LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

南京大学
NANJING UNIVERSITY 1902

基于符号的大规模随机优化

张利军
南京大学

第二十四届中国机器学习及其应用研讨会（MLA 2026）



目录

➢ 研究背景

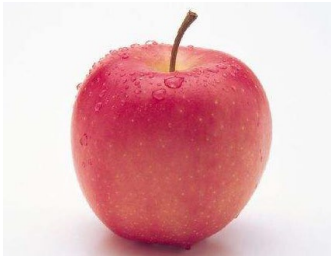
➢ 重尾噪声下的收敛理论

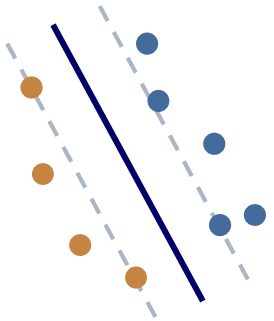
➢ 残差修正的子空间优化

➢ 总结与展望

机器学习：数据 → 模型

➢ 图像分类



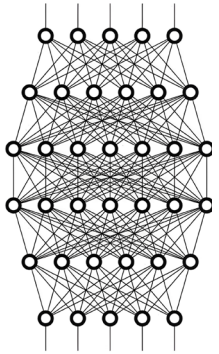


苹果 ✓

梨子 ×

➢ 大语言模型（LLM）

南京大学
位于□□



南

北

其他

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

经验风险最小化

面向机器学习的优化

➢ 有限求和优化问题

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}), \qquad n\text{为样本数量}$$

■ 确定性优化（全梯度）

$$\nabla F(\mathbf{w}_t) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{w}_t)$$

单次更新：n次样本梯度计算

■ 随机优化（小批量）

$$\mathbf{g}_t = \frac{1}{b}\sum_{i\in B_t} \nabla f_i(\mathbf{w}_t)$$

单次更新：b次样本梯度计算

随机优化在计算代价与收敛性能之间取得更好的平衡

自适应随机优化

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

SGD → AdaGrad → Adam

经典随机优化方法 | [JMLR 2011] | [ICLR 2015]

预设步长

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$$

$\mathbf{g}_t$为当前时刻的随机梯度

自适应步长

$$H_t = H_{t-1} + \mathbf{g}_t^2$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\mathbf{g}_t}{\sqrt{H_t} + \varepsilon}$$

动量与自适应步长

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

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon}$$

Adam及其变种成为训练大规模神经网络的主流优化算法

Duchi, Hazan, Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, JMLR 2011
Kingma, Ba. Adam: A Method for Stochastic Optimization, ICLR 2015

基于符号的随机优化

Sign SGD

[ICML 2018]

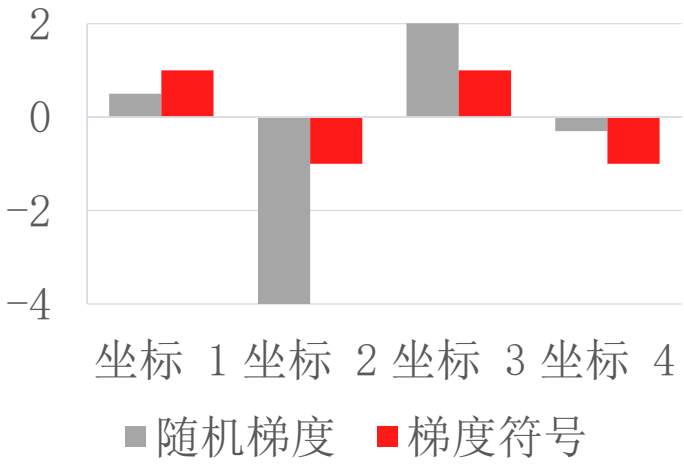
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathrm{Sign}(\mathbf{g}_t)$$

$$\mathrm{Sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

降低
通讯代价



抑制
噪声影响



Bernstein et al. signSGD: Compressed Optimisation for Non-Convex Problems, ICML 2018

基于符号的随机优化

Sign SGD → Lion

[ICML 2018] [NeurIPS 2023]

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

$$\mathrm{Sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

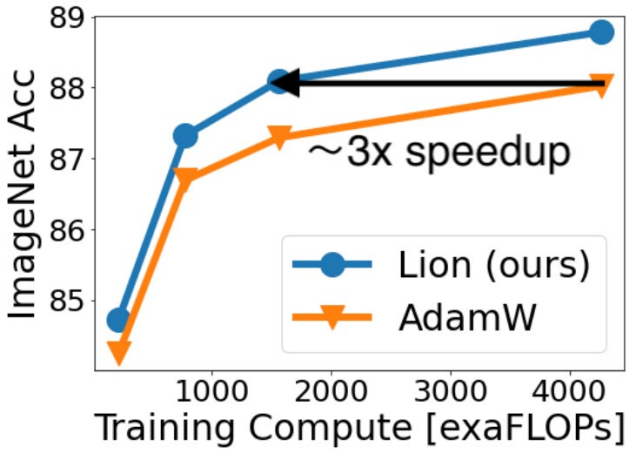
两组时间尺度

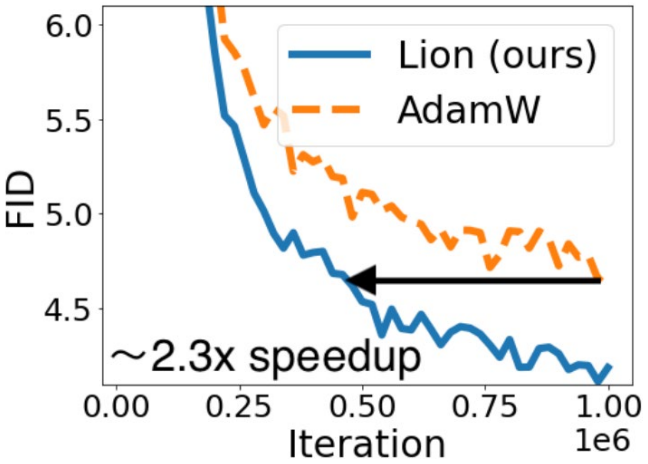
$$\mathbf{v}_t = (1-\beta_1)\mathbf{m}_{t-1} + \beta_1 \mathbf{g}_t$$

$$\mathbf{m}_t = (1-\beta_2)\mathbf{m}_{t-1} + \beta_2 \mathbf{g}_t$$

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

Google Brain





基于符号的随机优化

Sign SGD → Lion → Muon

[ICML 2018] [NeurIPS 2023] [Jordan 2024]

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

$$\mathrm{Sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

两组时间尺度

$$\mathbf{v}_t = (1-\beta_1)\mathbf{m}_{t-1} + \beta_1 \mathbf{g}_t$$

$$\mathbf{m}_t = (1-\beta_2)\mathbf{m}_{t-1} + \beta_2 \mathbf{g}_t$$

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

Google Brain

矩阵变量

$$M_t = \mu M_{t-1} + G_t$$

$$M_t = U_t \Sigma_t V_t^\top, \quad \text{SVD分解}$$

$$W_{t+1} = W_t - \eta U_t \mathrm{Sign}(\Sigma_t) V_t^\top$$
$$= W_t - \eta U_t V_t^\top$$

Jordan. Muon: An optimizer for hidden layers in neural networks, https://kellerjordan.github.io/posts/muon, 2024

Muon在工业界的应用

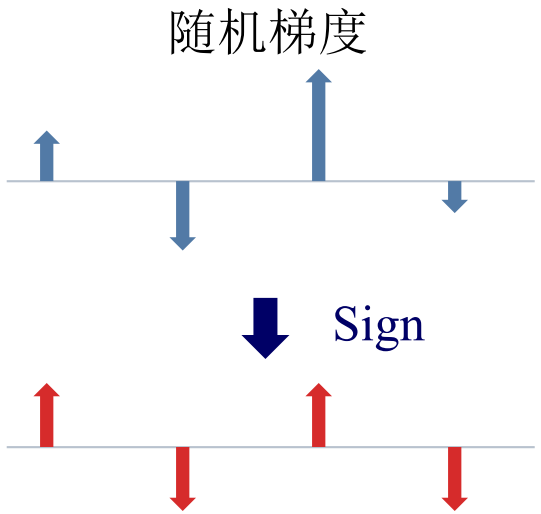
➢ 已用于多款大模型的预训练，覆盖百亿到万亿参数规模

模型	总参数	激活参数	预训练优化器	报告时间
Moonlight [1]	16B	3B	Muon	2025.02
Kimi K2 [2]	1T	32B	MuonClip	2025.07
Kimi K3 [3]	2.78T	104.2B	Per-Head Muon	2026.07
GLM-4.5 [4]	355B	32B	Muon	2025.08
GLM-5 [5]	744B	40B	Muon Split	2026.02
DeepSeek-V4-Pro [6]	1.6T	49B	Muon	2026.04

[1] Moonlight, https://arxiv.org/abs/2502.16982; [2] Kimi K2, https://arxiv.org/abs/2507.20534; [3] Kimi K3, https://arxiv.org/abs/2607.24653
[4] GLM-4.5, https://arxiv.org/abs/2508.06471; [5] GLM-5, https://arxiv.org/abs/2602.15763; [6] DeepSeek-V4, https://arxiv.org/abs/2606.19348

符号随机优化面临的问题

■ 收敛分析

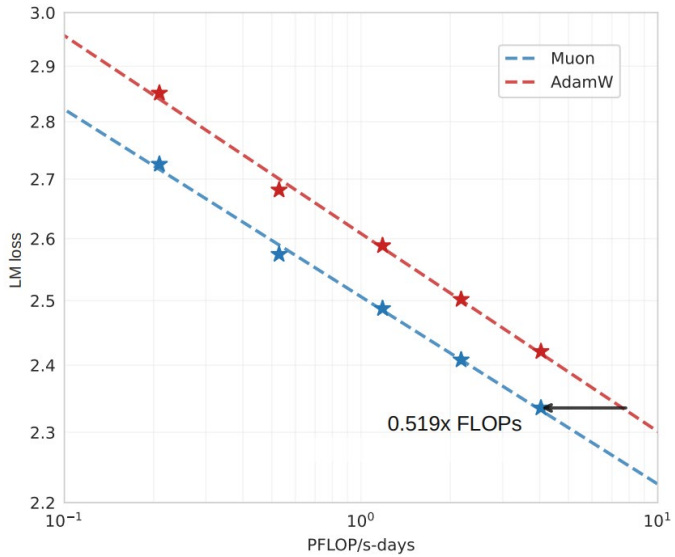


幅值丢失导致不收敛

[ICML 2019]

