(x_*)]] \end{aligned}$$

这蕴含着

$$\|x_{t+1} - x_*\| \leqslant \|\nabla^2 f(x_t)^{-1}\| \cdot \|\nabla^2 f(x_t)(x_t - x_*) - [\nabla f(x_t) - \nabla f(x_*)]\|$$

断言 24.2. $\|\nabla^2 f(x_t)(x_t - x_*) - [\nabla f(x_t) - \nabla f(x_*)]\| \leqslant \frac{\beta}{2} \|x^t - x_*\|^2$

Proof. 对$\nabla f(x_t)$应用带积分型余项的Taylor定理，有

$$\nabla f(x_t) - \nabla f(x_*) = \int_0^1 \nabla^2 f(x_t + \gamma(x_* - x_t)) \cdot (x_t - x_*) \, \mathrm{d}\gamma.$$

因此有

$$\begin{aligned} &\|\nabla^2 f(x_t)(x_t - x_*) - [\nabla f(x_t) - \nabla f(x_*)]\| \\ &= \left\| \int_0^1 [\nabla^2 f(x_t) - \nabla^2 f(x_t + \gamma(x_t - x_*))](x_t - x_*) \, \mathrm{d}\gamma \right\| \\ &\leqslant \int_0^1 \|\nabla^2 f(x_t) - \nabla^2 f(x_t + \gamma(x_t - x_*))\| \cdot \|x_t - x_*\| \, \mathrm{d}\gamma \\ &\leqslant \beta \left(\int_0^1 \gamma \, \mathrm{d}\gamma \right) \|x_t - x_*\|^2 \qquad (\nabla^2 f(x_t)\text{是}\beta\text{−Lipschitz的}) \\ &= \tfrac{\beta}{2} \|x_t - x_*\|^2 \end{aligned}$$

■

断言 24.3. Hessian阵$\nabla^2 f(x_t)$正定，并且满足$\|\nabla^2 f(x_t)^{-1}\| \leqslant \frac{2}{\alpha}$.

Proof. 由Wielandt-Hoffman定理[9]，有

$$\begin{aligned}|\lambda_{\min}(\nabla^2 f(x_t)) - \lambda_{\min}(\nabla^2 f(x_*))| &\leqslant \|\nabla^2 f(x_t) - \nabla^2 f(x_*)\| \\ &\leqslant \beta\|x_t - x_*\| \quad (\nabla^2 f(x)\text{是1-Lipschitz连续的})\end{aligned}$$

因此，由已知$\nabla^2 f(x_*) \succeq \alpha I$，对于$\|x_t - x_*\| \leqslant \frac{\alpha}{2\beta}$，这蕴含着

$$\lambda_{\min}(\nabla^2 f(x_t)) \geqslant \frac{\alpha}{2}.$$

所以$\nabla^2 f(x_t)$是正定的，并且因此，

$$\|\nabla^2 f(x_t)^{-1}\| \leqslant \frac{2}{\alpha}.$$

■

将这两个断言放在一起，有

$$\|x_{t+1} - x_*\| \leqslant \frac{2}{\alpha} \cdot \frac{\beta}{2}\|x_t - x_*\|^2 = \frac{\beta}{\alpha}\|x_t - x_*\|^2.$$

■

请注意证明中并不需要凸性. 如果当前迭代点已经在局部极小点x_*的局部邻域内，那么仅在$O(\log\log\frac{1}{\epsilon})$步就能达到$\epsilon$ 误差. 将这称作**二次收敛(quadratic convergence)**.

24.1 阻尼更新

通常，牛顿法有可能非常难以预测. 比如，考虑函数

$$f(x) = \sqrt{x^2 + 1},$$

这本质上是绝对值$|x|$的光滑版本. 很显然，函数在$x_* = 0$处取到最小值. 计算牛顿法所需要的导数，发现

$$\begin{aligned}f'(x) &= \frac{x}{\sqrt{x^2+1}} \\ f''(x) &= (1+x^2)^{-3/2}.\end{aligned}$$

[9]Hoffman-Wielandt不等式：设A, B是n 阶Hermite矩阵，矩阵的特征值排列为$\lambda_1(\cdot) \geqslant \cdots \geqslant \lambda_n(\cdot)$. 则对$p \geqslant 1$，有

$$\sum_{i=1}^{n} |\lambda_i(A) - \lambda_i(B)|^p \leqslant \|A - B\|_p^p$$

成立. 当$p = 2$ 时，Hoffman-Wielandt 不等式等价于

$$\mathrm{tr}(AB) \leqslant \sum_{i=1}^{n} \lambda_i(A)\lambda_i(B),$$

即von-Neumann迹不等式.

请注意$f(x)$的二阶导数是严格正的，从而是强凸的，并且是1-光滑的($|f''(x)| < 1$). 极小化$f(x)$的牛顿迭代是

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x)} = -x_t^3 .$$

该算法的行为与x_t的幅度有关. 特别地，有如下三个环境：(i) $|x_t| < 1$, 算法三次**(cubically)**收敛， (ii) $|x_t| = 1$, 算法在-1和1之间振荡， (iii) $|x_t| > 1$，算法发散. 该例表明即使对梯度是Lipschitz连续的强凸函数，也只能保证牛顿法是局部收敛的. 为了避免发散，使用**阻尼步长(damped step-size)**技术：

$$x_{t+1} = x_t - \eta_t \nabla^2 f(x_t)^{-1} \nabla f(x_t),$$

其中可用第 18 节的回溯Armijo线搜索选取步长η_t. 通常$\bar{\eta} = 1$是好的首次选取值，因为如果已经在收敛域内，则步长取 1 ，从而能保证得到二次收敛.

