# Friendly Sharpness-Aware Minimization

## Supplementary Material

## A1. Proof of Theorem 1

The proof is based on [4, 9]. We first introduce the Vector Bernstein inequality presented in [4].

**Theorem A1** (Vector Bernstein)**.** *Let $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$ be independent, zero-mean vector-valued random variables with common dimension d. We have the waeker Vector Berstein version as follows*

$$P\Big(\|\sum_{i=1}^{n} \boldsymbol{x_i}\| \geq \epsilon\Big) \leq \exp(-\frac{\epsilon^2}{8V} + \frac{1}{4}), \tag{1}$$

*where $V = \sum_{i=1}^{n} \mathbb{E}\Big[\|\boldsymbol{x_i}\|^2\Big]$ is the sum of the variances of the centered vectors $\boldsymbol{x_i}$.*

**Corollary A1.** *Let $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$ be independent, zero-mean vector-valued random variables with common dimension d. Assume $\mathbb{E}\Big[\|\boldsymbol{x_i}\|^2\Big] \leq M$. Let $w \in \Delta_n$ in the simplex. Then we have*

$$P\Big(\|\sum_{i=1}^{n} w_i \boldsymbol{x_i}\| \geq \epsilon\Big) \leq \exp(-\frac{\epsilon^2}{8M\|w\|_2^2} + \frac{1}{4}). \tag{2}$$

*Proof.* Simply apply Theorem A1 with $\hat{\boldsymbol{x}}_{\boldsymbol{i}} = w_i \boldsymbol{x_i}$. □

Now we can apply the above concentration results to prove Theorem 1.

*Proof.* First we separately bound the bias and variance, then use Corollary A1. The bias is:

$$\begin{aligned}
\left\|\sum_{t=1}^{T} w_t f\left(x_t\right) - f\left(x_T\right)\right\| &= \left\|\sum_{t=1}^{T} w_t \left(f\left(x_t\right) - f\left(x_T\right)\right)\right\| \qquad (3)\\
&\leq \sum_{t=1}^{T} w_t \left\|f\left(x_t\right) - f\left(x_T\right)\right\| \qquad (4)\\
&\leq \beta \sum_{t=1}^{T} w_t \left\|x_t - x_T\right\| \qquad (5)\\
&\leq \beta \sum_{t=1}^{T} w_t \sum_{s=t+1}^{T} \left\|x_s - x_{s-1}\right\| \qquad (6)\\
&\leq \gamma G \beta \sum_{t=1}^{T} w_t (T-t) \qquad (7)\\
&= \gamma G \beta \cdot \frac{1}{\sum_{t=1}^{T} \lambda^{T-t}} \cdot \sum_{t=1}^{T} \lambda^{T-t}(T-t). \qquad (8)
\end{aligned}$$

□

Note that by a well-known identity,

$$\sum_{t=1}^{T} \lambda^{T-t}(T-t) = \sum_{s=0}^{T-1} s\lambda^s \leq \sum_{s=0}^{\infty} s\lambda^s = \frac{\lambda}{(1-\lambda)^2}. \tag{9}$$

Hence, the bias is bounded by

$$\begin{align}
\gamma G\beta \cdot \frac{1}{\sum_{t=1}^{T} \lambda^{T-t}} \cdot \frac{\lambda}{(1-\lambda)^2} &= \gamma G\beta \cdot \frac{1-\lambda}{1-\lambda^T} \cdot \frac{\lambda}{(1-\lambda)^2} \tag{10}\\
&= \gamma G\beta \cdot \frac{1}{1-\lambda^T} \cdot \frac{\lambda}{1-\lambda} \tag{11}\\
&\leq G\beta \cdot \frac{\gamma}{(1-\lambda)\left(1-\lambda^T\right)}. \tag{12}
\end{align}$$

Applying Corollary A1 to $x_t = x_t - f\left(x_t\right)$, we have that

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T} w_t \left(y_t - f\left(x_t\right)\right)\right\| > k\right) \leq \exp(-\frac{\epsilon^2}{8M^2\|w\|_2^2} + \frac{1}{4}). \tag{13}$$

