## APPENDIX FOR

**Sharpness-Aware Minimization with Adaptive Regularization for Training Deep Neural Networks**

### A. Proof of Theorem 1

A significant conclusion (3) could be deduced from Assumption 2. If $\nabla f(\mathbf{x})$ has a Lipschitz gradient with constant $L > 0$, we have

$$
\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= -\eta_k \big\langle \nabla f(\mathbf{x}_k), (1-\lambda_k)\mathbf{g}(\mathbf{x}_k) + \lambda_k \mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k}) \big\rangle + \frac{L}{2}\eta_k^2 \big\|(1-\lambda_k)\mathbf{g}(\mathbf{x}_k) + \lambda_k \mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k})\big\|^2 \\
&= \underbrace{-\eta_k \big\langle \nabla f(\mathbf{x}_k), (1-\lambda_k)\mathbf{g}(\mathbf{x}_k) \big\rangle}_{:=\spadesuit} \underbrace{-\eta_k \big\langle \nabla f(\mathbf{x}_k), \lambda_k \mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k}) \big\rangle}_{:=\heartsuit} + \underbrace{\frac{L}{2}\eta_k^2 \big\|(1-\lambda_k)\mathbf{g}(\mathbf{x}_k) + \lambda_k \mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k})\big\|^2}_{:=\clubsuit}.
\end{aligned} \tag{3}
$$

We calculate each of the above three parts respectively

$$
\begin{aligned}
\spadesuit =& -\eta_k \big\langle \nabla f(\mathbf{x}_k), (1-\lambda_k)\mathbf{g}(\mathbf{x}_k) \big\rangle \\
=& -\eta_k (1-\lambda_k) \left[\big\langle \nabla f(\mathbf{x}_k), \mathbf{g}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \big\rangle\right] - \eta_k (1-\lambda_k) \big\|\nabla f(\mathbf{x}_k)\big\|^2, && (4) \\
\heartsuit =& -\eta_k \big\langle \nabla f(\mathbf{x}_k), \lambda_k \mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k}) \big\rangle \\
=& -\eta_k \lambda_k \big\langle \nabla f(\mathbf{x}_k), \mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k}) - \mathbf{g}(\mathbf{x}_k) \big\rangle - \eta_k \lambda_k \big\langle \nabla f(\mathbf{x}_k), \mathbf{g}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \big\rangle - \eta_k \lambda_k \big\|\nabla f(\mathbf{x}_k)\big\|^2 \\
\overset{(a)}{\leq}& \eta_k \lambda_k L \rho \|\nabla f(\mathbf{x}_k)\| - \eta_k \lambda_k \|\nabla f(\mathbf{x}_k)\|^2 - \eta_k \lambda_k \big\langle \nabla f(\mathbf{x}_k), \mathbf{g}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \big\rangle, && (5) \\
\clubsuit =& \frac{L}{2}\eta_k^2 \|(1-\lambda_k)\mathbf{g}(\mathbf{x}_k) + \lambda_k \mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2 \\
\overset{(b)}{\leq}& L\eta_k^2 \left[(1-\lambda_k)^2 \|\mathbf{g}(\mathbf{x}_k)\|^2 + \lambda_k^2 \|\mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2\right], && (6)
\end{aligned}
$$

where the inequality (a) uses Lipschitz continuity $\|\mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k}) - \mathbf{g}(\mathbf{x}_k)\| \leq L\|\boldsymbol{\epsilon}_{\mathbf{x}_k}\|$ and the Cauchy's inequality $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$. The inequality (b) depends on the inequality $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$. Taking expectation on (4), (5) and (6), we can further get

$$
\begin{aligned}
\mathbb{E}[\spadesuit] \overset{(c)}{=}& -\eta_k (1-\lambda_k) \mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big], && (7) \\
\mathbb{E}[\heartsuit] \leq& \eta_k \lambda_k L \rho \mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|\big] - \eta_k \lambda_k \mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big] \\
\overset{(d)}{\leq}& \frac{L}{2}\rho^2 + \frac{L}{2}\eta_k^2 \lambda_k^2 \big[\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|]\big]^2 - \eta_k \lambda_k \mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big] \\
\overset{(e)}{\leq}& \frac{L}{2}\rho^2 + \frac{L}{2}\eta_k^2 \lambda_k^2 \mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big] - \eta_k \lambda_k \mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big], && (8) \\
\mathbb{E}[\clubsuit] \leq& L\eta_k^2 \left[(1-\lambda_k)^2 \mathbb{E}[\|\mathbf{g}(\mathbf{x}_k)\|^2] + \lambda_k^2 \mathbb{E}[\|\mathbf{g}(\mathbf{x}_k + \boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2]\right] \\
\overset{(f)}{\leq}& L\eta_k^2 \Big[(1-\lambda_k)^2 \big(\sigma^2 + \mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\big) + \lambda_k^2 \big(2L^2\rho^2 + 2\sigma^2 + 2\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\big)\Big], && (9)
\end{aligned}
$$

