Practical Sharpness-Aware Minimization Cannot Converge All the Way to Optima

Dongkuk Si, Chulhee Yun

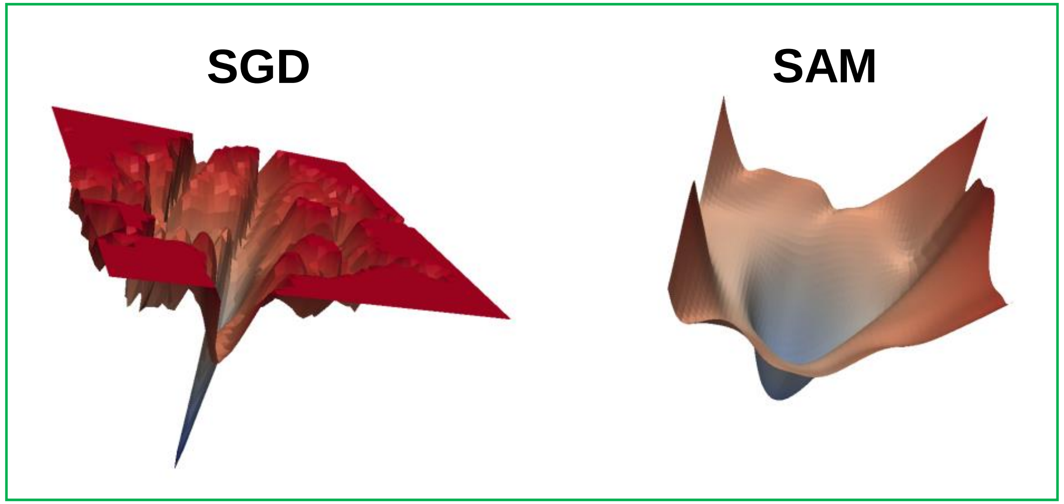
KAIST AI

Sharpness-Aware Minimization

Sharpness-Aware Minimization (SAM) [Foret et al., 2021]:

$$\min_x f^{\mathrm{SAM}}(x) = \min_x \max_{\|\epsilon\| \leq \rho} f(x+\epsilon)$$

$$x_{t+1} = x_t - \eta \nabla f\left(x_t + \rho \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}\right)$$



ρ is a perturbation size that is **fixed as a constant** in practice

[Foret et al., 2021] Sharpness-Aware Minimization for Efficiently Improving Generalization. ICLR 2021

Problem Setting

GOAL: Investigate SAM's convergence properties across various function classes

- **Deterministic SAM**:

$$x_{t+1} = x_t - \eta \nabla f \left(x_t + \rho \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|} \right)$$

- **Stochastic SAM**: given $f(x) = \mathbb{E}_\xi [l(x;\xi)]$ and $g(x) = \nabla_x l(x;\xi)$,

$$x_{t+1} = x_t - \eta g \left(x_t + \rho \frac{g(x_t)}{\|g(x_t)\|} \right)$$

- Focus on Practical Settings: **constant** ρ, **normalization** in ascending step

Problem Setting

GOAL: Investigate SAM's convergence properties across **various function classes**

For $f : \mathbb{R}^d \to \mathbb{R}$ and $\forall x, y \in \mathbb{R}^d$,

- **Smoothness**: $\exists \beta \geq 0$ such that $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$

- **Convexity**: $\lambda \in [0, 1]$, $f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$

- **Strong Convexity**: $\exists \mu > 0$ such that $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$