Now note that

$$\begin{align}
\|w\|_2^2 = \sum_{t=1}^{T} w_t^2 &= \frac{1}{\left(\sum_{t=1}^{T} \lambda^{T-t}\right)^2} \sum_{t=1}^{T} \left(\lambda^2\right)^{T-t} \tag{14}\\
&= \frac{(1-\lambda)^2}{\left(1-\lambda^T\right)^2} \sum_{t=1}^{T} \left(\lambda^2\right)^{T-t} \tag{15}\\
&= \frac{(1-\lambda)^2}{\left(1-\lambda^T\right)^2} \cdot \frac{1-\lambda^{2T}}{1-\lambda^2} \tag{16}\\
&= \frac{1-\lambda^{2T}}{\left(1-\lambda^T\right)^2} \cdot \frac{(1-\lambda)^2}{1-\lambda^2} \tag{17}\\
&= \frac{1+\lambda^T}{1-\lambda^T} \cdot \frac{1-\lambda}{1+\lambda} \tag{18}\\
&\leq \frac{2(1-\lambda)}{1-\lambda^T}. \tag{19}
\end{align}$$

Setting the right hand side of the high probability bound to $\delta$, we have concentration w.p. $1-\delta$ for $k$ satisfying

$$\delta \geq \exp\left(-\frac{\epsilon^2}{8M\|w\|_2^2} + \frac{1}{4}\right). \tag{20}$$

Rearranging, we find

$$\begin{align}
\log\left(e^{-\frac{1}{4}}/\delta\right) &\leq \frac{\epsilon^2}{8M\|w\|_2^2} \tag{21}\\
\Leftrightarrow \epsilon &\geq 2\sqrt{2}\sqrt{M}\|w\|_2 \cdot \sqrt{\log\left(e^{-\frac{1}{4}}/\delta\right)}. \tag{22}
\end{align}$$

Combining this with the triangle inequality,

$$\begin{align}
\left\|\sum_{t=1}^{T} w_t y_t - f\left(x_T\right)\right\| &\leq \left\|\sum_{t=1}^{T} w_t y_t - \sum_{t=1}^{T} w_t f\left(x_t\right)\right\| + \left\|\sum_{t=1}^{T} w_t \left(f\left(x_t\right) - f\left(x_T\right)\right)\right\| \tag{23}\\
&\leq 2\sqrt{2}\sqrt{M}\|w\|_2 \sqrt{\log(e^{-\frac{1}{4}}/\delta)} + G\beta \cdot \frac{\gamma}{(1-\lambda)\left(1-\lambda^T\right)} \tag{24}\\
&\leq 2\sqrt{2}\sqrt{\frac{2(1-\lambda)}{1-\lambda^T}}\sqrt{M}\sqrt{\log(e^{-\frac{1}{4}}/\delta)} + G\beta \cdot \frac{\gamma}{(1-\lambda)\left(1-\lambda^T\right)} \tag{25}\\
&\leq 4\sqrt{M}\frac{\sqrt{1-\lambda}}{\sqrt{1-\lambda^T}}\sqrt{\log(e^{-\frac{1}{4}}/\delta)} + G\beta \cdot \frac{\gamma}{(1-\lambda)\left(1-\lambda^T\right)}, \tag{26}
\end{align}$$

with probability $1-\delta$. Since $1/\sqrt{1-\lambda^T} \leq 1/\left(1-\lambda^T\right)$, this can further be bounded by

$$\left(4\sqrt{M} \cdot \sqrt{1-\lambda} \cdot \sqrt{\log\left(e^{-\frac{1}{4}}/\delta\right)} + G\beta \cdot \frac{\gamma}{1-\lambda}\right) \cdot \frac{1}{1-\lambda^T}. \tag{27}$$