where the equality (c) utilizes that $\mathbf{g}(\mathbf{x}_k)$ is the unbiased estimation of $f(\mathbf{x}_k)$. The inequality (d) leverages the basic inequality $\frac{a+b}{2} \geq \sqrt{ab}$. The inequality (e) comes from the nonnegative variance property $\mathrm{Var}(\mathbf{X}) = \mathbb{E}[\mathbf{X}^2] - \big[\mathbb{E}\,[\mathbf{X}]\big]^2 \geq 0$. A combination of the following results (11) and (13) leads to (f). After a simple mathematical derivation of $\|\mathbf{g}(\mathbf{x}_k)\|^2$, the following results could be derived:

$$
\begin{aligned}
\|\mathbf{g}(\mathbf{x}_k)\|^2 &= \|\mathbf{g}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)\|^2 \\
&= \|\mathbf{g}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2 + \|\nabla f(\mathbf{x}_k)\|^2 + 2\big\langle \mathbf{g}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_k) \big\rangle.
\end{aligned} \tag{10}
$$

Taking expectation on (10), then we derive

$$
\mathbb{E}[\|\mathbf{g}(\mathbf{x}_k)\|^2] \leq \sigma^2 + \mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2. \tag{11}
$$

After a plain derivation of $\|\mathbf{g}(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2$, we have

$$\begin{aligned}\|\mathbf{g}(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2 &= \|\mathbf{g}(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})-\mathbf{g}(\mathbf{x}_k)+\mathbf{g}(\mathbf{x}_k)\|^2\\ &\le 2\|\mathbf{g}(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})-\mathbf{g}(\mathbf{x}_k)\|^2+2\|\mathbf{g}(\mathbf{x}_k)\|^2\\ &\le 2L^2\rho^2+2\|\mathbf{g}(\mathbf{x}_k)\|^2. \end{aligned} \tag{12}$$

Taking expectation on (12) gives us

$$\mathbb{E}[\|\mathbf{g}(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2]\le 2L^2\rho^2+2\mathbb{E}[\|\mathbf{g}(\mathbf{x}_k)\|^2]. \tag{13}$$

Taking expectation on (3) combining (7), (8), (9), (11) and (13), we can get

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x_{k+1}})-f(\mathbf{x_k})\right] \le&\mathbb{E}[\spadesuit]+\mathbb{E}[\heartsuit]+\mathbb{E}[\clubsuit]\\ \le&-\eta_k(1-\lambda_k)\mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big]+\frac{L}{2}\rho^2+\frac{L}{2}\eta_k^2\lambda_k^2\mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big]-\eta_k\lambda_k\mathbb{E}\big[\|\nabla f(\mathbf{x}_k)\|^2\big]\\ &+L\eta_k^2\Big[(1-\lambda_k)^2\big(\sigma^2+\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\big)+\lambda_k^2\big(2L^2\rho^2+2\sigma^2+2\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\big)\Big]. \end{aligned} \tag{14}$$

Rearranging these terms and dividing the constant $\eta_k$ on both sides of (14) yields

$$\begin{aligned}\underbrace{\left[1-\frac{5}{2}L\eta_k\lambda_k^2-L\eta_k(1-\lambda_k)^2\right]}_{:=M_k}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\le&2L\eta_k\lambda_k^2\sigma^2+\frac{L\rho^2}{2\eta_k}+L\eta_k(1-\lambda_k)^2\sigma^2\\ &+2L^3\eta_k\lambda_k^2\rho^2+\frac{\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}_{k+1})]}{\eta_k}.\end{aligned}$$

It holds that $\lambda_k\in(0,1)$ according to the update rule in Algorithm 1. Thus we have $1-\frac{5}{2}L\eta_k\le M_k=1-\frac{5}{2}L\eta_k\lambda_k^2-L\eta_k(1-\lambda_k)^2\le 1-\frac{5}{7}L\eta_k$, $M_k\ge\nu=1-\frac{5L\eta_0}{2\sqrt{K}}$ with $\eta_k=\frac{\eta_0}{\sqrt{K}}$ and $\rho=\frac{\rho_0}{\sqrt{K}}$. Consequently, we further get

$$\begin{aligned}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\le&\frac{1}{\nu}\Big[2L\eta_k\sigma^2+\frac{L\rho^2}{2\eta_k}+L\eta_k\sigma^2+2L^3\eta_k\rho^2+\frac{\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}_{k+1})]}{\eta_k}\Big]\\ =&\frac{1}{\nu}\Big[\frac{2L\eta_0\sigma^2}{\sqrt{K}}+\frac{L\rho_0^2}{2\eta_0\sqrt{K}}+\frac{L\eta_0\sigma^2}{\sqrt{K}}+\frac{2L^3\eta_0\rho_0^2}{K^{\frac{3}{2}}}+\frac{\mathbb{E}[f(\mathbf{x}_k)-f(\mathbf{x}_{k+1})]}{\eta_0}\sqrt{K}\Big].\end{aligned} \tag{15}$$