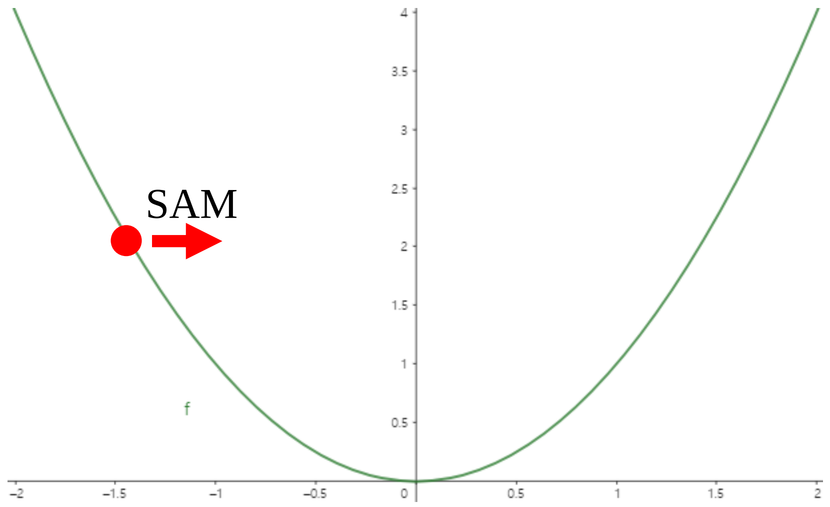
- **Lipschitzness**: $\exists L \geq 0$ such that $\|f(x) - f(y)\| \leq L \|x - y\|$

- **Bounded Gradient Variance**: $\exists \sigma \geq 0$ such that $\mathbb{E}_\xi \|\nabla f(x) - \nabla l(x; \xi)\|^2 \leq \sigma^2$

Virtual Loss

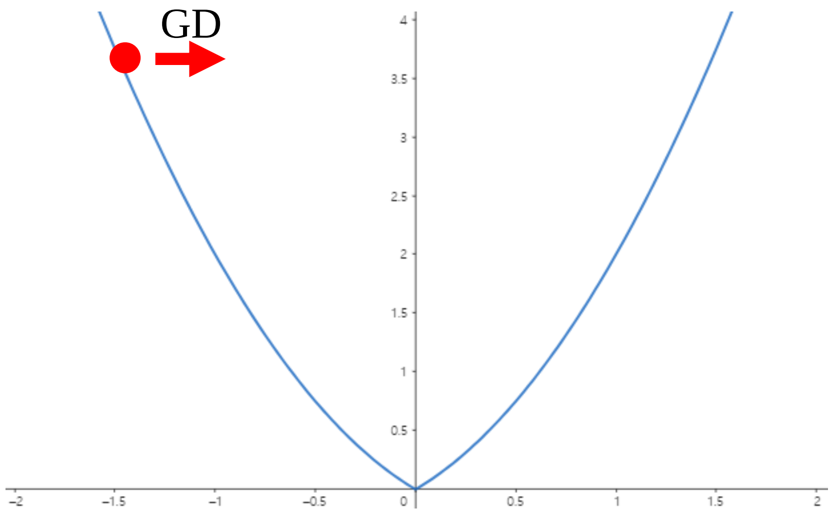
For continuous 1-dimensional function f, we can define the virtual loss J_f :

