优化理论与算法

第八章 梯度下降法

郭加熠|助理教授

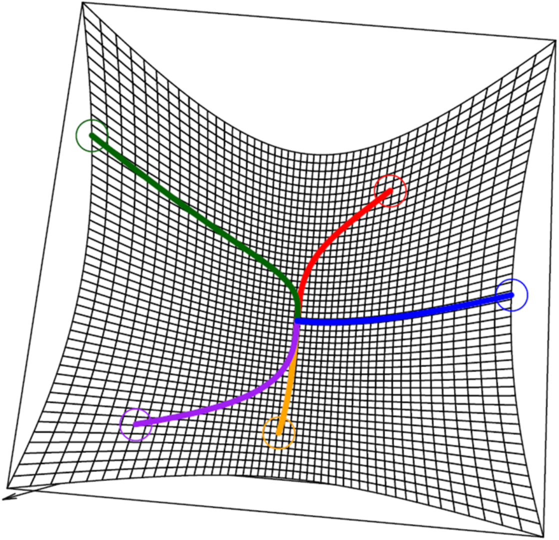
目录

讲　　员

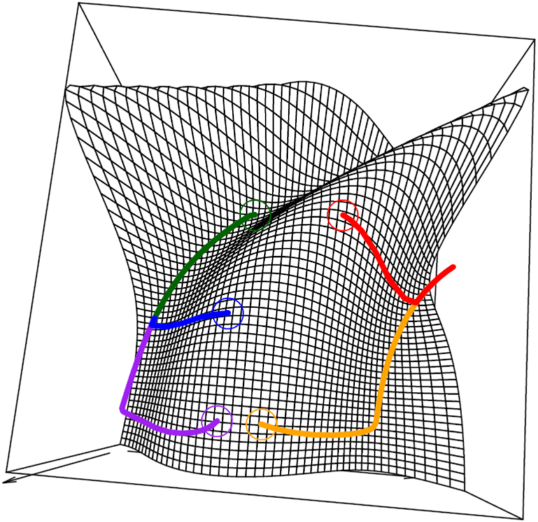
郭加熠，江波，刘慧康

如何走下迷雾笼罩的山峰

梯度下降法实验



梯度下降法实验



梯度下降法

考虑可微函数 f 的无约束优化问题：

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

选择初始点 x^0，重复迭代：

$$x^{k+1} = x^k - t_k \nabla f(x^k), \ \ k = 0, 1, 2, \ldots$$

- $t_k > 0$ 表示自定义的**步长**
- $-\nabla f(x_k)$ 称为最速**下降方向**

  在 x_k 处的泰勒展开：

$$f(x^{k+1}) = f(x^k) - t_k \|\nabla f(x^k)\|_2^2 + o(t_k).$$

算法

Algorithm 1 梯度下降法

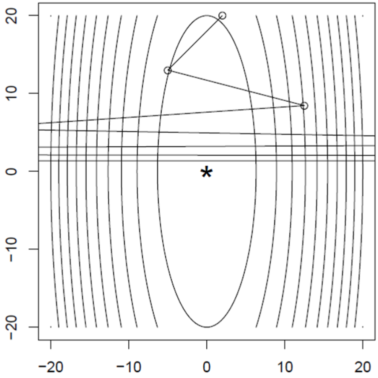
1: 初始点 x_0
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: 选取合适步长 t_k
4: 更新 $x^{k+1} = x^k - t_k \nabla f(x^k)$
5: **if** 满足终止条件 **then**
6: 停止
7: **end if**
8: **end for**

关键问题：如何选择步长？

过大步长的影响

过大步长可能导致**发散**。

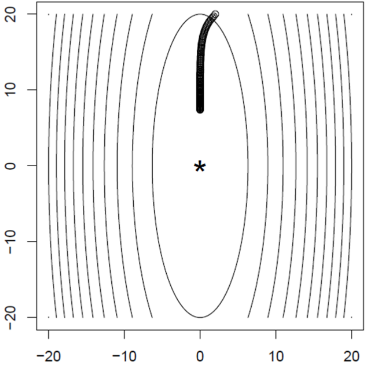
考虑函数 $f(x_1, x_2) = (x_1^2 + 10x_2^2)/2$，固定步长迭代 8 次。



