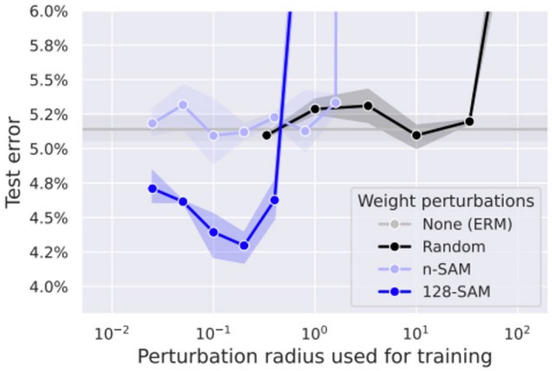
Towards Understanding Sharpness-Aware Minimization

Maksym Andriushchenko (EPFL), Nicolas Flammarion (EPFL)

1. ***m*-sharpness** matters in m-SAM

$$m\textbf{-SAM}: \min_{w\in\mathbb{R}^{|w|}} \sum_{\substack{\mathcal{S}\subset\mathcal{S}_{train},\\ |\mathcal{S}|=m}} \max_{\|\delta\|_2\le\rho} \sum_{i\in\mathcal{S}} \ell_i(w+\delta)$$

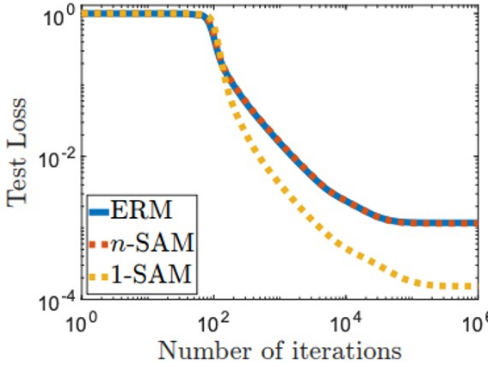
ResNet-18 on CIFAR-10



⚠️ The PAC-Bayes generalization bound doesn't explain this

2. The **implicit bias** of 1-SAM vs. n-SAM and ERM can be well understood for diagonal linear networks

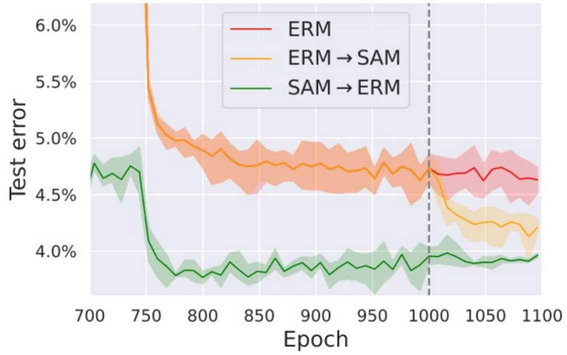
Diagonal linear network



💡 Simple models can be surprisingly predictive

3. m-SAM has some interesting effects: running ERM → SAM **gradually improves generalization**

ResNet-18 on CIFAR-10



❗ The same also happens for diagonal linear networks

Background: Sharpness-Aware Minimization

- Sharpness-Aware Minimization (SAM) [Foret et al., ICLR'21]:

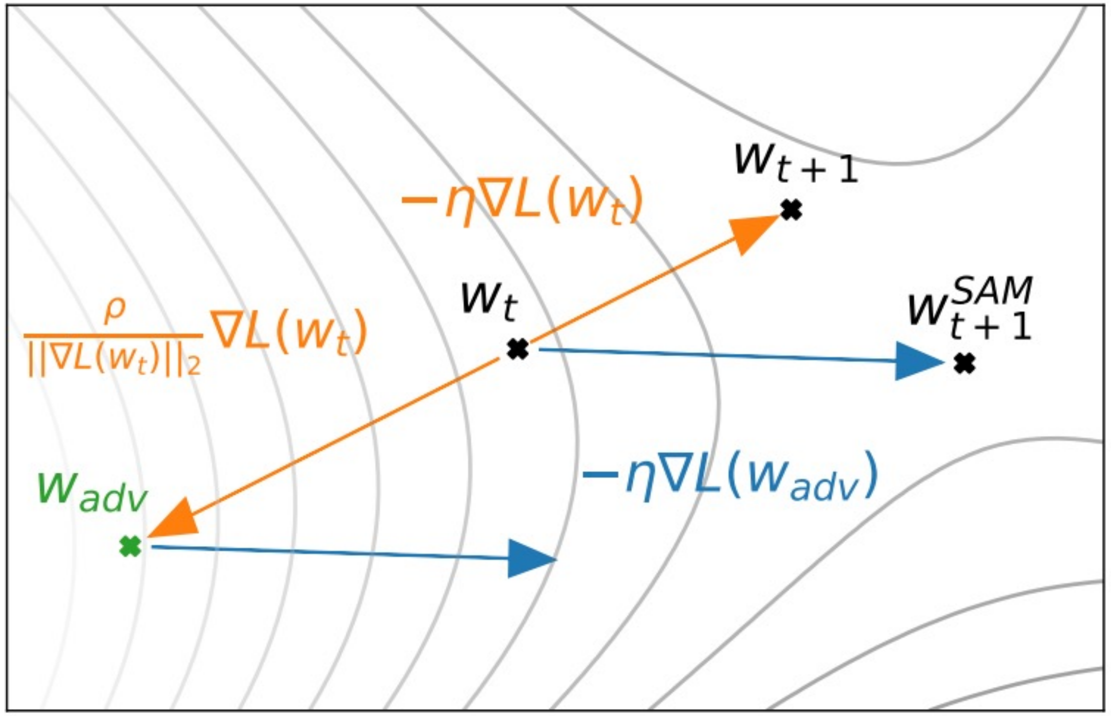
$$w_{t+1} = w_t - \frac{\gamma_t}{|I_t|} \sum_{i \in I_t} \nabla \ell_i \left(w_t + \frac{\rho_t}{|I_t|} \sum_{j \in I_t} \nabla \ell_j (w_t) \right)$$

- Foret et al., ICLR'21 motivate SAM by minimization of **sharpness:**

$$\min_{w \in \mathbb{R}^{|w|}} \max_{\|\delta\|_2 \leq \rho} \frac{1}{n} \sum_{i=1}^{n} \ell_i(w + \delta)$$

- SAM consistently **improves generalization** in the state-of-the-art settings (!) and has only 2x computational overhead

