Supplementary Materials

Anonymous Author(s)
Affiliation
Address
email

1 Missing Proofs

In this section, we will proof the main theorem in our paper. Before the detailed proof, we first recall the following assumptions that are commonly used for characterizing the convergence of non-convex stochastic optimization.

Assumption 1. *(Bounded Gradient.) It exists* $G \geq 0$ *s.t.* $||\nabla f(\boldsymbol{w})|| \leq G$.

Assumption 2. *(Bounded Variance.) It exists* $\sigma \geq 0$ *s.t.* $\mathbb{E}[||g(\boldsymbol{w}) - \nabla f(\boldsymbol{w})||^2] \leq \sigma^2$.

Assumption 3. *(L-smoothness.) It exists* $L > 0$ *s.t.* $||\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v})|| \leq L||\boldsymbol{w} - \boldsymbol{v}||, \forall \boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$.

1.1 Proof of Theorem 1

Based on the objective function of Sharpness-Aware Minimization (SAM), suppose we can obtain the noisy observation gradient $g(\boldsymbol{w})$ of true gradient $\nabla f(\boldsymbol{w})$, we can write the iteration of SAM:

$$\begin{cases} \boldsymbol{w}_{t+\frac{1}{2}} = & \boldsymbol{w}_t + \rho \cdot \frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||} \\ \boldsymbol{w}_{t+1} = & \boldsymbol{w}_t - \eta \cdot g(\boldsymbol{w}_{t+\frac{1}{2}}) \end{cases} \tag{1}$$

Lemma 1. *For any* $\rho > 0$, $L > 0$ *and the differentiable function* f*, we have the following inequality:*

$$\langle \nabla f(\boldsymbol{w}_t), \nabla f(\boldsymbol{w}_t + \rho \frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||}) \rangle \geq ||\nabla f(\boldsymbol{w}_t)||^2 - \rho L G$$

Proof. We first add and subtract a term $||\nabla f(\boldsymbol{w}_t)||$ to make use of classical inequalities bounding $\langle \nabla f(\boldsymbol{w}_1) - \nabla f(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle$ by $||\boldsymbol{w}_1 - \boldsymbol{w}_2||^2$ for smooth.

$$\begin{aligned} LHS =& \langle \nabla f(\boldsymbol{w}_t), \nabla f(\boldsymbol{w}_t + \rho \frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||}) - \nabla f(\boldsymbol{w}_t) \rangle + ||\nabla f(\boldsymbol{w}_t)||^2 \\ =& \frac{||\nabla f(\boldsymbol{w}_t)||}{\rho} \langle \frac{\rho}{||\nabla f(\boldsymbol{w}_t)||} \nabla f(\boldsymbol{w}_t), \nabla f(\boldsymbol{w}_t + \rho \frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||})) - \nabla f(\boldsymbol{w}_t) \rangle + ||\nabla f(\boldsymbol{w}_t)||^2 \\ \geq & - L \frac{||\nabla f(\boldsymbol{w}_t)||}{\rho} || \frac{\rho}{||\nabla f(\boldsymbol{w}_t)||} \nabla f(\boldsymbol{w}_t)||^2 + ||\nabla f(\boldsymbol{w}_t)||^2 \\ =& - ||\nabla f(\boldsymbol{w}_t)|| \rho L + ||\nabla f(\boldsymbol{w}_t)||^2 \\ \geq & - G \rho L + ||\nabla f(\boldsymbol{w}_t)||^2 \end{aligned}$$

where the first inequality is that

$$\langle \nabla f(\boldsymbol{w}_1) - \nabla f(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle \geq -L||\boldsymbol{w}_1 - \boldsymbol{w}_2||^2,$$

and the second inequality is the Assumption 1. □

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Lemma 2. *For $\rho > 0$, $L > 0$, the iteration 1 satisfies following inequality:*

$$\mathbb{E}\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \geq \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 - L^2\rho^2 - L\rho G$$

