第 3 章：无约束优化

(2)：下降法的理论

授课教师：曹语

课程主页：https://yucaoyc.github.io/math3806

背景和目标

我们已经见过梯度和最速下降法可以用于求解无约束凸优化问题。上一节我们已经知道梯度下降法 ‘‘怎么走’’，但还不知道它 ‘‘为什么会收敛’’、‘‘什么时候会慢’’。

特别是，第 1 部分里看到的之字形轨迹，这一节要用理论来解释。

上节课结束时，我们留下了两个问题：

Q1 下降方向怎么选？（已解决：在 L^2 意义下，负梯度是最速下降方向；最速下降在不同范数意义下具有不同形式）

Q2 步长怎么确定？（已解决：精确直线搜索或回溯直线搜索）

Q3 算法能否收敛到最优解？**【即收敛性】**

Q4 需要多少步才能达到精度要求？**【和条件数的关系】**

学习目标

学完本节后，希望能够：

- 用泰勒展开理解一步下降量为什么可以被控制
- 解释强凸性和条件数的含义
- 复述梯度下降法的线性收敛结论
- 说明为什么条件数越大，问题越难

复习回顾

- 上节实验中，$\log(e_N)$ 对 N 的图像大致是什么形状?
- 对于例子 $f(\boldsymbol{x}) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$，$\gamma$ 取什么时，梯度下降一般而言会变慢?

目录

泰勒展开：多项式近似的视角

核心思想： 用简单的多项式来近似复杂的函数。

以一维函数为例：

0 阶近似：$f(y) \approx f(x)$

1 阶近似：$f(y) \approx f(x) + f'(x)\,(y - x)$

2 阶近似：$f(y) \approx f(x) + f'(x)\,(y - x) + \dfrac{f''(x)}{2}(y - x)^2$

直观理解：

- 0 阶：用常数近似（只关心函数值）
- 1 阶：用直线近似（关心函数值和斜率）
- 2 阶：用抛物线近似（还关心曲率）

0 阶泰勒展开（均值定理）

定理 [0 阶泰勒展开的余项定理]

若 f 是一维函数，且导数存在，则对于任意的 x, y，存在 z（在前两者之间）使得

$$f(y) = f(x) + f'(z)(y - x)$$

证明： 考虑 $g(z) = f(z) - f(x) - \frac{f(y)-f(x)}{y-x}(z - x)$，可验证 $g(x) = g(y) = 0$，并使用罗尔中值定理可得存在某点 z 满足，$g'(z) = 0$，即 $f'(z) = \frac{f(y)-f(x)}{y-x}$。

直观含义： 在 x 和 y 之间，存在某点 z 使得该点的切线斜率等于连接 $(x, f(x))$ 和 $(y, f(y))$ 的直线斜率。

1 阶泰勒展开（含余项）

定理 [1 阶泰勒展开的余项定理]

若 f 是一维函数，且二阶导数存在，则对于任意的 x, y，存在 η（在前两者之间）使得

$$f(y) = f(x) + f'(x)(y - x) + \frac{f''(\eta)}{2}(y - x)^2$$

证明： 设余项为 R，即 $f(y) = f(x) + f'(x)(y - x) + \frac{R}{2}(y - x)^2$。构造

$$g(z) = f(z) - f(x) - f'(x)(z - x) - \frac{R}{2}(z - x)^2$$

则 $g(x) = g(y) = 0$ 且 $g'(x) = 0$。对 g 和 g' 分别用罗尔定理，得 $\exists \eta$ 使 $g''(\eta) = 0$，故 $R = f''(\eta)$。

关键点： 余项中的 η 是未知的，但 $f''(\eta)$ 被某个范围内的值所控制。

从一维到多维：n 维泰勒展开

我们把多维问题压缩到连接 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的一条线段上，于是就可以使用一维结论。

考虑 n 维函数 f，将问题转化为一维：

$$g(t) = f\big(\boldsymbol{x} + t\big(\boldsymbol{y} - \boldsymbol{x}\big)\big)$$

