标题 **title**

作者 *author*

2023 年 7 月 27 日

前言

目录

第一部分

科学的逻辑

第二部分

信息与数据

第三部分

决策与优化

第八章　不动点理论

考虑优化算法 A，它在函数 f 上的的收敛性如何？算法运行所产生的点列记为 $\{x_n\}$，它满足 $x_{n+1} = A(x_n)$. 如果关注序列 x_n 本身，那么这是一个数学分析的思路，通过寻找不同量之间的联系，来分析收敛性. 如果从算法 A 本身来看，这是一个算子法与泛函分析的思路，研究算法本身的性质，收敛性往往归结为吸收点的存在性. 后者是更加抽象且现代的思路。在本章中，我们将看到，从算子的角度来理解收敛性，最终问题就归结到了不动点理论。

不动点的定义是非常直接的，考虑一个集合 X 以及它到自身的映射 $f : X \to X$，元素 $a \in X$ 称为映射 $f : X \to X$ 的**不动点**，如果 $f(a) = a$. 本章将介绍三种不动点存在性定理。

§8.1　Banach 不动点定理

为了陈述不动点定理，我们需要引入一些数学概念。

定义 8.1 (度量与度量空间) 集合 X 上的**度量**（或**距离**）d 是映射

$$d : X \times X \to \mathbb{R}$$

满足条件

- 非负性：$d(x_1, x_2) \geq 0$，并且 $d(x_1, x_2) = 0 \iff x_1 = x_2$.
- 对称性：$d(x_1, x_2) = d(x_2, x_1)$.
- 三角不等式：$d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$.

其中 x_1, x_2, x_3 是 X 的任意元素.

此时，(X, d) 或 X 被称为**度量空间**.

下面给出一些度量的例子。

例 8.1 考虑实数集 $\mathbb{R}$，要成为度量空间，可以装备以下度量：

- 平凡的离散度量：$\forall x_1 \neq x_2\, d(x_1, x_2) \equiv 1, d(x, x) = 0.$
- $d(x_1, x_2) = |x_1 - x_2|.$

考虑向量空间 $\mathbb{R}^n$，要成为度量空间，可以装备以下度量：

- *Minkowski* 度量（L^p 度量）：$d(x_1, x_2) = (\sum_{i=1}^n |x_1^i - x_2^i|^p)^{1/p}$ $(p \geq 1).$
- *Manhattan* 度量（L^1 度量）：$d(x_1, x_2) = \sum_{i=1}^n |x_1^i - x_2^i|.$
- *Euclid* 度量（L^2 度量）：$d(x_1, x_2) = \sqrt{\sum_{i=1}^n |x_1^i - x_2^i|^2}.$
- *Chebyshev* 度量（L^∞ 度量）：$d(x_1, x_2) = \max_i |x_1^i - x_2^i| = \lim_{p\to\infty}(\sum_{i=1}^n |x_1^i - x_2^i|^p)^{1/p}.$

我们的目标是找到一类和实数集非常像的度量空间。实数集一个非常重要的性质是实数列收敛当且仅当它是 Cauchy 列。我们把这一性质抽象出来，就得到了如下定义：

例 8.2 度量空间 (X, d) 的点列 $\{x_n : n \in \mathbb{N}\}$ 称为 ***Cauchy* 列**，如果对于任何 $\epsilon > 0$，都可以找到序号 $N \in \mathbb{N}$，使得对于任何大于 N 的序号 $m, n \in \mathbb{N}$，$d(x_m, x_n) < \epsilon$ 成立.

度量空间 (X, d) 称为**完备的**，如果任意 *Cauchy* 列 $\{x_n : n \in \mathbb{N}\}$ 都收敛：$\exists a \in X, \lim_{n\to\infty} d(a, x_n) = 0.$

度量空间的任何收敛点列显然是 Cauchy 列，完备性本质上只是假设 Cauchy 收敛准则在该空间成立.

下面是一些完备度量空间的例子

例 8.3

- L^p 度量下下 $\mathbb{R}^n$ 是完备的.
- 使用度量 $d(x_1, x_2) = |x_1 - x_2|$，则 $X = \mathbb{R} \setminus \{0\}$ 不是完备度量空间. 考虑 $\{x_n = \frac{1}{n} : n \in \mathbb{N}\}$，它是 *Cauchy* 列，但该点列在 X 中没有极限（极限是 0）.
- $[0, 1]$ 到自身的连续函数空间 $C([0, 1])$ 在 L^∞ 度量下是完备的. 此时

$$d(f, g) = \sup_{x\in[0,1]} |f(x) - g(x)|,$$

完备性由一致收敛得到，函数空间是泛函分析中一个典型的研究对象.