Write $\alpha = 1-\lambda$. The inner part of the bound is optimized when

$$\begin{aligned} 4\sqrt{M} \cdot \sqrt{\alpha} \cdot \sqrt{\log\left(e^{-\frac{1}{4}}/\delta\right)} &= G\beta \cdot \frac{\gamma}{\alpha} && (28)\\ \Leftrightarrow \alpha^{3/2} &= \frac{G\beta\gamma}{4\sqrt{M}\sqrt{\log(e^{-\frac{1}{4}}/\delta)}} && (29)\\ \Leftrightarrow \alpha &= \frac{G^{2/3}\beta^{2/3}\gamma^{2/3}}{4^{2/3} \cdot M^{1/3} \cdot \left(\log\left(e^{-\frac{1}{4}}/\delta\right)\right)^{1/3}}, && (30) \end{aligned}$$

for which the overall inner bound is

$$2G\beta \cdot \frac{\gamma}{\alpha} = 2^{7/3} \cdot M^{1/3} \cdot \log\left(e^{-\frac{1}{4}}/\delta\right)^{1/3} \cdot (G\beta\gamma)^{1/3}. \tag{31}$$

If $T$ is sufficiently large, the $1/\left(1-\lambda^T\right)$ term will be less than 2 . In particular,

$$T > \frac{2}{\log(1+\alpha)} \Longrightarrow \frac{1}{1-(1-\alpha)^T} < 2.$$

Since $\log(1+\alpha) > \alpha/2$ for $\alpha < 1$, it suffices to have $T > 4/\alpha$.

## A2. Proof of Lemma 1

*Proof.*

$$\begin{aligned} \mathbb{E}\left\langle \widetilde{\operatorname{Proj}}^{\top}_{\nabla L(\boldsymbol{w}_t)} \nabla L_{\mathcal{B}}(\boldsymbol{w}_t), \nabla L(\boldsymbol{w}_t)\right\rangle &= \mathbb{E}\left\langle \nabla L_{\mathcal{B}}(\boldsymbol{w}_t) - \sigma \boldsymbol{m}_t, \nabla L(\boldsymbol{w}_t)\right\rangle && (32)\\ &= \|\nabla L(\boldsymbol{w}_t)\|^2(1-\sigma). && (33) \end{aligned}$$

Eqn. (33) uses the assumption that $\boldsymbol{m}_t$ is an unbiased estimator to $\nabla L(\boldsymbol{w}_t)$. Then with $\sigma = 1$, we complete the proof. □

## A3. Proof of Theorem 2

*Proof.* Denote $\boldsymbol{w}_{t+1/2} = \boldsymbol{w}_t + \rho \frac{\nabla L_{\mathcal{B}}(\boldsymbol{w}_t) - \boldsymbol{m}_t}{\|\nabla L_{\mathcal{B}}(\boldsymbol{w}_t) - \boldsymbol{m}_t\|}$. From Assumption 1, it follows that

$$\begin{aligned} L(\boldsymbol{w}_{t+1}) \leq & L(\boldsymbol{w}_t) + \nabla L(\boldsymbol{w}_t)^{\top}(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{\beta}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 && (34)\\ = & L(\boldsymbol{w}_t) - \gamma_t \nabla L(\boldsymbol{w}_t)^{\top} \nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}) + \frac{\gamma_t^2 \beta}{2}\|\nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2})\|^2 && (35)\\ = & L(\boldsymbol{w}_t) - \gamma_t \nabla L(\boldsymbol{w}_t)^{\top} \nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2})\\ & + \frac{\gamma_t^2 \beta}{2}\left(\|\nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}) - \nabla L(\boldsymbol{w}_t)\|^2 - \|\nabla L(\boldsymbol{w}_t)\|^2 + 2\nabla L(\boldsymbol{w}_t)^{\top} \nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2})\right) && (36)\\ = & L(\boldsymbol{w}_t) - \frac{\gamma_t^2 \beta}{2}\|\nabla L(\boldsymbol{w}_t)\|^2 + \frac{\gamma_t^2 \beta}{2}\|\nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}) - \nabla L(\boldsymbol{w}_t)\|^2 - (1-\beta\gamma_t)\gamma_t \nabla L(\boldsymbol{w}_t)^{\top} \nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}) && (37)\\ \leq & L(\boldsymbol{w}_t) - \frac{\gamma_t^2 \beta}{2}\|\nabla L(\boldsymbol{w}_t)\|^2 + \gamma_t^2 \beta\|\nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}) - \nabla L(\boldsymbol{w}_{t+1/2})\|^2\\ & + \gamma_t^2 \beta\|\nabla L(\boldsymbol{w}_{t+1/2}) - \nabla L(\boldsymbol{w}_t)\|^2 - (1-\beta\gamma_t)\gamma_t \nabla L(\boldsymbol{w}_t)^{\top} \nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}). && (38) \end{aligned}$$