对 $g(t)$ 应用 1 阶泰勒展开（取 $t = 1$）：

$$g(1) = g(0) + g'(0) \cdot 1 + \frac{g''(r)}{2} \cdot 1^2$$

从一维到多维：n 维泰勒展开（续）

把 $g(1), g(0), g'(0), g''(r)$ 分别写回原函数，就得到：

计算可得**多维泰勒展开公式**：

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x})^\top \nabla^2 f(\boldsymbol{z}) (\boldsymbol{y} - \boldsymbol{x})$$

其中 $\boldsymbol{z} = \boldsymbol{x} + r(\boldsymbol{y} - \boldsymbol{x})$，$r \in (0, 1)$。

泰勒展开的应用：凸函数的一阶条件

若 $\nabla^2 f \succeq \mathbf{0}_{n\times n}$（$f$ 是凸函数），则

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \qquad \forall \boldsymbol{x}, \boldsymbol{y}$$

这就是凸函数的一阶条件！

也就是说：函数图像总在任意切平面的上方。

目录

强凸：比普通凸函数更强的条件

定义 [强凸]

目标函数 f 在 S 上是m-强凸的，若

$$\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}_{n\times n}, \qquad \forall \boldsymbol{x} \in S$$

其中 $\boldsymbol{I}_{n\times n}$ 是单位矩阵，$m > 0$。

几何直观：

- 普通凸：函数图像向上弯
- 强凸：函数图像至少以 m 的曲率向上弯

强凸：等价的不等式形式

强凸也可以写成下面这个更有用的不等式：

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\top}(\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$

它比普通凸函数多出了一个二次项，因此能提供更强的下界。

强凸的性质

定理 [强凸的下界性质]

若 f 在 S 上是 m-强凸的，则

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in S \tag{1}$$

并且关于 $\boldsymbol{y}$ 求极小值，可得

$$p^\star \geq f(\boldsymbol{x}) - \frac{1}{2m}\|\nabla f(\boldsymbol{x})\|_2^2, \qquad \forall \boldsymbol{x} \in S \tag{2}$$

第一条给函数值一个二次下界；第二条把 ‘‘梯度小’’ 和 ‘‘接近最优’’ 联系起来。

【证明见课本 9.1.2，或见课内】

强凸性质的应用：（1）终止条件的理论保障

由公式(2)可知：

$$\|\nabla f(\boldsymbol{x})\|_2 \leq \sqrt{2m\epsilon} \implies f(\boldsymbol{x}) \leq p^\star + \epsilon$$

这解释了为什么可以用梯度大小作为停止准则！

也就是说：梯度足够小，就能保证函数值已经接近最优值。

强凸性质的应用：（2）下水平集有界

回顾下水平集 $S = \{\boldsymbol{x} \mid f(\boldsymbol{x}) \leq f(\boldsymbol{x}^{(0)})\}$。由公式(1)，取 $\boldsymbol{x} = \boldsymbol{x}^{(0)}$，$\boldsymbol{y} \in S$，则

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^{(0)}) + \nabla f(\boldsymbol{x}^{(0)})^{\top}(\boldsymbol{y} - \boldsymbol{x}^{(0)}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}^{(0)}\|_2^2$$

再用 $f(\boldsymbol{y}) \leq f(\boldsymbol{x}^{(0)})$，整理得

$$\|\boldsymbol{y} - \boldsymbol{x}^{(0)}\|_2^2 + \frac{2}{m}\nabla f(\boldsymbol{x}^{(0)})^{\top}(\boldsymbol{y} - \boldsymbol{x}^{(0)}) \leq 0$$

配平方后得到

$$\left\|\boldsymbol{y} - \boldsymbol{x}^{(0)} + \frac{\nabla f(\boldsymbol{x}^{(0)})}{m}\right\|_2^2 \leq \left\|\frac{\nabla f(\boldsymbol{x}^{(0)})}{m}\right\|_2^2$$

所以，所有满足 $f(\boldsymbol{y}) \leq f(\boldsymbol{x}^{(0)})$ 的点都被限制在某个球里，算法的迭代点不会在空间里无限乱跑。**故下水平集 S 有界。**

Hessian 上界

再假设 $\nabla^2 f$ 在 S 上连续。由于下水平集 S 有界且为闭集，因此 S 是紧集；根据极值定理，$\exists M$ 使得

$$\nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}_{n\times n}, \qquad \forall \boldsymbol{x} \in S$$

