最优化方法　第四讲

Weiwen Wang(王伟文)

暨南大学

2025 年 4 月 3 日

目录

无约束优化

考虑无约束优化问题

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \tag{1}$$

此处假设 $f:\mathbb{R}^d\to\mathbb{R}$ 可微.
采用如下迭代法求解该优化问题, 相应的迭代格式为

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+h_k\boldsymbol{d}_k,$$

其中 $\{h_k\}$ 为预设步长, $\boldsymbol{d}_k$ 为移动方向.

由 f 可微知 (Proposition 1.0.14, Lecture 1)

$$f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k + h_k \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + h_k \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k + o(h_k \|\boldsymbol{d}_k\|).$$

由 f 可微知 (Proposition 1.0.14, Lecture 1)

$$f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k + h_k \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + h_k \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k + o(h_k \|\boldsymbol{d}_k\|).$$

给定步长 h_k, 设 $\boldsymbol{d}_k$ 为单位向量, 即 $\|\boldsymbol{d}_k\| = 1$, 最优方向应该使得 $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})$ 尽可能地大, 即使得每一次迭代函数值以最快的速度下降, 此时应使得

$$-\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k$$

取最大值.

由 Cauchy-Schwartz 不等式

$$-\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k \leq \|\nabla f(\boldsymbol{x}_k)\| \cdot \|\boldsymbol{d}_k\| = \|\nabla f(\boldsymbol{x}_k)\|.$$

由 Cauchy-Schwartz 不等式

$$-\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k \le \|\nabla f(\boldsymbol{x}_k)\| \cdot \|\boldsymbol{d}_k\| = \|\nabla f(\boldsymbol{x}_k)\|.$$

当 $\nabla f(\boldsymbol{x}_k) = 0$ 或 $\boldsymbol{d}_k = -\frac{\nabla f(\boldsymbol{x}_k)}{\|\nabla f(\boldsymbol{x}_k)\|}$ 时等号成立，即最优方向 $\boldsymbol{d}_k$ 与 $-\nabla f(\boldsymbol{x}_k)$ 同向. 相应的迭代格式为

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \nabla f(\boldsymbol{x}_k),$$

称为梯度下降法.

若注意到 $\nabla f(\boldsymbol{x}_k) \neq 0$,

$$\lim_{t\to 0^+} \frac{f(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)) - f(\boldsymbol{x}_k)}{t} = f'(\boldsymbol{x}_k; -\nabla f(\boldsymbol{x}_k)) = -\nabla f(\boldsymbol{x}_k)^T \nabla f(\boldsymbol{x}_k) < 0$$

故存在步长 t_k, 使得 $f(\boldsymbol{x}_k - t_k\nabla f(\boldsymbol{x}_k)) < f(\boldsymbol{x}_k)$, 函数值随着迭代进行单调不增.

步长策略

迭代步长常用的三种选择包括,

- 常数步长 $h_k = h > 0$
- 精确线搜索 (exact line search)

$$h_k = \arg\min_{h>0} f(\boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k)).$$

- 回溯法 (backtracking)　设有参数 $s>0$, $\alpha\in(0,1),\beta\in(0,1)$. 取 $h_k=s$.
  若
  $$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_k-h_k\nabla f(\boldsymbol{x}_k))<\alpha h_k\|\nabla f(\boldsymbol{x}_k)\|^2$$
  取 $h_k=\beta h_k$. 即取步长 $h_k=\beta^{i_k}s$, 使之满足
  $$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_k-\beta^{i_k}s\nabla f(\boldsymbol{x}_k))\geq\alpha\beta^{i_k}s\|\nabla f(\boldsymbol{x}_k)\|^2,$$
  i_k 为步长首次 满足上述不等式的在第 k 次迭代中回溯法执行的次数. 回溯法能够在有限步内停止.(原因是什么?)