The last step using the fact that $\|a-b\|^2 \leq 2\|a-c\|^2 + 2\|c-b\|^2$. Then taking the expectation on both sides gives:

$$\mathbb{E}[L(\boldsymbol{w}_{t+1})] \leq \mathbb{E}[L(\boldsymbol{w}_t)] - \frac{\gamma_t^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \gamma_t^2\beta M + \rho_t^2\gamma_t^2\beta^3 - (1-\beta\gamma_t)\gamma_t\mathbb{E}[\nabla L(\boldsymbol{w}_t)^\top \nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2})]. \tag{39}$$

For the last term, we have:

$$\begin{aligned}\mathbb{E}[\nabla L(\boldsymbol{w}_t)^\top \nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2})] =&\mathbb{E}\left[\nabla L(\boldsymbol{w}_t)^\top\left(\nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}) - \nabla L_{\mathcal{B}}(\boldsymbol{w}_t) + \nabla L_{\mathcal{B}}(\boldsymbol{w}_t)\right)\right] \\ =&\mathbb{E}\left[\|\nabla L(\boldsymbol{w}_t)\|^2\right] + \mathcal{C}.\end{aligned} \tag{40}$$

where $\mathcal{C} = \mathbb{E}\left[\nabla L(\boldsymbol{w}_t)^\top\left(\nabla L_{\mathcal{B}}(\boldsymbol{w}_{t+1/2}) - \nabla L_{\mathcal{B}}(\boldsymbol{w}_t)\right)\right]$. Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathcal{C} \leq&\mathbb{E}\left[\frac{1}{2}\|\nabla L(\boldsymbol{w}_t)\|^2 + \frac{1}{2}\|\nabla L(\boldsymbol{w}_{t+1/2}) - \nabla L(\boldsymbol{w}_t)\|^2\right] \\ \leq&\frac{1}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \frac{\rho_t^2\beta^2}{2}.\end{aligned} \tag{41}$$

Plugging Eqn. (40) and (41) into Eqn. (39), we obtain:

$$\begin{aligned}\mathbb{E}[L(\boldsymbol{w}_{t+1})] \leq&\mathbb{E}[L(\boldsymbol{w}_t)] - \frac{\gamma_t^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \gamma_t^2\beta M + \rho_t^2\gamma_t^2\beta^3 - (1-\beta\gamma_t)\gamma_t\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 && (42)\\ &+ (1-\beta\gamma_t)\gamma_t\left(\frac{1}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \frac{\rho_t^2\beta^2}{2}\right) && (43)\\ \leq&\mathbb{E}[L(\boldsymbol{w}_t)] - \frac{\gamma_t}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \gamma_t^2\beta M + \frac{1}{2}\gamma_t\rho_t^2\beta^2(1+\beta\gamma_t). && (44)\end{aligned}$$

Taking summation over $T$ iterations, we have:

$$\frac{\gamma_0}{2\sqrt{T}}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \leq \mathbb{E}[L(\boldsymbol{w}_0)] - \mathbb{E}[L(\boldsymbol{w}_T)] + (\beta M + \frac{1}{2}\rho_0^2\beta^3)\gamma_0^2\sum_{t=1}^{T}\frac{1}{T} + \frac{1}{2}\rho_0^2\beta^2\gamma_0\sum_{t=1}^{T}\frac{1}{t}. \tag{45}$$

