矩阵论

2024年秋学期

第九讲

2024年10月12日

第4章 梯度分析与最优化

无约束最小化问题的梯度分析—实值目标函数的最速下降方向

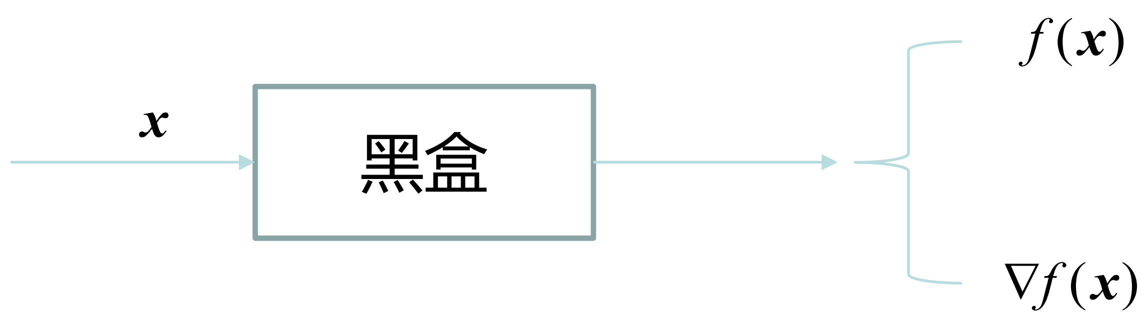
以复向量或者矩阵为变元的实值目标函数的平稳点存在两种选择

$$\left.\frac{\partial f(Z,Z^*)}{\partial Z}\right|_{Z=C}=O_{m\times n} \quad 或 \quad \left.\frac{\partial f(Z,Z^*)}{\partial Z^*}\right|_{Z=C}=O_{m\times n}$$

在设计优化迭代算法时，应该选哪一种梯度?

平滑凸优化的一阶算法——梯度法/最陡下降方法

$$\min_{x \in \mathcal{D}} f(\boldsymbol{x})$$



令 x_{opt} 表示 $\min f(x)$ 的最优解, 一阶黑盒优化 (first-order black-box optimization)就是只利用 $f(\boldsymbol{x})$ 和 $\nabla f(\boldsymbol{x})$ 求解向量 $\boldsymbol{y} \in Q$ 满足 $y: \ f(\boldsymbol{y}) - f\left(\boldsymbol{x}_{\text{opt}}\right) \leqslant \varepsilon$ → 给定的精度误差

$$\arg\min f(\boldsymbol{x})$$

Argument，函数取最小值时 $\boldsymbol{x}$ 的取值

平滑凸优化的一阶算法——梯度法/最陡下降方法

下降法是一种最简单的一阶优化算法

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mu_k \Delta \boldsymbol{x}_k, \quad k = 1, 2, \cdots$$

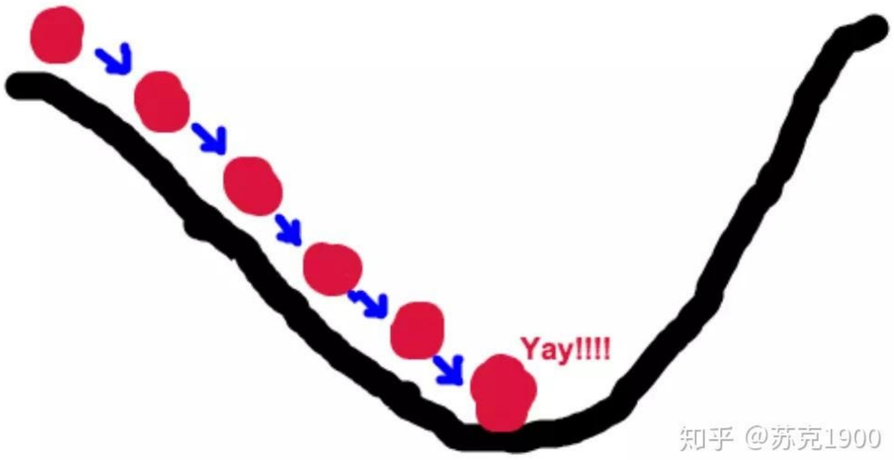
x_{opt}

搜索方向或者更新方向

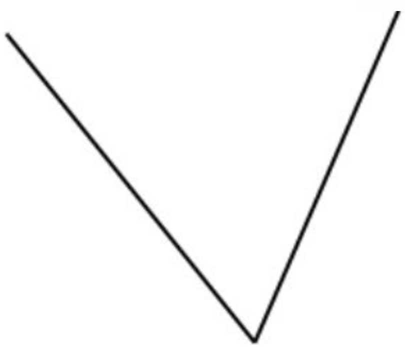
步长(学习率)
用于控制更新x寻优的步
伐

$$f\left(\boldsymbol{x}_{k+1}\right) < f\left(\boldsymbol{x}_k\right)$$ 迭代过程中目标函数下降

