第 5 章 1-3 节 梯度下降法

SMaLL

[1] 中国石油大学（华东）
SMaLL 课题组
small.sem.upc.edu.cn
liangxijunsd@163.com

2023

第 5 章 1-3 节 梯度下降法

1. 全局优化算法的复杂度

2. 优化算法构造思想

3. 梯度下降法

全局优化的复杂性边界 I

问题和假定.

考虑以下问题:

$$\min_{x \in B_n} f(x) \tag{1}$$

基本可行集:$B_n \subseteq R^n$:

$$B_n = \{x \in R^n \,|\, 0 \le x^{(i)} \le 1,\ \ i = 1, \ldots, n\}.$$

假定.

目标函数 $f(x)$ 在 B_n 上是 Lipschitz 连续的:

$$|f(x) - f(y)| \le L\|x - y\|_\infty, \quad \forall x, y \in B_n, \tag{2}$$

$L > 0$: Lipschitz 常数

全局优化的复杂性边界 II

均匀网格法

考虑简单的解决问题 (1) 的方法: 均匀网格法

方法 $\mathcal{G}(p)$

1. 构造 $(p+1)^n$ 个点
$$x_{(i_1,\ldots,i_n)} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \ldots, \frac{i_n}{p}\right)^T,$$
其中 $(i_1,\ldots,i_n) \in \{0,\ldots,p\}^n$.
2. 在所有点 $x_{(i_1,\ldots,i_n)}$ 中，找到点 $\bar{x}$, 使得目标函数值最小.
3. 返回结果: $(\bar{x}, f(\bar{x}))$.

全局优化的复杂性边界 III

注.

- 该方法在 B_n 内形成测试点的均匀网格, 计算该网格点上目标函数的最小值 $\to$ 返回该值作为问题 (1) 的近似解.
- 这是 一种零阶迭代方法, 不受累积信息对测试点序列的任何影响.

复杂度上限 I

定理 1

(复杂度上限) 设 f^* 是问题 (1) 的全局最优值. 那么

$$f(\bar{x}) - f^* \le \frac{L}{2p}.$$

复杂度上限 II

证明. 令 x^* 为问题 (1) 的全局最小值. 存在候选集 $(i_1, i_2, \ldots, i_n)$:

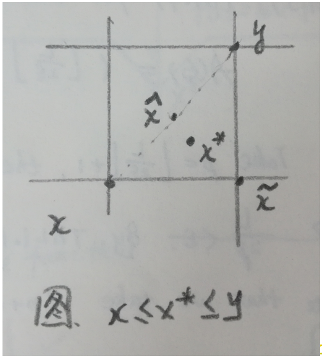
$$x \equiv x_{(i_1,i_2,\ldots,i_n)} \leq x^* \leq x_{(i_1+1,i_2+1,\ldots,i_n+1)} \equiv y$$

$x \leq y \quad \Leftrightarrow x^{(i)} \leq y^{(i)}, \ \forall i = 1, 2, \ldots, n.$

注:

$$y^{(i)} - x^{(i)} = \frac{1}{p}, \qquad \text{对于} i = 1, \ldots, n.$$

$$x^{*(i)} \in [x^{(i)}, y^{(i)}], i = 1, \ldots, n$$



复杂度上限 III

记:

$$\hat{x} = (x + y)/2.$$

考虑网络点:

$$\tilde{x}^{(i)} = \begin{cases} y^{(i)}, & \text{if } x^{*(i)} \geq \hat{x}^{(i)} \\ x^{(i)}, & \text{otherwise} \end{cases}$$

显然:

$$|\tilde{x}^{(i)} - x^{*(i)}| \leq \frac{1}{2p}.$$

$\Rightarrow \|\tilde{x} - x^*\|_\infty = \max_i |\tilde{x}^{(i)} - x^{*(i)}| \leq \frac{1}{2p}$. $\tilde{x}$ 属于网格之中 $\Rightarrow$

$$f(\bar{x}) - f(x^*) \leq f(\tilde{x}) - f(x^*) \leq L\|\tilde{x} - x^*\|_\infty \leq \frac{L}{2p}.$$

$\square$

复杂度上限 IV

分析复杂度.

