基于梯度符号的随机优化

张利军
南京大学

第二十届“中国人工智能基础年会”青年学者论坛

目录

- 研究背景

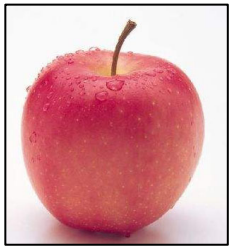
- 加速的梯度符号优化
  - 动量法
  - 方差约减

- 分布式场景

- 总结展望

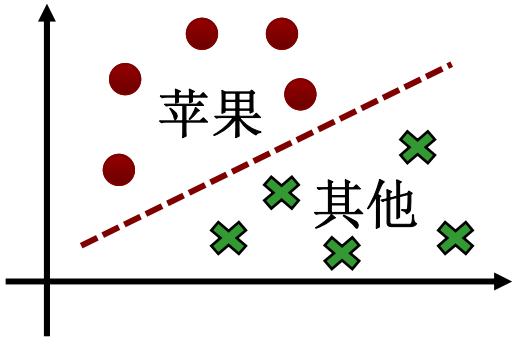
机器学习中的优化问题

➢ 监督学习



苹果

$$(\mathbf{x}_1, y_1)$$
$$\cdots$$
$$(\mathbf{x}_n, y_n) \Rightarrow y \approx h(\mathbf{x})$$



➢ 经验风险最小化

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i) + \phi(h)$$

- n是样本数量、$\mathcal{H}$为假设空间，其维度为d
- $\ell(\cdot,\cdot)$为损失函数、$\phi(\cdot)$为正则化因子

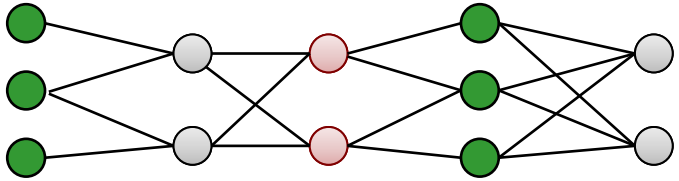
典型学习算法

- 最小二乘回归

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

  - 平方损失：$\ell(u,v) = (u-v)^2$
  - 线性函数空间：$\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}, \mathbf{w} \in \mathbb{R}^d\}$

- 深度学习
  - 交叉熵损失：$\ell(\mathbf{x}, y) = \sum_{c=1}^{K} y_c \log p_c(\mathbf{x})$
  - 函数空间：神经网络



优化算法

- 优化问题

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^{\top}\mathbf{x}_i, y_i)$$

- 确定优化——梯度下降（GD）

1. 计算梯度

$$\nabla F(\mathbf{w}_t) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell\left(\mathbf{w}_t^{\top}\mathbf{x}_i, y_i\right)$$

计算复杂度$O(nd)$

2. 更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla F(\mathbf{w}_t)$$

计算复杂度高×、收敛速率快✓

优化算法

- 优化问题

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^{\top}\mathbf{x}_i, y_i)$$

- 随机优化——随机梯度下降（SGD）

1. 随机采样 $(\mathbf{x}_t, y_t) \sim \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

2. 计算随机梯度 $\mathbf{g}_t = \nabla \ell\left(\mathbf{w}_t^{\top}\mathbf{x}_t, y_t\right)$ — 计算复杂度$O(d)$

3. 更新模型 $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$

计算复杂度低✓、收敛速率慢×

基于梯度符号的随机优化

- 随机梯度下降（SGD）
  1. 随机梯度 $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t^\top \mathbf{x}_t, y_t)$
  2. 更新模型 $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$

SGD及其变形（如Adam）广泛应用大规模训练

- 符号随机梯度下降（signSGD）
  1. 随机梯度 $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t^\top \mathbf{x}_t, y_t)$
  2. 利用梯度符号更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathrm{Sign}(\mathbf{g}_t)$$

降低通讯代价

大模型训练

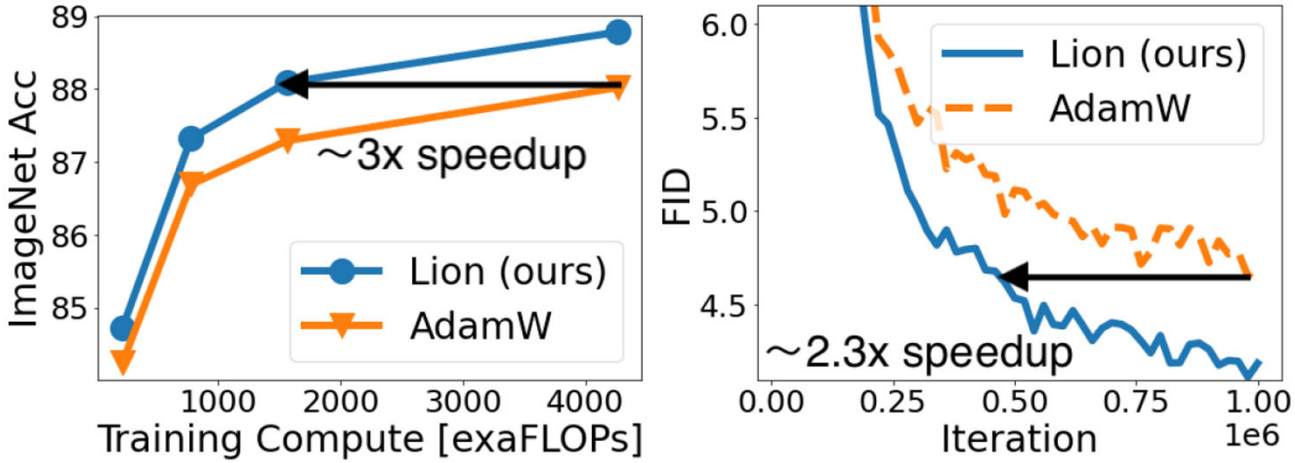
➢ Lion (EvoLved Sign Momentum) [Chen et al., 2023]

$$\mathbf{u}_t = \beta_1 \mathbf{u}_{t-1} + (1-\beta_1)\nabla \ell\left(\mathbf{w}_t^\top \mathbf{x}_t, y_t\right)$$

$$\mathbf{m}_t = \beta_2 \mathbf{u}_{t-1} + (1-\beta_2)\nabla \ell\left(\mathbf{w}_t^\top \mathbf{x}_t, y_t\right)$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t(\mathrm{Sign}(\mathbf{m}_t) + \lambda \mathbf{w}_t)$$

Google Brain



不仅提高通讯效率，还提升训练效率

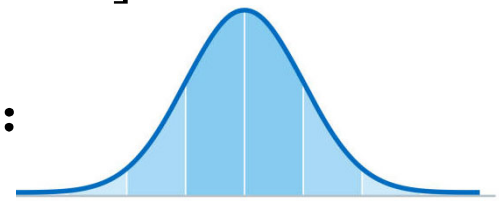
理论研究严重滞后

- 单机场景
  - 依赖大批量采样 [Bernstein et al., 2018]