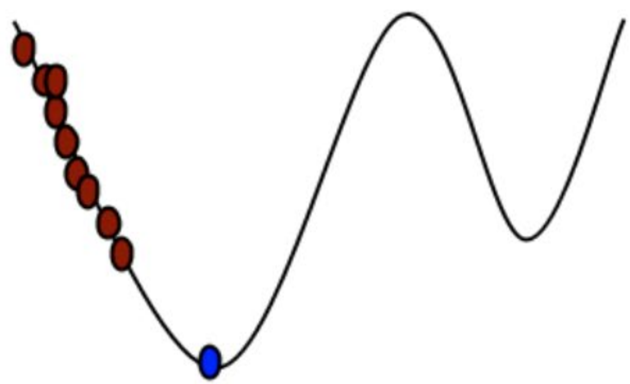
平滑凸优化的一阶算法——梯度法/最陡下降方法



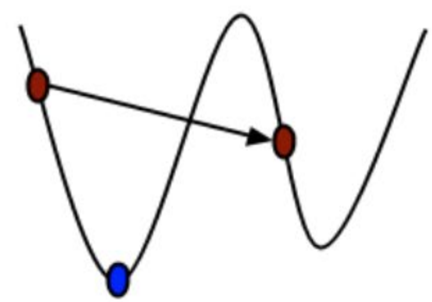
Gradient Descent lecture notes from UD262 Udacity Georgia Tech ML Course.



not differentiable at the corner



very small learning rate needs lots of steps



too big learning rate: missed the minimum

平滑凸优化的一阶算法——梯度法/最陡下降方法

梯度法的基本流程

初始化参数：随机选取，某些启发式算法

↓

选择学习率：算法的稳定性和收敛速度

↓

计算梯度：指向函数变化最快的方向

↓

更新参数

↓

检查收敛条件 —否→ 计算梯度

↓ 是

输出结果

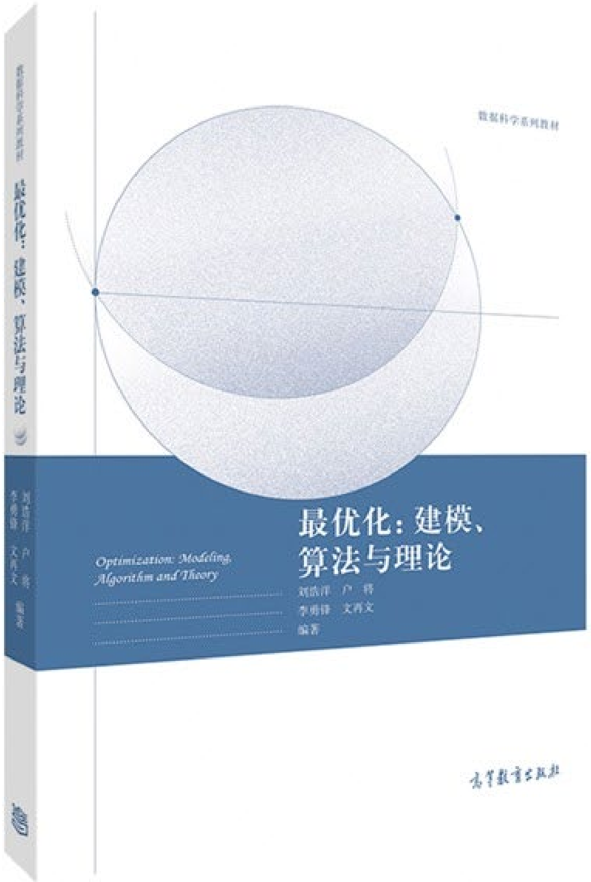
梯度消失和梯度爆炸：在深度学习模型训练中，由于层数过多，梯度可能在传播过程中逐渐变小或变大，导致训练难以进行。需要采取特定策略来解决这些问题。