Proof. We denote the deterministic values of $\boldsymbol{w}_{t+\frac{1}{2}}$ as $\hat{\boldsymbol{w}}_{t+\frac{1}{2}} = \boldsymbol{w}_t + \rho\frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||}$ in this section. After we add and subtract the term $g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}})$, we have the following equation:

$$\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle = \langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_t + \rho\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||}) - g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle + \langle g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}}), \nabla f(\boldsymbol{w}_t)\rangle$$

For the first term, we bound it by using the smoothness of $g(\boldsymbol{w})$:

$$\begin{aligned}
-\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_t + \rho\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||}) - g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle \leq& \frac{1}{2}||g(\boldsymbol{w}_t + \rho\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||}) - g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}})||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\
\leq& \frac{L^2}{2}||\boldsymbol{w}_t + \rho\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||} - \hat{\boldsymbol{w}}_{t+\frac{1}{2}}||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\
=& \frac{L^2}{2}||\boldsymbol{w}_t + \rho\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||} - (\boldsymbol{w}_t + \rho\frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||})||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\
=& \frac{L^2\rho^2}{2}||\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||} - \frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||}||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\
\leq& L^2\rho^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2
\end{aligned}$$

For the second term, by using the Lemma 1, we have:

$$\begin{aligned}
\mathbb{E}\langle g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}}), \nabla f(\boldsymbol{w}_t)\rangle =& \langle \nabla f(\hat{\boldsymbol{w}}_{t+\frac{1}{2}}), \nabla f(\boldsymbol{w}_t)\rangle \\
=& \langle \nabla f(\boldsymbol{w}_t + \rho\frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||}), \nabla f(\boldsymbol{w}_t)\rangle \\
\geq& ||\nabla f(\boldsymbol{w}_t)||^2 - \rho LG
\end{aligned}$$

Asesmbling the two inequalities yields to the result. □

Lemma 3. *For $\eta \leq \frac{1}{L}$, the iteration 1 satisfies for all $t > 0$:*

$$\mathbb{E}f(\boldsymbol{w}_{t+1}) \leq \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + \eta L^2\rho^2 + (1 - L\eta)\eta LG\rho$$

Proof. By the smoothness of the function f, we obtain

$$\begin{aligned}
f(\boldsymbol{w}_{t+1}) \leq& f(\boldsymbol{w}_t) - \eta\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle + \frac{L\eta^2}{2}||g(\boldsymbol{w}_{t+\frac{1}{2}})||^2 \\
=& f(\boldsymbol{w}_t) - \eta\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle + \frac{L\eta^2}{2}(||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_{t+\frac{1}{2}})||^2 - ||\nabla f(\boldsymbol{w}_t)||^2 + 2\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle) \\
=& f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_{t+\frac{1}{2}})||^2 - (1 - L\eta)\eta\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \\
\leq& f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + L\eta^2||g(\boldsymbol{w}_t) - g(\boldsymbol{w}_{t+\frac{1}{2}})||^2 \\
& - (1 - L\eta)\eta\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \\
\leq& f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + L\eta^2L^2||\boldsymbol{w}_t - \boldsymbol{w}_{t+\frac{1}{2}}||^2 \\
& - (1 - L\eta)\eta\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \\
=& f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + \eta^2L^3\rho^2 \\
& - (1 - L\eta)\eta\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle
\end{aligned}$$

Taking the expectation and using Lemma 2 we obtain

$$\begin{aligned}
\mathbb{E}f(\boldsymbol{w}_{t+1}) \leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\mathbb{E}||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + \eta^2 L^3\rho^2 \\
& - (1-L\eta)\eta\mathbb{E}\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \\
\leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + \eta^2 L^3\rho^2 \\
& - (1-L\eta)\eta\mathbb{E}\langle \nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \\
\leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + \eta^2 L^3\rho^2 \\
& - (1-L\eta)\eta\left[\frac{1}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 - L^2\rho^2 - L\rho G\right] \\
= & \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + \eta^2 L^3\rho^2 + (1-L\eta)\eta L^2\rho^2 + (1-L\eta)\eta L\rho G \\
= & \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + \eta L^2\rho^2 + (1-L\eta)\eta LG\rho
\end{aligned}$$

