

最优化方法 第十三讲

Weiwen Wang(王伟文)

暨南大学

2025 年 6 月 16 日

目录

- 1 非欧氏梯度下降
- 2 非欧氏临近梯度法

梯度下降法回顾

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

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

其中

$$\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|_2^2$$

度量空间 $(\mathbb{R}^d, \|\cdot\|_2)$

非欧空间的梯度更新

考虑函数 $f: \mathbb{E} \rightarrow \mathbb{R}$,

$$\min_{\mathbf{x} \in \mathbb{E}} f(\mathbf{x})$$

梯度更新

$$\mathbf{x}_{k+1} = \underbrace{\mathbf{x}_k}_{\in \mathbb{E}} - \underbrace{\gamma_k \nabla f(\mathbf{x}_k)}_{\in \mathbb{E}^*}$$

度量空间 $(\mathbb{R}^d, \|\cdot\|)$, 其中 $\|\cdot\|$ 为定义在 $\mathbb{E}$ 上的范数.

$\nabla f(\mathbf{x}_k)$ 在 $\mathbb{E}$ 的对项

设 $\mathbf{a} \in \mathbb{E}^*$, 定义

$$\Lambda_{\mathbf{a}} = \underset{\|\mathbf{x}\| \leq 1}{\operatorname{argmax}} \langle \mathbf{a}, \mathbf{x} \rangle$$

则 $\nabla f(\mathbf{x}_k)$ 在 $\mathbb{E}$ 的对项定义为 $\nabla f(\mathbf{x}^k)^\dagger$

$$\nabla f(\mathbf{x}^k)^\dagger \in \Lambda_{\nabla f(\mathbf{x}_k)}$$

修正后的梯度更新

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \underbrace{\gamma_k \|\nabla f(\mathbf{x}^k)\|_* \cdot \nabla f(\mathbf{x}^k)^\dagger}_{\text{scaled}}$$

引理 1 ($\Lambda_{\mathbf{a}}$ 的性质)

设 $\mathbf{a} \in \mathbb{E}^*$, 则

- (a) 若 $\mathbf{a} \neq \mathbf{0}$, 则 $\|\mathbf{a}^\dagger\| = 1, \forall \mathbf{a}^\dagger \in \Lambda_{\mathbf{a}}$;
- (b) 若 $\mathbf{a} = \mathbf{0}$, 则 $\Lambda_{\mathbf{a}} = \{\mathbf{x} \in \mathbb{E} : \|\mathbf{x}\| \leq 1\}$;
- (c) $\forall \mathbf{a}^\dagger \in \Lambda_{\mathbf{a}}, \langle \mathbf{a}, \mathbf{a}^\dagger \rangle = \|\mathbf{a}\|_*$;
- (d) $\Lambda_{\mathbf{a}} = \partial \|\mathbf{a}\|_*$.

证明.

(d)

$$\begin{aligned}
 \Lambda_{\mathbf{a}} &= \arg \max_{\|\mathbf{x}\| \leq 1} \langle \mathbf{x}, \mathbf{a} \rangle \\
 &= \arg \max_{\mathbf{x} \in \mathbb{E}} \left\{ \langle \mathbf{x}, \mathbf{a} \rangle - \delta_{B_{\|\cdot\|}[\mathbf{0}, 1]}(\mathbf{x}) \right\} \\
 &= \partial \left(\delta_{B_{\|\cdot\|}[\mathbf{0}, 1]} \right)^* (\mathbf{y})
 \end{aligned}$$

由因为

$$\begin{aligned}
 \left(\delta_{B_{\|\cdot\|}[\mathbf{0}, 1]} \right)^* &= \max_{\mathbf{x} \in \mathbb{E}} \left\{ \langle \mathbf{x}, \mathbf{a} \rangle - \delta_{B_{\|\cdot\|}[\mathbf{0}, 1]}(\mathbf{x}) \right\} \\
 &= \max_{\|\mathbf{x}\| \leq 1} \langle \mathbf{x}, \mathbf{a} \rangle \\
 &= \|\mathbf{a}\|_*
 \end{aligned}$$