Visual description of the SAM algorithm

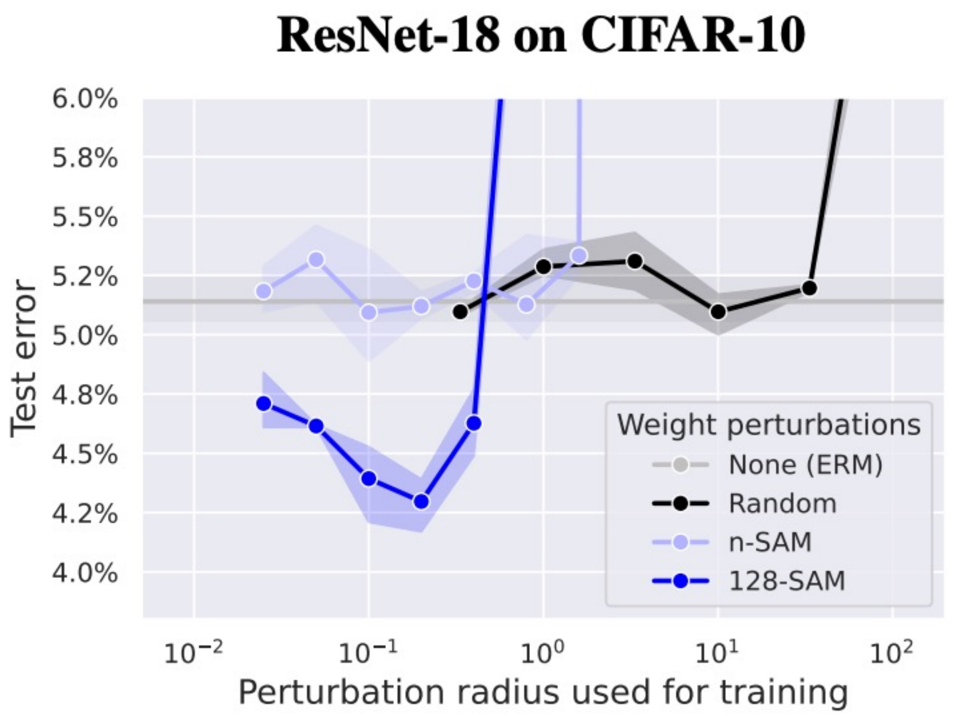


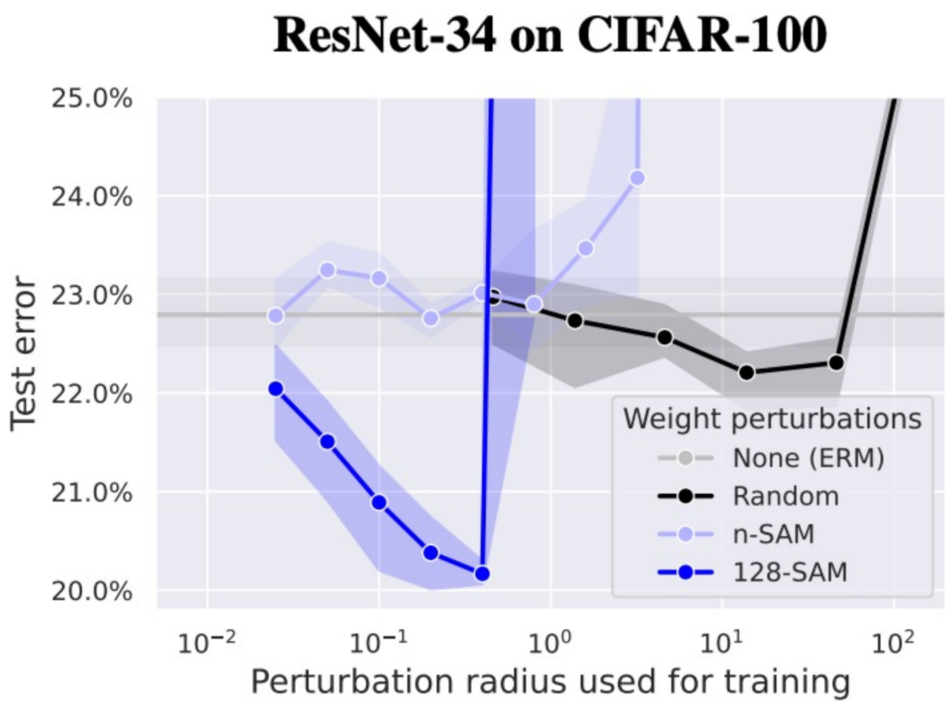
Source: Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR'21

Which components of SAM are crucial?

$$n\text{-SAM: } \min_{w\in\mathbb{R}^{|w|}} \max_{\|\delta\|_2\le\rho} \sum_{i=1}^{n} \ell_i(w+\delta) \quad \rightarrow \quad m\text{-SAM: } \min_{w\in\mathbb{R}^{|w|}} \sum_{\substack{\mathcal{S}\subset\mathcal{S}_{train},\\ |\mathcal{S}|=m}} \max_{\|\delta\|_2\le\rho} \sum_{i\in\mathcal{S}} \ell_i(w+\delta)$$

Worst-case weight perturbations, with a small m (aka $\boldsymbol{m}$**-sharpness**) are key!