充分下降引理

假设目标函数 f 满足一阶 Lipschitz 连续性质, 即

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

由 Theorem 2.0.8, Lecture 2, 立即可以得到以下充分下降引理

引理 1 (充分下降引理)

设 f 为 L-smooth, 则对任意 $\boldsymbol{x} \in \mathbb{R}^d$ 及 $h > 0$, 有

$$f(\boldsymbol{x}) - f(\boldsymbol{x} - h\nabla f(\boldsymbol{x})) \geq h(1 - \frac{Lh}{2})\|\nabla f(\boldsymbol{x})\|^2.$$

充分下降引理的最优步长

充分下降引理说明，若要获得足够大的目标函数下降量，应关于步长 h 最大化 $h(1-\frac{Lh}{2})$，此时 $h=\frac{1}{L}$，相应地，在梯度下降法中有

$$f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2.$$

不同步长策略的下降性质

引理 2

设 f 为 L-smooth, $\{\boldsymbol{x}_k\}$ 为梯度下降法产生的迭代序列, 步长策略取常数步长 $h \in (0, \frac{2}{L})$, 或精确线搜索, 或以 (s, α, β) 为参数的回溯法, 均有

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq M \|\nabla f(\boldsymbol{x}_k)\|^2$$

其中

$$M = \begin{cases} h(1 - \dfrac{Lh}{2}), & \text{常数步长;} \\ \dfrac{1}{2L}, & \text{精确线搜索;} \\ \alpha \min\{s, \dfrac{2(1-\alpha)\beta}{L}\}, & \text{回溯法.} \end{cases}$$

情形 1：常数步长 $h \in (0, \frac{2}{L})$ 充分下降引理的直接结论.

情形 1：常数步长 $h \in (0, \frac{2}{L})$ 充分下降引理的直接结论.

情形 2：精确线搜索 设 $\bar{h} = \arg\min_{h>0} f(\boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k))$, $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \bar{h}\nabla f(\boldsymbol{x}_k)$.

由 $\bar{h}$ 的最优性, 有

$$f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - \bar{h}\nabla f(\boldsymbol{x}_k)) \le f(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)),$$

情形 1: **常数步长** $h \in (0, \frac{2}{L})$ 充分下降引理的直接结论.

情形 2 : **精确线搜索** 设 $\bar{h} = \arg\min_{h>0} f(\boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k))$, $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \bar{h}\nabla f(\boldsymbol{x}_k)$.

由 $\bar{h}$ 的最优性, 有

$$f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - \bar{h}\nabla f(\boldsymbol{x}_k)) \le f(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)),$$

故

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \ge f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)) \ge \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2.$$

情形 3: 回溯法 若 $s(1-\frac{Ls}{2}) \geq \alpha s$, 由充分下降引理, 此时回溯法步长 $h_k = s$.

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \alpha s \|\nabla f(\boldsymbol{x}_k)\|^2.$$

情形 3: 回溯法 若 $s(1-\frac{Ls}{2}) \geq \alpha s$, 由充分下降引理, 此时回溯法步长 $h_k = s$.

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \alpha s \|\nabla f(\boldsymbol{x}_k)\|^2.$$

否则, 由回溯法的执行过程知

$$\beta^{i_k-1} s(1 - \frac{L\beta^{i_k-1}s}{2}) \|\nabla f(\boldsymbol{x}_k)\|^2 \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - \beta^{i_k-1} s \nabla f(\nabla \boldsymbol{x}_k)) < \alpha \beta^{i_k-1} \|\nabla f(\boldsymbol{x}_k)\|^2$$

解得 $\beta^{i_k} s > \frac{2(1-\alpha)\beta}{L}$,

情形 3: 回溯法 若 $s(1-\frac{Ls}{2})\geq\alpha s$, 由充分下降引理, 此时回溯法步长 $h_k=s$.