例子

若 $\mathbf{a} \neq \mathbf{0}$ 且 $\|\cdot\| \triangleq \|\cdot\|_2$, 则

$$\Lambda_{\mathbf{a}} = \left\{ \frac{\mathbf{a}}{\|\mathbf{a}\|} \right\}.$$

考虑度量空间 $(\mathbb{R}^n, \|\cdot\|_1)$, 其对偶度量空间为 $(\mathbb{R}^n, \|\cdot\|_\infty)$ 若 $\mathbf{a} \neq 0$, 则

$$\Lambda_{\mathbf{a}} = \partial \|\mathbf{a}\|_\infty = \left\{ \sum_{i \in I(\mathbf{a})} \lambda_i \text{sgn}(a_i) \mathbf{e}_i, \lambda \in \Delta_{|I(\mathbf{a})|} \right\}$$

其中 $I(\mathbf{a}) = \{k \in [d] : |a_k| = \|\mathbf{a}\|_\infty\}$

定理 2 (非欧氏梯度下降的充分下降引理)

设 $f: \mathbb{E} \rightarrow \mathbb{R}$ 为 $\|\cdot\|$ 下的 L -smooth 函数, 则有

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \gamma_k \left(1 - \frac{L\gamma_k}{2}\right) \|\nabla f(\mathbf{x}_k)\|_*^2$$

证明.

由 L -smooth 性质

$$\begin{aligned}
 f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
 &= f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \gamma_k \|\nabla f(\mathbf{x}_k)\|_* \nabla f(\mathbf{x}_k^\dagger) \rangle + \frac{L\gamma_k^2}{2} \|\nabla f(\mathbf{x}_k)\|^2 \|\nabla f(\mathbf{x}_k)^\dagger\|_*^2 \\
 &\leq f(\mathbf{x}_k) - \gamma_k \|\nabla f(\mathbf{x}_k)\|_*^2 + \frac{L\gamma_k^2}{2} \|\nabla f(\mathbf{x}_k)\|_*^2 \quad (\nabla f(\mathbf{x}_k)^\dagger \text{ 的最优性})
 \end{aligned}$$

移项整理得到

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \gamma_k \left(1 - \frac{L\gamma_k}{2}\right) \|\nabla f(\mathbf{x}_k)\|_*^2$$

□

假设 3

- $f: \mathbb{E} \rightarrow \mathbb{R}$ 为 L -smooth 的凸函数;
- 优化问题的最优解集非空;
- $D = \max_{\mathbf{x}, \mathbf{y} \in \mathbb{E}} \|\mathbf{x} - \mathbf{y}\|$

引理 4

设 f 满足假设 3, $\mathbf{x}^*$ 为最小化问题的最优解, 最优值记为 f^* , 取非欧式梯度下降法的步长 $\gamma_k = \frac{1}{L}$, 则迭代序列满足

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2LD^2} (f(\mathbf{x}_k) - f^*)^2$$

证明.

由充分下降引理

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

由凸函数性质及对偶范数定义

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle \leq \|\nabla f(\mathbf{x}_k)\|_* \|\mathbf{x}_k - \mathbf{x}^*\|$$

因为 $f(\mathbf{x}_k) \geq f(\mathbf{x}^*)$,

$$(f(\mathbf{x}_k) - f(\mathbf{x}^*))^2 \leq 2LD^2(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}))$$



引理 5

设 $\{a_k\}_{k \geq 0}$ 为非负实数序列, 若

$$a_k - a_{k+1} \geq \frac{1}{\gamma} a_k^2 \quad \forall k \geq 0,$$

则

$$a_k \leq \frac{\gamma}{k} \quad \forall k \geq 1.$$

定理 6

设 f 满足假设 3, $\mathbf{x}^*$ 为最小化问题的最优解, 最优值记为 f^* , 取非欧氏梯度下降法的步长 $\gamma_k = \frac{1}{L}$, 则迭代序列满足

$$f(\mathbf{x}_k) - f^* \leq \frac{2LD^2}{k} \quad \forall k \geq 1.$$

证明.

记 $\Delta_k = f(\mathbf{x}_k) - f^* \geq 0$. 由引理 4

$$\Delta_k - \Delta_{k+1} \geq \frac{1}{2LD^2} \Delta_k^2$$

由引理 5

$$f(\mathbf{x}_k) - f^* = \Delta_k \leq \frac{2LD^2}{k} \quad \forall k \geq 1.$$



在假设 3 下,

- 非欧氏梯度下降: $f(\mathbf{x}_k) - f^* \leq \frac{2LD^2}{k}$;
- 欧氏梯度下降: $f(\mathbf{x}_k) - f^* \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2k}$

例子

考虑最小化问题

$$\min_{\mathbf{x}} f(\mathbf{x}) \triangleq \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$$

其中 $\mathbf{A} \in \mathbb{S}_{++}^n$, $\mathbf{b} \in \mathbb{R}^n$.

