



清华大学 交叉信息研究院

Institute for Interdisciplinary Information Sciences, Tsinghua University



# The Crucial Role of Normalization in Sharpness-Aware Minimization

Yan Dai <sup>\*</sup> <sup>1</sup>, Kwangjun Ahn <sup>\*</sup> <sup>2</sup>, Suvrit Sra <sup>2</sup> <sup>3</sup>

<sup>1</sup> IIS, Tsinghua

<sup>2</sup> EECS, MIT


<sup>3</sup> TU Munich

Presented by Yan Dai

# Sharpness-Aware Minimization (SAM)

- Introduced by [Foret et al. \[2021\]](#) that performs sequential updates to loss function  $L$ :

$$w_{t+1} = w_t - \eta \nabla L \left( w_t + \frac{\rho \nabla L(w_t)}{\|\nabla L(w_t)\|_2} \right), \forall t = 1, 2, \dots \quad (1)$$


  
**Normalization Factor**

- Very impressive performance in training deep neural networks to generalize well

# Sharpness-Aware Minimization (SAM)

- Introduced by [Foret et al. \[2021\]](#) that performs sequential updates to loss function  $L$ :

$$w_{t+1} = w_t - \eta \nabla L \left( w_t + \frac{\rho \nabla L(w_t)}{\|\nabla L(w_t)\|_2} \right), \forall t = 1, 2, \dots \quad (1)$$

  
**Normalization Factor**

- Very impressive performance in training deep neural networks to generalize well
- Theoretical analyses were conducted towards characterizing SAM dynamics & properties, while most of them **removes normalization** [[Andriushchenko & Flammarion; 2022](#)] as:


$$w_{t+1} = w_t - \eta \nabla L(w_t + \rho \nabla L(w_t)), \forall t = 1, 2, \dots \quad (2)$$

- The simplified version (Unnormalized SAM, or **USAM**) gives elegant theoretical results

# Sharpness-Aware Minimization (SAM)

- Introduced by [Foret et al. \[2021\]](#) that performs sequential updates to loss function  $L$ :

$$w_{t+1} = w_t - \eta \nabla L \left( w_t + \frac{\rho \nabla L(w_t)}{\|\nabla L(w_t)\|_2} \right), \forall t = 1, 2, \dots \quad (1)$$

  
**Normalization Factor**

- Very impressive performance in training deep neural networks to generalize well
- Theoretical analyses were conducted towards characterizing SAM dynamics & properties, while most of them **removes normalization** [[Andriushchenko & Flammarion; 2022](#)] as:

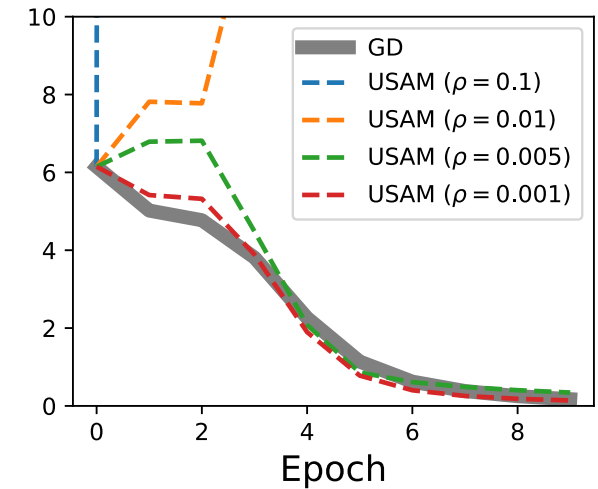
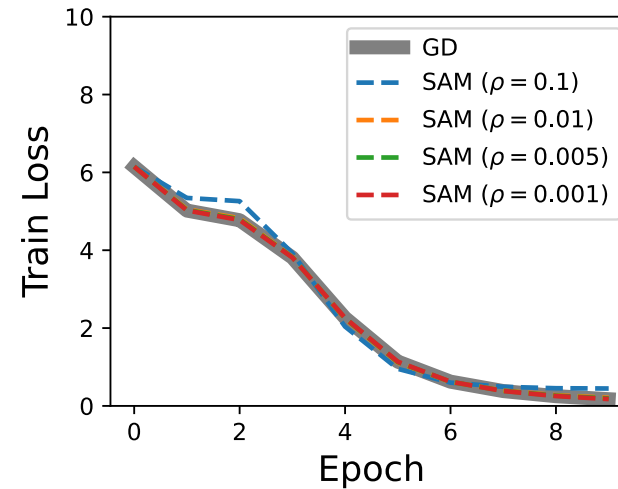
$$w_{t+1} = w_t - \eta \nabla L(w_t + \rho \nabla L(w_t)), \forall t = 1, 2, \dots \quad (2)$$

