Numerical Optimization, 2020 Fall
Homework 7

Due on 14:59 NOV 26, 2020
请尽量使用提供的 tex 模板, 若手写作答请标清题号并拍照加入文档.

1 收敛速率

分别构造具有次线性，线性，超线性和二阶收敛速率的序列的例子。[10 pts]

1. 次线性: $\{a_n\} = 1, \frac{1}{2}, \cdots, \frac{1}{n}, \cdots$
2. 线性: $\{a_n\} = 1, \frac{1}{2^2}, \cdots, \frac{1}{2^n}, \cdots$
3. 超线性: $\{a_n\} = 1, \frac{1}{2^2}, \cdots, \frac{1}{n^n}, \cdots$
4. 二阶: $\{a_n\} = 1, \frac{1}{2^4}, \cdots, \frac{1}{2^{2^n}}, \cdots$

2 梯度下降法的收敛性分析

考虑如下优化问题:

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

其中目标函数 f 满足一下性质:

- 对任意 $\boldsymbol{x}$, $f(\boldsymbol{x}) \geq \underline{f}$。
- ∇f 是 Lipschitz 连续的，即对于任意的 $\boldsymbol{x}, \boldsymbol{y}$，存在 $L > 0$ 使得

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

若采用梯度下降法求解问题(1)，记所产生的迭代点序列为 $\{\boldsymbol{x}^k\}$。迭代点的更新为 $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$。试证明以下问题。

(i) 在一点 $\boldsymbol{x}^k$ 处给定一个下降方向 $\boldsymbol{d}^k$，即 $\boldsymbol{d}^k$ 满足 $\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \rangle < 0$。试证明：对于充分小的 $\alpha > 0$，有 $f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) < f(\boldsymbol{x}^k)$ 成立。[10 pts]

(ii) 假设存在 $\delta > 0$ 使得 $-\frac{\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \rangle}{\|\nabla f(\boldsymbol{x}^k)\|_2 \|\boldsymbol{d}^k\|_2} > \delta$。证明回溯线搜索会有限步终止，并给出对应步长 α^k 的下界。[10 pts]

(iii) 根据上一问结果证明 $\lim_{k\to\infty}\|\nabla f(\boldsymbol{x}^k)\|_2 = \mathbf{0}$。[10 pts]

(iv) 令 $\boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k)$，采用固定步长 $\alpha^k \equiv \alpha = \frac{1}{L}$。试证明该设定下梯度下降法的全局收敛性。[20 pts]

(i) Applying the first order Taylor expansion of $f(\boldsymbol{x}+\alpha\boldsymbol{d}^k)$ at the point $\boldsymbol{x}^k$,

$$\begin{aligned} f(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k) =& f(\boldsymbol{x}^k) + \left\langle \nabla f(\boldsymbol{x}^k), \alpha\boldsymbol{d}^k \right\rangle + \frac{\alpha^2}{2}(\boldsymbol{d}^k)^T\nabla^2 f(\boldsymbol{x}^k+\theta\alpha\boldsymbol{d}^k)\boldsymbol{d}^k \\ \leq& f(\boldsymbol{x}^k) + \left\langle \nabla f(\boldsymbol{x}^k), \alpha\boldsymbol{d}^k \right\rangle + \frac{\alpha^2}{2}L\|\boldsymbol{d}^k\|_2^2, \end{aligned} \tag{2}$$

where $\theta \in (0,1)$, and the inequality follows by the Lipschitz continuity of ∇f. For a sufficiently small α, the term $\frac{\alpha^2}{2}(\boldsymbol{d}^k)^T\nabla^2 f(\boldsymbol{x}^k+\theta\alpha\boldsymbol{d}^k)\boldsymbol{d}^k$ can be ignored. Since $\left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \right\rangle < 0$ and $\alpha > 0$, it holds

$$f(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k) = f(\boldsymbol{x}^k) + \left\langle \nabla f(\boldsymbol{x}^k), \alpha\boldsymbol{d}^k \right\rangle < f(\boldsymbol{x}^k),$$

which completes the proof.

(ii) According to (i), for an arbitrary α, we achieve an upper bound of $f(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k)$ as shown in (2). By making α further satisfy the sufficient decrease condition, we have

$$\begin{aligned} f(\boldsymbol{x}^k+\alpha\boldsymbol{d}^k) \leq& f(\boldsymbol{x}^k) + \left\langle \nabla f(\boldsymbol{x}^k), \alpha\boldsymbol{d}^k \right\rangle + \frac{\alpha^2}{2}L\|\boldsymbol{d}^k\|_2^2 \\ \leq& f(\boldsymbol{x}^k) + c_1\left\langle \nabla f(\boldsymbol{x}^k), \alpha\boldsymbol{d}^k \right\rangle. \end{aligned}$$

Therefore, for any $\alpha \in [0, \frac{2(c_1-1)\left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \right\rangle}{L\|\boldsymbol{d}^k\|_2^2}]$, the sufficient decrease condition is satisfied. And then, the backtracking procedure must end up with

$$\alpha^k \geq 2\gamma\frac{(c_1-1)\left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \right\rangle}{L\|\boldsymbol{d}^k\|_2^2}, \tag{3}$$

where γ is the decay constant of the line search.

