An SDE for Modeling SAM: Theory and Insights

Enea Monzio Compagnoni, Luca Biggio, Antonio Orvieto,
Frank Norbert Proske, Hans Kersting, Aurelien Lucchi

June 28, 2023

Research Questions

SDEs have been used for optimal control of the learning rate, scaling rules (SGD, Adam, and RMSprop), exit times, and convergence bounds.

We use SDEs to address the following questions:

1. *How does the noise-curvature interaction help SAM escape sharp regions?*
2. *Are there any "traps" that could slow SAM down?*

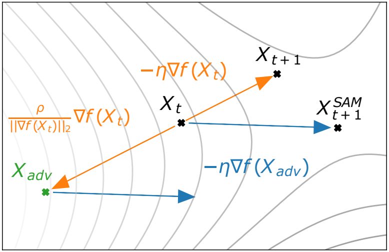
Recap on SAM

Sharpness-Aware Minimization (SAM) is a successful optimizer [3] which

1. Aims at minimizing the worst-case-sharpness:

$$\left[\max_{\|\epsilon\|_2 \leq \rho} f_{\mathcal{S}}(x+\epsilon) - f_{\mathcal{S}}(x)\right] \rightsquigarrow \textbf{Better Generalization}$$

2. Approximates the optimal perturbation with $\epsilon^*(x) \approx \rho \frac{\nabla f_{\mathcal{S}}(x)}{\|\nabla f_{\mathcal{S}}(x)\|}$



Recap on SAM

These are the SAM variants that we analyzed:

1. SAM:

$$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \frac{\nabla f_{\gamma_k}(x_k)}{\|\nabla f_{\gamma_k}(x_k)\|}\right)$$

2. USAM (Unnormalized SAM):

$$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}\left(x_k + \rho \nabla f_{\gamma_k}(x_k)\right)$$

Related Works

The understanding of the dynamics of SAM is drawing much-deserved attention and is in constant evolution. So far:

1. SAM **implicitly** minimizes a regularized loss which drives the dynamics toward **flatter** areas [2, 4]
2. Convergence results for different classes of functions [1]
3. ODE Approximations [1, 4]

Findings

For USAM, we have

$$dX_t = -\nabla \tilde{f}^{\text{USAM}}(X_t)dt + \left(I_d + \rho \nabla^2 f(X_t)\right)\left(\eta \Sigma^{\text{SGD}}(X_t)\right)^{1/2} dW_t,$$

where $\tilde{f}^{\text{USAM}}(x) := f(x) + \frac{\rho}{2}\|\nabla f(x)\|_2^2$

Findings

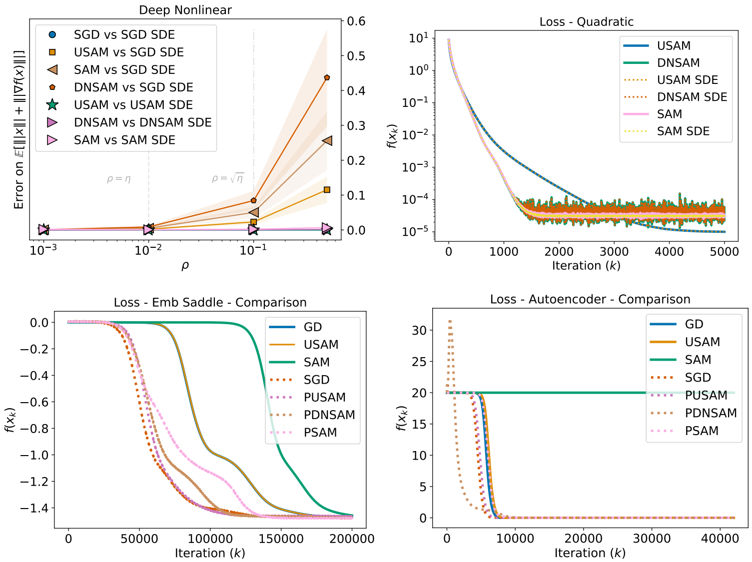
For SAM, we have

$$dX_t = -\nabla \tilde{f}^{\mathrm{SAM}}(X_t)dt + \sqrt{\eta \left(\Sigma^{\mathrm{SGD}}(X_t) + \rho H_t \left(\bar{\Sigma}(X_t) + \bar{\Sigma}(X_t)^\top\right)\right)}dW_t,$$

where

1. $H_t = \nabla^2 f(X_t)$
2. $\bar{\Sigma}(x) = \mathbb{E}\left[\left(\nabla f(x) - \nabla f_\gamma(x)\right) \cdot \left(\mathbb{E}\left[\frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right] - \frac{\nabla f_\gamma(x)}{\|\nabla f_\gamma(x)\|_2}\right)^\top\right]$
3. $\tilde{f}^{\mathrm{SAM}}(x) := f(x) + \rho \mathbb{E}\left[\|\nabla f_\gamma(x)\|_2\right]$

Experimental Validation



Conclusion

- **Implicit** regularization drives SAM towards **any** critical point
- The **implicit noise** of SAM scales with the local **curvature** $\rightsquigarrow$ **Helps to escape sharper areas**
- Might be **attracted by saddles** or at least, they might be **slower at escaping** them than SGD
- Much more to do on this topic!

References

[1] Maksym Andriushchenko and Nicolas Flammarion. "Towards understanding sharpness-aware minimization". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 639–668.

[2] Peter L Bartlett, Philip M Long, and Olivier Bousquet. "The Dynamics of Sharpness-Aware Minimization: Bouncing Across Ravines and Drifting Towards Wide Minima". In: *arXiv preprint arXiv:2210.01513* (2022).

[3] Pierre Foret et al. "Sharpness-aware minimization for efficiently improving generalization". In: *ICLR 2021* (2021).

[4] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. "How Does Sharpness-Aware Minimization Minimize Sharpness?" In: *ICLR 2023* (2023).