- The simplified version (Unnormalized SAM, or **USAM**) gives elegant theoretical results
- **Question:** What's the **role of normalization** (i.e., factor  $1/\|\nabla L(w_t)\|$ ) in SAM update (1)?
  - **In other words...** Whether the simplification in (2) can be safely adopted to simplify analysis?

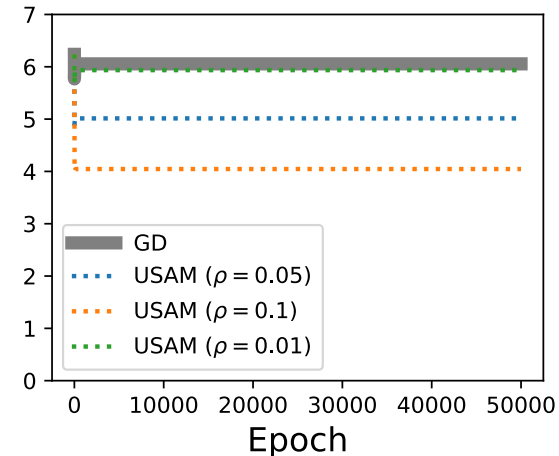
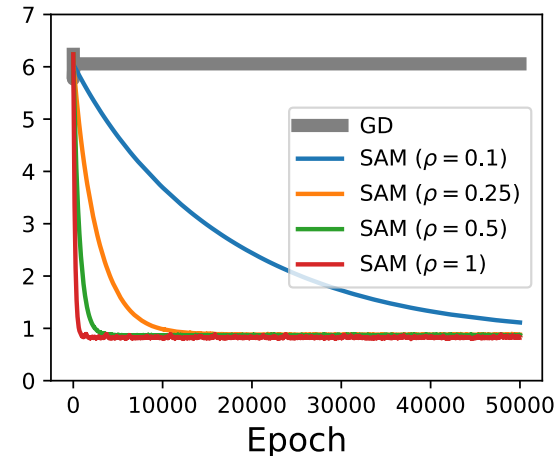
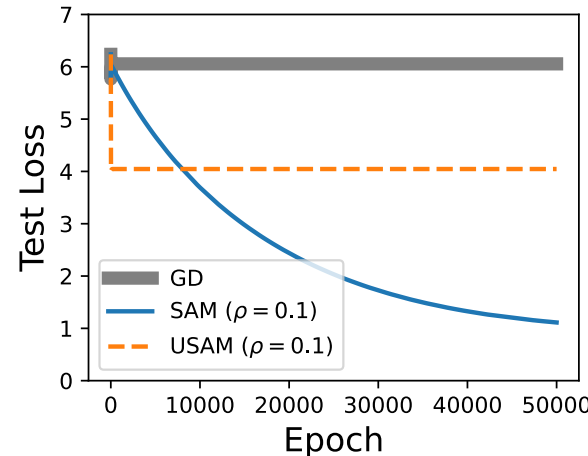
# Motivating Experiments

- **Setup:** over-parameterized matrix sensing problem [Li et al., 2018]

Same initialization (far from minimum);  
 Fix  $\eta$  (for which GD works) & adjust  $\rho$   
**SAM is much more stable than USAM!**



Same initialization (near minimum)  
**USAM gets stuck -- just like GD**  
**SAM w/ diff  $\rho$  finds same minimum**  
 (believed to be good for generalization  
 [Bartlett et al., 2022; Wen et al., 2023])

