**MATH 3011 运筹学** **秋季 2023**

# Lecture 10: 无约束优化 梯度下降法

*Lecturer:* 陈士祥 *Scribes:* 陈士祥

## 1 问题形式

无约束最优化问题

$$\min_{x\in\mathbb{R}^n} \quad f(x) \tag{10.1}$$

其目标函数 $f$ 是定义在 $\mathbb{R}^n$ 上的实值函数，决策变量 $x$ 的可取值之集合是全空间 $\mathbb{R}^n$. $f$ 是可微函数。

## 2 梯度下降方法

梯度下降法取负梯度作为迭代算法的搜索方向，其迭代格式为

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

**Algorithm 1** 梯度下降算法 GD

**Require:** 选取初始点 $x^0$, 设置终止误差 $\varepsilon > 0$, 令 $k := 0$.

1: **while** $\|\nabla f(x^k)\| > \epsilon$ **do**

2: 令 $d^k = -\nabla f(x^k)$, 并由一维搜索确定步长因子 $\alpha_k$ 使得 $f(x^k+\alpha_k d^k)$ 满足 Backtracking linesearch 或者 Wolfe-Powell 条件

3: 迭代更新 $x^{k+1} = x^k + \alpha_k d^k$, 置 $k := k+1$ 。

4: **end while**

## 3 梯度下降法全局收敛性定理

我们下面着重讲解，在非凸、凸函数、强凸三种情况下，梯度算法的收敛结果。

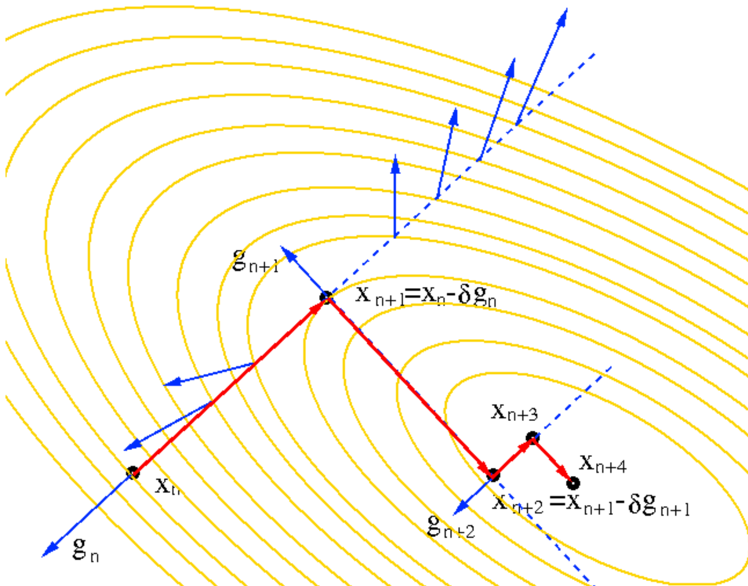


图 10.1: 梯度下降法求解二次问题的迭代示意图

### 3.1 非凸函数情况下的收敛

**Definition 10.1** 若给定函数 $f$ 是可微函数，并且对于任意定义域的点 $x, y$，梯度满足李氏连续性（*Lipschitz continuous*），即存在 $L > 0$，使得

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

则称 $f$ 是梯度李氏连续，或者李氏光滑（$L$-光滑）的。

**Lemma 10.1** 若 $f$ 是李氏光滑的，则 $f$ 有二次上界，即

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

**Proof:** 由 $f$ 可微，可得

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \nabla f(x + t(y - x))^T(y - x)dt \\ &= f(x) + \nabla f(x)^T(y - x) + \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T(y - x)dt \end{aligned}$$

因此，

$$\begin{aligned}
f(y)-f(x)-\nabla f(x)^T(y-x) &= \int_0^1 (\nabla f(x+t(y-x))-\nabla f(x))^T(y-x)dt \\
&\leq \int_0^1 \|\nabla f(x+t(y-x))-\nabla f(x)\|\|y-x\|dt \\
&\leq \int_0^1 Lt\|y-x\|^2 dt \\
&= \frac{L}{2}\|y-x\|^2.
\end{aligned}$$