24.2 拟牛顿法

将梯度下降法和牛顿法放到一起比较：

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \qquad \text{(梯度下降法)}$$
$$x_{t+1} = x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t). \qquad \text{(牛顿法)}$$

可将梯度下降法看作牛顿更新中用单位矩阵的伸缩来近似

$$\nabla^2 f(x_t)^{-1},$$

即当

$$\nabla^2 f(x_t)^{-1} = \eta_t I$$

时，梯度下降等价于牛顿法，这里的I是单位矩阵.

拟-牛顿法通过用某个其它矩阵近似Hessian阵以便取类似的步. 如此做的动机是避免每步昂贵的矩阵求逆. 想要一个近似

$$\widehat{f}_{B_t}(x) \approx f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2}(x - x_t) B_t^{-1} (x - x_t)$$

满足：

$$\nabla \widehat{f}_{B_t}(x_t) = \nabla f(x_t). \tag{24.1}$$

和

$$\nabla \widehat{f}_{B_t}(x_{t-1}) = \nabla f(x_{t-1}) \tag{24.2}$$

条件(24.1)表明该近似中的一阶项是精确的，这看起来是合理的. 条件(24.2)说明梯度在前一个迭代处也是正确的. 如果上两次的梯度是正确的，期待沿着方向$x_t - x_{t-1}$的Hessian阵近似是合理的. 称(24.2) 是**割线近似(secant approximation)**，可写作

$$\nabla \widehat{f}_{B_t}(x_{t-1}) = \nabla f(x_t) + B_t^{-1}(x_{t-1} - x_t) = \nabla f(x_{t-1}).$$

如果令

$$s_{t-1} = x_t - x_{t-1}$$
$$y_{t-1} = \nabla f(x_t) - \nabla f(x_{t-1}),$$

那么割线方程即

$$s_{t-1} = B_t y_{t-1}.$$

存在多个B_t满足该条件. 可以添加其它约束来缩小特定选取. 一个流行的要求是需要B_t是正定的、确信B_t对于某种度量尽可能接近B_{t-1}，或者要求B_t 是以前迭代的低秩更新，可通过Sherman–Morrison公式完成其中的更新. 这种实现中最成功的之一是BFGS和有限内存BFGS(L–BFGS)，这里的BFGS是以发明者的姓Broyden, Fletcher, Goldfarb, Shanno来命名的.

25 二阶方法的实验

本讲是一系列代码示例，在这里可以找到它们：

Lecture 24
(在你的浏览器中打开)

26 内点法入门

在上一讲，讨论了存粹牛顿法. 尽管它享有快的局部收敛保证，牛顿法的全局收敛是没有保证的. 本讲引入内点法，这可看作能保证全局收敛的扩展牛顿法[10]. 首先引入通用**障碍法(barrier methods)** 的主要思想.

26.1 障碍法

障碍法用所谓的**障碍函数(barrier function)** 代替不等式约束，并将它加到优化问题的目标函数中. 考虑如下优化问题：

$$\begin{array}{ll} \min_x & f(x) \\ \text{s.t.} & x \in \Omega, \\ & g_j(x) \leqslant 0\,, j = 1, \cdots, m\,, \end{array} \tag{26.1}$$

其中$f : \mathbb{R}^n \to \mathbb{R}$, $g_j : \mathbb{R}^n \to \mathbb{R}$ 是已知函数. f 是连续的，并且Ω是闭集. 本讲的剩余部分，假设g_j是连续凸函数，并且$\Omega = \mathbb{R}^n$. 用x_* 表示问题 (26.1)的最优解.

定义 26.1 (约束域的内部)**.** 约束域(相对于Ω)的内部定义为

$$S = \{x \in \Omega : g_j(x) < 0,\ j = 1, \cdots, m\}.$$

[10]使用数值技术使得存粹牛顿法具有全局收敛性的方法还有修正牛顿法和信赖域牛顿法

假设S非空. 所谓S上的障碍函数$B(x)$ 定义为连续的、并且当x趋于约束域的边界时趋于正无穷. 更正式的，

$$\lim_{g_j(x)\to 0-0} B(x) = +\infty,$$

两个最常见的例子是对数障碍函数和倒数障碍函数：

$$\text{对数：} \qquad B(x) = -\sum_{j=1}^{m} \ln(-g_j(x)), \tag{26.2}$$

$$\text{倒数：} \qquad B(x) = -\sum_{j=1}^{m} \frac{1}{g_j(x)}. \tag{26.3}$$

如果所有$g_j(x)$是凸的时，这两个均是凸函数.

