最优化方法　第十四讲

Weiwen Wang(王伟文)

暨南大学

2025 年 6 月 25 日

目录

有限和目标函数

考虑优化问题

$$\min_{\boldsymbol{x}} \ f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}),$$

其中 $f_i : \mathbb{R}^d \to \mathbb{R}$ 可微.

- 梯度下降法 (Gradient Descent, GD)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\boldsymbol{x})$$

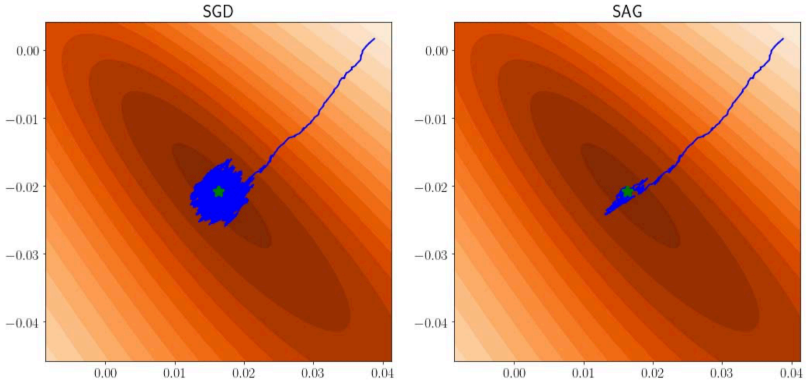
- 小批量梯度下降法 (Min-Batch, MBGD)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla f_i(\boldsymbol{x}) \quad \mathcal{B} \subseteq \{1, 2, \ldots, n\}.$$

- 随机梯度下降法 (Stochastic Gradient Descent, SGD)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k \nabla f_{i_k}(\boldsymbol{x}_k) \quad i_k \sim \texttt{Unif}(1, 2, \ldots, n)$$

SGD 不收敛: $\lim_{k\to\infty} \nabla f(\boldsymbol{x}_k) \neq 0$



Gower, Robert M., et al. "Variance-reduced methods for machine learning."
Proceedings of the IEEE 108.11 (2020): 1968-1983.

引理 1

设 $X \in R^d$ 为随机向量存在有限方差, 则有

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \mathbb{E}[\|X\|^2]$$

证明.

$$\begin{aligned}\mathbb{E}\left[\|X-\mathbb{E}X\|^2\right] &= \mathbb{E}\left[\|X\|^2 - 2\langle X, \mathbb{E}X\rangle + \|\mathbb{E}X\|^2\right] \\ &= \mathbb{E}[\|X\|^2] - \|\mathbb{E}X\|^2 \le \mathbb{E}[\|X\|^2].\end{aligned}$$

□

引理 2

设 f 为 L-*smooth* 凸函数, 则有

$$f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{1}{2L}\|\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y})\|^2\leq f(\boldsymbol{y}).$$

定义 3 (方差缩减性质 (Variance Reduction Property, VRP))

设随机序列 $\{\boldsymbol{g}_k\}_{k\geq 0}\subset\mathbb{R}^d$, 若有

$$\mathbb{E}\left[\|\boldsymbol{g}_k-\nabla f(\boldsymbol{x}_k)\|^2\right]\to 0 \quad \text{当} \quad k\to\infty$$

则称随机序列 $\{\boldsymbol{g}_k\}_{k\geq 0}$ 满足方差缩减性质.

如何构造满足方差缩减性质的随机序列 $\{\boldsymbol{g}_k\}_{k\geq 0}$?

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

原型方法-SGD$_*$

假设已知 $\boldsymbol{x}^*$ 为有限和优化问题的最优解, 则定义

$$\boldsymbol{g}_k = \nabla f_{i_k}(\boldsymbol{x}_k) - \nabla f_{i_k}(\boldsymbol{x}^*) \quad i_k \sim \texttt{Unif}(1,2,\ldots,n).$$

$\boldsymbol{g}_k$ 是 $\nabla f(\boldsymbol{x}_k)$ 的条件无偏估计.