    收敛速率：$O\left(\frac{1}{T^{1/4}}\right)$　　批量大小：$O\left(\frac{1}{\epsilon^2}\right)$

  - 对噪声添加额外假设 [Bernstein et al., 2019]

    收敛速率：$O\left(\frac{1}{T^{1/4}}\right)$　　单峰对称噪声：

  - 收敛速率慢 [Sun et al., 2023]

    收敛速率：$O\left(\frac{d}{T^{1/4}}\right)$，对维度$d$的依赖过高

理论研究严重滞后

- 分布式场景
  - 收敛速率慢 [Jin et al., 2021]

    收敛速率：$O\left(\frac{d^{3/8}}{T^{1/8}}\right)$，对迭代轮数$T$的依赖差

  - 无法收敛到0 [Sun et al., 2023]

    收敛速率：$O\left(\frac{d}{T^{1/4}}+\frac{d}{\sqrt{m}}\right)$，$m$为分布式节点数量

我们的工作

➢ 理论挑战

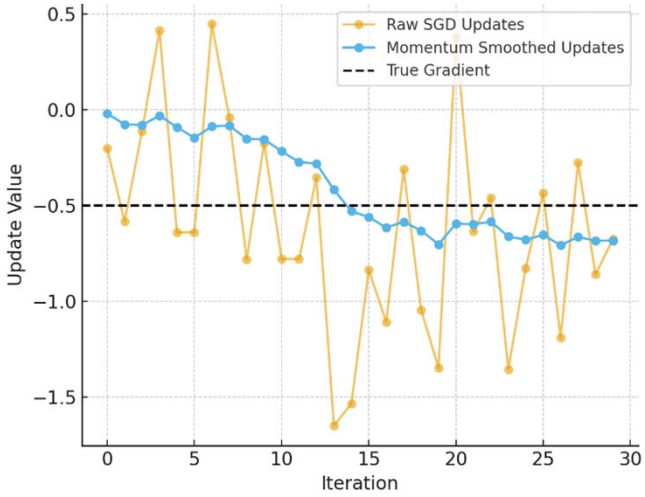
$\mathbf{g}_t = \nabla \ell(\mathbf{w}^\top \mathbf{x}_t, y_t)$

- 随机梯度方差大

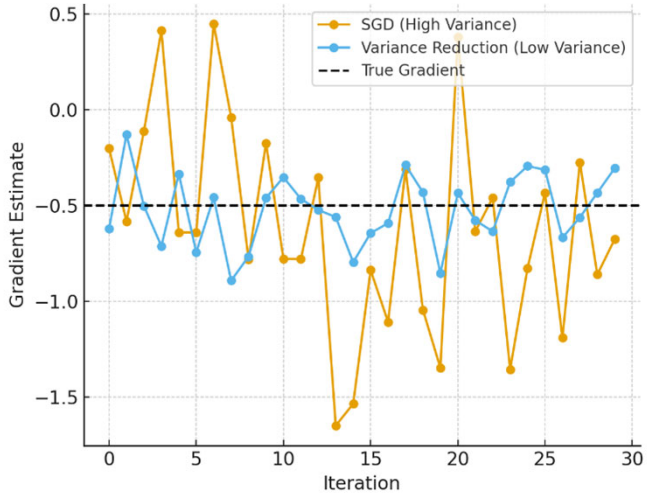
$\mathrm{Sign}(\mathbf{g}_t)$

- 符号操作引入误差

➢ 快速收敛的符号随机优化 [Jiang et al., 2024; 2025]



动量法（Momentum）



方差约减（Variance Reduction）

目录

- 研究背景
- 加速的梯度符号优化
  - 动量法
  - 方差约减
- 分布式场景
- 总结展望

1、结合动量法的符号随机优化

➢ 算法流程 [Jiang et al., 2025]

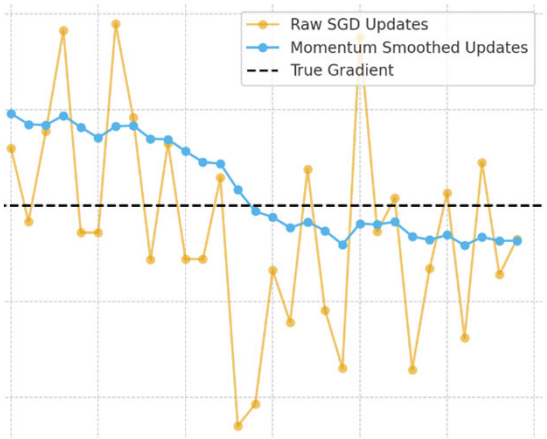
1. 随机梯度

$$\mathbf{g}_t = \nabla \ell\left(\mathbf{w}_t^{\top}\mathbf{x}_t, y_t\right)$$

2. 梯度动量

$$\mathbf{v}_t = (1-\beta)\mathbf{v}_{t-1} + \beta \mathbf{g}_t$$

考虑梯度的累积趋势



3. 利用符号更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathrm{Sign}(\mathbf{v}_t)$$

1、结合动量法的符号随机优化

➢ 理论保障 [Jiang et al., 2025]

假设1：函数平滑

$$\|\nabla F(\mathbf{u}) - \nabla F(\mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\|$$

假设2：方差有界

$$\mathbb{E}_{(\mathbf{x},y)}\left[\left\|\nabla F(\mathbf{w}) - \nabla \ell(\mathbf{w}^\top \mathbf{x}, y)\right\|^2\right] \leq \sigma^2$$

◘收敛速率

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_1] = O\left(\frac{\sqrt{d}}{T^{1/4}}\right)$$

