A Technical lemmas

Lemma A.1. ((Zhang et al., 2020) Descent Inequality) Suppose objective function $f(\cdot)$ satisfies Assumption 2, and $c > 0$ be a constant. For any $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$, as long as $||\mathbf{x}_k - \mathbf{x}_{k+1}|| \leq \frac{c}{L_1}$, we have

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + (\mathbf{x}_{k+1} - \mathbf{x}_k)^T \nabla f(\mathbf{x}_k) + \frac{AL_0 + BL_1 ||\nabla f(\mathbf{x}_k)||}{2} ||\mathbf{x}_{k+1} - \mathbf{x}_k||^2 \tag{1}$$

where $A = 1 + e^c - \frac{e^c - 1}{c}$, $B = \frac{e^c - 1}{c}$. Note that A and B are monotonically increasing functions w.r.t. $c > 0$.

Lemma A.2. (Lemma 3.1) For all $\mathbf{g} \in \mathbb{R}^d$, and random vector $\mathbf{s} \sim \mathcal{R}$ where $\mathcal{R}$ is the Rademacher distribution, i.e., each element $\mathrm{s} \sim \{+1, -1\}$ with equal chances and $\mathbb{E}_{\mathbf{s}\sim\mathcal{R}} ||\mathbf{s}||_2^2 = d$, then $\mathbb{E}_{\mathbf{s}\sim\mathcal{R}} |\langle \mathbf{g}, \mathbf{s} \rangle| \geq \frac{1}{\sqrt{2}} ||\mathbf{g}||_2$.

Proof.

$$|\langle \mathbf{g}, \mathbf{s} \rangle| = |\sum_{i=1}^{d} \mathrm{g}_i \mathrm{s}_i| \tag{2}$$

According to Khintchine inequality (Khintchine, 1923), i.e.,

$$A_p (\sum_{i=1}^{d} |\mathrm{g}_i|^2)^{\frac{1}{2}} \leq (\mathbb{E} |\sum_{i=1}^{d} \mathrm{g}_i \mathrm{s}_i|^p)^{\frac{1}{p}} \leq B_p (\sum_{i=1}^{d} |\mathrm{g}_i|^2)^{\frac{1}{2}}$$

where

$$A_p = \begin{cases} 2^{\frac{1}{2} - \frac{1}{p}} & 0 < p < p_0 \\ 2^{\frac{1}{2}} (\Gamma((p+1)/2)/\sqrt{\pi})^{\frac{1}{p}} & p_0 < p < 2 \\ 1 & 2 \leq p < \infty. \end{cases} \quad B_p = \begin{cases} 1 & 0 < p \leq 2 \\ 2^{\frac{1}{2}} (\Gamma((p+1)/2)/\sqrt{\pi})^{\frac{1}{p}} & 2 < p < \infty. \end{cases}$$

where $p_0 \approx 1.847$ and Γ is the Gamma function, we have

$$\frac{1}{\sqrt{2}} ||\mathbf{g}||_2 \leq \mathbb{E} |\sum_{i=1}^{d} \mathrm{g}_i \mathrm{s}_i| \leq ||\mathbf{g}||_2,$$

Combined with equation 2, we have

$$\frac{1}{\sqrt{2}} ||\mathbf{g}||_2 \leq \mathbb{E}_{\mathbf{s}\sim\mathcal{R}} |\langle \mathbf{g}, \mathbf{s} \rangle| \leq ||\mathbf{g}||_2.$$

This completes the proof. □

B Convergence analysis under the general smoothness assumption

B.1 Progressive bound of S2P

Lemma B.1. (Lemma 3.2) (Progressive bound) Suppose objective function $f(\cdot)$ satisfies Assumption 1 and $||\nabla f(\mathbf{x}_k)||_2 \geq \epsilon_g$. If we run algorithm 1 with step size $\alpha = \frac{\sqrt{2}\epsilon_g}{2Ld}$, we have following progressive bound $\mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) | \mathbf{x}_k] \leq -\Omega(\frac{\epsilon_g^2}{Ld})$, where $\mathbb{E}[\cdot | \mathbf{x}_k]$ denotes the conditional expectation w.r.t. $\mathbf{x}_k$.

