

最优化方法 第四讲

Weiwen Wang(王伟文)

暨南大学

2025 年 4 月 10 日

目录

- 1 Nesterov 加速梯度下降
- 2 Polyak-Lojasiewicz 不等式
- 3 坐标下降法
 - 随机坐标下降法
 - 重要性采样
 - 最速坐标下降法
 - ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Nesterov 加速梯度下降

取 $\mathbf{z}_0 = \mathbf{y}_0 = \mathbf{x}_0$

- ① $\mathbf{y}_{k+1} = \mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k)$
- ② $\mathbf{z}_{k+1} = \mathbf{z}_k - \frac{k+1}{2L} \nabla f(\mathbf{x}_k)$
- ③ $\mathbf{x}_{k+1} = \frac{k+1}{k+3} \mathbf{y}_{k+1} + \frac{2}{k+3} \mathbf{z}_{k+1}$

无加速梯度下降

$$\bullet \mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k)$$

Nesterov 加速梯度下降

定理 1

设 f 为可微凸函数且为 L -smooth, 记 $\mathbf{x}^*$ 为 f 的全局最小值点, 则 Nesterov 加速梯度下降法满足

$$f(\mathbf{y}_K) - f(\mathbf{x}^*) \leq \frac{2L\|\mathbf{z}_0 - \mathbf{x}^*\|^2}{K(K+1)}, \quad K > 0.$$

评述 Nesterov 加速梯度下降的算法复杂度为 $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$, 比 $\mathcal{O}(\frac{1}{\epsilon})$ 更好.

证明

证明思路

① 构造 potential 函数

$$\Phi(k) = k(k+1)(f(\mathbf{y}_k) - f(\mathbf{x}^*)) + 2L\|\mathbf{z}_k - \mathbf{x}^*\|^2.$$

② 证明 $\Phi(k+1) \leq \Phi(k)$, 若成立则有

$$K(K+1)(f(\mathbf{y}_K) - f(\mathbf{x}^*)) \leq \underbrace{K(K+1)(f(\mathbf{y}_K) - f(\mathbf{x}^*)) + 2L\|\mathbf{z}_K - \mathbf{x}^*\|^2}_{\Phi(K)} \leq \underbrace{2L\|\mathbf{z}_0 - \mathbf{x}^*\|^2}_{\Phi(0)}$$

为了证明 $\Phi(k+1) \leq \Phi(k)$, 只需证

$$\Delta := \frac{\Phi(k+1) - \Phi(k)}{k+1} \leq 0.$$

为了简化符号, 记 $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

为了证明 $\Phi(k+1) \leq \Phi(k)$, 只需证

$$\Delta := \frac{\Phi(k+1) - \Phi(k)}{k+1} \leq 0.$$

为了简化符号, 记 $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$

$$\begin{aligned} \Delta &= k(f(\mathbf{y}_{k+1}) - f(\mathbf{y}_k)) + 2(f(\mathbf{y}_{k+1}) - f(\mathbf{x}^*)) + \frac{2L}{k+1} \left(\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 - \|\mathbf{z}_k - \mathbf{x}^*\|^2 \right) \\ &= \cdots + \cdots + \frac{k+1}{2L} \|\mathbf{g}_k\|^2 - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle \end{aligned}$$

为了证明 $\Phi(k+1) \leq \Phi(k)$, 只需证

$$\Delta := \frac{\Phi(k+1) - \Phi(k)}{k+1} \leq 0.$$

为了简化符号, 记 $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$

$$\Delta = k(f(\mathbf{y}_{k+1}) - f(\mathbf{y}_k)) + 2(f(\mathbf{y}_{k+1}) - f(\mathbf{x}^*)) + \frac{2L}{k+1} \left(\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 - \|\mathbf{z}_k - \mathbf{x}^*\|^2 \right)$$

$$= \cdots + \cdots + \frac{k+1}{2L} \|\mathbf{g}_k\|^2 - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle$$

$$\leq \cdots + \cdots + (k+1)(f(\mathbf{x}_k) - f(\mathbf{y}_{k+1})) - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle \quad (\text{充分下降引理})$$

$$= k(f(\mathbf{x}_k) - f(\mathbf{y}_k)) + 2(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + (f(\mathbf{y}_{k+1}) - f(\mathbf{x}_k)) - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle$$