定义目标如下:

$$\text{找到 } \bar{x} \in B_n : \ f(\bar{x}) - f^* \le \epsilon. \tag{3}$$

推论 2

方法 $\mathcal{G}$ 的问题 (1), (2) 和 (3) 的分析复杂度：最多为

$$\mathcal{A}(\mathcal{G}) = \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2 \right)^n$$

证明. 令 $p = \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 1\right)$, 则 $p > \frac{L}{2\epsilon}$. 由定理 1知, $f(\bar{x}) - f^* \le \frac{L}{2p} < \epsilon$.

注. 我们构造了 $(p+1)^n$ 个点. □

复杂度上限 V

注.

- $\mathcal{A}(\mathcal{G})$ 给出了问题类的复杂度上限
- 问题 1. 我们的证明可能过于粗略，$\mathcal{G}(p)$ 的实际性能要好得多.
- 问题 2. 我们仍然不能确定 $\mathcal{G}(p)$ 是求解问题 (1) 的合理方法. 可能存在其他性能高得多的方案.

复杂度下限 I

需要给出问题 (1), (2) 和 (3) 的复杂度下界. 主要特征:

- 基于黑盒概念
- 这些边界对于所有合理的迭代方案都是有效的 $\rightarrow$ 提供了 问题类的分析复杂度的下限
- 这种下界通常基于对抗 oracle的思想

SMaLL Statistical Machine Learning Lyceum

复杂度下限 II

对抗 oracle 的概念

- 对抗 oracle: 为每种具体方法创建一个最坏的问题
- 它从一个“空”函数开始，“以最坏的方式”回应算法的每次提问
- 答案必须 与之前的答案和问题类别的描述相一致

复杂度下限 III

考虑对抗 Oracle 是如何解决这个问题 (1) 的. 考虑问题 $\mathcal{C}$:

模型	$\min_{x \in B_n} \ f(x)$ $f(x)$ 是在 B_n 上 l_∞-Lipschitz 连续的.
Oracle	零阶局部黑盒.
近似解	找到 $\bar{x} \in B_n : f(\bar{x}) - f^* \le \epsilon$.

复杂度下限 IV

定理 3

(复杂度下限) 对于 $\epsilon < \frac{L}{2}$, 对于零阶方法, $\mathcal{C}$ 的分析复杂度至少为 $\left(\lfloor \frac{L}{2\epsilon} \rfloor\right)^n$.

证明.

记 $p = \left(\lfloor \frac{L}{2\epsilon} \rfloor\right) \geq 1$. 假设存在一种方法需要 $N < p^n$ 次调用 oracle 来解决问题类 $\mathcal{C}$. 将算法应用于如下的对抗策略中:

$$\boxed{\text{在每一个测试点 } x \text{ 处，Oracle 返回 } f(x) = 0 \text{ .}}$$

这个方法找到的解 $\bar{x} \in B_n$ 满足 $f(\bar{x}) = 0$. 然而, 我们注意到存在 $\hat{x} \in B_n$, 满足

$$\hat{x} + \frac{1}{p}\mathbf{1} \in B_n, \quad \mathbf{1} = (1, 1, \ldots, 1)^T \in R^n$$

并且集合中没有测试点:

$$B = \{x \,|\, \hat{x} \leq x \leq \hat{x} + \frac{1}{p}\mathbf{1}\}$$

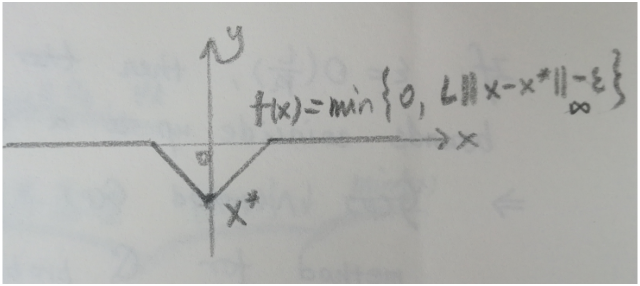
复杂度下限 V

记

$$x^* = \hat{x} + \frac{1}{2p}\mathbf{1}.$$

考虑函数:

$$f(x) = \min\{0, L\|x - x^*\|_\infty - \epsilon\}.$$



复杂度下限 VI

显然, $f(x)$ 是 L_∞-Lipschitz 连续函数，常数为 L, 其全局最优值为 $-\varepsilon$.

