北京航空航天大学
学院路校区

最优化方法讲义

刘红英

北京航空航天大学
数学科学学院
Email: liuhongying@buaa.edu.cn

September 7, 2023

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/.

这些讲义的算法部分的原始素材主要出自Berkeley于2018年春季课程*EE227C: Convex Optimization and Approximation*(凸优化和近似，https://ee227c.github.io/)，理论部分的素材主要出自“刘红英，夏勇，周水生. 数学规划基础，北京航空航天大学出版社，北京：2012.10”. 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),
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/)，和
5. L. Vandenberge教授2022 spring在 UCLA的ECE236C 课程使用的幻灯片“Optimization Methods for Large-scale Systems”(大规模问题的优化方法, http://www.seas.ucla.edu/~vandenbe/236C/).

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，也融入了相关领域最新的理论进展和应用. 请注意，讲义所有向量不分行或者列. 默认参与矩阵向量乘运算的向量为列向量.

目录

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, 有时是全局!)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)梯度下降法具有更快的速率.

1 凸性

本节给出凸集和凸函数中部分重要的内容，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$C,K\subseteq\mathbb{R}^n$均是非空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$b\in\mathbb{R}$使得 对所有$x\in C$, 有$\langle a,x\rangle\geqslant b$. 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). 下面给出这种情况下结论的证明.

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$(p,q)\in C\times K$ 是取到最小值的两个点. 过p和q的点且垂直于$q-p$ 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$S_+^n = \{A \in \mathbb{R}^{n\times n} \mid A^T = A, A \succeq 0\}$. 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:
$$S_+^n = \bigcap_{x\in\mathbb{R}^n\setminus\{0\}} \{A \in S_n \mid x^\top Ax \geqslant 0\}.$$

- 已知$x_i \in \mathbb{R}^n, i = 0, 1, \cdots, m$. 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. 比如$2-$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)
$$\Delta_n = \{x \in \mathbb{R}^n : x \geqslant 0, \sum_{i=1}^n x_i = 1\}.$$
将$e_0 = 0$加入$e_1, \cdots, e_n$，得到对应的n 维单纯形是
$$\Delta_n^+ = \{x \in \mathbb{R}^n : x \geqslant 0, \sum_{i=1}^n x_i \leqslant 1\}.$$

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

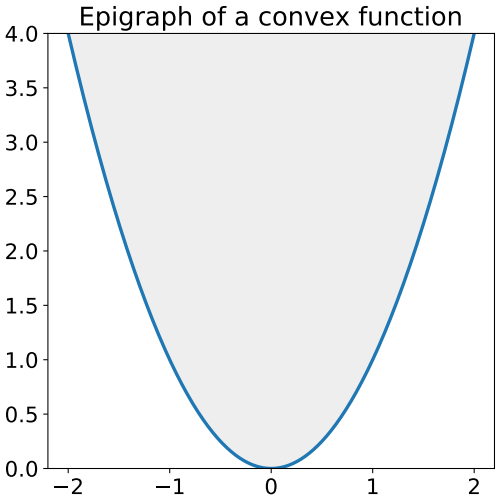


图 1.1: 函数的上图.

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

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

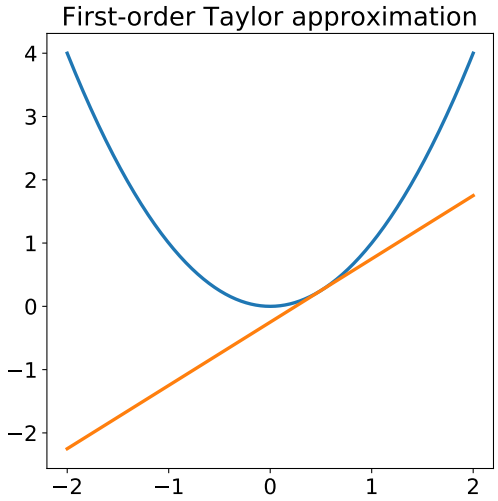


图 1.2: 函数$f(x) = x^2$在点0.5的Taylor近似.

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, $0 \neq d \in \mathbb{R}^n$. 考虑

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

那么,

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

特别地，$\phi'(0) = \nabla f(x)^\top d$.

对于可微函数而言，凸性等价于：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，几何直观如图1.2所示.

命题 **1.9** (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

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

证明. 首先，假设f是凸的，那么由定义

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

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，并相加，得到

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

由此得到凸性. ■

推论 **1.10.** 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

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

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.

Proof. 由凸函数的梯度不等式，充分性是显然的. 下面证明必要性. 任取$x\in\Omega$. 考虑定义在区间$[0,1]$上的一元函数

$$\phi(\theta)=f(x_*+\theta(x-x_*)),$$

因为Ω凸，所以$\forall\theta\in[0,1],x_*+\theta(x-x_*)\in\Omega$. 再由$x_*$的最优性，知$\phi$在$\theta=0$处取到它在区间$[0,1]$上的最小值，从而由一元函数的最优性条件，有

$$\phi'(0)=\nabla f(x_*)^T(x-x_*)\geqslant 0.$$

■

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 **1.11** (次梯度)**.** 设$f:\mathbb{R}^n\to\mathbb{R}$. 对于$x\in\mathrm{dom}f$，若存在向量$g\in\mathbb{R}^n$ 满足

$$f(y)\geqslant f(x)+g^\top(y-x)\ \ \forall y\in\mathrm{dom}f,$$

则称该向量为f在x处的次梯度(subgradient). 记次梯度向量组成的集合为$\partial f(x)$，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$x\in\mathrm{ri}\ \Omega$(表示集合Ω的相对内部，是集合在包含自己的仿射空间中的内部. 比如$\Omega=\{(1,y)\in\mathbb{R}^2:y\in(0,1]\}$，则$\mathrm{int}\Omega=\varnothing,\mathrm{ri}\Omega=\{(1,y)\in\mathbb{R}^2:y\in(0,1)\}$)，$\partial f(x)\neq\varnothing$. 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. 留作作业.

1.2.2 二阶刻画

将$f:\Omega\to\mathbb{R}$在点$x\in\Omega$处的Hessian阵定义为元素为二阶偏导数的矩阵:

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

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.

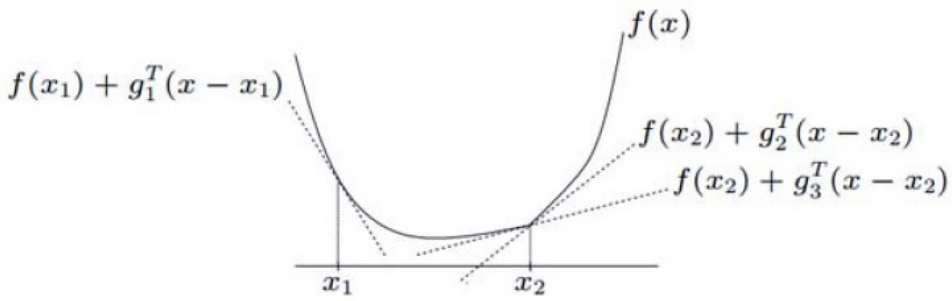


图 1.3: 函数f在x_1处是可微的，在x_2处是次可微的，g_2和g_3 均是f在x_2处的次梯度.

事实 **1.12.** 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$d = y - x$，考虑由式 (1.1)定义的一元函数. 那么

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

Taylor定理蕴含着如下命题.

命题 **1.13.** 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

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

证明. 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. ■

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

证明. 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 1.9表明

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

因此，由命题 1.13,

$$\begin{aligned}
0 &\leqslant \int_0^1 (1 - \gamma)\phi''(\gamma)\,\mathrm{d}\gamma \\
&= (1 - \bar{\gamma})\phi''(\bar{\gamma}) \quad \text{对某个} \bar{\gamma} \in (0, 1) && \text{(中值定理)} \\
&= (1 - \bar{\gamma})(y - x)^\top \nabla^2 f(x + \bar{\gamma}(y - x))(y - x)\,. && \text{(事实 1.12)}
\end{aligned}$$

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$\bar{\gamma}$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 1.9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以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已知优化问题，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]，

$$\min_{x\in D} c^\top x = \min_{x\in \mathrm{conv}(D)} c^\top x, \tag{1.4}$$

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(1.4)式的左边至少不小于右边. 用$\Delta_k$表示$k-1$维标准单纯形. 下面证明另外一个方向. 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

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

后者是在凸集上极小化线性函数，所以是凸问题.

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，如果$0 \in \partial f(x)$，那么x是f在Ω上的全局极小点.

[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.

Proof. 若$0 \in \partial f(x)$，由次梯度的定义，对于所有的$y \in \Omega$有$f(x) - f(y) \leqslant 0$. 由此知x是f在Ω上的全局极小点.

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. ■

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，并且大部分教材回避了它. 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 $\widehat{x}$ 满足

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

其中$\epsilon > 0$是某个正误差. 算法的代价与目标误差有关. 高度实用算法的代价通常和ϵ的多项式有关，诸如$O(1/\epsilon)$或者甚至$O(1/\epsilon^2)$. 一些算法理论上达到$O(\log(1/\epsilon))$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.

该课程尝试突出算法的理论性能和实际魅力. 此外，还将讨论诸如对噪声的鲁棒性等其他性能准则. 要确定一个算法有多好，这通常与碰到的应用有关，由单个准则几乎是不可能完成的.

2 梯度法

本节学习至关重要的基础“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. 先对目标函数(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

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

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不一定是最理想的.

2.2 光滑函数

将要遇到的第一个性质称作光滑性(smoothness). 光滑性的要点是能控制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 **2.1** (光滑性)**.** 称连续可微函数 f是$\beta-$光滑的，如果梯度$\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n$ 是β-Lipschitz连续的，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 **2.2.** 设函数 f在$\mathbb{R}^n$上是β-光滑的. 那么，对每个$x, y \in \mathbb{R}^n$,

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

证明. 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，

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

■

该引理的意义在于：选择

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

能得到

$$f(y) - f(x) \leqslant -\frac{1}{2\beta}\|\nabla f(x)\|^2. \tag{2.1}$$

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

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

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

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

这里的第一个不等式同时利用了梯度不等式和引理 2.2. 因此，由凸性和光滑性可以得到不等式. ■

将证明更新规则

$$x_{t+1} = x_t - \eta \nabla f(x_t), \tag{2.2}$$

即梯度下降法在$\beta-$光滑的条件下，能达到的收敛速率. 利用上面两个引理可以证明如下结论.

定理 **2.4** (针对光滑函数的梯度法)**.** 设f是$\mathbb{R}^n$上的$\beta-$光滑凸函数. 那么$\eta = 1/\beta$ 的梯度下降法 (2.2)满足

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

证明. 由更新规则 (2.2)和 引理 2.2 有

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

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

$$\delta_{t+1} \leqslant \delta_t - \frac{1}{2\beta}\|\nabla f(x_t)\|^2.$$

由凸性也有

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

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

$$\delta_{t+1} \leqslant \delta_t - \frac{1}{2\beta\|x_1 - x_*\|^2}\delta_t^2. \tag{2.3}$$

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

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

请注意，这里利用了$\{\delta_t\}$的单调递减性质. 为了结束证明，还需要证$\|x_t - x_*\|$关于t是递减的. 由引理 2.3, 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$\nabla f(x_*) = 0$得到

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

■

2.3 投影

当约束集 Ω不是整个 $\mathbb{R}^n$时，梯度更新可能跳出可行域 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 **2.5** (投影)**.** 点 y在集合Ω上的投影 **(projection)**定义为

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

例子 **2.6.** 在欧氏单位球B_2上的投影恰好是规范化:

$$\Pi_{B_2}(y) = \begin{cases} \frac{y}{\|y\|}, & \|y\|_2 > 1, \\ y, & \|y\|_2 \leqslant 1. \end{cases}$$

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

定理 **2.7** (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

$$\langle y - \pi, x - \pi\rangle \leqslant 0, \quad \forall x \in \Omega. \tag{2.4}$$

证明. 存在性：由$\Omega \neq \emptyset$，所以存在$\bar{x} \in \Omega$. 从而

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

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存在.

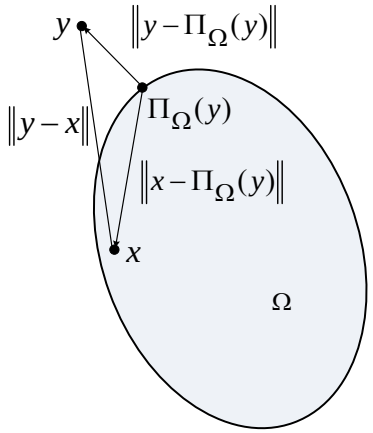


图 2.1: 投影的刻画与性质.

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

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

将右边展开，得

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

这等价于

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

因为对所有的$\theta \in (0,1)$，上式都成立，令$\theta \to 0+0$得到式(2.4).

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$x = \pi_2$，第二个不等式中取$x = \pi_1$，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与y在凸集上的投影比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 **2.8** (投影的压缩性)**.**

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

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

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

图 2.2: 投影梯度下降法

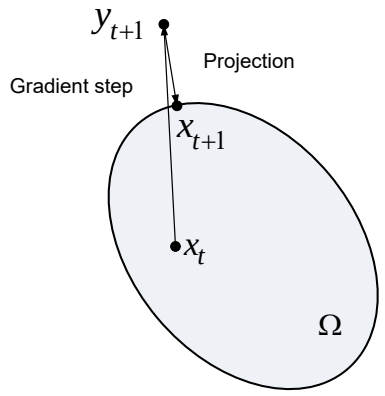


图 2.3: 投影梯度法.

2.4 次梯度投影法

因此，原来的步骤经修正后如 图 2.3所示，由投影步保证$x_{t+1} \in \Omega$. 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 2.6). 在后面的讲义中将会看到几个其它非平凡的例子.

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.

定义 **2.9** (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 continuous)连续的.

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

命题 **2.10.** 设$\mathcal{C} \subseteq \mathbb{R}^n$是开凸集，并且$f : \mathbb{R}^n \to \mathbb{R}$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不等式上控不等式的右边，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$\|g\| \leqslant L$. ■

现在能够证明梯度下降法的第一个收敛速率.

定理 **2.11** (投影梯度法的复杂性)**.** 假设$\Omega \subset \mathbb{R}^n$是闭凸集，并且$f$在在包含$\Omega$的开集上是$L$-Lipschitz 连续的凸函数 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$x_1, \cdots, x_t$是投影梯度法产生的$t$步迭代序列，其中步长$\eta_s = \frac{R}{L\sqrt{t}}$. 那么

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

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:

$$u^\top v = \tfrac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2). \tag{2.5}$$

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

定理 **2.11**的证明**.** 首先确定函数值之差$f(x_s) - f(x_*)$的上界.

$$\begin{aligned}
f(x_s) - f(x_*) &\leqslant g_s^\top (x_s - x_*) && (g_s \in \partial f(x_s)\text{和梯度不等式})\\
&= \tfrac{1}{\eta}(x_s - y_{s+1})^\top (x_s - x_*) && (\text{更新规则})\\
&= \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x_*\|^2\right) && (\text{(2.5)})\\
&= \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 - \|y_{s+1} - x_*\|^2\right) + \tfrac{\eta}{2}\|g_s\|^2 && (\text{更新规则})\\
&\leqslant \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 - \|y_{s+1} - x_*\|^2\right) + \tfrac{\eta L^2}{2} && (\text{Lipschitz条件})\\
&\leqslant \tfrac{1}{2\eta}\left(\|x_s - x_*\|^2 - \|x_{s+1} - x_*\|^2\right) + \tfrac{\eta L^2}{2} && (\text{引理 2.8})
\end{aligned}$$

现在，将这些函数值之差的不等式从 $s=1$到$s=t$求和:

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

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

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

■

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，强凸性将提供一个二次下界. 当将这两个假设放在一起时，能得到线性收敛速率，因此这两个假设相当强大. 非正式地，对光滑的强凸函数，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，如果对于每个已知的x，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，梯度∇f扮演着g的角色. 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:

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

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”.

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连续的概念. 称函数f是L-Lipschitz连续的，如果它“stretches”自己输入的大小由L上控决定:

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

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

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.

上述结论未对极小点的非退化性作假设，比如极小点可以不唯一，并且f的图形在X_*附近“有可能非常扁平”. 在额外的关于f强凸的假设下，能得到更好的收敛速率的结论.

定义 **3.1.** 称函数$f : \mathbb{R}^n \to \mathbb{R}$是$\alpha-$强凸的，如果

$$f((1-\theta)x_0 + \theta x_1) \leqslant (1-\theta)f(x_0) + \theta f(x_1) - \frac{\alpha}{2}\theta(1-\theta)\|x_0 - x_1\|_2^2,$$

其中$\alpha > 0$是强凸参数.

留做习题：f是$-\alpha$强凸的当且仅当$f(x) - \frac{\alpha}{2}\|x\|^2$是凸的. 下面是强凸函数的判别.

命题 **3.2.** 函数$f : \Omega \to \mathbb{R}$ 是α**-**强凸的**(α-strongly convex)**，那么有

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

设$f : \mathbb{R}^n \to \mathbb{R}$连续可微. 那么$f$是$\alpha$-强凸当且仅当

$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2}\|y - x\|^2 \ \ \forall x, y \in \Omega, \tag{3.1}$$

或者

$$[x - y]^T [\nabla f(x) - \nabla f(y)] \geqslant \alpha \|x - y\|_2^2 \ \ \forall x, y \in \Omega. \tag{3.2}$$

设$f : \mathbb{R}^n \to \mathbb{R}$二次连续可微. 那么$f$是$\alpha$-强凸当且仅当

$$\nabla^2 f(x) \succeq \alpha I \ \ \forall x \in \Omega..$$

易于证明可微函数强凸的充分必要条件(3.1)和(3.2)是等价的. 此外，对于可微的强凸函数，假设x_*是稳定点，由(3.1) 得到

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

这时，由函数值与最优值之间的误差可以控制点列误差.

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- 强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.

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

特别地，当A是对称正定矩阵时，二次函数

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

在整个空间是$\alpha = \lambda_{\min}(A)$强凸，$\beta = \lambda_{\max}(A)$光滑的，该函数的条件数$\kappa = \lambda_{\max}/\lambda_{\min}$与基于二次函数的Hessian阵的条件数是一致的，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: 将梯度法应用到各种函数类时，作为步数t的函数所得到的误差ϵ的上界.

	凸	强凸
Lipschitz连续	$\epsilon \leqslant O(1/\sqrt{t})$(定理 2.11)	$\epsilon \leqslant O(1/t)$ (定理 3.4)
光滑	$\epsilon \leqslant O(1/t)$(定理 2.4)	$\epsilon \leqslant e^{-\Omega(t)}$(定理 3.5)

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 强凸函数

从Lipschitz连续的强凸函数的收敛界开始，这种情况下也能得到收敛速率$O(1/t)$.

定理 **3.4** (针对强凸函数投影梯度法的复杂性)**.** 设$\Omega \subset \mathbb{R}^n$是闭凸集. 假设 f在包含Ω的开集上是α-强凸的. 设 x_* 是f 在Ω上的最小点, 并设x_s 是使用投影次梯度法在步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

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

定理蕴含着投影次梯度法对α-强凸函数的收敛速率与对β-光滑函数的类似，均为 $\epsilon \leqslant O(1/t)$. 为了证明 定理 3.4, 需要Jensen不等式(命题 1.7).

定理 3.4 的证明. 回忆投影次梯度的两步更新规则

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

[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$f(n) = \Omega(g(n))$.

首先，从探索函数值$f(x_s)$和$f(x_*)$之差开始.

$$
\begin{aligned}
& f(x_s)-f(x_*) \\
\leqslant\ & g_s^{\top}(x_s-x_*)-\tfrac{\alpha}{2}\|x_s-x_*\|^2 && \text{(命题 3.2)} \\
=\ & \tfrac{1}{\eta_s}(x_s-y_{s+1})^{\top}(x_s-x_*)-\tfrac{\alpha}{2}\|x_s-x_*\|^2 && \text{(更新规则)} \\
=\ & \tfrac{1}{2\eta_s}(\|x_s-x_*\|^2+\|x_s-y_{s+1}\|^2-\|y_{s+1}-x_*\|^2)-\tfrac{\alpha}{2}\|x_s-x_*\|^2 \\
& && \text{("最优化基本定理")} \\
=\ & \tfrac{1}{2\eta_s}(\|x_s-x_*\|^2-\|y_{s+1}-x_*\|^2)+\tfrac{\eta_s}{2}\|g_s\|^2-\tfrac{\alpha}{2}\|x_s-x_*\|^2 && \text{(更新规则)} \\
\leqslant\ & \tfrac{1}{2\eta_s}(\|x_s-x_*\|^2-\|x_{s+1}-x_*\|^2)+\tfrac{\eta_s}{2}\|g_s\|^2-\tfrac{\alpha}{2}\|x_s-x_*\|^2 && \text{(引理 2.8)} \\
\leqslant\ & \left(\tfrac{1}{2\eta_s}-\tfrac{\alpha}{2}\right)\|x_s-x_*\|^2-\tfrac{1}{2\eta_s}\|x_{s+1}-x_*\|^2+\tfrac{\eta_s L^2}{2} && \text{(Lipschitz连续性)}
\end{aligned}
$$

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

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

最后，能够找到定理 3.4中显示的由t步投影次梯度法得到的函数值的上界：

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

由此断定，用投影梯度下降法求解具有强凸性质的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连续性的凸函数的相比，这个更快. ■

3.4 光滑强凸函数

定理 **3.5** (针对光滑强凸函数梯度法的复杂性)**.** 假设$f\colon \mathbb{R}^n\to\mathbb{R}$是$\alpha$-强凸和 β-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 (2.2)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 3.5, 需要使用如下引理. 请注意 (3.3)与 (2.1)的联系与区别.

引理 **3.6.** 假设 f如定理 3.5中所描述. 那么 $\forall x,y\in\mathbb{R}^n$ 和更新方式

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

有

$$
f(x^+)-f(y)\leqslant\nabla f(x)^{\top}(x-y)-\tfrac{1}{2\beta}\|\nabla f(x)\|^2-\tfrac{\alpha}{2}\|x-y\|^2. \tag{3.3}
$$

[3]请注意，这里是点列收敛.

引理 **3.6** 的证明.

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

■

现在用引理 3.6能够证明定理 3.5.

定理 **3.5** 的证明.

$$\begin{aligned}
\|x_{t+1} - x_*\|^2 &= \|x_t - \tfrac{1}{\beta}\nabla f(x_t) - x_*\|^2 \\
&= \|x_t - x_*\|^2 - \tfrac{2}{\beta}\nabla f(x_t)^\top (x_t - x_*) + \tfrac{1}{\beta^2}\|\nabla f(x_t)\|^2 \\
&\leqslant \left(1 - \tfrac{\alpha}{\beta}\right)\|x_t - x_*\|^2 && (\text{使用引理 3.6，其中} y = x_*, x = x_t) \\
&\leqslant \left(1 - \tfrac{\alpha}{\beta}\right)^t \|x_1 - x_*\|^2 \\
&\leqslant \exp\left(-t\tfrac{\alpha}{\beta}\right)\|x_1 - x_*\|^2
\end{aligned}$$

■

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 **3.7** (光滑强凸函数投影梯度法的复杂性)**.** 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$\frac{1}{\beta}$时投影梯度法在步骤$t$ 的更新点，即使用更新规则

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

其中Π_Ω 是投影算子. 那么

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

像定理 3.5那样, 为了证明 定理 3.7， 需要使用类似于引理 3.6的如下结论.

引理 **3.8.** 假设f如定理 3.7 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

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

引理 **3.8** 的证明**.** 对于引理 3.8 中所定义的x, x^+和y，由投影定理(定理 2.7)，得

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

因此， 与引理 3.6的证明类似，有

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

■

再用投影梯度下降更新代替标准梯度的下降更新后， 用引理 3.8代替 引理 3.6，那么定理 3.7 的证明与定理 3.5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

前面分析了投影梯度法的收敛性，并且证明了(定理 2.11)：考虑闭凸集Ω上的$L-$Lipschitz连续凸函数f，当优化步长$\eta_s = R/(L\sqrt{t})$ 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$\sqrt{n/t}$收敛到$f$ 在$B_{2,n}$上的极小值. 当使用本节介绍的镜像下降法，得到的收敛速率是$\sqrt{\ln n/t}$.

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.)

现在想在一个非欧空间(特别是ℓ_1空间)里执行梯度法. 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 由梯度法到临近梯度法

先来看梯度法和投影梯度法的临近梯度法理解，为此考虑目标函数是可微凸函数与凸函数之和的优化问题：

$$\min_{x\in X} F(x) := f(x) + h(x), \tag{5.1}$$

其中f是凸的可微函数，并且$\mathrm{dom} f = \mathbb{R}^n$, h是凸函数，X是欧氏空间E中的闭凸子集[4]，为了分析方便，下面假设$X = \mathbb{R}^n$.

比如$F(x) = \frac{1}{2}\|Ax - b\|^2 + \tau\|x\|_1$，其中$A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$已知，$\tau > 0$是参数；如果令$h(x) = 0 \forall x \in \mathbb{R}^n$，能看到无约束优化问题是该问题的特例; 如果令

$$h(x) = \delta_\Omega(x) := \begin{cases} 0, & \text{如果 } x \in \Omega \\ +\infty, & \text{否则,} \end{cases}$$

其中$\Omega \subset \mathbb{R}^n$是闭凸集. 凸分析中称$\delta_\Omega$是集合$\Omega$的指示函数. 当$\Omega$是凸集时，易见其指示函数也是凸函数. 这时(5.1)就还原成集合Ω上极小化$f(x)$.

梯度法和投影梯度法是一种临近梯度法(Proximal gradient method): 可将x_{s+1}写作

$$\begin{aligned} x_{s+1} &= \Pi_\Omega(x_s - \alpha_s \nabla f(x_s)) \\ &= \operatorname*{argmin}_{x\in\Omega} \frac{1}{2\alpha_s}\|x - (x_s - \alpha_s \nabla f(x_s))\|_2^2 \\ &= \operatorname*{argmin}_{x\in\Omega} f(x_s) + \nabla f(x_s)^\top (x - x_s) + \tfrac{1}{2\alpha_s}\|x - x_s\|_2^2 \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^n} f(x_s) + \nabla f(x_s)^\top (x - x_s) + \tfrac{1}{2\alpha_s}\|x - x_s\|_2^2 + \delta_\Omega(x). \end{aligned}$$

将δ_Ω推广到一般的凸函数h，得到

$$x_{s+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n} f(x_s) + \nabla f(x_s)^\top (x - x_s) + \tfrac{1}{2\alpha_s}\|x - x_s\|_2^2 + h(x). \tag{5.2}$$

这就是求解(5.1)的临近梯度法(proximal gradient mathod)：新的迭代点极小化h与f在x_s附近的一个简单的二次局部模型函数之和.

具体地，在求解(5.1)时，每一步利用f的可微性，用f在x_s的一阶Taylor展式代替f. 如果函数是线性的，只需该线性近似项即可. 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$\|x - x_s\|_2^2$. 这里的步长决定了惩罚力度. 惩罚强迫x_{s+1}靠近x_s以便f的线性近似是精确的； 惩罚项也可以看作f在x_s的线性近似的误差项，从而新的迭代点极小化$h(x)$加上f在x_s附近的一个简单局部二次模型. 在(5.2)中取$h(x) = 0 \forall x \in \mathbb{R}^n$，即得最速下降法，取$h(x)$是凸集$\Omega$的指示函数，得到投影梯度法. 因此，临近梯度法是这两种方法的推广.

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.

5.2 Bregman散度

现在，想将$\frac{1}{2}\|x - x_s\|_2^2$用一个广义距离对应的量来替换，有两个潜在好处：一是使用f的一个在x_s附近更精确的模型或者可用另外的范数代替2-范数，希望获得限

[4]Nesterov的原始表示使用了集合约束$x \in X$和临近设置$(\omega, \|\cdot\|)$，其中ω是可微且强凸的，是距离生成函数

制性更少的Lipschitz 常数，由此有可能减少迭代次数；二是使得子问题更易于计算，由此减少每步的计算成本.

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$x^\top y \leqslant \|x\|\|y\|_*$. 用同样的思考方式，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 **5.1** (Bregman散度,1967)**.** 已知凸集$\mathcal{D}$上的可微函数Φ，假设它关于$\|\cdot\|$是$\alpha-$强凸的. 对$x, y \in \mathcal{D}$，定义

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

称它是与Φ关联的Bregman散度，这里x是临近中心. 称Φ是势函数，或者距离生成函数.

该散度表示Φ与Φ在x处的线性近似之差. 由强凸函数的二次上界，得

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

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.

例子 **5.2** (欧氏设置)**.** $\Phi(x) = \|x\|_2^2, \|\cdot\| = \|\cdot\|_2, \mathcal{D} = \mathbb{R}^n, \nabla\Phi(x) = 2x$. 这里

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

例子 **5.3** (ℓ_1设置)**.** $\|\cdot\| = \|\cdot\|_1, \mathcal{D} = \mathbb{R}^n_{++}, \Phi(x)$ 为负熵，即

$$\Phi(x) = \sum_{i=1}^n x_i \ln(x_i), \qquad \nabla\Phi(x) = (1 + \ln(x_1), \cdots, 1 + \ln(x_n)).$$

负熵生成的Bregman散度

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

这是x与y 间的Kullback-Leibler 散度(KL)，当$x, y \in \Delta_n$时，即为x与y 之间的相对熵(交叉熵与熵之差).

更多Bregman散度的例子参见作业.

命题 **5.4** (三点性质)**.** 已知凸集$\mathcal{D}$上的可微凸函数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. 由Bregman散度的定义，

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

■

定义 **5.5.** 如果$\|\cdot\|$为$\mathbb{R}^n$上的某种范数，其对偶范数记作$\|\cdot\|_*$，定义为

$$\|y\|_* = \max_{\|x\|\leqslant 1} y^T x.$$

由对偶范数的定义，有如下Cauchy-Schwartz不等式的推广：

$$|y^T x| \leqslant \|x\|\|y\|_* \quad \forall x, y \in \mathbb{R}^n. \tag{5.3}$$

例子 **5.6.** 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，$q = +\infty$. 这时，(5.3)还原成Hölder 不等式.

定义 **5.7** (Bregman投影)**.** 设Φ为$\mathcal{D} \subseteq \mathbb{R}^n$ 上的可微强凸函数. 定义$y \in \mathcal{D}$ 在Ω上关于Φ 的Bregman 投影为

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

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$y \in \mathcal{D}$. 那么

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

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

证明. 记$\pi := \Pi_\Omega^\Phi(y)$，并定义$h(t) = D_\Phi(\pi + t(x-\pi), y)$. 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

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

根据Bregman散度的定义，

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

因此，

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

再由三点性质，知道

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

因为$D_\Phi(\pi, y) \geqslant 0$，所以$D_\Phi(x, \pi) \leqslant D_\Phi(x, y)$. ■

5.3 用Bregman散度正则化：镜像下降法

如前所述，镜像下降算法是一种广义临近梯度法，即

$$x_{s+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n} f(x_s) + g_s^\top (x - x_s) + h(x) + \frac{1}{\eta_s} D_\Phi(x, x_s). \tag{5.4}$$

为了更具体深刻地理解镜像下降法，本小节假设$h(x) = \delta_\Omega(x)$，此时，这个子问题还原成Bregman投影，方法的的每一步中都应用Bregman散度将更新点投影到约束集上，对应的镜像镜像下降算法的正式描述见算法1，几何直观如 图5.1.

Algorithm 1 Mirror Descent algorithm

Require: $x_1 \in \arg\min_{\Omega\cap\mathcal{D}} \Phi(x)$, constant $\eta > 0$.