1、结合动量法的符号随机优化

➢ 理论优势

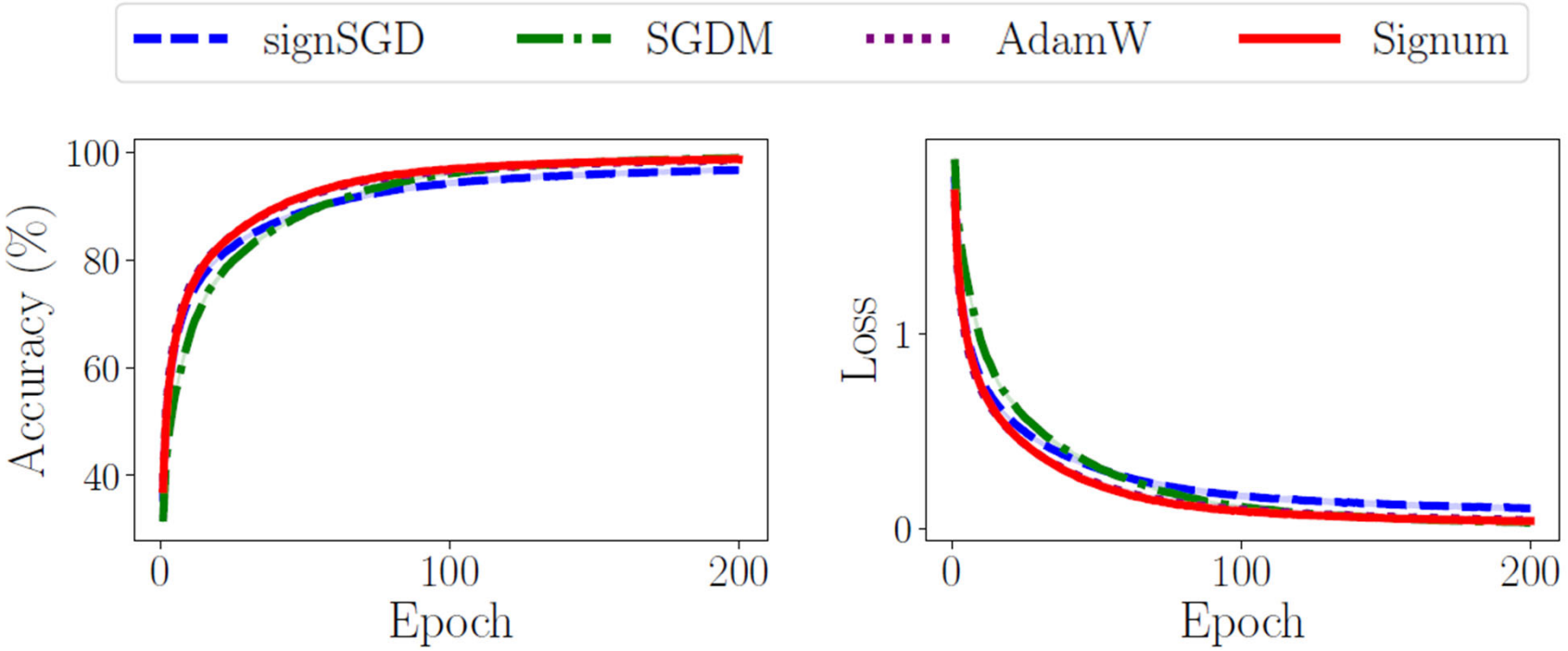
方法	收敛速率	额外假设
Bernstein et al., 2018	$O(1/T^{1/4})$	极大批量
Bernstein et al., 2019	$O(1/T^{1/4})$	单峰对称噪声
Sun et al., 2023	$O(d/T^{1/4})$	-
Jiang et al., 2025	$O\left(\sqrt{d}/T^{1/4}\right)$	-

注：前两个工作没有维度d依赖，是由于平滑性假设不同

实验结果

➢ 数据集：CIFAR-10

实值优化算法

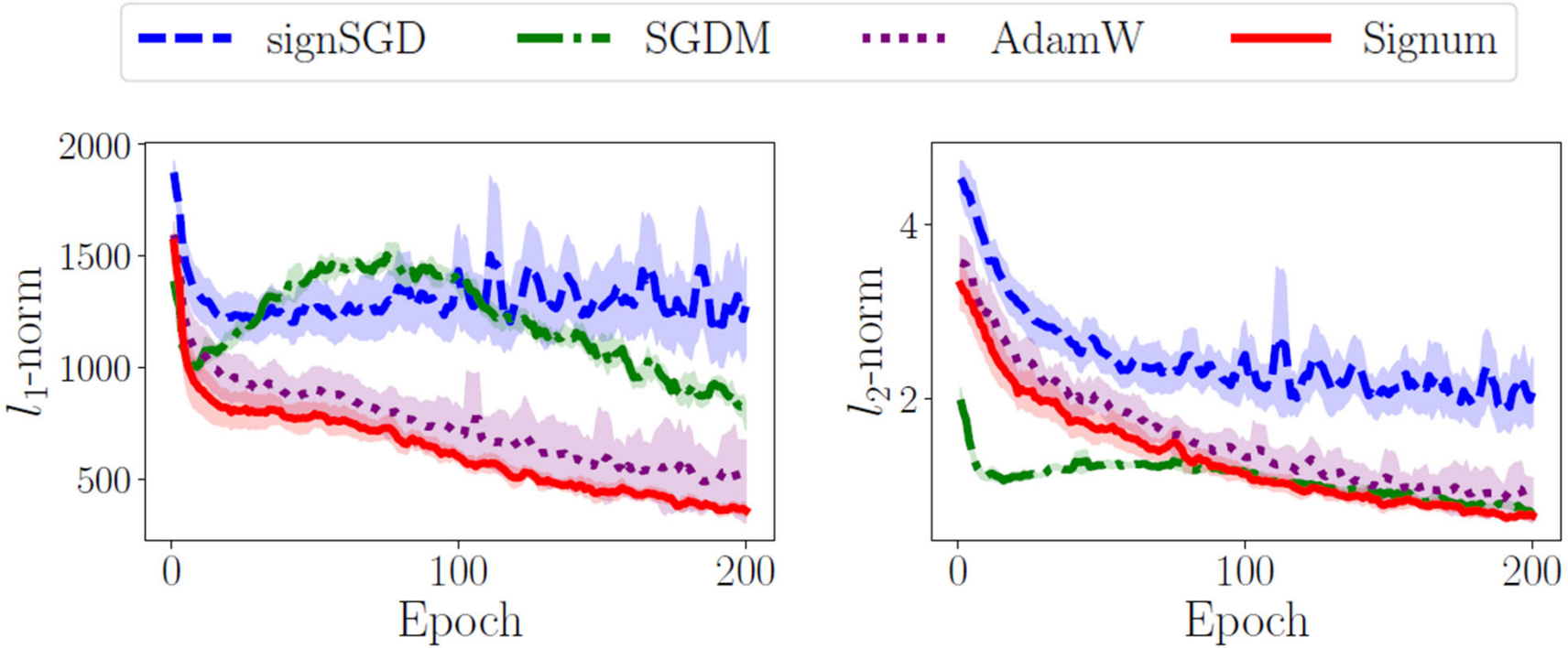


准确率　　损失函数

实验结果

➢ 数据集：CIFAR-10

实值优化算法



梯度ℓ_1范数　　梯度ℓ_2范数

2、结合方差约减的符号随机优化

➢ 算法流程 [Jiang et al., 2024]