$$\begin{aligned}\mathbb{E}[\boldsymbol{g}_k|\boldsymbol{x}_k] &= \mathbb{E}\left[\nabla f_{i_k}(\boldsymbol{x}_k) - \nabla f_{i_k}(\boldsymbol{x}^*)|\boldsymbol{x}_k\right] \\ &= \sum_{i=1}^{n} \frac{1}{n}\left(f_i(\boldsymbol{x}_k) - f_i(\boldsymbol{x}^*)\right) \\ &= \frac{1}{n}\sum_{i=1} f_i(\boldsymbol{x}_k) - \frac{1}{n}\sum_{i=1} f_i(\boldsymbol{x}^*) \\ &= \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}_k).\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|^2 | \boldsymbol{x}_k\right] &\overset{(i)}{\leq} \mathbb{E}\left[\|\boldsymbol{g}_k\|^2 | \boldsymbol{x}_k\right] \\ &\leq \mathbb{E}\left[\|\nabla f_{i_k}(\boldsymbol{x}_k) - \nabla f_{i_k}(\boldsymbol{x}^*)\|^2 | \boldsymbol{x}_k\right]\end{aligned}$$

故

$$\mathbb{E}\left[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|^2\right] \leq \mathbb{E}\left[\|\nabla f_{i_k}(\boldsymbol{x}_k) - \nabla f_{i_k}(\boldsymbol{x}^*)\|^2\right]$$

若 $\boldsymbol{x}_k \to \boldsymbol{x}_*$, 则 $\{\boldsymbol{g}_k\}$ 满足 VRP.

引理 4

设 f_i 为 L_i-*smooth* 的凸函数, 记 $L_{\max} = \max\limits_{i\in[n]} L_i$. 若 $i_k \sim \texttt{Unif}\{1,2,\ldots,n\}$, 则

$$\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(\boldsymbol{x}) - \nabla f_{i_k}(\boldsymbol{x}^*)\|^2\right] \le 2L_{\max}(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)) \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

证明.

由引理 2

$$\|\nabla f_{i_k}(\boldsymbol{x}) - \nabla f_{i_k}(\boldsymbol{x}^*)\|^2 \le 2L_{\max}\left(f_{i_k}(\boldsymbol{x}) - f_{i_k}(\boldsymbol{x}^*) - \langle \nabla f_{i_k}(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle\right) \quad \forall \boldsymbol{x} \in \mathbb{R}^d$$

不等式两侧同时关于 i_k 求期望

$$\begin{aligned}
\mathbb{E}_{i_k}\|\nabla f_{i_k}(\boldsymbol{x}) - \nabla f_{i_k}(\boldsymbol{x}^*)\|^2 &\le \mathbb{E}_{i_k}\left[2L_{\max}\left(f_{i_k}(\boldsymbol{x}) - f_{i_k}(\boldsymbol{x}^*) - \langle \nabla f_{i_k}(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle\right)\right] \\
&= 2L_{\max}\left(\frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x}) - \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x}^*) - \left\langle \frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\right\rangle\right) \\
&= 2L_{\max}(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)).
\end{aligned}$$

□

定理 5

设 f_i 为 L_i-*smooth* 凸函数, f 为可微 μ-强凸函数, 记 $L_{\max} = \max\limits_{i\in[n]} L_i$, 取步长 $\gamma \in (0, 1/L_{\max})$, *SGD*$_*$ 迭代满足

$$\mathbb{E}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right] \le (1-\gamma\mu)\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2\right] \quad \forall k \ge 0$$

和 *VRP.*

证明.

由迭代格式

$$\begin{aligned}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_k-\gamma\boldsymbol{g}_k-\boldsymbol{x}^*\|^2\\ &= \|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-2\langle\gamma\boldsymbol{g}_k,\boldsymbol{x}_k-\boldsymbol{x}^*\rangle+\gamma^2\|\boldsymbol{g}_k\|^2\\ &= \|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-2\gamma\langle\nabla f_{i_k}(\boldsymbol{x}_k)-\nabla f_{i_k}(\boldsymbol{x}^*),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle+\gamma^2\|\nabla f_{i_k}(\boldsymbol{x}_k)-\nabla f_{i_k}(\boldsymbol{x}^*)\|^2\end{aligned}$$

不等式两端在 $\boldsymbol{x}_k$ 条件下关于 i_k 求期望

$$\mathbb{E}_{i_k}\left[\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2|\boldsymbol{x}_k\right]=\|\boldsymbol{x}_k-\boldsymbol{x}_*\|^2-2\gamma\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle+\gamma^2\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(\boldsymbol{x}_k)-\nabla f_{i_k}(\boldsymbol{x}^*)\|^2|\boldsymbol{x}_k\right]$$

□

证明.

f 为 μ-强凸函数

$$\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^* \rangle \geq f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$

由引理 4

$$\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(\boldsymbol{x}_k) - \nabla f_{i_k}(\boldsymbol{x}^*)\|^2 | \boldsymbol{x}_k\right] \leq 2L_{\max}(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*))$$

综合上述不等式

$$\mathbb{E}_{i_k}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 | \boldsymbol{x}_k\right] \leq \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \gamma\mu\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - 2\gamma(1 - \gamma L_{\max})(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*))$$

□

证明.