■

很多常用的函数满足李氏光滑性, 例如 $f(x)=\frac{1}{2}\|Ax-b\|^2$. 我们有

$$\|\nabla f(y)-\nabla f(x)\| \leq \lambda_{\max}(A^TA)\|y-x\|,$$

这里 $\lambda_{\max}(A^TA)$ 是 $A^TA$ 的最大特征根。因此，二次函数的 Lipschitz 常数是 $L=\lambda_{\max}(A^TA)$. 通常来说, $L=\max_x \lambda_{\max}(\nabla^2 f(x))$, 即定义域内的所有 Hessian 矩阵的最大特征根。

**反例：** $f(x)=e^x, f(x)=x^3$.

**作业 10.1** 对于逻辑回归问题，$\min f(x)=\frac{1}{N}\sum_{i=1}^N \log(1+e^{-y_i a_i^T x})$, 这里 $y_i\geq 0, a_i$ 是已知的。估计 $\nabla f$ 李氏常数 $L$.

**梯度法的另一个理解：最大化最小化算法类**

构造 $f(x)$ 的一个二次函数上界.

**定义**: $q_x(y)$ 是 $f$ 的上界函数, 如果

- $q_x(y)=f(x)$
- $q_x(y)\geq f(y)$, for any $y$.

**最大化-最小化方法**:

$$x_{k+1}=\arg\min_y q_{x_k}(y)$$

我们有

$$f(x_{k+1})\leq q_{x_k}(x_{k+1})\leq q_{x_k}(x_k)=f(x_k)$$

**Theorem 10.1** *若 $f$ 是 $L-$ 光滑函数并且 $f$ 有最小值 $f^*$,则选取步长 $\alpha_k=1/L$, 我们有 $\lim_{k\to\infty}\|\nabla f(x_k)\|=0$ 并且并且对于任意正整数 $T$，有*

$$\min_{k=0,1,\ldots,T}\|\nabla f(x_k)\|^2\leq \frac{2L(f(x_0)-f^*)}{T}.$$

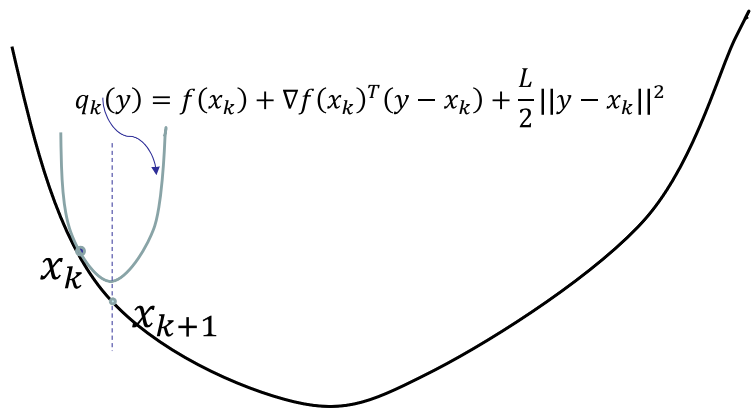


图 10.2: 最大化-最小化示例，在每个点 $x_k$ 出，用二次上界 $q_k(y)$ 作为原函数的逼近。这样可以通过最小化上界函数得到函数值下降的迭代点 $x_{k+1}$.