如何保证收敛性？

■ 机理解释

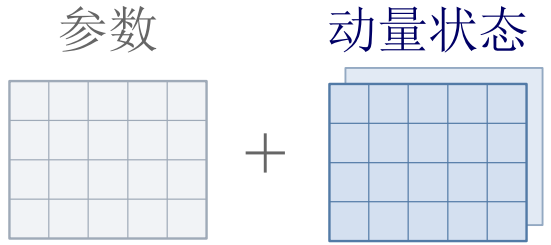


同等损失用更少算力

[Arxiv 2025]

为什么效果更好？

■ 高效计算



动量引入额外存储\计算

$M \rightarrow U \quad \Sigma \quad V^{\mathrm{T}}$

Muon中的高昂SVD分解

如何降低计算量？

Karimireddy et al. Error Feedback Fixes SignSGD and other Gradient Compression Schemes, ICML 2019
Liu et al. Muon is Scalable for LLM Training, https://arxiv.org/abs/2502.16982, 2025

我们的部分工作

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

收敛分析

- Jiang et al. Efficient Sign-Based Optimization: Accelerating Convergence via Variance Reduction, NeurIPS 2024
- Jiang et al. Convergence Analysis of the Lion Optimizer in Centralized and Distributed Settings, ICML 2026

机理解释

- Yu et al. Sign-Based Optimizers Under Heavy-Tailed Noise: On the Effectiveness of Lion and Muon, NeurIPS 2026
- Tao et al. When and Why SignSGD Outperforms SGD: A Theoretical Study Based on ℓ_1-norm Lower Bounds, NeurIPS 2026

高效计算

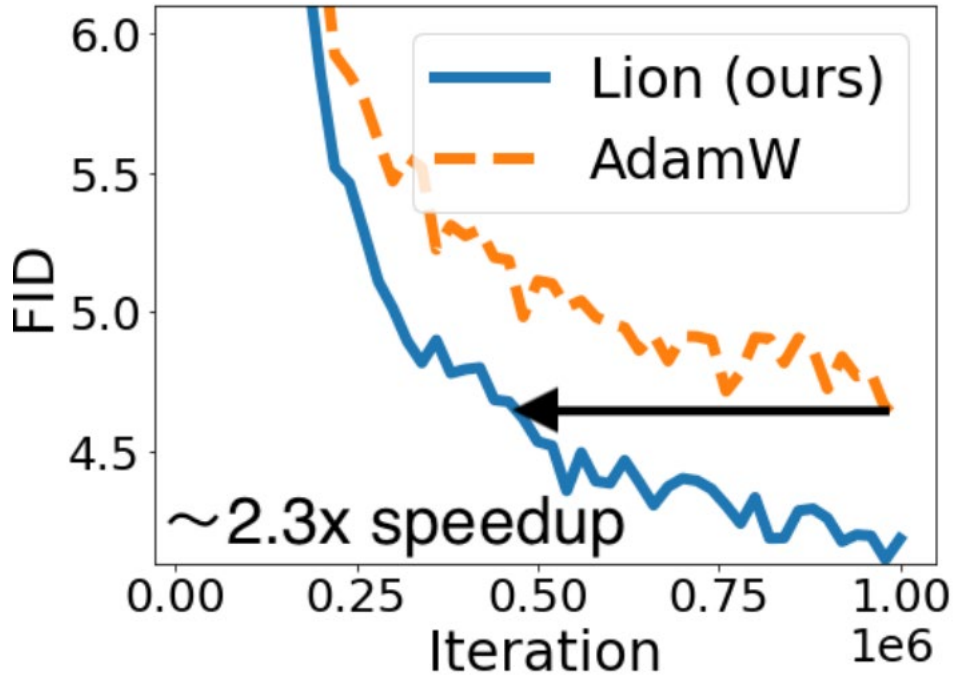
- Yang et al. Memory-Efficient Subspace Optimization with Residual Correction: From Momentum SGD to Muon, https://arxiv.org/abs/2610.19844, 2026

目录

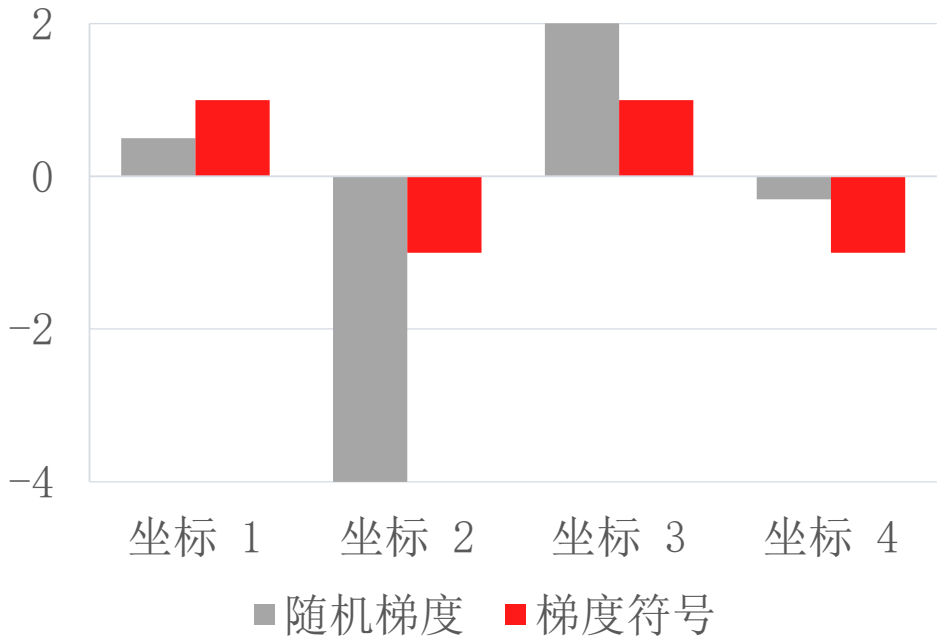
- 研究背景
- 重尾噪声下的收敛理论
- 残差修正的子空间优化
- 总结与展望

研究动机

■ 优异的经验表现



■ 符号的抗噪特性



从抗噪特性出发，理解符号优化的性能优势

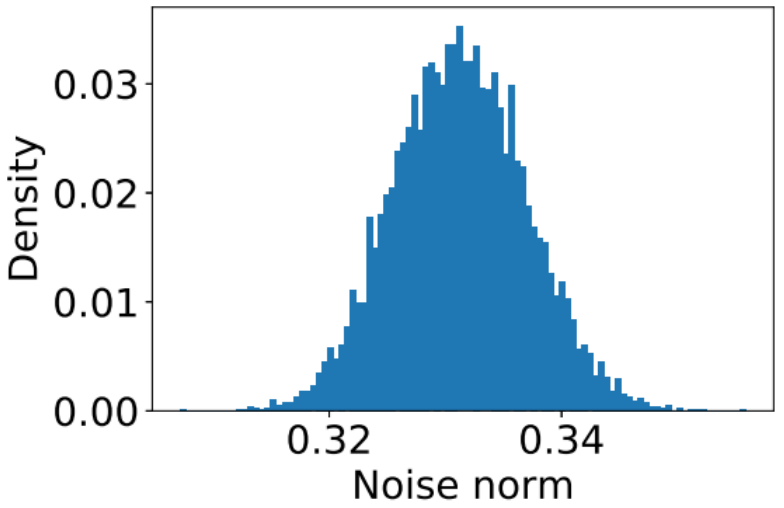
传统随机优化

- 经典噪声模型

$$\mathbb{E}_f[\|\nabla F(\mathbf{w}) - \nabla f(\mathbf{w})\|^2] \leq \sigma^2$$

噪声方差有界

ImageNet



- SGD的收敛速率 [SIAM Opt 2013]

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_2] = O\left(\frac{1}{T^{1/4}}\right)$$

SGD达到最优

- 一阶随机优化下界$\Omega\left(T^{-1/4}\right)$ [MP 2023]

Ghadimi, Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341–2368, 2013
Arjevani et al. Lower bounds for non-convex stochastic optimization. Mathematical Programming, 199(1): 165–214, 2023.

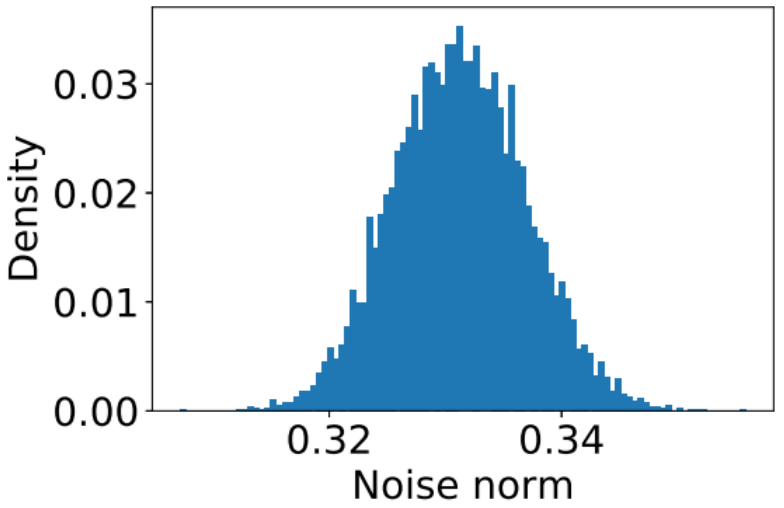
语言模型中的梯度噪声

- 经典噪声模型

$$\mathbb{E}_f[\|\nabla F(\mathbf{w}) - \nabla f(\mathbf{w})\|^2] \leq \sigma^2$$