下面我们给出与 Banach 不动点定理相关的概念：

定义 8.2 (压缩映射) 度量空间 (X,d) 到自身的映射 $f: X \to X$ 称为**压缩映射**，如果存在 $0 < q < 1$，使得不等式

$$d(f(x_1), f(x_2)) \leq q \cdot d(x_1, x_2)$$

对于 X 中的任何两个点 x_1, x_2 都成立.

用 $\delta - \epsilon$ 语言容易证明压缩映射一定是连续映射：

引理 8.1 压缩映射 $f: X \to X$ 是连续映射.

压缩映射一定有不动点，这就是 Banach 不动点定理：

定理 8.1 (Banach 不动点定理，压缩映像原理) 完备度量空间 (X,d) 到自身的压缩映射 $f: X \to X$ 具有唯一的不动点 a.

此外，对于任何点 $x_0 \in X$，迭代序列 $x_0, x_1 = f(x_0), \cdots, x_{n+1} = f(x_n), \cdots$ 收敛到 a. 收敛速度由以下估计给出：

$$d(a, x_n) \leq \frac{q^n}{1-q} d(x_1, x_0).$$

证明 首先证明存在性。$d(x_{n+1}, x_n) \leq q d(x_n, x_{n-1}) \leq \cdots \leq q^n d(x_1, x_0)$. 从而

$$\begin{aligned} d(x_{n+k}, x_n) &\leq d(x_n, x_{n+1}) + \cdots + d(x_{n+k-1}, x_{n+k}) \\ &\leq (q^n + \cdots + q^{n+k-1}) d(x_1, x_0) \leq \frac{q^n}{1-q} d(x_1, x_0). \end{aligned}$$

因此 $\{x_n\}$ 是 Cauchy 列，存在极限 $\lim_{n\to\infty} x_n = a \in X$. 结合压缩映射的连续性，有 $a = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n) = f(a)$.

然后证明唯一性。若 f 还有其他不动点 a_1, a_2，则 $0 \leq d(a_1, a_2) = d(f(a_1), f(a_2)) \leq q d(a_1, a_2)$. 而这当且仅当 $d(a_1, a_2) = 0$，即 $a_1 = a_2$ 时才可能成立. 最后证明收敛速度. 对 $d(x_{n+k}, x_n) \leq \frac{q^n}{1-q} d(x_1, x_0)$. 取 $k \to \infty$，得到 $d(a, x_n) \leq \frac{q^n}{1-q} d(x_1, x_0)$. □

例 8.4 (落在地面上的地图) 将一座公园的地图铺开在公园地面上，则地面上恰有唯一一点与地图上对应的点重合. 设公园可以用有界的面闭区域 Ω 表示. 设地图的压缩比是 $\lambda \in (0,1)$. 现在固定一个平面直角坐标系，把地图铺在区域 Ω 内，则从 Ω 内的点 x（公园中的地点）到地图上对应点 x' 的变换由下面的公式给出：

$$x' = f(x) := \lambda R x + b.$$

其中 R 和 b 分别为旋转和平移变换.

考虑 $\|\lambda R\| = \sup_{\|x\|=1} \|\lambda R x\| = \lambda < 1$, 由 *Banach* 不动点定理可知, 压缩映射 $f(x)$ 有唯一不动点 $a = f(a)$.

例 8.5 (梯度下降的收敛性) 我们优化目标是寻找二阶可微凸函数 $f(x), x \in \mathbb{R}^n$ 的最小值. 使用梯度下降方法: 每次往最小梯度方向移动. 假设对任意 $x \in \mathbb{R}^n$,

$$L \leq \lambda_{\min}(\nabla^2 f(x)) \leq \lambda_{\max}(\nabla^2 f(x)) \leq U.$$

其中 $\nabla^2 f(x)$ 是 f 的 *Hessian* 矩阵(二次导数), $U \geq L > 0$ 为给定的常数, $\lambda_{\min}(A), \lambda_{\max}(A)$ 表示矩阵 A 的最小、最大特征值.

我们要证明: 梯度下降能收敛到最小值点, 且具有指数收敛速度.

先看一下证明的思路, 我们要设法证明梯度下降算法是完备度量空间中的一个压缩映射。首先, 二阶可微凸函数的最小值点充分必要地满足 $\nabla f(x) = 0$. 其次, $\nabla f(x) = 0 \iff x \in \mathbb{R}^n$ 是梯度下降算子 $\mathcal{T}^{(\alpha)}$ 的不动点, 其中 $\mathcal{T}^{(\alpha)} : x \mapsto x - \alpha \nabla f(x)$, 这里 $\alpha \in \mathbb{R}_+$ 为步长. 最后, $\mathcal{T}^{(\alpha)}$ 是一个完备度量空间的压缩映射, 其压缩系数为 $q(\alpha) = 1 - L\alpha$. 因此梯度下降可以收敛至唯一的最小值点, 收敛速度可以由压缩系数估计.