Summing (15) from $k=0$ to $K-1$, we have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\le&\frac{1}{\nu}\Big[\frac{2L\eta_0\sigma^2}{\sqrt{K}}+\frac{L\rho_0^2}{2\eta_0\sqrt{K}}+\frac{L\eta_0\sigma^2}{\sqrt{K}}+\frac{2L^3\eta_0\rho_0^2}{K^{\frac{3}{2}}}+\frac{\mathbb{E}[f(\mathbf{x}_0)-f(\mathbf{x}_K)]}{\eta_0\sqrt{K}}\Big]\\ \le&\frac{1}{\nu}\Big[\frac{f(\mathbf{x}_0)-f_{\inf}}{\eta_0\sqrt{K}}+\frac{L\rho_0^2}{2\eta_0\sqrt{K}}+\frac{3L\eta_0\sigma^2}{\sqrt{K}}+\frac{2L^3\eta_0\rho_0^2}{K^{\frac{3}{2}}}\Big].\end{aligned} \tag{16}$$

As to $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2]$, which could be derived by leveraging on (16). After an ordinary derivation of $\|\nabla f(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2$, we are then led to

$$\begin{aligned}\|\nabla f(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2 &= \|\nabla f(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})-\nabla f(\mathbf{x}_k)+\nabla f(\mathbf{x}_k)\|^2\\ &\le 2\|\nabla f(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})-\nabla f(\mathbf{x}_k)\|^2+2\|\nabla f(\mathbf{x}_k)\|^2\\ &\le 2L^2\rho^2+2\|\nabla f(\mathbf{x}_k)\|^2.\end{aligned} \tag{17}$$

Utilizing (17), we have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\mathbf{x}_k+\boldsymbol{\epsilon}_{\mathbf{x}_k})\|^2] &\le 2L^2\rho^2+\frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2]\\ &\le\frac{2}{\nu}\left[\frac{f(\mathbf{x}_0)-f_{\inf}}{\eta_0\sqrt{K}}+\frac{L\rho_0^2}{2\eta_0\sqrt{K}}+\frac{3L\eta_0\sigma^2}{\sqrt{K}}+\frac{2L^3\eta_0\rho_0^2}{K^{\frac{3}{2}}}\right]+\frac{2L^2\rho_0^2}{K}.\end{aligned} \tag{18}$$

TABLE III

HYPERPARAMETERS SETTING USED TO PRODUCE THE RESULTS OF CIFAR10/CIFAR100

| Dataset | Model | Optimizer | Lr | $\rho$ | $\theta$ | $\gamma$ | $\chi$ |
|---|---|---|---|---|---|---|---|
| CIFAR10 | ResNet-34 | SAMAR | 0.3 | 0.10 | - | 1.550 | 1.100 |
| | | SGD | 0.3 | - | - | - | - |
| | | SAM | 0.3 | 0.10 | - | - | - |
| | | VaSSO | 0.3 | 0.10 | 0.9 | - | - |
| | Wide-Resnet-34-10 | SAMAR | 0.1 | 0.10 | - | 1.400 | 1.050 |
| | | SGD | 0.1 | - | - | - | - |
| | | SAM | 0.1 | 0.10 | - | - | - |
| | | VaSSO | 0.1 | 0.10 | 0.9 | - | - |
| CIFAR100 | ResNet-34 | SAMAR | 0.3 | 0.10 | - | 1.400 | 1.075 |
| | | SGD | 0.3 | - | - | - | - |
| | | SAM | 0.3 | 0.10 | - | - | - |
| | | VaSSO | 0.3 | 0.10 | 0.9 | - | - |
| | Wide-Resnet-34-10 | SAMAR | 0.3 | 0.15 | - | 1.500 | 1.000 |
| | | SGD | 0.3 | - | - | - | - |
| | | SAM | 0.3 | 0.15 | - | - | - |
| | | VaSSO | 0.3 | 0.15 | 0.9 | - | - |

### B. *Hyperparameters for experiements*

The hyperparameters throughout the experiment are noteworthy. Since every model is only trained for 100 epochs, we set a slightly larger initial learning rate to ensure that all models converge after training 100 epochs. Referring to the hyperparameters in the relevant literature [5, 6, 11–13, 16], and combining them with the changes in experimental results we observed while finetuning the hyperparameters, the hyperparameters in experiments are shown in Table III. Fololowing [12], VaSSO adopts $\theta = 0.9$. To reduce the effort of finetuning the hyperparameters, we take $\lambda_0 = 1, \delta = 0.01$ for SAMAR, and weight decay is 0.0005 in all experiments.