1. 随机梯度

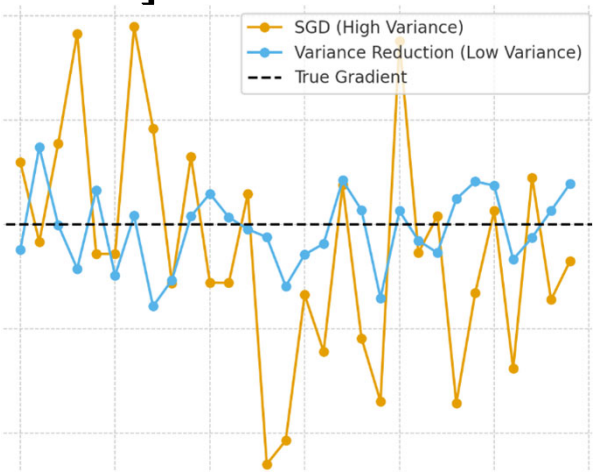
$$\mathbf{g}_t(\mathbf{w}_t) = \nabla \ell\left(\mathbf{w}_t^{\top}\mathbf{x}_t, y_t\right)$$

2. 方差约减 [Cutkosky and Orabona, 2019]

$$\mathbf{v}_t = (1-\beta)\mathbf{v}_{t-1} + \beta \mathbf{g}_t(\mathbf{w}_t)$$

$$+(1-\beta)\left(\mathbf{g}_t(\mathbf{w}_t) - \mathbf{g}_t(\mathbf{w}_{t-1})\right)$$

通过梯度作差降低偏差



3. 利用符号更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathrm{Sign}(\mathbf{v}_t)$$

2、结合方差约减的符号随机优化

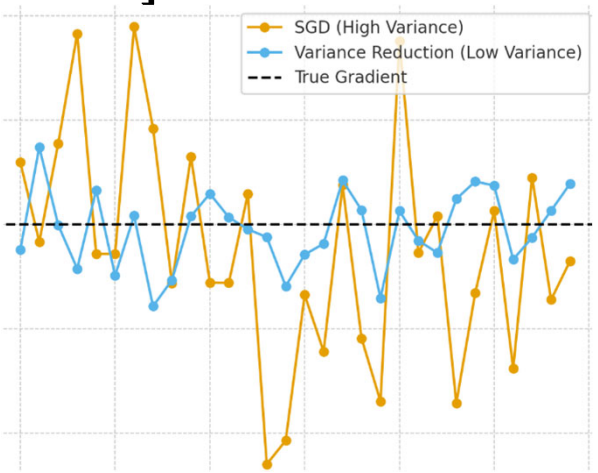
➢ 算法流程 [Jiang et al., 2024]

1. 随机梯度

$$\mathbf{g}_t(\mathbf{w}_t) = \nabla \ell\left(\mathbf{w}_t^\top \mathbf{x}_t, y_t\right)$$

2. 方差约减 [Cutkosky and Orabona, 2019]

$$\mathbf{v}_t = \mathbf{g}_t(\mathbf{w}_t) + (1-\beta)(\mathbf{v}_{t-1} - \mathbf{g}_t(\mathbf{w}_{t-1}))$$



通过梯度作差降低偏差

3. 利用符号更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathrm{Sign}(\mathbf{v}_t)$$

2、结合方差约减的符号随机优化

➢ 理论保障 [Jiang et al., 2024]

假设1：平均平滑

$$\mathbb{E}_{(\mathbf{x},y)}\left[\left\|\nabla \ell(\mathbf{u}^\top \mathbf{x}, y) - \nabla \ell(\mathbf{v}^\top \mathbf{x}, y)\right\|^2\right] \leq L^2 \|\mathbf{u} - \mathbf{v}\|^2$$

假设2：方差有界

$$\mathbb{E}_{(\mathbf{x},y)}\left[\left\|\nabla F(\mathbf{w}) - \nabla \ell(\mathbf{w}^\top \mathbf{x}, y)\right\|^2\right] \leq \sigma^2$$

◘收敛速率

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_1] = O\left(\frac{\sqrt{d}}{T^{1/3}}\right)$$

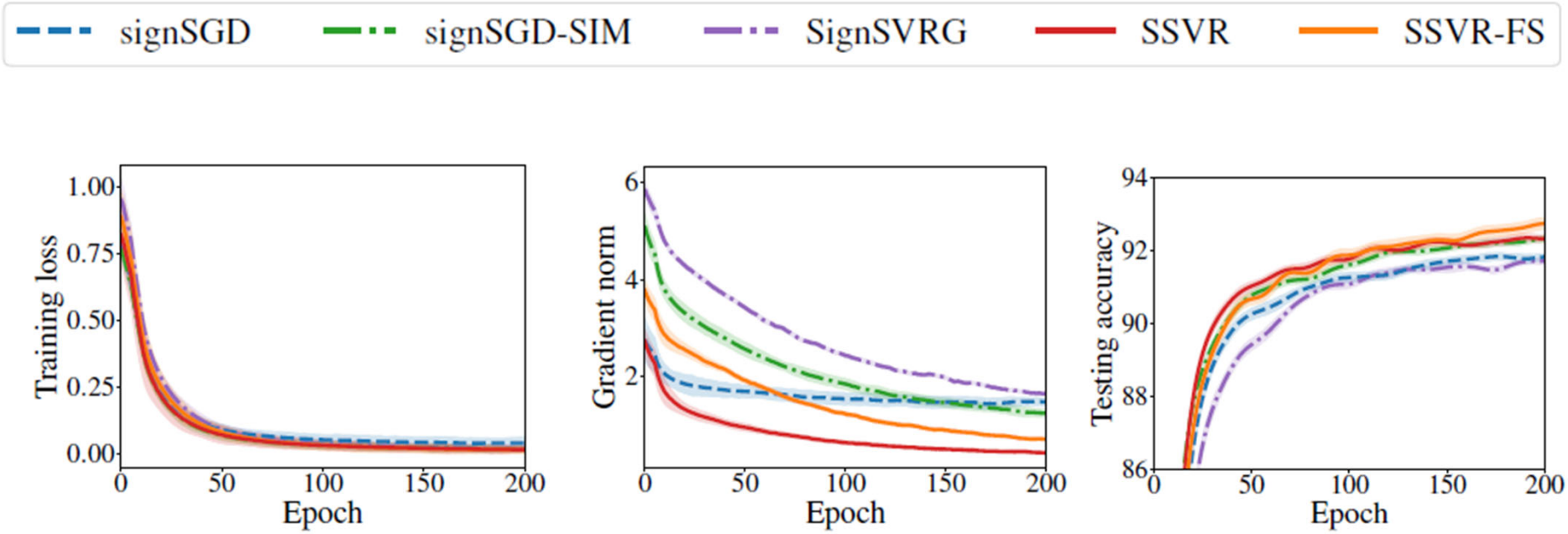
2、结合方差约减的符号随机优化

➢ 理论优势

方法	收敛速率	额外假设
Bernstein et al., 2018	$O(1/T^{1/4})$	极大批量
Bernstein et al., 2019	$O(1/T^{1/4})$	单峰对称噪声
Sun et al., 2023	$O(d/T^{1/4})$	-
Jiang et al., 2025	$O\left(\sqrt{d}/T^{1/4}\right)$	-
Jiang et al., 2024	$O\left(\sqrt{d}/T^{1/3}\right)$	平均光滑性