$$\leq k(f(\mathbf{x}_k) - f(\mathbf{y}_k)) + 2(f(\mathbf{x}_k) - f(\mathbf{x}^*)) - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle \quad (\text{充分下降引理})$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

$$\begin{aligned}\Delta &\leq k(f(\mathbf{x}_k) - f(\mathbf{y}_k)) + 2(f(\mathbf{x}_k) - f(\mathbf{x}^*)) - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle \\ &\leq k\langle \mathbf{g}_k, \mathbf{x}_k - \mathbf{y}_k \rangle + 2\langle \mathbf{g}_k, \mathbf{x}_k - \mathbf{x}^* \rangle - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle\end{aligned}$$

$$\begin{aligned}
 \Delta &\leq k(f(\mathbf{x}_k) - f(\mathbf{y}_k)) + 2(f(\mathbf{x}_k) - f(\mathbf{x}^*)) - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle \\
 &\leq k\langle \mathbf{g}_k, \mathbf{x}_k - \mathbf{y}_k \rangle + 2\langle \mathbf{g}_k, \mathbf{x}_k - \mathbf{x}^* \rangle - 2\langle \mathbf{g}_k, \mathbf{z}_k - \mathbf{x}^* \rangle \\
 &= \langle \mathbf{g}_k, (k+2)\mathbf{x}_k - k\mathbf{y}_k - 2\mathbf{z}_k \rangle. \\
 &= 0.
 \end{aligned}$$

因为

$$\mathbf{x}_k = \frac{k}{k+2}\mathbf{y}_k + \frac{2}{k+2}\mathbf{z}_k$$

定义 2 (Polyak-Lojasiewicz 不等式)

设 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 为可微函数, $\mathbf{x}^*$ 为 f 的全局最小值点, 若存在 $\mu > 0$, 使得对任意 $\mathbf{x} \in \mathbb{R}^d$, 有

$$\frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2 \geq f(\mathbf{x}) - f(\mathbf{x}^*)$$

则称 f 满足 Polyak-Lojasiewicz(PL) 不等式.

- 若 f 满足 PL 不等式, 则驻点一定是全局最小值点.

命题 3

设函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微, 且为 μ -强凸, 则 f 满足以 μ 为参数的 Polyak-Lojasiewicz 不等式

- PL 不等式不能推出强凸性质
- 甚至一些非凸函数, 如 $f(x) = x^2 + 3\sin^2(x)$ 也满足 PL 不等式.

证明

由可微、 μ -强凸函数的性质, 知 f 的全局最小值点 $\mathbf{x}^*$ 存在且唯一, 且对任意 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ 有

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \\ &\geq \min_{\mathbf{y}} \left\{ f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \right\} \\ &= f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2 \end{aligned}$$

证明

由可微、 μ -强凸函数的性质, 知 f 的全局最小值点 $\mathbf{x}^*$ 存在且唯一, 且对任意 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ 有

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \\ &\geq \min_{\mathbf{y}} \left\{ f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \right\} \\ &= f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2 \end{aligned}$$

故有

$$\min_{\mathbf{y}} f(\mathbf{y}) = f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2$$

PL 不等式与梯度下降法收敛性

定理 4

设 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微且满足以 μ 为参数的 *Polyak-Lojasiewicz* 不等式. 在梯度下降法中取步长 $h = \frac{1}{L}$, 则有

$$f(\mathbf{x}_K) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{L}\right)^K (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \quad K > 0$$

其中 $\mathbf{x}^*$ 为 f 的全局最小值点, $\mathbf{x}_0$ 为梯度下降法初始点.

证明

取步长 $h = \frac{1}{L}$ 时, 由充分下降引理

$$\frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}).$$

另一方面, 由 PL 不等式, 有 $\frac{1}{2\mu} \|\nabla f(\mathbf{x}_k)\|^2 \geq f(\mathbf{x}_k) - f(\mathbf{x}^*)$.