This gives

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \leq& \frac{2\left(\mathbb{E}[L(\boldsymbol{w}_0)] - \mathbb{E}[L(\boldsymbol{w}_T)]\right)}{\gamma_0\sqrt{T}} + \frac{2\beta M\gamma_0 + \rho_0^2\beta^3\gamma_0}{\sqrt{T}} + \frac{\rho_0^2\beta^2\log T}{\sqrt{T}} && (46)\\ \leq& \frac{2\left(\mathbb{E}[L(\boldsymbol{w}_0) - L(\boldsymbol{w}^*)]\right)}{\gamma_0\sqrt{T}} + \frac{2\beta M\gamma_0 + \rho_0^2\beta^3\gamma_0}{\sqrt{T}} + \frac{\rho_0^2\beta^2\log T}{\sqrt{T}} && (47)\\ =& \frac{2\Delta}{\gamma_0\sqrt{T}} + \frac{\Theta}{\sqrt{T}} + \frac{\Pi\log T}{\sqrt{T}}, && (48)\end{aligned}$$

where $\boldsymbol{w}^*$ is the optimal solution, $\Delta = \mathbb{E}[L(\boldsymbol{w}_0) - L(\boldsymbol{w}^*)]$, $\Theta = 2\beta M\gamma_0 + \rho_0^2\beta^3\gamma_0$, and $\Pi = \rho_0^2\beta^2$. □

## A4. Extension to SAM's variants

Since we only modify the perturbation of SAM, our modification can be straightforwardly extended into the SAM variants, such as ASAM [5] and FiserSAM [3]. For SAM variants, their min-max objectives can be written into a unified formulation:

$$\min_{\boldsymbol{w}} \max_{\|T_w^{-1}\boldsymbol{\epsilon}\| \leq \rho} L(\boldsymbol{w} + \boldsymbol{\epsilon}), \tag{49}$$

where $T_w$ is a normalization operator, e.g., $T_w = \|\boldsymbol{w}\|$ for ASAM. The inner maximization problem in Eq. (49) can then be solved via first-order approximation as follows:

$$\boldsymbol{\epsilon}_s = \rho\frac{T_w^2\nabla L_{\mathcal{B}}(\boldsymbol{w})}{\|T_w\nabla L_{\mathcal{B}}(\boldsymbol{w})\|_2}. \tag{50}$$

To incorporate our improvement, we can modify Eqn. (50) as follows:

$$\boldsymbol{\epsilon}_s = \rho\frac{T_w^2\left(\nabla L_{\mathcal{B}}(\boldsymbol{w}_t) - \sigma\boldsymbol{m}_t\right)}{\|T_w\left(\nabla L_{\mathcal{B}}(\boldsymbol{w}_t) - \sigma\boldsymbol{m}_t\right)\|_2}. \tag{51}$$

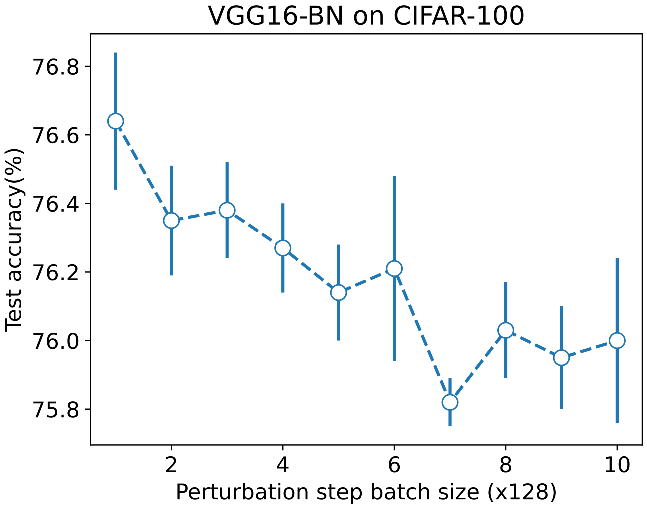

Figure A1. Results on enlarging the batch size of SAM's adversarial perturbation.

# A5. Investigation Details

## A5.1. Experimental Settings

We follow the training setting of our main experiments.