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\geq\alpha s\|\nabla f(\boldsymbol{x}_k)\|^2.$$

否则, 由回溯法的执行过程知

$$\beta^{i_k-1}s(1-\frac{L\beta^{i_k-1}s}{2})\|\nabla f(\boldsymbol{x}_k)\|^2\leq f(\boldsymbol{x}_k)-f(\boldsymbol{x}_k-\beta^{i_k-1}s\nabla f(\nabla\boldsymbol{x}_k))<\alpha\beta^{i_k-1}\|\nabla f(\boldsymbol{x}_k)\|^2$$

解得 $\beta^{i_k}s>\frac{2(1-\alpha)\beta}{L}$,
故 $h_k\geq\min\{s,\frac{2(1-\alpha)\beta}{L}\}$, 又因为回溯法步长满足

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\geq\alpha h_k\|\nabla f(\boldsymbol{x}_k)\|^2\geq\alpha\min\{s,\frac{2(1-\alpha)\beta}{L}\}\|\nabla f(\boldsymbol{x}_k)\|^2.$$

L-smooth 函数收敛性

定理 3

设 f 为 L-smooth, $\{\boldsymbol{x}_k\}$ 为梯度下降法产生的迭代序列, $\boldsymbol{x}^*$ 为无约束优化问题的全局最优解, 则

$$\min_{k=0,\ldots,K} \|\nabla f(\boldsymbol{x}_k)\| \le \sqrt{\frac{f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*)}{M(K+1)}},$$

其中

$$M=\begin{cases} h(1-\dfrac{Lh}{2}), & \text{常数步长;}\\ \dfrac{1}{2L}, & \text{精确线搜索;}\\ \alpha\min\{s,\dfrac{2(1-\alpha)\beta}{L}\}, & \text{回溯法.}\end{cases}$$

证明

由引理 2.2 有

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\geq M\|\nabla f(\boldsymbol{x}_k)\|^2$$

记 $\Delta_k=f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)$, 则

$$M\|\nabla f(\boldsymbol{x}_k)\|^2\leq\Delta_k-\Delta_{k+1},\quad k=0,1,\ldots,K$$

证明

由引理 2.2 有

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq M \|\nabla f(\boldsymbol{x}_k)\|^2$$

记 $\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$, 则

$$M\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \Delta_k - \Delta_{k+1}, \quad k = 0, 1, \ldots, K$$

对此 $K+1$ 项求和, 可得

$$(K+1)M\left(\min_{k=0,\ldots,K} \|\nabla f(\boldsymbol{x}_k)\|\right)^2 \leq M \sum_{k=0}^{K} \|\nabla f(\boldsymbol{x}_k)\|^2 \leq \Delta_0 - \Delta_{K+1} \leq \Delta_0 = f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*).$$

证明

由引理 2.2 有

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq M\|\nabla f(\boldsymbol{x}_k)\|^2$$

记 $\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$, 则

$$M\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \Delta_k - \Delta_{k+1}, \quad k = 0, 1, \ldots, K$$

对此 $K+1$ 项求和, 可得

$$(K+1)M\left(\min_{k=0,\ldots,K}\|\nabla f(\boldsymbol{x}_k)\|\right)^2 \leq M\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \Delta_0 - \Delta_{K+1} \leq \Delta_0 = f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*).$$

整理得到

$$\min_{k=0,\ldots,K}\|\nabla f(\boldsymbol{x}_k)\| \leq \sqrt{\frac{f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)}{M(K+1)}}.$$

凸函数收敛性

假设目标函数 f 为可微凸函数, 则有

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\leq\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle$$

凸函数收敛性

假设目标函数 f 为可微凸函数, 则有

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\le\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle$$

关于不等式右边内积项, 有

$$\begin{aligned}\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle&=\frac{1}{2h}\left(h^2\|\nabla f(\boldsymbol{x}_k)\|^2+\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_k-h\nabla f(\boldsymbol{x}_k)-\boldsymbol{x}^*\|^2\right)\\&=\frac{1}{2h}\left(h^2\|\nabla f(\boldsymbol{x}_k)\|^2+\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2\right)\end{aligned}$$