这说明曲率不会无限大。

Hessian 上界（续）

这给出了函数的上界：

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\top}(\boldsymbol{y} - \boldsymbol{x}) + \frac{M}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$

以及关于不等式两边对 $\boldsymbol{y}$ 求极小值可得：

$$p^{\star} \leq f(\boldsymbol{x}) - \frac{1}{2M}\|\nabla f(\boldsymbol{x})\|_2^2$$

逻辑上可以理解为：强凸给出曲率下界，连续 Hessian 在紧集上给出曲率上界，因此曲率被夹在 $[m, M]$ 之间。

条件数：问题难度的度量

结合上下界，我们有条件：

$$m\boldsymbol{I}_{n\times n} \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}_{n\times n}, \qquad \forall \boldsymbol{x} \in S$$

定义

比值 $\kappa = M/m$ 被称为该问题的**条件数**（condition number）。

直观理解：

- $\kappa = 1$：所有方向曲率相同 $\Rightarrow$ 等高线接近圆形，最容易优化
- $\kappa \gg 1$：不同方向曲率差异大 $\Rightarrow$ 等高线狭长，梯度下降会出现之字形摆动

这正是我们在第 1 部分实验中观察到的现象！

条件数的例子

例子 1：对角二次函数

$$f(\boldsymbol{x}) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$$

- 若 $\gamma > 1$，则 $m = 1$，$M = \gamma$，条件数 $\kappa = \gamma$
- 若 $\gamma < 1$，则 $m = \gamma$，$M = 1$，条件数 $\kappa = 1/\gamma$

因此 $\gamma \to 0$ 或 $\gamma \to \infty$ 时，$\kappa \to \infty$，问题变难。

条件数的例子（续）

例子 2： 一般二次函数

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + r$$

其中 $\boldsymbol{P}$ 对称正定，则 $m = \lambda_{\min}(\boldsymbol{P})$，$M = \lambda_{\max}(\boldsymbol{P})$，条件数为

$$\kappa = \frac{\lambda_{\max}(\boldsymbol{P})}{\lambda_{\min}(\boldsymbol{P})}$$

目录

精确直线搜索的收敛速度

定理 [梯度法 + 精确直线搜索]

设 f 为 m-强凸且 Hessian 有上界 M，使用精确直线搜索的梯度下降法满足

$$e_{n+1} \leq \left(1 - \frac{m}{M}\right) e_n, \qquad \forall n \tag{3}$$

其中 $e_n := f(\boldsymbol{x}^{(n)}) - p^\star$。

因此误差按固定比例衰减：

$$e_N \leq \left(1 - \frac{m}{M}\right)^N e_0 = \left(1 - \frac{1}{\kappa}\right)^N e_0 \leq e^{-N/\kappa} e_0 \tag{4}$$

线性收敛的意思是：误差每一步大约按固定比例缩小。

当 $N \to \infty$ 时，$e_N \to 0$，即算法收敛。

达到指定精度的迭代次数

先看结论的直观含义：要更高精度，或者问题更难，就需要更多步。

为达到误差阈值 ϵ，即 $e_N \le \epsilon$，可知当

$$N \ge \kappa \log\left(\frac{e_0}{\epsilon}\right)$$

误差一定小于阈值。

直观理解：

- 初始误差 e_0 越大 $\Rightarrow$ 需要更多迭代
- 误差阈值 ϵ 越小 $\Rightarrow$ 需要更多迭代
- 问题难度（即 κ）越大 $\Rightarrow$ 需要更多迭代

【证明见课本 9.3.1，或见课内】

例题：具体数值计算

问题： 对 $f(\boldsymbol{x}) = \frac{1}{2}(x_1^2 + 10x_2^2)$，初值 $\boldsymbol{x}_0 = \begin{bmatrix} 9, 2 \end{bmatrix}$，需要误差不超过 10^{-6}，精确直线搜索需要多少步?

先请判断：这个问题的 m、M 和 κ 分别是多少?

例题：具体数值计算

问题： 对 $f(\boldsymbol{x}) = \frac{1}{2}(x_1^2 + 10x_2^2)$，初值 $\boldsymbol{x}_0 = \begin{bmatrix} 9, 2 \end{bmatrix}$，需要误差不超过 10^{-6}，精确直线搜索需要多少步?