1: **for** $s = 1$ to $t-1$ **do**
2: $\quad \nabla\Phi(y_{s+1}) = \nabla\Phi(x_s) - \eta g_s$ 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 $x^\circ \in \operatorname*{argmin}_{x\in\{x_1,\cdots,x_t\}} f(x)$

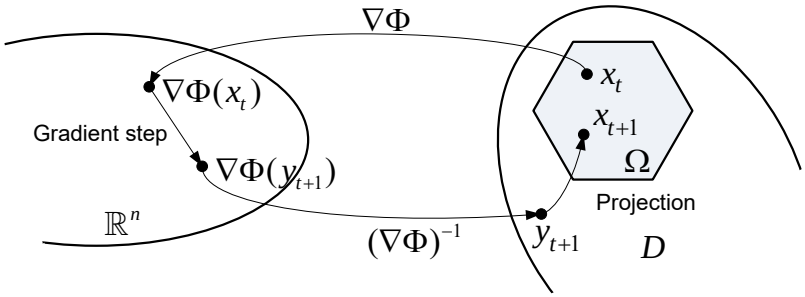


图 5.1: 镜像下降法.

在投影梯度法的复杂性证明中用到了欧氏范数如下的性质((2.5)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 5.4)表明Bregman 散度本质上表现为投影梯度法中欧氏范数的平方. 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)，则当$\eta = \frac{R}{L}\sqrt{\frac{2\alpha}{t}}$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$x_* \in \Omega \cap \mathcal{D}$. 与投影梯度法的证明类似，有：

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

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)直接由镜像下降算法得到. 上式中的(iii)根据三点性质(命题 5.4)得到. 对于不等式(iv)，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$x_* \in \Omega \cap \mathcal{D}$，有$D_\Phi(x_*, y_{s+1}) \geqslant D_\Phi(x_*, x_{s+1})$. 下面证明不等式(v).

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

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，$\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}) = \eta g_s$. 再利用不等式(5.3)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

$$\frac{1}{t}\sum_{s=1}^{t}[f(x_s) - f(x_*)] \leqslant \frac{\eta L^2}{2\alpha} + \frac{D_\Phi(x_*, x_1)}{t\eta}. \tag{5.5}$$

根据Bregman散度定义得到

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

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$\nabla\Phi(x_1)(x_* - x_1) \geqslant 0$. 再将不等式代入(5.5)，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

$$\frac{1}{t}\sum_{s=1}^{t}[f(x_s) - f(x_*)] \leqslant RL\sqrt{\frac{2}{\alpha t}}.$$

从而得到所需证明的结论. ■

由例5.2知道，用欧氏距离的平方作为势函数，即$\Phi(x) = \|x\|_2^2$，得到的Bregman散度$D_\Phi(x,y) = \|x-y\|_2^2$. 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

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

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法，得到的误差上界为$\frac{LR}{\sqrt{t}}$. 这恰好是定理 2.11的结论.

为了叙述方便，以下记

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

命题 **5.10.** 已知$y \in \mathbb{R}^n_{++}$. 则y在单纯形$\triangle_n$上关于KL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，给出这个Bregman投影子问题的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$\Pi^{\Phi}_{\Delta_n}(y) = \frac{y}{\|y\|_1}$.

有了以上准备，下面可以看一个异于梯度投影法的镜像下降法的完整例子.

例子 **5.11** (ℓ_1设置续)**.** 对于负熵$\Phi(x)$，由它确定的Bregman散度

$$D_\Phi(x,y)\sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{y_i}\right) - \sum_{i=1}^{n}(x_i - y_i).$$

此外，可以证明Φ在Δ_n 上关于$\|\cdot\|_1$是1−强凸的. 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.

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

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

和负熵的梯度，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

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

再由命题 5.10中KL散度的投影刻画，得

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

称该设置为指数梯度下降或者带有乘性权重的镜像下降. 综上，具有这种更新和投影的镜像下降算法为：

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

为了分析收敛速率，要研究Δ_n上的ℓ_1范数.

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$\eta = \frac{1}{L}\sqrt{\frac{2\ln n}{t}}$ 的镜像下降法得到

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

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，因为$f_j \in \{-1, 1\}$，所以$\|F\|_\infty \leqslant 1$. Boosting是一种集成学习方法，与之对应的经验风险极小化为

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

其中φ是$L-$Lipschitz连续的一元损失函数. 考虑用梯度下降法求解 (Boosting). 易见

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

是目标函数的一个次梯度.

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ℓ_2 范数的上界：

$$\|g\|_2 \leqslant \frac{L}{m}\sum_{i=1}^{m} \|F(X_i)\| \leqslant L\sqrt{M}.$$

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$RL/\sqrt{t}$ 变成$2L\sqrt{M}/\sqrt{t}$. 为了把误差控制在$1/m$以内，则需要$t \sim m^2 M$，由于M很大，从而投影梯度法性能很差. 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$t \sim m^2 \ln M$.

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). 从而如果利用距离生成函数Φ是负熵函数，得到

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

为了误差小于$1/m$，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$\sqrt{M}$. 这就是投影梯度法的速率为$\sqrt{M/t}$ 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$\sup -$ 范数给出其上界，而$\sup -$ 范数通常是投影梯度法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

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

则Φ关于ℓ_1范数是$c\sqrt{\ln n}-$强凸的.

6 条件梯度法

这讲讨论条件梯度法，也称作Frank-Wolfe (FW)算法 [FW56]. 方法的动机是在有些场景下，当投影梯度法中的投影步无法有效计算时，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. 考虑从某个点 $x_0 \in \Omega$ 出发, 置

$$x_{t+1} = x_t + \eta_t(y_t - x_t),$$

其中$\eta_t \in (0, 1]$,

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

该表达式简化成:

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

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，对应的几何直观如图6.2.

Starting from $x_0 \in \Omega$, repeat:

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

$$\eta_t \in \arg\min_{\eta\in[0,1]} f(x_t + \eta(y_t - x_t)), \qquad \text{(线搜索)}$$

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

图 6.1: 条件梯度法

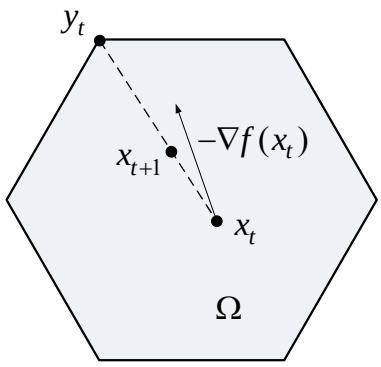


图 6.2: 条件梯度/Frank-Wolfe法的几何直观.

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$\Omega \subseteq \mathbb{R}^n$是闭凸集，$f\colon \Omega \to \mathbb{R}$是$\beta$-光滑的凸函数，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法满足

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

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

证明. 由光滑性知引理 2.2成立，从而

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

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:

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

根据定理 6.1中D的定义，并且观测到$\|y_t - x_t\|^2 \leqslant D^2$. 再由$y_t$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，也有梯度不等式

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

将此不等式代入上式，

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

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

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

从而，归纳假设对于基本情况成立.

归纳步. 按归纳法，假设不等式 (6.1)对所有不超过t 的正整数成立，下面证明 (6.1)对$t + 1$也成立. 由 (6.2),

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

这样，不等式 (6.1)对$t + 1$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 核范数球上的投影

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

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

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 rho并在0处截断使得平移截断值之和等于 1，这里的ρ需要求一个分段线性函数的零点. 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

$$\min_{X\in B_*} f(X) := \tfrac{1}{2}\|Y - P_O(X)\|_F^2, \tag{6.3}$$

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X中与观测值对应的元素保持作X的对应元素，其它元素补 0 ， 可显式表示为$P_O(X) = X \odot O$，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

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

可以使用投影梯度法求解问题(6.3)，但是使用Frank-Wolfe法会更有效，后者需要求解线性最优化oracle

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

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：

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

再结合凹函数在某集合上的最小值等于它在该集合凸包上的最小值的事实，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

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

它的另一种理解方式：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 $y_1 = A^\top x_1 / \|A^\top x_1\|$.
- From $s = 1$ to $t - 1$:
  - Put $x_{s+1} = \frac{Ay_s}{\|Ay_s\|}$
  - Put $y_{s+1} = \frac{A^\top x_{s+1}}{\|A^\top x_{s+1}\|}$
- 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的凸函数，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:

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

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

$$f(x) = \tfrac{1}{2} a x^2 - bx + c$$

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:

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

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.

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$x_* = A^{-1}b$，即x_*是线性方程组

$$Ax - b = 0$$

的解. 当$A \succ 0$时, 二次函数是严格凸的(strictly convex) 并且这个临界点是唯一的全局极小点.

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 3.7 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速率. 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$A \in \mathbb{R}^{m\times n}$. 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 3.5 知步长为$\frac{1}{\beta}$的梯度下降法以

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

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(QF). 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:

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

证明. 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，

$$\|x_{t+1} - x_*\|_2 \leqslant \left\| \prod_{s=1}^{t} (I - \eta_s A) \right\|_2 \|x_1 - x_*\|_2 . \tag{7.1}$$

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步长的选取, 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. 注意到由(7.1)，有

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

其中

$$p_t(A) := \prod_{s=1}^{t} (I - \eta_s A) , \|p_t(A)\|_2 = \max_{a \in \Lambda(A)} |p_t(a)| .$$

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 $p_t(0) = 1$. (保证多项式的常数项是 1)

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 $t = 3$和$t = 6, \alpha = 1$，$\beta = 10$ 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.

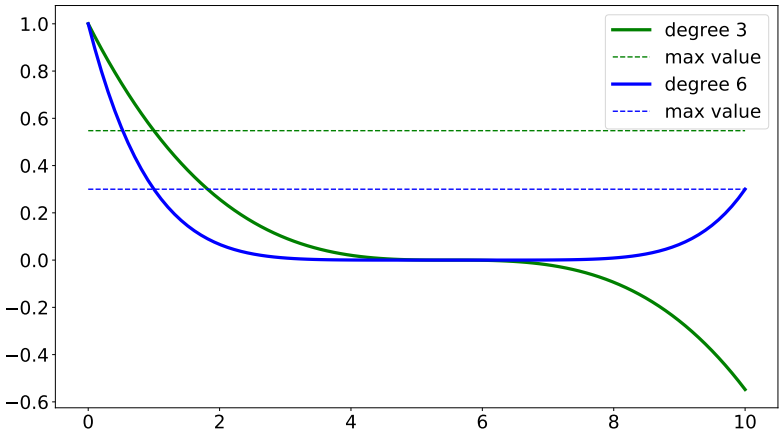


图 7.1: $t = 3$和$t = 6$的朴素多项式$p_t(1 - \eta a)^t$，其中$\alpha = 1, \beta = 10$.

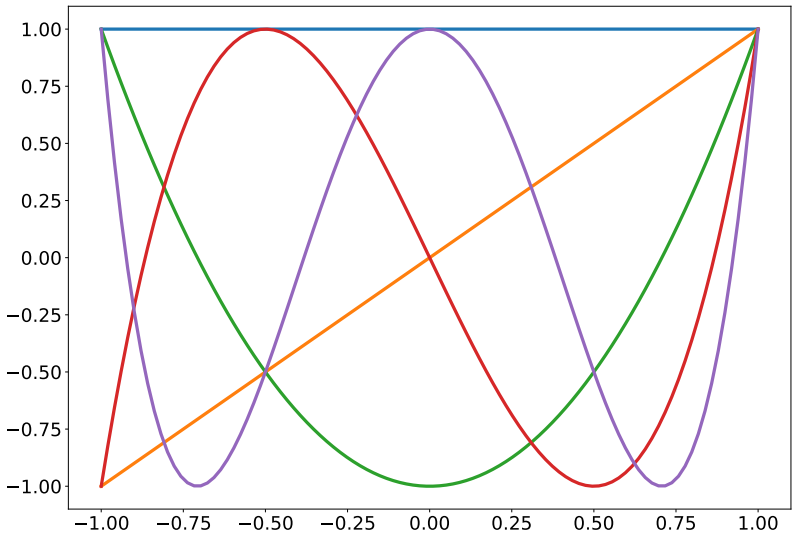


图 7.2: 前5个Chebyshev多项式：$T_0, T_1, \cdots, T_4$.

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:

$$T_0(a) = 1, \quad T_1(a) = a \tag{7.2}$$

$$T_{t+1}(a) = 2aT_t(a) - T_{t-1}(a), \text{ 当} t \geqslant 1\,. \tag{7.3}$$

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$[-1, 1]$.

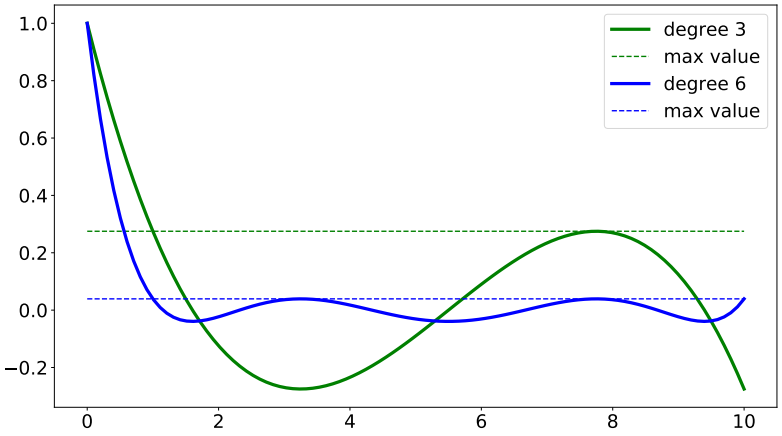


图 7.3: 重伸缩后的 3 次和 6 次Chebyshev多项式，其中$\alpha = 1, \beta = 10$.

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:

$$P_t(a) = \frac{T_t\left(\frac{\beta+\alpha-2a}{\beta-\alpha}\right)}{T_t\left(\frac{\beta+\alpha}{\beta-\alpha}\right)}. \tag{7.4}$$

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.

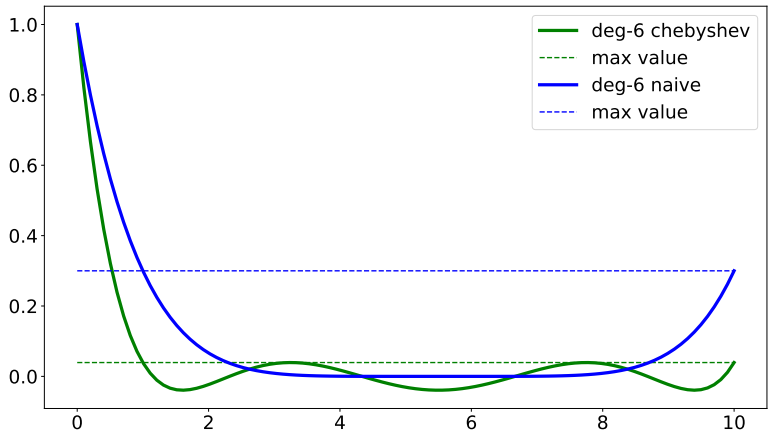


图 7.4: 重伸缩的Chebyshev多项式与朴素多项式

为此，思考多项式在区间$[\alpha,\beta]$上的模有多大？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:

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

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

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

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:

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

那么，对等的可以为算法的误差确定上界，所以有：

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

如此表明由Chebyshev多项式可以获得的误差界:

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

这意味着对于大的κ， 在误差指数地减小之前所需要的迭代次数获得二次节省. 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.

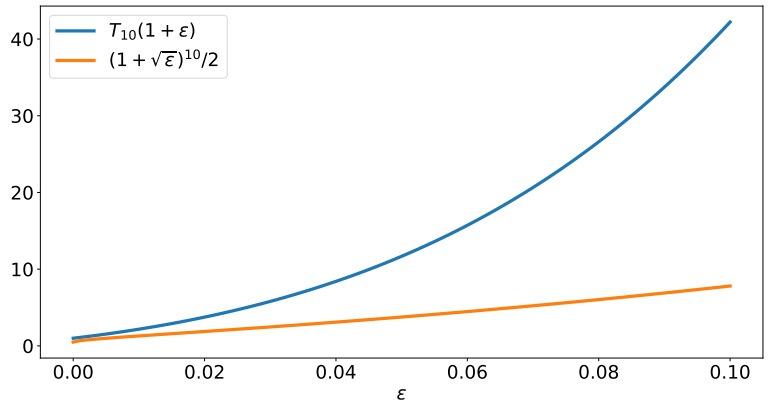


图 7.5: 朴素多项式和Chebyshev多项式的收敛

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(7.2)转换成重伸缩后的切比雪夫多项式(7.4)，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可以用递归定义算出来. 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. 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.

7.5 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 **7.4** (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 **7.5** (与多项式的联系)**.** Krylov子空间满足

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

其中$\deg(p)$表示多项式p的次数.

证明. 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, 假设对称正定矩阵 $A \in \mathbb{R}^{n\times n}$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

$$b = \alpha_1 u_1 + \cdots + \alpha_n u_n,$$

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

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

8 共轭梯度法

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——共轭梯度法.

8.1 共轭与椭球范数

已知 $A \in S_{++}^n$, $b \in \mathbb{R}^n$, 考虑二次函数$f(x) = \frac{1}{2}x^T Ax - b^T x$. 那么

$$\nabla f(x) = Ax - b,$$

并且f有唯一极小点 $x_* = A^{-1}b$. 易于看到，$\min_{x\in\mathbb{R}^n} f(x)$ 等价于 $Ax = b$.

定义 **8.1** (内积)**.** 已知正定矩阵A. 定义两个向量的内积为

$$\langle u, v\rangle_A = u^T Av \quad \forall u, v \in \mathbb{R}^n.$$

定义 **8.2** (椭球范数)**.** 由上述内积的定义，可以诱导出向量的椭球范数

$$\|u\|_A = \sqrt{u^T Au}.$$

对于二次函数(QF)，易于验证

$$f(x) - f(x_*) = \frac{1}{2}(x - x_*)^T A(x - x_*) = \frac{1}{2}\|(x - x_*)\|_A^2.$$

下面给出向量正交概念的推广.

定义 **8.3** (向量共轭)**.** 已知向量 $u, v \in \mathbb{R}^n$. 如果 $\langle u, v\rangle_A = u^T Av = 0$， 称$u$与$v$关于$A$是共轭的.

定义 **8.4.** 向量 $d_1, \cdots, d_k$ 关于矩阵 A 共轭, 如果 $d_i^T Ad_i = 0, \forall i \neq j$, 即任意两个互不相同的向量共轭.

事实 **8.5.** 易见当$A = I$时，这里的共轭概念就还原成向量正交的概念. 类似于正交向量组的性质，可以证明不含零的共轭向量组必线性无关.

8.2 共轭方向法

定义 **8.6** (共轭方向法)**.** 当把方法应用于以 A 为 Hessian 阵的二次函数(QF)时, 若方法产生的搜索方向关于 A 是共轭的, 则称该方法是共轭方向法**(Conjugate direction method)**.

线搜索法是一种重要的优化算法格式. 已知$x_0 \in \mathbb{R}^n, t = 0$. 更新

$$x_{t+1} = x_t + \eta_t d_t,$$

其中 d_t 是x_t 处的搜索方向, η_t 是步长. 精确步长指

$$\eta_t = \underset{\eta\in\mathbb{R}}{\operatorname{argmin}} f(x_t + \eta d_t).$$

求精确步长是关于一元函数的优化问题. 当目标函数函数可微时，由极值点的一阶必要条件，得到

$$r_{t+1}^T d_t = 0,$$

其中 $r_{t+1} = \nabla f(x_t + \eta_t d_t)$.

对于强凸二次函数, 上述步长具有解析表达式：$\eta_t = \frac{-r_t^T d_t}{d_t^T Ad_t}$, 其中 $r_t = \nabla f(x_t)$.

定理 **8.7** (共轭方向法的性质)**.** 应用精确线搜索法的共轭方向法 $(d_0, d_1, \cdots, d_{n-1}$ 关于 A 共轭$)$极小化 *Hessian* 阵为 A 的二次函数(QF), 设 $x_0 = 0$, 若$r_{t+1} \neq 0$，则

(i) r_{t+1} 与 $d_0, d_1, \cdots, d_t$ 正交;

(ii) x_{t+1} 是 $f(x)$ 在方向 $d_0, d_1, \cdots, d_t$ 生成的线性子空间 $\text{span}\{d_0, d_1, \cdots, d_t\}$ 上的极小点.

推论 **8.8.** 应用精确线搜索法的共轭方向法极小化 *Hessian* 阵为 A 的 n 元二次函数, 则方法至多迭代 n 步后终止于函数最小点.

8.3 由Gram-Schmidt过程构造共轭的搜索方向

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.

该方法由Hestense 与 Stiefel (1952年)提出，当时用于求解系数矩阵对称正定的大规模线性方程组的方法; 后来由 Fletcher-Reeves (1964 年), Polak 与 Ribiere (1969年)等将它推广求解二次函数的极小化问题

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

偏微分方程数值解, 信号处理, 参数估计和优化方法等相关计算问题中经常用该方法, 通常使用预条件共轭梯度法(PCG).

使用 Gram-Schmidt 过程, 用负梯度构造相互共轭的搜索方向. 具体地，置

$$d_0 = -r_0, \quad t = 0.$$

使用待定系数法, 令

$$d_{t+1} = -r_{t+1} + \sum_{s=0}^{t} \gamma_{ts} d_s.$$

由 $d_t^T A d_j = 0, j = 0, \cdots, t-1$, 解得

$$\gamma_{ts} = \begin{cases} 0, & s \leqslant t-1 \\ \frac{\|r_{t+1}\|_2^2}{\|r_t\|_2^2}, & s = t \end{cases},$$

从而

$$d_{t+1} = -r_{t+1} + \frac{\|r_{t+1}\|_2^2}{\|r_t\|_2^2} d_t;$$

这里的精确步长$\eta_t = \underset{\eta\in\mathbb{R}}{\text{argmin}}\, f(x_t + \eta d_t)$ 变成

$$\eta_t = \frac{\|r_t\|_2^2}{p_t^T A p_t}.$$

共轭梯度法的伪码见算法2.

Algorithm 2 共轭梯度法

$x_0 = 0, r_0 = \nabla f(x_0) = -b, d_0 = -\nabla f(x_0)$.
for $t \geqslant 0$ **do**
  $p_t^{\mathrm{mv}} = Ad_t$
  $\eta_t = \frac{\|r_t\|^2}{\langle d_t, p_t^{\mathrm{mv}} \rangle}$
  $x_{t+1} = x_t + \eta_t d_t$
  $r_{t+1} = r_t + \eta_t p_t^{\mathrm{mv}}$
  $\gamma_t = \frac{\|r_{t+1}\|^2}{\|r_t\|^2}$
  $d_{t+1} = -r_{t+1} + \gamma_t d_t$
end for
Ensure: x_t

引理 **8.9** (性质)**.** 对于共轭梯度法, 如下三个性质成立:

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

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

(iii) 搜索方向是共轭的: 对 $i \neq j$, $d_i^T A d_j = 0$.

引理 **8.10.** 共轭梯度法产生的点列满足

$$x_{t+1} = \underset{x \in K_t(A,b)}{\mathrm{argmin}} \|x - x_*\|_A.$$

证明. 由迭代格式, 有 $x_{t+1} \in K_t(A, b)$. 此外, $x \in K_t(A, b)$, 有

$$\begin{aligned}
&\|x - x_*\|_A^2 \\
=&\|x - x_{t+1} + x_{t+1} - x_*\|_A^2 \\
=&\|x - x_{t+1}\|_A^2 + \|x_{t+1} - x_*\|_A^2 + 2(x - x_{t+1})^T A(x_{t+1} - x_*) \\
=&\|x - x_{t+1}\|_A^2 + \|x_{t+1} - x_*\|_A^2 \geqslant \|x_{t+1} - x_*\|_A^2.
\end{aligned}$$

最后一个不等式是由引理 8.9(ii)得到的. ■

事实 **8.11.** 共轭梯度法具有二次终止性, 即 n 步之内可以得到 f 的极小点. 更确切地, 设 A 有 k 个互不相同特征值数目, 那么至多 k 步之内可以得到 f 的极小点.

事实 **8.12.** 共轭梯度法本质上在求解多项式逼近问题:

$$\min_{p:\deg(p) \leqslant t, p(0)=1} \|p(A)(x_0 - x_*)\|_A.$$

事实 **8.13.** 就 $\|\cdot\|_A$ 而言, 共轭梯度法至少和 *Chebyshev* 法一样快.

例子 **8.14.** 利用共轭梯度法解方程组 $Ax = b$, 其中

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

初始点 $x_0 = (0, 0, 0)^T$, 终止条件 $\|r_t\| \leqslant 10^{-6}$.

表 8.1: 共轭梯度法解例8.14的迭代数据

t	x_t	$\nabla f(x_t)$	γ_t	d_t	η_t
0	$(0,0,0)$	$(-1,-1,-1)$		$(1,1,1)$	$\frac{1}{2}$
1	$(\frac{1}{2},\frac{1}{2},\frac{1}{2})$	$(-\frac{1}{2},0,\frac{1}{2})$	$\frac{1}{6}$	$(\frac{2}{3},\frac{1}{6},-\frac{1}{3})$	$\frac{3}{5}$
2	$(\frac{9}{10},\frac{3}{5},\frac{3}{10})$	$(-\frac{1}{10},\frac{1}{5},-\frac{1}{10})$	$\frac{3}{25}$	$(\frac{9}{50},-\frac{9}{50},\frac{3}{50})$	$\frac{5}{9}$
3	$(1,\frac{1}{2},\frac{1}{3})$	$(0,0,0)$			

共轭梯度法的核心计算量是计算一个矩阵向量乘积，而且并不需要形成矩阵本身，只要能计算矩阵向量的乘积就足够了. 很多应用会利用这个特点. 比如考虑稀疏矩阵与向量的乘积，比如求 $y = Av$, 其中

$$A = \begin{bmatrix} 4 & 1 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}.$$

普通的矩阵向量乘需要6^2次乘法运算，根据该矩阵的特点，算法

Set $n = 6$;

for $i = 1, 2, \cdots, n$

$y_i = 4v_i$;

if $i > 1$ then $y_i \leftarrow y_i + v_{i-1}$;

if $i < n$ then $y_i \leftarrow y_i + v_{i+1}$;

end for

仅需要6 次乘法运算就可以得到矩阵向量乘积.

9 Nesterov快速梯度法

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

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

最后一项被解释为“动量".

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的快速算法，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]

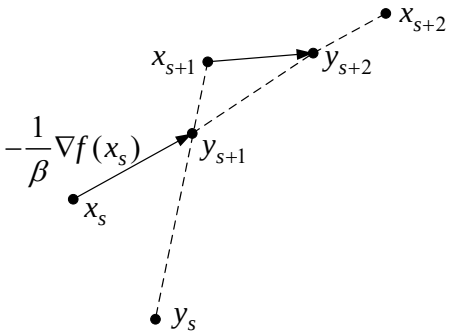


图 9.1: Nesterov的加速梯度法的说明

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$\beta-$光滑的$\alpha$-强凸函数，得到的收敛速率是 $\exp\left(-\Omega\left(\sqrt{\frac{\beta}{\alpha}}t\right)\right)$[5].

表 9.1比较了当将Nesterov方法和普通梯度下降法应用于性质不同的函数时，得到的误差$\epsilon(t)$的上界，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 2.4)	$O\left(\beta/t^2\right)$(定理 9.2)
β-光滑, α-强凸	$\exp\left(-\Omega(t\alpha/\beta)\right)$(定理 3.5)	$\exp\left(-\Omega(t\sqrt{\alpha/\beta})\right)$ (定理 9.1)

这里描述原始的Nesterov方法, 该方法达到了光滑凸优化的最佳oracle复杂度. 对于强凸和非强凸情形, 都给出了该方法的细节. 我们参考了Su et al.[2014][Su2]对该方法的微分方程的最新解释, 并参考了Allen-Zhu and Orecchia[2014][AZO17]关于其与镜像下降的关系(见第5讲).

9.1 光滑强凸情形

Nesterov加速梯度法的几何直观如图 9.1所示, 可以描述为:从任意初始点$x_1 = y_1$开始, 当$t \geqslant 1$时, 按如下方式迭代:

$$y_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t), \tag{9.1}$$

$$x_{t+1} = (1 + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})y_{t+1} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}y_t. \tag{9.2}$$

这里的主迭代序列是$\{y_t\}$，每次先将当前点y_t沿$y_t - y_{t-1}$外推至x_t，再在梯度步.