证明

取步长 $h = \frac{1}{L}$ 时, 由充分下降引理

$$\frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}).$$

另一方面, 由 PL 不等式, 有 $\frac{1}{2\mu} \|\nabla f(\mathbf{x}_k)\|^2 \geq f(\mathbf{x}_k) - f(\mathbf{x}^*)$.
结合上述两个不等式,

$$\frac{\mu}{L} (f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$$

整理得到

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{L}\right) (f(\mathbf{x}_k) - f(\mathbf{x}^*))$$

逐坐标 $\mathcal{L}$ -smooth

定义 5 (逐坐标 $\mathcal{L}$ -smooth)

设 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微. 若对于任意 $i \in \{1, \dots, d\}$, 存在 $L_i > 0$, 对于任意 $\mathbf{x} \in \mathbb{R}^d$ 满足

$$\|\nabla f(\mathbf{x} + \lambda \mathbf{e}_i) - \nabla f(\mathbf{x})\| \leq L_i |\lambda|, \forall \lambda \in \mathbb{R},$$

其中 $\mathbf{e}_i$ 为标准基. 则称 f 以参数 $\mathcal{L} = (L_1, \dots, L_d)$ 逐坐标 smooth(光滑). 若对所有 i , 有 $L_i = L$, 则记 $\mathcal{L} = L$.

- 逐坐标 $\mathcal{L}$ -smooth 函数满足

$$f(\mathbf{x} + \lambda \mathbf{e}_i) \leq f(\mathbf{x}) + \lambda \nabla_i f(\mathbf{x}) + \frac{L_i \lambda^2}{2}, \quad \forall \lambda \in \mathbb{R}.$$

逐坐标 $\mathcal{L}$ -smooth

L -smooth 函数一定是逐坐标 $\mathcal{L}$ -smooth, 后者提供更具体的函数光滑性信息.

- $f(x_1, x_2) = x_1^2 + 10x_2^2$ 为 L -smooth, $L = 20$. 同时 f 为逐坐标 $\mathcal{L}$ -smooth, $\mathcal{L} = (2, 20)$.

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

逐坐标 $\mathcal{L}$ -smooth

L -smooth 函数一定是逐坐标 $\mathcal{L}$ -smooth, 后者提供更具体的函数光滑性信息.

- $f(x_1, x_2) = x_1^2 + 10x_2^2$ 为 L -smooth, $L = 20$. 同时 f 为逐坐标 $\mathcal{L}$ -smooth, $\mathcal{L} = (2, 20)$.
- $f(x_1, x_2) = x_1^2 + x_2^2 + Mx_1x_2$ ($M > 0$), 容易知道 f 为逐坐标 $\mathcal{L}$ -smooth, $\mathcal{L} = (2, 2)$. 同时, f 为 L -smooth, 可直接计算得到 $L = M + 2$. 这个例子说明 L -smooth 函数在考虑逐坐标 $\mathcal{L}$ -smooth 时, $\max L_i$ 可以比 L 小.

坐标下降法

坐标下降法在迭代过程包含两个步骤

- 1 从 $[d] := \{1, 2, \dots, d\}$ 中选择坐标 i
- 2 $\mathbf{x}_{k+1} = \mathbf{x}_k - h_i \nabla_i f(\mathbf{x}_k) \mathbf{e}_i$

其中 h_i 表示更新第 i 个分量的步长, $\nabla_i f(\mathbf{x}_k)$ 为梯度 $\nabla f(\mathbf{x}_k)$ 的第 i 个分量.

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

逐坐标充分下降

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

引理 6

若函数 f 逐坐标 $\mathcal{L}$ -smooth, 参数 $\mathcal{L} = (L_1, \dots, L_d)$, 坐标下降法在更新第 i 个坐标分量时取步长 $h_i = \frac{1}{L_i}$, 则有

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k))^2$$

随机坐标下降法

- 1 从 $[d] := \{1, 2, \dots, d\}$ 中等概率地选择坐标 i
- 2 $\mathbf{x}_{k+1} = \mathbf{x}_k - h_i \nabla_i f(\mathbf{x}_k) \mathbf{e}_i$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

随机坐标下降法

定理 7

设函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微且有全局最小值点 $\mathbf{x}^*$. 若 f 逐坐标 L -smooth 且满足以 μ 为参数的 PL 不等式. 以 $\mathbf{x}_0$ 为初始点, 取步长 $h_i = \frac{1}{L}$, 随机坐标下降法有

$$\mathbb{E}[f(\mathbf{x}_K) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \quad K > 0.$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