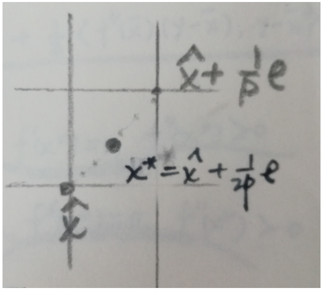
$\forall x_1, x_2 \in B$, $|f(x_1) - f(x_2)| \le |L\|x_1 - x^*\|_\infty - L\|x_2 - x^*\|_\infty| \le L\|x_1 - x_2\|_\infty$.

此外, $f(x)$ 只有在集合中不为 0

$$B' = \{x | \, \|x - x^*\|_\infty \le \frac{\epsilon}{L}\}$$

由 $2p \le \frac{L}{\epsilon}$ 得:

$$B' \subseteq B = \{x | \, \|x - x^*\|_\infty \le \frac{1}{2p}\}.$$



因此,

复杂度下限 VII

- $f(x)$ 在所有测试点处都等于 0.
- 这个方法所得结果的精度为 ϵ.

故有如下结论: 如果调用 oracle 的次数小于 p^n, 结果的精确度不会优于 ϵ.

$\square$

均匀网格法复杂度

$\mathcal{G}(p)$ **是问题类** $\mathcal{C}$ 的一个 最优方法

现在我们可以对均匀网格方法的性能做更多的介绍。将均匀网格法复杂度的上界与下界进行比较:

$$\begin{array}{ll} \text{上界} & \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n \\ \text{下界} & \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor\right)^n \end{array}$$

因此, 如果 $\epsilon = O(\frac{L}{n})$, 则下界和上界相同, 仅差一个常值乘数因子. 这表明均匀网格法 $\mathcal{G}(p)$ 是问题类 $\mathcal{C}$ 的一个最优方法.

一般的优化问题是不可解的

定理 1.1.2 支持我们最初的说法，即一般优化问题是不可解决的.

例 4

例题 1.1.4 考虑由以下参数定义的问题类 $\mathcal{F}$:

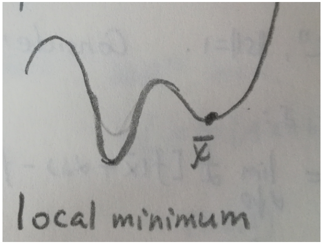
$$L = 2; \quad n = 10; \quad \epsilon = 0.01$$

此问题类的复杂度下限是 $\left(\frac{L}{2\epsilon}\right)^n$.

复杂度下界:	10^{20} 次 oracle 调用
单次 Oracle 调用的计算复杂度:	至少 n 次算术运算
总复杂度:	10^{21} 算术运算
计算机运算速度:	10^6 算术运算/秒
总时间:	10^{15} 秒
一年:	不到 3.2×10^7 秒.
我们需要:	32 000 000 年.

松弛和近似 I

广义非线性规划的**主要任务**: 找到可微函数的局部极小值.



大多数非线性规划方法都是基于松弛的思想:

我们称序列 $\{a_k\}_{k=0}^{\infty}$ 为一个 松弛序列如果 $a_{k+1} \leq a_k, \quad \forall\, k \geq 0.$

松弛和近似 II

考虑无约束最小化问题

$$\min_{x\in R^n} \ f(x), \tag{4}$$

其中 f 是一个光滑的函数.

我们生成一个松弛序列 $\{f(x_k)\}_{k=0}^{\infty}$:

$$f(x_{k+1}) \leq f(x_k), \quad k = 0, 1, \ldots$$

这个方法有以下 重要优势:

1. 如果 $f(x)$ 在 R_n 中有下界, 则序列 $\{f(x_k)\}_{k=0}^{\infty}$ 收敛.
2. 在任何情况下，目标函数的初始值总能得到改进.

松弛和近似 III

松弛思想要付诸实施，离不开另一个优化的基本原则：近似.