update of SAM
on $f(x)$



=

(sub)gradient descent
update on $J_f(x)$



Convergence of Deterministic SAM

Convergence of SAM with constant ρ after T steps

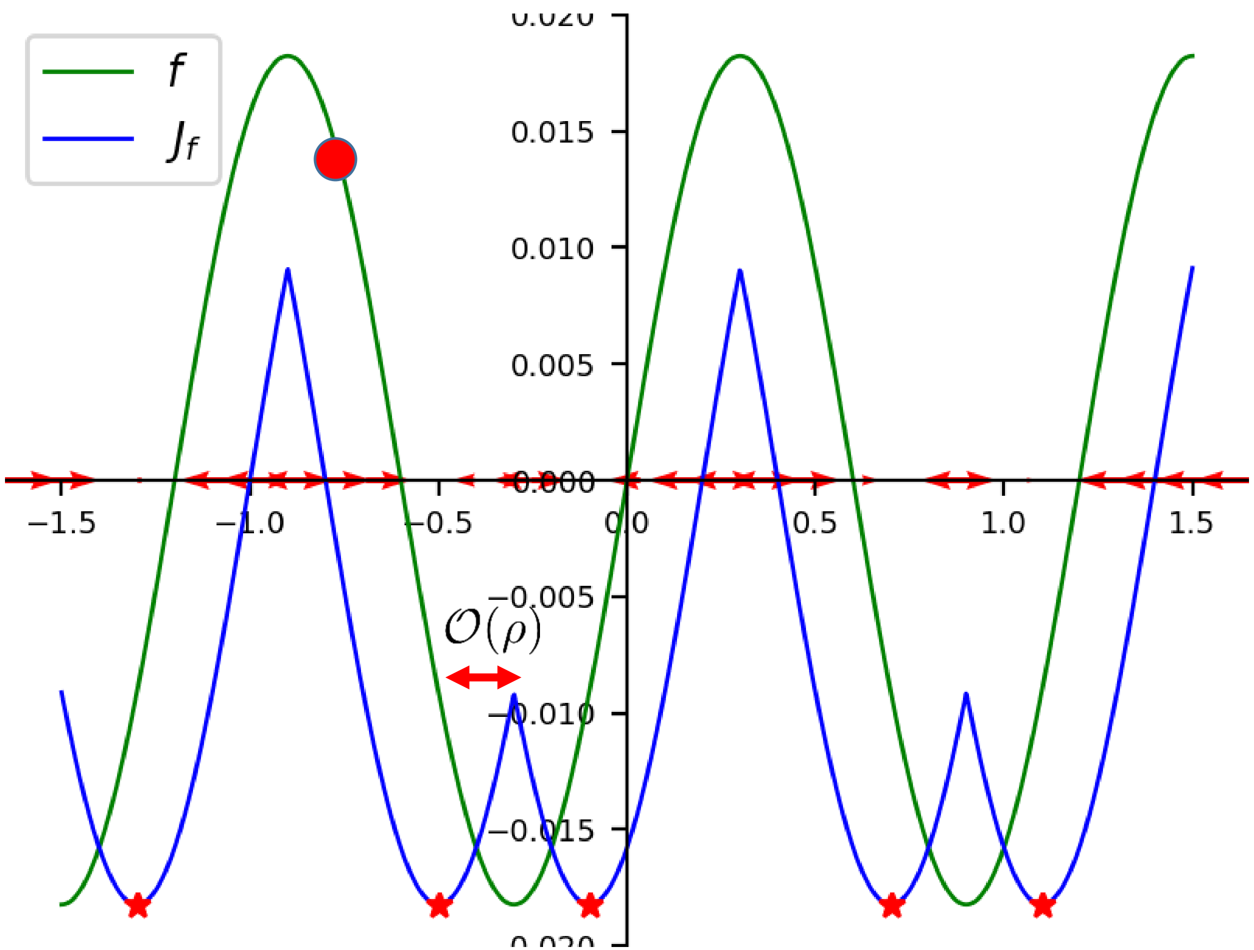
Optimizer	Function Class	Convergence Upper/Lower Bounds
Deterministic SAM	β-smooth, μ-strongly convex	$\min_{t\in\{0,\dots,T\}} f(\boldsymbol{x}_t) - f^* = \tilde{\mathcal{O}}\left(\exp(-T) + \frac{1}{T^2}\right)$
Deterministic SAM	β-smooth, μ-strongly convex	$\min_{t\in\{0,\dots,T\}} f(\boldsymbol{x}_t) - f^* = \Omega\left(\frac{1}{T^2}\right)$
Deterministic SAM	β-smooth, convex	$\frac{1}{T}\sum_{t=0}^{T-1} \\|\nabla f(\boldsymbol{x}_t)\\|^2 = \mathcal{O}\left(\frac{1}{T} + \frac{1}{\sqrt{T}}\right)$
Deterministic SAM	β-smooth	$\frac{1}{T}\sum_{t=0}^{T-1} \\|\nabla f(\boldsymbol{x}_t)\\|^2 \leq \mathcal{O}\left(\frac{1}{T}\right) + \beta^2\rho^2$

Additive factor $\mathcal{O}(\rho^2)$

Tight!