**Training from scratch.** We train the models for 200 epochs and set the initial learning rate as 0.05 with a cosine learning rate schedule. The momentum and weight decay are set to 0.9 and 0.0005 for SGD, respectively. SAM adopt the same setting except that the weight decay is set to 0.001 following [6, 8]. We apply standard random horizontal flipping, cropping, normalization, and cutout augmentation [1]. For SAM and its modified variants, we set the perturbation radius $\rho$ as 0.1 and 0.2 for CIFAR-10 and CIFAR-100 [6, 8].

**Transfer learning.** We use a Deit-small model [10] pre-trained on ImageNet. We use AdamW [7] as base optimizer and train the model for 10 epochs with batch size 128, weight decay $10^{-5}$ and initial learning rate of $10^{-4}$. We adopt $\rho = 0.075$ for SAM and its modified variants. We apply image resizing (to $224 \times 224$) and normalization for data preprocessing without extra augmentations.

**SAM's modified variants.** 1) SAM-full: we use full gradient $\nabla L(\boldsymbol{w})$ over the entire training dataset to calculate SAM's perturbation, i.e., $\boldsymbol{\epsilon}_s = \rho \frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}$; 2) SAM-db: we use an extra random batch data $\mathcal{B}'$ to calculate SAM's perturbation, i.e., $\boldsymbol{\epsilon}_s = \rho \frac{\nabla L_{\mathcal{B}'}(\boldsymbol{w})}{\|\nabla L_{\mathcal{B}'}(\boldsymbol{w})\|}$ 3) SAM-noise: we use residual projection direction w.r.t. the full gradient to calculate the perturbation, i.e., $\boldsymbol{\epsilon}_s = \rho \frac{\mathrm{Proj}^{\top}_{\nabla L(\boldsymbol{w})} \nabla L_{\mathcal{B}}(\boldsymbol{w})}{\|\mathrm{Proj}^{\top}_{\nabla L(\boldsymbol{w})} \nabla L_{\mathcal{B}}(\boldsymbol{w})\|}$. We align the gradient of the model parameters as a vector to perform gradient projection.

## A5.2. More Experiments on Effects of Full Gradient Component

To further substantiate the detrimental effects of strengthening the full gradient components, we conducted additional experiments on the CIFAR-100 dataset using the VGG16-BN architecture, as illustrated in Figure Fig. A1.

# A6. Training curves

In Fig. A2, we compare the training curves of SAM and F-SAM. We observe that F-SAM achieves a faster convergence than SAM especially on the initial stage. This is because at the initial stage, the proportion of the full gradient component in the minibatch gradient is more significant,and hence removing this component to facilitate convergence in F-SAM has a more pronounced effect. Moreover, as the perturbation radius grows (2x in this case), the magnitude of the full gradient component in $\boldsymbol{\epsilon}_s$ also grows. This can significantly hinder the convergence of SAM and degrade its performance. In contrast, F-SAM is able to mitigate this undesired effects and maintain a good performance.