□

Proposition 1. *Let $\eta_t = \frac{\eta_0}{\sqrt{t}}$ and perturbation amplitude ρ decay with square root of t,* e.g., *$\rho_t = \frac{\rho_0}{\sqrt{t}}$. For $\rho_0 \leq G\eta_0$ and $\eta_0 \leq \frac{1}{L}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 \leq C_1\frac{1}{\sqrt{T}} + C_2\frac{\log T}{\sqrt{T}},$$

where $C_1 = \frac{2}{\eta_0}(f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T))$ and $C_2 = 2(L\sigma^2\eta_0 + LG\rho_0)$.

Proof. By Lemma 3, we replace ρ and η with $\rho_t = \frac{\rho_0}{\sqrt{t}}$ and $\eta_t = \frac{\eta_0}{\sqrt{t}}$, we have

$$\mathbb{E}f(\boldsymbol{w}_{t+1}) \leq \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta_t}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta_t^2\sigma^2 + \eta_t L^2\rho_t^2 + (1-L\eta_t)\eta_t LG\rho_t.$$

Take telescope sum, we have

$$\mathbb{E}f(\boldsymbol{w}_T) - f(\boldsymbol{w}_0) \leq -\sum_{t=1}^{T}\frac{\eta_t}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\sum_{t=1}^{T}\frac{1}{t} + (L^2\eta_0\rho_0^2 - L^2G\eta_0^2\rho_0)\sum_{t=1}^{T}\frac{1}{t^{\frac{3}{2}}}$$

Under $\rho_0 \leq G\eta_0$, the last term will be less than 0, which means:

$$\mathbb{E}f(\boldsymbol{w}_T) - f(\boldsymbol{w}_0) \leq -\sum_{t=1}^{T}\frac{\eta_t}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\sum_{t=1}^{T}\frac{1}{t}.$$

With

$$\frac{\eta_T}{2}\sum_{t=1}^{T}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 \leq \sum_{t=1}^{T}\frac{\eta_t}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 \leq f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T) + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\sum_{t=1}^{T}\frac{1}{t},$$

we have

$$\begin{aligned}
\frac{\eta_0}{2\sqrt{T}}\sum_{t=1}^{T}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 \leq & f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T) + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\sum_{t=1}^{T}\frac{1}{t} \\
\leq & f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T) + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\log T.
\end{aligned}$$

Finally, we achieve the result:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 \leq \frac{2\cdot(f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T))}{\eta_0}\frac{1}{\sqrt{T}} + 2(L\sigma^2\eta_0 + LG\rho_0)\frac{\log T}{\sqrt{T}},$$

which shows that SAM can converge at the rate of $O(\log T/\sqrt{T})$. □

1.2 Proof of Theorem 2

Suppose we can obtain the noisy observation gradient $g(\boldsymbol{w}_t)$ of true gradient $\nabla f(\boldsymbol{w}_t)$, and the mask $\boldsymbol{m}$, we can write the iteration of SAM: Consider the iteration of Sparse SAM:

$$\begin{cases} \tilde{\boldsymbol{w}}_{t+\frac{1}{2}} = & \boldsymbol{w}_t + \rho \frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||} \odot \boldsymbol{m}_t \\ \boldsymbol{w}_{t+1} = & \boldsymbol{w}_t - g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}}) \end{cases} \tag{2}$$

Let us denote the difference as $\boldsymbol{w}_{t+\frac{1}{2}} - \tilde{\boldsymbol{w}}_{t+\frac{1}{2}} = \boldsymbol{e}_t$.