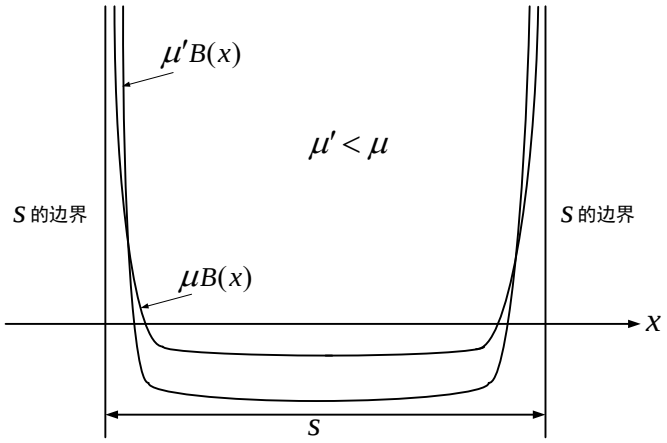


图 26.1: 障碍项的形状

已知障碍函数$B(x)$, 定义新的成本函数 $f_\mu(x) = f(x) + \mu B(x)$, 其中$\mu$是正参数，称作障碍因子，用来控制近似解距$S$ 边界的距离. 那么，可以在约束问题中删去不等式约束，得到如下问题：

$$\min_{x\in\Omega} f_\mu(x)$$

障碍项$\mu B(x)$的形状见图 26.2. 当μ变小时，f_μ的极小点会更接近可行域的边界. 从而，引入序列$\{\mu_t\}$ 来定义障碍法，该序列满足

$$0 < \mu_{t+1} < \mu_t, t = 0, 1, \cdots,$$

和$\mu_t \to 0$. 那么找到序列$\{x_t\}$使得

$$x_t \in \arg\min_{x\in S} f_{\mu_t}(x). \tag{B_{μ_t}}$$

例子 26.2 (对数障碍函数). 问题

$$\begin{array}{ll} \text{minimize}_x & x \\ \text{subject to} & g(x) := 1 - x \leqslant 0 \end{array} \tag{26.4}$$

的对数障碍函数为 $f_\mu(x) = x - \mu\ln(x-1)$, 图 26.2对一组 μ 的值给出了函数的图像，由此可以看到随着 $\mu_t \to 0, x_t \to x_*$.

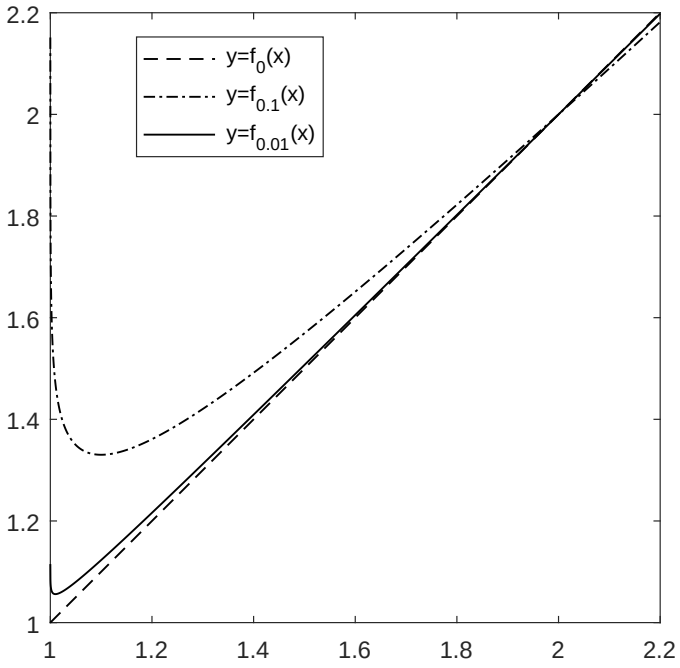


图 26.2: 障碍函数中递增的病态性

请注意对于所有内点$x \in S$，当$\mu_t \to 0$, 障碍项$\mu_t B(x)$ 趋于零，这允许x_t越来越接近边界. 因此，直观上，不管x_* 是在S的内部，还是在它的边界上，x_t 都应该逼近x_*. 它的收敛性的正式徐述见如下命题.

命题 26.3. 由障碍法产生的序列$\{x_t\}$的每个极限点是原始约束优化问题 (26.1) 的全局极小点.

Proof. 参见[Ber16]的命题5.1.1. ■

上面的命题表明障碍问题 (B_{μ_t})的全局最优解收敛到全局约束最优解. 但是如何求解这一系列的最优化问题？ 核心直观是：对足够大的μ_0，通常易于得到初始内点. 那么在每次迭代，能使用x_t作为初始点，用牛顿法找到x_{t+1}. 如果μ_t靠近μ_{t+1}, 期望x_t也靠近x_{t+1}. 因此，有理由认为x_t 在问题($B_{\mu_{t+1}}$)的牛顿法的局部二次收敛域内. 用这种方式，可将牛顿法的局部收敛保证延拓成全局性质.

从实践角度讲，需要解决以下三个问题. 首先，需要一种方法能找到严格可行点x_0来初始化算法. 其次，需要设计求解子问题 (B_{μ_t})的有效方法. 最后，需要指定障碍因子μ_t的更新方法. 而后面两个问题，通常是放在一起解决的.

26.2 线性规划

在勾勒出整个扩展的基本思想后，现在将裁剪对数障碍法，将其应用到如下定义的线性规划(Linear programming, LP)问题：

$$\begin{aligned} \min_{x} \quad & c^\top x \\ \text{s.t.} \quad & Ax \geqslant b \end{aligned} \tag{LP}$$

其中$A \in \mathbb{R}^{m\times n}, m \geqslant n$，并且$\mathrm{rank}(A) = n$. 记$x_*$是 (LP)问题的最优解.