Proof. Using L-gradient Lipschitz, we have (descent lemma)

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|\mathbf{x}_k] \\
&\quad \leq \mathbb{E}[\nabla f(\mathbf{x}_k)^T(\mathbf{x}_{k+1} - \mathbf{x}_k)|\mathbf{x}_k] + \frac{L}{2}\mathbb{E}[||\mathbf{x}_{k+1} - \mathbf{x}_k||^2] \\
&\quad = -\alpha\mathbb{E}|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{L\alpha^2}{2}\mathbb{E}||\mathbf{s}_k||_2^2 \quad \text{Take updating step} \\
&\quad = -\alpha\mathbb{E}|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{L\alpha^2 d}{2}
\end{aligned}$$

Lemma 2 shows that $\mathbb{E}_{\mathbf{s}_k\sim\mathcal{R}}|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| \geq \frac{1}{\sqrt{2}}||\nabla f(\mathbf{x}_k)||_2$, then

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|\mathbf{x}_k] &\leq -\frac{\alpha}{\sqrt{2}}||\nabla f(\mathbf{x}_k)||_2 + \frac{L\alpha^2 d}{2} \\
&\leq -\frac{\alpha}{\sqrt{2}}\epsilon_g + \frac{L\alpha^2 d}{2}
\end{aligned}$$

To guarantee convergence, $\alpha \sim [0, \frac{\sqrt{2}\epsilon_g}{Ld}]$, then suppose $\alpha = \frac{\sqrt{2}\epsilon_g}{2Ld}$, we have $\mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|\mathbf{x}_k] \leq -\frac{\epsilon_g^2}{4Ld}$ which completes the proof. □

B.2 Query complexity of S2P

Theorem B.1. (Theorem 3.1) (Query complexity) Suppose objective function $f(\cdot)$ satisfies Assumption 1. If we run algorithm 1 with step size strategy options 1 or 2, the algorithm returns in expectation an ϵ-first-order stationary point in $\mathcal{O}(\frac{d}{\epsilon^2})$ function evaluations.

Proof. Using L-gradient Lipschitz, we have (descent lemma)

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_{k+1})|\mathbf{x}_k] &\leq f(\mathbf{x}_k) + \mathbb{E}[\nabla f(\mathbf{x}_k)^T(\mathbf{x}_{k+1} - \mathbf{x}_k)|\mathbf{x}_k] + \frac{L}{2}\mathbb{E}[||\mathbf{x}_{k+1} - \mathbf{x}_k||^2] \\
&= f(\mathbf{x}_k) - \alpha\mathbb{E}|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{L\alpha^2}{2}\mathbb{E}||\mathbf{s}_k||_2^2 \quad \text{Take updating step} \\
&= f(\mathbf{x}_k) - \alpha\mathbb{E}|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{L\alpha^2 d}{2} \qquad (3)
\end{aligned}$$

Option 1. Stationary step size

Lemma 2 shows that $\mathbb{E}_{\mathbf{s}_k\sim\mathcal{R}}|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| \geq \frac{1}{\sqrt{2}}||\nabla f(\mathbf{x}_k)||_2$, then inequality (3) can be reformulated as

$$\mathbb{E}[f(\mathbf{x}_{k+1})|\mathbf{x}_k] \leq f(\mathbf{x}_k) - \frac{\alpha}{\sqrt{2}}||\nabla f(\mathbf{x}_k)||_2 + \frac{L\alpha^2 d}{2}$$

Taking expectations in the above inequality w.r.t. $\mathbf{s}_k$ conditional on $\mathbf{x}_k$, and denoting $\theta_k = \mathbb{E}[f(\mathbf{x}_{k+1})]$ and $g_k = \mathbb{E}[||\nabla f(\mathbf{x}_k)||_2]$, we have

$$\begin{aligned}
\theta_{k+1} &\leq \theta_k - \frac{\alpha g_k}{\sqrt{2}} + \frac{L\alpha^2 d}{2} \\
g_k &\leq \sqrt{2}(\frac{\theta_k - \theta_{k+1}}{\alpha} + \frac{L\alpha d}{2}) \\
\sum_{k=0}^{K} g_k &\leq \sqrt{2}(\frac{\theta_0 - \theta_{k+1}}{\alpha} + \frac{KL\alpha d}{4})
\end{aligned}$$

We can conclude that there exists an iteration $j \sim [0, K]$ such that

$$\begin{aligned}
g_j &\leq \sqrt{2}(\frac{\theta_0 - \theta_{k+1}}{\alpha K} + \frac{L\alpha d}{2}) \\
g_j &\leq \sqrt{2}(\frac{(f(\mathbf{x}_0) - f^\star)\sqrt{Kd}}{\alpha_0 K} + \frac{L\alpha_0\sqrt{d}}{2\sqrt{K}}) \text{ By taking } \alpha = \frac{\alpha_0}{\sqrt{Kd}} \\
g_j &\leq \frac{\sqrt{2d}}{\sqrt{K}}(\frac{(f(\mathbf{x}_0) - f^\star)}{\alpha_0} + \frac{L\alpha_0}{2})
\end{aligned}$$

Then let $\frac{\sqrt{2d}}{\sqrt{K}}(\frac{(f(\mathbf{x}_0)-f^\star)}{\alpha_0} + \frac{L\alpha_0}{2}) \leq \epsilon$, we have

$$K \geq \frac{2d}{\epsilon^2}(\frac{(f(\mathbf{x}_0) - f^\star)}{\alpha_0} + \frac{L\alpha_0}{2})^2,$$

, which completes the proof for option 1.