Lemma 4. *With $\rho > 0$, we have:*

$$\mathbb{E}\langle \nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}}) \rangle \geq \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 - 2L^2\rho^2 - L\rho G - L^2||\boldsymbol{e}_t||^2$$

Proof. Similar to Lemma 2, We denote the true gradient as $\hat{\boldsymbol{w}}_{t+\frac{1}{2}} = \boldsymbol{w}_t + \rho \frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||}$, and also add and subtract the item $g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})$:

$$\langle \nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}}) \rangle = \langle \nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}}) - g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}}) \rangle + \langle g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}}), \nabla f(\boldsymbol{w}_t) \rangle$$

For the first term, we bound it by using the smoothness of $g(\boldsymbol{w})$:

$$\begin{aligned} -\langle \nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}}) - g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}}) \rangle \leq & \frac{1}{2}||g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}}) - g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}})||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\ \leq & \frac{L^2}{2}||\tilde{\boldsymbol{w}}_{t+\frac{1}{2}} - \hat{\boldsymbol{w}}_{t+\frac{1}{2}}||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\ = & \frac{L^2}{2}||\boldsymbol{w}_{t+\frac{1}{2}} - \boldsymbol{e}_t - \hat{\boldsymbol{w}}_{t+\frac{1}{2}}||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\ \leq & L^2(||\boldsymbol{w}_{t+\frac{1}{2}} - \hat{\boldsymbol{w}}_{t+\frac{1}{2}}||^2 + ||\boldsymbol{e}_t||^2) + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\ = & L^2(\rho^2||\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||} - \frac{\nabla f(\boldsymbol{w}_t)}{||\nabla f(\boldsymbol{w}_t)||}||^2 + ||\boldsymbol{e}_t||^2) + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \\ \leq & 2L^2\rho^2 + L^2||\boldsymbol{e}_t||^2 + \frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 \end{aligned}$$

For the second term, we do the same in Lemma 2:

$$\mathbb{E}\langle g(\hat{\boldsymbol{w}}_{t+\frac{1}{2}}), \nabla f(\boldsymbol{w}_t) \rangle \geq ||\nabla f(\boldsymbol{w}_t)||^2 - \rho L G.$$

Assembling the two inequalities yields to the result. □

Lemma 5. *For $\eta \leq \frac{1}{L}$, the iteration 2 satisfies for all $t > 0$:*

$$\begin{aligned} \mathbb{E}f(\boldsymbol{w}_{t+1}) \leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + 2\eta L^2\rho^2 + (1 - L\eta)\eta L G\rho \\ & + (1 + L\eta)\eta L^2||\boldsymbol{e}_t||^2 \end{aligned}$$

Proof. By the smoothness of the function f, we obtain

$$
\begin{aligned}
f(\boldsymbol{w}_{t+1}) \leq & f(\boldsymbol{w}_t) - \eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle + \frac{L\eta^2}{2}||g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})||^2 \\
= & f(\boldsymbol{w}_t) - \eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle + \frac{L\eta^2}{2}(||\nabla f(\boldsymbol{w}_t) - g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})||^2 - ||\nabla f(\boldsymbol{w}_t)||^2 + 2\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle) \\
= & f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t) - g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})||^2 - (1 - L\eta)\eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle \\
\leq & f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + L\eta^2||g(\boldsymbol{w}_t) - g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})||^2 \\
& - (1 - L\eta)\eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle \\
\leq & f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + L\eta^2L^2||\boldsymbol{w}_t - \tilde{\boldsymbol{w}}_{t+\frac{1}{2}}||^2 \\
& - (1 - L\eta)\eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle \\
= & f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + \eta^2L^3||\boldsymbol{w}_t - \boldsymbol{w}_{t+\frac{1}{2}} + \boldsymbol{e}_t||^2 \\
& - (1 - L\eta)\eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle \\
\leq & f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + 2\eta^2L^3(||\boldsymbol{w}_t - \boldsymbol{w}_{t+\frac{1}{2}}||^2 + ||\boldsymbol{e}_t||^2) \\
& - (1 - L\eta)\eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle \\
= & f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + 2\eta^2L^3(\rho^2 + ||\boldsymbol{e}_t||^2) \\
& - (1 - L\eta)\eta\langle\nabla f(\boldsymbol{w}_t), g(\tilde{\boldsymbol{w}}_{t+\frac{1}{2}})\rangle
\end{aligned}
$$