首先，写出由对数障碍函数得到的增广成本函数，即

$$f_\mu(x) = c^\top x - \mu \sum_{j=1}^{m} \ln\left(a_j^\top x - b_j\right). \tag{26.5}$$

其中$a_j^\top$是矩阵A的第j行，b_j是向量b的第j个分量. 定义$x_\mu^* = \mathrm{argmin}_x f_\mu(x)$.

事实 26.4. 对任何$\mu > 0$，最优点x_μ^*存在并且唯一.

Proof. 易于检查$f_\mu(x)$是凸的(是两个凸函数之和). 已于说明f_μ有稳定点，它的Hessian阵是正定的，从而极小点唯一. 因此，极小点必定存在并且唯一[11].

为了证明f_μ的凸性，检查二阶导数，如后面(26.7)所示，它是正定的. ■

26.2.1 中心路径

集合$\left\{x_\mu^* : \mu > 0\right\}$描述了 (LP)问题的**中心路径**(central path)，几何直观如图 26.3所示. 目标是设计算法，其能近似地跟踪中心路径. 假设已经有一个“足够好”的初始点, 那么在每一步，应用单步牛顿法. 为了保证算法收敛，需要回答如下问题：

- 单步牛顿法在什么条件下能够工作？
- 应该如何更新μ?

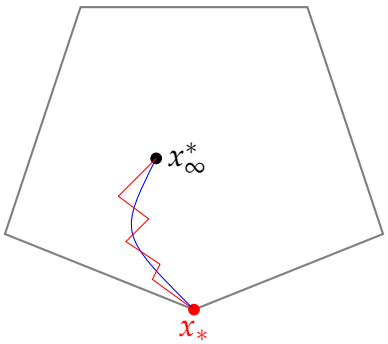


图 26.3: 蓝色曲线表示一个二维LP实例的中心路径

26.2.2 牛顿减量

为了应用牛顿法，首先需要求出f_μ的一阶导数和二阶导数. 请注意

$$\nabla f_\mu(x) = c - \mu \sum_{j=1}^{m} \frac{a_j}{a_j^\top x - b_j} \triangleq c - \mu A^\top S^{-1}\mathbf{1} \tag{26.6}$$

$$\nabla^2 f_\mu(x) = \mu A^\top S^{-2} A = \mu \sum_{j=1}^{m} \frac{a_j a_j^\top}{s_j^2} \tag{26.7}$$

[11]凸函数极小点的存在性：有稳定点，必有极小点.
极小点何时唯一？- Hessian阵正定，是严格凸函数，从而极小点唯一. 可将这两个问题设计成作业.

其中$\mathbf{1} = [1, 1, \cdots, 1]^\top \in \mathbb{R}^{m\times 1}$, 且 $S = \mathrm{Diag}\{s_1, \ldots, s_m\}$是由松弛变量 $s_j = a_j^\top x - b_j$ 构成的对角矩阵. 回忆牛顿更新

$$\bar{x} = x - [\nabla^2 f_\mu(x)]^{-1}\nabla f_\mu(x) = x - [\mu A^\top S^{-2} A]^{-1}\left(c - \mu A^\top S^{-1}\mathbf{1}\right).$$

这是通过令一阶近似为零，求所得方程组得到的解. 为了度量牛顿更新使得一阶近似减小的幅度，引入牛顿减量的概念.

定义**牛顿减量(Newton decrement)** $q(x, \mu)$ 为

$$q^2(x, \mu) = \nabla f_\mu(x)^\top [\nabla^2 f_\mu(x)]^{-1}\nabla f_\mu(x).$$

等价地，

$$\begin{aligned} q(x, \mu) &= \left\| [\nabla^2 f_\mu(x)]^{-1/2}\nabla f_\mu(x) \right\|_2 \\ &= \left\| \nabla^2 f_\mu(x)^{-1}\nabla f_\mu \right\|_{\nabla^2 f_\mu(x)}, \end{aligned}$$

其中 $\|x\|_H = \sqrt{x^\top H x}$. 最后的等式揭示了可将牛顿减量看作由Hessian 阵定义的**局部范数(local norm)** 来度量牛顿步的幅度.

请注意牛顿减量也将$f_\mu(x)$与其二阶近似的最小值之差关联起来：

$$\begin{aligned} & f_\mu(x) - \min_{\bar{x}} \left(f_\mu(x) + \nabla f_\mu(x)^\top(\bar{x} - x) + \tfrac{1}{2}(\bar{x} - x)^\top \nabla^2 f_\mu(x)(\bar{x} - x) \right) \\ = \;& f_\mu(x) - \left(f_\mu(x) - \frac{1}{2}\nabla f_\mu(x)^\top [\nabla^2 f_\mu(x)]^{-1}\nabla f_\mu(x) \right) \\ = \;& \frac{1}{2}\nabla f_\mu(x)^\top [\nabla^2 f_\mu(x)]^{-1}\nabla f_\mu(x) =: \frac{1}{2}q^2(x, \mu). \end{aligned} \tag{26.8}$$

将使用牛顿减量来找出保证算法收敛的条件.

26.2.3 短步路径跟踪算法的更新规则和收敛性

现在将提出一个障碍因子的更新规则，如果满足某些初始条件，就能保证收敛. 为了给出更新规则，首先引入如下命题.