(iii) Combining the sufficient decrease condition with (3),

$$\begin{aligned} f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k+1}) =& f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k+\alpha^k\boldsymbol{d}^k) \\ \geq& -c_1\alpha^k\left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \right\rangle \\ \geq& c_1\frac{2\gamma(1-c_1)\left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \right\rangle}{L\|\boldsymbol{d}^k\|_2^2}\left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \right\rangle \\ >& \frac{2\gamma c_1(1-c_1)\delta^2\|\nabla f(\boldsymbol{x}^k)\|_2^2}{L}, \end{aligned} \tag{4}$$

where the last inequality is by $-\frac{\left\langle \nabla f(\boldsymbol{x}^k), \boldsymbol{d}^k \right\rangle}{\|\nabla f(\boldsymbol{x}^k)\|_2\|\boldsymbol{d}^k\|_2} > \delta$.

By rearranging (4) and summing up both sides from 1 to K,

$$\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}^k)\|_2^2 < \frac{L}{c_1(1-c_1)\delta^2}\sum_{k=0}^{K}\left(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k+1})\right) \leq \frac{L}{2\gamma c_1(1-c_1)\delta^2}\left(f(\boldsymbol{x}^0) - \underline{f}\right),$$

completing the proof.

(iv) Substituting $\alpha^k \equiv \alpha = \frac{1}{L}$ and $\boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k)$ in (2), it holds

$$\begin{aligned} f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k) \leq& \left\langle \nabla f(\boldsymbol{x}^k), \alpha\boldsymbol{d}^k \right\rangle + \frac{\alpha^2}{2}L\|\boldsymbol{d}^k\|_2^2 \\ =& -\frac{1}{2L}\|\nabla f(\boldsymbol{x}^k)\|_2^2. \end{aligned} \tag{5}$$

Summing up both sides of (5) from 1 to K,

$$\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}^k)\|_2^2 \le 2L\left(f(\boldsymbol{x}^k)-f(\boldsymbol{x}^{k+1})\right) \le 2L\left(f(\boldsymbol{x}^0)-\underline{f}\right).$$

Thus, as $k\to\infty$, $\|\nabla f(\boldsymbol{x}^k)\|_2^2 \to 0$.

3 编程题

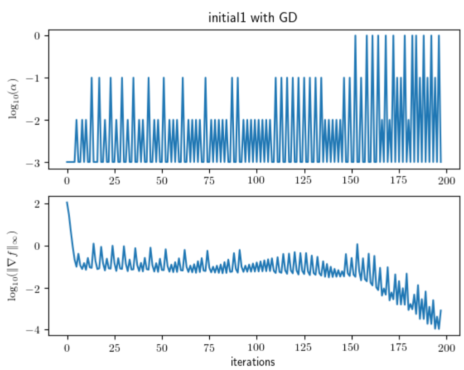
考虑求解如下优化问题:

$$\min_{x_1,x_2} \quad 100(x_2-x_1^2)^2+(1-x_1)^2. \tag{6}$$

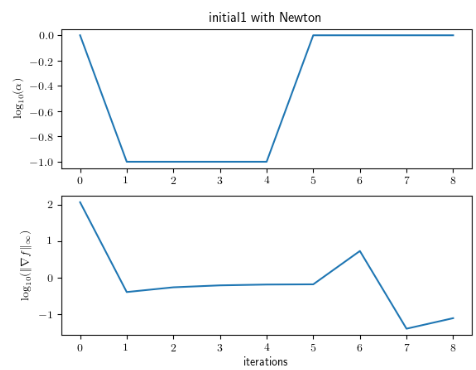
分别用**梯度下降法**和**牛顿法**结合 Armijo 回溯搜索编程求解该问题。分别考虑用 $\boldsymbol{x}^0=[1.2,1.2]^T$ 和 $\boldsymbol{x}^0=[-1.2,1]^T$(较困难) 作为初始点启动算法。

要求: 对于两种初始点，分别画出两种算法步长 α^k 和 $\|\nabla f(\boldsymbol{x}^k)\|_\infty$ 随迭代步数 k 变化的曲线。(编程可使用 matlab 或 python 完成，请将代码截图贴在该文档中。) [40pts]

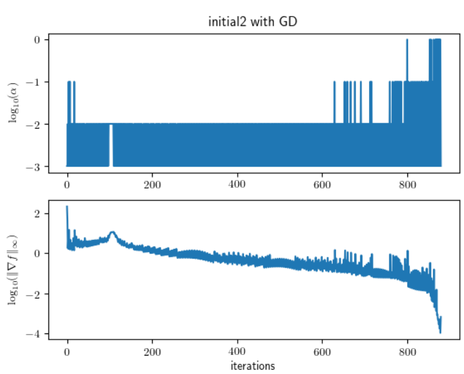
(Hint: 步长初始值 $\alpha_0=1$, 参数 c_1 可选为 10^{-4}，终止条件为 $\|\nabla f(\boldsymbol{x}^k)\|_\infty \le 10^{-4}$.)



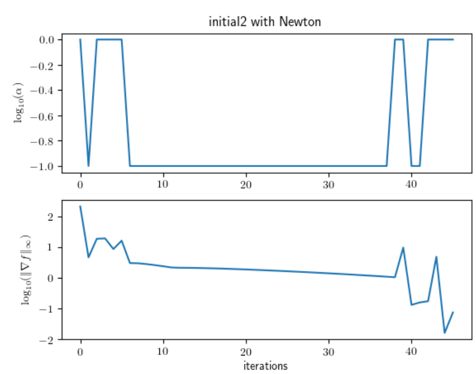
(a) 梯度下降法 & 初始点 $\boldsymbol{x}^0=[1.2,1.2]^T$



(b) 牛顿法 & 初始点 $\boldsymbol{x}^0=[1.2,1.2]^T$



(c) 梯度下降法 & 初始点 $\boldsymbol{x}^0=[-1.2,1]^T$



(d) 牛顿法 & 初始点 $\boldsymbol{x}^0=[-1.2,1]^T$