Taking the expectation and using Lemma 4 we obtain

$$
\begin{aligned}
\mathbb{E}f(\boldsymbol{w}_{t+1}) \leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\mathbb{E}||\nabla f(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)||^2 + 2\eta^2L^3(\rho^2 + ||\boldsymbol{e}_t||^2) \\
& - (1 - L\eta)\eta\mathbb{E}\langle\nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \\
\leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + 2\eta^2L^3(\rho^2 + ||\boldsymbol{e}_t||^2) \\
& - (1 - L\eta)\eta\mathbb{E}\langle\nabla f(\boldsymbol{w}_t), g(\boldsymbol{w}_{t+\frac{1}{2}})\rangle \\
\leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{L\eta^2}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + 2\eta^2L^3(\rho^2 + ||\boldsymbol{e}_t||^2) \\
& - (1 - L\eta)\eta\left[\frac{1}{2}||\nabla f(\boldsymbol{w}_t)||^2 - 2L^2\rho^2 - L\rho G - L^2||\boldsymbol{e}_t||^2\right] \\
= & \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + 2\eta L^2\rho^2 + (1 - L\eta)\eta LG\rho \\
& + (1 + L\eta)\eta L^2||\rho\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||} \odot \boldsymbol{m}_t - \rho\frac{g(\boldsymbol{w}_t)}{||g(\boldsymbol{w}_t)||}||^2 \\
= & \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta^2\sigma^2 + 2\eta L^2\rho^2 + (1 - L\eta)\eta LG\rho \\
& + (1 + L\eta)\eta L^2||\boldsymbol{e}_t||^2
\end{aligned}
$$

□

Proposition 2. *Let us $\eta_t = \frac{\eta_0}{\sqrt{t}}$ and perturbation amplitude ρ decay with square root of t,* e.g.*, $\rho_t = \frac{\rho_0}{\sqrt{t}}$. With $\rho_0 \leq G\eta_0/2$, we have:*

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}||\nabla f(\boldsymbol{w})||^2 \leq C_3\frac{1}{\sqrt{T}} + C_4\frac{\log T}{\sqrt{T}},
$$

where $C_3 = \frac{2}{\eta_0}(f(\boldsymbol{w}_0 - \mathbb{E}f(\boldsymbol{w}_T) + \eta_0L^2\rho^2(1 + \eta_0L)\frac{\pi^2}{6})$ and $C_4 = 2(L\sigma^2\eta_0 + LG\rho_0)$.

Proof. By taking the expectation and using Lemma 4, and taking the schedule to be $\eta_t = \frac{\eta_0}{\sqrt{t}}$, $\rho_t = \frac{\rho_0}{\sqrt{t}}$, we obtain:

$$\begin{aligned}\mathbb{E}f(\boldsymbol{w}_{t+1}) \leq & \mathbb{E}f(\boldsymbol{w}_t) - \frac{\eta_t}{2}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 + L\eta_t^2\sigma^2 + 2\eta_t L^2\rho_t^2 + (1 - L\eta_t)\eta_t LG\rho_t \\ & + (1 + L\eta_t)\eta_t L^2||\boldsymbol{e}_t||^2\end{aligned}$$