命题 26.5. 假设严格可行点x(即$Ax > b$)满足$q(x, \mu) < 1$. 那么有

$$c^\top x - c^\top x_* \leqslant 2\mu n.$$

特别地，如果维持x_t是满足$Ax_t > b$的内点，并且$q(x_t, \mu_t) < 1$, 那么当μ_t收敛到 0 时，$c^\top x_t$ 收敛到$c^\top x_*$，即x_t 收敛到全局最优点. 然而，条件$q(x_t, \mu_t) < 1$并不是平凡的.

命题 26.6. 已知障碍因子$\mu > 0$. 假设严格可行点x(即$Ax > b$)满足$q(x, \mu) < 1$. 那么纯粹牛顿迭代步$\bar{x}$ 满足

$$q(\bar{x}, \mu) \leqslant q(x, \mu)^2.$$

该结论表明，对障碍因子μ固定时，从满足$q(x,\mu) < 1$的严格可行点x出发，纯粹牛顿迭代法是良定的，并且二次收敛于中心路径上的点x^*_μ. 此外，希望对于某$\bar{\mu} < \mu$，也有$q(\bar{x},\bar{\mu}) < 1$成立.

命题 26.7. 假设正数μ和x满足$q(x,\mu) \leqslant \frac{1}{2}$ 和$Ax > b$. 置

$$\bar{\mu} = \left(1 - \tfrac{1}{6\sqrt{n}}\right)\mu,$$

那么有

$$q(\bar{x},\bar{\mu}) \leqslant \tfrac{1}{2}.$$

这些命题表明如下更新规则，

$$\begin{aligned} x_{t+1} &= x_t - \nabla^2 f_{\mu_t}(x)^{-1}\nabla f_{\mu_t}(x_t) \\ \mu_{t+1} &= \left(1 - \tfrac{1}{6\sqrt{n}}\right)\mu_t \end{aligned}$$

定理 26.8. 假设(x_0,μ_0)满足$Ax_0 > b$和 $q(x_0,\mu_0) \leqslant \frac{1}{2}$, 那么算法在$\mathcal{O}(\sqrt{n}\log(n/\epsilon))$次迭代内收敛到$\epsilon$ 误差，即在$t = \mathcal{O}(\sqrt{n}\log(n/\epsilon))$ 次迭代之后，有

$$c^\top x_t \leqslant c^\top x_* + \epsilon.$$

Proof. 因为牛顿步保持x_{t+1}在Ω的内部，使用上面的三个命题，有

$$\begin{aligned} c^\top x_t &\leqslant c^\top x_* + 2\mu_t n \\ &= c^\top x_* + 2\left(1 - \tfrac{1}{6\sqrt{n}}\right)^t n\mu_0 \\ &\leqslant c^\top x_* + 2\exp\left(-\tfrac{t}{6\sqrt{n}}\right) n\mu_0 \end{aligned} \tag{26.9}$$

因此，当 $t \geqslant 6\sqrt{n}\ln\frac{2n\mu_0}{\epsilon}$ 时，误差为ϵ. 那么能得到算法在 $\mathcal{O}(\sqrt{n}\ln(n/\epsilon))$ 次迭代后收敛误差达到ϵ. ■

上面陈述的是所谓的**短步**路径跟踪算法. 尽管收敛速率有理论上的保证， 但在实践中，μ 小的减少量和单步牛顿法的结合很慢. 与此相反，一个更实用的方法是所谓的长步法，其中每次迭代的μ 以更快的速率在减小，同时取好几个牛顿步.

这小节的内容选自 [Wri92]. 这里尚未说明如何选择初始障碍因子μ_0和获得满足$Ax_0 > b$ 和$q(x_0,\mu_0)$的初始点x_0.

27 原始-对偶内点法

前面讨论了针对不等式约束的线性规划问题 (LP) 的障碍法和所谓的短步法，并证明收敛是有保证的(尽管慢). 这里研究一种原始-对偶内点法(所谓的 "长步"路径跟踪算法), 它类似地在中心路径上寻找近似点. 与短步法不同，长步法考虑原始-对偶迭代，并且只要它位于中心路径的邻域内，由此寻找一个更大胆的步长.

27.1 得到对偶问题

设 $x \in \mathbb{R}^n$是决策变量，$A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$. 考虑标准形式的线性规划问题

$$\min\ c^\top x \quad \text{s.t.} \quad Ax = b, x \geqslant 0, \tag{LP$_\text{S}$}$$

其中$\geqslant$是逐分量的. 观察可将 (LP$_\text{S}$)等价表述为

$$\begin{aligned} & \min_{x\geqslant 0}\ c^\top x + \max_z z^\top(b - Ax) \\ = & \min_{x\geqslant 0}\max_z [c^\top x + z^\top(b-Ax)] \\ \geqslant & \max_z \min_{x\geqslant 0} z^\top b + (c - A^\top z)^\top x. \end{aligned}$$

由于

$$\min_{x\geqslant 0} z^\top b + (c - A^\top z)^\top x = \begin{cases} b^\top z, & A^\top z \leqslant c \\ -\infty, & \text{否则.} \end{cases}$$

因此， (LP$_\text{S}$)的对偶问题是

$$\max\ b^\top z \quad \text{s.t.} \quad A^\top z \leqslant c. \tag{LP$_\text{D}$}$$