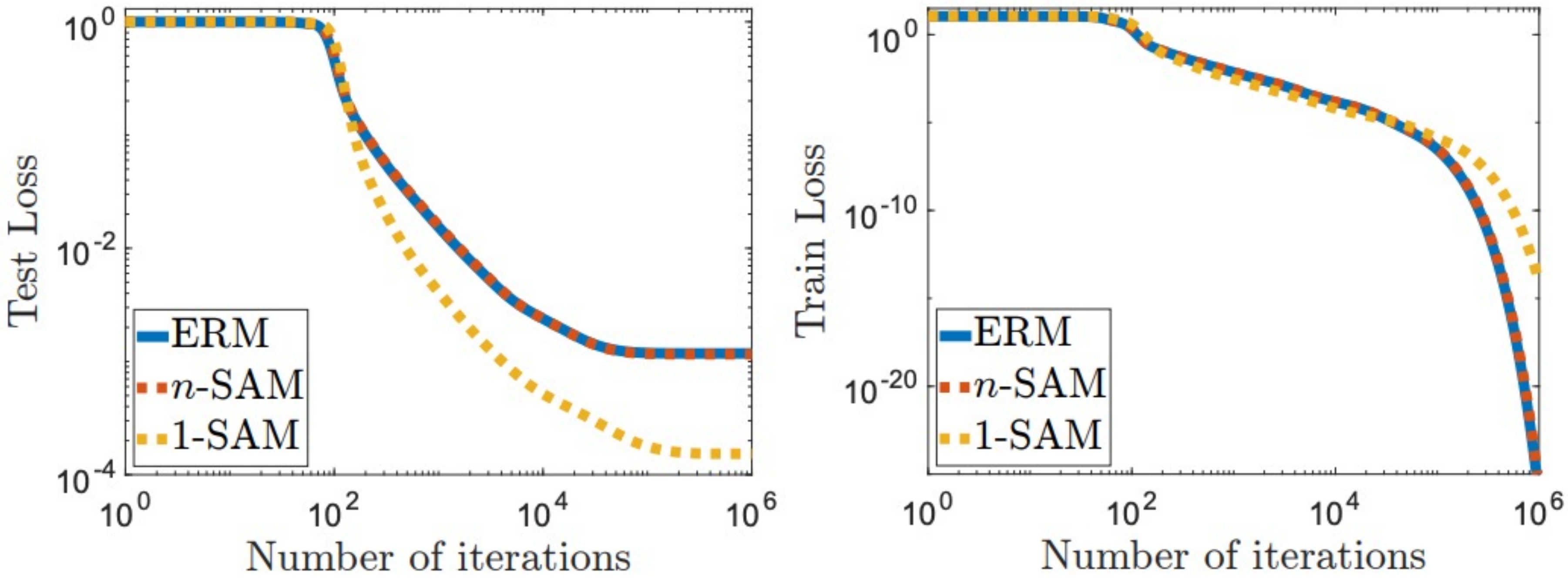




Understanding m-SAM on simple models

We will use **diagonal linear networks $f(x) = \langle x, u \odot v \rangle$** for sparse regression that shows different generalization depending on the initialization scale and SGD noise

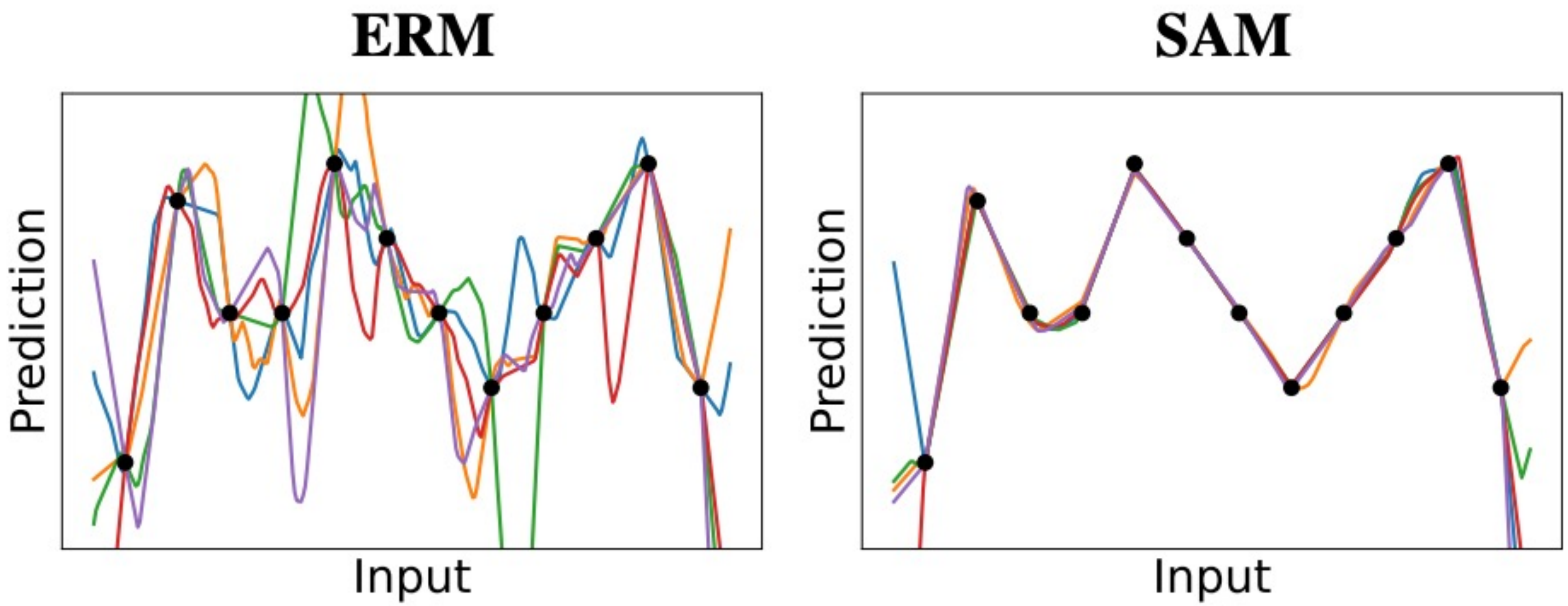
1-SAM for $f(x)$ generalizes significantly better than ERM and n-SAM!