[5]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 **9.1.** 设f是$\beta-$光滑的$\alpha-$强凸函数, 那么*Nesterov*加速梯度法满足

$$f(y_{t+1}) - f(x^*) \leqslant \frac{\alpha+\beta}{2}\|x_1 - x^*\|^2 \exp(-\frac{\sqrt{\kappa}}{)}.$$

证明. 通过归纳定义$\alpha-$强凸二次函数$\Phi_t, t \geqslant 1$如下:

$$\begin{aligned}\Phi_1(x) &= f(x_1) + \frac{\alpha}{2}\|x - x_1\|^2,\\ \Phi_{t+1}(x) &= (1 - \frac{1}{\sqrt{\kappa}})\Phi_t(x) + \frac{1}{\sqrt{\kappa}}[f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{\alpha}{2}\|x - x_t\|^2]. \end{aligned} \tag{9.3}$$

直观上, Φ_t在以下意义上成为f的越来越精细的近似值(从小到大):

$$\Phi_{t+1}(x) \leqslant f(x) + (1 - \frac{1}{\sqrt{\kappa}})^t(\Phi_1(x) - f(x)) \ \forall x \in \mathbb{R}^n. \tag{9.4}$$

利用$\alpha-$强凸性得到

$$f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{\alpha}{2}\|x - x_t\|^2 \leqslant f(x).$$

然后直接用归纳法可证明不等式(9.4). 然而不等式(9.4)本身并不能说明什么, 要想有用, 人们需要了解Φ_t比f小多少. 以下不等式回答了这个问题:

$$f(y_t) \leqslant \min_{x\in\mathbb{R}^n} \Phi_t(x). \tag{9.5}$$

下面证明(9.5)是正确的. 在开始之前， 首先看如何结合(9.4)和(9.5)来得到由定理给出的收敛速率(由$\beta-$光滑性，有$f(x) - f(x^*) \leqslant \frac{\beta}{2}\|x - x^*\|^2$):

$$\begin{aligned} f(y_{t+1}) - f(x_*) &\leqslant \Phi_{t+1}(x_*) - f(x_*)\\ &\leqslant \Big(1 - \frac{1}{\sqrt{\kappa}}\Big)^t(\Phi_1(x_*) - f(x_*))\\ &\leqslant \frac{\alpha+\beta}{2}\|x_1 - x_*\|^2\Big(1 - \frac{1}{\sqrt{\kappa}}\Big)^t. \end{aligned}$$

现在用归纳法证明(9.5)(注意, 由于$x_1 = y_1$, 所以$t = 1$时(9.5)成立). 令$\Phi_t^* = \min_{x\in\mathbb{R}^n} \Phi_t(x)$.利用$y_{t+1}$的定义、$\beta-$光滑性, 凸性和归纳假设, 可以得到

$$\begin{aligned} f(y_{t+1}) &\leqslant f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2\\ &= (1 - \frac{1}{\sqrt{\kappa}})f(y_t) + (1 - \frac{1}{\sqrt{\kappa}})(f(x_t) - f(y_t))\\ &\quad + \frac{1}{\sqrt{\kappa}}f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2\\ &\leqslant (1 - \frac{1}{\sqrt{\kappa}})\Phi_t^* + (1 - \frac{1}{\sqrt{\kappa}})\nabla f(x_t)^\top (x_t - y_t)\\ &\quad + \frac{1}{\sqrt{\kappa}}f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2. \end{aligned}$$

因此证明

$$(1-\frac{1}{\sqrt{\kappa}})\Phi_t^*+(1-\frac{1}{\sqrt{\kappa}})\nabla f(x_t)^\top(x_t-y_t)+\frac{1}{\sqrt{\kappa}}f(x_t)-\frac{1}{2\beta}\|\nabla f(x_t)\|^2\leqslant\Phi_{t+1}^* \tag{9.6}$$

是充分的. 为了证明这个不等式, 必须更好地理解函数Φ_t. 首先注意, $\nabla^2\Phi_{t+1}(x)=\alpha I_n$(直接由归纳和$\Phi_t$的定义产生). 因此, 存在$v_{t+1}\in\mathbb{R}^n$使得$\Phi_{t+1}$必须具有以下形式:

$$\Phi_{t+1}(x)=\Phi_{t+1}^*+\frac{\alpha}{2}\|x-v_{t+1}\|^2.$$

观察到, 通过微分(9.3)并使用Φ_{t+1}的上述形式, 可以得到

$$\nabla\Phi_{t+1}(x)=\alpha(1-\frac{1}{\sqrt{\kappa}})(x-v_t)+\frac{1}{\sqrt{\kappa}}\nabla f(x_t)+\frac{\alpha}{\sqrt{\kappa}}(x-x_t).$$

特别地, 根据定义, Φ_{t+1}在v_{t+1}取到最小值, 从而可以通过使用上述等式的归纳来定义, 准确地说:

$$v_{t+1}=(1-\frac{1}{\sqrt{\kappa}})v_t+\frac{1}{\sqrt{\kappa}}x_t-\frac{1}{\alpha\sqrt{\kappa}}\nabla f(x_t). \tag{9.7}$$

使用Φ_t和Φ_{t+1}的上述形式, 以及原始定义(9.3), 计算Φ_{t+1}在x_t的值，得到:

$$\Phi_{t+1}^*+\frac{\alpha}{2}\|x_t-v_{t+1}\|^2=(1-\frac{1}{\sqrt{\kappa}})\Phi_t^*+\frac{\alpha}{2}(1-\frac{1}{\sqrt{\kappa}})\|x_t-v_t\|^2+\frac{1}{\sqrt{\kappa}}f(x_t). \tag{9.8}$$

请注意, 由于(9.7), 有

$$\begin{aligned}\|x_t-v_{t+1}\|^2=(1-\frac{1}{\sqrt{\kappa}})^2\|x_t-v_t\|^2+\frac{1}{\alpha^2\kappa}\|\nabla f(x_t)\|^2\\-\frac{2}{\alpha\sqrt{\kappa}}(1-\frac{1}{\sqrt{\kappa}})\nabla f(x_t)^\top(v_t-x_t),\end{aligned}$$

将此代入(9.8)，并整理得到

$$\begin{aligned}(1-\frac{1}{\sqrt{\kappa}})\Phi_t^*+\frac{1}{\sqrt{\kappa}}f(x_t)+\frac{\alpha}{2\sqrt{\kappa}}(1-\frac{1}{\sqrt{\kappa}})\|x_t-v_t\|^2\\-\frac{1}{2\beta}\|\nabla f(x_t)\|^2+\frac{1}{\sqrt{\kappa}}(1-\frac{1}{\sqrt{\kappa}})\nabla f(x_t)^\top(v_t-x_t)=\Phi_{t+1}^*.\end{aligned}$$

最后, 用归纳法证明

$$v_t-x_t=\sqrt{\kappa}(x_t-y_t)$$

就完成了(9.6)的证明, 从而也完成了该定理的证明:

$$\begin{aligned}v_{t+1}-x_{t+1}&=(1-\frac{1}{\sqrt{\kappa}})v_t+\frac{1}{\sqrt{\kappa}}x_t-\frac{1}{\alpha\sqrt{\kappa}}\nabla f(x_t)-x_{t+1}\\&=\sqrt{\kappa}x_t-(\sqrt{\kappa}-1)y_t-\frac{\sqrt{\kappa}}{\beta}\nabla f(x_t)-x_{t+1}\\&=\sqrt{\kappa}y_{t+1}-(\sqrt{\kappa}-1)y_t-x_{t+1}\\&=\sqrt{\kappa}(x_{t+1}-y_{t+1}),\end{aligned}$$

其中第一个等式来自(9.7), 第二个来自归纳假设, 第三个来自y_{t+1}的定义, 最后一个来自x_{t+1}的定义. ■

9.2 光滑情形

在这一节中, 我们展示了如何在$\alpha = 0$的情况下, 使用主序列$\{y_t\}$中素的时变组合来调整Nesterov加速梯度法. 首先, 定义以下序列:

$$\lambda_0 = 0, \lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2}, 和\gamma_t = \frac{1 - \lambda_t}{\lambda_{t+1}}.$$

(注意$\gamma_t \leqslant 0$.)现在, 该算法简单地由以下等式定义, $x_1 = y_1$是任意初始点,

$$\begin{aligned} y_{t+1} &= x_t - \frac{1}{\beta}\nabla f(x_t), \\ x_{t+1} &= (1 - \gamma_t)y_{t+1} + \gamma_t y_t. \end{aligned}$$

定理 **9.2.** 令f是$\beta-$光滑的凸函数, 则*Nesterov*加速梯度法满足

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

这里遵循Beck and Teboulle [2009][BT09]的证明. 还参考了Tseng [2008][Tse08]的一个更简单的步长证明.

证明. 利用引理3.6(其中$\alpha = 0$)和y_s的定义

$$\begin{aligned} & f(y_{s+1}) - f(y_s) \\ & \leqslant \nabla f(x_s)^\top (x_s - y_s) - \frac{1}{2\beta}\|\nabla f(x_s)\|^2 \\ & = \beta(x_s - y_{s+1})^\top (x_s - y_s) - \frac{\beta}{2}\|x_s - y_{s+1}\|^2. \end{aligned} \tag{9.9}$$

类似地得到

$$f(y_{s+1}) - f(x_*) \leqslant \beta(x_s - y_{s+1})^\top (x_s - x_*) - \frac{\beta}{2}\|x_s - y_{s+1}\|^2. \tag{9.10}$$

现在给(9.9) 两边同时乘以$(\lambda_s - 1)$并将结果加到(9.10), 再由$\delta_s = f(y_s) - f(x_*)$得到,

$$\begin{aligned} & \lambda_s\delta_{s+1} - (\lambda_s - 1)\delta_s \\ & \leqslant \beta(x_s - y_{s+1})^\top (\lambda_s x_s - (\lambda_s - 1)y_s - x_*) - \frac{\beta}{2}\lambda_s\|x_s - y_{s+1}\|^2. \end{aligned}$$

给此不等式两边同时乘以λ_s, 并根据定义$\lambda_{s-1}^2 = \lambda_s^2 - \lambda_s$, 以及恒等式

$$2a^\top b - \|a\|^2 = \|b\|^2 - \|b - a\|^2$$

得到

$$\begin{aligned} & \lambda_s^2\delta_{s+1} - \lambda_{s-1}^2\delta_s \\ & \leqslant \frac{\beta}{2}(2\lambda_s(x_s - y_{s+1})^\top (\lambda_s x_s - (\lambda_s - 1)y_s - x_*) - \|\lambda_s(y_{s+1} - x_s)\|^2) \\ & = \frac{\beta}{2}(\|\lambda_s x_s - (\lambda_s - 1)y_s - x_*\|^2 - \|\lambda_s y_{s+1} - (\lambda_s - 1)y_s - x_*\|^2). \end{aligned} \tag{9.11}$$

接下来, 根据定义, 有

$$\begin{aligned} & x_{s+1} = y_{s+1} + \gamma_s(y_s - y_{s+1}) \\ \Leftrightarrow\ & \lambda_{s+1}x_{s+1} = \lambda_{s+1}y_{s+1} + (1-\lambda_s)(y_s - y_{s+1}) \\ \Leftrightarrow\ & \lambda_{s+1}x_{s+1} - (\lambda_{s+1}-1)y_{s+1} = \lambda_s y_{s+1} - (\lambda_s - 1)y_s. \end{aligned} \tag{9.12}$$

由(9.11)和(9.12)以及$u_s := \lambda_s x_s - (\lambda_s - 1)y_s - x_*$, 我们得到

$$\lambda_s^2\delta_{s+1} - \lambda_{s-1}^2\delta_s^2 \leqslant \frac{\beta}{2}(\|u_s\|^2 - \|u_{s+1}\|^2).$$

从$s = 1$到$s = t$, 将这些不等式相加, 可以得到:

$$\delta_{t+1} \leqslant \frac{\beta}{2\lambda_t^2}\|u_1\|^2.$$

用归纳法容易得到$\lambda_t \geqslant \frac{t}{2}$, 从而完成证明. ■

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连续	次梯度法	$RL/\sqrt{t}$ (定理 2.11)
强凸、Lipschitz连续	次梯度法	$L^2/(\alpha t)$(定理 3.4)
凸、光滑	加速梯度法	$\beta R^2/t^2$(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.

一般来说, 黑箱过程是从“历史”映射到下一个查询点, 也就是说它映射

$$(x_1, g_1, \ldots, x_t, g_t)$$

(这里$g_s \in \partial f(x_s)$)到x_{t+1}. 为了简化表示和参数, 形式上，定义黑箱方法如下.

定义 **10.1** (黑箱程序)**.** 黑箱程序产生点列 $\{x_t\}$使得

$$x_{t+1} \in x_1 + \operatorname{span}\{g_1, \cdots, g_t\}, \tag{10.1}$$

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，梯度法是一个黑箱程序. 的确，把迭代展开，可将x_{t+1}表示为

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

现在证明针对任何黑箱程序的收敛速率的下界. 第一个定理是关于两种非光滑凸和强凸函数的情况. 定理来自[Nes83], 但是陈述遵循[Nes04].

定理 **10.2** (非光滑凸函数)**.** 设 $t \leqslant n$, $L, R > 0$. 存在 L-Lipschitz连续的凸函数 f， 使得任何满足 (10.1)的黑箱程序而言，

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2(R)} f(x) \geqslant \frac{RL}{2(1+\sqrt{t})}. \tag{10.2}$$

也存在L-Lipschitz连续的α-强凸函数 f使得

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2\left(\frac{L}{2\alpha}\right)} f(x) \geqslant \frac{L^2}{8\alpha t}. \tag{10.3}$$

证明策略就是构造一个凸函数f， 以便对任何黑箱程序， $\operatorname{span}\{g_1, \cdots, g_t\} \subset \operatorname{span}\{e_1, \cdots, e_t\}$, 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.

证明. 考虑函数[6]

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

这里$\alpha > 0, \gamma > 0$是待定参数. 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连续的.

考虑 $x_* \in \mathbb{R}^n$ 使得

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

[6]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.

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.

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

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

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连续的情况下，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连续的，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连续的，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，证明策略是构造一个病态凸函数. 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 使得黑箱方法满足

$$\min_{1\leqslant s\leqslant t} f(x_s) - f(x_*) \geqslant \frac{3\beta\|x_1 - x_*\|_2^2}{32(t+1)^2}. \tag{10.4}$$

证明. 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是有向链图的Laplace矩阵.[7] 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

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

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

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

如果$x_1 = 0$, 类似于定理 10.2的证明, 由f的构造可以证明

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

那么对任何黑箱程序，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$s \leqslant t$,

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

[7] 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$s \leqslant t$,

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

■

为了简化下一个定理的证明, 将考虑$n \to +\infty$的极限情况. 更准确地说, 假设现在在

$$l_2 = x = (x(n))_{n\in\mathcal{Z}_{++}} : \sum_{i=1}^{+\infty} x(i)^2 < +\infty$$

中进行而不是在$\mathbb{R}^n$. 注意本章中证明的所有定理实际上在任意希尔伯特空间$\mathcal{H}$ 中都是有效的. 这里选择在$\mathbb{R}^n$中论述只是为了阐述的清晰性.

定理 **10.4.** 令$\kappa > 1$.存在一个$\beta-$光滑的$\alpha-$强凸函数$f : l_2 \to \mathbb{R}$, 使得$\kappa = \beta/\alpha$，且对任意$t \geqslant 1$和满足 (10.1)的黑箱过程，

$$f(x_t) - f(x_*) \geqslant \frac{\alpha}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(t-1)}\|x_1 - x_*\|^2.$$

注意到对很大的条件数κ,

$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(t-1)} \approx \exp\left(-\frac{4(t-1)}{\sqrt{\kappa}}\right).$$

证明. 整个论证类似于定理 10.3的证明. 设$A : l_2 \to l_2$为线性算子, 对应的是无限三对角矩阵, 对角线元素为2, 上下对角线元素为-1. 现在考虑函数:

$$f(x) = \frac{\alpha(\kappa-1)}{8}(\langle Ax, x\rangle - 2\langle e_1, x\rangle) + \frac{\alpha}{2}\|x\|^2.$$

已经证明的事实$0 \preceq A \preceq 4I$蕴含着f是$\alpha-$强凸和$\beta-$光滑的. 和往常一样, 关键是观察到：这个函数, 由于我们假设的是黑箱过程, 一定有$x_t[i] = 0, \forall i \geqslant t$. 这蕴含着:

$$\|x_t - x_*\|^2 \geqslant \sum_{i=t}^{+\infty} x_*[i]^2.$$

进一步, 因为f是$\alpha-$强凸函数,

$$f(x_t) - f(x_*) \geqslant \frac{\alpha}{2}\|x_t - x_*\|^2.$$

因此只需要计算x_*. 这可以通过对f求导并将梯度设为0来实现, 这就得到了下面的无穷方程组

$$1 - 2\frac{\kappa+1}{\kappa-1}x_*[1] + x_*[2] = 0,$$
$$x_*[k-1] - 2\frac{\kappa+1}{\kappa-1}x_*[k] + x_*[k+1] = 0, \forall k \geqslant 2.$$

很容易验证定义为$x_*[i] = (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^i$的$x_*$满足这个无限方程, 然后通过简单计算得到该定理的结论. ■

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. 这小节处理这种情况.

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

$$0 \leqslant f(x) - [f(y) + \langle \nabla f(y), x - y\rangle] \leqslant \frac{\beta}{2}\|x - y\|^2. \tag{10.5}$$

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 $x \in \Omega$， (10.5)式精确成立. 一个非精确oracle，返回一对使得 (10.5)对某个松弛的 δ 成立.

定义 **10.5** (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

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

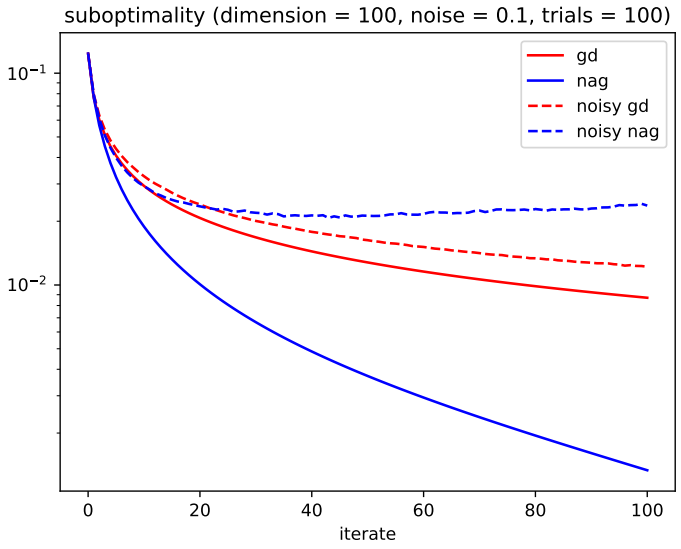


图 10.1: 将梯度法(GD)和Nesterov加速梯度法(NAG)应用到$n = 100$的世界上最差的函数时，最优性间隙随迭代的变化. 请注意，加速极大地得益于精确梯度. 然而，当给梯度增加均匀半径$\delta = 0.1$的球面噪声后，随机梯度法(noisy GD) 仍然表现稳健，而随机加速梯度法(noisy NAG)会累积误差. 随机方法的数据是100 次试验结果的平均.

考虑返回δ-非精确oracle的梯度法. 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，任何黑箱法如果在非精确情况下被加速，它必定会累积误差.

定理 **10.6** ([DGN14]，定理6)**.** 考虑收敛速率为$O\left(\frac{\beta R^2}{t^p}\right)$的黑箱法使用非精确oracle的情况. 用$\delta$-非精确oracle, 假设算法获得速率

$$f(x_t) - f(x_*) \leqslant O\left(\frac{\beta R^2}{t^p}\right) + O\left(t^q \delta\right), \tag{10.6}$$

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经验风险很有效.

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，这在经验风险最小化的应用中很常见. 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$f(x) = \mathbb{E}_I[f_I(x)]$. 所以，(11.1)可看作随机优化 (SO)的特例. 后面为了区别一般的随机优化问题，将 (11.1)称作外部随机优化问题.

11.1 风险极小化与经验风险极小化

本小节介绍机器学习中训练问题的随机优化表述. 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. 现在定义非负实值损失函数 $\ell(\widehat{y}, y)$ 来度量预测值 $\widehat{y}$和真实结果y之间的差异.

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

$$R(f) = \mathop{\mathbb{E}}_{(X,Y)\sim P_Z} \ell(f(X), Y)\,.$$

学习算法的目标是在函数类$\mathcal{F}$中找到极小化$R(f)$ 的预测函数

$$f_* \in \arg\min_{f\in\mathcal{F}} R(f).$$

实际中，每个学习算法会将函数类$\mathcal{F}$参数化，即$f(X) = f_w(X), w \in \Omega$，其中$w$是确定函数$f$的参数. 由此得到$\ell(w, Z) := \ell(f_w(X), Y)$. 学习算法的终极目标是求解优化变量为$w$的风险极小化问题

$$w_* = \arg\min_{w\in\Omega} R(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, Z) \tag{RM}$$

易见(RM)形如 (SO)，是随机优化问题.

通常，因为联合分布未知，所以无法计算风险$R(f)/R(w)$. 实际中能有的只是数据，可以建模为：已知从P_Z抽取的m个独立同分布样例 $S = ((X_1, Y_1), \ldots, (X_m, Y_m))$. 因此，取而代之的是极小化所谓的经验风险(empirical risk)，其定义为损失函数在训练集上的平均值:

$$\widehat{R}(f) = \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i),$$

参数化的经验风险最小点(empirical risk minimizer, ERM)是f_{w^*}，其中

$$w_* = \arg\min_{w\in\Omega} \widehat{R}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, Z_i), \tag{ERM}$$

这里$\ell(w, Z_i) = \ell(f_w(x_i), y_i)$. 易见经验风险极小化问题形如 (11.1).

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 algorithm)**从点$x_0 \in \Omega, t = 0$ 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

$$x_* = \frac{1}{m}\sum_{i=1}^{m} p_i.$$

现在，用步长$\eta_t = 1/t$，并按循环顺序执行随机梯度法，即$i_t = t$ 和 $x_0 = 0$:

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

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 $w_1 \in \mathbb{R}^n$，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) 的经验风险极小化问题执行SGM.

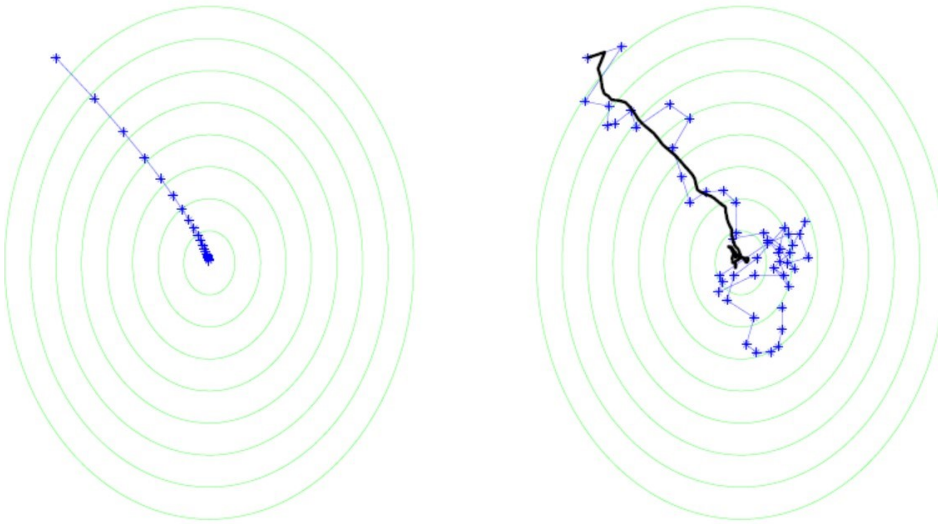
例子 **11.5** (SVM)**.** 已知带标签的点 $((x_1, y_1), \cdots, (x_m, y_m)) \in (\mathbb{R}^n \times \{-1, 1\})^m$, SVM正则化的经验风险极小化目标函数:

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

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

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求解 (SO)，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 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 3.



(a) 梯度下降法　　(b) 随机梯度下降法

图 11.1: 求解外部随机化问题 (11.1)的梯度下降法和随机梯度法

Algorithm 3 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: $\quad y_{s+1} = x_s - \eta_s \widehat{g}_s, \ \ \widehat{g}_s \in \partial \ell(x_s, Z_s)$
3: $\quad x_{s+1} = \Pi_\Omega(y_{s+1})$
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 $\mathbb{E}[\widehat{g}_s|x_s] \in \partial f(x_s)$. 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，一个基本假设方差有界(variance bounded)：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$\mathrm{diam}(\Omega) \leqslant R$. 假设凸函数$f(x) = \mathbb{E}[\ell(x, Z)]$在$x_* \in \Omega$ 处取到它在Ω 上的最小值，还假设已知Z，$\ell(x, Z)$ 关于x是凸的，并且方差有界假设(VB)成立. 那么，如果$\eta_s \equiv \eta = R/(L\sqrt{t})$,

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

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

$$f_s(x_s) - f_s(x_*) \leqslant \widehat{g}_s^\top (x_s - x_*). \tag{11.2}$$

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

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

■

在风险极小化中，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. 目标是基于这$m$ 个观测数据,

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

其中$f_\theta = \sum_{j=1}^{M} \theta_j h_j(\cdot)$, 这里$h_j : X \to \{-1, 1\}$是已知函数,

$$\Delta_M := \{\theta \in \mathbb{R}^M : \theta \geqslant 0, \sum_j \theta_j = 1\}.$$

已经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$\mathbb{E}[e^{-Yf_\theta(X)}]$. 特别地，对于每对数据，最多只能利用一次. 称作关于数据执行了一轮(one pass).

也可以尝试最小化经验风险：

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

具体地，产生k 个在$\{1,\cdots,m\}$ 上均匀分布的独立随机变量$I_1,\cdots,I_k$，并在每次迭代中使用一个随机变量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，在Δ_M上最小化

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

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.

11.4 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 4 Stochastic Mirror Descent algorithm

Require: $x_1 \in \operatorname{argmin}_{\Omega\cap\mathcal{D}} \Phi(x)$, independent random variables $Z_1,\cdots,Z_t$ with distribution P_Z.
  for $s = 1,\cdots,t$ **do**
    $\nabla\Phi(y_{s+1}) = \nabla\Phi(x_s) - \eta_s\widehat{g}_s$ 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 **11.7.** 假设Φ是$\Omega\cap\mathcal{D}$上关于$\|\cdot\|$ 的可微$\alpha-$ 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

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

证明. 本质上是在重复镜像下降算法的证明. 由$\widehat{g}_s \in \partial f_s(x_s)$，从而 (11.2)成立. 因此

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

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

$$\mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t}[f(x_s) - f(x_*)]\right] \leqslant \frac{\eta L^2}{2\alpha} + \frac{D_\Phi(x_*, x_1)}{t\eta}. \tag{11.3}$$

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}.

序列决策结束后，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，

$$\begin{aligned} v_s[i] &= w_{s-1}[i]e^{-\eta f_s[i]} && \text{(指数权重更新)} \\ w_s &= v_s/(\textstyle\sum_i v_s[i]) && \text{(归一化)} \end{aligned}$$

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.

在线凸优化

$$\min_{w\in\Delta_n}\sum_{s=1}^{t}\langle w, f_s\rangle, \tag{11.4}$$

表示时刻s，仅有信息$f_1,\cdots,f_{s-1}$，据此信息做出决策. 可将乘性权重解释成用镜像下降法求解在线凸优化(11.4)， 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

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

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

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

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_∞范数的上界是 1 . 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，能得到平均遗憾的上界

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

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 .

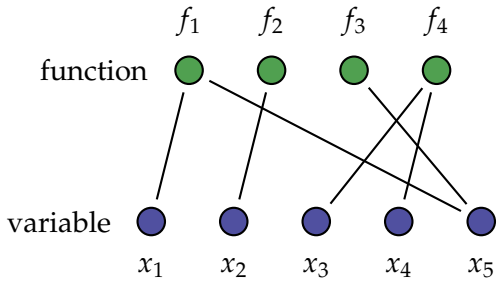


图 12.1: 由函数$f : \mathbb{R}^n \to \mathbb{R}^m$的群稀疏结构诱导的函数分量$f_i$和变量$x_j$间的二分图例子. f_i和x_j之间的边传达了第i个分量函数与输入的第j个坐标有关.

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f是$\mathbb{R}^n$上可微的凸$L-$Lipschitz连续函数. 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：在每步中需要计算梯度的所有分量$\nabla_i f$，以便更新每个分量. 随机坐标下降(random coordinate descent, RCD)算法的思路：在每步中均匀地随机选取一个方向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RCD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 5 Random Coordinate Descent (RCD) 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**
    $y_{s+1} = x_s - \eta_s n \nabla_{I_s} f(x_s) e_{I_s}$
    $x_{s+1} = \Pi_\Omega(y_{s+1})$
  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)中，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 $\mathbb{E}[\widehat{g}_s | x_s] = \nabla f(x_s)$, 但是

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

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

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

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$u, v \in \mathbb{R}$, 存在$\beta_i > 0$满足

$$|\nabla_i f(x + ue_i) - \nabla_i f(x + ve_i)| \leqslant \beta_i |u - v|,$$

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

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

所以这一般并不等价于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 2.4的随机版)**.** 已知$\gamma > 0, \beta \in \mathbb{R}^n_{++}$. 定义加权范数

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

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

证明. 由引理 2.2后面的说明 (2.1)，知道针对一般的β_φ-光滑凸函数 φ,

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

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

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

从而，如果 $i_t \sim p^\gamma$,

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

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

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

此外，由该不等式，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

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

回忆这就是在非随机-情况下，即定理 2.4用来证明收敛性的递归关系 (2.3). ■

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

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

证明. 需要如下引理:

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

证明. 从$\alpha-$强凸的定义开始，对任意的x，得

$$\begin{aligned}
f(x) - f(x_*) &\leqslant \nabla f(x)^\top (x - x_*) - \tfrac{\alpha}{2}\|x - x_*\|_2^2 \\
&\leqslant \|\nabla f(x)\|_* \|x - x_*\| - \tfrac{\alpha}{2}\|x - x_*\|_2^2 \\
&\leqslant \max_{t\geqslant 0} \|\nabla f(x)\|_* t - \tfrac{\alpha}{2} t^2 \\
&= \tfrac{1}{2\alpha}\|\nabla f(x)\|_*^2 .
\end{aligned}$$

■

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，并且由 $\delta_t = \mathbb{E}[f(x_t)] - f(x_*)$可得结论. ■

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并不成立. 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 $w \in \Omega \subseteq \mathbb{R}^n$来指定“模型”.
- "损失函数"记作 $\ell\colon \Omega \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, 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

$$R(w) = \mathop{\mathbb{E}}_{Z \sim D}[\ell(w, Z)]. \tag{13.1}$$

目的是找到极小化 $R(w)$的模型.

达成该目标的一种方法是直接使用随机梯度法极小化总体目标 (13.1)：

$$w_{t+1} = w_t - \eta \nabla \ell(w_t, Z_t), \quad Z_t \sim D$$

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$R(w)$.

13.1 经验风险和推广误差

考虑有限样本. 假设

$$S = ((X_1, Y_1), \cdots, X_m, Y_m) \in (\mathcal{X} \times \mathcal{Y})^m,$$

其中$Z_i = (X_i, Y_i)$代表第i个带标签的样例. 经验风险(empirical risk)定义为

$$\widehat{R}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, Z_i)\,.$$

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 $\widehat{R}(w) \approx R(w)$. 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$\widehat{R}(w)$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

$$\epsilon_{\text{gen}}(w) = R(w) - \widehat{R}(w)\,.$$

请注意如下尽管是同义反复，然而却很重要的等式:

$$R(w) = \widehat{R}(w) + \epsilon_{\text{gen}}(w). \tag{13.2}$$

特别地，这表明：如果通过优化设法使经验风险$\widehat{R}(w)$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Z'_i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价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 $\mathbb{E}[\epsilon_{\text{gen}}(A)] = \Delta(A)$.

证明. 请注意，由定义

$$\mathbb{E}[\epsilon_{\text{gen}}(A)] = \mathbb{E}\left[R(A(S)) - \widehat{R}(A(S))\right],$$
$$\mathbb{E}[R(A(S))] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(A(S), Z_i')\right],$$
$$\mathbb{E}\left[\widehat{R}(A(S))\right]] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(A(S), Z_i)\right].$$

同时，由于Z_i和Z_i'是同分布，并且与其它样例独立，所以

$$\mathbb{E}[\ell(A(S), Z_i)] = \mathbb{E}[\ell(A(S^{(i)}), Z_i')].$$

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

$$\mathbb{E}[R(A(S)) - \widehat{R}(A(S))] = \Delta(A).$$

■

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 [TVW+17].

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连续的. 设

$$\widehat{w}_S = \arg\min_{w\in\Omega} \widehat{R}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, Z_i)$$

表示经验风险最小点(empirical risk minimizer, ERM). 那么 ERM 满足

$$\Delta_{\sup}(\mathrm{ERM}) \leqslant \frac{4L^2}{\alpha m}.$$

证明. 由已知，$\widehat{w}_S$表示与样本 S 关联的经验风险最小点. 已知容量为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$\alpha-$强凸函数的算术平均也是$\alpha-$强凸的知$\widehat{R}$关于w是$\alpha-$强凸的. 进一步由强凸性

$$\widehat{R}(\widehat{w}_{S^{(i)}}) - \widehat{R}(\widehat{w}_S) \geqslant \frac{\alpha}{2}\|\widehat{w}_S - \widehat{w}_{S^{(i)}}\|^2. \tag{13.3}$$

另一方面，

$$\begin{aligned}
&\widehat{R}(\widehat{w}_{S^{(i)}}) - \widehat{R}(\widehat{w}_S)\\
=&\tfrac{1}{m}[\ell(\widehat{w}_{S^{(i)}}, Z_i) - \ell(\widehat{w}_S, Z_i)] + \tfrac{1}{m}\sum_{j\neq i}[\ell(\widehat{w}_{S^{(i)}}, Z_j) - \ell(\widehat{w}_S, Z_j)]\\
=&\tfrac{1}{m}[\ell(\widehat{w}_{S^{(i)}}, Z_i) - \ell(\widehat{w}_S, Z_i)] + \tfrac{1}{m}[\ell(\widehat{w}_S, Z_i') - \ell(\widehat{w}_{S^{(i)}}, Z_i')] + \left[\widehat{R}_{S^{(i)}}(\widehat{w}_{S^{(i)}}) - \widehat{R}_{S^{(i)}}(\widehat{w}_S)\right]\\
\leqslant&\tfrac{1}{m}|\ell(\widehat{w}_{S^{(i)}}, Z_i) - \ell(\widehat{w}_S, Z_i)| + \tfrac{1}{m}|(\ell(\widehat{w}_S, Z_i') - \ell(\widehat{w}_{S^{(i)}}, Z_i'))|\\
\leqslant&\tfrac{2L}{m}\|\widehat{w}_{S^{(i)}} - \widehat{w}_S\|.
\end{aligned} \tag{13.4}$$

这里的第一个不等式利用了 $\widehat{R}_{S^{(i)}}(\widehat{w}_{S^{(i)}}) - \widehat{R}_{S^{(i)}}(\widehat{w}_S)) \leqslant 0$, 第二个不等式利用了损失函数$\ell$是$L-$Lipschitz连续的事实.

将 (13.3)和 (13.4)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连续性，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函数

$$f(w) = \widehat{R}(w) + \frac{\alpha}{2}\|w\|^2 = \frac{1}{m}\sum_{i=1}^{m}\ell(w, Z_i) + \frac{\alpha}{2}\|w\|^2$$

正则化损失$f(w)$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，由前一个定理，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，算法本身在求解正则化目标，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 连续的. 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 $\eta_s \approx 1/m$那么 $\Delta_{\sup}(\mathrm{SGM}_t) = O(\ln(t)/m)$, 并且对于$\eta_s \approx 1/\sqrt{m}$和$t = O(m)$ 那么$\Delta_{\sup}(\mathrm{SGM}_t) = O(1/\sqrt{m})$. 证明参见[HRS15].