By taking sum and bound ρ with $\frac{G\eta_0}{2}$, we have:

$$\begin{aligned}\frac{\eta_0}{2\sqrt{T}}\sum_{t=1}^{T}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 \leq & f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T) + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\sum_{t=1}^{T}\frac{1}{t} \\ & + \sum_{t=1}^{T}(1 + L\eta_t)\eta_t L^2||\boldsymbol{e}_t||^2 \\ \leq & f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T) + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\sum_{t=1}^{T}\frac{1}{t} \\ & + \eta_0 L^2\rho_0^2\sum_{t=1}^{T}\frac{1}{t^{\frac{3}{2}}} + \eta_0^2 L^3\rho^2\sum_{t=1}^{T}\frac{1}{t^2} \\ \leq & f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T) + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\sum_{t=1}^{T}\frac{1}{t} \\ & + \eta_0 L^2\rho_0^2\sum_{t=1}^{T}\frac{1}{t^2} + \eta_0^2 L^3\rho_0^2\sum_{t=1}^{T}\frac{1}{t^2} \\ \leq & f(\boldsymbol{w}_0) - \mathbb{E}f(\boldsymbol{w}_T) + (L\sigma^2\eta_0^2 + LG\rho_0\eta_0)\log T \\ & + \eta_0 L^2\rho_0^2(1 + \eta_0 L)\frac{\pi^2}{6}\end{aligned}$$

Finally, we achieve the result:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}||\nabla f(\boldsymbol{w}_t)||^2 \leq & \frac{2(f(\boldsymbol{w}_0 - \mathbb{E}f(\boldsymbol{w}_T) + \eta_0 L^2\rho^2(1 + \eta_0 L)\frac{\pi^2}{6})}{\eta_0}\frac{1}{\sqrt{T}} \\ & + 2(L\sigma^2\eta_0 + LG\rho_0)\frac{\log T}{\sqrt{T}}\end{aligned}$$

□

So far, we have completed the proof of the theory in the main text.

2 More Experimets

VGG on CIFAR10. To further confirm the model-agnostic characteristic of our Sparse SAM, we test the VGG-style architecture on CIFAR10. Following [**?**], we test SSAM training the VGG11-BN on CIFAR10 and the results are shown in the following Table 1. The perturbation magnitude ρ is set to 0.05.

Table 1: Test accuracy of VGG11-BN on CIFAR10 with proposed Sparse SAM.

Model	Dataset	Optimizer	Sparsity	Accuracy
VGG11-BN	CIFAR10	SGD	/	93.42%
		SAM	0%	93.87%
		SSAM-F/SSAM-D	50%	**94.03%**/93.79%
			80%	93.83%/**93.95%**
			90%	93.76%/93.85%
			95%	93.77%/93.48%
			98%	93.54%/93.54%
			99%	93.47%/93.33%

SAM with different perturbation magnitude ρ. We determine the perturbation magnitude ρ by using grid search. We choose ρ from the set $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ for CIFAR, and choose ρ from $\{0.01, 0.02, 0.05, 0.07, 0.1, 0.2\}$ for ImageNet. We show the results when varying ρ in Table 2 and Table 3. From this table, we can see that the $\rho = 0.1$, $\rho = 0.2$ and $\rho = 0.07$ is sutiable for CIFAR10, CIFAR100 and ImageNet respectively.

Table 2: Test accuracy of ResNet18 and WideResNet28-10 on CIFAR10 and CIFAR100 with different perturbation magnitude ρ.

Dataset	SAM ρ	0.01	0.02	0.05	0.1	0.2	0.5
CIFAR10	ResNet18	96.58%	96.54%	96.68%	**96.83%**	96.32%	93.16%
	WideResNet28-10	97.26%	97.34%	97.31%	**97.48%**	97.29%	95.13%
CIFAR100	ResNet18	79.56%	79.98%	80.71%	80.65%	**81.03%**	77.57%
	WideResNet28-10	82.25%	83.04%	83.47%	83.47%	**84.20%**	84.03%

Table 3: Test accuracy of ResNet50 on ImageNet with different perturbation magnitude ρ.

datasets	SAM ρ	0.01	0.02	0.05	0.07	0.1	0.2
ImageNet	ResNet50	76.63%	76.78%	77.12%	**77.25%**	77.00%	76.37%

