第 3 章：无约束优化

(2)：下降法的理论

曹语

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

背景和目标

我们已经看到了梯度和最速下降法可以被应用于求解无约束的凸优化问题

待解决的理论问题：

Q3 为何下降法能保证 $f(\boldsymbol{x}^{(k)}) \leq p^\star + \epsilon$，当迭代步数 k 足够大的时候？

【即收敛性】

Q4 该方法的效率是多少？即需要多少步 k 才能实现 $f(\boldsymbol{x}^{(k)}) \leq p^\star + \epsilon$？

【主要我们要看和条件数的关系，即问题本身的难度】

练习

- 请描述梯度下降法收敛的速度，即描述误差 e_N 和迭代次数 N 的关系；
- 并且描述如何在实际问题中，通过数值验证相应的关系（即如何画图验证）。

目录

回顾

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 $f'(z) = \frac{f(y)-f(x)}{y-x}$。

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： 固定 x, y，通过表达式 $f(y) = f(x) + f'(x)(y-x) + \frac{R}{2}(y-x)^2$ 来定义 R (此时它是一个数)。令 $g(z) = f(z) - f(x) - f'(x)(z-x) - \frac{R}{2}(z-x)^2$ 是一个关于 z 的函数。显然 $g(x) = g(y) = 0$，且 $g'(x) = 0$，因此通过罗尔中值定理存在某个数 $g'(\xi) = 0$ (ξ 在 x 和 y 之间)；再次使用罗尔中值定理可得存在某个数 η，使得 $g''(\eta) = 0$，因此得到 $0 = g''(\eta) = f''(\eta) - R$。

对于一般维数

对于 n 维函数 f，考虑 $g(t) = f\big(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\big)$

通过对于一维函数 g 的 1 阶泰勒展开的余项定理，存在某个 r 使得

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

计算可得

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\big(\mathbf{y} - \mathbf{x}\big)^\top \nabla^2 f(\mathbf{z})\big(\mathbf{y} - \mathbf{x}\big), \quad \mathbf{z} = \mathbf{x} + r\big(\mathbf{y} - \mathbf{x}\big)$$

这个公式值得记住！

应用

若 $\nabla^2 f \succeq \mathbf{0}_{n\times n}$（等价于 f 是凸函数，见凸函数的二阶条件），则

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

这其实就是凸函数的一阶条件的形式。

提示： 此处我们用了二阶条件来证明一阶条件（即使用了更强的假设），之前介绍的证明方法更加一般

目录

强凸（Hessian 具有下界）

在分析部分，我们会额外假设目标函数在 S 上是强凸的：即 $\exists m > 0$ 使得

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

其中 $\boldsymbol{I}_{n\times n}$ 是单位矩阵。

定理

若 f 在 S 上是 m-强凸的，则

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

并且关于左右两边对于 $\boldsymbol{y}$ 求极小值，我们可以进一步得知

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

【证明见课本 9.1.2，或见课内】

上述公式的用处

- 公式(2)的推论：只要梯度足够小 $\|\nabla f(\boldsymbol{x})\|_2 \leq \sqrt{2m\epsilon}$，则 $f(\boldsymbol{x}) \leq p^\star + \epsilon$。
  1. 即之前提到的终止条件不仅直观上合理，且具有一定数学上的保障
  2. 这是一种概念性的保证，不是算法级别的保证（m 可能未知）
- 公式(1)的推论：下水平集 S 是有界的

证明：$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^{(0)}) + \nabla f(\boldsymbol{x}^{(0)})^\top(\boldsymbol{y} - \boldsymbol{x}^{(0)}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}^{(0)}\|_2^2$；由于 $\boldsymbol{y} \in S$ 满足 $f(\boldsymbol{y}) \leq f(\boldsymbol{x}^{(0)})$，因此

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

通过配平方 $\left\|\boldsymbol{y} - \boldsymbol{x}^{(0)} + \frac{\nabla f(\boldsymbol{x}^{(0)})}{m}\right\|_2 \leq \frac{\|\nabla f(\boldsymbol{x}^{(0)})\|_2}{m}$，因此可知该下水平集 S 有界。

Hessian 上界

由于 S 有界，且是闭集（我们之前直接假设了），根据极值定理，$\exists M$ 使得

$$\nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}_{n\times n}$$

由此可知

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{M}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \tag{3}$$

并且关于左右两边对于 $\boldsymbol{y}$ 求极小值，我们可以进一步得知

$$p^\star \leqslant f(\boldsymbol{x}) - \frac{1}{2M}\|\nabla f(\boldsymbol{x})\|_2^2 \tag{4}$$