Part IV
对偶方法

14 Lagrange对偶

14.1 引例

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 1.10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

$$\begin{aligned} \nabla f(x_*) + A^T\lambda^* &= 0 \\ Ax_* &= b. \end{aligned}$$

这就是将著名的**Karush-Kuhn-Tucker**条件**(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

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} c \\ b \end{bmatrix}.$$

下面将该结论拓展到包含非线性不等式约束，且f不可微的情况.

14.2 对偶问题

考虑一般的约束优化问题:

$$\begin{aligned} f_* = \inf_{x\in X} \quad & f(x) \\ \text{s.t.} \quad & h(x) = 0 \\ & g(x) \leqslant 0 \end{aligned} \tag{MP}$$

其中 $X \subseteq \mathbb{R}^n$是凸集，$g = (g_1, \cdots, g_m) : \mathbb{R}^n \mapsto \mathbb{R}^m, h = (h_1, \cdots, g_\ell) : \mathbb{R}^n \mapsto \mathbb{R}^\ell$. 感兴趣的问题假设已经有可行解$x_*$, 那么它是最优解的条件(必要条件、充分条件、充分必要条件)是什么? 对于一般光滑的问题，仅能利用微积分刻画局部解的条件. 凸规划有可验证的关于全局最优性的局部充分条件. 本节使用凸择一定理作为工具，给出全局解的充要条件. 因为这里不需要函数可微，从而也称得到的结论是(MP)最优解的零阶最优性条件.

称

$$L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x)$$

是(MP)的Lagrange 函数. 定义$q : \mathbb{R}^m \times \mathbb{R}^\ell \mapsto [-\infty, \infty), q(\lambda, \mu) = \inf_{x\in X} L(x, \lambda, \mu)$，称

$$\begin{aligned} q^* = \quad & \sup_{\lambda\in\mathbb{R}^m, \mu\in\mathbb{R}^\ell} \quad q(\lambda, \mu) \\ & \text{s.t.} \quad \lambda \geqslant 0 \end{aligned} \tag{D_{MP}}$$

是(MP)的对偶问题. 需要指出的是，这里q的值域包括$-\infty$. 实际的表示和计算是在q的定义域

$$\operatorname{dom} q = \{(\lambda, \mu) : q(\lambda, \mu) > -\infty\}$$

和$\mathbb{R}^m_+ \times \mathbb{R}^\ell$的交集上极小化$q$，从而有些对偶问题的显式表示中，除不等式约束非负外，还会出现保证$q(\lambda, \mu) > -\infty$的显式约束.

命题 **14.1** (弱对偶性)**.** 设 x 是(MP)的可行解, $(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^\ell$, 那么 $f(x) \geqslant q(\lambda, \mu)$. 于是 $f_* \geqslant q^*, f_* \geqslant q(\lambda, \mu), f(x) \geqslant q^*$.

证明. 由x 是(MP)的可行解($x \in X, g(x) \leqslant 0, h(x) = 0$和$\lambda \geqslant 0$)和$\lambda \geqslant 0$，有

$$L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x) \leqslant f(x).$$

再由对偶函数的定义有

$$q(\lambda, \mu) = \inf_{x\in X} L(x, \lambda, \mu) \leqslant f(x).$$

■

例子 **14.2** (0对偶间隙)**.** 问题

$$\begin{aligned} f_* = \min_{(x_1,x_2)\geqslant 0} \quad & e^{-\sqrt{x_1 x_2}} \\ \text{s.t.} \quad & x_1 \leqslant 0, \end{aligned}$$

满足 $f_* = 1$. 但对于 $\lambda \geqslant 0$,

$$q(\lambda) = 0 = \inf_{x_1\geqslant 0,\ x_2\geqslant 0} e^{-\sqrt{x_1 x_2}} + \lambda x_1.$$

所以$q^* = 0$. 从而$f_* - q^* > 0$.

例子 **14.3** (0对偶间隙，但对偶问题取不到最优值)**.** 考虑问题

$$\begin{aligned}\inf_{x\in\mathbb{R}} \quad & f(x)=x\\ \text{s.t.} \quad & g(x)=x^2\leqslant 0.\end{aligned}$$

易见 $x_*=0, f_*=0,$

$$\begin{aligned}q(\lambda)&=\inf_{x\in\mathbb{R}} x+\lambda x^2\\ &=\begin{cases}-\frac{1}{4\lambda}, & \lambda>0\\ -\infty, & \lambda=0\end{cases}\end{aligned}$$

因此 $0=\sup_{\lambda\geqslant 0} q(\lambda)$. 从而$f_*-q^*=0$，但是 $\nexists\lambda^*\geqslant 0$ 使得 $q(\lambda^*)=0.$

弱对偶性表明对偶间隙$f_*-q^*\geqslant 0$. 例 14.2表明:原始对偶问题的最优值的对偶间隙可以为正；例 14.3表明：即对偶间隙为 0，对偶问题的最优解也不一定存在. 对偶间隙为0时，称强调偶成立. 此时若对偶问题的最优解存在，可由对偶问题的解恢复出原始问题的解. 得到的结论也称作(MP)最优解的零阶条件.

14.3 强对偶性

研究**Lagrange**强对偶性的工具是凸择一定理. 为了表述简洁，突出本质，本小节仅针对不等式约束问题

$$\begin{aligned}\min_{x\in X} \quad & f(x)\\ \text{s.t.} \quad & g(x)\leqslant 0,\end{aligned}\tag{ICP}$$

并且假设它是凸的，即X 是$\mathbb{R}^n$ 的凸子集，$f(\cdot), g_1(\cdot), \cdots, g_m(\cdot)$是$X$上的实值凸函数，没有等式约束. 尽管可以允许有线性等式约束, 但这并不具备普适性.

现在考虑如何验证系统

$$\begin{aligned}&f(x)<c\\ &g_j(x)\leqslant 0, \quad j=1,\cdots,m\\ &x\in X\end{aligned}\tag{I}$$

无解. 答案：假设存在非负权重$\lambda_j, j=1,\cdots,m,$ 使得不等式

$$f(x)+\sum_j \lambda_j g_j(x)<c$$

在X上无解，即

$$\exists\lambda_j\geqslant 0:\ \inf_{x\in X}\left[f(x)+\sum_j\lambda_j g_j(x)\right]\geqslant c.$$

那么(I)无解.

定理 **14.4** (凸择一定理)**.** 考虑关于x的约束系统(I)和连同关于λ的约束系统

$$\begin{aligned}\inf_{x\in X}\left[f(x)+\sum_j\lambda_j g_j(x)\right]&\geqslant c\\ \lambda_j&\geqslant 0, \quad j=1,\cdots,m.\end{aligned}\tag{II}$$

(a) *[平凡部分]*如果(II)有解, 那么(I)无解.

(b) *[非平凡部分]* 如果(I)无解，系统(I)是凸的(即X是凸集; $f, g_1, \cdots, g_m$ 是X上的实值凸函数), 并且子系统

$$\begin{array}{c} g_j(x) < 0, \quad j = 1, \cdots, m, \\ x \in X \end{array}$$

有解*[Slater*条件*]*. 那么(II)有解.

证明. 非平凡部分：假设(I)无解. 考虑$\mathbb{R}^{m+1}$ 上的两个集合:

$$\begin{aligned} T &:= \left\{u \in \mathbb{R}^{m+1} : \exists x \in X \text{ s.t.} f(x) \leqslant u_0, g_1(x) \leqslant u_1, \cdots, g_m(x) \leqslant u_m\right\}, \\ S &:= \left\{u \in \mathbb{R}^{m+1} : u_0 < c, \ u_1 \leqslant 0, \cdots, \ u_m \leqslant 0\right\}. \end{aligned}$$

观察到S, T是非空凸集；S与T不相交(否则(I)有解). 从而由凸集分离定理 1.2知S和T是可分离的:

$$\exists (a_0, \cdots, a_m) \neq 0 : \inf_{u \in T} a^T u \geqslant \sup_{u \in S} a^T u.$$

即$\exists (a_0, \cdots, a_m) \neq 0$使得

$$\begin{aligned} & \inf_{x \in X} \inf_{u_0, u_1, \cdots, u_m} \left\{a^T u : u_0 \geqslant f(x), u_1 \geqslant g_1(x), \cdots, u_m \geqslant g_m(x)\right\} \\ \geqslant & \sup_{u_0, u_1, \cdots, u_m} \left\{a^T u : u_0 < c, u_1 \leqslant 0, \cdots, u_m \leqslant 0\right\}. \end{aligned}$$

从而

$$\inf_{x \in X} [a_0 f(x) + a_1 g_1(x) + \cdots + a_m g_m(x)] \geqslant a_0 c \tag{14.4}$$

并且由S的结构知$a \geqslant 0$. 即存在$a \geqslant 0, a \neq 0$使得(14.4)成立.

进一步观察到$a_0 > 0$. 的确，否则$0 \neq (a_1, \cdots, a_m) \geqslant 0$，并且

$$\inf_{x \in X} [a_1 g_1(x) + \cdots + a_m g_m(x)] \geqslant 0,$$

这与$\exists \bar{x} \in X$:对所有j有$g_j(\bar{x}) < 0$矛盾. 由$a_0 > 0$和(14.4)得到

$$\inf_{x \in X} \left[f(x) + \sum_j \underbrace{\left[\frac{a_j}{a_0}\right]}_{\lambda_j \geqslant 0} g_j(x)\right] \geqslant c.$$

■

按语： 该定理中的Slater条件可以替换成松弛Slater条件：存在可行点$\bar{x}$使得非线性不等式约束严格满足，即存在$\bar{x} \in X$对所有i满足$g(\bar{x}) \leqslant 0$，并且对那些不是放射函数的g_j满足$g_j(\bar{x}) < 0$.

定理 **14.5** (强对偶性)**.** 如果(ICP)是凸的(*X*是凸集; $f, g_1, \cdots, g_m$ 是X上的实值凸函数), 有下界, 并且满足松弛*Slater*条件, 那么存在$\lambda^* \geqslant 0$使得 $f_* = q^*$.

证明. 因为(ICP)是凸的, 有下界, 并且满足松弛Slater条件. 从而凸系统

$$f(x) < f_*,\ \ g_j(x) \leqslant 0, j = 1, \cdots, m, x \in X$$

无解, 同时问题

$$g_j(x) < 0, j = 1, \cdots, m, x \in X$$

有解. 根据凸择一定理(定理14.4，其中$c = f_*$)，

$$\exists \lambda^* \geqslant 0 : f(x) + \sum_j \lambda_j^* g_j(x) \geqslant f_*\ \ \forall x \in X,$$

再根据$q(\cdot)$的定义，有

$$q(\lambda^*) \geqslant f_*.$$

结合弱对偶性有

$$q^* = q(\lambda^*) = f_*.$$

■

请注意就等价关系而言，拉格朗日函数把(ICP)与它的对偶问题的关系牢记在心. 这种关系充其量为等价关系. 的确，

$$q^* = \sup_{\lambda \geqslant 0} \inf_{x \in X} L(x, \lambda)$$

是由拉格朗日函数给出的. 现在考虑函数

$$\overline{L}(x) = \sup_{\lambda \geqslant 0} L(x, \lambda) = \begin{cases} f(x), & g_j(x) \leqslant 0, j \leqslant m; \\ +\infty, & \text{否则}. \end{cases}$$

问题(ICP)显然等价于在$x \in X$上最小化$\overline{L}(x)$的问题，即

$$f_* = \inf_{x \in X} \sup_{\lambda \geqslant 0} L(x, \lambda).$$

14.4 鞍点与最优性条件

设$X \subset \mathbb{R}^n, \wedge \subset \mathbb{R}^m$是非空集合, $F(x, \lambda)$是$X \times \wedge$上的实值函数, 这个函数产生两个优化问题:

$$\text{Opt}(P) = \inf_{x \in X} \overbrace{\sup_{\lambda \in \wedge} F(x, \lambda)}^{\overline{F}(x)} \tag{P}$$

$$\text{Opt}(D) = \sup_{\lambda \in \wedge} \underbrace{\inf_{x \in X} F(x, \lambda)}_{\underline{F}(\lambda)} \tag{D}$$

二人零和博弈解释. 玩家 I 选择$x \in X$, 玩家 II 选择$\lambda \in \wedge$. 对于玩家选取的x, λ, 玩家 I 向玩家 II 支付$F(x, \lambda)$. 为了最优化他们的财富，玩家们该如何博弈?

如果玩家 I 先选择x, 玩家II知道该如何选择, 此时他将选择λ 以最大化自己利润, 而玩家 I 的损失将是$\overline{F}(x)$. 为了最小化该损失, 玩家 I 将求解问题(P), 以确保自己的损失是$\mathrm{Opt}(P)$或更少.

如果玩家 II 先选择λ, 并且玩家 I 做决策时，也知道玩家II的选择. 因此玩家 I 会最小化他的损失, 从而玩家 II 的利润将是$\underline{F}(\lambda)$. 玩家 II 为了最大化利润, 将求解问题(D), 确保其利润是$\mathrm{Opt}(D)$ 或更多.

观察到第二种情况看起来对玩家 I 更有利(在已有λ的先验信息下进行决策), 因此自然而然地猜测: 他的预期损失在该情况下小于等于他在第一种情况下所预期的损失:

$$\mathrm{Opt}(D) \equiv \sup_{\lambda\in\wedge}\inf_{x\in X} F(x,\lambda) \leqslant \inf_{x\in X}\sup_{\lambda\in\wedge} F(x,\lambda) \equiv \mathrm{Opt}(P).$$

这的确是事实. 假定$\mathrm{Opt}(P) < \infty$(否则该不等式关系显然成立),

$$\begin{aligned}
&\forall(\epsilon>0):\exists x_\epsilon \in X : \sup_{\lambda\in\wedge} F(x_\epsilon,\lambda) \leqslant \mathrm{Opt}(P)+\epsilon\\
&\Rightarrow \forall\lambda\in\wedge : \underline{F}(\lambda) = \inf_{x\in X} F(x,\lambda) \leqslant F(x_\epsilon,\lambda) \leqslant \mathrm{Opt}(P)+\epsilon\\
&\Rightarrow \mathrm{Opt}(D) \equiv \sup_{\lambda\in\wedge}\underline{F}(\lambda) \leqslant \mathrm{Opt}(P)+\epsilon\\
&\Rightarrow \mathrm{Opt}(D) \leqslant \mathrm{Opt}(P).
\end{aligned}$$

当玩家们同时做选择时, 他们该如何决策呢? 能够回答该问题的一种“良好情况"——F具有鞍点.

定义 **14.6.** 称点$(x_*,\lambda^*)\in X\times\wedge$是$F$的鞍点(saddle point)，如果

$$F(x,\lambda^*) \geqslant F(x_*,\lambda^*) \geqslant F(x_*,\lambda)\ \ \forall(x\in X,\lambda\in\wedge).$$

例子 **14.7.** 设$X=\mathbb{R},\Lambda=\mathbb{R}$,

$$F(x,\lambda) = -x\lambda, F(x,\lambda) = x^2-\lambda^2+x\lambda.$$

易见两种情况下，均有$F(x,0)\geqslant F(0,0)\geqslant F(0,\lambda)$. 所以$(0,0)$是二者的鞍点.

用博弈的术语说, 鞍点就是平衡点**(equilibrium)**——在该点, 倘若对手保持自己的选择不变的话，玩家无法增加自己的财富.

命题 **14.8** (鞍点集合的结构)**.** F有鞍点当且仅当(P)和(D)都是可解的, 且二者的最优值相等. 此时, F 的所有鞍点恰好是点对(x_*,λ^*), 其中x_* 是(P)的最优解, λ^*是(D) 的最优解.

证明 $\Rightarrow$: 假设(x_*,λ^*)是F的鞍点. 由鞍点定义得

$$F(x,\lambda^*) \geqslant F(x_*,\lambda^*) \geqslant F(x_*,\lambda)\ \ \forall(x\in X,\lambda\in\wedge)$$

随之有

$$\begin{aligned}
\mathrm{Opt}(P) &\leqslant \overline{F}(x_*) = \sup_{\lambda\in\wedge} F(x_*,\lambda) = F(x_*,\lambda^*),\\
\mathrm{Opt}(D) &\geqslant \underline{F}(\lambda^*) = \inf_{x\in X} F(x,\lambda^*) = F(x_*,\lambda^*).
\end{aligned}$$

由于$\mathrm{Opt}(P) \geqslant \mathrm{Opt}(D)$, 从而证明了不等式链:

$$\mathrm{Opt}(P) \leqslant \overline{F}(x_*) = F(x_*, \lambda^*) = \underline{F}(\lambda^*) \leqslant \mathrm{Opt}(D)$$

中的不等号都取等号. 这样, x_*求解(*P*), λ^* 求解(*D*), 并且$\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

$\Leftarrow$: 假设(*P*)和(*D*)的最优解分别为x_*和λ^*, 并且$\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

由x_*和λ^*的最优性,

$$\begin{aligned} \mathrm{Opt}(P) = \overline{F}(x_*) &= \sup_{\lambda \in \wedge} F(x_*, \lambda) \geqslant F(x_*, \lambda^*), \\ \mathrm{Opt}(D) = \underline{F}(\lambda^*) &= \inf_{x \in X} F(x, \lambda^*) \leqslant F(x_*, \lambda^*). \end{aligned} \tag{14.5}$$

由于$\mathrm{Opt}(P) = \mathrm{Opt}(D)$, 那么(14.5)中所有不等号都取等号, 因此

$$\sup_{\lambda \in \wedge} F(x_*, \lambda) = F(x_*, \lambda^*) = \inf_{x \in X} F(x, \lambda^*).$$

所以(x_*, λ^*)是F的鞍点. □

定理 **14.9** (凸规划鞍点形式的最优性条件)**.** 设$x_* \in X$.

(a) [充分条件] 倘若能将x_*扩展$\lambda^* \geqslant 0$得到拉格朗日函数在$X \times \{\lambda \geqslant 0\}$ 上的鞍点:

$$L(x, \lambda^*) \geqslant L(x_*, \lambda^*) \geqslant L(x_*, \lambda) \ \ \forall (x \in X, \lambda \geqslant 0), \tag{14.6}$$

那么x_*是(ICP)的最优解, 并且$\lambda_j^* g_j(x_*) = 0, j = 1, \cdots, m$.

(b) [必要条件] 假设(ICP)是凸的, 并且满足松弛*Slater*条件. 如果x_*是(ICP) 的最优解, 那么可将x_* 扩展$\lambda^* \geqslant 0$ 后得到拉格朗日函数在$X \times \{\lambda \geqslant 0\}$ 上的鞍点.

证明 $\Rightarrow$: 假设 $x_* \in X$, $\exists \lambda^* \geqslant 0$满足(14.6). 显然,

$$\sup_{\lambda \geqslant 0} L(x_*, \lambda) = \begin{cases} +\infty, & x_* \text{是可行的}; \\ f(x_*), & \text{其它}. \end{cases}$$

因此, $\lambda^* \geqslant 0$并且$L(x_*, \lambda^*) \geqslant L(x_*, \lambda) \ \ \forall \lambda \geqslant 0$ 等价于

$$g_j(x_*) \leqslant 0 \ \forall j \ \text{ 和 } \ \lambda_j^* g_j(x_*) = 0 \ \forall j.$$

结果有$L(x_*, \lambda^*) = f(x_*)$, 随之由

$$L(x, \lambda^*) \geqslant L(x_*, \lambda^*) \ \forall x \in X$$

得到

$$L(x, \lambda^*) \geqslant f(x_*) \ \forall x. \tag{14.7}$$

由于对于$\lambda \geqslant 0$及所有可行的x, 有$f(x) \geqslant L(x, \lambda)$, 因此(14.7)蕴含着

$$x \text{是可行的} \ \ \Rightarrow \ \ f(x) \geqslant f(x_*).$$

所以x_*是(ICP)的最优解.

$\Leftarrow$: 假设x_*是凸问题(ICP)的最优解, 并且该凸问题满足松弛Slater条件. 由拉格朗日对偶定理(定理14.5), $\exists \lambda^* \geqslant 0$:

$$f(x_*) = q(\lambda^*) \equiv \inf_{x \in X} \left[f(x) + \sum_j \lambda_j^* g_j(x) \right]. \tag{14.8}$$

由于x_*是可行解, 从而有

$$\inf_{x \in X} \left[f(x) + \sum_j \lambda_j^* g_j(x) \right] \leqslant f(x_*) + \sum_j \lambda_j^* g_j(x_*) \leqslant f(x_*).$$

根据式(14.8), 上式中最后的"$\leqslant$"取"=", 即$\lambda_j^* g_j(x_*) = 0 \ \forall j$, $\lambda^* \geqslant 0$, 由此得

$$f(x_*) = L(x_*, \lambda^*) \geqslant L(x_*, \lambda) \quad \forall \lambda \geqslant 0. \tag{14.9}$$

由(14.8)有

$$L(x, \lambda^*) \geqslant f(x_*) = L(x_*, \lambda^*).$$

再结合(14.9)得

$$F(x, \lambda^*) \geqslant F(x_*, \lambda^*) \geqslant F(x_*, \lambda), (x, \lambda) \in X \times \{\lambda \geqslant 0\}.$$

所以(x_*, λ^*)是L 的鞍点. □

定理 **14.10** (凸规划的KKT条件)**.** 在(ICP)中, 设X凸, f, g_j是凸的, x_* 是它的可行解, 函数$f, g_1, \cdots, g_m$在x_*处可微, 那么

(a) [充分条件] 如果存在拉格朗日乘子$\lambda^* \geqslant 0$ 使得

$$\langle \nabla f(x_*) + \sum_j \lambda_j^* \nabla g_j(x_*), x - x_* \rangle \geqslant 0 \quad \forall x \in X, \tag{14.10}$$

$$\lambda_j^* g_j(x_*) = 0, \ \ j \leqslant r \ \ [\text{互补松弛性}] \tag{14.11}$$

那么x_*是(ICP)的最优解.

(b) [充要条件] 如果(ICP)满足松弛*Slater*条件, 即$\exists \bar{x} \in X$满足$g(\bar{x}) \leqslant 0$ 且对任何非线性约束g_j满足$g_j(\bar{x}) < 0$, 那么x_*是(ICP)的最优解当且仅当它满足*KKT*条件.

证明. $\Rightarrow$: 设(ICP)是凸的, x_*是它的可行解, 并且f, g_j在x_*处可微. 假设此时KKT 条件也成立, 即存在$\lambda^* \geqslant 0$ 使得(14.10)和(14.11)成立.

的确, 互补松弛条件和$\lambda^* \geqslant 0$确保(14.9)成立. 进一步, $L(x, \lambda^*)$ 在$x \in X$上凸, 且在$x_* \in X$ 处可微, 因此(14.10)意味着

$$L(x, \lambda^*) \geqslant L(x_*, \lambda^*) \quad \forall x \in X.$$

这样, 可将x_*扩充λ^*得到拉格朗日函数的鞍点, 从而x_*是(ICP)的最优解.

$\Leftarrow$: [在**Slater**条件下] 设(ICP)是凸的且满足Slater条件, x_*是最优的, 并且f, g_j在x_*处可微.

根据鞍点最优性条件, 由x_*的最优性得到: $\exists \lambda^* \geqslant 0$使得$(x_*, \lambda^*)$是$L(x,\lambda)$在$X \times \mathbb{R}^m_+$ 上的鞍点, 这等价于

$$\lambda_j^* g_j(x_*) = 0 \ \ \forall j$$

和

$$\min_{x\in X} L(x,\lambda^*) = L(x_*,\lambda^*). \tag{14.12}$$

由于函数$L(x,\lambda^*)$在X上关于x是凸的, 在$x_* \in X$处可微, 关系式(14.12)意味着(14.10)成立. ■

例子 **14.11** (应用举例)**.** 假定$a_i > 0$, $p \geqslant 1$, 考虑求解问题

$$\min_{x>0} \left\{ \sum_i \frac{a_i}{x_i} : \ \sum_i x_i^p \leqslant 1 \right\}.$$

假定$x_* > 0$是该问题的解, 且满足$\sum\limits_i (x_i^*)^p = 1$, 由KKT条件得

$$\begin{aligned} \nabla_x \left\{ \sum_i \tfrac{a_i}{x_i} + \lambda \Big(\sum_i x_i^p - 1 \Big) \right\} \ &= \ 0 \ \Leftrightarrow \ \tfrac{a_i}{x_i^2} = p\lambda x_i^{p-1} \\ \sum_i x_i^p \ &= \ 1 \end{aligned}$$

随之得$x_i = c(\lambda) a_i^{\frac{1}{p+1}}$. 由于$\sum\limits_i x_i^p$应该等于 1, 得到

$$x_i^* = \frac{a_i^{\frac{1}{p+1}}}{\left(\sum\limits_j a_j^{\frac{p}{p+1}} \right)^{\frac{1}{p}}}.$$

该点是可行的, 问题是凸的, 所以该点满足KKT条件，所以x^* 是最优的!

下面用两个例子说明线性情况下如何构造对偶问题.

例子 **14.12.** 考虑原始问题

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \geqslant b. \end{aligned}$$

首先，原始问题的对偶函数

$$\begin{aligned} q(\lambda) &= \inf_{x\in\mathbb{R}^n} L(x,\lambda) = c^T x + \lambda^T (b - Ax) \\ &= \inf_{x\in\mathbb{R}^n} \left(c - A^T\lambda \right)^T x + \lambda^T b \\ &= \begin{cases} b^T\lambda, & c - A^T\lambda = 0 \\ -\infty, & \text{否则.} \end{cases} \end{aligned}$$

进而得对偶问题

$$\max_{\lambda \geqslant 0} \ q(\lambda).$$

它等价于

$$\begin{aligned} \max \quad & b^T\lambda \\ \text{s.t.} \quad & A^T\lambda = c, \lambda \geqslant 0. \end{aligned}$$

例子 **14.13.** 确定问题

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & b - Ax = 0, \\ & x \geqslant 0 \end{aligned}$$

的对偶问题. 首先，得到对偶函数

$$\begin{aligned} q(\mu) &= \inf_{x \geqslant 0} L(x, \mu) = c^T x + \mu^T (b - Ax) \\ &= \inf_{x \geqslant 0} \left(c - A^T \mu \right)^T x + \mu^T b \\ &= \begin{cases} b^T \mu, & c - A^T \mu \geqslant 0 \\ -\infty, & \text{否则.} \end{cases} \end{aligned}$$

进而得对偶问题

$$\max_{\mu \in \mathbb{R}^m} \ q(\mu).$$

它等价于

$$\begin{aligned} \max \quad & b^T \mu \\ \text{s.t.} \quad & A^T \mu \leqslant c. \end{aligned}$$

例子 **14.14.** 考虑二次优化中的二次规划问题

$$\begin{aligned} \min \quad & c^T x + \tfrac{1}{2} x^T Q x \\ \text{s.t.} \quad & Ax \geqslant b, \end{aligned}$$

其中$Q \in \mathbb{R}^{n \times n}$对称半正定，$c \in \mathbb{R}^n, b \in \mathbb{R}^m, A \in \mathbb{R}^{m \times n}$.

首先，得对偶函数

$$\begin{aligned} q(\lambda) &= \inf_{x \in \mathbb{R}^n} L(x, \lambda) = c^T x + \lambda^T (b - Ax) + \frac{1}{2} x^T Q x \\ &= \inf_{x \in \mathbb{R}^n} \left(c - A^T \lambda \right)^T x + \lambda^T b + \frac{1}{2} x^T Q x \\ &= \begin{cases} b^T \lambda - \frac{1}{2} x^T Q x, & c - A^T \lambda + Qx = 0 \\ -\infty, & \text{否则,} \end{cases} \end{aligned}$$

其中的等式约束是由$\nabla_x L(x, \lambda) = 0$ 得到的. 因此得对偶问题:

$$\max_{\lambda \geqslant 0} \ q(\lambda).$$

它等价于

$$\begin{aligned} \max \quad & b^T \lambda - \tfrac{1}{2} x^T Q x \\ \text{s.t.} \quad & c - A^T \lambda + Qx = 0 \\ & \lambda \geqslant 0. \end{aligned}$$

若 Q 可逆, 由等式约束解得 $x = Q^{-1}(A^T \lambda - c)$, 这样得到由$\lambda$显式表示的对偶问题:

$$\max_{\lambda \geqslant 0} b^T \lambda - \frac{1}{2} (A^T \lambda - c)^T Q^{-1} (A^T \lambda - c).$$

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

考虑仿射等式约束优化问题

$$\begin{aligned} \underset{x\in X}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & Ax = b, \end{aligned} \tag{15.1}$$

它的**Lagrange**函数

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

对偶函数(dual function)

$$q(\lambda) := \inf_{x\in X} L(x,\lambda).$$

对偶问题(dual problem)是

$$\sup_{\lambda\in\mathbb{R}^m} q(\lambda).$$

请注意，这里为了讨论方便，仅把等式约束作为显式约束，从而这里的$\lambda_i \in \mathbb{R}$. 如果用$Ax \geqslant b$代换$Ax = b$，对偶问题就需要添加约束$\lambda_i \geqslant 0$.

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X不是凸的)对偶函数总是凹的.

证明. 对任何 $x \in X$, $L(x,\lambda)$关于λ是线性函数，因此$q(\lambda)$ 是一族线性函数的逐点下确界，因此是凹的. ■

由于对偶函数q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$\bar{\lambda}$使得

$$X(\bar{\lambda}) := \arg\inf_{x\in X} L(x,\lambda) \neq \emptyset,$$

那么$-q$在λ处是次可微的，并且

$$b - Ax(\lambda) \in \partial(-q(\lambda)),$$

其中$x(\lambda) \in X(\bar{\lambda})$. 如果$X(\bar{\lambda})$是单点集，那么$q$在$\bar{\lambda}$可微，并且

$$\nabla q(\bar{\lambda}) = Ax(\lambda) - b,$$

其中$x(\bar{\lambda}) = \arg\inf_{x\in X} L(x,\bar{\lambda})$.

15.2 对偶梯度上升法

对偶函数$g(\lambda)$的凹性确保$-q(\lambda)$的次梯度是存在的，因此可用(次)梯度法来极小化$-q(\lambda)$，即用 (次)梯度法极大化q，称对应的方法是求解原始问题 (15.1)的 对偶梯度上升(dual gradient ascent)法:

从初始点λ_0开始. 对所有$t \geqslant 0$:

$$x_t = \arg\inf_{x\in X} L(x, \lambda_t)$$

$$\lambda_{t+1} = \lambda_t + \eta(Ax_t - b)$$

由次梯度法(定理 2.11)知该算法的收敛速率是$O(1/\sqrt{t})$.