证明

由充分下降引理有

$$\frac{1}{2L} (\nabla_i f(\mathbf{x}_k))^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$$

在给定 $\mathbf{x}_k$ 条件下, 不等式两端关于 i 求期望

$$\frac{1}{2L} \mathbb{E}_i \left[(\nabla_i f(\mathbf{x}_k))^2 | \mathbf{x}_k \right] \leq \mathbb{E}_i [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k] = \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k].$$

即

$$\frac{1}{2dL} \|\nabla f(\mathbf{x}_k)\|^2 = \frac{1}{2L} \sum_{i=1}^d \frac{1}{d} (\nabla_i f(\mathbf{x}_k))^2 \leq \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k].$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

$$\frac{1}{2dL} \|\nabla f(\mathbf{x}_k)\|^2 = \frac{1}{2L} \sum_{i=1}^d \frac{1}{d} (\nabla_i f(\mathbf{x}_k))^2 \leq \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k].$$

结合 PL 不等式

$$\frac{1}{2\mu} \|\nabla f(\mathbf{x}_k)\|^2 \geq f(\mathbf{x}_k) - f(\mathbf{x}^*),$$

整理得到

$$\mathbb{E} [f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) | \mathbf{x}_k] \leq \left(1 - \frac{\mu}{dL}\right) (f(\mathbf{x}_k) - f(\mathbf{x}^*))$$

不等式两端关于 $\mathbf{x}_k$ 求期望, 由期望的塔式法则

$$\mathbb{E} [f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\mu}{dL}\right) \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}^*)]$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

$$\mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\mu}{dL}\right) \mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)]$$

利用上述递推关系得到

$$\mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\mu}{dL}\right)^{k+1} (f(\mathbf{x}_0) - f(\mathbf{x}^*)).$$

重要性采样

在 L_i 不一致的情形下

① 从 $[d] := \{1, 2, \dots, d\}$ 中依概率 $\frac{L_i}{\sum_{j=1}^d L_j}$ 地选择坐标 i

② $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L_i} \nabla_i f(\mathbf{x}_k) \mathbf{e}_i$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

重要性采样

定理 8

设函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微且有全局最小值点 $\mathbf{x}^*$. 若 f 逐坐标 $\mathcal{L}$ -smooth 且满足以 μ 为参数的 PL 不等式. 其中 $\mathcal{L} = (L_1, \dots, L_d)$, 令 $\bar{L} = \sum_{i=1}^d L_i/d$. 以 $\mathbf{x}_0$ 为初始点, 取步长 $h_i = \frac{1}{L_i}$, 基于重要性采样的随机坐标下降法满足

$$\mathbb{E}[f(\mathbf{x}_K) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^K (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \quad K > 0.$$

证明.

由充分下降引理

$$\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}).$$

在给定 $\mathbf{x}_k$ 条件下, 不等式两端同时关于 i 求期望

$$\mathbb{E}_i \left[\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \right] \leq \mathbb{E}_i [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k] = \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k]$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

证明.

由充分下降引理

$$\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}).$$

在给定 $\mathbf{x}_k$ 条件下, 不等式两端同时关于 i 求期望

$$\mathbb{E}_i \left[\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \right] \leq \mathbb{E}_i [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k] = \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k]$$

$$\mathbb{E}_i \left[\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \right] = \sum_{i=1}^d \frac{1}{2L_i} \frac{L_i}{d\bar{L}} (\nabla_i f(\mathbf{x}_k))^2 = \frac{1}{2d\bar{L}} \|\nabla f(\mathbf{x}_k)\|^2.$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

证明.

由充分下降引理

$$\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}).$$

在给定 $\mathbf{x}_k$ 条件下, 不等式两端同时关于 i 求期望

$$\mathbb{E}_i \left[\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \right] \leq \mathbb{E}_i [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k] = \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) | \mathbf{x}_k]$$

$$\mathbb{E}_i \left[\frac{1}{2L_i} (\nabla_i f(\mathbf{x}_k)) \right] = \sum_{i=1}^d \frac{1}{2L_i} \frac{L_i}{d\bar{L}} (\nabla_i f(\mathbf{x}_k))^2 = \frac{1}{2d\bar{L}} \|\nabla f(\mathbf{x}_k)\|^2.$$

由 PL 不等式

$$\|\nabla f(\mathbf{x}_k)\|^2 \geq 2\mu (f(\mathbf{x}_k) - f(\mathbf{x}^*))$$



Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

续.