为了使 $q(\alpha)$ 确实一个压缩系数, 我们需要 $\alpha < \min L^{-1}$. $\mathcal{T}^{(\alpha)}$ 的不动点恰好满足 $\nabla f(x) = 0$, 因此是最小值点. 我们只需要证明 $\mathcal{T}^{(\alpha)}$ 是压缩映射, 并给出压缩系数

由有限增量原理:

$$\left\| \mathcal{T}^{(\alpha)} x - \mathcal{T}^{(\alpha)} y \right\| \leq \sup_{z \in [x,y]} \left\| I - \alpha \nabla^2 f(z) \right\|_2 \cdot \|x - y\|_2 .$$

最后, 注意到 $\left\| I - \alpha \nabla^2 f(z) \right\|_2$ 等于 $I - \alpha \nabla^2 f(z)$ 特征值的最大模, 根据条件可知特征值的最大模 $\leq 1 - L\alpha$.

§8.2 Brouwer 不动点定理

下面我们考虑更一般的度量空间中的不动点定理, 为此我们需要引入连续映射的概念。

定义 8.3 设 X 和 Y 是度量空间 $(X, d_X), (Y, d_Y)$, 映射 $f : X \to Y$ 在点 $a \in X$ **连续**, 指的是

$$\forall \epsilon > 0 \exists \delta > 0 \forall x \in X (d_X(a, x) < \delta \Rightarrow d_Y(f(a), f(x)) < \epsilon).$$

如果它在每个点 $x \in X$ 连续, 称 f 为**连续映射**. X 到 Y 的连续映射的集合记为 $C(X, Y)$.

当度量空间为欧氏空间时，连续映射的定义与欧氏空间中连续映射的定义相同. 接下来，我们还需要几个集合的概念。

定义 8.4 (开集、闭集和紧集) *考虑度量空间 (X,d), $a \in X$ 的邻域 $B(a,\delta) := \{x \in X | d(a,x) < \delta\}$.*

- *集合 G 是*开集*G 指的是对于任何点 $x \in G$，满足 $B(x,\delta) \subset G$ 的邻域 $B(x,\delta)$ 存在.*
- *集合 F 是*闭集*，如果它的补集 $X \setminus F$ 是 (X,d) 中的开集.*
- *集合 K 是*紧集*，如果从 X 中任何覆盖 K 的开集族中可以选出有限个开集来覆盖 K.*

当度量空间为欧氏空间时，开集和紧集的定义与欧氏空间中的定义相同，紧集等价于有界闭集. 后一条在一般的度量空间不一定成立！

有了上面的准备，我们就可以叙述 Brouwer 不动点定理了：

定理 8.2 (Brouwer 不动点定理) *设 $M \subset \mathbb{R}^n$ 是一个非空紧凸集，而 $F : M \to M$ 是一个连续函数. 则存在 $x \in M$ 使得 $F(x) = x$ 成立.*

Brouwer 不动点定理可以通过该实际的例子来理解：将一张白纸平铺在桌面上，再将它揉成一团（不撕裂），放在原来白纸所在的地方，那么只要它不超出原来白纸平铺时的边界，那么白纸上一定有一点在水平方向上没有移动过. 这个断言依据 Brouwer 不动点定理在 $\mathbb{R}^2$ 的情况，因为把纸揉皱是一个连续的变换过程.

另一个例子：大商场等地方可以看到的平面地图，上面标有“您在此处”的红点. 如果标注足够精确，那么这个点就是把实际地形映射到地图的连续函数的不动点.

下面我们看一个 Brouwer 不动点定理的应用例子。首先引入矩阵**不可约**的概念：对于 n 阶方阵 A 而言，如果存在一个置换矩阵（通过交换单位阵的列获得）P 使得 $P^{\mathsf{T}}AP$ 为一个分块上三角阵，我们就称矩阵 A 是可约的，否则就称该矩阵是不可约的.

定理 8.3 (Perron-Frobenius 定理) *设 $A = (a_{ij})$ 为 $n \times n$ 不可约实矩阵，所有元素均非负，$a_{ij} \geq 0$，则下列结论成立.*

- *存在一个实特征值 r，其他特征值 λ 的模均不超过 r，即 $|\lambda| \leq r$.*
- *存在一个与 r 对应的特征向量，其所有元素恒正.*
- $\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}$.