对偶梯度法的主要优点是简单. 此外，对于特殊问题由它易于得到可并行化的更新规则，即对偶分解法. 假设能将原始问题 (15.1)的自变量x剖分成大小为$(n_i)_{i=1}^N$的块，并且满足

$$x = (x^{(1)}, \cdots, x^{(N)}) \qquad x^{(i)} \in \mathbb{R}^{n_i}, \sum_{i=1}^N n_i = n$$

$$X = X^{(1)} \times \cdots \times X^{(N)} \qquad X^{(i)} \subset \mathbb{R}^{n_i}$$

$$A = [A_1 \cdots A_N] \qquad Ax = \sum_{i=1}^N A_i x^{(i)}$$

$$f(x) = \sum_{i=1}^N f_i(x^{(i)})$$

那么，Lagrange函数

$$L(x,\lambda) = \sum_{i=1}^N (\underbrace{f_i(x^{(i)}) + \lambda^T A_i x^{(i)} - \tfrac{1}{N}\lambda^T b}_{L_i(x^{(i)},\lambda)}) = \sum_{i=1}^N L_i(x^{(i)}, \lambda)$$

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: $x_t^{(i)} = \arg\inf_{x^{(i)}\in X^{(i)}} L_i(x^{(i)}, \lambda_t)$
- 在主节点: $\lambda_{t+1} = \lambda_t + \eta(Ax_t - b)$

例子 **15.4** (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 **15.5** (网络效用最大化)**.** 假设拥有的网络共有L条链路，第ℓ条链路的容量是c_ℓ. 感兴趣的是为K 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$x_k \in \mathbb{R}$代表分给流k的数量，$U_k : \mathbb{R} \to \mathbb{R}$是凹效用函数，其返回流$k$数量为$x_k$时所得到的效用大小. 最优化问题是

$$\max_{x \geqslant 0} \sum_{k=1}^{K} U_k(x_k) : Rx \leqslant c$$

其中R是$L \times K$的路由矩阵，如果流k 的路由通过链路ℓ，它的第(ℓ, k)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:

$$\min_{x \geqslant 0} -\sum_{k} U_k(x_k) : Rx \leqslant c.$$

那么对偶问题是

$$\max_{\lambda \geqslant 0} \min_{x \geqslant 0} \sum_{k} -U_k(x_k) + \lambda^T (Rx - c)$$

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$\lambda^T (\sum_{k=1}^{K} [R_k x_k - c/K])$, 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$k$ 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

$$x_t^{[k]} \in \arg\max_{x_k \geqslant 0} U_k(x_k) - \lambda_t^T R_k x_k$$

主节点计算

$$\lambda_{t+1} = [\lambda_t + \eta(Rx_t - c)]_+$$

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的值可解释成经济学中资源的“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.

15.3 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):

$$\lambda_{t+1} = \mathrm{prox}_{\eta_t q}(\lambda_t) = \arg\sup_{\lambda} \underbrace{\underbrace{\inf_{x \in X} f(x) + \lambda^T (Ax - b)}_{q(\lambda)} - \underbrace{\frac{1}{2\eta_t} \|\lambda - \lambda_t\|^2}_{\text{proximal term}}}_{h(\lambda)}$$

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X是凸集时应

用Sion定理)那么可以重写

$$\sup_{\lambda} \inf_{x\in X} f(x) + \lambda^T(Ax-b) - \tfrac{1}{2\eta_t}\|\lambda-\lambda_t\|^2 \tag{15.2}$$

$$= \inf_{x\in X} \sup_{\lambda} f(x) + \lambda^T(Ax-b) - \tfrac{1}{2\eta_t}\|\lambda_t-\lambda\|^2 \tag{15.3}$$

$$= \inf_{x\in X} f(x) + \lambda_t^T(Ax-b) + \tfrac{\eta_t}{2}\|Ax-b\|^2 \tag{15.4}$$

这里的第二个等式是因为内部sup的最优解具有闭合式

$$\lambda(x) = \lambda_t + \eta_t(Ax-b), \tag{15.5}$$

将其代入，可以得到第二个等式. 为了单独考虑剩下的关于x的优化问题，做如下定义.

定义 **15.6** (增广Lagrange函数)**.** 问题 (15.1)的增广**Lagrange**函数**(augmented Lagrangian)**

$$L_\eta(x,\lambda) = f(x) + \lambda^T(Ax-b) + \tfrac{\eta}{2}\|Ax-b\|^2$$

综上，从对偶临近点的角度出发，新迭代点λ_{t+1}是 MaxiMin问题(15.2)解中的λ部分，如果强对偶成立的话，就是MiniMax问题(15.3)解中的λ部分，而这个鞍点问题中的x问题为(15.4)，记该问题的解记为x_t，代入(15.5)，就得到鞍点问题中的λ，也即此处的新迭代λ_{t+1}. 由此得增广**Lagrange**函数法(Augmented Lagrangian Method, ALM)，亦称乘子法(Method of Multipliers, MM)的更新方式:

$$x_t = \arg\inf_{x\in X} L_{\eta_t}(x,\lambda_t)$$

$$\lambda_{t+1} = \lambda_t + \eta_t(Ax_t - b)$$

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

$$q(\lambda_t) - q^* \geqslant -\tfrac{\|\lambda^*\|^2}{2\eta t}.$$

15.4 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. 考虑

$$\begin{aligned} \underset{x,y}{\text{minimize}} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax + By = c, \end{aligned}$$

这里目标和约束能拆分成x和y两块. 该问题的增广Lagrange函数

$$L_\eta(x,y,\lambda_t) = f(x) + g(y) + \lambda_t^T(Ax+By-c) + \frac{\eta}{2}\|Ax+By-c\|_2^2.$$

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:

$$(x_{t+1}, y_{t+1}) = \inf_{x,y} L_\eta(x, y, \lambda_t)$$
$$\lambda_{t+1} = \lambda_t + \eta(Ax_{t+1} + By_{t+1} - c).$$

请注意，这里的二次惩罚项$\|Ax + By - c\|_2^2$使得增广Lagrange函数关于x, y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. 具体地， ADMM在关于x和y优化增广Lagrange函数之间交替地("ADMM" 中的A)进行一次:

$$x_{t+1} = \inf_x L_\eta(x, y_t, \lambda_t)$$
$$y_{t+1} = \inf_y L_\eta(x_{t+1}, y, \lambda_t)$$
$$\lambda_{t+1} = \lambda_t + \eta(Ax_{t+1} + By_{t+1} - c)$$

这与乘子法不同，后者不能并行化，因为在y_{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

$$L_\eta(x, y, \lambda) = f(x) + g(y) + \lambda^T(Ax + By - c)$$

具有鞍点x_*, y_*, λ^*, 即

$$\forall x, y, \lambda : L(x_*, y_*, \lambda) \leqslant L(x_*, y_*, \lambda^*) \leqslant L(x, z, \lambda^*).$$

那么，随着$t \to \infty$, ADMM 满足$f(x_t) + g(y_t) \to p^*, \lambda_t \to \lambda^*$，其中$p_* = f(x_*) + g(y_*)$.

例子 **15.8** (求解LASSO的ADMM)**.** 考虑$\|\cdot\|_1$正则化的线性回归问题

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \tau\|x\|_1, \tag{15.6}$$

其中$A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$已知，$\tau > 0$是参数. 可将该问题等价表述成共识优化问题

$$\begin{aligned} \underset{x,y \in \mathbb{R}^n}{\text{minimize}} \quad & \tfrac{1}{2}\|Ax - b\|^2 + \tau\|y\|_1 \\ \text{subject to} \quad & x = y = 0. \end{aligned} \tag{15.7}$$

这是一个可分离问题，它的增广Lagrange函数

$$L_\eta(x, y\lambda) = \frac{1}{2}\|Ax - b\|^2 + \tau\|y\|_1 + \lambda^T(x - y) + \frac{1}{2\eta}\|x - y\|_2^2,$$

易于验证求解(15.7)的ADMM：已知$y_0, u_0, t = 0$,

$$x_{t+1} = \left(A^TA + \frac{1}{\eta}I\right)^{-1}\left(A^Tb + \frac{1}{\eta}(y_t - u_t)\right),$$
$$y_{t+1} = S_{\eta\tau}(x_{t+1} + u_t),$$
$$u_{t+1} = u_t + (x_{t+1} - y_{t+1}),$$

其中$u = \eta\lambda, S_\rho : \mathbb{R} \to \mathbb{R}$是软阈值算子

$$S_\rho(a) = \begin{cases} a-\rho, & a > \rho \\ 0, & |a| \leqslant \rho \\ a+\rho, & a < -\rho \end{cases}$$

请注意，迭代公式中的软阈值算子是逐分量运算.

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

$$f^*(p) := \sup_{x \in \mathrm{dom} f} px - f(x)$$

是f的Legendre变换. 对于已知的p，设这里的上确界在$x(p)$处取到. 图16.1给出了f与f^*关系的几何直观. Fenchel共轭是Legendre 变换对非凸和(或者)不可微函数的扩展. 当函数是可微凸函数时，Fenchel共轭就还原成Legendre变换.

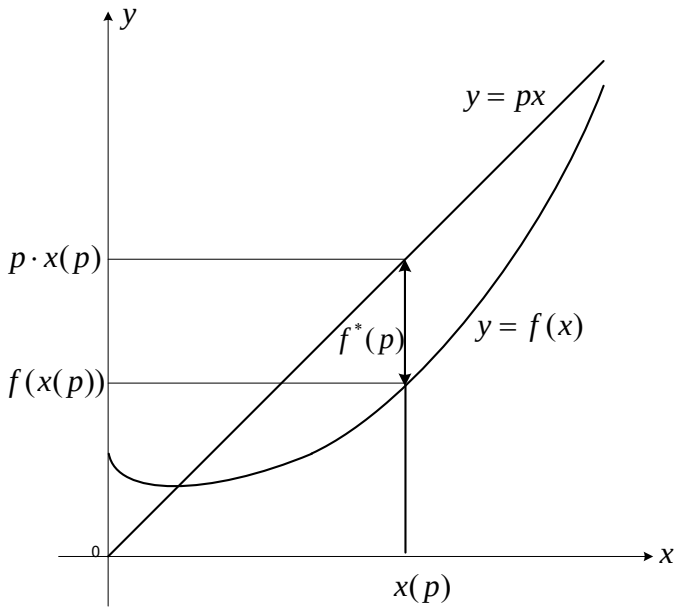


图 16.1: Legendre变换. 函数$f^*(p)$是为了让直线$y = px$在$(x, f(x))$ 与f的图形(上图) 相切，而需要垂直平移f上图的大小，即将epi$f$垂直平移$f*(p)$，那么直线$y = px$支撑epi$f$.

定义 **16.1** (Fenchel共轭)**.** 函数$f : \mathbb{R}^n \to \mathbb{R}$的**Fenchel**共轭(Fenchel Conjugate)是

$$f^*(p) = sup_{x \in \mathrm{dom} f} \langle p, x \rangle - f(x)$$

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$\partial f^*(p) = \{x : p \in \partial f(x)\}$.

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

$$\min_{x \in \mathrm{dom} f \cap \mathrm{dom} g} f(x) - g(x) = \max_{p \in \mathrm{dom} g^* \cap \mathrm{dom} f^*} g^*(p) - f^*(p)$$

其中g^*是g的凹共轭，定义为$g^*(p) := \inf_x \langle p, x\rangle - g(x)$.

图16.2给出了一维情况时Fenchel对偶性的几何直观.

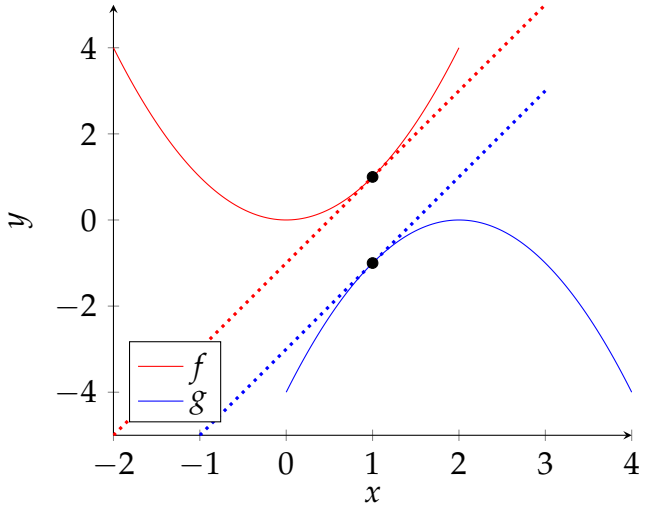


图 16.2: 一维中的Fenchel对偶性

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：

$$\begin{aligned} \underset{x,y \in \mathbb{R}^n}{\text{minimize}} \quad & f(x) - g(y) \\ \text{subject to} \quad & x = y. \end{aligned} \tag{16.1}$$

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线性函数的系数向量w), x_i 看作训练集中第i 个样例的特征，对应输出是$\langle w, x_i\rangle$. $\phi_i(\langle \cdot, x_i\rangle)$ 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 $g(w) = \frac{1}{2}\|w\|_2^2$，超参数$\lambda > 0$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

$$\begin{aligned}
D(\alpha) &= \min_{w,z}\frac{1}{m}\sum_{i=1}^{m}\phi_i(z_i) + \lambda g(w) + \frac{1}{m}\alpha^\top(z - X^\top w) \\
&= \min_{w,z}\frac{1}{m}\sum_{i=1}^{m}[\phi_i(z_i) + \alpha_i z_i] + \lambda g(w) - \frac{1}{m}\alpha^\top X^\top w \\
&= -\left[\max_{w,z} -\sum_{i=1}^{m}\frac{1}{m}[\phi_i(z_i) + \alpha_i z_i] + \frac{1}{m}(X\alpha)^\top w - \lambda g(w)\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\max_{z_i}[-\phi_i(z_i) - \alpha_i z_i] - \lambda\max_{w}[(\tfrac{X\alpha}{m\lambda})^\top w - g(w)] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\phi_i^*(-\alpha_i) - \lambda g^*(\tfrac{X\alpha}{m\lambda})
\end{aligned}$$

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. 在这种情况下对偶问题 (16.4)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，这表明最优线性函数的系数向量属于数据生成的子空间. 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 极大化对偶函数. 算法的伪码描述如下:

Algorithm 6 stochastic dual coordinate ascent method (SDCA)

Require: $\{(x_i, y_i)\}_{i=1}^m$, parameter $\lambda > 0$ and initialization $\alpha_0, t = 1$.

1: Start from $w^0 := w(\alpha^0)$.
2: **for** $t = 1$ to T **do**
3: Randomly pick $i \in [m]$.
4: Find $\Delta\alpha_i$ which maximizes

$$-\phi_i^*\left(-(\alpha_i^{t-1} + \Delta\alpha_i)\right) - \frac{m\lambda}{2}\left\|w^{t-1} + \frac{1}{m\lambda}x_i\Delta\alpha_i\right\|^2 \tag{16.6}$$

5: Update the dual solution: $\alpha^t = \alpha^{t-1} + \Delta\alpha_i e_i$.
6: Update the primal solution: $w^t = w^{t-1} + \frac{1}{m\lambda}\Delta\alpha_i x_i$.
7: **end for**

Ensure: Either $\bar{w} = \frac{1}{T}\sum_{t=1}^T w_t$ or $w^\circ \in \operatorname{argmin}_{w \in \{w_1, \cdots, w_T\}} P(w)$

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:

$$\Delta\alpha_i = \frac{y_i - x_i^T w^{t-1} - 0.5\alpha_i^{t-1}}{0.5 + \|x\|^2/(\lambda m)}.$$

请注意算法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

$$\epsilon_D^t \leqslant P(w^t) - D(\alpha^t).$$

进一步，

$$D(\alpha^t) - D(\alpha^{t-1}) = \epsilon_D^{t-1} - \epsilon_D^t.$$

给该等式两边取期望，并应用引理 16.9, 得到

$$\begin{aligned}\mathbb{E}[\epsilon_D^{t-1} - \epsilon_D^t] &= \mathbb{E}[D(\alpha^t) - D(\alpha^{t-1})] \\ &\geqslant \frac{s}{m}\mathbb{E}[P(w^{t-1}) - D(\alpha^{t-1})] \\ &\geqslant \frac{s}{m}\mathbb{E}[\epsilon_D^{t-1}].\end{aligned}$$

重新整理，并递归地应用前面的讨论产生

$$\mathbb{E}[\epsilon_D^t] \leqslant \left(1 - \frac{s}{m}\right)\mathbb{E}[\epsilon_D^{t-1}] \leqslant \left(1 - \frac{s}{m}\right)^t \epsilon_D^0.$$

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：

$$\begin{aligned}\mathbb{E}[P(w^t) - P(w^*)] &\leqslant \mathbb{E}[P(w^t) - D(\alpha^t)] \\ &\leqslant \frac{m}{s}\mathbb{E}[D(\alpha^{t+1}) - D(\alpha^t)] \\ &\leqslant \frac{m}{s}\mathbb{E}[\epsilon_D^t],\end{aligned}$$

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Lagrange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Lagrange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- $z_1, \ldots, z_d$

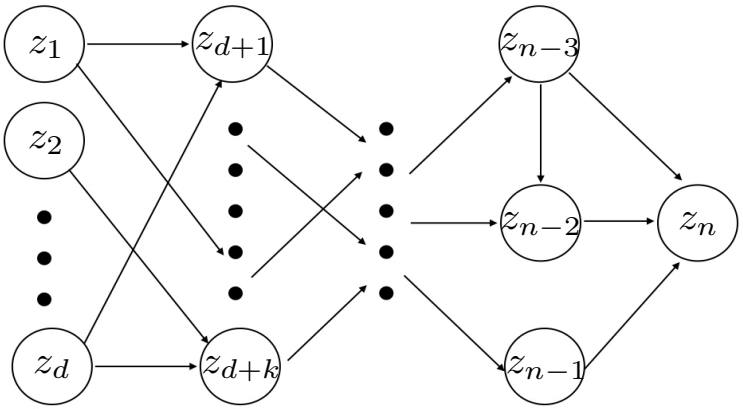


图 17.1: 计算图

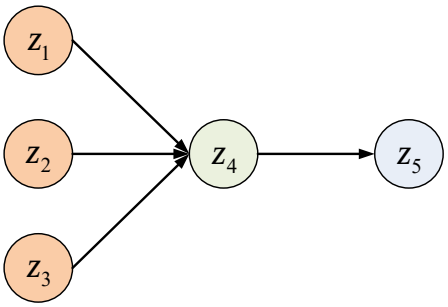


图 17.2: 热身例子的计算图，其中$d = 3, L = 3, n = 5$

– 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Lagrange函数是

$$\mathcal{L}(z, \lambda) = z_n + \sum_{i=1}^{n} \lambda_i (f_i(z_{\alpha(i)}) - z_i).$$

图17.2是热身例子的计算图. 像该例中做的那样，置 $\nabla \mathcal{L} = 0$. 可将这看作一个两步算法:

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.

证明. 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.

基本情况: $\lambda_n = 1 = \frac{\partial f_n(x)}{\partial z_n} = \frac{\partial z_n}{\partial z_n}$.

归纳: 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，

$$\begin{aligned}
\lambda_j &= \sum_{i \in \beta(j)} \lambda_i \frac{\partial f_i(z_{\alpha(i)})}{\partial z_j} && \text{(BP)} \\
&= \sum_{i \in \beta(j)} \frac{\partial f(x)}{\partial z_i} \frac{\partial z_i}{\partial z_j} && \text{(因为}z_{\beta(j)}\text{属于层}p+1\text{由归纳假设和 (BP))} \\
&= \frac{\partial f(x)}{\partial z_j} && \text{(由多元链式法则).}
\end{aligned}$$

■

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.** $\Lambda_1 = \Lambda B^T, \Lambda_2 = A^T\Lambda$

证明.

$$f = \sum_{i,j} \lambda_{ij} C_{ij} = \sum_{i,j} (AB)_{ij} = \sum_{i,j} \lambda_{ij} \Sigma_k a_{ik} b_{kj}.$$

如此，由Lagrange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

$$(\Lambda_2)_{pq} = (A^T\Lambda)_{pq}$$

■

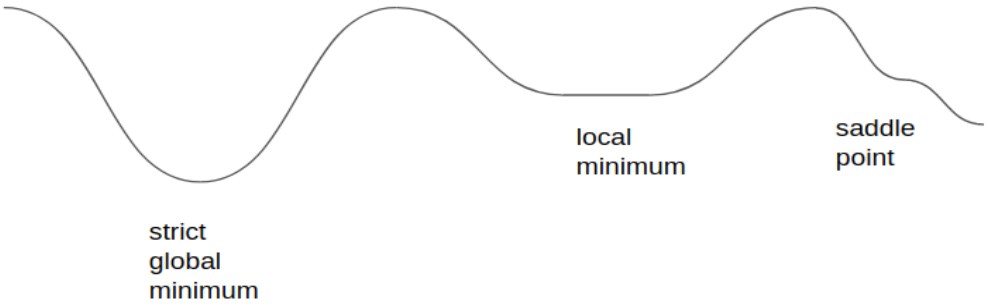
Part V
非凸优化：无约束问题

18 非凸问题

设$\Omega \subset \mathbb{R}^n, f : \mathbb{R}^n \to \mathbb{R}$，如果$\Omega$和$f$中至少有一个非凸，那么

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

就是非凸问题. 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

首先着手讨(SCO)的属于Ω内部，也即f的无约束局部极小点的必要和充分条件.

定义 **18.1** (局部极小点)**.** 称点$x_* \in \Omega$是(SCO)的局部极小点**(local minimum)**，如果存在$\delta > 0$使得对于所有满足$\|x - x_*\| < \delta$的$x \in \Omega$有 $f(x_*) \leqslant f(x)$成立.

定义 **18.2** (全局极小点)**.** 称点$x_* \in \Omega$是(SCO)的全局极小点，如果对于所有$x \in \Omega$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$x_* \in \operatorname{Int}\Omega$是(SCO)的局部极小点，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) 进一步，如果f在包含x_*的开集上是二次连续可微的，那么$\nabla^2 f(x_*) \succeq 0$.

证明. 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)**，如果梯度在x_*消失，即 $\nabla f(x_*) = 0$.

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 . 比如对一维的情况对于$f(x) = x^3$而言，$x_* = 0$ 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)**.** 假设$f \colon \mathbb{R}^n \to \mathbb{R}$在$x^*$处二阶连续可微. 如果$x_* \in \operatorname{Int} \Omega$ 使得$\nabla f(x_*) = 0$并且$\nabla^2 f(x_*) \succ 0$ (正定). 那么 x_* 是(SCO)的严格局部极小点.

证明. 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

$$f(x,y)=\frac{1}{2}x^2+\frac{1}{4}y^4-\frac{1}{2}y^2,$$

它的梯度

$$\nabla f(x,y)=\begin{bmatrix} x \\ y^3-y \end{bmatrix},$$

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 线搜索与**Armijo**法则

本节假设Ω是开集，及考虑求f的无约束极小点的算法. 为此，不妨设$X=\mathbb{R}^n$. 针对非凸问题(SCO)，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，可以保证它收敛到驻点.

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$d^T\nabla f(x)$对于η而言是常数. 因此，充分小的正步长η 能使得$f(x')<f(x)$. 接下来讨论用灵活的方法挑选一个步长.

已知下降方向d (比如 $d=-\nabla f(x)$), 设步长

$$\eta_*\in\operatorname*{argmin}_{\eta\in\mathbb{R}}\phi(\eta):=f(x+\eta d).$$

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. 一般的函数都需要数值方法来求解. 数值方法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，称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

$$\phi(\eta)\leqslant\phi(0)+\rho\phi'(0)\eta \tag{18.5}$$

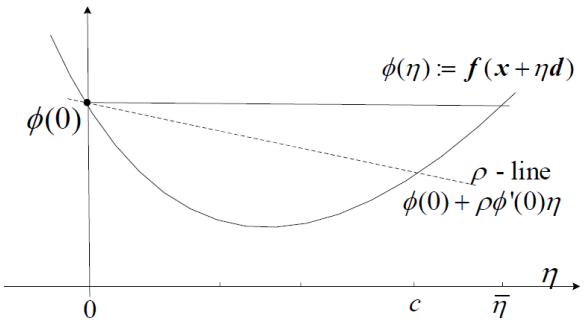


图 18.1: 非精确线搜索示意图

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(直线$y = \phi(0) + \rho\phi'(0)\eta$)的图示. 这里的Armijo条件 (18.5)要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$\gamma^{-1}\bar{\eta}, \gamma^{-2}\bar{\eta}, \gamma^{-3}\bar{\eta}, \cdots$，直到当前值首次不满足 (18.5)时终止，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

$$\gamma = \tfrac{1}{2}, \quad \rho = \tfrac{1}{100}, \quad \bar{\eta} = 1\,.$$

18.3 最速下降法的全局收敛性

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.

证明. 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$\widehat{\eta}_t = \frac{\eta_t}{\gamma}$，由 (18.9)可继续进行如下推导

$$\begin{aligned}
&\tfrac{f(x_t) - f(x_t - \widehat{\eta}_t \nabla f(x_t))}{\widehat{\eta}_t} < \rho\|\nabla f(x_t)\|^2 \\
\Rightarrow &\nabla f(x_t - \theta_t \nabla f(x_t))^T \nabla f(x_t) < \rho\|\nabla f(x_t)\|^2, \text{其中}\theta_t \in (0, \widehat{\eta}_t) \qquad (18.10)\\
\Rightarrow &\|\nabla f(x_t)\|^2 \leqslant \rho\|\nabla f(x_t)\|^2 \qquad (18.11)
\end{aligned}$$

不等式(18.10)是由Lagrange中值定理(Mean Value Theorem, MVT) 得到的. 取极限，由于$\eta_t \to 0 \Rightarrow \widehat{\eta}_t \to 0$ 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，如果非精确线搜索确定的步长满足Armijo法则，就能保证梯度下降法收敛到驻点.

19 逃离鞍点

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？ 为此，首先进一步将驻点分成不同类别. 鞍点是一类重要的驻点.

定义 **19.1** (鞍点)**.** 假设$f \in C^1$. 称f的非局部最优(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 **19.2.** 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.

19.1 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 19.1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$\nabla f(x)$连续可微的情况下，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.

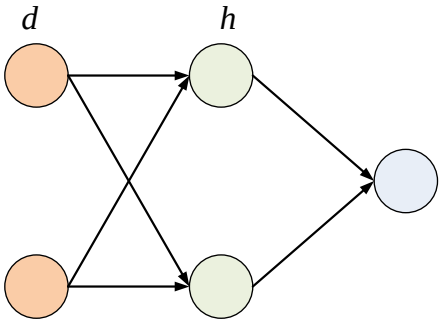


图 19.1: 两层全连接神经网络

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]

下面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]. 先给出严格鞍点的定义.

定义 **19.3** (严格鞍点)**.** 针对二次连续可微函数 f, 驻点x_*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，则称x_*是f的严格鞍点**(strict saddle point)**.

19.2 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.

19.3 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.

19.4 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. ■

命题 **19.10.** 令 T 在开域 D 上 *Fréchet* 可微. 令 $x \in D$ 且对所有 α, $0 \leqslant \alpha \leqslant 1$ 有 $x + \alpha h \in D$, 则

$$\|T(x+h) - T(x)\| \leqslant \|h\| \sup_{0<\alpha<1} \|T'(x+\alpha h)\|.$$

证明. 令 y^* 为 Y^* 的一个非零元素, 与元素 $T(x+h) - T(x)$ 对齐. 函数 $\varphi(\alpha) = y^*[T(x+\alpha h)]$ 定义在区间 $[0,1]$ 上, 由链式法则得到导数为

$$\varphi'(\alpha) = y^*[T'(x+\alpha h)h].$$

根据实变量函数的均值定理可知

$$\varphi(1) - \varphi(0) = \varphi'(\alpha_0), \ \ 0 < \alpha_0 < 1,$$

于是

$$|y^*[T(x+h) - T(x)] \leqslant \|y^*\| \sup_{0<\alpha<1} \|T'(x+\alpha h)\|\|h\|,$$

由于 y^* 与 $T(x+h) - T(x)$共线, 所以

$$\|T(x+h) - T(x)\| \leqslant \|h\| \sup_{0<\alpha<1} \|T'(x+\alpha h)\|.$$

■

20 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

$$x_{t+1} = x_t - \frac{\varphi(x_t)}{\varphi'(x_t)}.$$

牛顿法的几何直观就是利用当前点处的切线作为函数图形的近似，将切线与横轴的交点(一阶Taylor展式的零点)作为下一个迭代点，具体如图 20.1所示.

20.1 二次收敛

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

$$x_{t+1} = x_t - J_F^{-1}(x_t)F(x_t).$$

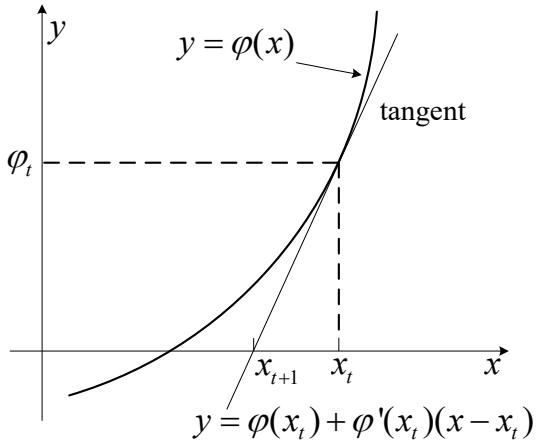


图 20.1: 单变量函数求根的牛顿法的几何直观

.

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 **20.1.** 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

$$\|x_{t+1} - x_*\| \leqslant \tfrac{\beta}{\alpha} \|x_t - x_*\|^2.$$

证明. 先证明如下断言， 即初始点充分靠近x_*时，牛顿法是良定义的.

断言 **20.2.** Hessian阵$\nabla^2 f(x_t)$正定，并且满足$\|\nabla^2 f(x_t)^{-1}\| \leqslant \frac{2}{\alpha}$.

证明. 由Wielandt-Hoffman定理[8]，有

$$\begin{aligned} |\lambda_{\min}(\nabla^2 f(x_t)) - \lambda_{\min}(\nabla^2 f(x_*))| &\leqslant \|\nabla^2 f(x_t) - \nabla^2 f(x_*)\| \\ &\leqslant \beta \|x_t - x_*\| \quad (\nabla^2 f(x)\text{是}\beta\text{-Lipschitz连续的}) \end{aligned}$$

[8]Hoffman-Wielandt不等式：设A, B是n 阶Hermite矩阵，矩阵的特征值排列为$\lambda_1(\cdot) \geqslant \cdots \geqslant \lambda_n(\cdot)$.

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

$$\lambda_{\min}(\nabla^2 f(x_t)) \geqslant \tfrac{\alpha}{2}.$$

所以$\nabla^2 f(x_t)$是正定的，并且因此，

$$\|\nabla^2 f(x_t)^{-1}\| \leqslant \tfrac{2}{\alpha}.$$

■

因为$\nabla^2 f(x_t)$正定，从而牛顿法有定义，考虑基本牛顿迭代与局部极小点之间的差向量，再结合已知$\nabla f(x_*) = 0$, 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 **20.3.** $\|\nabla^2 f(x_t)(x_t - x_*) - [\nabla f(x_t) - \nabla f(x_*)]\| \leqslant \frac{\beta}{2}\|x^t - x_*\|^2$

证明. 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

$$\begin{aligned} &\|\nabla^2 f(x_t)(x_t - x_*) - [\nabla f(x_t) - \nabla f(x_*)]\| \\ &= \left\| \int_0^1 [\nabla^2 f(x_t) - \nabla^2 f(x_t + \gamma(x_t - x_*))](x_t - x_*) \, \mathrm{d}\gamma \right\| \\ &\leqslant \int_0^1 \|\nabla^2 f(x_t) - \nabla^2 f(x_t + \gamma(x_t - x_*))\| \cdot \|x_t - x_*\| \, \mathrm{d}\gamma \\ &\leqslant \beta \left(\int_0^1 \gamma \, \mathrm{d}\gamma \right) \|x_t - x_*\|^2 \qquad (\nabla^2 f(x_t)\text{是}\beta\text{−Lipschitz的}) \\ &= \tfrac{\beta}{2}\|x_t - x_*\|^2 \end{aligned}$$

■

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.

将这两个断言放在一起，有

$$\|x_{t+1} - x_*\| \leqslant \tfrac{2}{\alpha} \cdot \tfrac{\beta}{2}\|x_t - x_*\|^2 = \tfrac{\beta}{\alpha}\|x_t - x_*\|^2.$$

■

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)**.

20.2 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

$$f'(x) = \frac{x}{\sqrt{x^2+1}}$$
$$f''(x) = (1+x^2)^{-3/2}.$$

请注意$f(x)$的二阶导数是严格正的，从而是严格凸的，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.

