Stability Analysis of Sharpness-Aware Minimization

Hoki Kim
Chung-Ang Univ.
hokikim@cau.ac.kr

Jinseong Park
KIAS
jinseong@kias.re.kr

Yujin Choi
UNIST, NTU
uznhigh@snu.ac.kr

Jaewook Lee
Seoul National Univ.
jaewook@snu.ac.kr

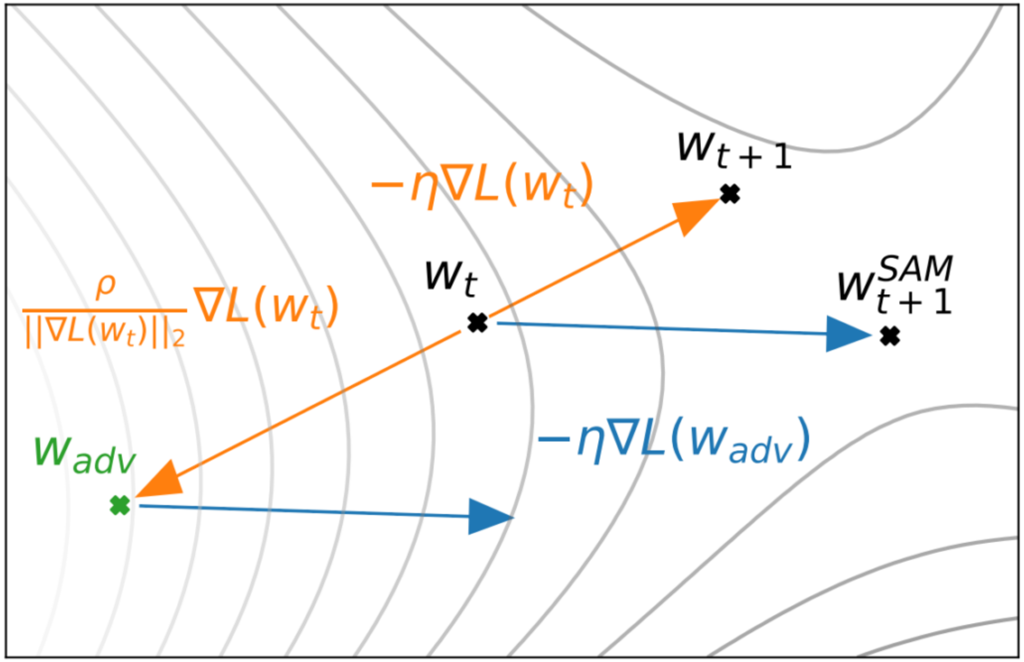
ICML 2026

ICML
International Conference
On Machine Learning

Presentation powered by Kokoro (https://github.com/hexgrad/kokoro)

1 Introduction

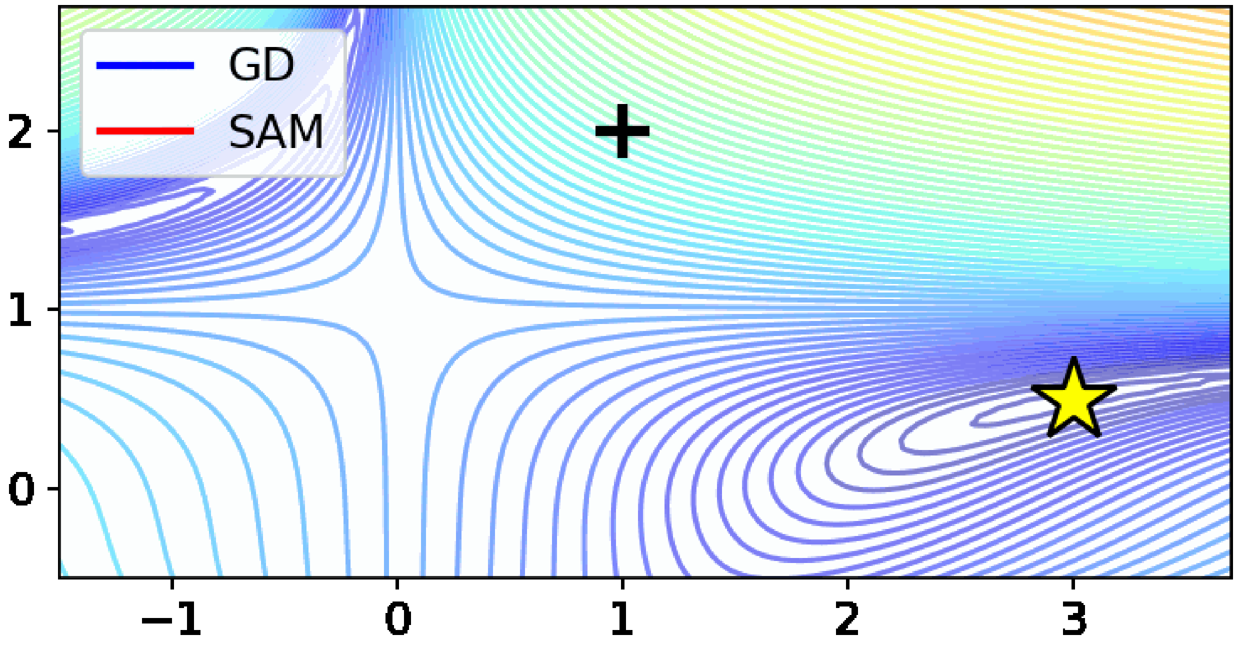
- Sharpness-aware minimization (SAM)
  - It seeks flat minima by minimizing the worst-case loss within a local parameter neighborhood, achieving strong performance across domains.