近似 意味着用一个简化的函数代替复杂的函数，足够接近原始函数.

第 5 章 1-3 节 梯度下降法

1. 全局优化算法的复杂度

2. 优化算法构造思想

3. 梯度下降法

梯度的两个重要性质. I

用 $\mathcal{L}_f(\alpha)$ 表示$f(x)$ 的水平集:

$$\mathcal{L}_f(\alpha) = \{x \in R^n \mid f(x) \le \alpha\}.$$

考虑在 $\bar{x}$ 处与 $\mathcal{L}_f(f(\bar{x}))$ 相切的方向集:

$$S_f(\bar{x}) = \left\{ s \in R^n \mid s = \lim_{y_k \to \bar{x}, f(y_k) = f(\bar{x})} \frac{y_k - \bar{x}}{\|y_k - \bar{x}\|} \right\}.$$

性质 1. 梯度矢量垂直于水平集的“切线方向”.

引理 5

如果 $s \in S_f(\bar{x})$, 那么 $\langle f'(\bar{x}), s \rangle = 0$.

证明. 由于

$$f(y_k) = f(\bar{x}) + \langle f'(\bar{x}), y_k - \bar{x} \rangle + o(\|y_k - \bar{x}\|) = f(\bar{x}).$$

因此, $\langle f'(\bar{x}), y_k - \bar{x} \rangle + o(\|y_k - \bar{x}\|) = 0$. 将该方程除以 $\|y_k - \bar{x}\|$ 并取 $y_k \to \bar{x}$ 中的极限, 得到结果. $\square$

梯度的两个重要性质. II

性质 2. 负梯度 $-f'(\bar{x})$ 是 $f(x)$ 在点 $\bar{x}$ 处局部下降最快的方向.

证明. 设 s 为 R^n 中的一个方向, $\|s\| = 1$. 考虑 $f(x)$ 沿 s 的局部递减:

$$\Delta(s) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha}[f(\bar{x} + \alpha s) - f(\bar{x})].$$

注意 $f(\bar{x} + \alpha s) - f(\bar{x}) = \alpha \langle f'(\bar{x}), s \rangle + o(\alpha)$. 因此

$$\Delta(s) = \langle f'(\bar{x}), s \rangle.$$

利用 Cauchy-Schwartz 不等式:

$$-\|x\| \cdot \|y\| \leq \langle x, y \rangle \leq \|x\| \cdot \|y\|,$$

得到: $\Delta(s) = \langle f'(\bar{x}), s \rangle \geq -\|f'(\bar{x})\|$. 令 $\bar{s} = -f'(\bar{x})/\|f'(\bar{x})\|$. $\Rightarrow$
$\Delta(\bar{s}) = -\langle f'(\bar{x}), f'(\bar{x}) \rangle / \|f'(\bar{x})\| = -\|f'(\bar{x}\|$. $\Rightarrow$ 方向 $-f'(\bar{x})$ (负梯度) 是 $f(x)$ 在点 $\bar{x}$ 处局部下降最快的方向. $\square$

梯度下降法 I

我们已经知道负梯度方向是可微函数局部下降最快的方向. 由于我们要找到这种函数的 局部极小值 , 下面是尝试的第一个方法:

梯度法
选择 $x_0 \in R^n$.
迭代 $x_{k+1} = x_k - h_k f'(x_k),\ k = 0, 1, \ldots.$

$h_k > 0$: 步长.

重要的步长策略. I

1. 序列 $\{h_k\}_{k=0}^{\infty}$ 是 提前选定的.例如,

$$h_k = h > 0, \qquad \text{(恒定步长)}$$

$$h_k = \frac{h}{\sqrt{k+1}}.$$

2. 完全松弛.

$$h_k = \quad \mathrm{argmin}_{h \geq 0} f(x_k - hf'(x_k)).$$

3. Goldstein-Armijo 准则: 寻找 $x_{k+1} = x_k - hf'(x_k)$ 满足

$$f(x_k) - f(x_{k+1}) \geq \alpha \langle f'(x_k), x_k - x_{k+1} \rangle \tag{5}$$

$$f(x_k) - f(x_{k+1}) \leq \beta \langle f'(x_k), x_k - x_{k+1} \rangle \tag{6}$$

其中 $0 < \alpha < \beta < 1$ 是固定参数.