实验结果

➢ 数据集：CIFAR-10



损失函数　　梯度范数　　准确率

目录

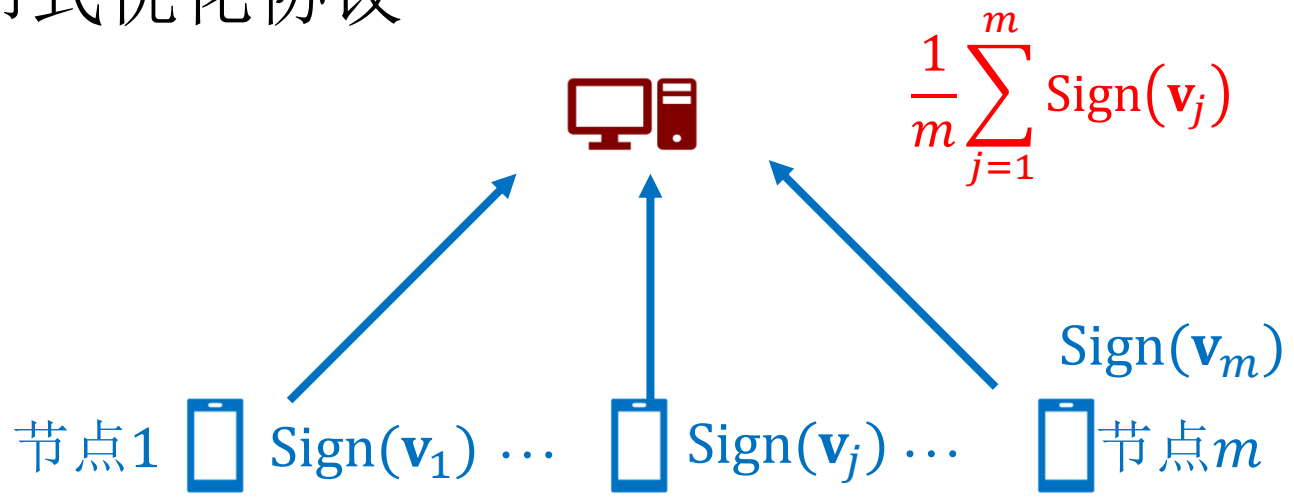
- 研究背景
- 加速的梯度符号优化
  - 动量法
  - 方差约减
- 分布式场景
- 总结展望

分布式场景

- 每个函数F_j分布在不同的节点上

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{m}\sum_{j=1}^{m} F_j(\mathbf{w}), \quad F_j(\mathbf{w}) = \mathbb{E}_{\xi^j \sim \mathcal{D}^j}\left[\ell(\mathbf{w}; \xi^j)\right]$$

- 分布式优化协议



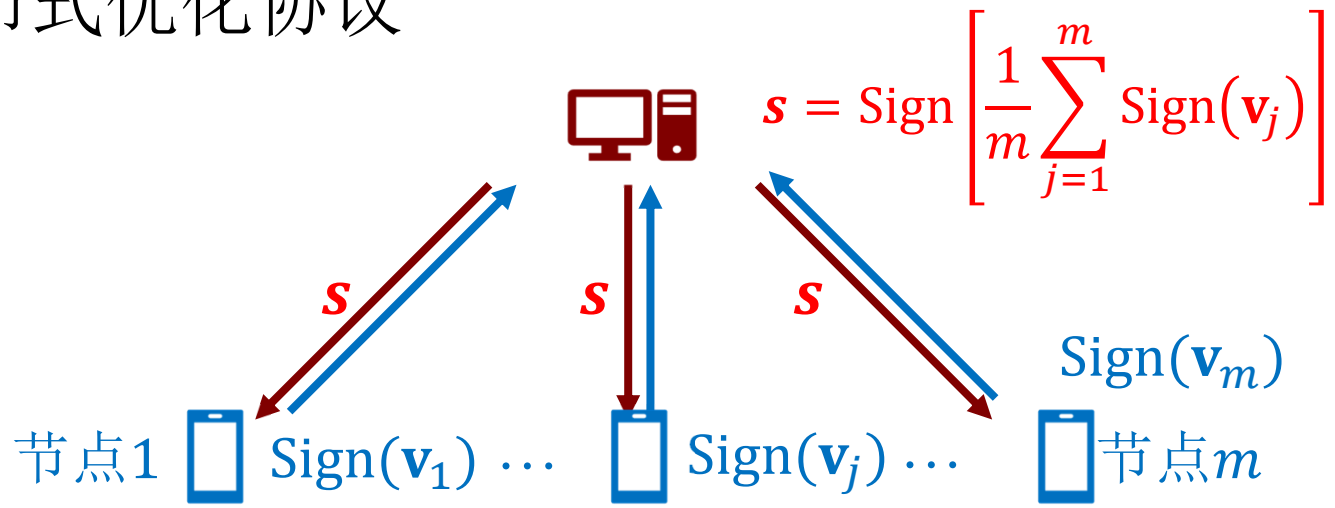
$$F_1(\mathbf{w}) = \mathbb{E}_{\xi^1 \sim \mathcal{D}^1}\left[\ell(\mathbf{w}; \xi^1)\right] \qquad F_m(\mathbf{w}) = \mathbb{E}_{\xi^m \sim \mathcal{D}^m}\left[\ell(\mathbf{w}; \xi^m)\right]$$

分布式场景

- 每个函数F_j分布在不同的节点上

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{m}\sum_{j=1}^{m} F_j(\mathbf{w}), \quad F_j(\mathbf{w}) = \mathbb{E}_{\xi^j \sim \mathcal{D}^j}\left[\ell(\mathbf{w}; \xi^j)\right]$$

- 分布式优化协议



$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{s}$$

无偏符号函数

- 分布式下的新挑战：两次Sign函数的应用
  - 误差叠加

$$\boldsymbol{s} = \mathrm{Sign}\left[\frac{1}{m}\sum_{j=1}^{m}\mathrm{Sign}(\mathbf{v}_j)\right]$$

- 无偏符号函数
  - 对于向量$\mathbf{v}$，其中$\|\mathbf{v}\|_\infty \le R$，定义

$$[\mathrm{S}_R(\mathbf{v})]_k = \begin{cases} 1, & \text{概率为} \dfrac{R+v_k}{2R} \\ -1, & \text{概率为} \dfrac{R-v_k}{2R} \end{cases}$$