Option 2. Dynamic step size

Taking expectations in the above inequality (3) w.r.t. $\mathbf{s}_k$ conditional on $\mathbf{x}_k$, and denoting $\theta_k = \mathbb{E}[f(\mathbf{x}_{k+1})]$, we have

$$\theta_{k+1} \leq \theta_k - \alpha|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{L\alpha^2 d}{2} \tag{4}$$

We know that the best $\alpha_k^{opt} = \frac{|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k|}{Ld}$, and we can approximate the best step size with $\alpha_k = \frac{|f(\mathbf{x}+\rho\mathbf{s}_k)-f(\mathbf{x}-\rho\mathbf{s}_k)|}{2\rho Ld}$ (or $\alpha_k = \alpha_0\frac{|f(\mathbf{x}+\rho\mathbf{s}_k)-f(\mathbf{x}-\rho\mathbf{s}_k)|}{2\rho}$ where $\alpha_0 = \frac{1}{Ld}$) where ρ is a scalar.

Before continuing working on the inequality (4), we estimate the error between the best step size and the approximated step size, $|\delta_k| := |\alpha_k - \alpha_k^{opt}|$, firstly.

$$\begin{aligned}
|\delta_k| &= \frac{1}{2\rho Ld}\big||f(\mathbf{x}+\rho\mathbf{s}_k) - f(\mathbf{x}-\rho\mathbf{s}_k)| - 2\rho|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k|\big| \\
&\leq \frac{1}{2\rho Ld}|f(\mathbf{x}+\rho\mathbf{s}_k) - f(\mathbf{x}-\rho\mathbf{s}_k) - 2\rho\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| \qquad (5) \\
&= \frac{1}{2\rho Ld}|(f(\mathbf{x}+\rho\mathbf{s}_k) - f(\mathbf{x}) - \rho\nabla f(\mathbf{x}_k)^T\mathbf{s}_k) - (f(\mathbf{x}-\rho\mathbf{s}_k) - f(\mathbf{x}) + \rho\nabla f(\mathbf{x}_k)^T\mathbf{s}_k)| \\
&\leq \frac{1}{2\rho Ld}(\frac{L}{2}\rho^2||\mathbf{s}_k||^2 + \frac{L}{2}\rho^2||\mathbf{s}_k||^2) \qquad (6) \\
&\leq \frac{\rho}{2} \qquad (7)
\end{aligned}$$

Note that inequality (5) applied reverse triangle inequality and inequality (6) applied the equivalent definitions of L-smooth function $|f(\mathbf{x}+\rho\mathbf{s}_k) - f(\mathbf{x}) - \rho\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| \leq \frac{L}{2}||\rho\mathbf{s}_k||^2$.

Suppose we do take $\alpha_k = \frac{|f(\mathbf{x}+\rho\mathbf{s}_k)-f(\mathbf{x}-\rho\mathbf{s}_k)|}{2\rho Ld}$ and substitute $\alpha_k = \alpha_k^{opt} + \delta_k$, inequality (4) can be reformulated as

$$\begin{aligned}
\theta_{k+1} &\leq \theta_k - (\alpha_k^{opt} + \delta_k)|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{L(\alpha_k^{opt} + \delta_k)^2 d}{2} \\
&= \theta_k - \frac{|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k|^2}{Ld} - \delta_k|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k|^2}{2Ld} + \delta_k|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| + \frac{Ld\delta_k^2}{2} \\
&= \theta_k - \frac{|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k|^2}{2Ld} + \frac{Ld\delta_k^2}{2} \\
&\leq \theta_k - \frac{|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k|^2}{2Ld} + \frac{Ld\rho^2}{8} \quad \text{Apply inequality (7)} \\
&\leq \theta_k - \frac{||\nabla f(\mathbf{x}_k)||^2}{4Ld} + \frac{Ld\rho^2}{8} \quad \text{Apply Lemma 2} \qquad (8)
\end{aligned}$$

Note that it actually put requirement on ρ to guarantee convergence, i.e., for ρ_k in each iterations, we need $0 < \rho \leq \frac{\sqrt{2}||\nabla f(\mathbf{x}_k)||}{Ld}$.