Foret, Pierre, et al. "Sharpness-aware Minimization for Efficiently Improving Generalization." International Conference on Learning Representations.

1 Introduction

- Sharpness-aware minimization (SAM)
  - ○ This paper analyzes SAM’s convergence instability near saddle points through dynamical systems theory.



1 Introduction

- Contributions:
    - We prove that saddle points can become attractors under SAM dynamics and show, via SAM diffusion, that SAM can escape saddle points less effectively than vanilla gradient descent.
    - We further demonstrate that momentum and batch size help mitigate this instability and improve generalization.
    - Experiments on standard optimization problems and benchmark tasks validate our theoretical and empirical findings.

2

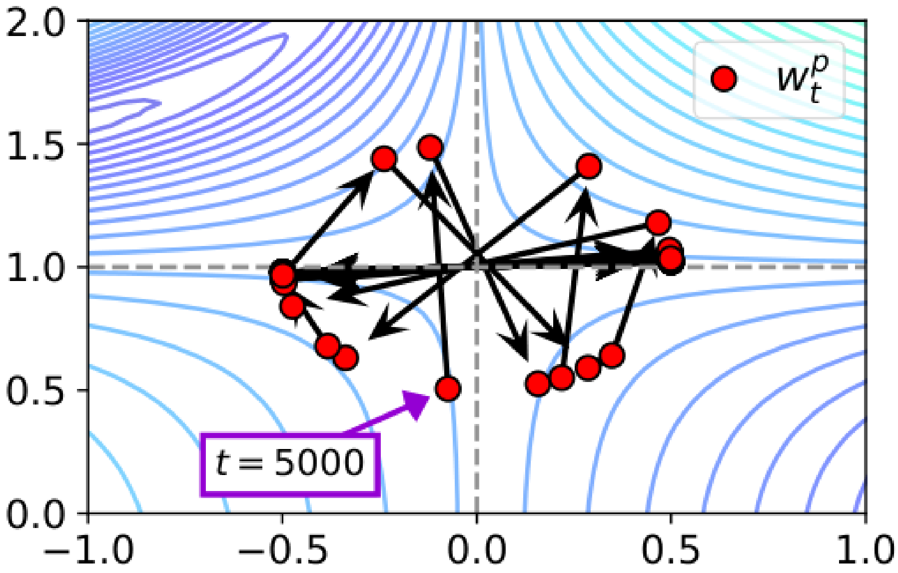
Saddle Point Becomes Attractor

-

- Preliminary
  - **SAM:**
    - (Inner max) $w_t^p = w_t + \rho \nabla \ell(w_t)$
    - (Outer min) $w_{t+1} = w_t - \eta \nabla \ell(w_t^p)$
  - **Beale function:**
    - $f(x) = (1.5 - x_1 + x_1 x_2)^2 + (2.25 - x_1 + x_1 x_2^2)^2 + (2.625 - x_1 + x_1 x_2^3)^2$