【证明见课本 9.1.2，或见课内】

条件数

前面已经说明了（通过假设或者通过极值定理），

$$m\boldsymbol{I}_{n\times n} \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}_{n\times n}$$

比值 $\kappa = M/m$ 被称为该问题的条件数

例子 1：$f(\boldsymbol{x}) = \frac{1}{2}(x_1^2 + \gamma x_2^2)$，若假设 $\gamma > 1$，则 $m = 1$, $M = \gamma$，因此条件数 $\kappa = \gamma$；若 $\gamma < 1$，则 $m = \gamma$, $M = 1$，条件数 $\kappa = \frac{1}{\gamma}$；因此 $\gamma \to 0$ 和 $\gamma \to \infty$ 的趋势可以用条件数 $\kappa \to \infty$ 这一个指标来衡量。

例子 2：对于一般的 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + r$，我们可以得到

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

目录

定理 (梯度法 + 精确直线搜索)

- 考虑精确直线搜索，每次迭代我们具有误差关系 $e_n := f(\boldsymbol{x}^{(n)}) - p^*$

$$e_{n+1} \le (1 - \frac{m}{M}) e_n, \qquad \forall n$$

因此误差 $e_N \le \left(1 - \frac{m}{M}\right)^N e_0 \le e^{-\frac{N}{\kappa}} e_0$。故而当 $N \to \infty$，$e_N \to 0$。

- 为了实现误差小于阈值 ϵ，N 的取值满足如下关系即可保证

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

初值误差 e_0 越大，精度要求越高（ϵ 越小），问题越困难（即条件数 κ 越大），迭代所需要的次数越多。

【证明见课本 9.3.1，或者课内】

问题：对于问题 $f(\boldsymbol{x})=\frac{1}{2}\left(x_1^2+\gamma x_2^2\right)$，$\gamma=10$，初值选成了 $\boldsymbol{x}_0=\left[9.0,2.0\right]$，若需要误差 $f(\boldsymbol{x}^{(N)})-p^*$ 不超过 10^{-6}，则通过精确直线搜索多少步必然能保证该误差?

答案：$N\geq 152.2$，因此 $N\geq 153$。实际所需的值是小于该数的。

定理 (梯度法 + 回溯直线搜索)

- 每次迭代我们具有误差关系

$$e_{n+1} \le c e_n, \qquad \forall n$$

其中 $c = 1 - \min\{2m\alpha, 2\beta\alpha\frac{m}{M}\} < 1$。因此收敛性得到保证。

- 为了实现误差小于阈值 ϵ，N 的取值满足如下关系即可保证

$$N \ge \frac{\log\left(\frac{e_0}{\epsilon}\right)}{\log\left(\frac{1}{c}\right)} \quad \rightarrow$$

初值误差 e_0 越大，精度要求越高（ϵ 越小），问题越困难（即条件数 κ 越大），迭代所需要的次数越多。

【证明见课本 9.3.1，或者课内】

在此处分析中，我们看到

$$f(\boldsymbol{x}^{(N)}) - p^* \leq c^N e_0, \qquad e_0 = f(\boldsymbol{x}^{(0)}) - p^*$$

其中 c 对于精确直线搜索和回溯直线搜索的情况具有不同值。另外一种合理的误差量化是 $\left\|\boldsymbol{x}^{(N)} - \boldsymbol{x}^*\right\|_2$。

在公式(1)，我们可以得到 $f(\boldsymbol{y}) \geq p^* + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}^*\|_2^2$。因此，

$$\left\|\boldsymbol{x}^{(N)} - \boldsymbol{x}^*\right\|_2 \leq \frac{2}{m}\left(f(\boldsymbol{x}^{(N)}) - p^*\right) \leq \frac{2}{m} c^N e_0$$

为了实现 L^2 误差小于 ϵ，我们仅需要取

$$N \geq \frac{\log(2/m) + \log(e_0/\epsilon)}{\log(1/c)}$$

额外量 $\log(2/m)$ 对于极高精度要求时 $(\epsilon \ll 1)$ 相比于 $\log(e_0/\epsilon)$ 是一个小量。

目录

总结

- 知道强凸的概念，以及泰勒展开的余项定理
- 能够复述收敛性的结论，有余力的话建议知道收敛性的证明

阅读作业 & 参考资料:

- 课本第 9.1 - 9.4 章