噪声方差有界

ImageNet



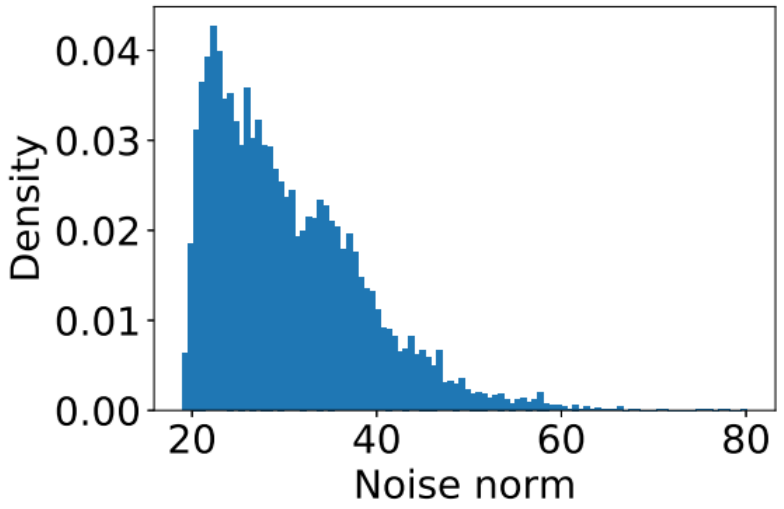
- 重尾噪声模型

$$\mathbb{E}_f[\|\nabla F(\mathbf{w}) - \nabla f(\mathbf{w})\|^{p}] \leq \sigma^{p}$$

p阶矩有界，$p \in (1,2]$

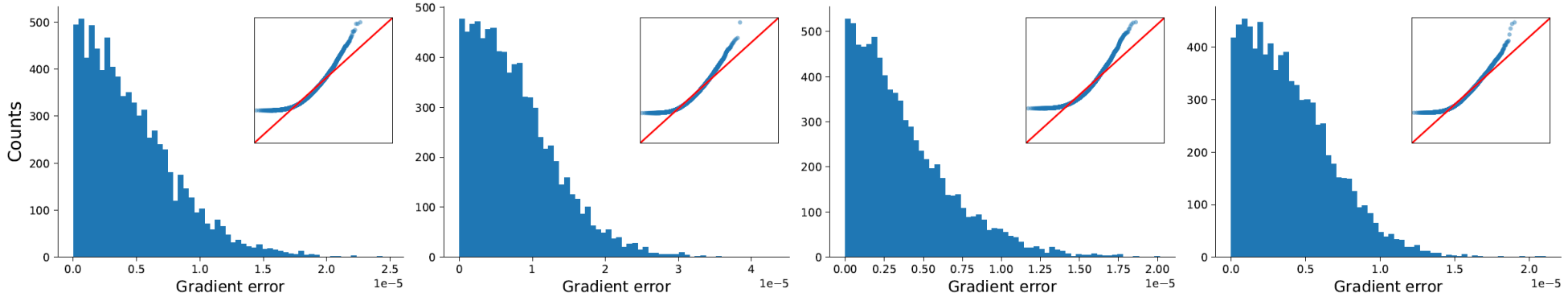
Bert

[NeurIPS 2020]

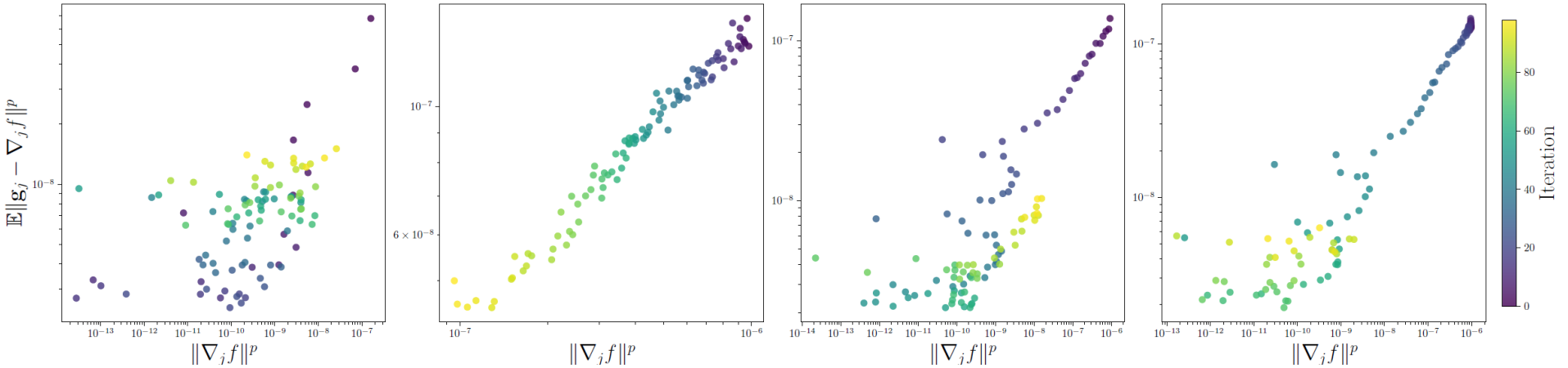


语言模型中的梯度噪声

- 梯度噪声呈现重尾特性



- 噪声强度与梯度范数相关



在C4上训练GPT-2观测的噪声特性

理论贡献

1. 广义重尾噪声模型

$$\mathbb{E}_f[\|\nabla F(\mathbf{w}) - \nabla f(\mathbf{w})\|^p] \leq \sigma_0^p + \sigma_1^p \|\nabla F(\mathbf{w})\|$$

σ_0^p：重尾特性　　$\sigma_1^p \|\nabla F(\mathbf{w})\|$：范数相关性

2. 广义平滑性假设

$$F(\mathbf{w}') \leq F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{1}{2}\|\mathbf{w}' - \mathbf{w}\|^2_{\mathbf{h}(\mathbf{w})}$$ → 局部依赖性

➢ 系统的收敛保障

Sign SGD（带动量）、Lion、Muon、Muonlight

Lion的收敛保障

Lion

[NeurIPS 2023]

两组时间尺度

$$\mathbf{v}_t = (1-\beta_1)\mathbf{m}_{t-1} + \beta_1 \mathbf{g}_t$$

$$\mathbf{m}_t = (1-\beta_2)\mathbf{m}_{t-1} + \beta_2 \mathbf{g}_t$$

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

Google Brain

➢收敛速率

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_1] = O\left(T^{-\frac{p-1}{3p-2}}\right)$$

✓支持重尾分布，$p \in (1,2]$

✓当$p = 2$，恢复经典结果

最优的$O\left(\frac{1}{T^{1/4}}\right)$

✓不直接依赖维度d

比现有少$d^{\frac{3p-2}{2(p-1)}}$

Muon的收敛保障

Muon

[Jordan 2024]

矩阵变量

$$M_t = \mu M_{t-1} + G_t$$

$$M_t = U_t \Sigma_t V_t^\top, \quad \text{SVD分解}$$

$$W_{t+1} = W_t - \eta U_t \mathrm{Sign}(\Sigma_t) V_t^\top$$
$$= W_t - \eta U_t V_t^\top$$