过小步长的影响

过小步长可能导致**收敛缓慢**。

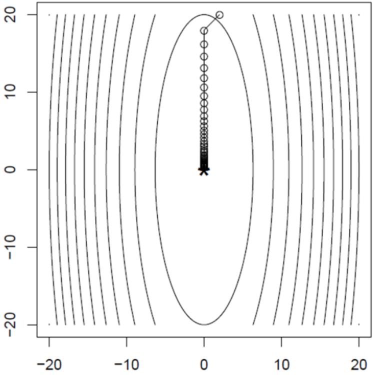
考虑函数 $f(x_1, x_2) = (x_1^2 + 10x_2^2)/2$，固定步长迭代 100 次。



合适步长的选择

合适步长可达到**最佳**效果。

考虑函数 $f(x_1, x_2) = (x_1^2 + 10x_2^2)/2$，固定步长迭代 40 次。



步长选择规则

小步长：

- 优点：大概率下降方向
- 缺点：迭代次数多，计算成本高

大步长：

- 优点：单步改进显著
- 缺点：可能发散或震荡

实用方法：

- 固定步长：$t_k = \alpha$（常数）
- 精确线搜索
- 非精确线搜索：回溯线搜索（最实用）
- 递减步长

精确线搜索

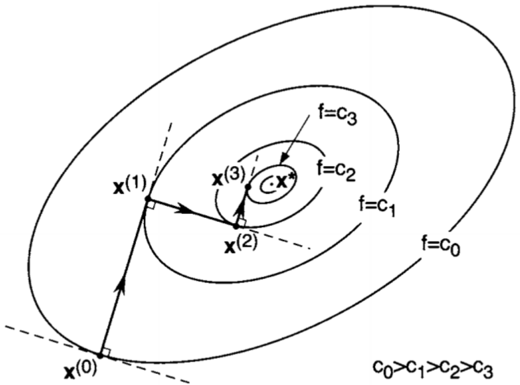
精确线搜索：通过精确最小化确定步长

$$t_k = \operatorname*{argmin}_{t \geq 0} f(x^k - t\nabla f(x^k))$$

适用场景：

- 简单函数，如二次函数
- 函数值计算便宜但梯度计算昂贵

精确线搜索性质



Proposition 8.1 *If* $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ *is a steepest descent sequence for a given function* $f : \mathbb{R}^n \to \mathbb{R}$, *then for each* k *the vector* $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$ *is orthogonal to the vector* $\boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)}$. □

示例：精确线搜索

考虑二维二次优化问题：

$$\min_x f(x) = \frac{1}{2}\left(x_1^2 + 10x_2^2\right)$$

梯度下降迭代公式：

$$x^{k+1} = x^k - t_k \nabla f(x^k) = \left((1-t_k)x_1^k, (1-10t_k)x_2^k\right)$$

取初始点 $x^0 = (10, 1)$：

- 第一步：计算步长

$$t_0 = \operatorname*{argmin}_{t \geq 0} f(x^0 - t\nabla f(x^0)) = \frac{1}{2}\left(100(1-t)^2 + 10(1-10t)^2\right) = \frac{2}{11}$$

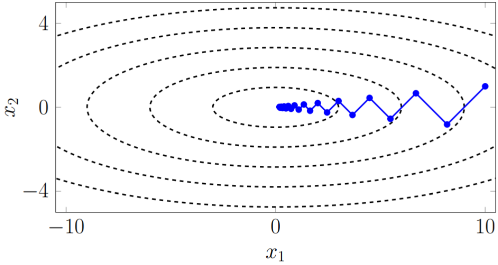
得到 $x^1 = \frac{9}{11}(10, -1)$

- 第二步：计算步长

$$t_1 = \operatorname*{argmin}_{t \geq 0} f(x^1 - t\nabla f(x^1)) = \frac{81}{242}\left(100(1-t)^2 + 10(1-10t)^2\right) = \frac{2}{11}$$