21 拟牛顿法

将梯度下降法和牛顿法放到一起比较：

$$x_{k+1} = x_k - \eta_t \nabla f(x_k), \qquad \text{(梯度下降法)}$$
$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1}\nabla f(x_k). \qquad \text{(牛顿法)}$$

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 $\nabla^2 f(x_k)^{-1}$, 即当

$$\nabla^2 f(x_k)^{-1} = \eta_k I$$

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. 想要确定二次近似

$$\widehat{f}_{B_{k+1}}(x) \approx f(x_{k+1}) + \nabla f(x_{k+1})^\top (x - x_{k+1}) + \frac{1}{2}(x - x_{k+1}) B_{k+1}(x - x_{k+1})$$

满足：

$$\nabla \widehat{f}_{B_{k+1}}(x_{k+1}) = \nabla f(x_{k+1}) \tag{21.1}$$

和

$$\nabla \widehat{f}_{B_{k+1}}(x_k) = \nabla f(x_k). \tag{21.2}$$

条件(21.1)表明该近似中的一阶项是精确的，这看起来是合理的. 条件(21.2)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$x_{k+1} - x_k$的Hessian阵近似是合理的. 称(21.2) 是割线近似**(secant approximation)**，可写作

$$\nabla \widehat{f}_{B_{k+1}}(x_k) = \nabla f(x_{k+1}) + B_{k+1}(x_k - x_{k+1}) = \nabla f(x_k).$$

如果令

$$\begin{aligned} p_k &= x_{k+1} - x_k \\ q_k &= \nabla f(x_{k+1}) - \nabla f(x_k), \end{aligned}$$

那么割线近似即

$$B_{k+1} p_k = q_k,$$

或者

$$H_{k+1} q_k = p_k, \tag{21.3}$$

其中$H_{k+1} = B_{k+1}^{-1}$. 存在多个B_{k+1}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_{k+1}是正定的、确信B_{k+1}对于某种度量尽可能接近B_k，或者要求B_{k+1} 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.

21.1 低秩校正

称(21.3)是割线方程(secant equation). 有不同的更新方式可以满足割线方程.

秩**-1**更新：

$$H_{k+1} = H_k + a_k z_k z_k^T,$$

其中 $a_k \in \mathbb{R}$, $z_k \in \mathbb{R}^n$. 希望选取a_k, z_k使得H_{k+1}满足(21.3)：

$$p_k = H_{k+1} q_k = H_k q_k + a_k \left(z_k^T q_k \right) z_k,$$

所以 z_k与$p_k - H_k q_k$共线. 若选取 $z_k = p_k - H_k q_k$, 并强制比例系数是1，即

$$1 = a_k \left(z_k^T q_k \right) = a_k (p_k - H_k q_k)^T q_k \implies a_k = \frac{1}{(p_k - H_k q_k)^T q_k},$$

得更新方式:

$$H_{k+1} = H_k + \frac{(p_k - H_k q_k)(p_k - H_k q_k)^T}{(p_k - H_k q_k)^T q_k}.$$

秩-**2**更新：对于 $a, b \in \mathbb{R}$ 以及 $u, v \in \mathbb{R}^n$ 有

$$H_{k+1} = H_k + auu^T + bvv^T.$$

割线方程(21.3)蕴含着

$$p_k = H_{k+1} q_k = H_k q_k + a(u^T q_k)u + b(v^T q_k)v.$$

如果选取 $u = p_k$ 以及 $v = H_k q_k$, 并强制

$$a(p_k^T q_k) = 1 \implies a = \frac{1}{p_k^T q_k}$$
$$b(q_k^T H_k q_k) = -1 \implies b = -\frac{1}{q_k^T H_k q_k},$$

那么，得到 **Davidon-Fletcher-Powell(DFP)**法, 其更新方式为

$$H_{k+1}^{\mathrm{DFP}} = H_k + \frac{p_k p_k^T}{p_k^T q_k} - \frac{H_k q_k q_k^T H_k}{q_k^T H_k q_k}.$$

这是第一个拟牛顿法.

引理 **21.1** (Cauchy-Schwartz不等式)**.** 对于 $c, d \in \mathbb{R}^n$, 有 $|c^T d| \leqslant \|c\|\|d\|$, 当且仅当 c, d共线时，等式成立.

定理 **21.2.** 对所有 $k \geqslant 0$, 若 $p_k^T q_k > 0$, 那么*DFP*法得到的所有 H_k 是正定对称阵.

证明. 用归纳法. 对于 $k = 0$ 时, 由假设 $H_0 \succ 0$, 结论显而易见. 假设对 $k \geqslant 0$ 有 $H_k \succ 0$. 已知 $u \neq 0$, 则

$$u^T H_{k+1}^{\mathrm{DFP}} u = u^T H_k u + \frac{(p_k^T u)^2}{p_k^T q_k} - \frac{(q_k^T H_k u)^2}{q_k^T H_k q_k}.$$

令 $c = H_k^{1/2} u$ 以及 $d = H_k^{1/2} q_k$, 则由引理 21.1得

$$\begin{aligned} u^T H_{k+1} u &= \|c\|^2 - \frac{(c^T d)^2}{\|d\|^2} + \frac{(p_k^T u)^2}{p_k^T q_k} \\ &= \frac{\|c\|^2\|d\|^2 - (c^T d)^2}{\|d\|^2} + \frac{(p_k^T u)^2}{p_k^T q_k} \geqslant 0. \end{aligned}$$

如果c和d不共线，由引理 21.1易见$u^T H_{k+1} u > 0$. 假设c和d共线. 又因为H_k非奇异，所以u和q_k共线，从而存在 $\lambda \neq 0$使得$u = \lambda q_k$. 于是 $p_k^T u = \lambda q_k^T p_k \neq 0$. 所以$H_{k+1} \succ 0$. ■

对于 $\alpha_k > 0$, 如何保证条件

$$0 < q_k^T p_k = (g_{k+1} - g_k)^T(\alpha_k d_k) = \alpha_k(g_{k+1}^T d_k - g_k^T d_k).$$

成立? 强制要求$g_{k+1}^T d_k > g_k^T d_k$ 就够了. 在 $0 < \rho < \sigma < 1$ 的情况下, Wolfe-Powell 线搜索便是这种非精确线搜索的一个范例. 特别地, 它满足以下条件:

(1) $f(x_k + \alpha_k d_k) \leqslant f(x_k) + \alpha_k \rho g_k^T d_k$.

(2) $g_{k+1}^T d_k \geqslant \sigma g_k^T d_k > g_k^T d_k$.

其它的秩**-2**更新方法: 尝试如下迭代方案

$$x_{k+1} = x_k - \alpha_k B_k^{-1} g_k, \;\; B_k \approx \nabla^2 f(x_k),$$

其中 B_{k+1} 通过秩-2更新公式

$$B_{k+1}^{\mathrm{BFGS}} = B_k + \frac{q_k q_k^T}{q_k^T p_k} - \frac{B_k p_k p_k^T B_k}{p_k^T B_k p_k}$$

得到，其满足$B_{k+1} p_k = q_k$, 这是著名的**Broyden-Fletcher-Goldfarb-Shannon (BFGS)** 更新. 大量数值结果表明，这是目前最好的拟牛顿法.

命题 **21.3** (Sherman-Morrison公式)**.** n阶方阵A的秩$m(\leqslant n)$修正通常可以写作 $A = B + USV^T$, 其中 $S \in \mathbb{R}^{m\times m}$, $A, B \in \mathbb{R}^{n\times n}$ 非奇异, $U, V \in \mathbb{R}^{n\times m}$. 若 $P = S^{-1} + V^T S^{-1} U$ 是非奇异的, 则

$$A^{-1} = B^{-1} - B^{-1} U P^{-1} V^T B^{-1}.$$

利用 Sherman-Morrison 公式，其中 $A = B_{k+1}^{\mathrm{BFGS}}$, $B = B_k$, $U = [q_k, B_k p_k]$, $V = U$ 以及

$$S = \begin{bmatrix} \frac{1}{p_k^T q_k} & 0 \\ 0 & -\frac{1}{p_k^T B_k p_k} \end{bmatrix}$$

可得到B_{k+1}^{BFGS}的逆矩阵，即

$$H_{k+1}^{\mathrm{BFGS}} = \left(B_{k+1}^{\mathrm{BFGS}}\right)^{-1} = H_k + \left(1 + \frac{q_k^T H_k q_k}{q_k^T p_k}\right)\frac{p_k p_k^T}{p_k^T q_k} - \frac{p_k q_k^T H_k + H_k q_k p_k^T}{q_k^T p_k},$$

Broyden族算法**:** 设 $\phi = \phi_k \in \mathbb{R}$,

$$H_{k+1}^{\phi} = (1-\phi) H_{k+1}^{\mathrm{DFP}} + \phi H_{k+1}^{\mathrm{BFGS}} = H_{k+1}^{\mathrm{DFP}} + \phi v_k v_k^T,$$

其中

$$v_k = (q_k^T H_k q_k)^{1/2}\left(\frac{p_k}{p_k^T q_k} - \frac{H_k q_k}{q_k^T H_k q_k}\right).$$

定理 **21.4.** 若 $H_k \succ 0$, $p_k^T q_k > 0$, $\phi \geqslant 0$, 则 $H_{k+1}^{\phi} \succ 0$.

21.2 收敛性

定理 **21.5.** 设 $f(x)=\frac{1}{2}x^TQx-b^Tx$, 其中$Q$对称正定. 若对每个 $k\geqslant 0$有$g_k\neq 0$，则

(i) $H_{k+1}^{\phi}q_j=p_j, j=0,1,\cdots,k.$

(ii) $p_j^TQp_i=0, 0\leqslant i<j\leqslant k.$

(iii) $p_0,p_1,\cdots,p_k$ 不含零向量.

于是, 该方法迭代 $m\leqslant n$ 次后终止. 若 $m=n$, 则 $H_n=Q^{-1}$.

由于 $q_j=Qp_j$, 因此 $H_{k+1}q_j=p_j\implies(H_{k+1}Q)p_j=q_j, j=0,1,\cdots,k$, 所以 $H_{k+1}Q$ 就像特定子空间上的单位算子. 特别地, 对所有 $x\in[p_0,\cdots,p_k]$, $(H_{k+1}Q)x=x$.

定理 **21.6.** 若 $H_0=I$, 则通过 *Broyden*族拟牛顿法 *(*用精确线搜索*)* 产生的迭代与共轭梯度法产生的相同.

定理 **21.7** (一般 f 的收敛性)**.** 令 $f:\mathbb{R}^n\to\mathbb{R}\in C^2(\mathbb{R}^n)$ 以及 $x_0\in\mathbb{R}^n$ 使得$S=\{x\in\mathbb{R}^n:f(x)\leqslant f(x_0)\}$ 是有界凸集，且$\nabla^2f(x)\succ 0,\forall x\in S$. 设 $\{x_k\}$ 为 *Broyden*族拟牛顿法产生的序列：

$$x_{k+1}=x_k-\alpha_kH_k^{\phi_k}g_k,$$

其中 $\phi_k\in[0,1]$, $H_0=I$, 根据*Wolfe-Powell*规则选取 α_k, 且第一次尝试步长 $\alpha_k=1$. 则在

$$\lim_{k\to\infty}\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|}=0$$

的意义下，$\{x_k\}$超线性收敛于f 在 S 上的唯一全局极小点x^*.

现在针对非二次函数的极小化问题，比较拟牛顿法和共轭梯度法. 拟牛顿法的一个优点是当线搜索可以精确执行时，算法不仅生成了共轭方向，而且生成了Hesse矩阵的逆矩阵的近似值. 因此，当收敛到具有正定Hesse矩阵的局部极小点时，它类似于牛顿法，具有较快的收敛速率. 重要的是，这个性质并不依赖于初始矩阵H_0. 因此，这个算法并不需要像共轭梯度法那样，利用最速下降步重新开始新的计算.

第二个优点是牛顿法对线搜索的精确度并不像共轭梯度法那么敏感. 这一点已经被大量的实验所证实. 一种可能的解释是，这种方法生成正定矩阵H_k以及下降方向并不受线搜索精确度的影响.

为了进一步比较共轭梯度法和拟牛顿法，考虑当n很大时，这两种算法在每一步的计算量. 共轭梯度法每次迭代需要计算函数值及其梯度(在线搜索中为了确定步长，可能需要计算多次), 计算搜索方向和下一个迭代点需要$O(n)$次运算. 牛顿法需要大体相同的计算量来计算目标函数的值和梯度，计算搜索方向和下一个迭代点需要$O(n^2)$次运算. 如果目标函数值的计算量要远大于或者相当于$O(n^2)$次运算的计算量，那么牛顿法比起共轭梯度法在每步迭代中仅需要稍微多一点的计算量，却可以具有上面提到的优点. 对于目标函数及其梯度计算量需要的计算时间远小于$O(n^2)$次运算的问题，共轭梯度法更合适. 在n非常大的的一些离散控制问题中，目标函数和梯度的计算量通常需要$O(n)$次运算，这时共轭梯度法就更适用于这些问题.

一般来说，共轭梯度法和拟牛顿法的每一步迭代需要的计算量都少于牛顿法. 牛顿法每次迭代的计算量包括计算目标函数值、梯度和Hesse矩阵, 和计算牛顿方向($O(n^3)$). 这些计算开销抵消了牛顿法的快速收敛. 然而，在一些问题中，利用特殊的结构来高效地计算牛顿方向，比如在有些最优控制问题的牛顿法中，每次迭代仅需要$O(n)$次计算，远远低于拟牛顿法的$O(n^2)$. 详细参加Bertsekas的1.9节**[?]**.

21.3 有限内存拟牛顿法

拟牛顿法的一般公式

$$\begin{aligned}\phi(H,p,q) &= H + \left(1+\frac{q^THq}{p^Tq}\right)\frac{pp^T}{p^Tq} - \frac{pq^TH+Hqp^T}{p^Tq}\\ &= \left(I-\frac{pq^T}{p^Tq}\right)H\left(I-\frac{qp^T}{p^Tq}\right)+\frac{pp^T}{p^Tq},\end{aligned}$$

特别地, $H_k^{\text{BFGS}} = \phi(H_{k-1}, p_{k-1}, q_{k-1})$. 有限内存类方法的思想是存储最新的$m$对数据对 $(p_i, q_i), i = k-1, \cdots, i, k-m$, 并通过如下步骤递归生成 H_k:

步 **1.** $H = H_0^k$ (简单地, $H = I$)

步 **2.** 置 $H \leftarrow \phi(H, p_i, q_i), i = k-m, \cdots, k-1.$

步 **3.** $H_k = H.$

该方案易于计算 $H_k g_k$, 最大的优势是不用存储中间矩阵H，从而适用于大规模问题. 该算法的完整描述如下.

Algorithm 7 Limited memory quasi-Newton methods (For computing $H_k g_k$)

Require: $k, m, (p_i, q_i)$ for $i = k-1, \cdots, k-m$.

$u = g_k$
for $i = k-1$ to $k-m$ **do**
  $\alpha_k = \frac{g_k^T g_k}{d_k^T Q d_k}$
  $u = u - \alpha_i q_i$
end for
$r = H_0^k u$
for $i = k-m$ to $k-1$ **do**
  $\beta_i = \frac{q_i^T r}{p_i^T q_i}$
  $r = r + (\alpha_i - \beta_i) p_i$
end for

Ensure: r

22 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)

23 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)

24 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)

Part VI
非凸优化：约束问题

考虑一般的约束优化问题(MP)，感兴趣的问题假设已经有可行解x_*, 那么它是最优解的条件(必要条件、充分条件、充分必要条件)是什么?

事实是除了凸规划外, 还没有可验证的关于全局最优性的局部充分条件. 然而存在关于局部(也因此关于全局)最优性的可验证必要条件和关于局部最优性的充分条件. 另一个事实是关于局部最优性的现有(一阶和二阶)条件都假设$x_* \in \mathrm{int}X$. 从x_*的局部最优性角度讲, 这与$X = \mathbb{R}^n$ 的描述完全一样. 所以，在讨论局部最优性时均假设$X = \mathbb{R}^n$，即考虑

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & g_i(x) \leqslant 0,\ \ i = 1, \cdots, m \\ & h_j(x) = 0,\ \ j = 1, \cdots, \ell \end{array} \tag{P}$$

其中 $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$. 问题(P)的可行域:

$$\Omega = \{x \in \mathbb{R}^n : g_i(x) \leqslant 0,\ i \leqslant m,\ h_j(x) = 0,\ j \leqslant \ell\}.$$

在局部理论中，使用微积分作为工具，仅能给出局部最优解的一阶和二阶条件.

定义 **24.1.** 称 $x_* \in \Omega$ 是(P) 的局部极小点, 如果存在 $\delta > 0$ 使得对所有满足 $\|x - x_*\| < \delta$ 的 $x \in \Omega$ 有 $f(x_*) \leqslant f(x)$.

定义 **24.2.** 称 $x_* \in \Omega$ 是(P) 的全局极小点, 如果对所有的 $x \in \Omega$ 有 $f(x_*) \leqslant f(x)$ 成立.

如果上述定义中的不等式对于 $x \neq x_*$ 是严格的, 分别称作"严格局部极小点"和"严格全局极小点". 下面当$f, g_i, h_j \in C^1$，或者 $f, g_i, h_j \in C^2$ 时, 刻画问题(P) 的局部极小点，得到的条件分别称作一阶和二阶最优性条件.

25 一阶最优性条件

25.1 切锥与几何最优性条件

定义 **25.1** (切锥)**.** 设 $\Omega \subseteq \mathbb{R}^n$ 非空, $\bar{x} \in \mathrm{cl}\,\Omega$. 若存在点列 $\{x_t\} \subseteq \Omega \setminus \{\bar{x}\}$ 和正标量序列 $\{\delta_t\}$, $\delta_t \to 0$ 满足

$$d = \lim_{t\to\infty} \frac{x_t - \bar{x}}{\delta_t}, \tag{25.1}$$

称 d 是 Ω 在 $\bar{x}$ 的切方向. 称 Ω 在 $\bar{x}$ 的切向量的全体是 Ω 在 $\bar{x}$ 的切锥(**tangent cone**)，记作$T_\Omega(\bar{x})$.

命题 **25.2** (局部极小点的几何最优性条件)**.** 设 x_* 是(P)的局部极小点, 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

$$d^T \nabla f(x_*) \geqslant 0,\ \ \forall d \in T_\Omega(x_*).$$

证明. 任取 $0 \neq d \in T_\Omega(x_*)$. 由切向量的定义, 存在点列 $\{x_t\} \subseteq \Omega \setminus \{x_*\}$ 和正标量序列 $\{\delta_t\}$, $\delta_t \to 0$ 满足

$$d = \lim_{t\to\infty} \frac{x_t - x_*}{\delta_t}. \tag{25.2}$$

由一阶 Taylor 展式得

$$f(x_t) = f(x_*) + \nabla f(x_*)^T (x_t - x_*) + o(\|x_t - x_*\|).$$

从而

$$\frac{f(x_t) - f(x_*)}{\delta_t} = \nabla f(x_*)^T \frac{x_t - x_*}{\delta_t} + \frac{o(\delta_t)}{\delta_t}$$

因为 x_* 是局部极小点, 且$x_t \to x_*$, 所以对充分大的 t 有 $f(x_t) - f(x_*) \geqslant 0$. 令 $t \to \infty$, 由(25.2)得 $d^T \nabla f(x_*) \geqslant 0$. ■

定义 **25.3** (积极集)**.** 已知(P)的可行点 $\bar{x} \in \Omega$. 定义 $\bar{x}$ 处的积极集为

$$\mathcal{A}(\bar{x}) = \{1, \cdots, \ell\} \cup \mathcal{J}(\bar{x}),$$

其中 $\mathcal{J}(\bar{x}) = \{i : g_i(\bar{x}) = 0,\ i \leqslant m\}$.

定义了点 $\bar{x}$ 处的积极约束. 因此，如果 $\bar{x}$ 在可行域的边界上, 则确定对应边界的约束就是积极的. 如果 $\bar{x}$ 可行, 显然有 $\mathcal{J}(\bar{x}) \subseteq \mathcal{A}(\bar{x})$. 问题(P)的解 x_* 处的积极约束特别重要. 如果事先知道这个集合, 忽略其余的非积极约束, 而将 $\mathcal{A}(x_*)$ 中的约束变成等式，则 x_* 是这个新问题的局部解. 而且，将 $i \notin \mathcal{A}(x_*)$ 的约束进行微小的扰动并不影响 x_* 的局部最优性, 但这个事实对积极约束通常是不成立的. 如果读者完成课后对应的练习题后，将会对积极约束有更深刻的理解.

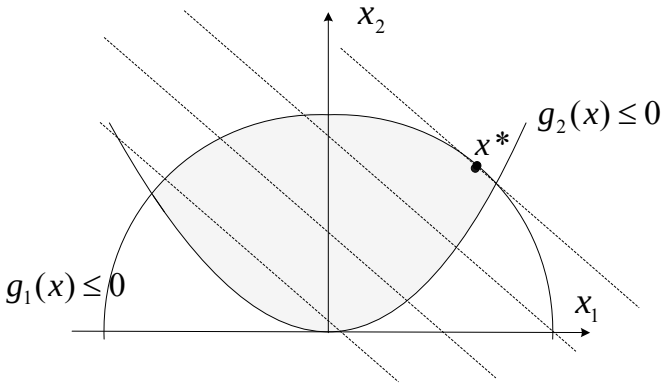


图 25.1: 积极约束和非积极约束.

例子 **25.4** (积极约束)**.** 考虑

$$\begin{array}{ll} \underset{x\in\mathbb{R}^2}{\text{minimize}} & f(x) = -x_1 - x_2 \\ \text{subject to} & g_1(x) = x_1^2 - x_2 \leqslant 0 \\ & g_2(x) = x_1^2 + x_2^2 - 1 \leqslant 0. \end{array} \tag{25.3}$$

由图 25.1 知解 $x_* = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$，且积极集 $\mathcal{A}(x_*) = \{2\}$, 即圆约束 $g_2(x)$ 是积极的，抛物线约束 $g_1(x)$ 是非积极的, 对其进行扰动或者去掉它均不会改变解 x_*. 当考虑二阶最优性条件时，需要将积极约束的定义进一步精细化为强积极和弱积极约束, 详见图 26.1.

当g_i可微时，图 25.2 给出了$\bar{x}$处一个积极约束的梯度向量$\nabla g(\bar{x})$的几何直观，它也是该不等式约束的边界所对应曲线在这点的法向量.

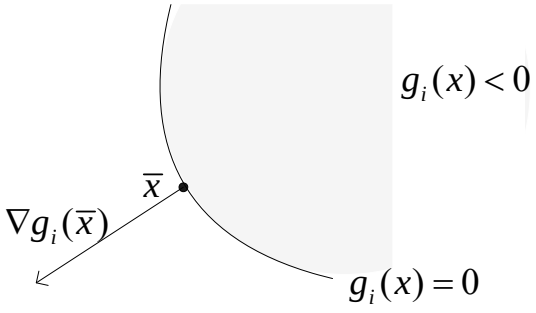


图 25.2: 积极约束的法向量.

25.2 一阶必要条件

集合的切锥仅由可行域决定，反应了可行域的本质特征, 但不容易计算. 下面引入线性化可行方向锥的概念.

定义 **25.5** (线性化可行方向锥)**.** 定义 $\bar{x}$ 处的线性化可行方向锥为

$$F(\bar{x}) = \left\{ d \in \mathbb{R}^n : \begin{array}{ll} d^T \nabla h_j(\bar{x}) = 0, & j \leqslant \ell \\ d^T \nabla h_i(\bar{x}) \leqslant 0, & i \in \mathcal{J}(\bar{x}) \end{array} \right\}.$$

例子 **25.6.** 求 $\Omega = \{x \in \mathbb{R}^2 : x_2 \leqslant x_1^3,\ x_2 \geqslant 0\}$ 在 $\bar{x} = (0,0)$ 处的线性化可行方向锥 $F(\bar{x})$.

解答**:** 这里$\mathcal{J}(\bar{x}) = \{1,2\}$. 进而

$$\nabla g_1(\bar{x}) = (0,1), \nabla g_2(\bar{x}) = (0,-1), F(\bar{x}) = \{(d_1,0) \in \mathbb{R}^2 : d_1 \in \mathbb{R}\}.$$

下面讨论集合在某点的切锥与线性化可行方向锥的关系.

命题 **25.7** (切锥与线性化可行方向锥的关系)**.** 设 $\bar{x} \in \Omega$, 且假设 g_i, h_j 在包含 $\bar{x}$ 的开集上连续可微 $(g_i, h_j \in C^1)$, 那么 $T_\Omega(\bar{x}) \subseteq F(\bar{x})$.

证明. 任取 $0 \neq d \in T_\Omega(\bar{x})$. 由切向量的定义, 存在点列 $\{x_t\} \subseteq \Omega \setminus \{\bar{x}\}$ 和正标量序列 $\{\delta_t\}$, $\delta_t \to 0$ 满足(25.1).

对 $j \leqslant \ell$, 由一阶 Taylor 展式得

$$\frac{h_j(x_t) - h_j(\bar{x})}{\delta_t} = \nabla h_j(\bar{x})^T \frac{x_t - \bar{x}}{\delta_t} + \frac{o(\delta_t)}{\delta_t}.$$

令 $t \to \infty$, 由 x_t 可行和(25.1)得 $d^T \nabla h_j(\bar{x}) = 0$. 对 $i \in \mathcal{J}(\bar{x})$, 由一阶 Taylor 展式得

$$\frac{g_i(x_t) - g_i(\bar{x})}{\delta_t} = \nabla g_i(\bar{x})^T \frac{x_t - \bar{x}}{\delta_t} + \frac{o(\delta_t)}{\delta_t}.$$

令 $t \to \infty$, 由 x_t 可行, $g_i(\bar{x}) = 0$ 和(25.1)得 $d^T \nabla g_i(\bar{x}) \leqslant 0$. ■

由上述定义看到，集合在某点的线性化可行方向锥的优点是易于计算, 但有可能反应不了可行域的本质特征. 此外，它还受可行域 Ω 的代数表示方式的影响.

例 25.6(续) $\Omega = \{x \in \mathbb{R}^2 : x_2 \leqslant x_1^3,\ x_2 \geqslant 0\}$, $\bar{x} = (0,0)$, 有

$$F(\bar{x}) = \{(d_1, 0) : d_1 \in \mathbb{R}\}.$$

由图 25.3中展示的几何直观，得

$$T_\Omega(\bar{x}) = \{(d_1, 0) : d_1 \geqslant 0\}.$$

因此 $T_\Omega(\bar{x}) \subsetneq F(\bar{x})$. 请注意这里是真包含关系.

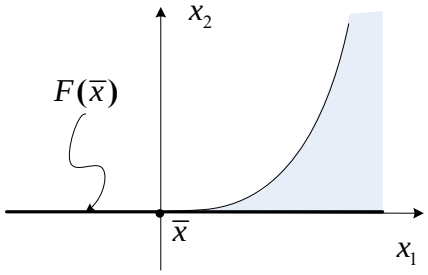


图 25.3: 切锥与线性化可行方向锥的关系，这里切锥真包含于线性化可行方向锥.

下面引入约束品性(constraint quality, CQ)的概念，它是一族条件，用以保证 $F(\bar{x}) \subseteq T_\Omega(\bar{x})$ 的充分条件. 后面会看到有各种各样的约束品性. 当某种 CQ 成立时, 可用 $F(\bar{x})$ 代替 $T_\Omega(\bar{x})$刻画最优解.

下面是最有名，也最常用的约束品性，即积极约束梯度线性无关约束品性.

定义 **25.8** (线性无关约束品性)**.** 设 $\bar{x} \in \Omega$. 如果 $\bar{x}$ 处积极约束的梯度线性无关, 即 $\nabla h_1(\bar{x}), \cdots, \nabla h_\ell(\bar{x}), \nabla g_i(\bar{x}),\ i \in \mathcal{J}(\bar{x})$, 线性无关, 则称 $\bar{x}$ 处线性无关约束品性 **(linear independence constraint quality, LICQ)** 成立.

命题 **25.9.** 假设 $g_i, h_j \in C^1$. 如果 $\bar{x} \in \Omega$ 处 *LICQ* 成立, 那么 $T_\Omega(\bar{x}) = F(\bar{x})$.

证明. 根据命题 25.7，仅需证明$F(\bar{x}) \subset T_\Omega(\bar{x})$. 任取 $d \in F(\bar{x})$, 下面利用隐函数定理，构造点列 $\{x_t\} \subseteq \Omega \setminus \{\bar{x}\}$ 和正标量序列 $\{\delta_t\}$, $\delta_t \to 0$ 满足(25.1).

不妨设 $\mathcal{A}(\bar{x}) = \{1, \cdots, \ell\} \cup \{1, \cdots, k\}$, 即

$$g_i(\bar{x}) = 0,\ i \leqslant k,\ g_i(\bar{x}) < 0,\ i > k.$$

因为 $\bar{x}$ 处 LICQ 成立, 所以 $\ell + k \leqslant n$，且可将积极约束的梯度向量扩充$b_1, \cdots, b_{n-\ell-k}$得到$\mathbb{R}^n$的基. 构造方程组:

$$\begin{aligned}
&h_j(x) - \theta d^T \nabla h_j(\bar{x}) = 0,\ j \leqslant \ell,\\
&g_i(x) - \theta d^T \nabla g_i(\bar{x}) = 0,\ i \leqslant k,\\
&b_i^T (x - \bar{x}) - \theta d^T b_i = 0, i = 1, \cdots, n - \ell - k.
\end{aligned}$$

该方程组共有n 个方程，$n+1$ 个未知数 (x, θ)；且$(\bar{x}, 0)$ 满足方程. 此外，由极大线性无关组扩充知, 方程组在 $(\bar{x}, 0)$ 处 Jacobi 矩阵的前 n 列线性无关. 根据隐函数定理, 存在 $\bar{x}$ 的开邻域 $N_{\bar{x}}$ 和 $\theta = 0$ 的开邻域 N_0 满足

(i) 对任意 $\theta \in N_0$, 该方程组有唯一解 $x(\theta) \in N_{\bar{x}}$, 并且

(ii) $x(\theta)$ 关于 θ 连续可微.

记

$$p = \frac{\mathrm{d}x(\theta)}{\mathrm{d}\theta}|_{\theta=0},$$

将 $x(\theta)$ 代入方程组, 由隐函数求导法则和由极大线性无关组扩充知, 得 $p = d$. 从而, 取 $\delta_t \in N_0$, $\delta_t > 0$ 且 $\delta_t \to 0$, 由 (i) 知 $x_t := x(\delta_t)$ 满足方程组; 进一步由 $d \in F(\bar{x})$和$\delta_t > 0$ 知 $\{x_t\} \subseteq \Omega$.

由 (ii) 有

$$\lim_{t\to\infty} \frac{x_t - \bar{x}}{\delta_t} = \lim_{t\to\infty} \frac{x(\delta_t) - \bar{x}}{\delta_t} = d.$$

从而 $d \in T_\Omega(\bar{x})$. ■

定义 **25.10** (线性约束品性)**.** 设 $\bar{x} \in \Omega$. 如果 $\bar{x}$ 处积极约束都是线性的, 即 $h_1, \cdots, h_\ell, g_i$, $i \in \mathcal{J}(\bar{x})$ 是线性函数, 称 $\bar{x}$ 处线性约束品性 (LCQ) 成立.

命题 **25.11.** 设 $\bar{x} \in \Omega$. 如果 $\bar{x}$ 处 *LCQ* 成立, 那么 $T_\Omega(\bar{x}) = F(\bar{x})$.

证明. 任取 $d \in F(\bar{x})$, 令 $x_t = \bar{x} + \frac{1}{t}d$, $\delta_t = \frac{1}{t}$, 易见 $\{\delta_t\}$ 是正标量序列，且 $\delta_t \to 0$;

再由 $F(\bar{x})$ 的定义和 LCQ 成立, 可验证 $\{x_t\} \subseteq \Omega \setminus \{\bar{x}\}$, 并且(25.1)成立. 从而 $d \in T_\Omega(\bar{x})$. ■

应用**:** 对只含线性约束的优化问题, 比如线性规划和二次规划, 自然有 $T_\Omega(\bar{x}) = F(\bar{x})$.