➢ 收敛速率

$$\mathbb{E}[\|\nabla F(W_\tau)\|_*] = O\left(T^{-\frac{p-1}{3p-2}}\right)$$

✓ 支持重尾分布，$p \in (1,2]$

✓ 当$p = 2$，恢复经典结果

最优的$O\left(\frac{1}{T^{1/4}}\right)$

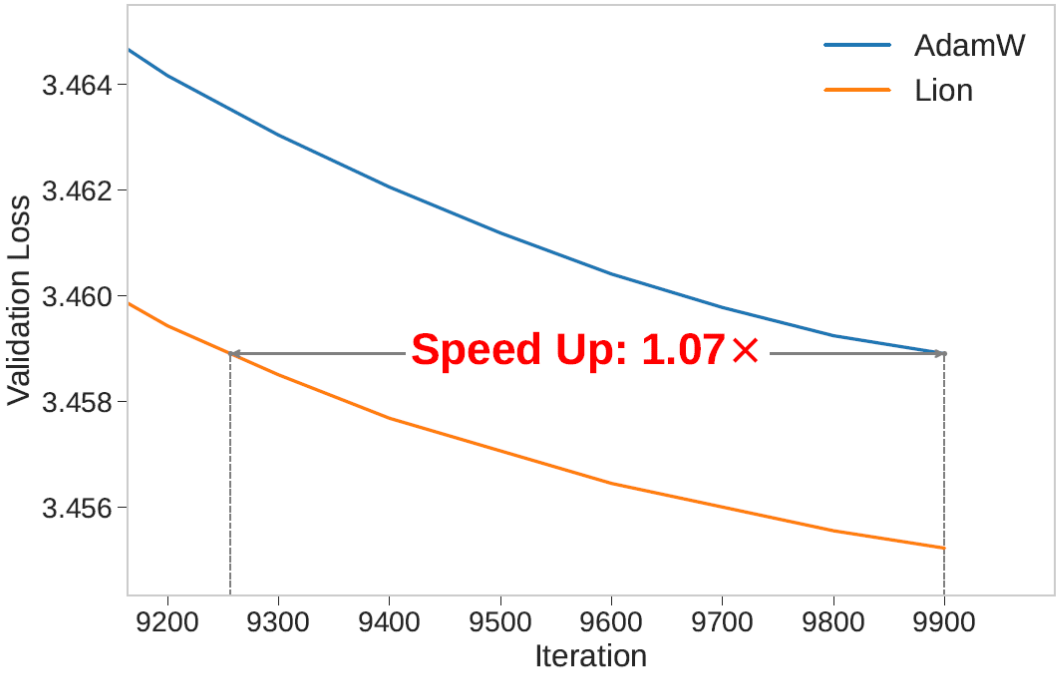
✓ 不直接依赖矩阵秩r

$$\|V\|_* \le \sqrt{r}\|V\|_F$$

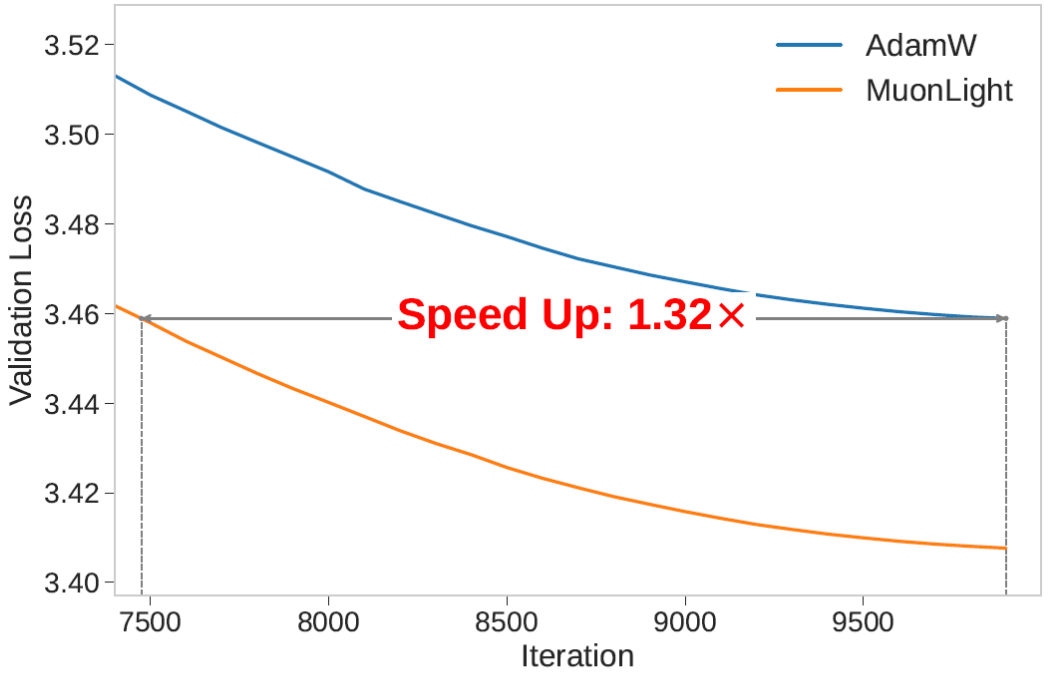
语言模型预训练实验

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

➢ 在C4上训练GPT-2的收敛性对比实验



(a) Lion: 1.07× iteration speedup

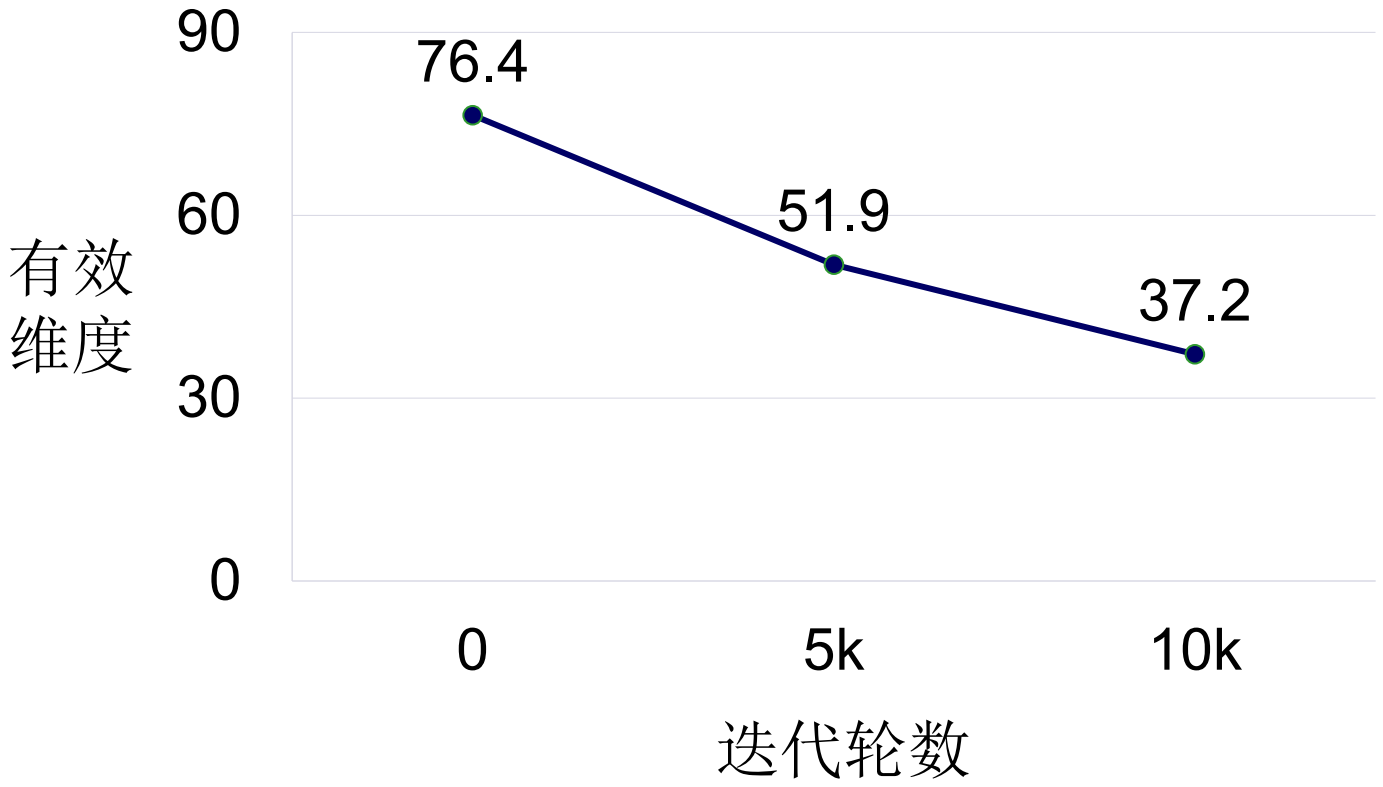


(b) Muon: 1.32× iteration speedup

语言模型预训练实验

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

➢ GPT-2 注意力模块（2304 × 768）的噪声有效维度



- 理论维度768
- 有效维度37.2

$$\sqrt{\frac{37.2}{768}} \approx 0.22$$

目录

- 研究背景
- 重尾噪声下的收敛理论
- 残差修正的子空间优化
- 总结与展望

研究动机

➢ 有限求和优化问题

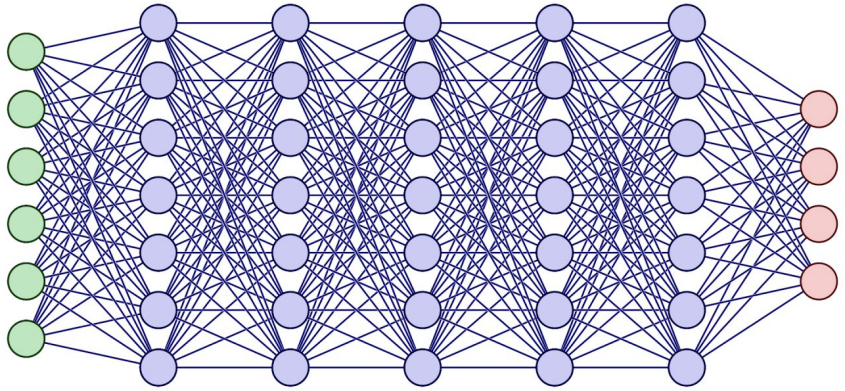
$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}), \qquad n\text{为样本数量}$$

■ 大数据：样本数量n大



随机采样，降低单步计算量

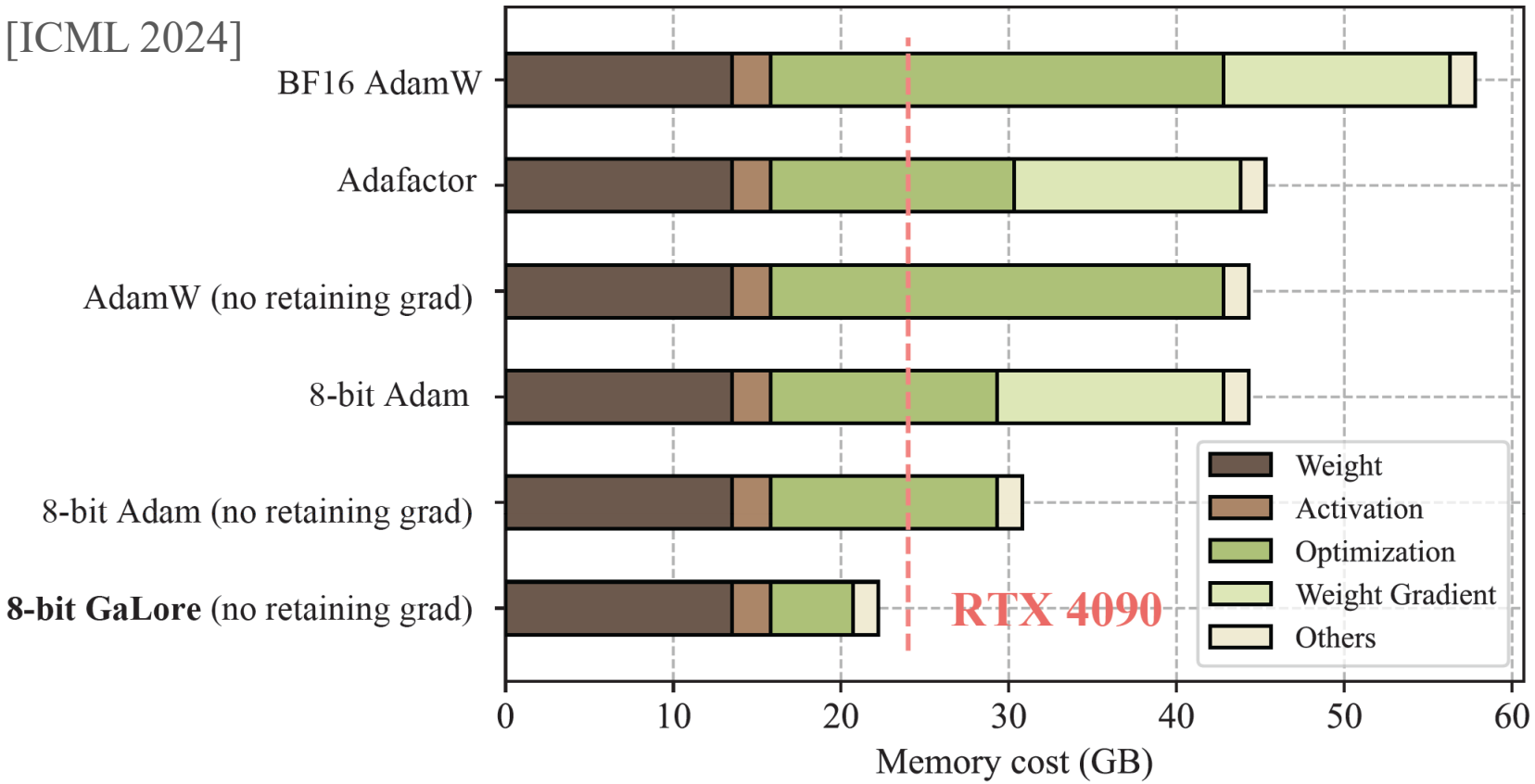
■ 大模型：参数维度d大



采样后，仍需处理高维梯度

研究动机

- LLaMA 7B预训练 [ICML 2024]
  - ✓模型　14GB
  - ✓优化器 42GB
  - ✓激活值　2GB
  - ✓共计　58GB



优化器的内存开销成为制约大模型训练的重要瓶颈

计算瓶颈

Adam

动量与自适应步长

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

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon}$$

Lion

两组时间尺度

$$\mathbf{v}_t = (1-\beta_1)\mathbf{m}_{t-1} + \beta_1 \mathbf{g}_t$$

$$\mathbf{m}_t = (1-\beta_2)\mathbf{m}_{t-1} + \beta_2 \mathbf{g}_t$$

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

Muon

矩阵变量

$$M_t = \mu M_{t-1} + G_t$$

$$M_t = U_t \Sigma_t V_t^\top, \quad \text{SVD分解}$$

$$W_{t+1} = W_t - \eta U_t \mathrm{Sign}(\Sigma_t) V_t^\top$$
$$= W_t - \eta U_t V_t^\top$$