整理得到

$$\mathbb{E} [f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) | \mathbf{x}_k] \leq \left(1 - \frac{\mu}{d\bar{L}}\right) (f(\mathbf{x}_k) - f(\mathbf{x}^*))$$

不等式两端同时关于 $\mathbf{x}_k$ 求期望, 得到递推关系

$$\mathbb{E} [f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\mu}{d\bar{L}}\right) \mathbb{E} [f(\mathbf{x}_k) - f(\mathbf{x}^*)]$$



最速坐标下降法

- 1 选择坐标 i 使得 $i = \arg \max_{j \in [d]} |\nabla_j f(\mathbf{x}_k)|$
- 2 $\mathbf{x}_{k+1} = \mathbf{x}_k - h_i \nabla_i f(\mathbf{x}_k) \mathbf{e}_i$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

定理 9

设函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微且有全局最小值点 $\mathbf{x}^*$. 若 f 逐坐标 L -smooth 且满足以 μ 为参数的 PL 不等式. 以 $\mathbf{x}_0$ 为初始点, 取步长 $h_i = \frac{1}{L}$, 最速坐标下降法有

$$f(\mathbf{x}_K) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \quad K > 0.$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

定理 9

设函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微且有全局最小值点 $\mathbf{x}^*$. 若 f 逐坐标 L -smooth 且满足以 μ 为参数的 PL 不等式. 以 $\mathbf{x}_0$ 为初始点, 取步长 $h_i = \frac{1}{L}$, 最速坐标下降法有

$$f(\mathbf{x}_K) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \quad K > 0.$$

证明.

利用不等式 $\max_{i \in [d]} |\nabla_i f(\mathbf{x})|^2 \geq \frac{1}{d} \sum_i |\nabla_i f(\mathbf{x})|^2$, 余下部分与之前证明相同. □

算法复杂度分析

迭代次数 K 与误差 ϵ 的关系

- 定理 4: 梯度下降法 $K = \mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$
- 定理 9: 最速坐标下降法 $K = \mathcal{O}\left(\frac{Ld}{\mu} \log \frac{1}{\epsilon}\right)$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

算法复杂度分析

迭代次数 K 与误差 ϵ 的关系

- 定理 4: 梯度下降法 $K = \mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$
- 定理 9: 最速坐标下降法 $K = \mathcal{O}\left(\frac{Ld}{\mu} \log \frac{1}{\epsilon}\right)$

算法复杂度的公平比较

- 梯度下降法一次迭代需要计算 d 次微分, 故在公平比较的情形下, 梯度下降法的计算复杂度为 $d \cdot \mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\epsilon}\right) = \mathcal{O}\left(\frac{Ld}{\mu} \log \frac{1}{\epsilon}\right)$
- 最速坐标下降法一次迭代同样需要计算 d 次微分, 故计算复杂度为 $d \cdot \mathcal{O}\left(\frac{Ld}{\mu} \log \frac{1}{\epsilon}\right) = \mathcal{O}\left(\frac{Ld^2}{\mu} \log \frac{1}{\epsilon}\right)$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Q: 坐标下降法中算法复杂度 $\mathcal{O}\left(\frac{Ld}{\mu} \log \frac{1}{\epsilon}\right)$ 的 d 能否被消除?

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

定义 10 (ℓ_∞ -Polyak-Lojasiewicz 不等式)

设 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 为可微函数, $\mathbf{x}^*$ 为 f 的全局最小值点, 若存在 $\mu > 0$, 使得对任意 $\mathbf{x} \in \mathbb{R}^d$, 有

$$\frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_\infty^2 \geq f(\mathbf{x}) - f(\mathbf{x}^*)$$

则称 f 满足 ℓ_∞ -Polyak-Lojasiewicz(PL) 不等式.