即

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\le\frac{1}{2h}\left(h^2\|\nabla f(\boldsymbol{x}_k)\|^2+\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2\right).$$

取 $k=0,1,\ldots,K$, 求 $K+1$ 项和.

$$\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right) \le \frac{h}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{1}{2h}\left(\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{K+1}-\boldsymbol{x}^*\|^2\right)$$
$$\le \frac{h}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{1}{2h}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2$$

取 $k=0,1,\ldots,K$, 求 $K+1$ 项和.

$$\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right) \le \frac{h}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{1}{2h}\left(\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{K+1}-\boldsymbol{x}^*\|^2\right)$$
$$\le \frac{h}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{1}{2h}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2$$

若有 $\|\nabla f(\boldsymbol{x}_k)\|\le B$, $\|\boldsymbol{x}_0-\boldsymbol{x}^*\|\le R$.

取 $k=0,1,\ldots,K$, 求 $K+1$ 项和.

$$\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right)\le\frac{h}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{1}{2h}\left(\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{K+1}-\boldsymbol{x}^*\|^2\right)$$
$$\le\frac{h}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{1}{2h}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2$$

若有 $\|\nabla f(\boldsymbol{x}_k)\|\le B$, $\|\boldsymbol{x}_0-\boldsymbol{x}^*\|\le R$.

$$\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right)\le\frac{h(K+1)B^2}{2}+\frac{R^2}{2h}$$

不等式两端同时除 $K+1$

$$\frac{1}{K+1}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right)\le\frac{hB^2}{2}+\frac{R^2}{2h(K+1)}$$

$$\frac{1}{K+1}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right) \le \frac{hB^2}{2}+\frac{R^2}{2h(K+1)}$$

不等式右端关于步长 h 最小化，此时取 $h=\frac{R}{B\sqrt{K+1}}$，

$$\frac{1}{K+1}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right) \le \frac{RB}{\sqrt{K+1}}$$

综合上述讨论，可得到以下定理.

定理 4

设目标函数 f 为可微凸函数且 $\|\nabla f(\boldsymbol{x})\| \leq B$，记 $\boldsymbol{x}^*$ 为 f 的全局最小值点，若梯度下降法初始点 $\boldsymbol{x}_0$ 满足 $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq R$，取步长

$$h = \frac{R}{B\sqrt{K+1}}, K \in \mathbb{N},$$

则有

$$\frac{1}{K+1}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)\right) \leq \frac{RB}{\sqrt{K+1}}.$$

综合上述讨论，可得到以下定理.

定理 4

设目标函数 f 为可微凸函数且 $\|\nabla f(\boldsymbol{x})\| \le B$，记 $\boldsymbol{x}^*$ 为 f 的全局最小值点，若梯度下降法初始点 $\boldsymbol{x}_0$ 满足 $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \le R$，取步长

$$h = \frac{R}{B\sqrt{K+1}}, K \in \mathbb{N},$$

则有

$$\frac{1}{K+1}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)\right) \le \frac{RB}{\sqrt{K+1}}.$$

上述定理表明若迭代次数 $K+1 \ge \frac{R^2B^2}{\epsilon^2}$，则迭代过程产生的平均误差

$$\frac{1}{K+1}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)\right) \le \epsilon.$$

L-smooth 凸函数收敛性

定理 5

设目标函数 f 为 L-smooth 凸函数，记 $\boldsymbol{x}^*$ 为 f 的全局最小值点，梯度下降法取步长 $h = \frac{1}{L}$，则有

$$f(\boldsymbol{x}_{K+1}) - f(\boldsymbol{x}^*) \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2(K+1)}, \quad K \in \mathbb{N}.$$

证明

回忆关于凸函数有

$$f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\leq\frac{1}{2h}\left(h^2\|\nabla f(\boldsymbol{x}_k)\|^2+\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2\right).$$

其中 h 为步长.