$f(\mathbf{x})$ 为 L -smooth 且

$$L \triangleq \|\mathbf{A}\|_{p,q} = \max\{\|\mathbf{A}\mathbf{x}\|_q : \|\mathbf{x}\|_p \leq 1\}$$

p 和 q 满足 $\frac{1}{p} + \frac{1}{q} = 1$ 且 $p, q \geq 0$.

$$\nabla f(\mathbf{x}) = \mathbf{Ax} + \mathbf{b} = \left(\mathbf{a}_i^T \mathbf{x} + b_i \right)_{i \in [n]}$$

- 若 $p = q = 2$, $L_2 = \max |\lambda(\mathbf{A})|$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L_2} (\mathbf{Ax}_k + \mathbf{b})$$

- 若 $p = 1, q = \infty, L_\infty = \max |\mathbf{A}_{i,j}|$

$$\begin{aligned}
 i_k &\in \operatorname{argmax}_{i \in [n]} |\mathbf{a}_i^T \mathbf{x} + b_i| \\
 \mathbf{x}_{k+1} &= \mathbf{x}_k - \frac{|\mathbf{a}_{i_k}^T \mathbf{x} + b_{i_k}|}{L_\infty} \operatorname{sgn}(\mathbf{a}_{i_k}^T \mathbf{x} + b_{i_k}) \mathbf{e}_{i_k} \\
 &= \mathbf{x}_k - \frac{1}{L_\infty} (\mathbf{a}_{i_k}^T \mathbf{x} + b_{i_k}) \mathbf{e}_{i_k}
 \end{aligned}$$

- 收敛速度快慢一定程度上取决于 $\frac{L_2}{L_\infty}$.

梯度更新

- "+" 操作数 $\mathcal{O}(n^2)$

$$\mathbf{g}_{k+1} = \mathbf{A}\mathbf{x}_{k+1} + \mathbf{b}$$

- "+" 操作数 $\mathcal{O}(n^2)$

$$\begin{aligned}\mathbf{g}_{k+1} &= \mathbf{g}_k - \frac{1}{L_\infty} (\mathbf{a}_{i_k}^T \mathbf{x} + b_{i_k}) \mathbf{A} \mathbf{e}_{i_k} \\ &= \mathbf{g}_k - \frac{1}{L_\infty} (\mathbf{a}_{i_k}^T \mathbf{x} + b_{i_k}) \mathbf{a}_{i_k}\end{aligned}$$

考虑最小化问题

$$\min_{\mathbf{x} \in \mathbb{E}} F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})$$

临近梯度法

$$\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x} \in E} \left\{ f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + g(\mathbf{x}) + \frac{1}{2\gamma_k} \|\mathbf{x} - \mathbf{x}_{k+1}\|^2 \right\}$$

Bregman 散度

$$\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x} \in E} \left\{ f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + g(\mathbf{x}) + \frac{1}{\gamma_k} B_\omega(\mathbf{x}, \mathbf{x}_k) \right\}$$

假设 7

- $g: \mathbb{E} \rightarrow \mathbb{R}$ 为凸函数
- $f: \mathbb{E} \rightarrow \mathbb{R}$ 为 L -smooth 凸函数;
- 最优解集 X^* 非空.

假设 8

- $\omega: \mathbb{E} \rightarrow \mathbb{R}$ 为可微的 1-强凸函数.

迭代格式

$$\begin{aligned}\mathbf{x}_{k+1} &= \arg \min_{\mathbf{x} \in E} \left\{ f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + g(\mathbf{x}) + \frac{1}{\gamma_k} B_\omega(\mathbf{x}, \mathbf{x}_k) \right\} \\ &= \arg \min_{\mathbf{x} \in E} \left\{ \langle \gamma_k \nabla f(\mathbf{x}_k) - \nabla \omega(\mathbf{x}_k), \mathbf{x} \rangle + \gamma_k g(\mathbf{x}) + \omega(\mathbf{x}) \right\}\end{aligned}$$

上述迭代格式是良态的, 即 $\mathbf{x}_{k+1}$ 存在且唯一.