回顾命题 25.2的几何最优性条件: 设 $f \in C^1$, x_* 是(P) 的局部极小点可，那么

$$d^T \nabla f(x_*) \geqslant 0, \ \ \forall d \in T_\Omega(x_*).$$

这等价于

$$\{d \in \mathbb{R}^n : d^T \nabla f(x_*) < 0\} \cap T_\Omega(x_*) = \emptyset. \tag{25.4}$$

然后引入了约束品性. 现在假设局部极小点处某种约束品性成立, 从而 $T_\Omega(x_*) = F(x_*)$. 综合这两点，若x_* 是(P) 的局部极小点，则(25.4)成立, 此即系统

$$\begin{aligned} &d^T \nabla f(x_*) < 0, \\ &d^T \nabla h_j(x_*) = 0, \ \ j \leqslant \ell \\ &d^T \nabla g_i(x_*) \leqslant 0, \ \ i \in \mathcal{J}(x_*) \end{aligned}$$

无解. 该事实等价于

$$\begin{aligned} 0 \in \operatorname{argmin} \ \ & d^T \nabla f(x_*) \\ \text{s.t.} \ \ & d^T \nabla h_j(x_*) = 0, \ \ j \leqslant \ell \\ & d^T \nabla g_i(x_*) \leqslant 0, \ \ i \in \mathcal{J}(x_*). \end{aligned}$$

由前面凸优化的最优性条件，存在 $\mu_j^*, j \leqslant \ell, \lambda_i^* \geqslant 0, i \in \mathcal{J}(x_*)$ 使得

$$\nabla f(x_*) + \sum_{j=1}^{\ell} \mu_j^* \nabla h_j(x_*) + \sum_{i\in\mathcal{J}(x_*)} \lambda_i^* \nabla g_i(x_*) = 0.$$

为非积极约束补充乘子: $\lambda_i^* = 0, i \notin \mathcal{J}(x_*)$. 则上式等价于存在 $\lambda_i^* \geqslant 0, i \leqslant m$, μ_j^*, $j \leqslant \ell$ 满足

$$\nabla f(x_*) + \sum_{j=1}^{\ell} \mu_j^* \nabla h_j(x_*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x_*) = 0,$$

和互补松弛条件 $\lambda_i^* g_i(x_*) = 0, i \leqslant m$. 综上，得到了著名的最优性的一阶必要条件，也称作Karush-Kuhn-Tucker条件，简称KKT条件.

定理 **25.12** (一阶必要条件)**.** 设 $f, g_i, h_j \in C^1$, x_* 是(P)的局部极小点，且 x_* 处 *LICQ* 或者 *LCQ* 成立, 那么存在 λ_i^*, μ_j^* 满足

$$\begin{aligned} \nabla f(x_*) + \sum_{j=1}^{\ell} \mu_j^* \nabla h_j(x_*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x_*) = 0, & \quad \text{对偶可行性} \\ \lambda_i^* \geqslant 0, \quad i \leqslant m & \quad \text{对偶可行性} \\ h_j(x_*) = 0, \quad j \leqslant \ell & \quad \text{原始可行性} \\ g_i(x_*) \leqslant 0, \quad i \leqslant m & \quad \text{原始可行性} \\ \lambda_i^* g_i(x_*) = 0, \quad i \leqslant m. & \quad \text{互补松弛条件} \end{aligned} \tag{25.5}$$

称满足 KKT 条件的三元对 (x_*, λ^*, μ^*) 是 KKT 对, 称其中的 x_* 是问题(P)的 KKT 点, (λ^*, μ^*) 是与 x_* 对应的 Lagrange 乘子. 当对每个 $i \leqslant m$, λ_i^* 与 $g_i(x_*)$ 中仅有一个等于零时, 称严格互补松弛条件成立. LICQ 成立时, 与 x_* 对应的 Lagrange 乘子唯一.

按语 **25.13.** 当 *CQ* 不成立, 局部极小点不一定是 *KKT* 点. 考虑以下两个问题:

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & x_2 \\ \text{s.t.} \quad & x_2 \leqslant x_1^3, \ x_2 \geqslant 0, \end{aligned}$$

和

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & x_1 \\ \text{s.t.} \quad & x_2 \leqslant x_1^3, \ x_2 \geqslant 0. \end{aligned} \tag{P2}$$

易见$x_* = (0,0)$ 是二者的最优解，且$(0,0)$ 是第一个的 *KKT* 点, 但它不是第二个的 *KKT* 点.

定理对分析与求解约束优化问题很重要(比如可以利用它为优化方法设计停止准则). 下面用几个例子来熟悉和理解KKT条件.

例子 **25.14** (单个等式约束)**.** 考虑问题

$$\underset{x \in \mathbb{R}^2}{\text{minimize}} \ x_1 + x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 - 2 = 0, \tag{25.6}$$

其有两个变量和一个等式约束(见图25.4 (a)). 用(P)的表述, $f(x) = x_1 + x_2, \ell = 1, m = 0, \mathcal{A} = \{1\}$, 且 $h_1(x) = x_1^2 + x_2^2 - 2$. 易见该问题的可行域是中心在原点, 半径为 $\sqrt{2}$ 的圆周. 解 $x_* = (-1,-1)$. 从该圆周上的任一其它点, 易找到一条前进轨线, 它在使 f 减小的同时保持可行, 即保持在圆上. 例如, 从点 $x = (\sqrt{2}, 0)$ 围绕这个圆周沿顺时针方向移动既能保持可行又能使目标值减小.

从图25.4 (a)也看到, 解 x_* 处约束的法向量 ∇h_1^* 与目标函数的梯度向量 $\nabla f(x_*)$ 是平行的, 即存在标量 $\mu_1^* = \frac{1}{2}$ 使得

$$-\nabla f(x_*) = \mu_1^* \nabla h_1(x_*), \quad \nabla h_1(x_*) = 2\begin{pmatrix}-1\\-1\end{pmatrix}, \quad \nabla f(x_*) = \begin{pmatrix}1\\1\end{pmatrix}. \tag{25.7}$$

所以$x_* = (1,1)$满足KKT条件，对应Lagrange乘子$\mu_1^* = -\frac{1}{2}$.

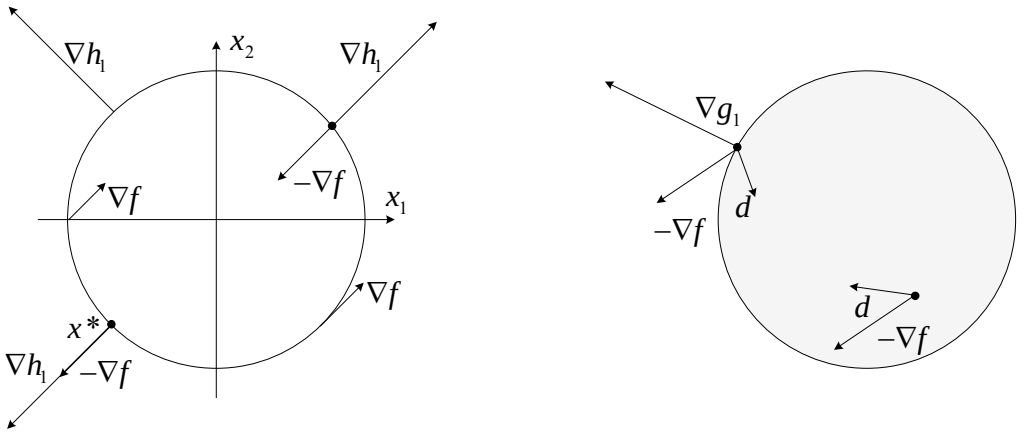


图 25.4: 可行点处约束函数和目标函数的梯度

例子 **25.15** (单个不等式约束)**.** 稍微修正例25.14, 即将其中的等式约束用不等式约束代替, 得

$$\underset{x\in\mathbb{R}^2}{\text{minimize}}\ x_1 + x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 - 2 \leqslant 0. \tag{25.8}$$

可行域由问题(25.6)的圆周和它的内部组成(见图25.4 (b)). 注意对圆周上的每个点, 约束法向量 ∇g_1 指向可行域的外部. 易见 $(-1,-1)$仍然是解, 且等式(25.7)对 $\lambda_1^* = \frac{1}{2}$ 仍然成立. 然而, 这个不等式约束问题与例25.14中的问题(25.6)是不同的，这里需要Lagrange乘子是非负的, 且h_1变成g_1. 此时，可验证$(1,1)$不再是KKT点.

例子 **25.16** (两个不等式约束)**.** 给问题(25.8)再加一个约束得到

$$\underset{x\in\mathbb{R}^2}{\text{minimize}}\ x_1 + x_2 \quad \text{subject to} \quad x_1^2 + x_2^2 \leqslant 2, \quad x_2 \geqslant 0. \tag{25.9}$$

可行域如图25.5中所示的半圆盘，易见解 $x_* = (-\sqrt{2}, 0)$, 该点处两个约束都是积极的. 在解 x_* 处, 有

$$\nabla f(x_*) = \begin{pmatrix}1\\1\end{pmatrix}, \qquad \nabla g_1(x_*) = \begin{pmatrix}-2\sqrt{2}\\0\end{pmatrix}, \qquad \nabla g_2(x_*) = \begin{pmatrix}0\\-1\end{pmatrix}.$$

因此, 当选取 $\lambda^* = (\frac{1}{2\sqrt{2}}, 1)$ 时, 易验证 $\nabla_x L(x_*, \lambda^*) = 0$. 需要注意的是这里 λ^* 的两个分量都是正的.

现在考虑问题(25.9)的一些非解的可行点. 对于点 $x = (\sqrt{2}, 0)$, 两个约束同样均是积极的. 在该点, 易见目标函数的负梯度 $-\nabla f$ 不再位于 ∇g_1和∇g_2张成的锥内，见图25.5 (b).

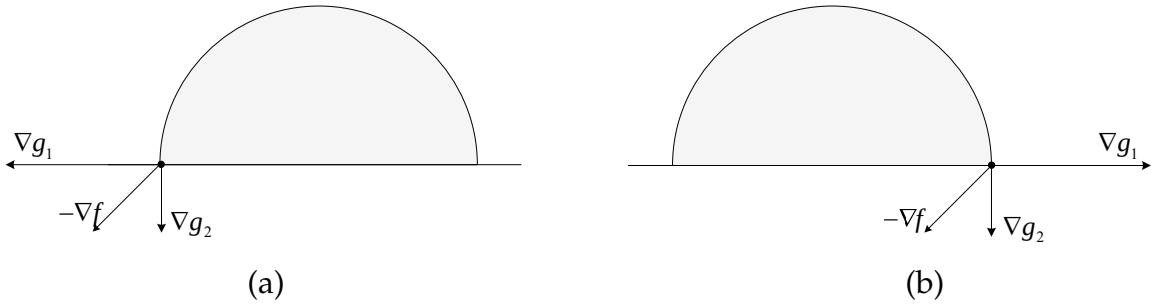


图 25.5: 问题(25.9)在不同可行点处的约束函数和目标函数的梯度

最后考虑点 $x = (1,0)$, 此时第二个约束是积极的. 在该点, 因为 $g_1(x) < 0$, 首先由互补松弛条件必有 $\lambda_1 = 0$. 因此, 在尝试满足 $\nabla_x L(x,\lambda) = 0$ 时, 要寻找 λ_2 使得 $\nabla f(x) + \lambda_2 \nabla g_2(x) = 0$. 由于不存在这样的 λ_2, 故该点不满足KKT条件.

至此只考虑了一阶(由一阶导数表示的)条件. 还需要叙述二阶条件, 它给出了目标函数与约束函数在局部极小点处曲率的信息. 这方面的内容放在 第26讲讨论. 当问题是凸规划时, 还能给出更强的结果, 见25.3节.

例子 **25.17** (灵敏度分析)**.** 考虑

$$\begin{aligned} \min_{x\in\mathbb{R}^2} \quad & -x_1 - x_2 \\ \text{s.t.} \quad & x_1^2 + x_2^2 - 1 \leqslant 0. \end{aligned}$$

解得$x_* = (\sqrt{2}/2, \sqrt{2}/2), \lambda^* = \sqrt{2}/2$. 对约束进行扰动, 得扰动问题为

$$\begin{aligned} \omega(\epsilon) = \min_{x\in\mathbb{R}^2} \quad & -x_1 - x_2 \\ \text{s.t.} \quad & x_1^2 + x_2^2 - 1 \leqslant \epsilon. \end{aligned}$$

易于验证$\omega(\epsilon) = -\sqrt{2(\epsilon+1)}$, $\omega'(0) = -\sqrt{2}/2 = -\lambda^*$. 灵敏度问题的实例请参见HW6.

以一般的二维问题为例. 设

$$\begin{aligned} x_* = \operatorname*{argmin}_{x\in\mathbb{R}^2} \quad & f(x) \\ \text{s.t.} \quad & g(x) \leqslant 0, \end{aligned}$$

其中 $g(x) : \mathbb{R}^2 \mapsto \mathbb{R}$. 设 x_* 是 KKT 点, $g(x_*) = 0$, 且与 x_* 对应的 Lagrange 乘子 $\lambda^* > 0$. 对应的扰动问题:

$$\begin{aligned} \omega(\epsilon) = \min_{x\in\mathbb{R}^2} \quad & f(x) \\ \text{s.t.} \quad & g(x) \leqslant \epsilon, \end{aligned}$$

它的

$$L(x,\lambda,\epsilon) = f(x) + \lambda(g(x) - \epsilon).$$

记扰动问题的解和乘子分别为 $x(\epsilon)$ 和 $\lambda(\epsilon)$, 有

$$\omega(\epsilon) := f(x(\epsilon)) = \mathcal{L}(x(\epsilon), \lambda(\epsilon), \epsilon).$$

则 $\omega'(0) = -\lambda^*$. 从而

$$\omega(\epsilon) \approx \omega(0) - \lambda^*\epsilon.$$

综上分析，可将Lagrange 乘子解释为最优值关于约束的灵敏度，即当约束右端项增加一个单位时, 最优值改变量的相反数!

和无约束优化中的驻点必要条件一样，KKT 条件只是必要的！比如

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^2} & x^3 \\ \text{s.t.} & -1 \leqslant x \leqslant 1. \end{array} \tag{P3}$$

易见 $x_* = \mathbf{0}$ 是 KKT 点, 但不是局部极小点. 下面的讨论表明，对于凸规划，KKT点就是问题的全局极小点.

25.3 凸规划

凸规划(优化)指在凸集 $\Omega \subseteq \mathbb{R}^n$ 上极小化凸函数 $f(x)$. 由第一次课就知道，凸规划问题的最有名性质是：局部极小点也全局极小点. 要利用这个性质，就必须会判断所面对的问题是否是凸优化，但是凸优化的定义不易于用来检验实践中出现的问题是否为凸的. 回忆第2讲凸函数的如下性质.

引理 **25.18.** 凸函数 f 的(下)水平集 $L_\gamma = \{x \in \text{dom } f : f(x) \leqslant \gamma\}$ 是凸集, 其中 $\gamma \in \mathbb{R}$.

这样，如下问题是凸优化:

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & g_i(x) \leqslant 0,\ i = 1, \cdots, m \\ & a_j^T x = b_j,\ j = 1, \cdots, \ell, \end{array} \tag{25.10}$$

其中X是凸集，$f, g_i\ \forall i$ 是凸函数, $a_j \in \mathbb{R}^n$, $b_j \in \mathbb{R}$, $j = 1, \cdots, \ell$.

从而，线性规划是凸规划; 当二次规划中目标函数的 Hessian 阵半正定时, 也是凸规划.

定理 **25.19.** 在(25.10)中，假设 $X = \mathbb{R}^n, f, g_i\ \forall i$ 是凸函数. 设 x_* 是(25.10) 的 *KKT* 点, 那么 x_* 是(25.10)的全局极小点.

例子 **25.20.** 考虑

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^2} & \left(x_1 - \frac{3}{2}\right)^2 + \left(x_2 - \frac{1}{2}\right)^4 \\ \text{s.t.} & x_1 + x_2 - 1 \leqslant 0 \\ & x_1 - x_2 - 1 \leqslant 0 \\ & -x_1 + x_2 - 1 \leqslant 0 \\ & -x_1 - x_2 - 1 \leqslant 0 \end{array}$$

则$x_* = (1, 0)$, $\mathcal{J}(x_*) = \{1, 2\}$，且

$$\nabla f(x_*) = \begin{bmatrix} -1 \\ -\frac{1}{2} \end{bmatrix},\ \nabla g_1(x_*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ \nabla g_2(x_*) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

得 Lagrange 乘子 $\lambda^* = (3/4, 1/4, 0, 0)$. 因为待求问题是凸规划, 所以 x_* 是全局解.

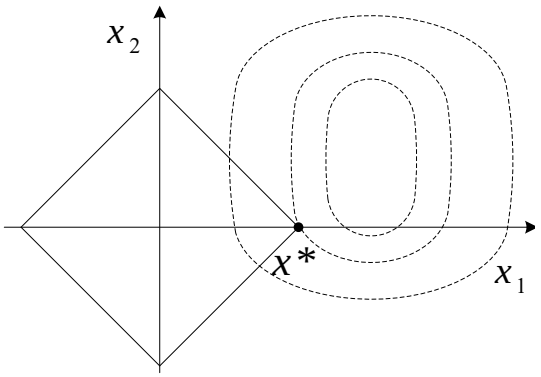


图 25.6: 凸规划的KKT点

例子 **25.21** (缺乏 CQ 时 KKT 条件失效)**.** 问题

$$\begin{aligned} \min_{x\in\mathbb{R}^2} \quad & x_1^2 + x_2^2 \\ \text{s.t.} \quad & (x_1-1)^2 + (x_2-1)^2 \leqslant 1 \\ & (x_1-1)^2 + (x_2+1)^2 \leqslant 1 \end{aligned}$$

是凸规划. 也可验证$(1,0)$是最优解，但不是KKT点. 该例表明，如果约束品性不成立，即使是凸规划，它的局部极小点也不一定是 KKT 点.

26 二阶最优性条件

本节考虑约束问题(P)的局部极小点的二阶最优性条件.

定理 **26.1** (二阶必要条件)**.** 设 $f, g_i, h_j \in C^2$, x_* 是(P)的局部极小点且x_* 处 *LICQ* 成立. 那么存在 $\lambda_i^* \geqslant 0, i \in \mathcal{J}(x_*)$, μ_j^* 满足(25.5)，并且

$$d^T \nabla_{xx}^2 L(x_*, \lambda^*, \mu^*) d \geqslant 0 \ \ \forall d \in T(x_*, \lambda^*),$$

其中

$$T(x_*, \lambda^*) = \{d \in \mathbb{R}^n : d^T \nabla h_j(x_*) = 0,\ j \leqslant \ell,\ d^T \nabla g_i(x_*) = 0,\ i \in \mathcal{J}(x_*)\}.$$

证明. 任取 $0 \neq d \in T(x_*, \lambda^*)$，则由$x_*$处LICQ成立，由命题 25.9的证明知，可由隐函数定理构造点列 $\{x_t\} \subseteq \Omega \setminus \{x_*\}$ 和正标量序列 $\{\delta_t\}$, $\delta_t \to 0$ 满足(25.2). 这里特别需要指出的是，因为这个$d \in T(x_*, \lambda^*)$，由证明中构造的方程组知$g_j(x_t) = 0$. 从而$f(x_t) = L(x_t, \lambda^*, \mu^*)$； 再由$x_*$是KKT点和$L(\cdot, \mu^*, \lambda^*)$在$x_*$的二阶 Taylor 展式，

$$f(x_t) - f(x_*) = \frac{1}{2}(x_t - x_*)^T \nabla_{xx}^2 L(x_*, \lambda^*, \mu^*)(x_t - x_*) + o(\|(x_t - x_*)\|^2).$$

最后，由x_*的局部最优性，x_t可行且$x_t \to x_*$(当$t \to \infty$)知对充分大的t，

$$\frac{1}{2}(x_t - x_*)^T \nabla_{xx}^2 L(x_*, \lambda^*, \mu^*)(x_t - x_*) + o(\|(x_t - x_*)\|^2) \geqslant 0.$$

给两边同时除以$\|(x_t - x_*)\|^2$, 令$t \to \infty$，由(25.2)和极限的保号性得待证结论. ■

事实 **26.2.** 设 $W \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{k\times n}$, 并且 A 的行线性无关. 设 $z_1, \cdots, z_{n-k}$ 是 null A 的一个基. 记 $Z = [z_1, \cdots, z_{n-k}] \in \mathbb{R}^{n\times(n-k)}$. 则

$$d^T W d \geqslant 0 \ \ \forall d \in \{d \in \mathbb{R}^n : Ad = 0\}$$

当且仅当 $Z^T W Z$ 半正定.

定理 **26.3** (二阶充分条件)**.** 设 $f, g_i, h_j \in C^2$, x_* 是(P)的*KKT*点，即存在 $\lambda_i^* \geqslant 0$, 满足(25.5)，并且

$$d^T \nabla^2_{xx} L(x_*, \lambda^*, \mu^*) d > 0 \ \ \forall 0 \neq d \in T^+(x_*, \lambda^*),$$

其中

$$T^+(x_*, \lambda^*) = \{d : d^T \nabla h_j(x_*) = 0,\ j \leqslant \ell,\ d^T \nabla g_i(x_*) = 0,\ \lambda_i^* > 0,\ i \in \mathcal{J}(x_*)\}.$$

那么 x_* 是(P)的严格局部极小点.

证明. 用反证法. 假设 x_* 不是严格局部极小点. 那么对任意正整数 t, 存在 $x_t \in \Omega \setminus \{x_*\}$, $\|x_t - x_*\| \leqslant \frac{1}{t}$ 满足

$$f(x_t) \leqslant f(x_*). \tag{26.1}$$

一方面, 令 $\delta_t = \|x_t - x_*\|$, 则 $\delta_t > 0, \delta_t \to 0$, 并且 $\left\{\frac{x_t - x_*}{\delta_t}\right\}$ 有收敛子列. 不妨设(25.2). 从而 $d \neq 0$且$d \in T_\Omega(x_*)$. 再由命题 25.7, $T_\Omega(x_*) \subseteq F(x_*)$. 因此 $d \in F(x_*)$. 进一步，类似于命题 25.2的证明，由(26.1)，和f的一阶 Taylor 展式得$d^T \nabla f(x_*) \leqslant 0$. 再结合KKT 条件中的对偶可行性和 $d \in F(x_*)$ 得 $d \in T^+(x_*, \lambda^*)$.

另一方面，由x_t可行得$f(x_t) \geqslant L(x_t, \lambda^*, \mu^*)$； 再由$x_*$是KKT点和$L(\cdot, \mu^*, \lambda^*)$在$x_*$的二阶 Taylor 展式，得

$$f(x_t) - f(x_*) \geqslant \frac{1}{2}(x_t - x_*)^T \nabla^2_{xx} L(x_*, \lambda^*, \mu^*)(x_t - x_*) + o(\|x_t - x_*\|^2)$$

再结合(26.1)，得

$$\frac{1}{2}(x_t - x_*)^T \nabla^2_{xx} L(x_*, \lambda^*, \mu^*)(x_t - x_*) + o(\|(x_t - x_*)\|^2) \leqslant 0.$$

给两边同时除以$\|(x_t - x_*)\|^2$, 令$t \to \infty$，由(25.2)和极限的保号性得

$$d^T \nabla^2_{xx} L(x_*, \lambda^*, \mu^*) d \leqslant 0.$$

这和 $0 \neq d \in T^+(x_*, \lambda^*)$ 与已知条件矛盾. 从而假设错误, 原命题得证. ■

下面讨论一个不同的二阶最优性条件. 它是借助于临界锥(critical cone)

$$C(x_*, \lambda^*) = \{d \in F(x_*) : d^T \nabla g_i(x_*) = 0,\ i \in \mathcal{J}(x_*) \text{ with } \lambda_i^* > 0\}$$

的概念给出. 可进一步等价表示为

$$C(x_*, \lambda^*) = \left\{ d \in \mathbb{R}^n : \begin{array}{c} d^T \nabla h_j(x_*) = 0,\ j \leqslant \ell \\ d^T \nabla g_i(x_*) = 0,\ i \in \mathcal{J}(x_*) \text{ with } \lambda_i^* > 0 \\ d^T \nabla g_i(x_*) \leqslant 0,\ i \in \mathcal{J}(x_*) \text{ with } \lambda_i^* = 0 \end{array} \right\}.$$

它与之前二阶条件中定义的两个集合的关系:

$$T(x_*, \lambda^*) \subseteq C(x_*, \lambda^*) \subseteq T^+(x_*, \lambda^*).$$

仅有等式约束, 或者严格互补松弛条件满足时，$T(x_*, \lambda^*) = C(x_*, \lambda^*) = T^+(x_*, \lambda^*)$.

按语 **26.4.** 设 x_* 是(P) 的 *KKT* 点, 且$d \in F(x_*)$. 那么

$$d^T \nabla f(x_*) = 0 \Leftrightarrow d \in C(x_*, \lambda^*).$$

当严格互补条件成立时, 局部二阶必要/充分最优性条件与下述问题的相同:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) = 0,\ i \in \mathcal{J}(x_*) \\ & h_j(x) = 0,\ j = 1, \cdots, \ell. \end{aligned}$$

当严格互补松弛条件不成立时，需要将积极约束进一步细分为弱积极约束**(**乘子等于**0)**与强积极约束**(**乘子大于**0)**讨论二阶最优性条件. 这些概念对应的几何直观如图 26.1所示.

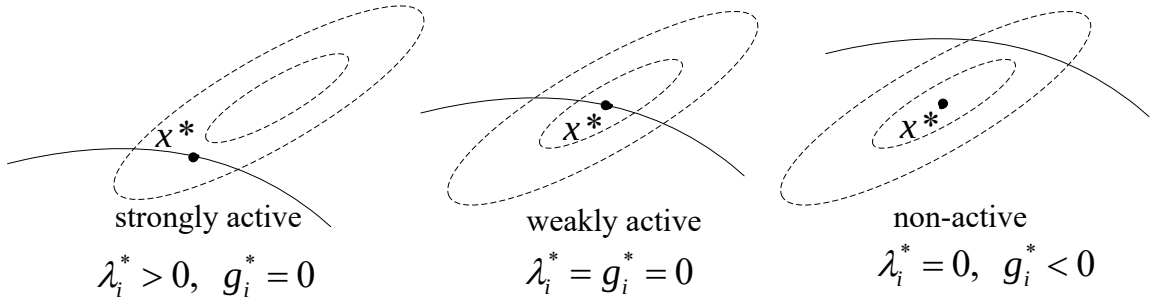


图 26.1: 积极约束、弱积极约束和强积极约束的几何直观.

定理 **26.5** (二阶必要条件)**.** 设 $f, g_i, h_j \in C^2$, x_* 是(P)的局部极小点且 x_* 处 *LICQ* 成立, 那么存在 λ_i^*, μ_j^* 使得 x_* 满足 *KKT* 条件, 并且

$$d^T \nabla_{xx}^2 L(x_*, \lambda^*, \mu^*) d \geqslant 0,\ \ \forall d \in C(x_*, \lambda^*).$$

定理 **26.6** (二阶充分条件)**.** 设 $f, g_i, h_j \in C^2$, x_* 处 *LICQ* 成立. 如果存在 λ_i^*, μ_j^* 使得 x_* 满足 *KKT* 条件, 并且

$$d^T \nabla_{xx}^2 L(x_*, \lambda^*, \mu^*) d > 0,\ \ \forall 0 \neq d \in C(x_*, \lambda^*),$$

那么 x_* 是(P)的严格局部极小点.

例子 **26.7.** 讨论参数 β 取何值时, $x_* = (0,0)$ 是

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & f(x) = \tfrac{1}{2}(x_1 - 1)^2 + \tfrac{1}{2}x_2^2 \\ \text{s.t.} \quad & h(x) = x_1 - \beta x_2^2 = 0 \end{aligned} \tag{26.2}$$

的局部极小点.

该问题的目标函数的等值线和$\beta = 1$与$1/4$的可行域如图 26.2所示. 易于验证 $\nabla f(x_*) = (-1, 0),\ \nabla h(x_*) = (1, 0)$, 所以 x_* 是 KKT 点, 且 $\mu^* = 1$. 进而, 计算可得

$$W^* = \nabla_{xx}^2 L(x_*, \mu^*) = \nabla^2 f(x_*) + \mu^* \nabla^2 h(x_*) = \begin{bmatrix} 1 & 0 \\ 0 & 1 - 2\beta \end{bmatrix},$$

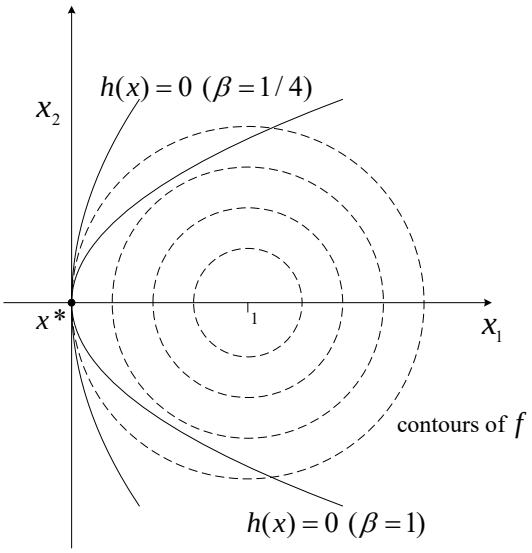


图 26.2: 问题(26.2)的几何直观.

$T(x_*,\lambda^*)=T^+(x_*,\lambda^*)=\{\alpha(0,1):\alpha\in\mathbb{R}\},\quad d\in T(x_*,\lambda^*)\Rightarrow d^TW^*d=\alpha^2(1-2\beta).$
从而 $\beta<\frac{1}{2}$ 时, x_* 是严格局部极小点; 当 $\beta>\frac{1}{2}$ 时, x_* 不是严格局部极小点; 当 $\beta=\frac{1}{2}$ 时, 用二阶条件无法判断.

例子 **26.8.** 考虑图26.3所示问题

$$\begin{array}{ll}\min & x_1\\ \text{s.t.} & 1-(x_1-1)^2-x_2^2\geqslant 0,\\ & x_2\geqslant 0.\end{array}\tag{26.3}$$

不难得到其解 $x_*=(0,0)$, 积极集 $\mathcal{J}(x_*)=\{1,2\}$, 以及唯一的 Lagrange 乘子 $\lambda^*=(0.5,0)$. 由于x_* 处积极约束在的梯度分别是 $(-2,0)$ 和 $(0,-1)$, LICQ成立. 所以对应的乘子唯一，且线性化可行方向集

$$F(x_*)=\{d\in\mathbb{R}^2:d\geqslant 0\},$$

对应的

$$T(x_*,\lambda^*)=\{(0,0)\},\ C(x_*,\lambda^*)=\{(0,d_2)^T:d_2\geqslant 0\},\ T^+(x_*,\lambda^*)=\{(0,d_2):d_2\in\mathbb{R}\}.$$

例子 **26.9** (最小特征值的变分刻画)**.** 已知 A 是 $n\times n$ 对称矩阵. 考虑

$$\begin{array}{ll}\min\limits_{x\in\mathbb{R}^2} & f(x)\equiv x^TAx\\ \text{s.t.} & h(x)\equiv 1-x^Tx=0.\end{array}$$

证明该问题的最优值是 A 的最小特征值, 最优解是与该特征值对应的单位特征向量.

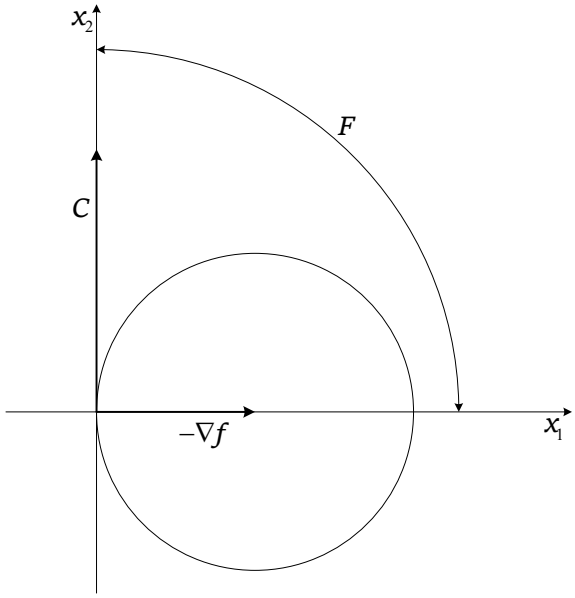


图 26.3: 问题(26.3)的 $F(x_*)$ 和 $C(x_*, \lambda^*)$.