这等价于

$$\max\ b^\top z \quad \text{s.t.} \quad A^\top z + s = c,\ s \geqslant 0,$$

其中引入了松弛变量$s \in \mathbb{R}^n$. 如果(x,z,s)仅仅是**可行的(feasible)**, 那么

$$Ax = b,\ A^\top z + s = c,\ x, s \geqslant 0.$$

此外，对于可行的(x,z,s)可以计算

$$0 \leqslant \langle x, s\rangle = \langle x, c - A^\top z\rangle = \langle x, c\rangle - \langle Ax, z\rangle = \langle x, c\rangle - \langle b, z\rangle .$$

这是弱对偶性的证明，即对于任何可行的x和z，有 $\langle x, c\rangle \geqslant \langle b, z\rangle$ 成立. 因此

$$\langle x_*, c\rangle \geqslant \langle b, z^*\rangle.$$

此外，如果存在原始-对偶可行对(x_*, z^*, s^*)满足$\langle x_*, s^*\rangle = 0$，那么有

$$\langle x_*, c\rangle = \langle b, z^*\rangle.$$

这是强对偶性.

对偶性对于上控次优性间隙非常有用，因为事实上，如果(x,z,s) 是原始-对偶可行对，那么

$$\langle x, s\rangle = \langle x, c\rangle - \langle b, z\rangle \geqslant \langle x, c\rangle - \langle x_*, c\rangle = \langle x - x_*, c\rangle.$$

27.2 沿着中心路径的原始-对偶迭代

定义严格原始-对偶可行集

$$\mathcal{F}^{\mathrm{o}} := \{(x,z,s) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n : Ax = b, A^\top z + s = c, x, s > 0\}.$$

对$(x,z,s) \in \mathcal{F}^{\mathrm{o}}$，定义

$$\bar{\mu} = \bar{\mu}(x,s) := \frac{\langle x,s\rangle}{n} = \frac{\langle x,c\rangle - \langle b,z\rangle}{n} \geqslant \frac{\langle x - x_*, c\rangle}{n} \tag{27.1}$$

上面的讨论让人想到如下方法：在线性约束集$\mathcal{F}^{\mathrm{o}}$上极小化**双线性(bilinear)** 目标函数$x^\top s$. 为此，已知$(x_t,z_t,s_t) \in \mathcal{F}^{\mathrm{o}}$, 目的是产生迭代$(x_{t+1},z_{t+1},s_{t+1}) \in \mathcal{F}^{\mathrm{o}}$ 使得

$$\bar{\mu}_{t+1} \leqslant (1 - C(n))\bar{\mu}_t,$$

其中$\bar{\mu}_t = \bar{\mu}(x_t,s_t), \bar{\mu}_{t+1} = \bar{\mu}(x_{t+1},s_{t+1}), C(n) \in (0,1)$是与$n$有关的常数.

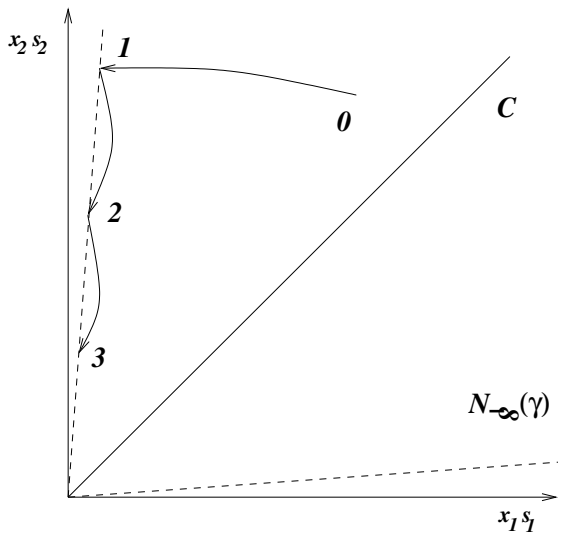


图 27.1: 要求迭代停留在中心路径的某个邻域内. 想要逐分量乘积$x_1s_1, \cdots, x_ns_n$相差不要太大

目标是找到三元对$(x,z,s) \in \mathcal{F}^{\mathrm{o}}$使得$\bar{\mu} \approx 0$. 为此，考虑如下方法. 定义

$$F_\mu(x,z,s) := \begin{bmatrix} Ax - b \\ A^\top z + s - c \\ x \circ s - \mu \mathbf{1} \end{bmatrix},$$

这里$\circ$表示两个矩阵的Hardmard逐分量乘积. 那么，目标是在$\mathcal{F}^{\mathrm{o}}$上近似求解

$$F_0(x,z,s) = \mathbf{0}.$$

观察发现，看到通过计算

$$F_\mu(x,z,s) = \mathbf{0}$$

的解 (x_μ, z_μ, s_μ) 可以达到目的. 回忆曲线$\mu \mapsto (x_\mu, z_\mu, s_\mu)$ 定义了"原始-对偶中心路径". 请注意，在原始-对偶中心路径上，对于某个$\mu > 0$ 有$x_i s_i = \mu$ 成立. 为了确保离中心路径很近，考虑

$$\mathcal{N}_{-\infty}(\gamma) \;:=\; \{(x,z,s) \in \mathcal{F}^{\mathrm{o}} : \min_i x_i s_i \geqslant \gamma \bar{\mu}(x,s)\}.$$

对于恰当的常数γ，已知$(x_t, z_t, s_t) \in \mathcal{N}_{-\infty}(\gamma)$. 想要选取迭代$(x_{t+1}, z_{t+1}, s_{t+1}) \in \mathcal{N}_{-\infty}(\gamma)$，并且$\bar{\mu}(x_{t+1}, s_{t+1})$严格变小. 这里属于$\mathcal{N}_{-\infty}(\gamma)$ 的要求确保了非负约束. 图27.1 给出了$\mathcal{N}_{-\infty}(\gamma)$ 和迭代过程的几何直观.