**证明：** 由李氏光滑性，$q_{x_k}(y) = f(x_k) + \nabla f(x_k)^T(y - x_k) + \frac{1}{2\alpha}\|y - x_k\|^2$ 为一个上界函数。梯度法迭代满足

$$x_{k+1} = \arg\min_y q_{x_k}(y) = x_k - 1/L\nabla f(x_k).$$

所以

$$f(x_{k+1}) \le q_{x_k}(x_{k+1}) = q_{x_k}(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 = f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2. \tag{10.2}$$

因此

$$\sum_{k=0}^{\infty} \frac{1}{2L}\|\nabla f(x_k)\|^2 < \infty.$$

故

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

并且对于任意正整数 $T$，有 $\min_{k=0,1,\dots,T} \|\nabla f(x_k)\|^2 \le \frac{2L(f(x_0)-f^*)}{T}$.

### 3.2 凸函数情况下的收敛

**Lemma 10.2** *设函数 $f(x)$ 是 $\mathbb{R}^n$ 上的凸可微函数, 则以下结论等价:*

1. *$f$ 的梯度为 $L$ -连续的;*

2. *$\nabla f(x)$ 有余强制性, 即对任意的 $x, y \in \mathbb{R}^n$, 有*

$$(\nabla f(x) - \nabla f(y))^T (x-y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \tag{10.3}$$

**我们只证明: (1)⇒ (2)** 定义函数 $\phi(y) = f(y) - \nabla f(x)^T y$. 函数 $\phi(y)$ 是凸函数，并且也是 $L-$ 光滑的。因为 $\nabla \phi(x) = 0$, 故 $x$ 是 $\phi$ 的最小值。根据 $L-$ 光滑，

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

由 $\nabla\phi(y) = \nabla f(y) - \nabla f(x)$ 可得

$$\phi(y) - \phi(x) \geq \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2.$$

即

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2.$$

交换上面 $x, y$，得到的不等式与上述不等式相加，即可得到结论。

**Theorem 10.2** *若 $f$ 是 $L-$ 光滑的凸函数，并且 $f$ 有最小值 $f^*$，则选取步长 $\alpha_k = 1/L$, 对于任意 $T \geq 1$,*

$$f(x_T) - f^* \leq \frac{L}{2T}\|x_0 - x^*\|^2.$$

**证明:** 由(10.2)和 $f$ 是凸函数可得

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\
&\leq f^* + \nabla f(x_k)^T (x_k - x^*) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\
&= f^* + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right) \\
&= f^* + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)
\end{aligned} \tag{10.4}$$

(10.4)表明梯度法中，函数值和最小值的差是严格减小的。对上式 $k=0,1,...,T$ 相加可得

$$\begin{aligned}\sum_{k=1}^{T}(f(x_k)-f^*) &\leq \frac{L}{2}\sum_{k=1}^{T}\left(\|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2\right)\\ &= \frac{L}{2}\left(\|x_0-x^*\|^2-\|x_{T+1}-x^*\|^2\right)\\ &\leq \frac{L}{2}\|x_0-x^*\|^2.\end{aligned}$$

因为 $f(x_k)$ 单调递减，所以

$$f(x_T)-f^* \leq \frac{L}{2T}\|x_0-x^*\|^2.$$

**结论：$f(x_k)-f^*$ 收敛的速度是次线性的**。收敛到 $f(x_k)-f^*\leq\epsilon$ 的速度是 $\mathcal{O}(1/k)$.

**一阶算法的收敛下界：**

**Definition 10.2 一阶方法**：任何选择 $x_{k+1}$ 在集合中的迭代算法

$$x_0+span\{\nabla f(x_0),\nabla f(x_1),\ldots,\nabla f(x_k)\}$$

**问题类**：满足 L 式光滑和凸假设的任何函数。
**定理 (Nesterov)**：对于每个整数 $k\leq\frac{n-1}{2}$ 和每个 $x_0$，存在在问题类中的函数，对于任何一阶方法

$$f(x_k)-f^*\geq\frac{3L\left\|x_0-x^*\right\|^2}{32(k+1)^2}$$

- 表明梯度方法的 $\frac{1}{k}$ 速率不是最优的。
- Nesterov's 加速梯度方法有 $\frac{1}{k^2}$ 的收敛性。

该定理见 Yu. Nesterov, Lectures on Convex Optimization (2018), section 2.1.（Theorem 2.1.7 in the book.）Nesterov 加速梯度算法，感兴趣也参考此书 section 2.2.