We are also able to capture it **theoretically**: 1-SAM promotes **sparsity** in terms of the linear predictor $u \odot v$ (and much more than n-SAM)

m-SAM for 2-layer ReLU networks: sparsity bias

For **non-linear** networks, we can observe some interesting properties empirically

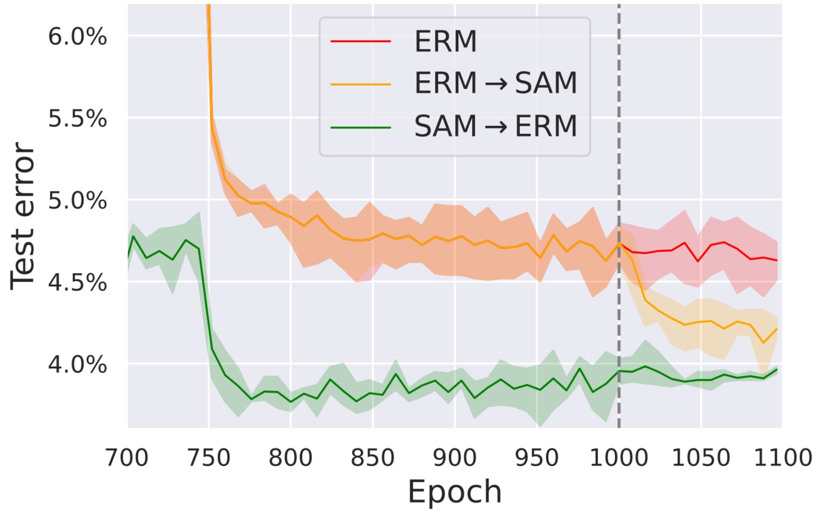


Using SAM for 2-layer ReLU networks on simple 1D regression also leads to a **sparsifying effect** but in terms of the **ReLU kinks**

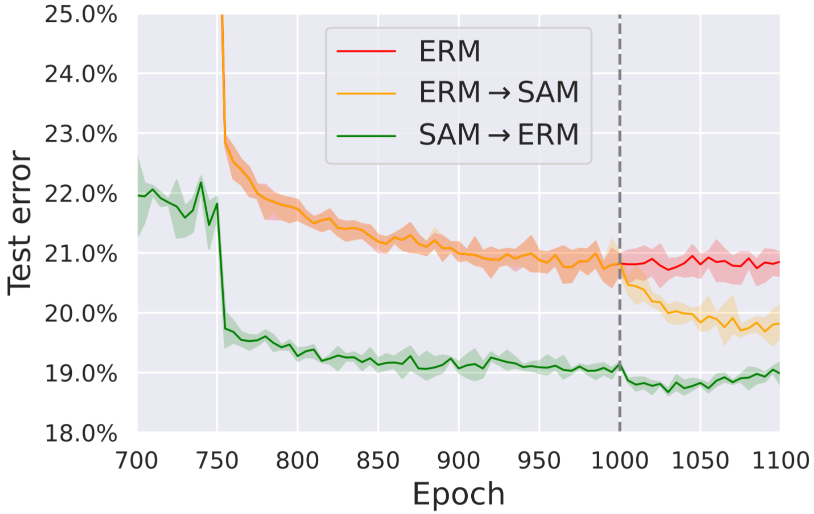
m-SAM for deep networks: interesting properties