证明

回忆关于凸函数有

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{1}{2h}\left(h^2 \|\nabla f(\boldsymbol{x}_k)\|^2 + \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right).$$

其中 h 为步长.
取 $h = \frac{1}{L}$, 上式可以表示为

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right).$$

由充分下降引理

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2$$

得到控制单步梯度的上界

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})$$

由充分下降引理

$$f(\boldsymbol{x}_{k+1}) \le f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2$$

得到控制单步梯度的上界

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 \le f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})$$

代入

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \le \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right).$$

化简为

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \le \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right).$$

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \le \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right).$$

上述不等式取 $k = 0, 1, \ldots, K$, 求 $K+1$ 项和.

$$\begin{aligned}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*)\right) &\le \frac{L}{2}\left(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{K+1} - \boldsymbol{x}^*\|^2\right)\\ &\le \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2\end{aligned}$$

$$f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}^*) \le \frac{L}{2}\left(\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2\right).$$

上述不等式取 $k=0,1,\ldots,K$, 求 $K+1$ 项和.

$$\begin{aligned}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}^*)\right) &\le \frac{L}{2}\left(\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{K+1}-\boldsymbol{x}^*\|^2\right)\\ &\le \frac{L}{2}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2\end{aligned}$$

由充分下降引理, 序列 $\{f(\boldsymbol{x}_k)\}$ 单调不增, 故

$$(K+1)\cdot(f(\boldsymbol{x}_{K+1})-f(\boldsymbol{x}^*)) \le \sum_{k=0}^{K}\left(f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}^*)\right) \le \frac{L}{2}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2,$$

整理得到

$$f(\boldsymbol{x}_{K+1})-f(\boldsymbol{x}^*) \le \frac{L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{2(K+1)}$$

L-smooth 及 μ-强凸函数的收敛性

定理 6

设目标函数 f 为 L-smooth 且 μ-强凸函数，记 $\boldsymbol{x}^*$ 为 f 的全局最小值点，梯度下降法取步长 $h=\frac{1}{L}$，则有

$$\|\boldsymbol{x}_K-\boldsymbol{x}^*\|^2 \le \exp(-K\mu/L)\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2, \quad K\in\mathbb{N}_+$$

证明

由 μ 强凸函数的定义

$$\frac{\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2 \le f(\boldsymbol{x}^*)-f(\boldsymbol{x}_k)+\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle$$

同时

$$\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle=\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{L}{2}\left(\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2\right)$$

证明

由 μ 强凸函数的定义

$$\frac{\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2 \le f(\boldsymbol{x}^*)-f(\boldsymbol{x}_k)+\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle$$

同时

$$\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle=\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{L}{2}\left(\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2\right)$$

结合上述两个不等式, 整理得到

$$\frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 \le f(\boldsymbol{x}^*)-f(\boldsymbol{x}_k)+\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{L-\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2$$

$$\frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 \le f(\boldsymbol{x}^*)-f(\boldsymbol{x}_k)+\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{L-\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2$$

由充分下降引理, 得到控制单步梯度的上界

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 \le f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})$$

$$\frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 \leq f(\boldsymbol{x}^*)-f(\boldsymbol{x}_k)+\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{L-\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2$$

由充分下降引理, 得到控制单步梯度的上界

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})$$

故有

$$\begin{aligned}\frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 &\leq f(\boldsymbol{x}^*)-f(\boldsymbol{x}_{k+1})+\frac{L-\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2\\ &\leq \frac{L-\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2\end{aligned}$$

整理得到

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 \leq \left(1-\frac{\mu}{L}\right)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2.$$

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le \left(1 - \frac{\mu}{L}\right) \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2.$$