2 Saddle Point Becomes Attractor

- On the Beale function

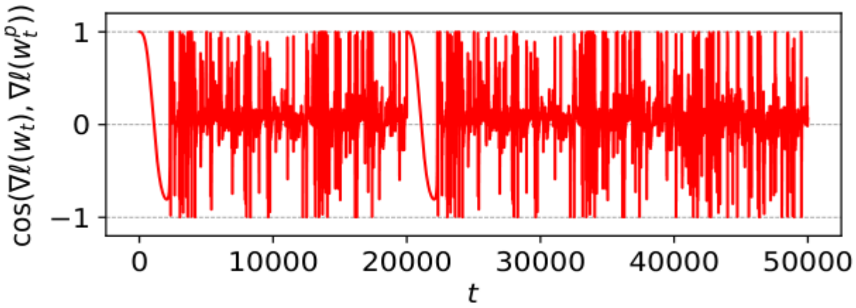


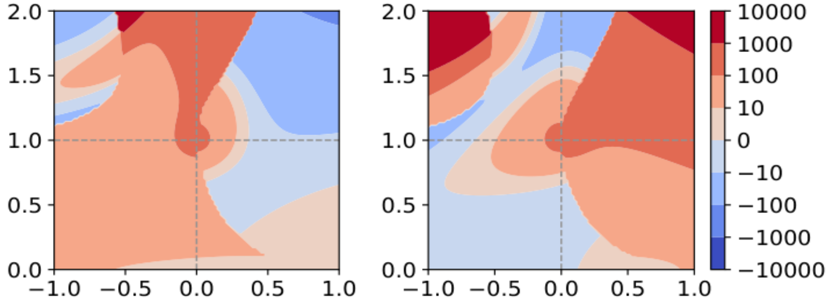
  - ○ Trajectory of w_t^p after the optimization step t = 5000, when SAM is beginning to become stuck in the saddle point during SAM optimization.

2

Saddle Point Becomes Attractor

- On the Beale function





(a) $\lambda_1 + \rho\lambda_1^2$ *(b)* $\lambda_2 + \rho\lambda_2^2$

- (Left) Gradient oscillation during SAM optimization. The line corresponds to $\cos(\nabla\ell(w_t), \nabla\ell(w_t^p))$ for optimization step t.
- (Right) Value of $\lambda + \rho\lambda^2$ for the Beale function. Near the saddle point at (0, 1), for both eigenvalues of the Hessian, $\lambda + \rho\lambda^2$ is positive, which indicates that the saddle point becomes an attractor.

2

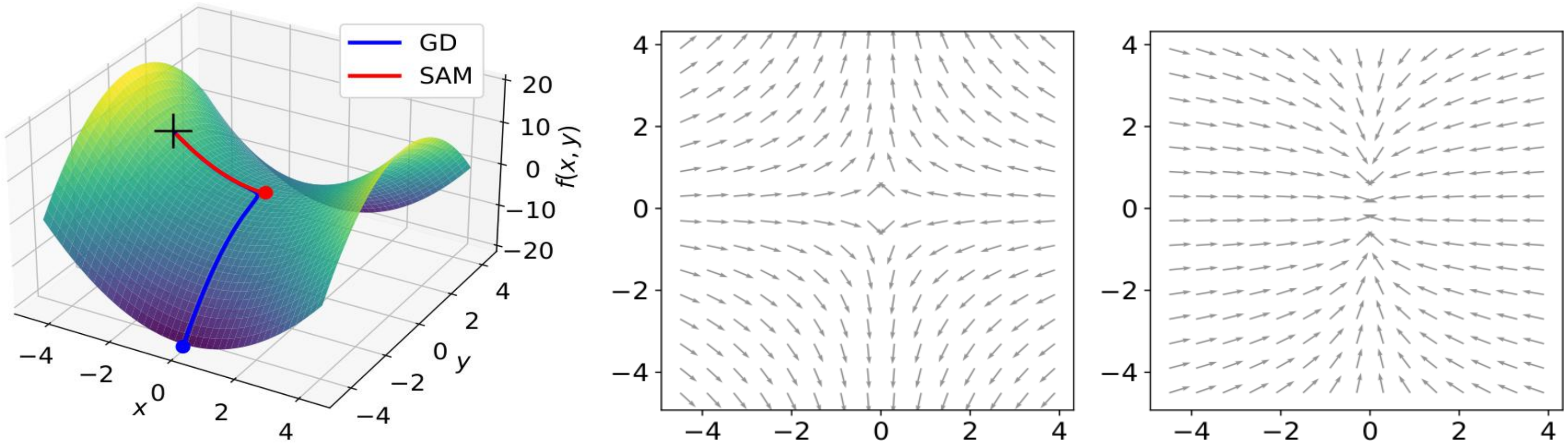
Saddle Point Becomes Attractor

- **Theorem 1.**
  - Let d be an index-one saddle point of system under SAM dynamics with a negative eigenvalue λ_1 of the Hessian matrix $H_\ell(d)$ of the loss function ℓ. Then**, the saddle point $\boldsymbol{d}$ is an attractor of SAM dynamics if $\boldsymbol{\rho > -1/\lambda_1}$, i.e., $\boldsymbol{\lambda_1 + \rho\lambda_1^2 > 0}$.**

2

Saddle Point Becomes Attractor

- **On a simple function $f(x,y) = x^2 - y^2$**



  - Optimization with the saddle point at (0,0). Initial point = (−3, −0.01).
  - Divergence of the gradient flow of GD near the saddle point.
  - Convergence of the gradient flow of SAM with ρ = 1.0 near the saddle point.