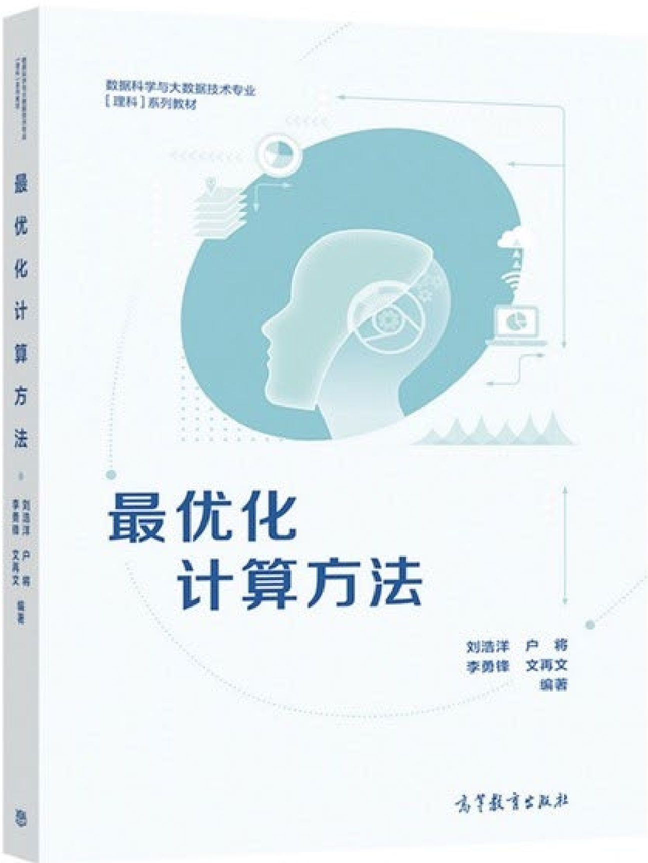
拓展——梯度下降的变体

- 批量梯度下降（Batch Gradient Descent）：每一步更新都使用所有训练样本来计算梯度。精度高但计算量大，对大数据集不够高效。
- 随机梯度下降（Stochastic Gradient Descent, SGD）：每一步更新只使用一个训练样本来计算梯度。计算速度快，但更新过程中有较多噪声。
- 小批量梯度下降（Mini-batch Gradient Descent）：每一步更新使用一小批训练样本来计算梯度。兼顾了批量梯度下降的精度和随机梯度下降的速度，是实际应用中的常用选择。

推荐参考书

http://faculty.bicmr.pku.edu.cn/~wenzw/optbook.html





二阶优化算法：牛顿型迭代算法

最速下降方向 $\Delta \boldsymbol{x} = -\nabla f(\boldsymbol{x})$ 只使用目标函数 $f(\boldsymbol{x})$ 的一阶梯度信息。如果能够再利用目标函数的二阶梯度即 Hessian 矩阵 $\nabla^2 f(\boldsymbol{x}_k)$，则有望找到更好的下降方向。此时，最优下降方向 $\Delta \boldsymbol{x}$ 应该是使 $f(\boldsymbol{x})$ 的二阶 Taylor 逼近函数最小化问题的解

$$\min_{\Delta \boldsymbol{x}} f(\boldsymbol{x} + \Delta \boldsymbol{x}) = f(\boldsymbol{x}) + (\nabla f(\boldsymbol{x}))^{\mathrm{T}} \Delta \boldsymbol{x} + \frac{1}{2} (\Delta \boldsymbol{x})^{\mathrm{T}} \nabla^2 f(\boldsymbol{x}) \Delta \boldsymbol{x} \tag{4.4.10}$$

平稳点

在最优点，相对于参数向量 $\Delta \boldsymbol{x}$ 的梯度必须等于零，即

$$\begin{aligned} \frac{\partial f(\boldsymbol{x} + \Delta \boldsymbol{x})}{\partial \Delta \boldsymbol{x}} &= \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x}) \Delta \boldsymbol{x} = \mathbf{0} \quad \text{牛顿方程} \\ \Longleftrightarrow \Delta \boldsymbol{x}_{\mathrm{nt}} &= -\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x}) \end{aligned} \tag{4.4.11}$$

其中 $\Delta \boldsymbol{x}_{\mathrm{nt}}$ 称为 Newton 步或 Newton 下降方向，相应的寻优方法称为 Newton 法。Newton 法也称 Newton-Raphson 法。

二阶优化算法：牛顿型迭代算法

算法 4.4.1　梯度下降算法及其变型

初始化　选择一个起始点 $\boldsymbol{x}_1 \in \operatorname{dom} f$ 和允许精度 $\varepsilon > 0$，并且令 $k = 1$。

步骤 1　计算目标函数在点 $\boldsymbol{x}_k$ 的梯度 $\nabla f(\boldsymbol{x}_k)$ (以及 Hessian 矩阵 $\nabla^2 f(\boldsymbol{x}_k)$)，并选择下降方向

$$\Delta \boldsymbol{x}_k = \begin{cases} -\nabla f(\boldsymbol{x}_k) & \text{(最速下降法)} \\ -\left(\nabla^2 f(\boldsymbol{x}_k)\right)^{-1} \nabla f(\boldsymbol{x}_k) & \text{(Newton 法)} \end{cases}$$

步长为1，经典牛顿法

步骤 2　选择步长 $\mu_k > 0$。

步骤 3　进行更新

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mu_k \Delta \boldsymbol{x}_k \tag{4.4.12}$$

步骤 4　判断停止准则是否满足：若 $|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)| \leqslant \varepsilon$，则停止迭代，并输出 $\boldsymbol{x}_k$；若不满足，则令 $k \leftarrow k+1$，并返回步骤 1，进行下一轮迭代，直至停止准则满足为止。