解答：

- 这里 $\gamma = 10$，所以 $m = 1$，$M = 10$，$\kappa = 10$
- 收敛常数 $c = 1 - m/M = 1 - 1/10 = 0.9$
- 初值误差 $e_0 = f(9, 2) - 0 = \frac{1}{2}(81 + 40) = 60.5$
- 解不等式 $c^N e_0 \leq 10^{-6}$:

$$0.9^N \cdot 60.5 \leq 10^{-6} \implies N \geq \frac{\log\left(10^{-6}/60.5\right)}{\log(0.9)} \approx 170.07$$

因此 $N \geq 171$ 步必然保证误差要求。用公式(4)来估计也可。

例题：从这个结果学到什么?

这个例子说明：

- 即使只是二维二次函数，达到很高精度也可能需要很多步
- 收敛速度很大程度上受条件数控制
- 这解释了为什么上一节实验里狭长等高线会导致慢收敛

回溯直线搜索的收敛速度

定理 [梯度法 + 回溯直线搜索]

设 f 为 m-强凸且 Hessian 有上界 M，回溯直线搜索参数满足 $\alpha \in (0, 1/2)$，$\beta \in (0, 1)$。则梯度下降法满足

$$e_{n+1} \leq c\, e_n, \qquad \forall n \tag{5}$$

其中 $c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1$。

虽然常数 c 的表达式更复杂，但结论和精确直线搜索一样：误差仍按固定比例衰减。为达到精度 ϵ，需要

$$N \geq \frac{\log(e_0/\epsilon)}{\log(1/c)}$$

结论与精确直线搜索类似：收敛速度由条件数 κ 决定。 【证明不要求】

回溯直线搜索：结论解读

虽然回溯直线搜索没有精确找到最佳步长，但它仍然有两个优点：

- 每一步都能保证函数值下降
- 在线性收敛意义下，结论和精确直线搜索同类型

因此它在实践中常常更容易使用。

L^2 误差的分析

除了函数值误差 $e_N = f(\boldsymbol{x}^{(N)}) - p^\star$，我们也可以关注点列误差 $\|\boldsymbol{x}^{(N)} - \boldsymbol{x}^\star\|_2$。

由强凸性(1)，取 $\boldsymbol{x} = \boldsymbol{x}^\star$。由于无约束可微最优解满足 $\nabla f(\boldsymbol{x}^\star) = 0$，因此

$$f(\boldsymbol{y}) \geq p^\star + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}^\star\|_2^2$$

因此

$$\|\boldsymbol{x}^{(N)} - \boldsymbol{x}^\star\|_2 \leq \sqrt{\frac{2}{m}\big(f(\boldsymbol{x}^{(N)}) - p^\star\big)} \leq \sqrt{\frac{2c^N e_0}{m}}$$

这说明：不仅函数值会靠近最优值，迭代点本身也会靠近最优点。

L^2 误差：直观理解

这个结果说明：

- 不仅函数值会越来越接近最优值
- 迭代点本身也会越来越接近最优点
- 因此我们既可以研究函数值误差，也可以研究点列误差

理论与实践的桥梁

理论告诉我们的：

- 梯度下降法对于强凸函数是线性收敛的
- 收敛速度由条件数 κ 决定
- κ 越大，问题越难

实践中的启示：

- 条件数大 $\Rightarrow$ 可考虑预处理或坐标变换
- 条件数未知 $\Rightarrow$ 可通过实验观察收敛行为
- 画 $\log(e_N)$ 对 N 的图，可以验证理论预测

速查练习

- 若 κ 变大，梯度下降通常会变快还是变慢？为什么？
- 若 $\|\nabla f(\boldsymbol{x})\|$ 很小，可以说明什么？

目录

总结

本节要点：

- **泰勒展开**：是收敛分析的基础工具
- **强凸性**：提供更强的 Hessain 下界
- **条件数**：解释了问题为什么有时会很难
- **收敛性**：梯度下降对强凸函数是线性收敛的

阅读作业 & 参考资料：

- 课本第 9.1 - 9.4 章