Continually, inequality (8) further can be re-formulated as

$$\begin{aligned} ||\nabla f(\mathbf{x}_k)||^2 &\leq 4Ld(\theta_k - \theta_{k+1}) + \frac{\rho^2}{2} \\ \sum_{k=0}^{K} ||f(\mathbf{x}_k)||^2 &\leq 4Ld(\theta_0 - \theta_{k+1}) + \frac{K\rho^2}{2} \end{aligned}$$

We can conclude that there exists an iteration $j \sim [0, K]$ such that

$$||f(\mathbf{x}_j)||^2 \leq \frac{4Ld(\theta_0 - \theta_{k+1})}{K} + \frac{\rho^2}{2} \leq \frac{4Ld(f(\mathbf{x}_0) - f^\star)}{K} + \frac{\rho^2}{2}$$

which further concludes that we need

$$K \geq \frac{4Ld(f(\mathbf{x}_0) - f^\star)}{\epsilon^2 - \frac{\rho^2}{2}}, \tag{9}$$

iterations to reach ϵ-first-order stationary point ($||f(\mathbf{x}_j)|| \leq \epsilon$).

Meanwhile, we require that $0 < \rho_k \leq \frac{\sqrt{2}||\nabla f(\mathbf{x}_k)||}{Ld}$ for ρ_k in each iterations, and it can be set to a small value universally. E.g., $0 < \rho \leq \frac{\sqrt{2}\epsilon}{Ld}$, then we have $K \geq \frac{4Ld(f(\mathbf{x}_0) - f^\star)}{\epsilon^2(1 - \frac{1}{L^2 d^2})}$.

Then, we can safely conclude that the algorithm returns in expectation an ϵ-first-order stationary point in $\mathcal{O}(\frac{d}{\epsilon^2})$ function evaluations, which completes the proof for option 2. □

C Convergence analysis under the relaxed smoothness assumption

C.1 Progressive bound of S2P

Lemma C.1. (Lemma 3.3) (Progressive bound) Suppose objective function $f(\cdot)$ satisfies Assumption 2 and $||\nabla f(\mathbf{x}_k)||_2 \geq \epsilon_g$. If we run algorithm 1 with step size $\alpha = \frac{\sqrt{2}\epsilon_g}{2(AL_0 + BL_1\epsilon_g)d}$, we have following progressive bound $\mathbb{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|\mathbf{x}_k] \leq -\Omega(\frac{\epsilon_g^2}{(AL_0 + BL_1\epsilon_g)d})$, where $\mathbb{E}[\cdot|\mathbf{x}_k]$ denotes the conditional expectation w.r.t. $\mathbf{x}_k$, and constants $A = 1.01, B = 1.01$.

Proof. Give the decent lemma inequality (1), we have

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_{k+1})] &\leq f(\mathbf{x}_k) - \alpha\mathbb{E}[\mathbf{g}_k^T \nabla f(\mathbf{x}_k)] + \frac{AL_0 + BL_1||\nabla f(\mathbf{x}_k)||}{2}\mathbb{E}[\alpha^2||\mathbf{g}_k||^2] \\ &= f(\mathbf{x}_k) - \alpha\mathbb{E}[|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)|] + \frac{AL_0 + BL_1||\nabla f(\mathbf{x}_k)||}{2}\mathbb{E}[\alpha^2||\mathbf{s}_k||^2] \quad \text{Take updating step} \\ &\leq f(\mathbf{x}_k) - \frac{\alpha}{\sqrt{2}}||\nabla f(\mathbf{x}_k)|| + \alpha^2\frac{AL_0 + BL_1||\nabla f(\mathbf{x}_k)||}{2}d \quad \text{Lemma 2} \end{aligned} \tag{10}$$

Suppose $||\nabla f(\mathbf{x}_k)|| \geq \epsilon_g$, and to guarantee convergence $\alpha \in [0, \frac{\sqrt{2}\epsilon_g}{(AL_0 + BL_1\epsilon_g)d}]$. Let $\alpha = \frac{\sqrt{2}\epsilon_g}{2(AL_0 + BL_1\epsilon_g)d}$, we have

$$\mathbb{E}[f(\mathbf{x}_{k+1})] \leq f(\mathbf{x}_k) - \frac{\epsilon_g^2}{4(AL_0 + BL_1\epsilon_g)d}.$$

which completes the proof.