注意到 $\gamma < 1/L_{\max}$ 且 $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \geq 0$. 故

$$\mathbb{E}_{\boldsymbol{x}_{k+1}}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 | \boldsymbol{x}_k\right] = \mathbb{E}_{i_k}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 | \boldsymbol{x}_k\right] \leq (1 - \mu\gamma)\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$

两端同时关于 $\boldsymbol{x}_k$ 求期望, 由塔公式法则

$$\mathbb{E}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right] \leq (1 - \mu\gamma)\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2\right]$$

□

证明.

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|^2 | \boldsymbol{x}_k\right] &\leq \mathbb{E}\left[\|\nabla f_{i_k}(\boldsymbol{x}_k) - \nabla f_{i_k}(\boldsymbol{x}^*)\|^2 | \boldsymbol{x}_k\right] \\ &\leq \mathbb{E}\left[L_i^2 \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 | \boldsymbol{x}_k\right] \\ &\leq L_{\max}^2 \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2\end{aligned}$$

故

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|^2\right] &\leq L_{\max}^2 \mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2\right] \\ &\leq L_{\max}^2 (1 - \mu\gamma)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2\end{aligned}$$

□

定义辅助变量 $\boldsymbol{v}_k^{(i)}$

$$\boldsymbol{v}_k^{(j)} = \begin{cases} \nabla f_j(\boldsymbol{x}_k), & j = i_k \\ \boldsymbol{v}_{k-1}^{(i)}, & j \neq i_k \end{cases} \qquad i_k \sim \texttt{Unif}(1,2,\ldots,n)$$

初始化为 $\boldsymbol{v}_0^{(i)} = \boldsymbol{0}, \forall i \in [n]$.

$$\begin{aligned} \boldsymbol{g}_k &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_k^{(i)} = \frac{1}{n}\sum_{j;j\neq i_k} \boldsymbol{v}_k^{(j)} + \frac{1}{n}\boldsymbol{v}_k^{(i_k)} = \frac{1}{n}\sum_{j;j\neq i_k} \boldsymbol{v}_{k-1}^{(j)} + \frac{1}{n}\boldsymbol{v}_k^{(i_k)} \\ &= \boldsymbol{g}_{k-1} - \frac{1}{n}\boldsymbol{v}_{k-1}^{(i_k)} + \frac{1}{n}\nabla f_{i_k}(\boldsymbol{x}_k) \end{aligned}$$

$\boldsymbol{g}_k$ 不是 $\nabla f(\boldsymbol{x}_k)$ 的一个无偏估计

$$\begin{aligned}\mathbb{E}[\boldsymbol{g}_k|\boldsymbol{x}_k] &= \mathbb{E}\left[\boldsymbol{g}_{k-1} - \frac{1}{n}\boldsymbol{v}_{k-1}^{(i_k)} + \frac{1}{n}\nabla f_{i_k}(\boldsymbol{x}_k)|\boldsymbol{x}_k\right] \\ &= \nabla f(\boldsymbol{x}_k) + \boldsymbol{g}_{k-1} - \frac{1}{n}\boldsymbol{g}_{k-1}\end{aligned}$$

给定一组向量 $\{\boldsymbol{v}_i\}_{i\in[n]}$,

$$\nabla f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(\boldsymbol{x}) - \boldsymbol{v}_i + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}_i\right)$$

记

$$\begin{aligned}\nabla f_i(\boldsymbol{x};\boldsymbol{v}_i) &= \nabla f_i(\boldsymbol{x}) - \boldsymbol{v}_i + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}_i \\ &= \nabla f_i(\boldsymbol{x}) - \boldsymbol{v}_i + \mathbb{E}[\boldsymbol{v}_i]\end{aligned}$$

定义

$$\boldsymbol{g}_k = \nabla f_{i_k}(\boldsymbol{x}_k, \boldsymbol{v}_{i_k}) = \nabla f_{i_k}(\boldsymbol{x}) - \boldsymbol{v}_{i_k} + \mathbb{E}[\boldsymbol{v}_{i_k}] \quad i_k \sim \texttt{Unif}(1,2,\ldots,n)$$