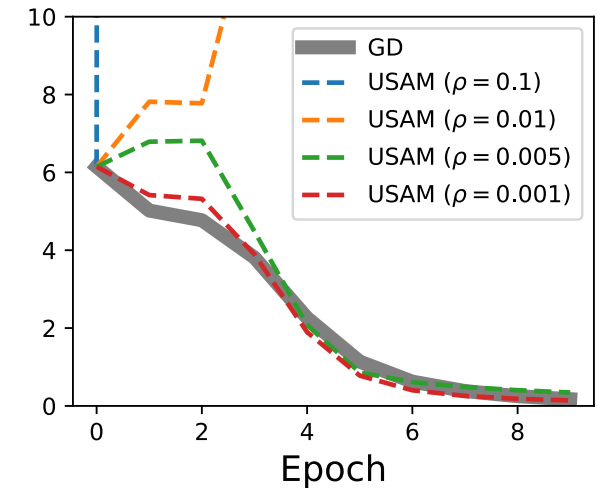
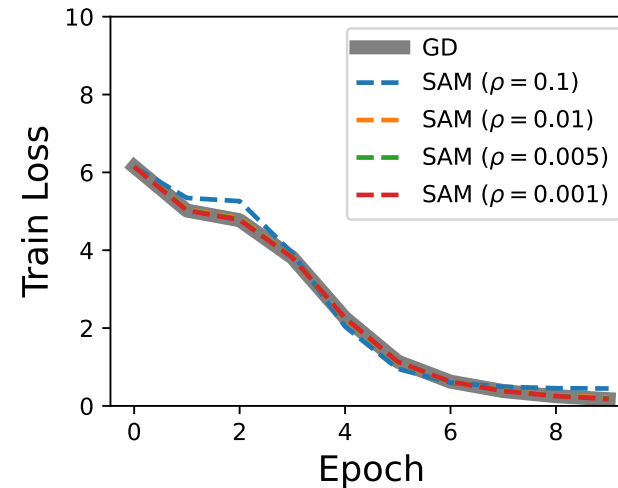


# Motivating Experiments

- **Setup:** over-parameterized matrix sensing problem [Li et al., 2018]

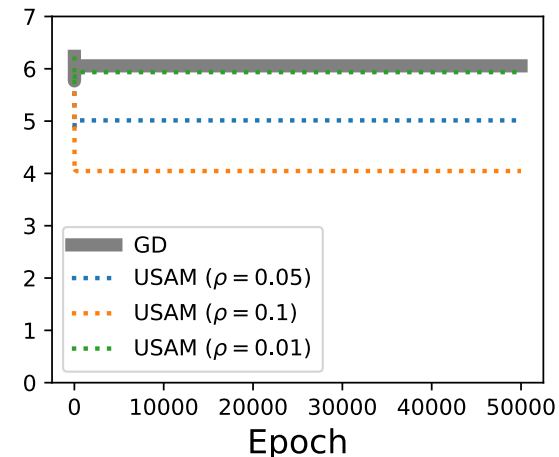
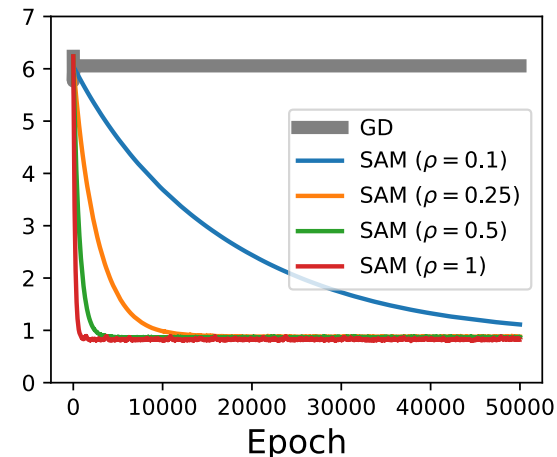
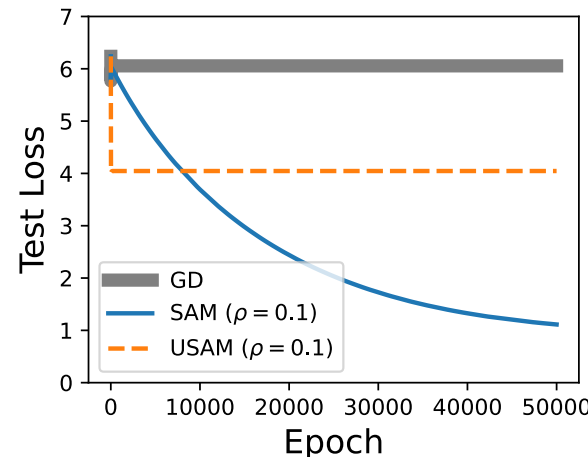
- **Normalization helps with stability**