1、结合动量法的符号随机优化

➢ 算法流程 [Jiang et al., 2025]

1. 各节点j计算随机梯度

$$\mathbf{g}_t^j = \nabla \ell\left(\mathbf{w}_t^\top \mathbf{x}_t^j, y_t^j\right)$$

2. 各节点j维护梯度动量

$$\mathbf{v}_t^j = (1-\beta)\mathbf{v}_{t-1}^j + \beta \mathbf{g}_t^j$$

3. 各节点j传递无偏符号$\mathrm{S}_R(\mathbf{v}_t^j)$

4. 服务器传递聚合符号

$$\boldsymbol{s}_t = \mathrm{Sign}\left(\frac{1}{m}\sum_{j=1}^{m} \mathrm{S}_R\left(\mathbf{v}_t^j\right)\right)$$

5. 各节点j更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{s}_t$$

1、结合动量法的符号随机优化

➢ 理论保障 [Jiang et al., 2025]

假设1：各节点函数平滑

$$\left\|\nabla F_j(\mathbf{u}) - \nabla F_j(\mathbf{v})\right\| \leq L\|\mathbf{u}-\mathbf{v}\|$$

假设2：各节点方差有界

$$\mathbb{E}_{(\mathbf{x}^j, y^j)}\left[\left\|\nabla F_j(\mathbf{w}) - \nabla \ell\left(\mathbf{w}^\top \mathbf{x}^j, y^j\right)\right\|^2\right] \leq \sigma^2$$

◘收敛速率

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_1] = O\left(\frac{\sqrt{d}}{\sqrt{T}} + \frac{d}{\sqrt{m}}\right)$$

1、结合动量法的符号随机优化

➢ 理论优势

方法	收敛速率	额外假设
Sun et al., 2023	$O\left(d/T^{1/4}+d/\sqrt{m}\right)$	-
Jiang et al., 2025	$O\left(\sqrt{d}/\sqrt{T}+d/\sqrt{m}\right)$	-

1、结合动量法的符号随机优化

➢ 算法流程 [Jiang et al., 2025]

1. 各节点j计算随机梯度

$$\mathbf{g}_t^j = \nabla \ell\left(\mathbf{w}_t^\top \mathbf{x}_t^j, y_t^j\right)$$

2. 各节点j维护梯度动量

$$\mathbf{v}_t^j = (1-\beta)\mathbf{v}_{t-1}^j + \beta \mathbf{g}_t^j$$

3. 各节点j传递无偏符号$\mathrm{S}_R(\mathbf{v}_t^j)$

4. 服务器传递无偏聚合符号

$$\boldsymbol{s}_t = \mathrm{S}_1\left(\frac{1}{m}\sum_{j=1}^{m} \mathrm{S}_R\left(\mathbf{v}_t^j\right)\right)$$

5. 各节点j更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{s}_t$$

1、结合动量法的符号随机优化

➢ 理论保障 [Jiang et al., 2025]

假设1：各节点函数平滑

$$\left\|\nabla F_j(\mathbf{u}) - \nabla F_j(\mathbf{v})\right\| \leq L\|\mathbf{u}-\mathbf{v}\|$$

假设2：各节点方差有界

$$\mathbb{E}_{(\mathbf{x}^j, y^j)}\left[\left\|\nabla F_j(\mathbf{w}) - \nabla \ell\left(\mathbf{w}^\top \mathbf{x}^j, y^j\right)\right\|^2\right] \leq \sigma^2$$

◘收敛速率

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_1] = O\left(\frac{d^{1/4}}{T^{1/4}} + \frac{d^{1/10}}{T^{1/5}}\right)$$

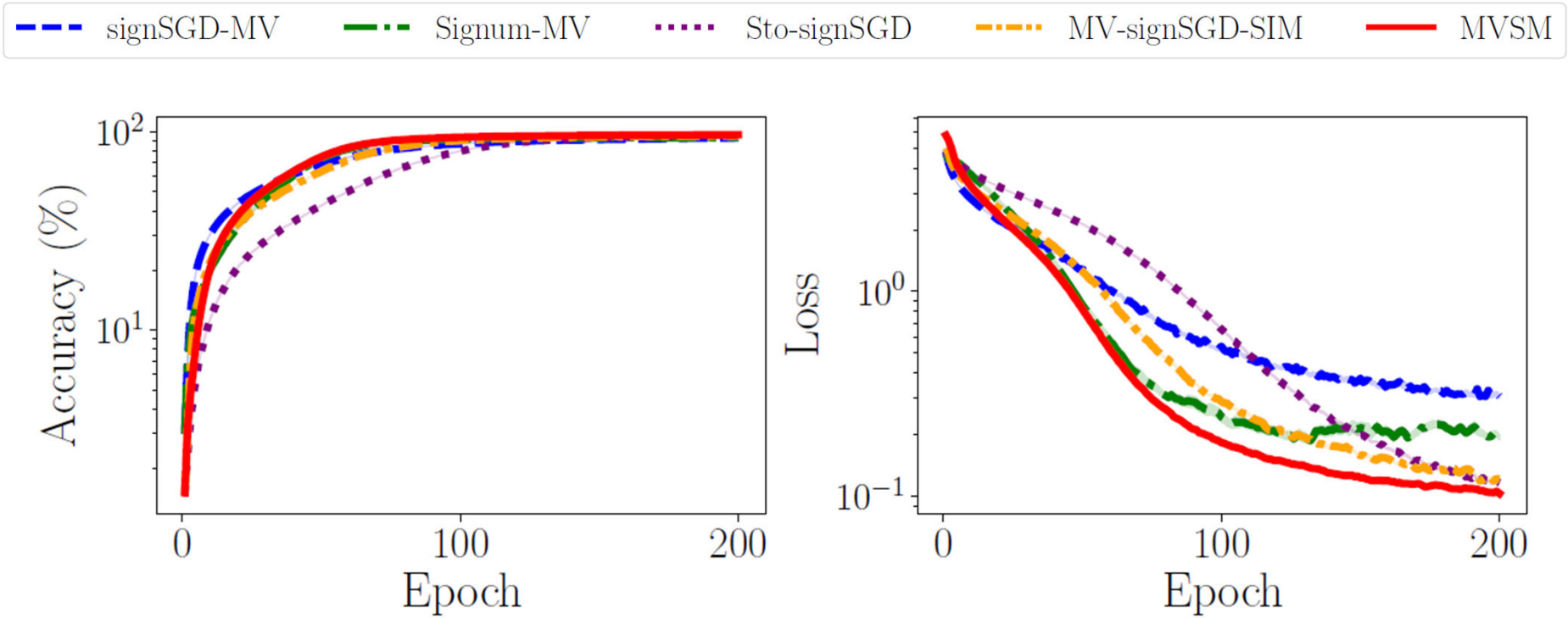
1、结合动量法的符号随机优化

➢ 理论优势

方法	收敛速率	额外假设
Sun et al., 2023	$O\left(d/T^{1/4}+d/\sqrt{m}\right)$	-
Jiang et al., 2025	$O\left(\sqrt{d}/\sqrt{T}+d/\sqrt{m}\right)$	-
Jin et al., 2021	$O\left(d^{3/8}/T^{1/8}\right)$	-
Jiang et al., 2025	$O\left(d^{1/4}/T^{1/4}+d^{1/10}/T^{1/5}\right)$	-