重要的步长策略. II

注意

- 第一种策略 (i.e, 提前选定) 是最简单的. 它是实际应用中，特别是凸优化情况下最常见的方法.
- 第二种策略 (i.e., 完全松弛) 是理论上的方法.
- 第三种策略 (i.e., Goldstein-Armijo 准则) 在许多非线性规划算法中广泛使用.

重要的步长策略. III

Goldstein-Armijo 准则的几何解释

固定 $x \in R^n$. 考虑关于变量的一元函数

$$\phi(h) = f(x - hf'(x)), \quad h \geq 0.$$

则该策略可接受的步长值位于两个线性函数之间:

$$\phi_1(h) = f(x) - \alpha h \|f'(x)\|^2,$$

$$\phi_2(h) = f(x) - \beta h \|f'(x)\|^2.$$

其中 $\phi(0) = \phi_1(0) = \phi_2(0)$ 且 $\phi'(0) < \phi_2'(0) < \phi_1'(0) < 0$.

因此, 只要 $\phi(h)$ 的下界存在，可接受的值一定存在.

重要的步长策略. IV

- 如果我们表示为:
$$\phi(h) = f(x^k + hd^k) - f(x^k),$$
$$\phi_1(h) = \beta \nabla f(x^k)^T d^k \cdot h$$
$$\phi_2(h) = \alpha \nabla f(x^k)^T d^k \cdot h,$$
那么 Goldstein-Armijo 准则可以重新定义为
$$\phi_1(h) \le \phi(h) \le \phi_2(h).$$
- 不等式 (6) $f(x_k) - f(x_{k+1}) \le \beta \langle f'(x_k), x_k - x_{k+1} \rangle$ 意味着步长 h 应该有一个下界.

几何递减的步长不合理

- 几何递减步长 (提前选定的)

$$h_k = h \cdot \omega^k, \qquad \omega \in (0,1).$$

  例如 $h_k = 0.5^k$, $k = 0, 1, \cdots$.
- $h_k = 0.5^k$, $k = 0, 1, \cdots \Rightarrow \sum_{k=0}^{\infty} 0.5^k = 2$
- $\Rightarrow$ 可能的搜索区域有限: $\{x_k\} \subseteq \mathbf{B}(x_0, 2)$
- $\Rightarrow$ 无法到达 x^* 如果 $\|x^* - x_0\| \geq 2$.

评估梯度法的性能 (讲解) I

考虑问题

$$\min_{x \in R^n} f(x),$$

其中$f \in C_L^{1,1}(R^n)$. 假定 $f(x)$ 在 R^n 上有下界.

评估一个梯度步长的目标函数的下降量

考虑 $y = x - hf'(x)$. 则有

$$\begin{aligned} f(y) &\leq f(x) + \langle f'(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \\ &= f(x) - h\|f'(x)\| + \frac{h^2}{2}L\|f'(x)\|^2 \\ &= f(x) - h(1 - \frac{h}{2}L)\|f'(x)\|^2. \end{aligned} \tag{7}$$

评估梯度法的性能 (讲解) II

为了获得目标函数下降量的最佳估计:

$$\Delta(h) = -h(1 - \frac{h}{2}L) \longrightarrow \min_h .$$

计算此函数的导数 $\longrightarrow$ 最佳步长必须满足以下方程:

$$\Delta'(h) = hL - 1 = 0.$$

因此, 在 $\Delta''(h) = L > 0$ 的情况下，可以得到 $\Delta(h)$ 的最优解为 $h^* = \frac{1}{L}$.
因此, 我们证明了梯度下降法一次迭代后目标函数值的下降量:

$$f(y) \leq f(x) - \frac{1}{2L}\|f'(x)\|^2.$$

评估梯度法的性能 (讲解) III

注.