Note that for the specific value of A and B, we have $A = 1 + e^c - \frac{e^c - 1}{c}$, $B = \frac{e^c - 1}{c}$ and $||\mathbf{x}_{k+1} - \mathbf{x}_k|| = ||\alpha\mathbf{s}_k|| = \frac{\sqrt{2}\epsilon_g}{2(AL_0 + BL_1\epsilon_g)\sqrt{d}} \leq \frac{c}{L_1} \rightarrow c \geq \frac{\sqrt{2}L_1\epsilon_g}{2(AL_0 + BL_1\epsilon_g)\sqrt{d}} \rightarrow c \geq \frac{1}{\sqrt{2d}B} \rightarrow e^c \geq 1 + \frac{1}{\sqrt{2d}}$. It is easy to see that such c exists, we can safely consider $A = 1.01, B = 1.01$ for simplicity (under large d) since A and B are expected to be small values. □

C.2 Query complexity of S2P

Theorem C.1. (Theorem 3.2) (Query complexity) Suppose objective function $f(\cdot)$ satisfies Assumption 2. If we run algorithm 1 with step size strategy options 3 or 4, the algorithm returns in expectation an ϵ-first-order stationary point in $\mathcal{O}(\frac{d}{\epsilon^2})$ function evaluations.

Proof. Give the decent lemma inequality (1), we have

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_{k+1})] &\le f(\mathbf{x}_k) - \alpha\mathbb{E}[\mathbf{g}_k^T\nabla f(\mathbf{x}_k)] + \frac{AL_0 + BL_1||\nabla f(\mathbf{x}_k)||}{2}\mathbb{E}[\alpha^2||\mathbf{g}_k||^2] \\
&= f(\mathbf{x}_k) - \alpha\mathbb{E}[|\mathbf{s}_k^T\nabla f(\mathbf{x}_k)|] + \alpha^2\frac{AL_0 + BL_1||\nabla f(\mathbf{x}_k)||}{2}\mathbb{E}[||\mathbf{s}_k||^2] \;\; \text{Take updating step}
\end{aligned} \tag{11}$$

Option 1. Stationary step size

Lemma 2 shows that $\mathbb{E}_{\mathbf{s}_k\sim\mathcal{R}}|\nabla f(\mathbf{x}_k)^T\mathbf{s}_k| \ge \frac{1}{\sqrt{2}}||\nabla f(\mathbf{x}_k)||_2$, then inequality (11) can be reformulated as

$$\mathbb{E}[f(\mathbf{x}_{k+1})] \le f(\mathbf{x}_k) - \frac{\alpha}{\sqrt{2}}||\nabla f(\mathbf{x}_k)|| + \alpha^2\frac{AL_0 + BL_1||\nabla f(\mathbf{x}_k)||}{2}d$$

Taking expectations in the above inequality w.r.t. $\mathbf{s}_k$ conditional on $\mathbf{x}_k$, and denoting $\theta_k = \mathbb{E}[f(\mathbf{x}_{k+1})]$ and $g_k = \mathbb{E}[||\nabla f(\mathbf{x}_k)||]$, we have

$$\begin{aligned}
\theta_{k+1} &\le \theta_k - \frac{\alpha}{\sqrt{2}}g_k + \alpha^2\frac{AL_0 + BL_1 g_k}{2}d \\
g_k(\frac{\sqrt{2}\alpha - B\alpha^2L_1d}{2}) &\le \theta_k - \theta_{k+1} + \frac{A\alpha^2L_0d}{2} \\
g_k &\le \frac{2(\theta_k - \theta_{k+1})}{\sqrt{2}\alpha - B\alpha^2L_1d} + \frac{A\alpha^2L_0d}{\sqrt{2}\alpha - B\alpha^2L_1d} \\
\sum_{k=0}^{K} g_k &\le \frac{2(\theta_0 - \theta_{k+1})}{\sqrt{2}\alpha - B\alpha^2L_1d} + \frac{KA\alpha^2L_0d}{\sqrt{2}\alpha - B\alpha^2L_1d}
\end{aligned}$$

We can conclude that there exists an iteration $j \sim [0, K]$ such that

$$\begin{aligned}
g_j &\le \frac{2(\theta_0 - \theta_{K+1})}{(\sqrt{2}\alpha - B\alpha^2L_1d)K} + \frac{A\alpha^2L_0d}{\sqrt{2}\alpha - B\alpha^2L_1d} \\
&\le \frac{2(f(\mathbf{x}_0) - f^\star)}{(\sqrt{2}\alpha - B\alpha^2L_1d)K} + \frac{A\alpha^2L_0d}{\sqrt{2}\alpha - B\alpha^2L_1d}
\end{aligned} \tag{12}$$

Suppose $\alpha = \frac{\sqrt{2}}{BL_1\sqrt{dK}}$, inequality (12) can be reformulated as

$$g_j \le \frac{B(f(\mathbf{x}_0) - f^\star)L_1\sqrt{d}}{\sqrt{K} - \sqrt{d}} + \frac{AL_0\sqrt{d}}{BL_1(\sqrt{K} - \sqrt{d})}.$$

Under this setting, we can see that the g_j can be continually decreased with at least $K > d$, which further shows that it need

$$K \ge (\sqrt{d} + \frac{AL_0\sqrt{d} + BL_1(f(\mathbf{x}_0) - f^\star)\sqrt{d}}{\epsilon})^2$$

iterations to reach ϵ-first-order stationary point. Then, we can safely conclude that the algorithm returns in expectation an ϵ-first-order stationary point in $\mathcal{O}(\frac{d}{\epsilon^2})$ function evaluations, which completes the proof for option 1.