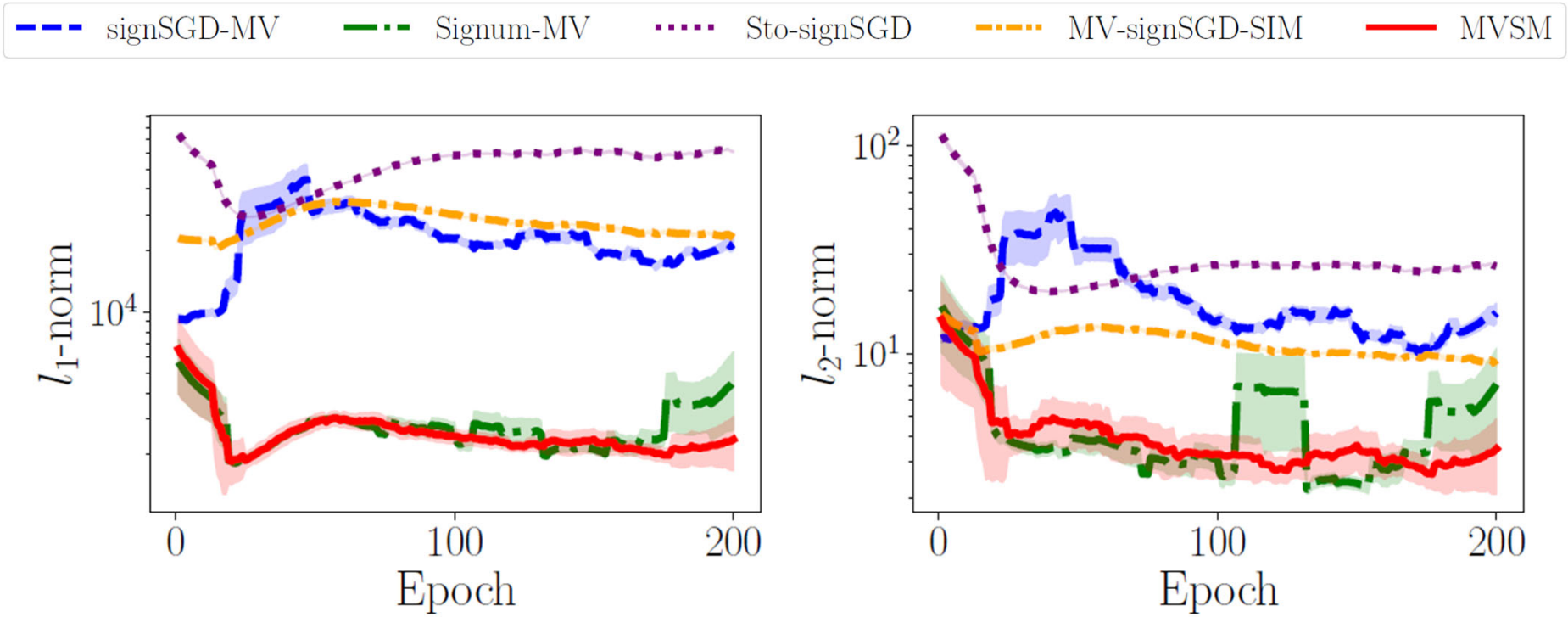
实验结果

➢ 数据集：CIFAR-100



实验结果

➢ 数据集：CIFAR-100



梯度ℓ_1范数　　梯度ℓ_2范数

2、结合方差约减的符号随机优化

➢ 算法流程 [Jiang et al., 2024]

1. 各节点j计算随机梯度

$$\mathbf{g}_t^j(\mathbf{w}_t) = \nabla \ell\left(\mathbf{w}_t^\top \mathbf{x}_t^j, y_t^j\right)$$

2. 各节点j执行方差约减

$$\mathbf{v}_t^j = \mathbf{g}_t^j(\mathbf{w}_t) + (1-\beta)\left(\mathbf{v}_{t-1}^j - \mathbf{g}_t^j(\mathbf{w}_{t-1})\right)$$

3. 各节点j传递无偏符号$\mathrm{S}_R(\mathbf{v}_t^j)$

4. 服务器传递无偏聚合符号

$$\boldsymbol{s}_t = \mathrm{S}_1\left(\frac{1}{m}\sum_{j=1}^{m} \mathrm{S}_R\left(\mathbf{v}_t^j\right)\right)$$

5. 各节点j更新模型

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \boldsymbol{s}_t$$

2、结合方差约减的符号随机优化

➢ 理论保障 [Jiang et al., 2024]

假设1：各节点平均平滑

$$\mathbb{E}_{(\mathbf{x}^j,y^j)}\left[\left\|\nabla\ell\left(\mathbf{u}^\top\mathbf{x}^j,y^j\right)-\nabla\ell\left(\mathbf{v}^\top\mathbf{x}^j,y^j\right)\right\|^2\right]\leq L^2\|\mathbf{u}-\mathbf{v}\|^2$$

假设2：各节点方差有界

$$\mathbb{E}_{(\mathbf{x}^j,y^j)}\left[\left\|\nabla F_j(\mathbf{w})-\nabla\ell\left(\mathbf{w}^\top\mathbf{x}^j,y^j\right)\right\|^2\right]\leq\sigma^2$$

◘收敛速率

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_1]=O\left(\frac{d^{1/4}}{T^{1/4}}\right)$$