- 迭代格式: $y = x - hf'(x)$.
- 对 $f \in C_L^{1,1} R^n$, 一次迭代后的目标函数下降量 至少为 $h(1 - \frac{h}{2}L)\|f'(x)\|^2$
- 令 $h = 1/L \to \to$ 一次迭代目标函数值下降量的最佳估计: $\frac{1}{2L}\|f'(x)\|^2$

评估梯度法的性能 (讲解) IV

考虑具体的步长策略:

设 $x_{k+1} = x_k - h_k f'(x_k)$

1. 对于恒定步长策略 $h_k \equiv h$:
$$f(x_k) - f(x_{k+1}) \geq h_k(1 - \frac{h_k}{2}L)\|f'(x_k)\|^2.$$
最优步长选择: $h_k = 1/L$.

2. 对完全松弛策略，有
$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|f'(x_k)\|^2$$
上面的单步下降量并不比 $h_k = \frac{1}{L}$ 的下降量更大

评估梯度法的性能 (讲解) V

3. 基于 Goldstein-Armijo 准则 (6):

$$f(x_k) - f(x_{k+1}) \le \beta \langle f'(x_k), x_k - x_{k+1} \rangle = \beta h_k \|f'(x_k)\|^2$$

由公式 (7) $f(y) \le f(x) - h(1 - \frac{2}{Lh})\|f'(x)\|^2$:

$$f(x_k) - f(x_{k+1}) \ge h_k(1 - \frac{h_k}{2}L)\|f'(x_k)\|^2.$$

$\Rightarrow h_k \ge \frac{2}{L}(1-\beta)$ 利用公式 (12) $f(x_k) - f(x_{k+1}) \ge \alpha \langle f'(x_k), x_k - x_{k+1} \rangle$:

$$f(x_k) - f(x_{k+1}) \ge \alpha \langle f'(x_k), x_k - x_{k+1} \rangle = \alpha h_k \|f'(x_k)\|^2.$$

与前面的不等式结合 $\rightarrow$

$$f(x_k) - f(x_{k+1}) \ge \frac{2}{L}\alpha(1-\beta)\|f'(x_k)\|^2.$$

$\rightarrow$ 基于各种步长规则 $(w > 0)$:

$$f(x_k) - f(x_{k+1}) \ge \frac{w}{L}\|f'(x_k)\|^2. \tag{8}$$

- $f \in C_L^{1,1}(R^n)$
- 不等式 (6): $f(x_k) - f(x_{k+1}) \le \beta \langle f'(x_k), x_k - x_{k+1} \rangle$

$\Rightarrow$ 步长 h 有下界 $h_k \ge 2L(1-\beta)$.

评估梯度法的性能. I

将不等式 (8):$f(x_k) - f(x_{k+1}) \geq \frac{w}{L}\|f'(x_k)\|^2$. 关于 $k = 0, \dots, N$ 进行累加:

$$\frac{w}{L}\sum_{k=0}^{N}\|f'(x_k)\|^2 \leq f(x_0) - f(x_{N+1}) \leq f(x_0) - f^*, \tag{9}$$

其中 f^* 是问题 (1.2.1) 的最优值 $\Rightarrow$

$$\|f'(x_k)\| \to 0, \qquad \text{当} \quad k \to \infty.$$

收敛速度

记 $g_N^* = \min_{0\leq k\leq N} g_k$, 其中 $g_k = \|f'(x_k)\| \Rightarrow$

$$g_N^* \leq \frac{1}{\sqrt{N+1}}\left[\frac{L}{w}(f(x_0) - f^*)\right]^{1/2}. \tag{10}$$

不等式右边描述了序列 $\{g_N^*\}$ 收敛到 0 的速度. 但是，这并不是序列 $\{f(x_k)\}$ 和 $\{x_k\}$ 的收敛速度.

梯度法只能找到一个稳定点 (讲解) I

在一般的非线性优化中，我们的目标是相当宽松的：我们想要 找到问题的局部极小值. 然而，即使是这个目标梯度方法也是无法实现的.