27 内点法入门

在上一讲，讨论了基本牛顿法. 尽管它享有快的局部收敛保证，但是不能保证全局收敛. 本讲引入内点法，这可看作能保证全局收敛的扩展牛顿法. 首先介绍常规障碍法**(barrier methods)** 的主要思想.

27.1 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：

$$\begin{array}{ll} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & x \in \Omega, \\ & g_j(x) \leqslant 0\,, j = 1, \cdots, m\,, \end{array} \tag{27.1}$$

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$X \subseteq \mathbb{R}^n$是闭凸集. 本讲的剩余部分，假设g_j是连续凸函数，并且$X = \mathbb{R}^n$. 用x_* 表示问题 (27.1)的最优解.

定义 **27.1** (约束域的内部)**.** 约束域(相对于X)的内部定义为

$$\Omega = \{x \in X \in\colon g_j(x) < 0,\ j = 1, \cdots, m\}.$$

假设Ω非空. 所谓Ω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:

$$\text{对数:} \qquad B(x) = -\sum_{j=1}^{m} \ln(-g_j(x)), \tag{27.2}$$

$$\text{倒数:} \qquad B(x) = -\sum_{j=1}^{m} \frac{1}{g_j(x)}. \tag{27.3}$$

如果所有$g_j(x)$是凸的时，这两个均是凸函数.

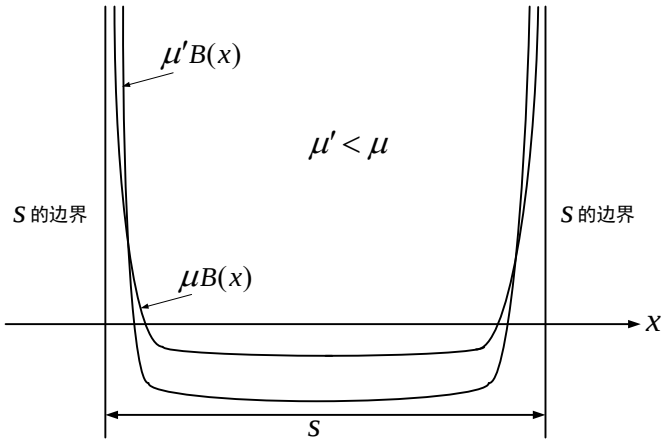


图 27.1: 障碍项的形状(图形中的S应为Ω).

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(当成无约束优化，在Ω的内部求极小. 求解时，从Ω的内部出发，做线搜索时将迭代保持在Ω内，即从Ω的内部逼近问题的最优解，所以称之为内点法)，得到如下问题:

$$\min_{x\in\Omega} f_\mu(x)$$

障碍项$\mu B(x)$的形状见图 27.1. 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 **27.2** (对数障碍函数)**.** 问题

$$\begin{array}{ll} \text{minimize}_{x\in\mathbb{R}} & x \\ \text{subject to} & g(x) := 1 - x \leqslant 0 \end{array} \tag{27.4}$$

的对数障碍函数 $f_\mu(x) = x - \mu\ln(x-1)$, 图 27.2对一组 μ 的值给出了函数f_μ的图形，由此可以看到随着 $\mu_t \to 0, x_t \to x_*$.

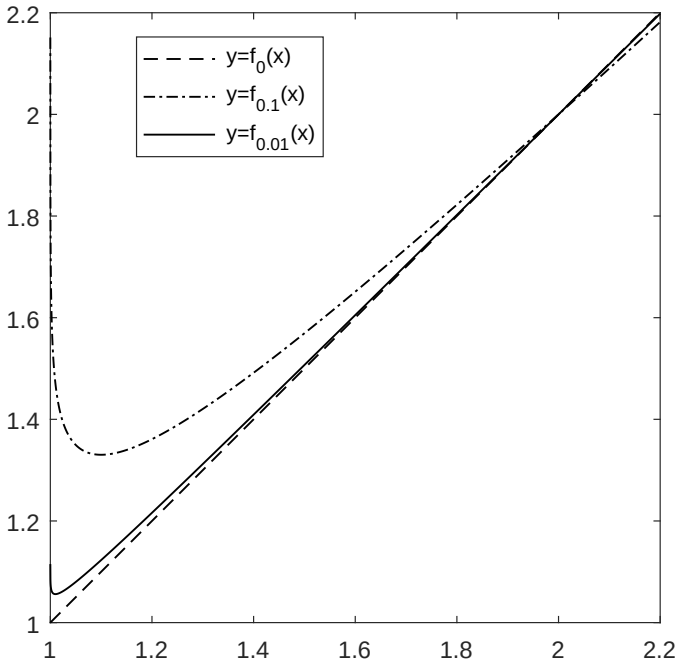


图 27.2: 障碍函数中递增的病态性

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_*. 它的收敛性的正式陈述如下.

命题 **27.3.** 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 (27.1) 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.

27.2 线性规划

在勾勒出常规障碍法的基本思想后，现在将对数障碍法应用线性规划(Linear programming, LP)问题：

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^\top x \\ \text{s.t.} \quad & Ax \geqslant b \end{aligned} \tag{LP}$$

其中$A \in \mathbb{R}^{m \times n}, m \geqslant n$，并且$\operatorname{rank}(A) = n$. 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

$$f_\mu(x) = c^\top x - \mu \sum_{j=1}^{m} \ln\left(a_j^\top x - b_j\right). \tag{27.5}$$

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$x_\mu^* = \operatorname{argmin}_{x\in\mathbb{R}^n} f_\mu(x)$.

事实 **27.4.** 对任何$\mu > 0$，f_μ的最小点x_μ^*存在并且唯一.

证明. 易于检查$f_\mu(x)$是凸的(是两个凸函数之和，或者检查(27.7)中的Hessian阵，说明它是正定的.). 易于说明f_μ有稳定点，它的Hessian阵是正定的，从而极小点唯一.[9].

为了证明f_μ的凸性， ■

27.2.1 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 27.3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?

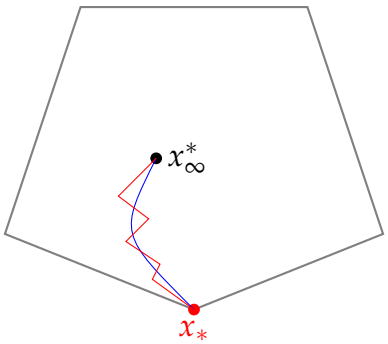


图 27.3: 蓝色曲线表示一个二维LP实例的中心路径

27.2.2 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

$$\nabla f_\mu(x) = c - \mu \sum_{j=1}^{m} \frac{a_j}{a_j^\top x - b_j} \triangleq c - \mu A^\top S^{-1}\mathbf{1} \tag{27.6}$$

$$\nabla^2 f_\mu(x) = \mu A^\top S^{-2} A = \mu \sum_{j=1}^{m} \frac{a_j a_j^\top}{s_j^2} \tag{27.7}$$

[9]可将这两个问题设计成作业：凸函数有稳定点，必有极小点. Hessian阵正定的是严格凸函数，从而极小点唯一.

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∇f_μ在x的一阶近似为零，求所得到的线性方程组得到的解. 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：

$$\begin{aligned} & f_\mu(x) - \min_{\bar{x}} \left(f_\mu(x) + \nabla f_\mu(x)^\top(\bar{x} - x) + \tfrac{1}{2}(\bar{x} - x)^\top \nabla^2 f_\mu(x)(\bar{x} - x) \right) \\ = \;& f_\mu(x) - \left(f_\mu(x) - \frac{1}{2}\nabla f_\mu(x)^\top [\nabla^2 f_\mu(x)]^{-1}\nabla f_\mu(x) \right) \\ = \;& \frac{1}{2}\nabla f_\mu(x)^\top [\nabla^2 f_\mu(x)]^{-1}\nabla f_\mu(x) =: \frac{1}{2}q^2(x, \mu). \end{aligned} \tag{27.8}$$

将使用牛顿减量来找出保证算法收敛的条件.

27.2.3 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 **27.5.** 假设严格可行点x(即$Ax > b$)满足$q(x, \mu) < 1$. 那么

$$c^\top x - c^\top x_* \leqslant 2\mu n.$$

特别地，如果维持x_t是满足$Ax_t > b$，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 **27.6.** 已知障碍因子$\mu > 0$. 假设严格可行点x(即$Ax > b$)满足$q(x, \mu) < 1$. 那么基本牛顿步$\bar{x}$ 满足

$$q(\bar{x}, \mu) \leqslant q(x, \mu)^2.$$

该结论表明，当障碍因子μ固定时，从满足$q(x, \mu) < 1$的严格可行点x出发，基本牛顿迭代法是良定义的，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 **27.7.** 假设正数μ和x满足$q(x,\mu) \leqslant \frac{1}{2}$ 和$Ax > b$. 置

$$\bar{\mu} = \left(1 - \frac{1}{6\sqrt{n}}\right)\mu,$$

那么有

$$q(\bar{x},\bar{\mu}) \leqslant \frac{1}{2}.$$

这些命题表明如下更新规则，

$$\begin{aligned} x_{t+1} &= x_t - \nabla^2 f_{\mu_t}(x)^{-1}\nabla f_{\mu_t}(x_t) \\ \mu_{t+1} &= \left(1 - \frac{1}{6\sqrt{n}}\right)\mu_t \end{aligned}$$

定理 **27.8.** 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$O(\sqrt{n}\log(n/\epsilon))$次迭代内得到的误差为$\epsilon$，即在$t = O(\sqrt{n}\ln(n/\epsilon))$ 次迭代之后，有

$$c^\top x_t \leqslant c^\top x_* + \epsilon.$$

证明. 因为牛顿步保持x_t在Ω的内部，使用上面的三个命题，有

$$\begin{aligned} c^\top x_t &\leqslant c^\top x_* + 2\mu_t n \\ &= c^\top x_* + 2\left(1 - \frac{1}{6\sqrt{n}}\right)^t n\mu_0 \\ &\leqslant c^\top x_* + 2\exp\left(-\frac{t}{6\sqrt{n}}\right) n\mu_0 \end{aligned} \tag{27.9}$$

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 $O(\sqrt{n}\ln(n/\epsilon))$ 次迭代后收敛误差达到ϵ. ■

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.

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.

28 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), 类似地，它在中心路径附近寻找近似点. 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， 籍此寻找一个更大胆的步长.

28.1 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

$$\min\ c^\top x \quad \text{s.t.} \quad Ax = b, x \geqslant 0, \tag{LP$_S$}$$

其中$\geqslant$是逐分量的. 观察可将 ($\mathrm{LP_S}$)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， ($\mathrm{LP_S}$)的对偶问题是

$$\max\ b^\top z \quad \text{s.t.} \quad A^\top z\leqslant c. \tag{$\mathrm{LP_D}$}$$

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)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

$$\langle x,s\rangle=\langle x,c\rangle-\langle b,z\rangle\geqslant\langle x,c\rangle-\langle x_*,c\rangle=\langle x-x_*,c\rangle.$$

28.2 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

$$\bar{\mu}=\bar{\mu}(x,s):=\frac{\langle x,s\rangle}{n}=\frac{\langle x,c\rangle-\langle b,z\rangle}{n}\geqslant\frac{\langle x-x_*,c\rangle}{n}. \tag{28.1}$$

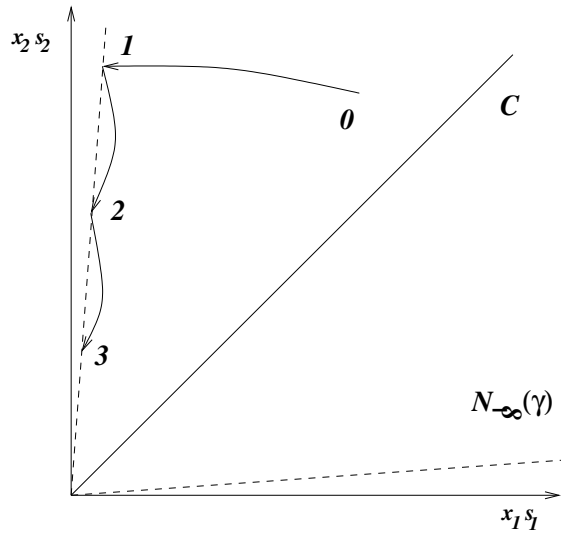


图 28.1: 要求迭代停留在中心路径的某个邻域内，以便逐分量乘积$x_1 s_1, \cdots, x_n s_n$相差不要太大

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.

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. 为了确保迭代离中心路径很近，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28.1给出了$\mathcal{N}_{-\infty}(\gamma)$ 和迭代过程的几何直观.

Algorithm 8 长步路径跟踪算法

Require: Parameters $\gamma \in (0,1)$, $0 < \sigma_{\min} \leqslant \sigma_{\max} < 1$.

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 $\sigma_t \in [\sigma_{\min}, \sigma_{\max}]$ and let $\mu_t = \sigma_t \bar{\mu}_t$.
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**

28.3 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 28.2. 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.

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$\mu = \sigma\bar{\mu} > 0$, 其中$\bar{\mu} = x^T s/n$. 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

$$F_\mu(x, z, s) = \begin{bmatrix} Ax - b \\ A^\top z + s - c \\ x \circ s - \mu\mathbf{1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ x \circ s - \mu\mathbf{1} \end{bmatrix}.$$

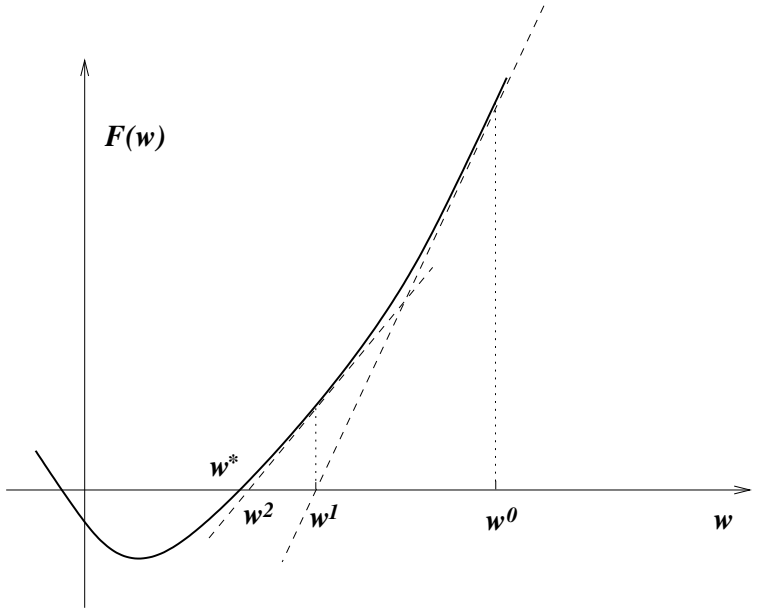


图 28.2: 回忆牛顿法迭代地求得实值函数根(或者零点)的更好近似.

因此，牛顿增量满足

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^\top & I \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \\ \Delta s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -x \circ s + \mu \mathbf{1} \end{bmatrix},$$

即

$$A\Delta x = 0 \tag{28.2}$$

$$A^\top \Delta z + \Delta s = 0 \tag{28.3}$$

$$s \circ \Delta x + x \circ \Delta s = -x \circ s + \mu \mathbf{1}. \tag{28.4}$$

(28.2)和(28.3)蕴含着$\forall \alpha$,

$$(x^+, z^+, s^+) := (x + \alpha\Delta x, z + \alpha\Delta z, s + \alpha\Delta s)$$

满足

$$Ax^+ - b = 0 \quad \text{和} \quad A^\top z^+ + s^+ - c = 0.$$

因此，对充分小的$\alpha > 0$，$(x^+, z^+, s^+) \in \mathcal{F}^{\mathrm{o}}$.

现在分析新迭代点的对偶性度量(算法8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(28.2)和(28.3)可知上式中的最后一项消失，因为

$$0 = \Delta x^\top (A^\top \Delta z + \Delta s) = (A\Delta x)^\top \Delta z + \langle \Delta x, \Delta s \rangle = \langle \Delta x, \Delta s \rangle.$$

此外，将(28.4)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算法8的第3行和$\bar{\mu}$的定义 (28.1)，即

$$n\mu = n\sigma\bar{\mu} = \sigma\langle x, s\rangle.$$

因此，

$$n\bar{\mu}(x + \alpha\Delta x, s + \alpha\Delta x) = n\bar{\mu}(x, s)[1 - (1 - \sigma)\alpha]$$

从而，如果能够证明存在某个与维数相关的常数$C(n) > 0$使得

$$(1 - \sigma)\alpha \geqslant C(n),$$

那么

$$\bar{\mu}(x_{t+1}, s_{t+1}) \leqslant (1 - C(n))^t \bar{\mu}(x_0, s_0)$$

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$\alpha = \Omega(1/n)$.

29 非凸目标函数与凸松弛

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:

$$\min_{\|x\|_0 \leqslant s} \|Ax - y\|_2^2, \tag{29.1}$$

其中$y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times d}$ 是已知的, $x \in \mathbb{R}^d, s \in \mathbb{Z}_{++}$, $\|x\|_0$表示x的非零元素个数. 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，其存在有效凸松弛.

压缩感知和稀疏线性回归问题均是带稀疏约束的最小二乘问题 (29.1)的特例，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，使得能够从更少的测量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$Ax = y$中不存在噪声时，问题 (29.1)可简化成ℓ_0极小化问题:

$$\begin{array}{ll} \text{minimize} & \|x\|_0 \\ \text{subject to} & Ax = y \end{array} \tag{29.2}$$

29.1 难度

将证明精确3-覆盖是ℓ_0-极小化问题 (29.2) 的特例，而精确3-覆盖是NP- 完全的. 从而这个简化版的非凸问题也是NP-难的. 这里的证明源自[FR13].

定义 **29.1. 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 $\cup_{j\in z} T_j = [m]$，且 $T_i \cap T_j = \emptyset, i, j \in z, i \neq j$? 称满足该条件的集合$z$ 是$[m]$ 的精确覆盖**(exact cover)**.

定义 **29.2.** 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 **29.3.** 针对一般(A, y)的ℓ_0-极小化问题 (29.2) 是*NP*-难的.

证明. 设y是分量全为 1 的向量. 已知$[m]$的3-元素子集$\{T_1, \cdots, T_d\}$. 定义$m \times d$的矩阵A为

$$A_{ij} = \begin{cases} 1, & \text{如果} i \in T_j \\ 0, & \text{否则,} \end{cases}$$

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

$$\|x\|_0 \geqslant \frac{\|y\|_0}{3} = \frac{m}{3}.$$

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：

(a) 如果$\|\hat{x}\|_0 = \frac{m}{3}$, 那么$z = \operatorname{supp}(\hat{x})$ 是精确3-覆盖.

(b) 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$\|\hat{x}\|_0 = \frac{m}{3}$，这与$\hat{x}$ 的最优性矛盾.

这样，由于通过ℓ_0-极小化问题 (29.2)能求解精确 3-覆盖，所以ℓ_0- 极小化问题 (29.2) 必是NP- 难的. ■

29.2 凸松弛

尽管ℓ_0-极小化问题 (29.2)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- 极小化问题：

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = y. \end{array} \tag{29.3}$$

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 **29.4.** 支集为$S \subset [d]$的近似稀疏向量

$$C(S) = \{\Delta \in \mathbb{R}^d : \|\Delta_{\bar{S}}\|_1 \leqslant \|\Delta_S\|_1\},$$

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 **29.5.** 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 **29.6.** 已知 $A \in \mathbb{R}^{m\times d}$和$y \in \mathbb{R}^m$，设$x^*$是 ℓ_0- 极小化问题 (29.2)的解，且矩阵A关于x^*的支集S满足受限零空间性质. 设 (29.3)的解是$\hat{x}$. 那么$\hat{x} = x^*$.

证明. 注意到，由定义x^*和$\hat{x}$都满足约束条件$Ax = y$. 记$\Delta = \hat{x} - x^*$是差向量，

$$A\Delta = A\hat{x} - Ax^* = 0,$$

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$\hat{x}$是ℓ_1优化的最优解，从而

$$\|\hat{x}\|_1 \leqslant \|x^*\|_1.$$

那么

$$\begin{aligned}
\|x_S^*\|_1 = \|x^*\|_1 &\geqslant \|\hat{x}\|_1 \\
&= \|x^* + \Delta\|_1 \\
&= \|x_S^* + \Delta_S\|_1 + \|x_{\overline{S}}^* + \Delta_{\overline{S}}\|_1 && \text{通过拆分}\ell_1\text{ 范数,} \\
&= \|x_S^* + \Delta_S\|_1 + \|\Delta_{\overline{S}}\|_1 && \|x^*\|_1\text{支集的假设,} \\
&\geqslant \|x_S^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{\overline{S}}\|_1.
\end{aligned}$$

因此$\|\Delta_S\|_1 \geqslant \|\Delta_{\overline{S}}\|_1$, 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 **29.7.** 已知$0 < \delta < 1$. 称矩阵A享有(s, δ)-RIP，如果对所有s-稀疏向量x ($\|x\|_0 \leqslant s$),

$$(1 - \delta)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1 + \delta)\|x\|_2^2.$$

上述要求希望测量具有稳定的能量性质，不等式表明它要保持s个重要分量的长度. 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 **29.8.** 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

$$(A_S)_{ij} = \begin{cases} A_{ij} & \text{如果 } j \in S \\ 0 & \text{否则.} \end{cases} \tag{29.4}$$

现在证明RIP蕴含着受限零空间性质.

定理 **29.9.** 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.

证明. 设$x \in \text{null}A$是任意的非零向量. 那么必须针对任意满足$|S| \leqslant s$ 的指标集合S证明

$$x \notin C(S).$$

特别地，记向量x的前s大元素的指标集为S_0. 证明

$$\|x_{S_0}\|_1 < \|x_{\overline{S}_0}\|_1 \tag{29.5}$$

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$\bar{S}_0 = \{1,\cdots,d\} \setminus S_0$进行剖分：

$$\overline{S}_0 = \bigcup_{j=1}^{\lceil \frac{d}{s} \rceil - 1} S_j$$

其中

- S_1 是$\bar{S}_0$中与x的前s大元素对应的指标集
- S_2 是$\bar{S}_0 \setminus S_1$中与x的前s大元素对应的指标集
- S_3是$\bar{S}_0 \setminus (S_1 \cup S_2)$ 中与x的前s 大元素对应的指标集
- 以此类推$\cdots$

因此$x = x_{S_0} + \sum\limits_j x_{S_j}$. 已经将$x$分解成大小为$s$的块. 也称这是剥壳(shelling). 由RIP,

$$\|x_{S_0}\|_2^2 \leqslant \tfrac{1}{1-\delta}\|Ax_{S_0}\|_2^2.$$

由假设$x \in \text{null}A$，

$$A(x_{S_0} + \sum_{j\geqslant 1} x_{S_j}) = 0 \implies Ax_{S_0} = -\sum_{j\geqslant 1} Ax_{S_j}.$$

因此

$$\begin{aligned}
\|x_{S_0}\|_2^2 &\leqslant \tfrac{1}{1-\delta}\|Ax_{S_0}\|_2^2 \\
&= \tfrac{1}{1-\delta}\langle Ax_{S_0}, Ax_{S_0}\rangle \\
&= \tfrac{1}{1-\delta}\sum_{j\geqslant 1}\langle Ax_{S_0}, -Ax_{S_j}\rangle \\
&= \tfrac{1}{1-\delta}\sum_{j\geqslant 1}[\langle Ax_{S_0}, -Ax_{S_j}\rangle + \langle x_{S_0}, x_{S_j}\rangle] && (\text{由于}\langle x_{S_0}, x_{S_j}\rangle = 0) \\
&= \tfrac{1}{1-\delta}\sum_{j\geqslant 1}[\langle A_{S_0}x_{S_0}, -A_{S_j}x_{S_j}\rangle + \langle I_{S_0}x_{S_0}, I_{S_j}x_{S_j}\rangle] && (\text{由于}Ax_{S_j} = A_{S_j}x_{S_j}) \\
&= \tfrac{1}{1-\delta}\sum_{j\geqslant 1}\langle x_{S_0}, (I_{S_0}I_{S_j} - A_{S_0}^\top A_{S_j})x_{S_j}\rangle \\
&\leqslant \tfrac{1}{1-\delta}\sum_{j\geqslant 1}\|x_{S_0}\|_2\|I_{S_0}I_{S_j} - A_{S_0}^\top A_{S_j}\|_2\|x_{S_j}\|_2 \\
&\leqslant \tfrac{1}{1-\delta}\sum_{j\geqslant 1}\|x_{S_0}\|_2\|I_{S_0\cup S_j} - A_{S_0\cup S_j}^\top A_{S_0\cup S_j}\|_2\|x_{S_j}\|_2 \\
&\leqslant \tfrac{\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_0}\|_2\|x_{S_j}\|_2 \quad . && (\text{引理 29.8})
\end{aligned}$$

从而

$$\|x_{S_0}\|_2 \leqslant \tfrac{\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_j}\|_2. \tag{29.6}$$

由构造方式知，针对每个$j \geqslant 1$,

$$\|x_{S_j}\|_\infty \leqslant \tfrac{1}{s}\|x_{S_{j-1}}\|_1$$

并且因此

$$\|x_{S_j}\|_2 \leqslant \tfrac{1}{\sqrt{s}}\|x_{S_{j-1}}\|_1. \tag{29.7}$$

代入 (29.6), 得到

$$\begin{aligned}
\|x_{S_0}\|_1 &\leqslant \sqrt{s}\|x_{S_0}\|_2 \\
&\leqslant \tfrac{\sqrt{s}\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_j}\|_2 && (\text{由 (29.6)式}) \\
&\leqslant \tfrac{\delta}{1-\delta}\sum_{j\geqslant 1}\|x_{S_{j-1}}\|_1 && (\text{由(29.7)式}) \\
&= \tfrac{\delta}{1-\delta}(\|x_{s_0}\|_1 + \sum_{j>1}\|x_{S_{j-1}}\|_1)
\end{aligned}$$

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(29.5)成立. ■

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 **29.10.** 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$\frac{1}{\sqrt{m}} A$ 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 **29.11.** 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**:

$$\min \|Ax - y\|_2^2 + \lambda \|x\|_1 \tag{29.8}$$

  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).

30 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，典型实例是 (29.2)，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，典型实例是 (29.1).

求解非凸优化问题 (29.2)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 (29.3). 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 (29.1)， 有

非凸约束→凸松弛→ 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，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.

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.

Algorithm 9 Iterative Hard Thresholding (ITH) Algorithm

Require: Parameters y, A, t, s.
1: Initialize $x^1 = 0$.
2: **for** $i = 1$ to t **do**
3: $\quad \tilde{x}^{i+1} \leftarrow x^i - A^\top(Ax^i - y)$.
4: $\quad x^{i+1} \leftarrow \Pi_s(\tilde{x}^{i+1})$.
5: **end for**
Ensure: $\widehat{x} \leftarrow x^{t+1}$.

定义 **30.1.** 称矩阵A满足(s,δ)**-RIP**，如果对所有的s-稀疏向量，

$$(1-\delta)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1+\delta)\|x\|_2^2.$$

引理 29.8表明：对于大小为s的支集S，

$$\|I_s - A_S^\top A_S\|_2 \leqslant \delta,$$

其中A_S是A在S上的限制，定义见(29.4).

定理 **30.2.** 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，

$$\|x^{i+1} - x_*\|_2 \leqslant \tfrac{1}{2}\|x^i - x_*\|_2 + 2\max_{|S|\leqslant 3s}\|A_S^\top e\|_2.$$

证明. 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

$$\begin{aligned}
\|x^{i+1} - x_*\|_2 &\leqslant \|x^{i+1} - \tilde{x}_{S'}^{i+1}\|_2 + \|\tilde{x}_{S'}^{i+1} - x_*\|_2 && \text{(三角不等式)}\\
&\leqslant 2\|\tilde{x}_{S'}^{i+1} - x_*\|_2 \\
&= 2\|x_{S'}^i - A_{S'}^\top(Ax^i - y) - x_*\|_2 && (\tilde{x}_{S'}^{i+1}\text{的定义})\\
&= 2\|x^i - A_{S'}^\top(A_{S'}x^i - A_{S'}x_* - e) - x_*\|_2 && \text{(代入}y\text{，并由}S'\text{的定义)}\\
&= 2\|x^i - x_* - A_{S'}^\top A_{S'}(x^i - x_*) + A_{S'}^\top e\|_2 \\
&\leqslant 2\|[I_{S'} - A_{S'}^\top A_{S'}](x^i - x_*)\|_2 + 2\|A_{S'}^\top e\|_2 \\
&\leqslant 2\delta\|x^i - x_*\|_2 + 2\max_{|S|\leqslant 3s}\|A_S^\top e\|_2. && \text{(RIP)}
\end{aligned}$$

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$\tilde{x}_{S'}^{i+1}$的最好s- 稀疏逼近，因此有：

$$\|x^{i+1} - \tilde{x}_{S'}^{i+1}\|_2 \leqslant \|x_* - \tilde{x}_{S'}^{i+1}\|_2.$$

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 **30.3.** 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. 因此在$t = \ln \frac{\|x_*\|_2}{\epsilon}$次迭代后

$$\|\widehat{x} - x_*\|_2 \leqslant \epsilon + \sqrt{5}\|e\|_2.$$

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.

命题 **30.4.** 假设f是$\beta-$光滑的. 那么对于所有$x, \Delta \in \mathbb{R}^d$

$$f(x+\Delta) \leqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{\beta}{2}\|\Delta\|_2^2.$$

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命题，得

$$\frac{1}{2}\|A(x+\Delta) - y\|_2^2 \leqslant \frac{1}{2}\|Ax - y\|_2^2 + \Delta^\top A^\top (Ax - y) + \frac{\beta}{2}\|\Delta\|_2^2.$$

将上式展开，整理后得

$$\frac{1}{2}\Delta^\top A^\top A\Delta \leqslant \frac{\beta}{2}\|\Delta\|_2^2.$$

因此二次函数的$\beta-$光滑性等价于 $\|A\Delta\|_2^2 \leqslant \beta\|\Delta\|_2^2$. 对于强凸性有类似结论.

命题 **30.5.** 假设f是$\alpha-$强凸的. 那么对所有$x, \Delta \in \mathbb{R}^d$

$$f(x+\Delta) \geqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{\alpha}{2}\|\Delta\|_2^2.$$

成立.

易于得到二次函数的α强凸性等价于$\|A\Delta\|_2^2 \geqslant \alpha\|\Delta\|_2^2$.

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RIP将光滑性和强凸性中的“对于所有Δ” (针对)替换成“对于所有s- 稀疏的Δ”.

定义 **30.6.** 称f是受限$\alpha-$强凸**(Restricted Strongly Convex, RSC)** 的，如果对于所有s- 稀疏的Δ，

$$f(x+\Delta) \geqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{\alpha}{2}\|\Delta\|_2^2.$$

类似地可以定义受限$\beta-$光滑**(Restricted $\beta-$ smooth, RM)**. 可以说

$$\text{RIP} = \text{RSC} + \text{RS}.$$

因此，如果矩阵A满足(s, δ)-RIP，取$\beta = 1 + \delta, \alpha = 1 - \delta$，上面的讨论等同于

$$\beta/\alpha = \frac{1+\delta}{1-\delta} \approx 1,$$

即函数f具有非常好的条件数β/α (因此常数步长$\approx 1/\beta$). 有大量的工作在削弱该假设，并进一步限制集合.

凸松驰牵扯最优时，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. 能将凸松驰与非凸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.

References

[AZO17] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proc. 8th ITCS*, 2017.

[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.

[BT09] Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems, 2009.

[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.

[Su2] A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.

[Tse08] Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.

[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.