## 3.3 强凸函数收敛性

**Definition 10.3** 可微函数是 $\mu-$ 强凸函数，如果

$$f(y)\geq f(x)+\nabla f(x)^T(y-x)+\frac{\mu}{2}\|y-x\|^2,\forall x,y\in\mathrm{dom}f.$$

假设中增加强凸性后，我们可以得到更好的结果。强凸性意味着最小值点唯一。

**Lemma 10.3** 设函数 $f(x)$ 是 $\mathbb{R}^n$ 上的 $\mu-$ 强凸可微函数，则有如下不等式：

$$(\nabla f(x)-\nabla f(y))^T(x-y)\geq\frac{\mu L}{L+\mu}\|x-y\|^2+\frac{1}{L+\mu}\|\nabla f(x)-\nabla f(y)\|^2,\forall x,y\in\mathrm{dom}f \tag{10.5}$$

**证明：** 记 $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$. 则 $\phi(x)$ 是凸函数, 并且是 $L-\mu$ 李氏光滑。由余强制性(10.3), 可得

$$(\nabla\phi(x) - \nabla\phi(y))^T(x-y) \geq \frac{1}{L-\mu}\|\nabla\phi(x) - \nabla\phi(y)\|^2, \forall x, y \in\in \mathrm{dom} f.$$

带入 $\nabla\phi(x) = \nabla f(x) - \mu x$，可得(10.5).

如果 $x^+ = x - \alpha\nabla f(x)$ 且 $0 < \alpha \leq \frac{2}{\mu+L}$：

$$\begin{aligned}
\left\|x^+ - x^*\right\|^2 &= \|x - \alpha\nabla f(x) - x^*\|^2 \\
&= \|x - x^*\|^2 - 2\alpha\nabla f(x)^T(x - x^*) + \alpha^2\|\nabla f(x)\|^2 \\
&\leq (1 - \alpha\frac{2\mu L}{\mu+L})\|x - x^*\|^2 + \alpha(\alpha - \frac{2}{\mu+L})\|\nabla f(x)\|^2 \\
&\leq (1 - \alpha\frac{2\mu L}{\mu+L})\|x - x^*\|^2
\end{aligned}$$

$$\|x_k - x^*\|^2 \leq c^k\|x_0 - x^*\|^2$$

其中 $c = 1 - \alpha\frac{2\mu L}{\mu+L}$。

- 这意味着 $x_k$ 线性收敛至最优值 $x^*$。
- 对于 $\alpha = \frac{2}{\mu+L}$，我们得到 $c = \left(\frac{\kappa-1}{\kappa+1}\right)^2$ 其中 $\kappa = \frac{L}{\mu}$ 被称为条件数。例如，正定矩阵 $A$ 的条件数是其最大特征根与最小特征根比值。矩阵条件数大，意味着问题是病态的。

$$f(x_k) - f^* \leq \frac{L}{2}\|x_k - x^*\|^2 \leq c^k\frac{L}{2}\|x_0 - x^*\|^2$$

结论：达到 $f(x_k) - f^* \leq \epsilon$ 所需的迭代次数是 $O(\log(1/\epsilon))$。

# 4 梯度下降法总结

| 问题类型 | 收敛描述 | 迭代复杂度 |
|---|---|---|
| Nonconvex $L$-smooth | $\|\nabla f(x)\| \leq \epsilon$ | $O\left(\frac{1}{\epsilon^2}\right)$ |
| Convex $L$-smooth | $f(x_k) - f^* \leq \epsilon$ | $O\left(\frac{1}{\epsilon}\right)$ |
| Strongly convex $\mu$-smooth | $\|x_k - x^*\|^2 < \epsilon$ | $O\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$ |

表 10.1: Convergence for gradient method under function properties