例 6

考虑二元函数:

$$f(x) = f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2.$$

该函数的梯度为：

$$f'(x) = (x_1, x_2^3 - x_2)^T.$$

因此，该函数的局部极小解之优可能在下述的三个点中取到：

$$(0,0), \quad (0,-1), \quad (0,1).$$

SMaLL Statistical Machine Learning Lyceum

梯度法只能找到一个稳定点 (讲解) II

计算该函数的 Hessian 矩阵

$$f''(x) = \begin{bmatrix} 1 & 0 \\ 0 & 3x_2^2 - 1 \end{bmatrix}$$

- 点 $(0,1)$ 和点 $(0,-1)$ 是孤立的局部极小值但是 $x_1^* = (0,0)$ 只是我们函数的一个驻点.
- 对于充分小的 ϵ，有 $f(x_1^*) = 0$ 和 $f(x_1^* + \epsilon e_2) = \frac{\epsilon^4}{4} - \frac{\epsilon^2}{2} < 0$.
- 考虑梯度下降法自 $x_0 = (1,0)$ 开始的迭代点列的轨迹. 注意这个点的第二个坐标是零。所以 $f'(x_0)$ 的第二个坐标也是零.
- 同样，x_1 的第二个坐标也是零$\cdots\to$ 由梯度下降法生成的迭代序列中的点第二个坐标都是 0 $\to$ 这个序列收敛于 $x_1^* = (0,0)$

复杂度上限 I

- 不等式 (10) 提供了极小化序列的收敛速度.
- 收敛速度提供了问题类的复杂度的上界
- 如果存在一种算法，它的复杂度上界与问题类的复杂度下界成正比，称这种方法为最优方法.

复杂度上限 II

例 7

考虑以下问题类:

模型	1. 无约束最小化. 2. $f \in C_L^{1,1}(R^n)$ 3. $f(x)$ 是下界.
Oracle	一阶黑盒.
ϵ-近似解	$f(\bar{x}) \leq f(x_0)$, $\|f'(\bar{x})\| \leq \epsilon$

不等式 (10) $\Rightarrow$

$$g_N^* \leq \frac{1}{\sqrt{N+1}}\left[\frac{1}{w}L(f(x_0)-f^*)\right]^{1/2} \leq \epsilon.$$

如果 $N+1 \geq \frac{L}{w\epsilon^2}(f(x_0)-f^*) \Rightarrow g_N^* \leq \epsilon.$

$\Rightarrow$ 该问题类复杂度上限 $\frac{L}{w\epsilon^2}(f(x_0)-f^*)$

- 对比定理 1.1.2 的结果 $\rightarrow$ 它的效果更好
- 这个问题类的复杂度下限是未知的.

梯度法的局部收敛性 I

考虑无约束最小化问题

$$\min_{x \in R^n} \ f(x)$$

假设:

1. $f \in C_M^{2,2}(R^n)$.
2. 存在函数 f 的局部极小值，在该极小值处 Hessian 矩阵是正定的.
3. 我们知道：Hessian 矩阵在最优解 x^* 处有界:
$$lI_n \preceq f''(x^*) \preceq LI_n. \tag{11}$$
4. 起始点 x_0 充分靠近 x^*.

梯度法的局部收敛性 II

考虑过程:

$$x_{k+1} = x_k - h_k f'(x_k).$$

注意 $f'(x^*) = 0$. 因此,

$$\begin{aligned} f'(x_k) = f'(x_k) - f'(x^*) &= \int_0^1 f''(x^* + \tau(x_k - x^*))(x_k - x^*)\, d\tau \\ &= G_k(x_k - x^*) \end{aligned}$$

其中 $G_k = \int_0^1 f''(x^* + \tau(x_k - x^*))\, d\tau$. 因此,

$$x_{k+1} - x^* = x_k - x^* - h_k G_k(x_k - x^*) = (I - h_k G_k)(x_k - x^*).$$

梯度法的局部收敛性 III

定理 8

设函数 $f(x)$ 满足我们的假设且初始点 x_0 足够接近局部极小值:

$$r_0 = \|x_0 - x^*\| < \bar{r} = \frac{2l}{M}.$$

那么按式选取步长 $h_k = \frac{2}{L+l}$ 的梯度下降法满足:

$$\|x_k - x^*\| \le \frac{\bar{r} r_0}{\bar{r} - r_0} \left(1 - \frac{2l}{L+3l}\right)^k.$$

这种收敛速度称为线性收敛速度.