- 维护动量
- SVD分解

子空间Muon

1. 梯度$G_t \in \mathbb{R}^{m\times n}$投影 [ICML 2024]

$$R_t = P^\top G_t \in \mathbb{R}^{r\times n}$$

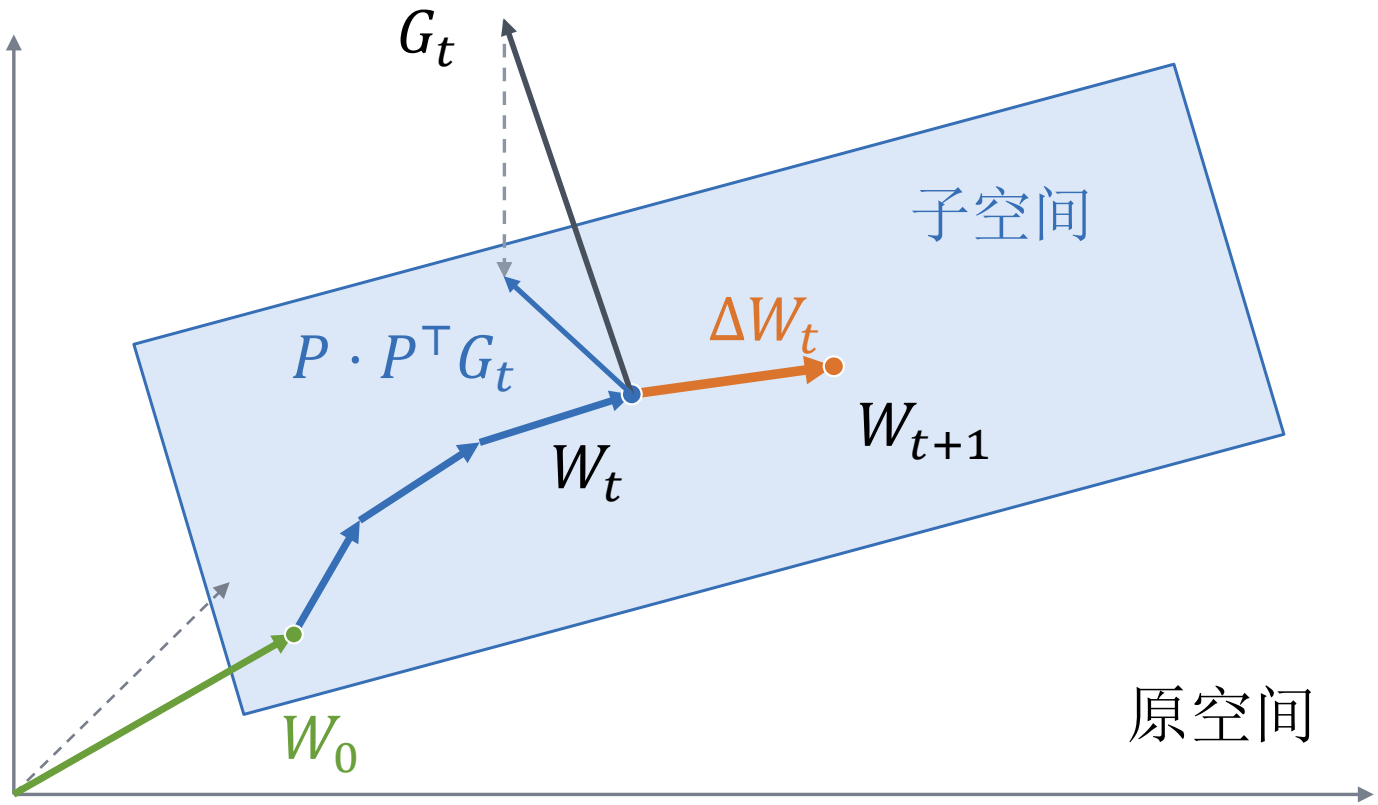
2. 子空间内优化

$$U_t = \text{Muon}(R_t) \in \mathbb{R}^{r\times n}$$

$$\Delta W_t = -\eta P \cdot U_t \in \mathbb{R}^{m\times n}$$

3. 更新模型

$$W_{t+1} = W_t + \Delta W_t$$



子空间Muon

Muon

$$M_t = \mu M_{t-1} + G_t \in \mathbb{R}^{m\times n}$$

$$M_t = U_t \Sigma_t V_t^\top \in \mathbb{R}^{m\times n}$$

$$W_{t+1} = W_t - \eta U_t V_t^\top$$

Projected Muon

$$R_t = P^\top G_t \in \mathbb{R}^{r\times n}$$

$$M_t = \mu M_{t-1} + R_t \in \mathbb{R}^{r\times n}$$

$$M_t = U_t \Sigma_t V_t^\top \in \mathbb{R}^{r\times n}$$

$$W_{t+1} = W_t - \eta P \cdot U_t V_t^\top$$

- 优化器空间： $mn \longrightarrow (m+n)r$
- SVD分解： $O(m^2 n) \longrightarrow O(r^2 n)$

不收敛

[ICML 2025]