收敛性分析

定理 9

设 $\mathbf{x}^*$ 为复合最小化问题的最优解, 记最优值为 F^* , 取 $\gamma_k = \frac{1}{L_f}$. 非欧临近梯度法的迭代序列 $\{\mathbf{x}_k\}_{k \geq 0}$ 满足

(a) $\{F(\mathbf{x}_k)\}_{k \geq 0}$ 单调非增;

(b)

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) \leq \frac{LB_\omega(\mathbf{x}^*, \mathbf{x}_0)}{k} \quad \forall k \geq 1.$$

证明.

(a) 由最优性条件

$$0 \in \gamma_k \nabla f(\mathbf{x}_k) - \nabla \omega(\mathbf{x}_k) + \gamma_k \partial g(\mathbf{x}_{k+1}) + \nabla \omega(\mathbf{x}_{k+1})$$

即

$$\frac{1}{\gamma_k} (\nabla \omega(\mathbf{x}_k) - \nabla \omega(\mathbf{x}_{k+1})) - \nabla f(\mathbf{x}_k) \in \partial g(\mathbf{x}_{k+1})$$

故

$$\begin{aligned} g(\mathbf{x}_k) &\geq g(\mathbf{x}_{k+1}) + \left\langle \frac{1}{\gamma_k} (\nabla \omega(\mathbf{x}_k) - \nabla \omega(\mathbf{x}_{k+1})) - \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_{k+1} \right\rangle \\ &\geq g(\mathbf{x}_{k+1}) + \frac{1}{\gamma_k} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 - \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_{k+1} \rangle \end{aligned}$$



续.

移项得到

$$g(\mathbf{x}_{k+1}) \leq g(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_{k+1} \rangle - \frac{1}{\gamma_k} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2$$

由 L -smooth 得到

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

两不等式相加得到

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \left(\frac{1}{\gamma_k} - \frac{L}{2} \right) \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2$$

若 $\gamma_k \in (0, \frac{2}{L})$, 序列 $\{F(\mathbf{x}_k)\}_{k \geq 0}$ 单调非增.

续.

(b) 由 (a)

$$\frac{1}{\gamma_k} (\nabla \omega(\mathbf{x}_k) - \nabla \omega(\mathbf{x}_{k+1})) - \nabla f(\mathbf{x}_k) \in \partial g(\mathbf{x}_{k+1})$$

故

$$\begin{aligned} g(\mathbf{x}^*) &\geq g(\mathbf{x}_{k+1}) + \left\langle \frac{1}{\gamma_k} (\nabla \omega(\mathbf{x}_k) - \nabla \omega(\mathbf{x}_{k+1})) - \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_{k+1} \right\rangle \\ &= g(\mathbf{x}_{k+1}) + L \langle \nabla \omega(\mathbf{x}_k) - \nabla \omega(\mathbf{x}_{k+1}), \mathbf{x}^* - \mathbf{x}_{k+1} \rangle - \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k + \mathbf{x}_k - \mathbf{x}_{k+1} \rangle \\ &= g(\mathbf{x}_{k+1}) + LB_\omega(\mathbf{x}^*, \mathbf{x}_{k+1}) + LB_\omega(\mathbf{x}_{k+1}, \mathbf{x}_k) - LB_\omega(\mathbf{x}^*, \mathbf{x}_k) \\ &\quad - \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle - \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_{k+1} \rangle \end{aligned}$$



续.

因此

$$\begin{aligned}
 & LB_{\omega}(\mathbf{x}^*, \mathbf{x}_k) - LB_{\omega}(\mathbf{x}^*, \mathbf{x}_{k+1}) + g(\mathbf{x}^*) + f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle \\
 & \geq g(\mathbf{x}_{k+1}) + f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + LB_{\omega}(\mathbf{x}_{k+1}, \mathbf{x}_k) \\
 & \stackrel{(i)}{\geq} g(\mathbf{x}_{k+1}) + f(\mathbf{x}_{k+1})
 \end{aligned}$$

即

$$F(\mathbf{x}_{k+1}) - F(\mathbf{x}^*) \leq LB_{\omega}(\mathbf{x}^*, \mathbf{x}_k) - LB_{\omega}(\mathbf{x}^*, \mathbf{x}_{k+1})$$

其中不等式 (i) 源于 L -smooth

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

且 $B_{\omega}(\mathbf{x}^*, \mathbf{x}_{k+1}) \geq \frac{1}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$.