Non-Convergence Example

The additive factor is **tight** in terms of ρ



Convergence of Deterministic SAM

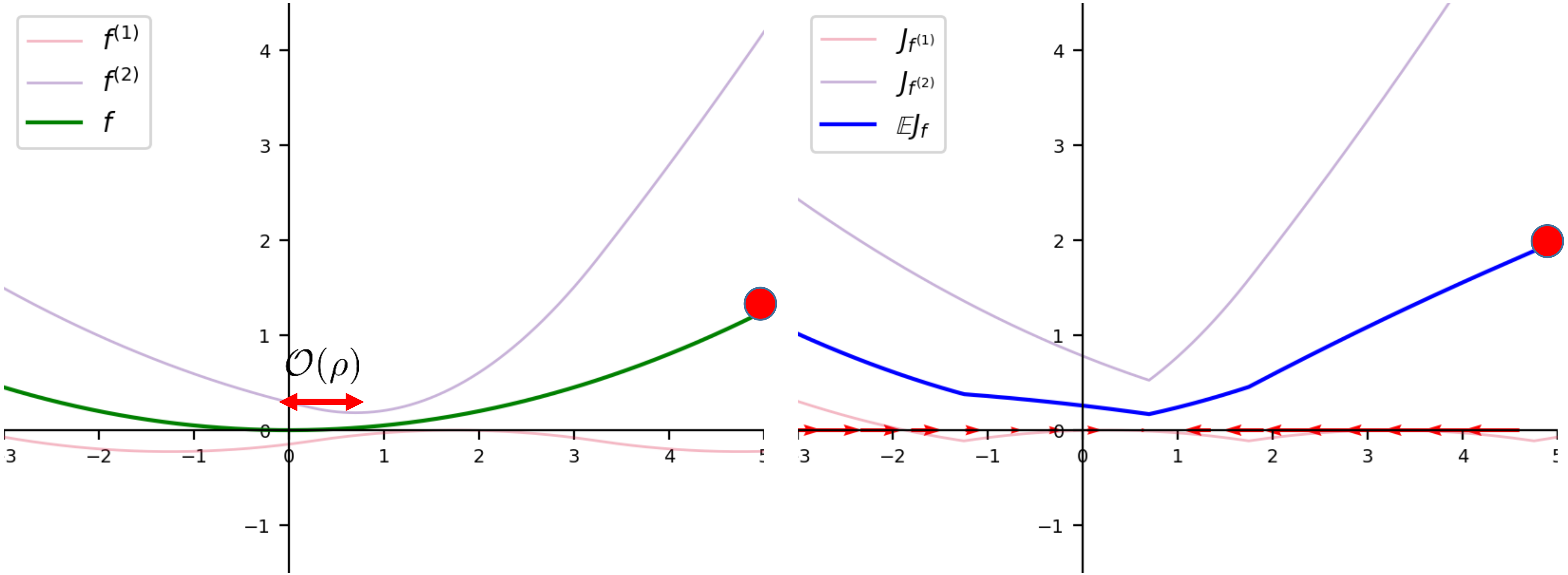
Convergence of SAM with constant ρ after T steps

Optimizer	Function Class	Convergence Upper/Lower Bounds
Stochastic SAM	β-smooth, μ-strongly convex	$\mathbb{E} f(\boldsymbol{x}_T) - f^* \leq \tilde{\mathcal{O}}\left(\exp(-T) + \frac{[\sigma^2 - \beta^2\rho^2]_+}{T}\right) + \boxed{\frac{2\beta^2\rho^2}{\mu}}$
Stochastic SAM	β-smooth, convex	$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{x}_t)\|^2 \leq \mathcal{O}\left(\frac{1}{T} + \frac{\sqrt{[\sigma^2 - \beta^2\rho^2]_+}}{\sqrt{T}}\right) + \boxed{4\beta^2\rho^2}$
Stochastic SAM	β-smooth, L-Lipschitz	$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[(\|\nabla f(\boldsymbol{x}_t)\| - \beta\rho)^2] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \boxed{5\beta^2\rho^2}$

Additive factors $\mathcal{O}(\rho^2)$

Non-Convergence Examples

The additive factors are **tight** in terms of ρ



Summary

- The convergence bound of deterministic / stochastic SAM suffers an **inevitable** additive term $\mathcal{O}(\rho^2)$, indicating convergence only up to neighborhoods of optima.

- Discover even more intriguing results in our paper!

Link to our paper