证明 首先证明 A 存在一个正的特征值 $r > 0$. 考虑单纯形 $S := \{x \in \mathbb{R}^n | x \geq 0, \sum_i x_i = 1\}$. 则 $\forall x \in S$，有 $Ax \geq 0$.

断言 $Ax > 0$，若不然，A 存在某一列全 0（由 $x \geq 0$ 和 A 非负可得）. 此时可通过置换阵将该 0 列交换到第一列，则得到的矩阵为分块上三角，与不可约性矛盾.

可以在 S 上定义映射

$$T(x) = \frac{1}{\rho(x)} Ax,$$

其中 $\rho(x) > 0$ 使得 $T(x) \in S$.

显然 $T(x)$ 是 $S \to S$ 的连续映射. S 是一个有界凸闭集. 由 Brouwer 不动点定理，存在 $x_0 \in S, x_0 = T(x_0) = \frac{1}{\rho(x_0)} Ax_0$.

令 $r = \rho(x_0)$，则可得 r 为 A 的一个正的特征值.

我们接下来证明，与 r 对应的特征向量所有元素恒正. 由之前的证明，与 r 对应的特征向量 $x_0 \in S$，则 $x_0 \geq 0$. 我们证明 $x_0 > 0$.

考虑 $A = PBP^{-1}$，其中 P 是置换矩阵，则

$$PBP^{-1}x_0 = rx_0 \implies B(P^{-1}x_0) = r(P^{-1}x_0).$$

记 $\tilde{x}_0 = P^{-1}x_0$. 取 B 使得 $\tilde{x}_0 = (\xi, 0)^\mathsf{T}, \xi > 0$. 则

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} \xi \\ 0 \end{pmatrix} = \begin{pmatrix} r\xi \\ 0 \end{pmatrix}.$$

此时 $B_{21}\xi = 0$，由 $\xi > 0$ 可得 $B_{21} = 0$. 这与不可约矛盾，因此 $x_0 > 0$.

然后我们证明：若 α 是 A 的任意特征值，有 $|\alpha| \leq r$. 设 $0 \leq B \leq A, By = \beta y$. 记 $y^\star = |y| = (|y_i|)_i$. 于是有

$$|\beta| y^\star = |\beta y| = |By| \leq By^\star \leq Ay^\star.$$

由 A^T 不可约，存在特征值 $r_1 > 0$ 和特征向量 $x_1 > 0, A^\mathsf{T}x_1 = r_1 x_1$. 因此有

$$|\beta| x_1^\mathsf{T} y^\star \leq x_1^\mathsf{T} A y^\star = r_1 x_1^\mathsf{T} y^\star.$$

由 $x_1^\mathsf{T} y^\star > 0$，则 $|\beta| \leq r_1$. 令 $B = A$ 可得 $|\alpha| \leq r_1$，特别地 $r \leq r_1$. 同样有 $r_1 \leq r$，故 $r = r_1$.

最后证明：

$$\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}.$$

以这样的方式获得 $\tilde{A}$：将 A 的每一行都扩增（不减小某个元素），使得每一行都达到 $\max_i \sum_j a_{ij}$. 此时 $\max_i \sum_j a_{ij}$ 成为 A 的一个正特征值，且特征向量 $\tilde{x}_0 = \frac{1}{n} \cdot \mathbf{1} \in S$. 由之前

的结论，假设 $0 \leq A \leq \tilde{A}$，可以得到 $\tilde{A}$ 的正特征值 $\tilde{r} \geq r$. 因此 $r \leq \max_i \sum_j a_{ij}$. 同理缩小 A 可得 $\min_i \sum_j a_{ij} \leq r$. □

Perron-Frobenius 定理在 Markov 链中的有非常重要应用。回忆 Markov 链的平稳分布：满足矩阵方程 $\pi = \pi P$ 和 $\sum_i \pi_i = 1$. 设该 Markov 链状态有限且对应的转移矩阵 P 是（非负实）不可约方阵. 由 Perron-Frobenius 定理，P 存在一个特征值 $1 = \min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij} = 1$，对应一个正特征向量 $x_0 \in S = \{x \in \mathbb{R}^n | x \geq 0, \sum_i x_i = 1\}$. 因此不可约有限状态 Markov 链必然存在平稳遍历分布.

§8.3 不动点的一般视角

第四部分

逻辑与博弈

第五部分

认知逻辑

参考文献

[Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[Inf] Information | Etymology, origin and meaning of information by etymonline. https://www.etymonline.com/word/information.

[Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.

[McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.

[Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.

[Uff22] Jos Uffink. Boltzmann's Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.

[李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

索引