Note that for the specific value of A and B, we have $A = 1+e^c-\frac{e^c-1}{c}$, $B = \frac{e^c-1}{c}$ and $||\mathbf{x}_{k+1}-\mathbf{x}_k|| = ||\alpha\mathbf{s}_k|| = \frac{\sqrt{2}}{BL_1\sqrt{K}} \le \frac{c}{L_1} \to c \ge \frac{\sqrt{2}}{B\sqrt{K}} \to e^c \ge 1 + \sqrt{\frac{2}{K}}$. It is easy to see that such c exists, we can safely consider $A = 1.01, B = 1.01$ for simplicity (under large d) since A and B are expected to be small values.

Option 2. Dynamic step size

Taking expectations in the above inequality (11) w.r.t. $\mathbf{s}_k$ conditional on $\mathbf{x}_k$, and denoting $\theta_k = \mathbb{E}[f(\mathbf{x}_{k+1})]$, we have

$$\begin{aligned}\theta_{k+1} &\leq \theta_k - \alpha|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)| + \alpha^2 \frac{AL_0 + BL_1||\nabla f(\mathbf{x}_k)||}{2} d \\ &\leq \theta_k - \alpha|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)| + \alpha^2 \frac{AL_0 + \sqrt{2}BL_1|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)|}{2} d. \end{aligned} \tag{13}$$

It is easy to know that $\alpha_k^{opt} = \frac{|\mathbf{s}^T \nabla f(\mathbf{x}_k)|}{(AL_0 + \sqrt{2}BL_1|\mathbf{s}^T \nabla f(\mathbf{x}_k)|)d}$. Let $|\gamma_k| = \frac{|f(\mathbf{x}_k + \rho \mathbf{s}_k) - f(\mathbf{x}_k - \rho \mathbf{s}_k)|}{2\rho}$, and we approximate the best step size with $\alpha_k = \frac{|\gamma_k|}{(AL_0 + \sqrt{2}BL_1|\gamma_k|)d}$ and denote the approximation error as $|\delta_k| := |\alpha_k - \alpha_k^{opt}|$.

Before we continue working on the inequality (13), we derive the upper bound of $|\delta_k|$ for our following analysis. Firstly, we denote $|\epsilon_\rho| := \big||\mathbf{s}^T \nabla f(\mathbf{x}_k)| - |\gamma_k|\big| = \big||\mathbf{s}^T \nabla f(\mathbf{x}_k)| - \frac{|f(\mathbf{x}_k + \rho \mathbf{s}_k) - f(\mathbf{x}_k - \rho \mathbf{s}_k)|}{2\rho}\big| = \mathcal{O}(\rho^2 d^{3/2})$ (Taylor expansion). So that, we can define $|\epsilon_\rho| \leq \xi \rho^2 d^{3/2}$ where ξ is a constant associated with third-order property of f. Note $d^{3/2}$ is the compensation of normalizing $\mathbf{s}$.

Specifically, we try to prove $|\delta_k| \leq |\epsilon_\rho|$. We define a new function $g(x) = \frac{x}{AL_0 + \sqrt{2}BL_1 x}$, then to prove $|\delta_k| \leq |\epsilon_\rho|$ is equivalent to prove $|g(|\mathbf{s}^T \nabla f(\mathbf{x}_k)|) - g(|\gamma_k|)| \leq d\big||\mathbf{s}^T \nabla f(\mathbf{x}_k)| - |\gamma_k|\big|$, further it is equivalent to prove $g'(x) = \frac{AL_0}{(AL_0 + \sqrt{2}BL_1 x)} \leq d$ when $x \geq 0$, which is obviously true. Overall, we have approximation error $|\delta_k| \leq \xi \rho^2 d^{3/2}$.