梯度法的局部收敛性 IV

- 收敛半径
$$\bar{r} = \frac{2l}{M}$$
与 l 成比例，且与 M 成反比, 其中
  1. l 是 $f''(x^*)$ 的最小特征值;
  2. M 是 $f''(x)$ 的 Lischitz 常数.
- 收敛速度取决于 L 和 l ,$f''(x^*)$ 的最大和最小的特征值. 在极端情况下 $L = l$ (注意 $L \geq l$),
$$\|x_k - x^*\| \leq \frac{\bar{r}r_0}{\bar{r} - r_0}(\frac{1}{2})^k.$$
然而, 当 $L >> l$ 时, 收敛明显较慢. 比如 $L = 1000l$, 则
$$\|x_k - x^*\| \leq \frac{\bar{r}r_0}{\bar{r} - r_0}(\frac{1001}{1003})^k.$$
注意 $(\frac{1001}{1003})^{300} \approx 0.55$, 这意味着与理想情况 $L = l$ 相比，为了达到相同精度，在这种情况下迭代次数将是 300 次.

总结 I

1. **均匀网格法.** 在可行集内形成测试点的统一网格, 并计算该网格上目标函数的最小值.
   - 上界: $\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n$
   - 下界: $\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor\right)^n$
     - 均匀网格方法是一种优化方法，适用于解决: $\min_{x\in B_n} f(x)$, 其中 f 是 Lipschitz 连续的.
     - 一般的优化问题是不可解的.
2. **松弛和近似.** 一般非线性规划方法的构造主要基于两个手段
   - 松弛：如果 $\{a_k\}_{k=0}^{\infty}$ 满足 $a_{k+1} \leq a_k, \quad \forall\, k \geq 0$ 则称其为松弛序列
   - 近似：用一个足够接近原始物体的简化物体来代替最初的复杂物体.

3. **梯度下降法.**

$$x_{k+1} = x_k + h_k \cdot (-f'(x_k)), k = 0, 1, 2, \cdots$$

- 步长策略:
  1. 提前选定: $h_k \equiv h$; $h_k = \frac{h}{\sqrt{k+1}}$;
  2. 完全松弛
  3. Golstein-Armijo 准则
- 梯度法的复杂性
  - 梯度下降法一次迭代后目标函数值的可能的下降量为:

$$f(x_k) - f(x_{k+1}) \geq \frac{w}{L}\|f'(x_k)\|^2.$$

  - 梯度法的复杂度

$$\frac{w}{L}\sum_{k=0}^{N}\|f'(x_k)\|^2 \leq f(x_0) - f(x_{N+1}) \leq f(x_0) - f^*,$$

$$\rightarrow \lim_{k\rightarrow\infty}\|f'(x_k)\| = 0$$

  - 收敛速度

    记 $g_N^* = \min_{0\leq k\leq N}\|f'(x_k)\|$, 则

$$g_N^* \leq \frac{1}{\sqrt{N+1}}[\frac{L}{w}(f(x_0) - f^*)]^{1/2}.$$

- $f \in C_L^{1,1}(R^n)$, f 有下界, 梯度法复杂度上界:

$$\frac{L}{w\epsilon^2}(f(x_0) - f^*)$$

- 局部收敛: 线性收敛速度

$$\min_{x \in R^n} f(x)$$

在给定假设下, 按 $h_k^* = \frac{2}{L+l}$ 选取步长的梯度下降法收敛如下:

$$\|x_k - x^*\| \leq \frac{\bar{r} r_0}{\bar{r} - r_0}\left(1 - \frac{2l}{L+3l}\right)^k.$$

作业题

1. 编程题：第 1 部分第 1 节中，有练习题：自行选取二分类数据集，编程实现梯度下降法求解 Logistic 回归模型，并对数据集进行分类，计算分类准确率。请在此基础上，使用 Armijo 准则确定每一次迭代的步长：

   Armijo 准则: 寻找 $x_{k+1} = x_k - hf'(x_k)$ 满足

   $$f(x_k) - f(x_{k+1}) \geq \alpha \langle f'(x_k), x_k - x_{k+1} \rangle \tag{12}$$

   其中 $0 < \alpha < 1$ 是固定参数.

SMaLL Statistical Machine Learning Lyceum