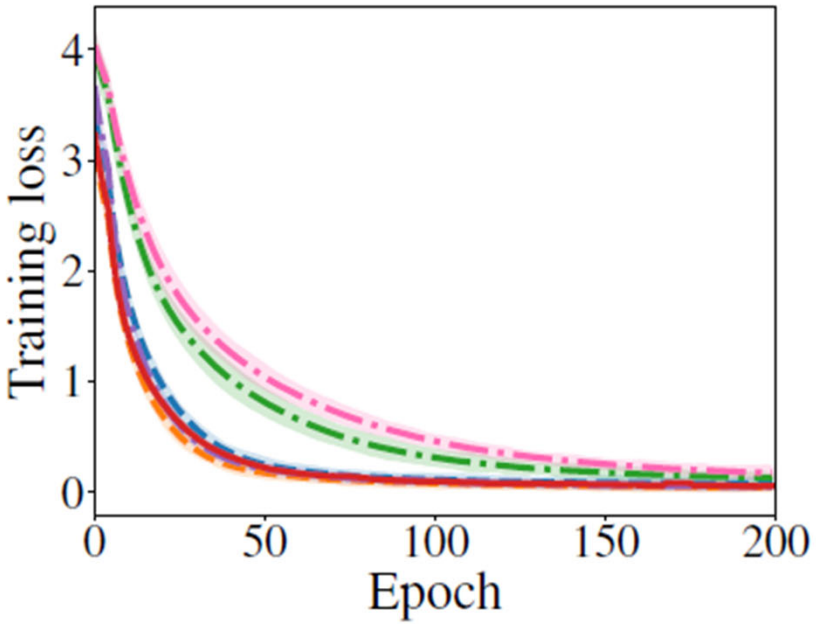
2、结合方差约减的符号随机优化

➢ 理论优势

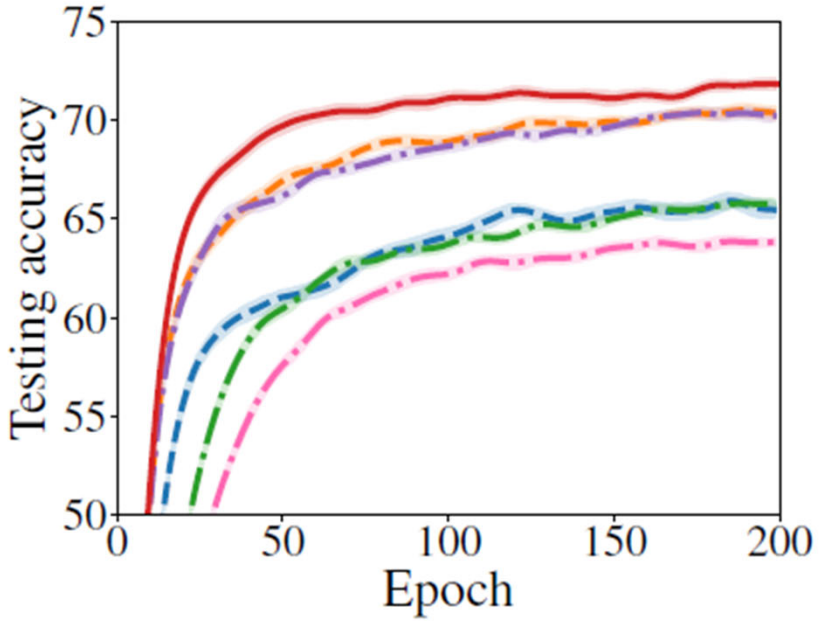
方法	收敛速率	额外假设
Sun et al., 2023	$O\left(d/T^{1/4}+d/\sqrt{m}\right)$	-
Jiang et al., 2025	$O\left(\sqrt{d}/\sqrt{T}+d/\sqrt{m}\right)$	-
Jin et al., 2021	$O\left(d^{3/8}/T^{1/8}\right)$	-
Jiang et al., 2025	$O\left(d^{1/4}/T^{1/4}+d^{1/10}/T^{1/5}\right)$	-
Jiang et al., 2024	$O\left(d^{1/4}/T^{1/4}\right)$	平均光滑性

实验结果

➢ 数据集：CIFAR-100



损失函数



准确率

目录

- 研究背景
- 加速的梯度符号优化
  - 动量法
  - 方差约减
- 分布式场景
- 总结展望

工作总结

- 单机场景下的梯度符号优化

Jiang et al., 2025	$O\left(\sqrt{d}/T^{1/4}\right)$	动量法
Jiang et al., 2024	$O\left(\sqrt{d}/T^{1/3}\right)$	方差约减

- 分布式场景下的梯度符号优化

Jiang et al., 2025	$O\left(\sqrt{d}/\sqrt{T}+d/\sqrt{m}\right)$	动量法 无偏符号
	$O\left(d^{1/4}/T^{1/4}+d^{1/10}/T^{1/5}\right)$	
Jiang et al., 2024	$O\left(d^{1/4}/T^{1/4}\right)$	方差约减 无偏符号

未来展望

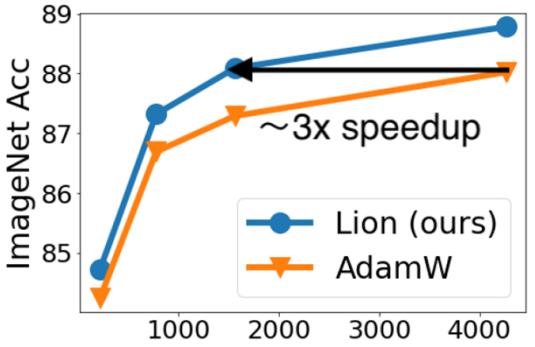
- 分布式场景下的收敛速率尚有提升空间
- 拓展如Lion [Jiang and Zhang, 2025]、Adam、Muon

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\nabla f_t(\mathbf{w}_t)$$

$$\text{Adam} \quad \mathrm{V}_t = \beta_2 \mathrm{V}_{t-1} + (1-\beta_2)\mathrm{diag}(\nabla f_t(\mathbf{w}_t)[\nabla f_t(\mathbf{w}_t)]^\top)$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\alpha}{\sqrt{t}} V_t^{-1/2}\, \mathbf{m}_t$$

- 符号优化的额外优势
  - 优化速率与实值类似
  - 实际表现有时候更好



敬请批评指正！

参考文献

- W. Jiang, S. Yang, W. Yang, and L. Zhang. Efficient Sign-Based Optimization: Accelerating Convergence via Variance Reduction, NeurIPS 2024.
- Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, and Lijun Zhang. Improved Analysis for Sign-based Methods with Momentum Updates, https://arxiv.org/abs/2507.12091.
- Wei Jiang, and Lijun Zhang. Convergence Analysis of the Lion Optimizer in Centralized and Distributed Settings, https://arxiv.org/abs/2508.12327.
- Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic Discovery of Optimization Algorithms, NeurIPS 2023.
- Jeremy Bernstein, Yu-XiangWang, Kamyar Azizzadenesheli, and Anima Anandkumar. SIGNSGD: Compressed Optimisation for Non-Convex Problems, ICML 2018.
- Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD with Majority Vote is Communication Efficient And Fault Tolerant, ICLR 2019.
- Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum Ensures Convergence of SIGNSGD under Weaker Assumptions, ICML 2023.
- Richeng Jin, Yufan Huang, Xiaofan He, Huaiyu Dai, and Tianfu Wu. Stochastic-Sign SGD for Federated Learning with Theoretical Guarantees, https://arxiv.org/abs/2002.10940.
- Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD, NeurIPS 2019.