子空间优化中的梯度残差

1. 梯度$G_t \in \mathbb{R}^{m\times n}$投影

$$R_t = P^\top G_t \in \mathbb{R}^{r\times n}$$

$$G_t = P \cdot P^\top G_t + (I - PP^\top)G_t$$

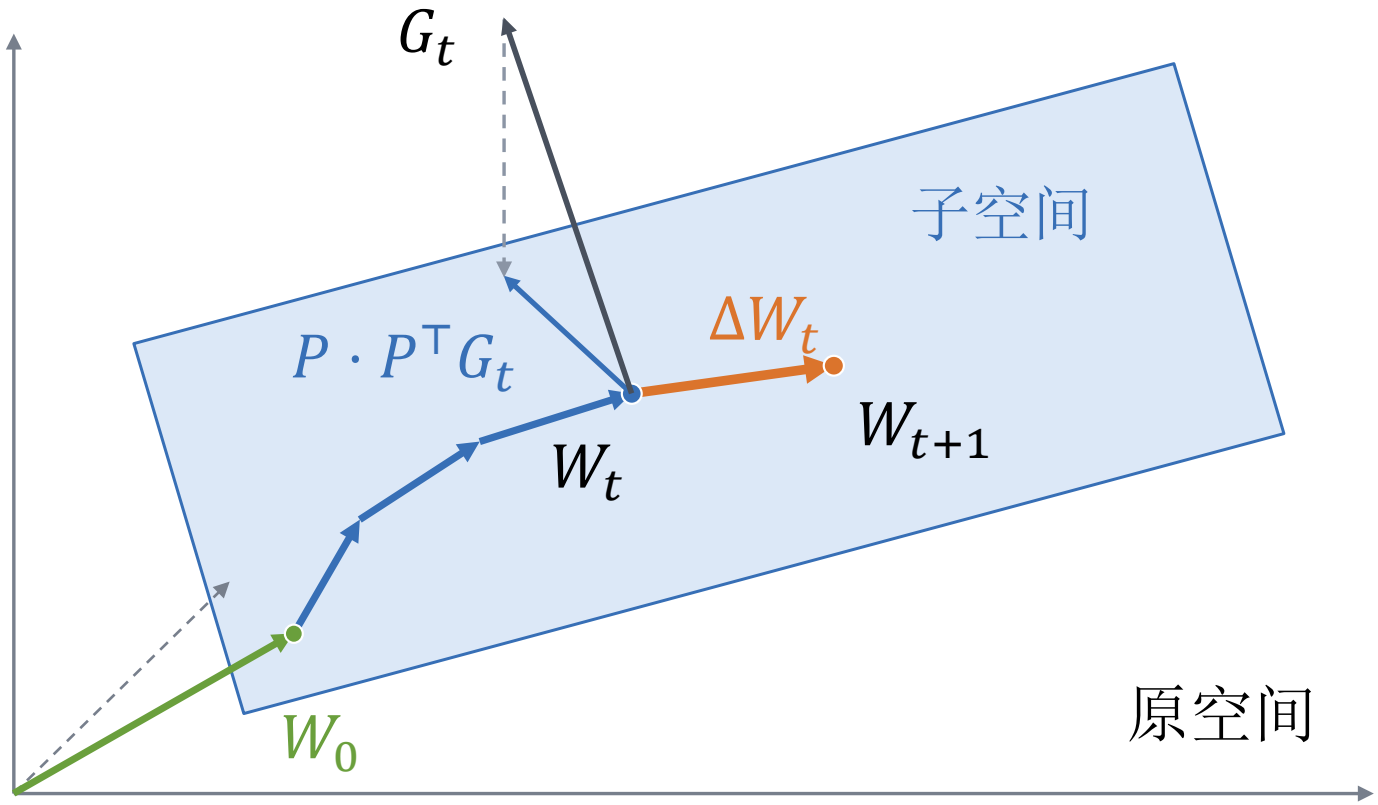
2. 子空间内优化

$$U_t = \mathrm{Muon}(R_t) \in \mathbb{R}^{r\times n}$$

$$\Delta W_t = -\eta P \cdot U_t \in \mathbb{R}^{m\times n}$$

3. 更新模型

$$W_{t+1} = W_t + \Delta W_t$$



子空间优化中的梯度残差

1. 梯度$G_t \in \mathbb{R}^{m\times n}$投影

$$R_t = P^\top G_t \in \mathbb{R}^{r\times n}$$

2. 子空间内优化

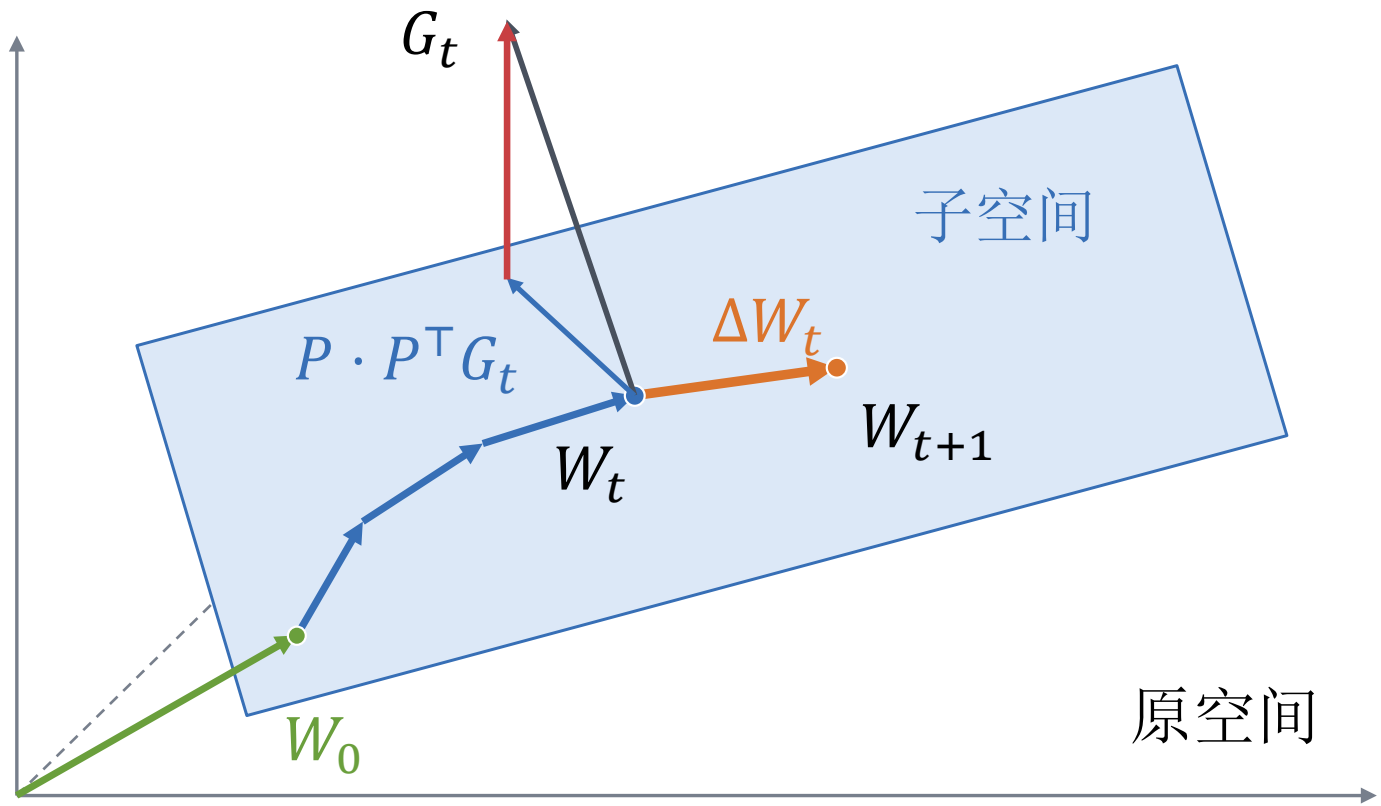
$$U_t = \mathrm{Muon}(R_t) \in \mathbb{R}^{r\times n}$$

$$\Delta W_t = -\eta P \cdot U_t \in \mathbb{R}^{m\times n}$$

3. 更新模型

$$W_{t+1} = W_t + \Delta W_t$$

$$G_t = P \cdot \underbrace{P^\top G_t}_{\text{利用}} + \underbrace{(I - PP^\top)G_t}_{\text{丢弃}}$$



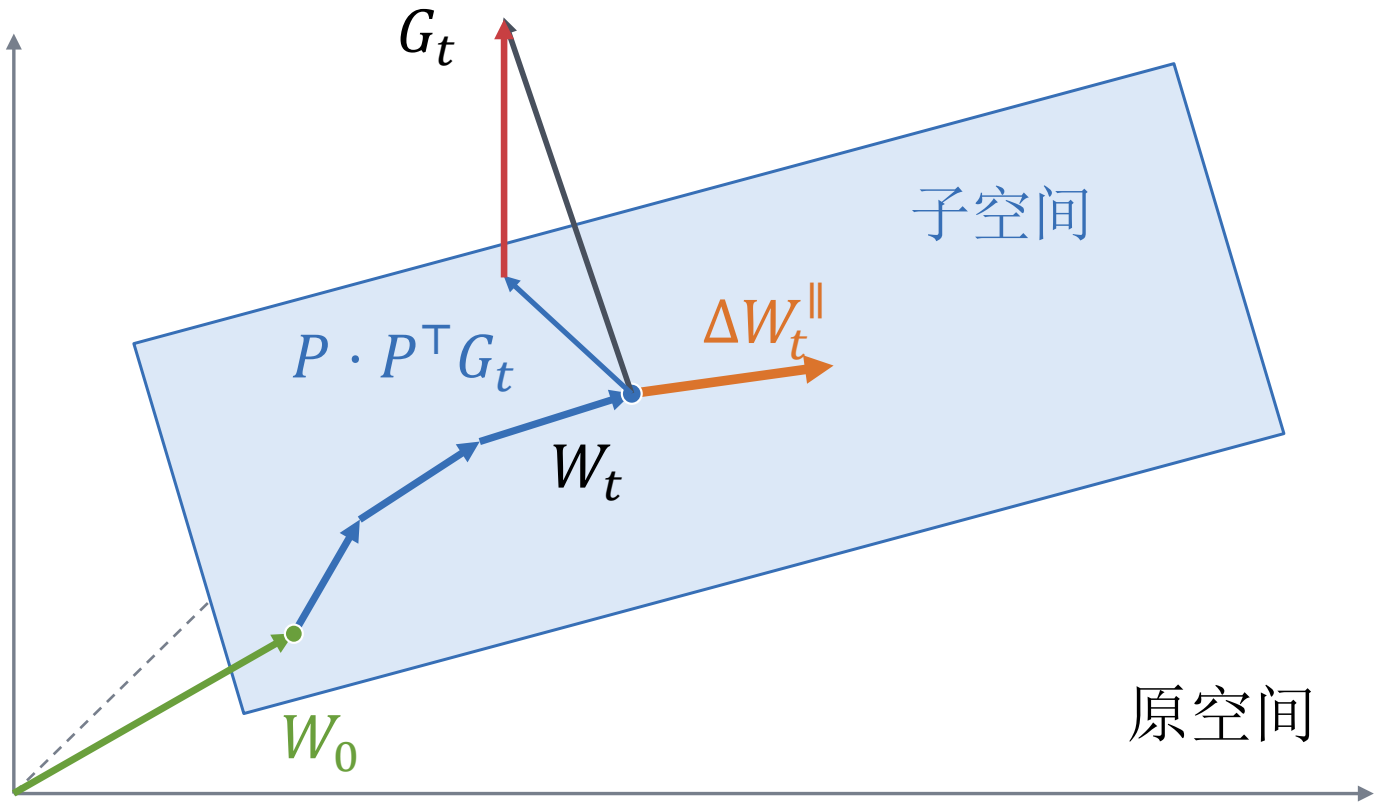
残差修正的子空间Muon

2. 子空间内更新

$$U_t = \mathrm{Muon}(R_t) \in \mathbb{R}^{r\times n}$$

$$\Delta W_t^{\parallel} = -\eta P \cdot U_t \in \mathbb{R}^{m\times n}$$

$$G_t = P \cdot \underbrace{P^{\top} G_t}_{\text{利用}} + (I - PP^{\top})G_t$$



残差修正的子空间Muon

2. 子空间内更新

$$U_t = \mathrm{Muon}(R_t) \in \mathbb{R}^{r\times n}$$

$$\Delta W_t^{\parallel} = -\eta P \cdot U_t \in \mathbb{R}^{m\times n}$$

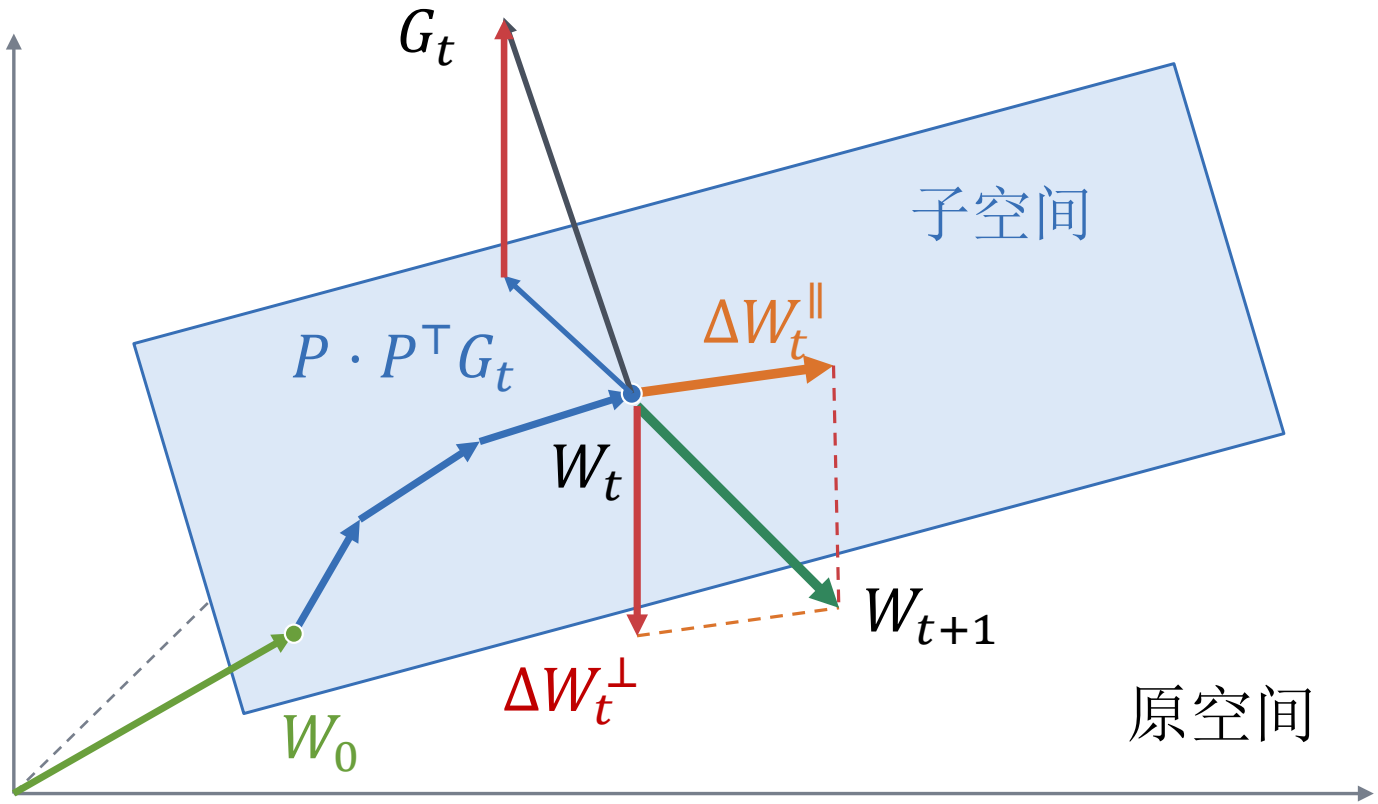
3. 补空间残差更新

$$\Delta W_t^{\perp} = -\gamma (I - PP^{\top}) G_t$$

4. 共同更新

$$W_{t+1} = W_t + \Delta W_t^{\parallel} + \Delta W_t^{\perp}$$

$$G_t = P \cdot \underbrace{P^{\top} G_t}_{\text{利用}} + \underbrace{(I - PP^{\top}) G_t}_{\text{利用}}$$