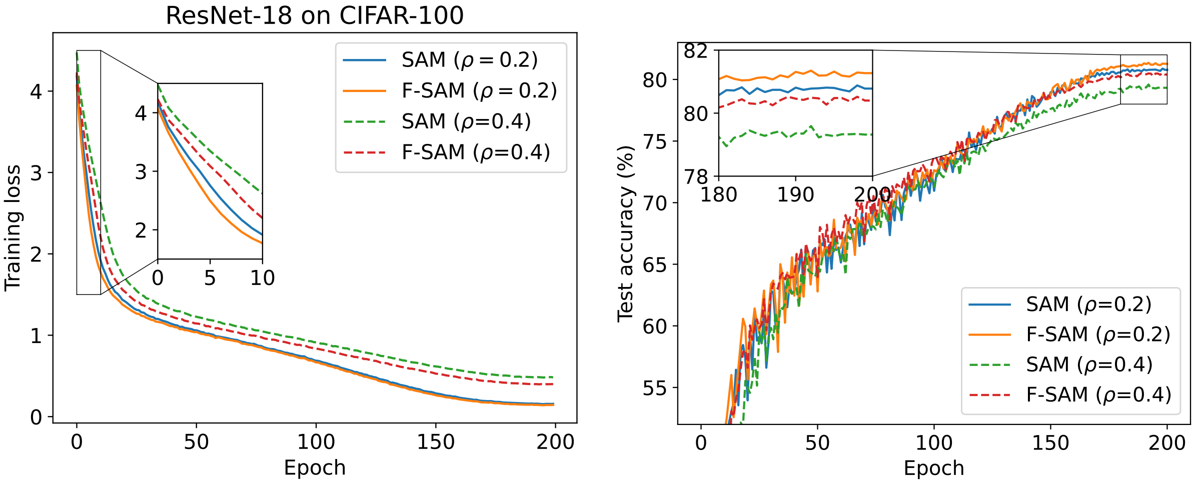

Figure A2. Training curve comparison on CIFAR-100 with ResNet-18.

## A7. Hessian Spectrum

In Fig. A3, we compare the Hessian eigenvalues of ResNet-18 trained with SAM and F-SAM. We focus on the largest eigenvalue $\lambda_1$ and the ratio of the largest to the fifth largest eigenvalue $\lambda_1/\lambda_5$. We approximate the calculation for Hessian spectrum using the Lanczos algorithm [2]. We observe that F-SAM achieves a smaller largest eigenvalue and smaller eigenvalue ratio compared with SAM. This confirms that F-SAM converges to a flatter solution and achieves better generalization by removing the undesirable full gradient component in adversarial perturbation.

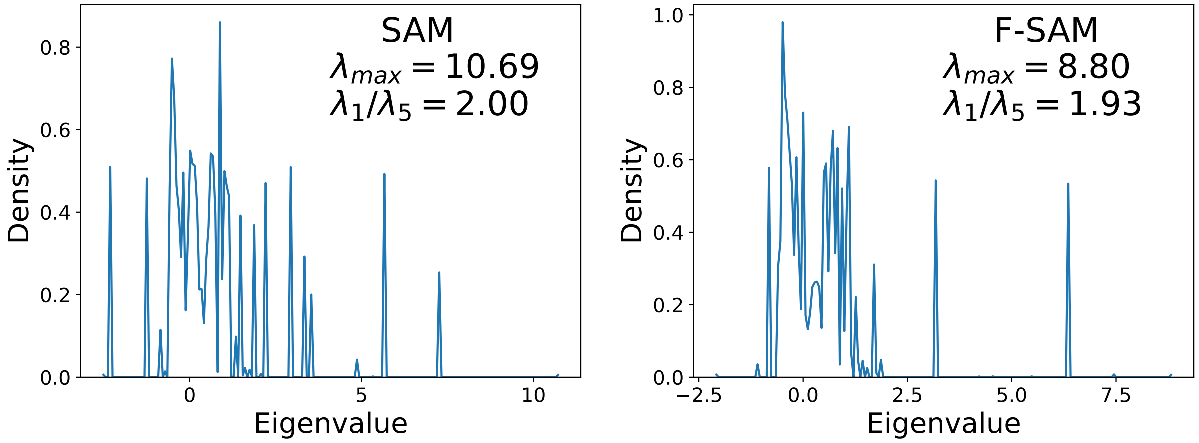

Figure A3. Hessian spectrum comparison on CIFAR-10 with ResNet-18.

## References

[1] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 5

[2] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, 2019. 6

[3] Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning (ICML)*, 2022. 4

[4] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR, 2017. 1

[5] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2021. 4

[6] Bingcong Li and Georgios B Giannakis. Enhancing sharpness-aware optimization through variance suppression. *arXiv preprint arXiv:2309.15639*, 2023. 5

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[8] Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: A sparsified perturbation approach. *arXiv preprint arXiv:2210.05177*, 2022. 5

[9] Matthew Staib, Sashank Reddi, Satyen Kale, Sanjiv Kumar, and Suvrit Sra. Escaping saddle points with adaptive gradient methods. In *International Conference on Machine Learning (ICML)*, 2019. 1

[10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021. 5