Algorithm 6 长步路径跟踪算法

Require: Parameters $\gamma \in (0,1)$, $0 < \sigma_{\min} < \sigma_{\max} < 1$.

1: Initialization $(x_0, z_0, s_0) \in \mathcal{N}_{-\infty}(\gamma)$ and get $\bar{\mu}_0 = \frac{1}{n}\langle x_0, s_0\rangle$.
2: **for** $t = 1$ to $\cdots$ **do**
3: Choose $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ and let $\mu_t = \sigma \bar{\mu}_t$.
4: Run Newton step on F_{μ_t} (to be defined).
5: Let $(\Delta x_t, \Delta z_t, \Delta s_t)$ denote the Newton step

$$(\Delta x_t, \Delta z_t, \Delta s_t) = -\nabla^2 F_{\mu_t}(w_t)^{-1} \cdot \nabla F_{\mu_t}(w_t),$$

where $w_t = (x_t, z_t, s_t)$.
6: Let $\alpha_t \in (0,1]$ be the largest step such that the iteration remains in $\mathcal{N}_\infty(\gamma)$, i.e.

$$\alpha_t = \max\{\alpha \in (0,1] : (x_t, z_t, s_t) + \alpha(\Delta x_t, \Delta z_t, \Delta s_t) \in \mathcal{N}_\infty(\gamma)\}.$$

7: Set $(x_{t+1}, z_{t+1}, s_{t+1}) \leftarrow (x_t, z_t, s_t) + \alpha_t(\Delta x_t, \Delta z_t, \Delta s_t)$.
8: Set $\bar{\mu}_{t+1} = \frac{1}{n}\langle x_{t+1}, s_{t+1}\rangle$
9: **end for**

27.3 用牛顿步生成迭代

考虑求解方程组 $F(w) = 0$的牛顿步. 的确

$$F(w+d) = F(w) + J_F(w) \cdot d + o(\|d\|).$$

牛顿法选取 $w \leftarrow w + d$, 其中d满足 $J_F(w)d = -F(w)$, 这蕴含着对于充分接近方程组的解w的向量$w + d$有

$$F(w+d) = o(\|d\|).$$

由此给出快速收敛. 牛顿迭代的几何直观如图 27.2. 请注意，如果F 是线性映射，那么事实上一个牛顿步是充分的. 这可由Taylor展式得到.

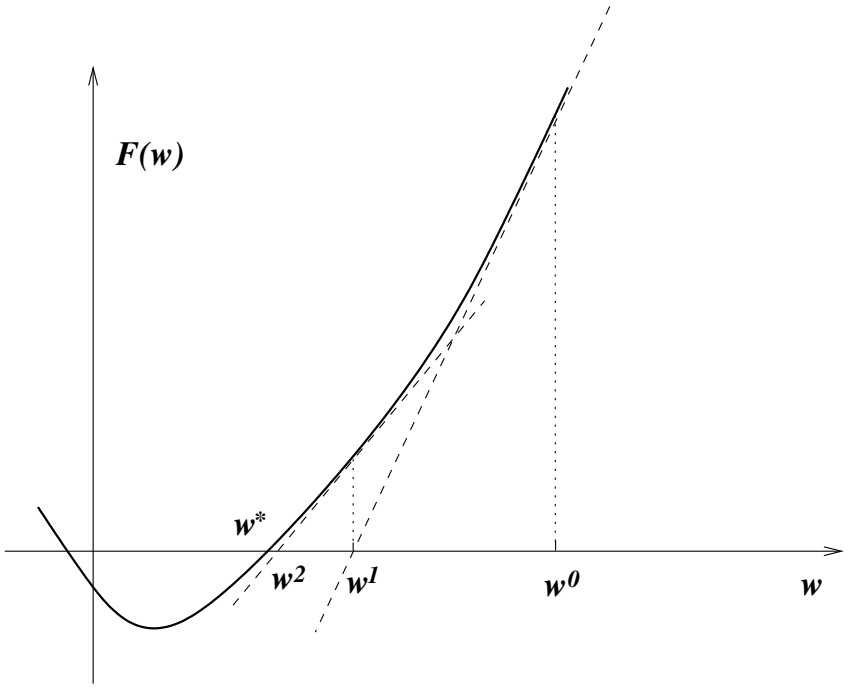


图 27.2: 回忆牛顿法迭代地求得实值函数根(或者零点)的更好近似.