残差修正的子空间Muon

Muon

$$M_t = \mu M_{t-1} + G_t \in \mathbb{R}^{m\times n}$$

$$M_t = U_t \Sigma_t V_t^\top \in \mathbb{R}^{m\times n}$$

$$W_{t+1} = W_t - \eta U_t V_t^\top$$

Projected Muon with Correction

$$R_t = P^\top G_t \in \mathbb{R}^{r\times n}$$

$$M_t = \mu M_{t-1} + R_t \in \mathbb{R}^{r\times n}$$

$$M_t = U_t \Sigma_t V_t^\top \in \mathbb{R}^{r\times n}$$

$$W_{t+1} = W_t - \eta P \cdot U_t V_t^\top - \gamma(I - PP^\top)G_t$$

- 优化器空间： $mn \longrightarrow (m+n)r$
- SVD分解： $O(m^2 n) \longrightarrow O(r^2 n)$

收敛

理论保证

1. 经典噪声模型

$$\mathbb{E}_f\left[\|\nabla F(W) - G(W)\|_F^2\right] \leq \sigma^2$$ 噪声方差有界

2. 标准平滑性假设

$$\|\nabla F(W) - \nabla F(W')\|_F \leq L\|W - W'\|_F$$

3. 标准正交基

$$P^\top P = I_r \in \mathbb{R}^{r\times r}$$

➢ 收敛速率

$$\mathbb{E}\left[\left\|P^\top \nabla F(W_\tau)\right\|_* + \left\|(I - PP^\top)\nabla F(W_\tau)\right\|_F\right] = O\left(\frac{1}{T^{1/4}}\right)$$ 最优速率

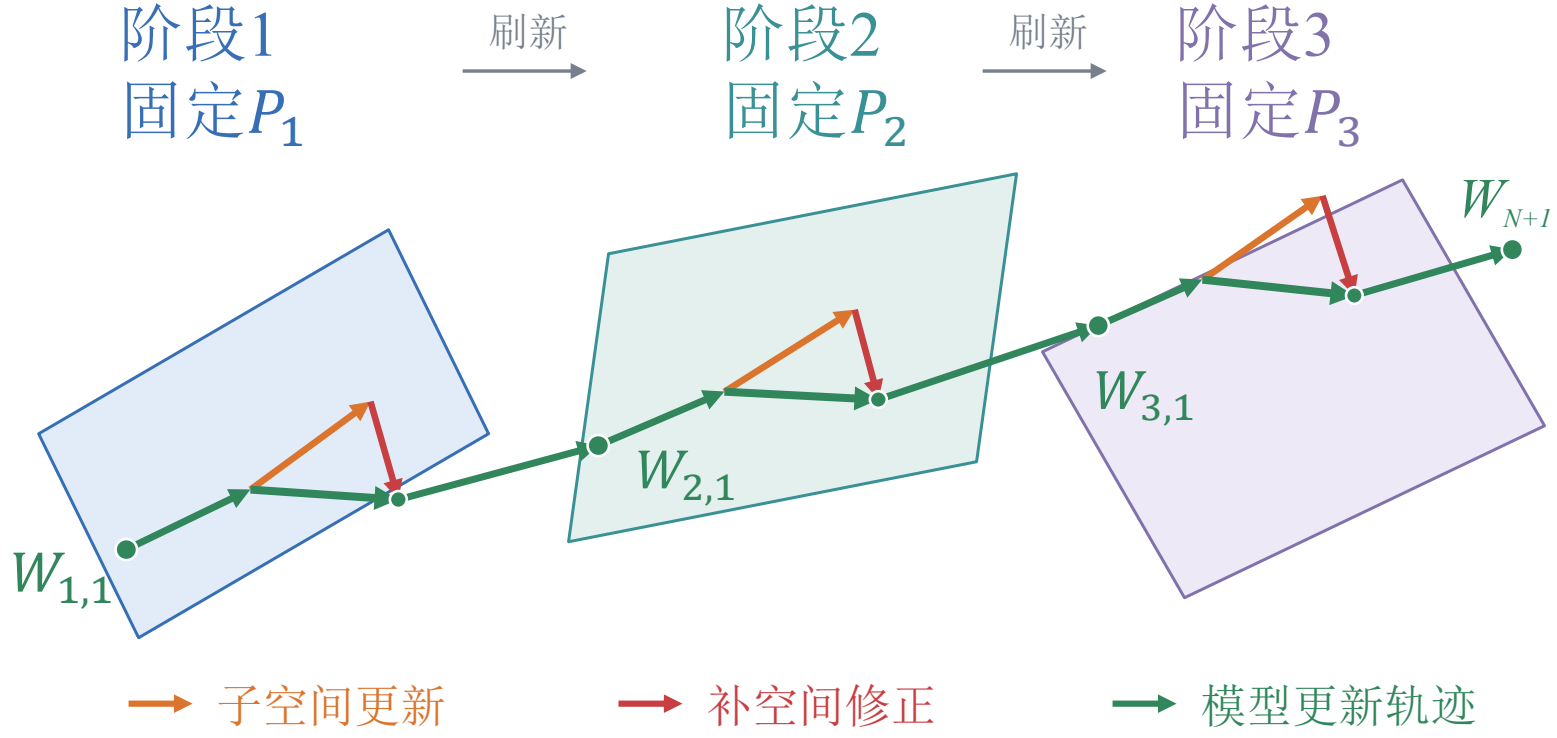
多阶段扩展

➢ 梯度的主导子空间随训练变化，需要周期性刷新投影矩阵P_k

1. 阶段内
   - ✓固定投影矩阵
   - ✓联合空间更新
2. 阶段间
   - ✓刷新投影矩阵
   - ✓继承模型参数

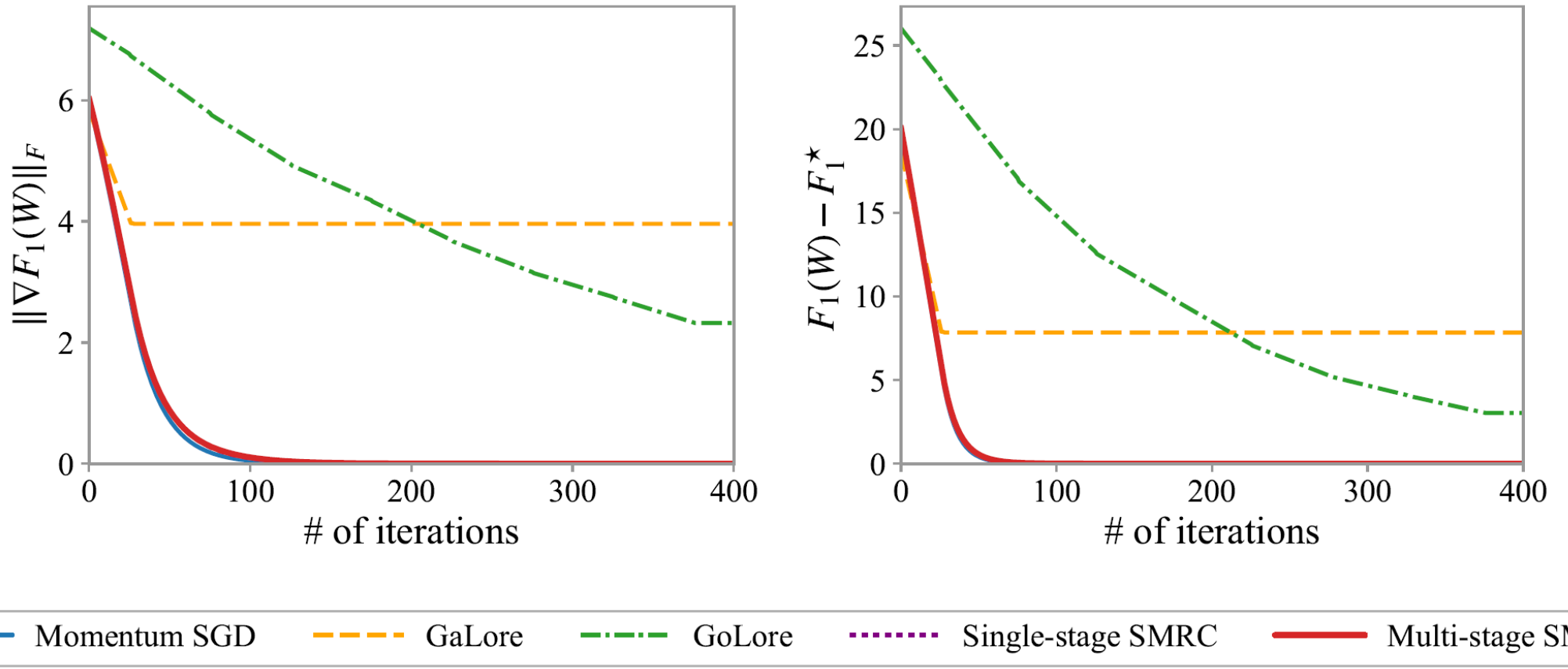
➢ 收敛速率

$$O\left(\frac{1}{T^{1/4}}\right)$$



收敛性验证实验

➢ 残差修正可以有效保证算法收敛



语言模型预训练实验

➢ 预训练LLaMA模型（60M、150M、350M、1.3B）

Loss Type	Memory-Efficient	Optimizer	60M	150M	350M	1.3B
Training Loss	×	AdamW	3.445	3.156	2.843	2.551
	×	Muon	3.383	3.086	2.703	2.414
	✓	GaLore	3.484	3.180	2.972	2.842
	✓	GoLore	3.516	3.196	3.003	3.080
	✓	Fira	3.601	3.180	2.852	2.657
	✓	GUM	3.461	3.141	2.871	2.609
	✓	**PMRC (Ours)**	**3.406**	**3.117**	**2.769**	**2.500**
Validation Loss	×	AdamW	3.421	3.134	2.892	2.613
	×	Muon	3.352	3.067	2.746	2.495
	✓	GaLore	3.462	3.202	2.980	2.875
	✓	GoLore	3.494	3.204	3.015	3.110
	✓	Fira	3.576	3.161	2.909	2.730
	✓	GUM	3.437	3.156	2.881	2.656
	✓	**PMRC (Ours)**	**3.380**	**3.095**	**2.786**	**2.570**