得到 $x^2 = \left(\frac{9}{11}\right)^2 (10, 1)$

- 第 k 步：$x^k = \left(\frac{9}{11}\right)^k (10, (-1)^k)$



示例：一般情况分析

考虑二维二次优化问题：

$$\min_{x} f(x) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$$

其中 $\gamma \geq 1$。取初始点 $x^0 = (\gamma, 1)$，应用精确线搜索可得：

$$\frac{\|x_k - x^*\|_2}{\|x_0 - x^*\|_2} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^k$$

- 梯度下降法常呈现" 锯齿" 运动轨迹
- 收敛速度严重依赖缩放比例

  二次函数 $f(x) = \frac{1}{2}x^T Qx + b^T x$，其中 $\gamma = \lambda_{\max}(Q)/\lambda_{\min}(Q)$

回溯线搜索（Armijo 准则）

- **下降方法**：$x^{k+1} = x^k + t_k p_k$
- **非精确线搜索**：寻找 $t_k \approx \mathrm{argmin}_{t\geq 0} f(x^k - t\nabla f(x^k))$。
  特别是，给定参数 $0 < \alpha < 0.5$，寻找 t_k 满足

$$f(x^k + t_k p_k) \leq f(x^k) + \alpha t_k \nabla f(x^k)^T p_k$$

当 $p_k = -\nabla f(x)$ 时：

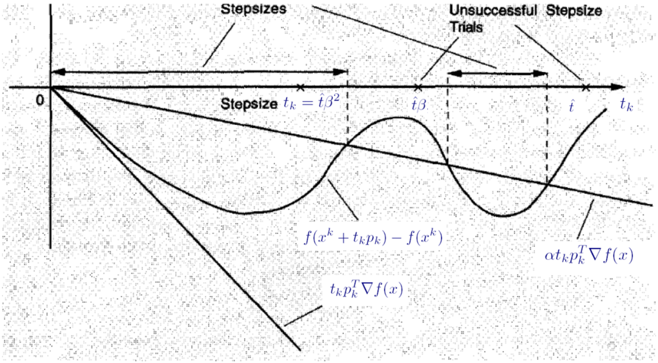
$$f(x^k + t_k p_k) \leq f(x^k) - \alpha t_k \|\nabla f(x^k)\|_2^2$$

数学依据（t 足够小时）：

$$f(x + tp) \approx f(x) + t\nabla f(x)^T p = f(x) - t\alpha\|\nabla f(x)\|_2^2 < f(x)$$

回溯线搜索图示

$$f(x^k + t_k p_k) \leq f(x^k) + \alpha t_k \nabla f(x^k)^T p_k$$



$$令\ g(t) = f(x^k + tp_k)，\ 则$$

$$g'(t) = \nabla f(x^k + tp_k)^T p_k, \quad g'(0) = \nabla f(x^k)^T p_k$$

算法

Algorithm 2 基于 Armijo 准则回溯线搜索

1: **输入**：x^k, p_k，**初始值** $\hat{t} > 0, \alpha > 0, \beta \in (0,1)$

2: **for** $s = 0, 1, 2, \ldots$ **do**

3: 设置 $t_k = \hat{t}\beta^s$

4: **if** 满足 $f(x^k + t_k p_k) \leq f(x^k) + \alpha t_k \nabla f(x^k)^T p_k$ **then**

5: 停止

6: **end if**

7: **end for**

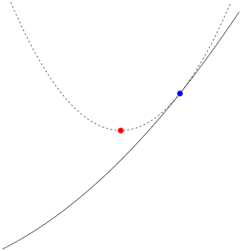
8: **输出**：t_k

- 初始值 $\hat{t} > 0$（例如 $\hat{t} = 1$）
- 要求：$0 < \alpha < 0.5, 0 < \beta < 1$（例如，$\alpha = 10^{-4}$，$\beta = 1/2$）

步长的另一种解释

令 $x_+ = x - t\nabla f(x)$.