Same initialization (far from minimum);  
 Fix  $\eta$  (for which GD works) & adjust  $\rho$   
**SAM is much more stable than USAM!**



- **Normalization permits moving along minima**

Same initialization (near minimum)  
**USAM gets stuck -- just like GD**  
**SAM w/ diff  $\rho$  finds same minimum**  
 (believed to be good for generalization  
 [Bartlett et al., 2022; Wen et al., 2023])





# Theoretical Results (Informal)

(more empirical results are contained in our paper)

- **Normalization helps with stability: a “large”  $\rho$  causes USAM to diverge**
- **Theorem 1.** For *strongly convex & smooth*  $L$ , **SAM converges** w/ configuration  $(\eta, \rho)$  as long as  $\eta < 2/\beta$  (i.e., **GD converges**), but **USAM diverges** a.s. if  $\eta > 2/(\beta + \rho\beta^2)$ !
- **Theorem 2.** For *scalar factorization*  $L(x, y) = (xy^2)/2$  with  $\eta = o(1)$ , **SAM finds an  $\mathcal{O}(\rho)$ -neighborhood** of  $(0,0)$  when  $\rho = \mathcal{O}(1)$ , but **USAM diverges** once  $\rho \approx \eta = o(1)$ !
- **Normalization permits moving along minima**



# Theoretical Results (Informal)

(more empirical results are contained in our paper)

- **Normalization helps with stability: a “large”  $\rho$  causes USAM to diverge**
- **Theorem 1.** For *strongly convex & smooth*  $L$ , **SAM converges** w/ configuration  $(\eta, \rho)$  as long as  $\eta < 2/\beta$  (i.e., **GD converges**), but **USAM diverges** a.s. if  $\eta > 2/(\beta + \rho\beta^2)$ !
- **Theorem 2.** For *scalar factorization*  $L(x, y) = (xy^2)/2$  with  $\eta = o(1)$ , **SAM finds an  $\mathcal{O}(\rho)$ -neighborhood** of  $(0,0)$  when  $\rho = \mathcal{O}(1)$ , but **USAM diverges** once  $\rho \approx \eta = o(1)$ !
- **Normalization permits moving along minima: a “small”  $\rho$  makes USAM stuck**
- **Theorems 3-5.** For *single-neuron linear net*  $L(x, y) = \ell(xy)$  [Ahn et al., 2023a] initied @  $(x_0, y_0)$ , **GD** finds  $(0, y_\infty)$  w/  $y_\infty^2 \approx \min(y_0^2 - x_0^2, 2/\eta)$  [Ahn et al., 2023a] (**so  $y_\infty^2 \gg 0$** ), **USAM** finds  $(0, y_\infty)$  w/  $(1 + \rho y_\infty^2)y_\infty^2 \approx 2/\eta$  (**again  $y_\infty^2 \gg 0$** ), **SAM** finds  $y_\infty^2 = o(1)$ !
- **Theorem 6.** For *PL & smooth*  $L$ , the distance **USAM travels along manifold is bounded!**
- **Main Takeaway: USAM is sensitive to  $(\eta, \rho)$ -choice & behaves differently from SAM!**



清华大学 交叉信息研究院

Institute for Interdisciplinary Information Sciences, Tsinghua University



# Thanks for Listening!

Email: [yan-dai20@mails.tsinghua.edu.cn](mailto:yan-dai20@mails.tsinghua.edu.cn); [kjahn@mit.edu](mailto:kjahn@mit.edu); [suvrit@mit.edu](mailto:suvrit@mit.edu)

## References

- Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In International Conference on Learning Representations, 2021.
- Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In International Conference on Machine Learning, pages 639–668. PMLR, 2022.
- Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In Conference On Learning Theory, pages 2–47. PMLR, 2018.
- Peter L Bartlett, Philip M Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. arXiv preprint arXiv:2210.01513, 2022.
- Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? In International Conference on Learning Representations, 2023.
- Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the “edge of stability”. NeurIPS 2023 (arXiv:2212.07469), 2023a.