Yang et al. Memory-Efficient Subspace Optimization with Residual Correction: From Momentum SGD to Muon, https://arxiv.org/abs/2610.19844, 2026

语言模型微调实验

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

➢有监督微调Qwen3-8B、LLaMA3.1-8B、OLMo-2-1124-7B

	IFEval			GSM8K		
Optimizer	**Qwen3-8B**	**LLaMA-3.1-8B**	**OLMo-2-1124-7B**	**Qwen3-8B**	**LLaMA-3.1-8B**	**OLMo-2-1124-7B**
AdamW	65.83	50.12	38.97	90.14	72.10	77.63
Muon	62.11	46.16	36.45	89.76	72.78	78.01
GaLore	66.67	47.24	<u>38.13</u>	89.16	<u>71.27</u>	77.86
GoLore	<u>68.35</u>	<u>47.96</u>	35.73	88.93	70.66	77.10
GUM	62.47	43.65	36.21	<u>89.61</u>	69.74	<u>78.32</u>
PMRC (Ours)	**69.06**	**50.24**	**38.97**	**90.30**	**72.48**	**78.62**

指令跟随任务　　数学推理任务

算法整体内存对比

➢ 预训练LLaMA模型（1.3B）的整体内存

Model	AdamW	Muon	GaLore	GoLore	Fira	GUM	PMRC
1.3B	57.77	55.52	54.00	54.00	54.00	53.85	**53.44**

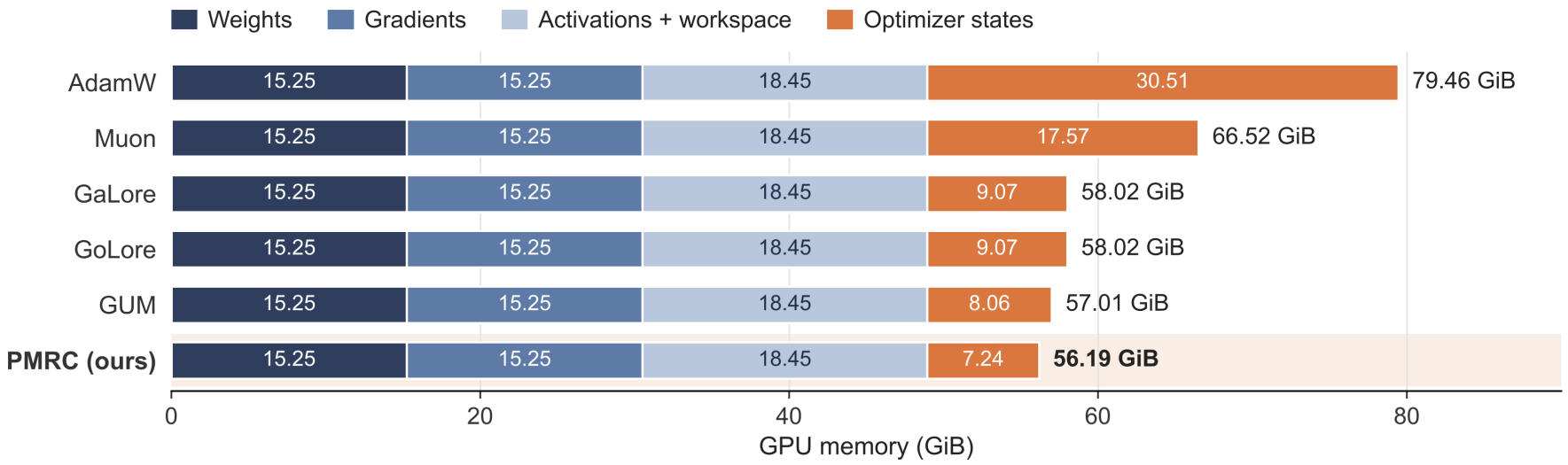
➢ 有监督微调的整体内存

Task	Model	AdamW	Muon	GaLore	GoLore	GUM	PMRC
IFEval	Qwen3-8B	89.16	76.23	67.71	67.71	66.01	**65.88**
	LLaMA-3.1-8B	78.32	65.32	56.63	56.63	54.90	**54.81**
	OLMo-2-1124-7B	69.74	57.43	49.23	49.23	47.79	**47.75**
GSM8K	Qwen3-8B	79.46	66.52	58.02	58.02	57.00	**56.18**
	LLaMA-3.1-8B	74.98	61.98	53.30	53.30	52.35	**51.48**
	OLMo-2-1124-7B	66.77	54.71	46.54	46.54	45.80	**45.03**

优化器内存对比

➢ 有监督微调优化器内存

Task	Model	AdamW	Muon	GaLore	GoLore	GUM	PMRC
IFEval	Qwen3-8B	30.51	17.57	9.07	9.07	7.37	**7.24**
	LLaMA-3.1-8B	29.92	16.92	8.23	8.23	6.50	**6.42**
	OLMo-2-1124-7B	27.19	15.13	6.96	6.96	5.49	**5.45**
GSM8K	Qwen3-8B	30.51	17.57	9.07	9.07	8.06	**7.24**
	LLaMA-3.1-8B	29.92	16.92	8.23	8.23	7.28	**6.42**
	OLMo-2-1124-7B	27.19	15.13	6.96	6.96	6.22	**5.45**



Yang et al. Memory-Efficient Subspace Optimization with Residual Correction: From Momentum SGD to Muon, https://arxiv.org/abs/2610.19844, 2026

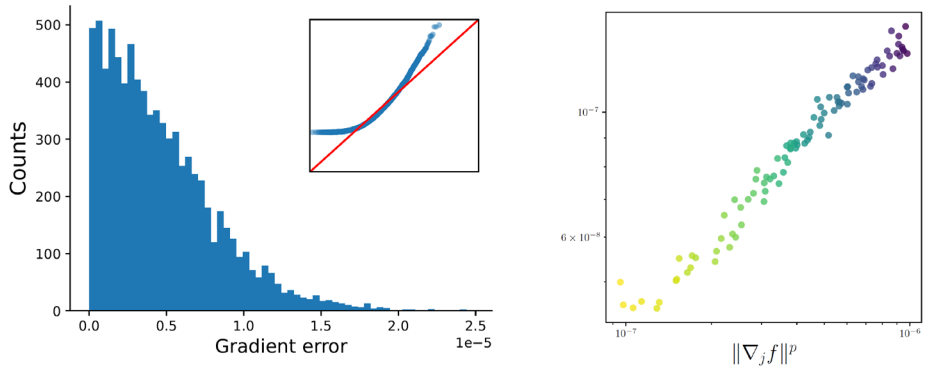
目录

- 研究背景
- 重尾噪声下的收敛理论
- 残差修正的子空间优化
- 总结与展望

总结

➢ 重尾噪声下的收敛理论

✓广义重尾噪声模型

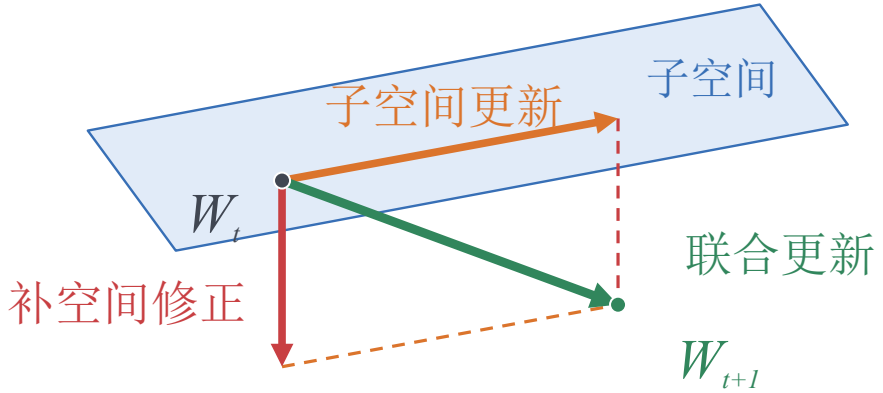


✓最优收敛速率

$$\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_1] = O\left(T^{-\frac{p-1}{3p-2}}\right)$$

➢ 残差修正的子空间优化

✓子空间更新、补空间修正



✓最优收敛速率

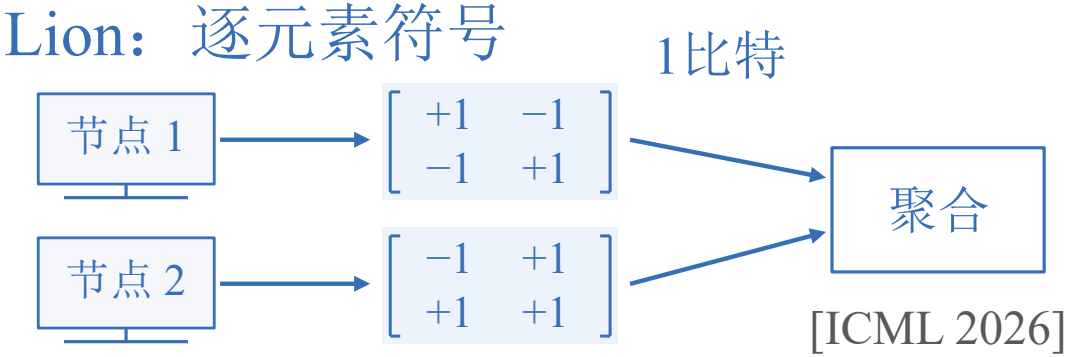
$$\mathbb{E}\left[\|\nabla^{\parallel}\|_* + \|\nabla^{\perp}\|_F\right] = O\left(\frac{1}{T^{1/4}}\right)$$

Yu et al. Sign-Based Optimizers Under Heavy-Tailed Noise: On the Effectiveness of Lion and Muon, NeurIPS 2026
Yang et al. Memory-Efficient Subspace Optimization with Residual Correction: From Momentum SGD to Muon, https://arxiv.org/abs/2610.19844, 2026

展望

通信高效的分布式优化

Lion：逐元素符号



[ICML 2026]

Muon：矩阵符号



优化器的自动选择与组合