A curious property of SAM: if we finetune an ERM model with SAM <u>on the same dataset</u>, we get a **significant generalization improvement**

ResNet-18 on CIFAR-10



ResNet-34 on CIFAR-100



And it's not so mysterious: **the same** is observed also for diagonal linear networks (see the paper)!

Additional results in the paper

- We provide a convergence proof for SAM with **constant** inner step sizes

- For deep networks, we show that SAM with both **constant** and **gradient-normalized** inner step sizes has a similar behavior (zero training error and same generalization)

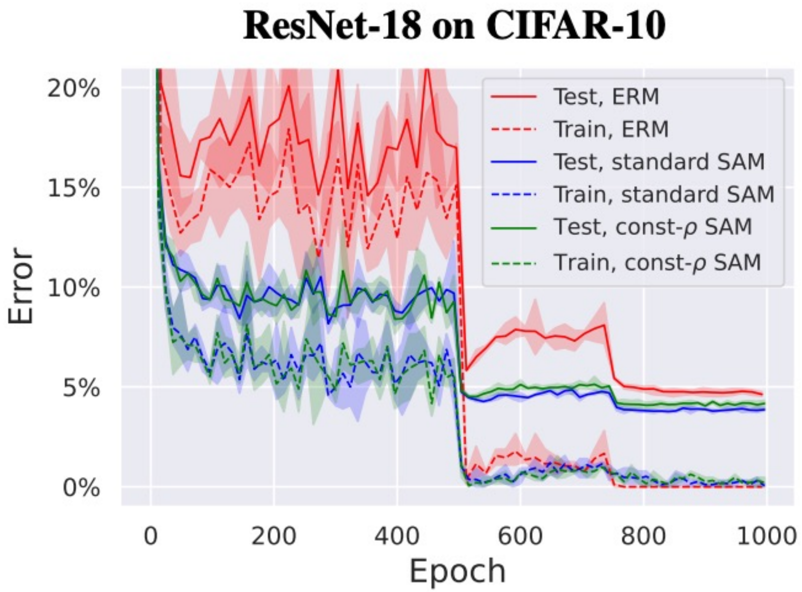
- Finally, convergence of SAM to global minima observed in practice can also have a **negative impact** → e.g., SAM overfits similarly to ERM when trained on noisy labelled datasets

Theorem 2. *Assume (A1) and (A2) for the iterates (4). Then for any number of iterations $T \geq 0$, batch size b, and step sizes $\gamma_t = \frac{1}{\sqrt{T}\beta}$ and $\rho_t = \frac{1}{T^{1/4}\beta}$, we have:*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla L(w_t)\|^2\right] \leq \frac{4\beta}{\sqrt{T}}(L(w_0)-L_*) + \frac{8\sigma^2}{b\sqrt{T}},$$

In addition, under (A3), with step sizes $\gamma_t = \min\{\frac{8t+4}{3\mu(t+1)^2}, \frac{1}{2\beta}\}$ and $\rho_t = \sqrt{\gamma_t/\beta}$:

$$\mathbb{E}[L(w_T)] - L_* \leq \frac{3\beta^2(L(w_0)-L_*)}{\mu^2T^2} + \frac{22\beta\sigma^2}{\mu^2 bT}.$$



Thanks for your attention!

Happy to answer your questions (in-person or virtually) :)

Paper: https://arxiv.org/abs/2206.06232
Code: https://github.com/tml-epfl/understanding-sam