这里的函数F_μ几乎是线性的，但不完全是. 已知$(x,z,s) \in \mathcal{F}^{\mathrm{o}}$和$\mu > 0$. 现在计算牛顿步. 观察到Jacobi 矩阵是线性算子

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^\top & I \\ S & 0 & X \end{bmatrix},$$

其中$S = \mathrm{Diag}(s), X = \mathrm{Diag}(x)$分别表示以向量$s$和$x$作对角线元素对应的对角矩阵. 此外，由于$(x,z,s) \in \mathcal{F}^{\mathrm{o}}$, 有

$$F_\mu(x,z,s) = \begin{bmatrix} Ax - b \\ A^\top z + s - c \\ x \circ s - \mu \mathbf{1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ x \circ s - \mu \mathbf{1} \end{bmatrix}.$$

那么，可以验证牛顿增量满足

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^\top & I \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \\ \Delta s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -x \circ s + \mu \mathbf{1} \end{bmatrix}.$$

一些按语：

(i) $A\Delta x = 0$, 且$A^\top \Delta z + \Delta s = 0$. 这蕴含着

$$(x^+, z^+, s^+) := (x + \Delta x, z + \Delta z, s + \Delta s)$$

满足

$$Ax^+ - b = 0 \quad \text{和} \quad A^\top z^+ + s^+ - c = 0. \tag{27.2}$$

(ii) 也有

$$s \circ \Delta x + x \circ \Delta s = -x \circ s + \mu \mathbf{1}. \tag{27.3}$$

这样

$$\begin{aligned} x^+ \circ s^+ &= x \circ s + (s \circ \Delta x + x \circ \Delta s) + \Delta x \circ \Delta s \\ &= x \circ s - x \circ s + \mu \mathbf{1} + \Delta x \circ \Delta s \end{aligned}$$

(iii) 这样，

$$F_\mu(x^+, z^+, s^+) = \begin{bmatrix} 0 \\ 0 \\ \Delta x \circ \Delta s \end{bmatrix}. \tag{27.4}$$

换句话说，如果能论证出$\Delta x \circ \Delta s$ 是“可忽略的”，那么牛顿步产生一个几乎精确的解.

更具体的，分析新得到迭代点的临近(中心路径)性度量(算法6的第8行)

$$\begin{aligned} n\bar{\mu}(x + \alpha\Delta x, s + \alpha\Delta x) &= \langle x + \alpha\Delta x, s + \alpha\Delta s\rangle \\ &= \langle x, s\rangle + \alpha\left[\langle x, \Delta s\rangle + \langle s, \Delta x\rangle\right] + \alpha^2\langle \Delta s, \Delta x\rangle. \end{aligned}$$

由上面的事实(i)-(iii)，可知上式中的最后一项消失，因为

$$0 = \Delta x^\top (A^\top \Delta z + \Delta s) = (A\Delta x)^\top \Delta z + \langle \Delta x, \Delta s\rangle = \langle \Delta x, \Delta s\rangle,$$

其中第一和第二个等式是因为(i). 此外，将 (27.3)中的n个方程求和，得到

$$\langle x, \Delta s\rangle + \langle s, \Delta x\rangle = -\langle x, s\rangle + \mu n = -(1-\sigma)\langle x, s\rangle$$

其中最后一个等式使用了行使用(算法6的第2行和$\bar{\mu}$的定义 (27.1))了

$$n\mu = n\sigma\bar{\mu} = \sigma\langle x, s\rangle.$$

因此，

$$n\bar{\mu}(x + \alpha\Delta x, s + \alpha\Delta x) = n\bar{\mu}(x, s)[1 - (1-\sigma)\alpha]$$

因此，如果能够证明存在某个与维数相关的常数α 使得

$$(1-\sigma)\alpha \geqslant C(n) > 0,$$

那么看到

$$n\bar{\mu}(x_{t+1}, s_{t+1}) \leqslant (1 - C(n))^t n\bar{\mu}(x_0, s_0)$$

给出了最优性度量的减小速率. 用技术性更强的分析能够证明$\alpha = \Omega(1/n)$，同时保持$\mathcal{N}_{-\infty}(\gamma)$是不变的.

References

[AZO17] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proc. 8th ITCS*, 2017.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[Ber16] D.P. Bertsekas. *Nonlinear Programming*. Athena scientific optimization and computation series. Athena Scientific, 2016.

[Bou14] Jean Bourgain. *An Improved Estimate in the Restricted Isometry Problem*, pages 65–70. Springer International Publishing, 2014.

[BS83] Walter Baur and Volker Strassen. The complexity of partial derivatives. *Theoretical computer science*, 22(3):317–330, 1983.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CT06] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Information Theory*, 52(12):5406–5425, 2006.

[DGN14] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

[DJL+17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proc. 25th ICML*, pages 272–279. ACM, 2008.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[FW56] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. *CoRR*, abs/1503.02101, 2015.

[HR15] Ishay Haviv and Oded Regev. The restricted isometry property of subsampled fourier matrices. *CoRR*, abs/1507.01768, 2015.

[HRS15] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *CoRR*, abs/1509.01240, 2015.

[Lax07] Peter D. Lax. *Linear Algebra and Its Applications*. Wiley, 2007.

[LPP+17] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *CoRR*, abs/1710.07406, 2017.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

[Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady AN SSSR (translated as Soviet Mathematics Doklady)*, 269:543–547, 1983.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Programming. Volume I: A basic course*. Kluwer Academic Publishers, 2004.

[RM51] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[Ros58] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386, 1958.

[RV07] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2007.

[SSSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

[SSZ13] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[TD97] Lloyd N. Trefethen and David Bau, III. *Numerical Linear Algebra*. SIAM, 1997.

[TVW+17] Stephen Tu, Shivaram Venkataraman, Ashia C Wilson, Alex Gittens, Michael I Jordan, and Benjamin Recht. Breaking locality accelerates block gauss-seidel. In *Proc. 34th ICML*, 2017.

[Wri92] M. H. Wright. Interior methods for constrained optimization. *Acta Numerica*, 1:341–407, 1992.