方差缩减性质

$$\mathbb{E}_{i_k}[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|^2] = \mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(\boldsymbol{x}) - \boldsymbol{v}_{i_k} + \mathbb{E}[\boldsymbol{v}_{i_k}] - \nabla f(\boldsymbol{x}_k)\|^2\right]$$
$$\leq \boldsymbol{E}_{i_k}\left[\|\nabla f_{i_k}(\boldsymbol{x}_k) - \boldsymbol{v}_{i_k}\|^2\right]$$

若 $\nabla f_{i_k}(\boldsymbol{x}_k) \underset{k\to\infty}{\longrightarrow} \boldsymbol{v}_{i_k}$, $\mathbb{E}_{i_k}[\|\boldsymbol{g}_k - \nabla f(\boldsymbol{x}_k)\|^2] \to 0$.

近似梯度

定义

$$\boldsymbol{v}_k^{(j)} = \nabla f_j(\boldsymbol{x}_{\setminus k})$$

其中 $\nabla f_j(\boldsymbol{x}_{\setminus k})$ 表示第 k 次迭代前最近一次计算的 f_j 的梯度.
迭代格式的近似梯度定义为

$$\boldsymbol{g}_k = \nabla f_{i_k}(\boldsymbol{x}_k) - \boldsymbol{v}_k^{(i_k)} + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_k^{(i)} \quad i_k \sim \texttt{Unif}(1,2,\ldots,n)$$

高效计算

记 $\bar{\boldsymbol{g}}_{k-1} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_k^{(i)}$

- $i_k \in \texttt{Unif}(1,2,\ldots,n)$
- $\boldsymbol{g}_k = \nabla f_{i_k}(\boldsymbol{x}_k) - \boldsymbol{v}_k^{(i_k)} + \bar{\boldsymbol{g}}_{k-1}$
- $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k \boldsymbol{g}_k$
- $\bar{\boldsymbol{g}}_k = \bar{\boldsymbol{g}}_{k-1} - \frac{1}{n}\boldsymbol{v}_k^{(i_k)} + \frac{1}{n}\nabla f_{i_k}(\boldsymbol{x}_k)$

初始化取 $\boldsymbol{v}_0^{(i)} = \boldsymbol{0}, \forall i \in [n],\ \bar{\boldsymbol{g}}_0 = 0.$

SAG 和 SAGA 需要维持向量组 $\{\boldsymbol{v}_i\}$, 空间复杂度 $\mathcal{O}(nd)$.

SVRG 定义近似梯度

$$\boldsymbol{g}_k = \nabla f_{i_k}(\boldsymbol{x}_k) - \nabla f_{i_k}(\tilde{\boldsymbol{x}}_k) + \nabla f(\tilde{\boldsymbol{x}}_k)$$

空间复杂度 $\mathcal{O}(d)$.

```python
import numpy as np

def stochastic_variance_reduced_gradient(theta, step_size, sample_num,
    outer_max, inner_max):
    for o in range(out_max):
        # compute full gradient of the finite sum objective function at theta
        full_grad = gradient_of_obj(theta)
        inner_theta = theta
        for i in range(inner_max):
          idx = np.random.randint(0, sample_num)
          # compute individual gradient with index idx at inner_theta
          grad1 = gradient_of_obj(inner_theta, idx)
          # compute individual gradient  with index idx at theta
          grad2 = gradient_of_obj(theta, idx)
          # gradient estimation
          grad = grad1 - grad2 + full_grad
          inner_theta += -step_size * grad

        theta = inner_theta

  return theta
```

定理 6

设 f_i 为 L_i-*smooth* 凸函数, f 为可微 μ-强凸函数, 记 $L_{\max} = \max\limits_{i\in[n]} L_i$, 取步长 $\gamma \in (0, \frac{1}{2L_{\max}})$, 若 *SVRG* 内循环迭代次数 m 充分大使得

$$\rho \triangleq \frac{1}{\mu\gamma(1-2L_{\max}\gamma)m} + \frac{2L_{\max}\gamma}{1-2L_{\max}\gamma} < 1$$

则

$$\mathbb{E}\left[f(\boldsymbol{x}_o) - f(\boldsymbol{x}^*)\right] \le \rho^o (f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)).$$

这里 $\boldsymbol{x}_o = \frac{1}{m}\sum\limits_{i=0}^{m-1} \boldsymbol{x}_{o_i}$, 其中 $\boldsymbol{x}_{o_i}$ 表示第 o 次外循环中的第 i 次内循环更新结果.