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

引理 11

设 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 为可微函数, 若存在 $\mu > 0$, 使得 $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_1^2$$

则 f 满足 ℓ_∞ -PL 不等式.

证明.

因为 $\forall \mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_1 \geq \|\mathbf{u}\|_2$, 故

$$\begin{aligned} f(\mathbf{y}) &\geq \underbrace{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_1^2}_{g(\mathbf{y})} \\ &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2. \end{aligned}$$

即 f 为 μ -强凸函数, 因此 f 存在唯一全局最小值点 $\mathbf{x}^*$.

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

证明.

因为 $\forall \mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_1 \geq \|\mathbf{u}\|_2$, 故

$$\begin{aligned}
 f(\mathbf{y}) &\geq \underbrace{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_1^2}_{g(\mathbf{y})} \\
 &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2.
 \end{aligned}$$

即 f 为 μ -强凸函数, 因此 f 存在唯一全局最小值点 $\mathbf{x}^*$.记 $i = \arg \max_{j \in [d]} |\nabla_j f(\mathbf{x})|$, 取 $\bar{\mathbf{y}} = \mathbf{x} - \frac{1}{\mu} \nabla_i f(\mathbf{x}) \mathbf{e}_i$, 有 $\mathbf{0} \in \partial g(\bar{\mathbf{y}})$, 故由最优性条件

$$\begin{aligned}
 g(\mathbf{y}) &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_1^2 \\
 &\geq g(\bar{\mathbf{y}}) = f(\mathbf{x}) - \frac{1}{\mu} (\nabla_i f(\mathbf{x}))^2 + \frac{1}{2\mu} (\nabla_i f(\mathbf{x}_k))^2
 \end{aligned}$$

□

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

续.

$$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_1^2 \geq f(\mathbf{x}) - \frac{1}{\mu} (\nabla_i f(\mathbf{x}))^2 + \frac{1}{2\mu} (\nabla_i f(\mathbf{x}_k))^2$$

故

$$\min_{\mathbf{y}} f(\mathbf{y}) = f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2\mu} (\nabla_i f(\mathbf{x}_k))^2$$

因为 $i = \arg\max_{j \in [d]} |\nabla_j f(\mathbf{x})|$, 故 $(\nabla_i f(\mathbf{x}_k))^2 = \|\nabla f(\mathbf{x})\|_\infty^2$.

□

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

定理 12

设函数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 可微且有全局最小值点 $\mathbf{x}^*$. 若 f 逐坐标 L -smooth 且满足以 μ 为参数的 ℓ_∞ -PL 不等式. 以 $\mathbf{x}_0$ 为初始点, 取步长 $h_i = \frac{1}{L}$, 最速坐标下降法有

$$f(\mathbf{x}_K) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{L}\right)^T (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \quad K > 0.$$

证明.

由充分下降引理

$$\frac{1}{2L} (\nabla_i f(\mathbf{x}_k))^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}),$$

其中 $i = \arg\max_{j \in [d]} |\nabla_j f(\mathbf{x}_k)|$, 因此

$$(\nabla_i f(\mathbf{x}_k))^2 = \|\nabla f(\mathbf{x}_k)\|_\infty^2.$$

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

证明.

由充分下降引理

$$\frac{1}{2L} (\nabla_i f(\mathbf{x}_k))^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}),$$

其中 $i = \arg\max_{j \in [d]} |\nabla_j f(\mathbf{x}_k)|$, 因此

$$(\nabla_i f(\mathbf{x}_k))^2 = \|\nabla f(\mathbf{x}_k)\|_\infty^2.$$

由 ℓ_∞ -PL 不等式

$$\|\nabla f(\mathbf{x}_k)\|_\infty^2 \geq 2\mu (f(\mathbf{x}_k) - f(\mathbf{x}^*)).$$

从而

$$\frac{\mu}{L} (f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}),$$

□

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

ℓ_∞ -PL 不等式

Nesterov 加速梯度下降

Polyak-Lojasiewicz 不等式

坐标下降法

随机坐标下降法

重要性采样

最速坐标下降法

 ℓ_∞ -PL 不等式

续.

$$\frac{\mu}{L} (f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}),$$

整理得到递推关系

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{L}\right) (f(\mathbf{x}_k) - f(\mathbf{x}^*))$$