由此不等式递推得到

$$\|\boldsymbol{x}_K - \boldsymbol{x}^*\|^2 \le \left(1 - \frac{\mu}{L}\right)^K \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 \le \exp(-K\mu/L)\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \quad K \in \mathbb{N}_+$$

总结

- L-smooth 函数

$$\min_{k=0,\ldots,K} \|\nabla f(\boldsymbol{x}_k)\| \leq \sqrt{\frac{f(\boldsymbol{x}_0)-f(\boldsymbol{x}^*)}{M(K+1)}},$$

$K=\mathcal{O}(\frac{1}{\epsilon^2})$

- 凸函数

$$\frac{1}{K+1}\sum_{k=0}^{K}\left(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\right) \leq \frac{RB}{\sqrt{K+1}}.$$

$K=\mathcal{O}(\frac{1}{\epsilon^2})$

- L-smooth 凸函数

$$f(\boldsymbol{x}_{K+1}) - f(\boldsymbol{x}^*) \le \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2(K+1)}, \quad K \in \mathbb{N}.$$

$K = \mathcal{O}(\frac{1}{\epsilon})$

- L-smooth 且 μ-强凸函数

$$\|\boldsymbol{x}_K - \boldsymbol{x}^*\|^2 \le \exp(-K\mu/L)\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \quad K \in \mathbb{N}_+$$

$K = \mathcal{O}(\log\frac{1}{\epsilon})$

收敛阶与复杂度上界

- 称 $\mathcal{O}(\log\frac{1}{\epsilon})$ 为线性收敛阶, 固定每走某常数步, 误差精度提升 1 位有效数字.
- 称 $\mathcal{O}\left((\frac{1}{\epsilon})^r\right)$ 为次线性收敛阶, 为了使得误差精度提升 1 位有效数字, 所需的迭代次数与当前的迭代总次数同阶.

收敛阶与复杂度上界

- 称 $\mathcal{O}(\log\frac{1}{\epsilon})$ 为线性收敛阶，固定每走某常数步，误差精度提升 1 位有效数字.
- 称 $\mathcal{O}\left((\frac{1}{\epsilon})^r\right)$ 为次线性收敛阶，为了使得误差精度提升 1 位有效数字，所需的迭代次数与当前的迭代总次数同阶.

- 复杂度上界回答了迭代次数达到何值能够保证误差被 ϵ 控制 (充分性).
- 但没有回答误差被 ϵ 控制至少需要多少次迭代 (必要性).

两位算法专家的对话

算法专家 A: 我找到了一个算法 $\mathscr{A}$, 在解函数类 $\mathscr{P}$ 时, 至多需要 $f(\epsilon)$ 步就能找到 ϵ-近似解.

算法专家 B: 你的算法 $\mathscr{A}$ 能再进一步提升计算效率吗? 也就是说, 进一步改进算法 $\mathscr{A}$ 使得存在一个函数 g, 满足在忽略与 ϵ 无关项的情形下 $g < f$, 至多需要 $g(\epsilon)$ 次迭代就能找到 ϵ-近似解.

算法专家 A: 不能, 这里有一个属于函数类 $\mathscr{P}$ 的函数 P, 在迭代次数少于 $f(\epsilon)$ 时还没有找到 ϵ-近似解. 换句话说, 用算法 $\mathscr{A}$ 解问题 P 时, 迭代次数至少要 $f(\epsilon)$ 才有可能找到 ϵ-近似解.

算法专家 B: 那我明白了, 你的算法 $\mathscr{A}$ 在求解函数类 $\mathscr{P}$ 时不能表现得比 $f(\epsilon)$ 更好, 因为函数类 $\mathscr{P}$ 中有一个问题 P, 迭代次数至少要达到 $f(\epsilon)$ 才有可能找到 ϵ-近似解.

算法专家 A: 是的, 问题 P 的存在阻碍了算法 $\mathscr{A}$ 在求解函数类 $\mathscr{P}$ 时进一步提升的可能.