Ablations of Masking Strategy. For further verification of our masking strategy, we perform more ablations in this paragraph. For the mask update in SSAM-F, the parameters with largest fisher information are selected. Compared with SSAM-F, we consider the random mask, *i.e.*, the mask is randomly generated to choose which parameters are perturbated. For the mask update in SSAM-D, we first drop the flattest weights and then random grow some weights. Compared with SSAM-D, we experiment the SSAM-D which drops randomly or drops the sharpest weights, *i.e.*, the weights with large gradients. The results of ablations are shown in Table 4. The results show that random strategies are less effective than our SSAM. The performance of SSAM-D dropping sharpest weights drops a lot even worse than random strategy, which is consistent with our conjecture.

Influence of hyper-parameters We first examine the effect of the number of sample size N_F of SSAM-F in Table 5. From it we can see that a certain number of samples is enough for the approximation of data distribution in SSAM-F, *e.g.*, $N_F = 128$, which greatly saves the computational cost of SSAM-F. In Table 6, we also report the influence of the mask update interval on SSAM-F and SSAM-D. The results show that the performance degrades as the interval becom longer, suggesting that dense mask updates are necessary for our methods. Both of them are ResNet18 on CIFAR10.

Table 4: Ablation of different masking strategy.

Model	Dataset	Optimizer	Strategy	Accuracy
ResNet50	ImageNet	SGD	/	76.67%
		SAM	/	77.25%
		Sparse SAM	Random Mask	77.08%(-0.17)
		SSAM-F	Topk Fisher Information	77.31%(+0.06)
		SSAM-D	Random Drop	77.08%(-0.17)
			Drop Sharpnest weights	76.68%(-0.57)
			Drop Flattest weights	77.25%(-0.00)

Table 5: Results of ResNet18 on CIFAR10 with different number of samples N_F in SSAM-F. 'Time' reported in table is the time cost to calculate Fisher Information based on N_F samples.

Sparsity	N_F	Acc	Time
0.5	16	96.77%	1.49s
	128	96.84%	4.40s
	512	96.67%	15.35s
	1024	96.83%	30.99s
	2048	96.68%	56.23s
	4096	96.66%	109.31s
0.9	16	96.79%	1.47s
	128	96.50%	5.42s
	512	96.43%	15.57s
	1024	96.75%	29.24s
	2048	96.62%	57.72s
	4096	96.59%	110.65s

Table 6: Results of ResNet18 on CIFAR10 with different T_m intervals of update mask. The left of '/' is accuracy of SSAM-F, while the right is SSAM-D.

Sparsity	T_m	Acc
0.5	1	96.81%/96.74%
	2	96.51%/96.74%
	5	96.83%/96.60%
	10	96.71%/96.73%
	50	96.65%/96.75%
	Fixed	96.57%/96.52%
0.9	1	96.70%/96.65%
	2	93.75%/96.63%
	5	96.51%/96.69%
	10	96.67%/96.74%
	50	96.64%/96.66%
	Fixed	96.21%/96.46%

3 Limitation and Societal Impacts

Limitation. Our method Sparse SAM is mainly based on sparse operation. At present, the sparse operation that has been implemented is only 2:4 sparse operation. The 2:4 sparse operation requires that there are at most two non-zero values in four contiguous memory, which does not hold for us. To sum up, there is currently no concrete implemented sparse operation to achieve training acceleration. But in the future, with the development of hardware for sparse operation, our method has great potential to achieve truly training acceleration.

Societal Impacts. In this paper, we provide a Sparse SAM algorithm that reduces computation burden and improves model generalization. In the future, we believe that with the development of deep learning, more and more models need the guarantee of generalization and also the efficient training. Different from the work on sparse networks, our proposed Sparse SAM does not compress the model for hardware limited device, but instead accelerates model training. It's helpful for individuals or laboratories which are lack computing resources.