$$
\begin{aligned}
x_+ &= \operatorname*{argmin}_y \frac{1}{2t}\|y-(x-t\nabla f(x))\|_2^2 \\
&= \operatorname*{argmin}_y \frac{1}{2t}\|(y-x)+t\nabla f(x))\|_2^2 \\
&= \operatorname*{argmin}_y \frac{t}{2}\|\nabla f(x)\|_2^2 + \nabla f(x)^T(y-x) + \frac{1}{2t}\|y-x\|_2^2 \\
&= \operatorname*{argmin}_y f(x) + \nabla f(x)^T(y-x) + \frac{1}{2t}\|y-x\|_2^2
\end{aligned}
$$



几何解释：最小化 f 在 x^k 处的线性逼近与保持 x_+ 接近 x 的正则项

目录

梯度下降法

收敛性分析

讲　　　　员

郭加熠，江波，刘慧康

梯度下降法收敛性分析

分析梯度下降法的收敛性

$$x^{k+1} = x^k - t_k \nabla f(x^k)$$

假设：

- f 可微且定义域为 $\mathbb{R}^n$
- 最优值有下界

即将呈现：函数 f 的好性质与收敛速度的关系

- L-光滑函数
- 凸且 L-光滑函数
- μ-强凸且 L-光滑函数
- 不可微函数（?）

Lipschitz 连续性与等价形式

以下性质满足 $1 \Rightarrow 2 \Leftrightarrow 3$，且当 f 凸时 $1 \Leftarrow 2$：

1 ∇f **L-Lipschitz 连续**：对任意 $x, y \in \operatorname{dom} f$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2$$

2 f **L-光滑**（二次上界）：

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2$$

3 **二阶条件**：$\nabla^2 f(x) \preceq LI$ 或 $\frac{L}{2} x^T x - f(x)$ 是凸函数

可得：

$$f^* \leq f(x^k) - \frac{1}{2L} \|\nabla f(x^k)\|_2^2 \quad \text{or} \quad \|\nabla f(x^k)\|_2^2 \leq 2L(f(x^k) - f^*)$$

(资源:https://xingyuzhou.org/blog/notes/Lipschitz-gradient)

一般函数（非凸）收敛性分析

- 求出最优解要求太高了！
- 寻求 ϵ-**稳定点**：$\|\nabla f(x)\| \le \epsilon$

定理：若 f 满足 L-**光滑**、有下界，采用**固定步长** $0 < t < 2/L$，则梯度下降法满足：

$$\min_{i=0,\ldots,k} \|\nabla f(x^i)\|_2 \le \sqrt{\frac{f(x^0) - f^*}{(k+1)M}}$$

其中 $M = t(1 - \frac{Lt}{2})$

- 收敛速率 $O(1/\sqrt{k})$，需 $O(1/\epsilon^2)$ 次迭代找到 ϵ-稳定点
- 线搜索方法具有类似结论（常数项略有不同）

证明

- L-光滑性：

$$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$$

- **充分下降性**：取 $x \leftarrow x^k$, $y \leftarrow x^{k+1}$, and $x^{k+1} = x^k - t\nabla f(x^k)$.

$$f(x^{k+1}) \le f(x^k) - t(1-\frac{Lt}{2})\|\nabla f(x^k)\|_2^2 = f(x^k) - M\|\nabla f(x^k)\|_2^2$$

- 当 $t \in (0, 2/L)$ 时，有 $M > 0$。移项得：

$$\|\nabla f(x^k)\|_2^2 \le \frac{1}{M}\left(f(x^k) - f(x^{k+1})\right)$$

- 累加并取下界得：

$$\min_{i=0,1,\dots,k}\|\nabla f(x^i)\|_2^2 \cdot (k+1) \le \sum_{i=0}^{k}\|\nabla f(x^i)\|_2^2 \le \frac{1}{M}\left(f(x^0) - f(x^{k+1})\right)$$