Then, we continue our analysis. Suppose we do take step size $\alpha_k = \frac{|\gamma_k|}{(AL_0 + \sqrt{2}BL_1|\gamma_k|)d}$ and substitute $\alpha_k = \alpha_k^{opt} + \delta_k$, then inequality (13) can be re-formulate as

$$\begin{aligned}\theta_{k+1} &\leq \theta_k - (\alpha_k^{opt} + \delta_k)|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)| + (\alpha_k^{opt} + \delta_k)^2 \frac{AL_0 + \sqrt{2}BL_1|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)|}{2} d \\ &= \theta_k - \frac{||\mathbf{s}^T \nabla f(\mathbf{x}_k)||^2}{(AL_0 + \sqrt{2}BL_1|\mathbf{s}^T \nabla f(\mathbf{x}_k)|)d} - |\mathbf{s}^T \nabla f(\mathbf{x}_k)|\delta_k + \frac{||\mathbf{s}^T \nabla f(\mathbf{x}_k)||^2}{2(AL_0 + \sqrt{2}BL_1|\mathbf{s}^T \nabla f(\mathbf{x}_k)|)d} \\ &\quad + \frac{AL_0 + \sqrt{2}BL_1|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)|}{2} d\delta_k^2 + |\mathbf{s}^T \nabla f(\mathbf{x}_k)|\delta_k \\ &\leq \theta_k - \frac{||\mathbf{s}^T \nabla f(\mathbf{x}_k)||^2}{2(AL_0 + \sqrt{2}BL_1|\mathbf{s}^T \nabla f(\mathbf{x}_k)|)d} + \frac{(AL_0 + \sqrt{2}BL_1|\mathbf{s}_k^T \nabla f(\mathbf{x}_k)|)d}{2} \delta_k^2 \\ &\leq \theta_k - \frac{||\nabla f(\mathbf{x}_k)||^2}{4(AL_0 + \sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||)d} + \frac{(AL_0 + \sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||)d}{2} \delta_k^2 \quad \text{Apply Lemma 2} \end{aligned} \tag{14}$$

Condition 1

Suppose $1 - \sqrt{2}BL_1 \geq 0$ and $||\nabla f(\mathbf{x}_k)|| \geq AL_0 + \sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||$, inequality (14) can be reformulated as

$$\theta_{k+1} \leq \theta_k - \frac{||\nabla f(\mathbf{x}_k)||}{4d} + \frac{||\nabla f(\mathbf{x}_k)||d}{2} \delta_k^2$$

Meanwhile, suppose $|\delta_k| \leq \xi \rho^2 d^{3/2} \leq \frac{1}{2d}$, we have

$$||\nabla f(\mathbf{x}_k)|| \leq 8d(\theta_k - \theta_{k+1})$$

$$\sum_{k=0}^{K} ||\nabla f(\mathbf{x}_k)|| \leq 8d(\theta_0 - \theta_{k+1})$$

We can conclude that there exists an iteration $j \sim [0, K]$ such that

$$||\nabla f(\mathbf{x}_j)|| \leq \frac{8d(\theta_0 - \theta_{k+1})}{K}$$

$$||\nabla f(\mathbf{x}_j)|| \leq \frac{8d(f(\mathbf{x}_0) - f^\star)}{K}$$

which concludes that we need

$$K \geq \frac{8d(f(\mathbf{x}_0) - f^\star)}{\epsilon} \tag{15}$$

iterations to reach ϵ-first-order stationary point.

Condition 2

Suppose $1 - \sqrt{2}BL_1 \geq 0$ and $||\nabla f(\mathbf{x}_k)|| \leq AL_0 + \sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||$, we have $||\nabla f(\mathbf{x}_k)|| \leq \frac{AL_0}{1-\sqrt{2}BL_1}$. Meanwhile, suppose $|\delta_k| \leq \xi\rho^2 d^{3/2} \leq \frac{||\nabla f(\mathbf{x}_k)||}{2(AL_0+\sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||)d}$, then inequality (14) can be reformulated as

$$\theta_{k+1} \leq \theta_k - \frac{||\nabla f(\mathbf{x}_k)||^2}{8(AL_0 + \sqrt{2}BL_1\frac{AL_0}{1-\sqrt{2}BL_1})d}$$

$$||\nabla f(\mathbf{x}_k)||^2 \leq (\theta_k - \theta_{k+1})\frac{8AL_0d}{1-\sqrt{2}BL_1}$$

$$\sum_{k=0}^{K} ||\nabla f(\mathbf{x}_k)||^2 \leq (\theta_0 - \theta_{k+1})\frac{8AL_0d}{1-\sqrt{2}BL_1}$$

We can conclude that there exists an iteration $j \sim [0, K]$ such that

$$||\nabla f(\mathbf{x}_j)||^2 \leq \frac{8AL_0d(\theta_0 - \theta_{k+1})}{(1-\sqrt{2}BL_1)K}$$

$$||\nabla f(\mathbf{x}_j)|| \leq \sqrt{\frac{8AL_0d(f(\mathbf{x}_0) - f^\star)}{(1-\sqrt{2}BL_1)K}},$$

which concludes that we need

$$K \geq \frac{8AL_0d(f(\mathbf{x}_0) - f^\star)}{(1-\sqrt{2}BL_1)\epsilon^2}$$

iterations to reach ϵ-first-order stationary point.