3 Stochastic Behavior of SAM

- SAM Diffusion
  - SAM Dynamics: $dw = -\nabla \ell(w^p)dt + [\eta C(w^p)]^{\frac{1}{2}} dW_t$.
  - Fokker-Planck Equation:
    $\frac{\partial P(w,t)}{\partial t} = \nabla \cdot [P(w,t)\nabla \ell(w^p)] + \nabla \cdot \nabla[D(w^p)P(w,t)]$.
- Theorem 2 & Corollary 1
  - Let us consider a saddle point $d = w_0$ as the initial parameter. (…) Let the mean squared displacement of SAM $\Delta_{SAM} := \langle \Delta w_j^2(t) \rangle_{SAM}$ and SGD $\Delta_{SGD} := \langle \Delta w_j^2(t) \rangle_{SGD}$.

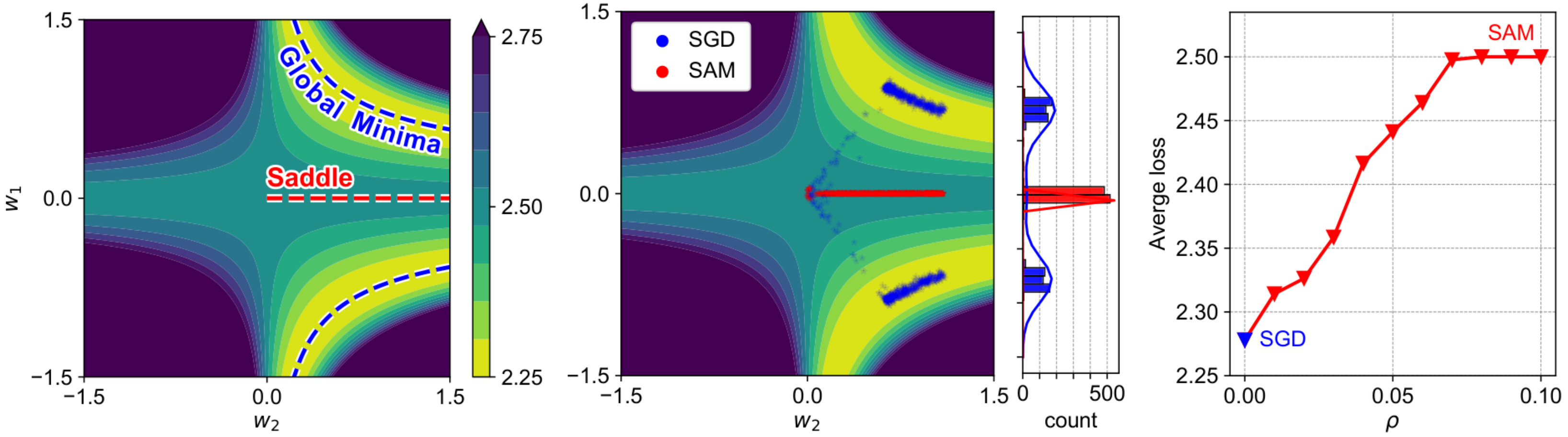
$$\Delta_{SGD} - \Delta_{SAM} = \frac{2\eta t^2 |\lambda_j|^3}{B}\rho + \mathcal{O}(B^{-1}\eta t^3 \lambda_j^4).$$

  - As $2\eta t^2 |\lambda_j|^3 \rho / B$ is always positive, the result implies that SAM escapes saddle points more slowly than SGD.

3

Stochastic Behavior of SAM

- 2-Layer Neural Network (Ziyin et al. , 2022)



(a) Loss landscape *(b)* Convergence distribution *(c)* Average loss for different ρ

  - ○ (Left) Loss landscape for different values of each neuron.
  - ○ (Middle) Distributions of converged points for SGD and SAM(ρ=0.1) with the marginal distribution of the parameter w_1.
  - ○ (Right) Average loss of converged points for different ρ, where ρ=0 indicates SGD.

3

Stochastic Behavior of SAM

- Theorem 3
  - **Momentum and Batch-size**
  - Use higher batch-size!
  - Use higher momentum!