凸且 L 光滑函数的收敛性

- 寻求 ϵ-**次优解**：$|f(x^k) - f(x^*)| \le \epsilon$

定理：若 f **凸**且 L-**光滑**且有下界，采用**固定步长** $0 < t \le 1/L$，则梯度下降法有：

$$f(x^k) - f^* \le \frac{\left\|x^0 - x^*\right\|_2^2}{2kt}$$

- 收敛速率 $O(1/k)$
- 需 $O(1/\epsilon)$ 次迭代找到 ϵ-次优解

证明

梯度下降法，固定步长 t：

$$x^{k+1} = x^k - t\nabla f(x^k)$$

充分下降性：

$$f(x^{k+1}) \le f(x^k) - t(1 - \frac{Lt}{2})\|\nabla f(x^k)\|_2^2$$

由于 $0 < t \le 1/L$，可推出 $(1 - \frac{Lt}{2}) \ge \frac{1}{2}$。因此，

$$\begin{aligned}
f(x^{k+1}) &\le f(x^k) - \frac{t}{2}\left\|\nabla f(x^k)\right\|_2^2 \\
\text{(由于凸性)} &\le f(x^*) + \nabla f(x^k)^T(x^k - x^*) - \frac{t}{2}\left\|\nabla f(x^k)\right\|_2^2 \\
&= f(x^*) + \frac{1}{2t}(\left\|x^k - x^*\right\|_2^2 - \left\|(x^k - x^*) - t\nabla f(x)\right\|_2^2) \\
&= f(x^*) + \frac{1}{2t}(\left\|x^k - x^*\right\|_2^2 - \left\|x^{k+1} - x^*\right\|_2^2)
\end{aligned}$$

累加前 k 次迭代:

$$\begin{aligned}\sum_{i=1}^{k}\left(f(x^i)-f(x^*)\right) &\leq \frac{1}{2t}\sum_{i=1}^{k}(\|x^{i-1}-x^*\|_2^2-\|x^i-x^*\|_2^2) \\ &= \frac{1}{2t}(\|x^0-x^*\|_2^2-\|x^k-x^*\|_2^2) \\ &\leq \frac{1}{2t}\left\|x^0-x^*\right\|_2^2\end{aligned}$$

由于 $f(x^i)$ 是非递增函数，有下式成立

$$f(x^k)-f(x^*)\leq\frac{1}{k}\sum_{i=1}^{k}f(x^i)-f(x^*)\leq\frac{1}{2kt}\left\|x^0-x^*\right\|_2^2$$

总之，达到 $f(x^k)-f(x^*)\leq\epsilon$ 的迭代次数是 $O(1/\epsilon)$.

有界梯度与回溯线搜索

- **有界梯度**: $\|\nabla f(x^k)\|_2^2 \le 2L(f(x^k) - f^*) \le \dfrac{L}{kt}\left\|x^0 - x^*\right\|_2^2$
- 类似的收敛性结论对**回溯线搜索** (第 17 页) 仍然成立。原因是当 $0 < t_k \le 1/L$ 且 $0 < \alpha < 1/2$ 时，结合第 27 页结果可得

$$f(x^{k+1}) = f(x^k) - \frac{t_k}{2}\left\|\nabla f(x^k)\right\|^2 \le f(x^k) - \alpha t_k \left\|\nabla f(x^k)\right\|^2$$

因此回溯线搜索终止时，要么 $t_k = 1$（初始值），要么 $t_k \ge \frac{\beta}{L}$。

综上，

$$f(x^k + t_k p_k) \le f(x^k) - \min\{\alpha, \alpha\beta/L\}\|\nabla f(x^k)\|_2^2$$

我们能做得更好吗？

一阶方法: 其迭代方法有更新规则

$$x^0 + \mathrm{span}\{\nabla f(x^0), \nabla f(x^1), ..., \nabla f(x^{k-1})\}$$

定理： 对于任意给定的迭代次数 k 和初始点 x^0，存在一个凸且 L-光滑的函数 f，使得任意一阶方法满足：

$$f(x^k) - f(x^*) \geq \frac{3L}{32(k+1)^2} \|x^0 - x^*\|_2^2$$

- 能否达到 $O(1/k^2)$ 的收敛速度?
- 可以。使用动量法或 Nesterov 加速梯度法

强凸函数与等价表述

以下性质等价:

- μ-**强凸性**: 若函数 f 关于参数 $\mu > 0$ 强凸，则

$$f(x) - \frac{\mu}{2} x^T x \text{ 是凸函数}$$

- 一阶条件（**二次下界**）

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|x - y\|_2^2$$

- 二阶条件: $\nabla^2 f(x) \succeq \mu I$

可推导得:

$$f^* \geq f(x^k) - \frac{1}{2\mu} \|\nabla f(x^k)\|_2^2 \quad \text{或} \quad \|\nabla f(x^k)\|_2^2 \geq 2\mu (f(x^k) - f^*)$$

(资源:https://xingyuzhou.org/blog/notes/strong-convexity)

收敛性分析：强凸且 L-光滑函数

定理： 若 f 是 μ-**强凸**且 L-**光滑**函数，在 $\mathbb{R}^n$ 上有下界，并采用回溯线搜索，则梯度下降法满足：

$$f(x^k) - f^* \leq c^k (f(x^0) - f^*)$$

其中收敛率系数 $c = 1 - \min\{2\alpha\mu, 2\alpha\beta\mu/L\} < 1$。

- 线性收敛速率
- 可在 $O(\log(1/\epsilon))$ 次迭代中找到 ϵ-次优点

证明

- 回溯线搜索条件（第 29 页）:

$$f(x^{k+1}) \le f(x^k) - \min\{\alpha, \alpha\beta/L\}\|\nabla f(x^k)\|_2^2$$

- μ-强凸性成立有如下性质:

$$\|\nabla f(x^k)\|_2^2 \ge 2\mu(f(x^k) - f^*)$$

- 因此有:

$$\begin{aligned} f(x^{k+1}) - f^* &\le (f(x^k) - f^*) - \min\{\alpha, \alpha\beta/L\}\|\nabla f(x^k)\|_2^2 \\ &\le c(f(x^k) - f^*) \end{aligned}$$

- 最终得到:

$$f(x^k) - f^* \le c^k(f(x^0) - f^*)$$

固定步长情形类似

定理： 若 f 是 μ-**强凸**且 L-**光滑**函数，在 $\mathbb{R}^n$ 上有下界，并采用固定步长 $0 < t < \frac{2}{\mu+L}$，则梯度下降法满足：

$$\|x^k - x^*\|_2^2 \leq \left(1 - t\left(\frac{2\mu L}{\mu + L}\right)\right)^k \|x^0 - x^*\|_2^2$$

证明关键步骤：

- 定义 $h(x) = f(x) - \frac{\mu}{2}x^T x$ 为凸且 $(L-\mu)$-光滑函数
- 证明不等式：
  $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{\mu L}{\mu+L}\|x-y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2$
- 从 $\|x^{k+1} - x^*\|_2^2 = \|(x^k - x^*) - t\nabla f(x)\|_2^2$ 展开推导

条件数

- 达到 $f(x^k) - f^* \leq \epsilon$ 所需迭代次数下限为:

$$\frac{\log((f(x^0) - f^*)/\epsilon)}{\log(c^{-1})} \approx \frac{\log((f(x^0) - f^*)/\epsilon)}{\log((1 - \mu/L)^{-1})} \approx \frac{L}{\mu} \log\left(\frac{f(x^0) - f^*}{\epsilon}\right)$$

  其中 L/μ 称为**条件数**

- 条件数是影响梯度下降速度的**主要因素**。以第 16 页例子为例，当 $\gamma = 10000$ 且 $k = 100$ 时，有 $\left(\frac{\gamma-1}{\gamma+1}\right)^k = 0.98$
- 当 L/μ 较大时，问题称为**病态条件问题**

总结：是否选择梯度下降法

优点

- 设计简单，每次迭代计算量小
- 无需计算二阶导数

缺点

- 对病态条件问题收敛缓慢

思考方向...

- 若函数 f 不可微时如何应对?
- 若可行域为 C 而非 $\mathbb{R}^n$ 时如何处理?
- 如何加速方法以达到最优收敛率?
- 二阶信息能带来哪些优势?