Condition 3

Suppose $1 - \sqrt{2}BL_1 \leq 0$ and $||\nabla f(\mathbf{x}_k)||^2 \leq (\frac{AL_0}{1-\sqrt{2}BL_1})^2$. Meanwhile, suppose $|\delta_k| \leq \xi\rho^2 d^{3/2} \leq \frac{||\nabla f(\mathbf{x}_k)||}{2(AL_0+\sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||)d}$, then inequality (14) can be reformulated as

$$\theta_{k+1} \leq \theta_k - \frac{||\nabla f(\mathbf{x}_k)||^2}{8(AL_0 + \sqrt{2}BL_1|\frac{AL_0}{1-\sqrt{2}BL_1}|)d}$$

$$||\nabla f(\mathbf{x}_k)||^2 \leq (\theta_k - \theta_{k+1})\frac{8AL_0d(2\sqrt{2}BL_1 - 1)}{\sqrt{2}BL_1 - 1}$$

$$\sum_{k=0}^{K} ||\nabla f(\mathbf{x}_k)||^2 \leq (\theta_0 - \theta_{k+1})\frac{8AL_0d(2\sqrt{2}BL_1 - 1)}{\sqrt{2}BL_1 - 1}$$

We can conclude that there exists an iteration $j \sim [0, K]$ such that

$$||\nabla f(\mathbf{x}_j)||^2 \leq \frac{8AL_0d(\theta_0 - \theta_{k+1})(2\sqrt{2}BL_1 - 1)}{(\sqrt{2}BL_1 - 1)K}$$

$$||\nabla f(\mathbf{x}_j)|| \leq \sqrt{\frac{8AL_0d(f(\mathbf{x}_0) - f^\star)(2\sqrt{2}BL_1 - 1)}{(\sqrt{2}BL_1 - 1)K}},$$

which concludes that we need

$$K \geq \frac{8AL_0 d(f(\mathbf{x}_0) - f^\star)(2\sqrt{2}BL_1 - 1)}{(\sqrt{2}BL_1 - 1)\epsilon^2} \tag{16}$$

iterations to reach ϵ-first-order stationary point.

Condition 4

Suppose $1 - \sqrt{2}BL_1 \leq 0$ and $||\nabla f(\mathbf{x}_k)||^2 \geq (\frac{AL_0}{1-\sqrt{2}BL_1})^2$. Meanwhile, suppose $\delta_k \leq \xi\rho^2 d^{3/2} \leq \frac{||\nabla f(\mathbf{x}_k)||}{2(AL_0+\sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||)d}$, then inequality (14) can be reformulated as

$$\theta_{k+1} \leq \theta_k - \frac{(\frac{AL_0}{1-\sqrt{2}BL_1})^2}{8(AL_0 + \sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||)d} \tag{17}$$

Since $\frac{(\frac{AL_0}{1-\sqrt{2}BL_1})^2}{8(AL_0+\sqrt{2}BL_1||\nabla f(\mathbf{x}_k)||)d}$ is a monotone decreasing function w.r.t. $||\nabla f(\mathbf{x}_k)||$, then we can conclude that the loss function cannot be indicator of reaching ϵ-first-order stationary points. However, with an appropriate selection of parameters, the loss function can be minimized. I.e.,

$$\begin{aligned}
\theta_{k+1} &\leq \theta_k - \frac{(\frac{AL_0}{\sqrt{2}BL_1-1})^2}{8(AL_0 + \sqrt{2}BL_1\frac{AL_0}{\sqrt{2}BL_1-1})d} \\
\theta_{k+1} &\leq \theta_k - \frac{AL_0}{8(2\sqrt{2}BL_1 - 1)(\sqrt{2}BL_1 - 1)d} \\
\theta_{k+1} &\leq \theta_0 - (K+1)\frac{AL_0}{8(2\sqrt{2}BL_1 - 1)(\sqrt{2}BL_1 - 1)d} \\
f(\mathbf{x}_k) - f^\star &\leq f(\mathbf{x}_0) - f^\star - K\frac{AL_0}{8(2\sqrt{2}BL_1 - 1)(\sqrt{2}BL_1 - 1)d},
\end{aligned}$$

which concludes that we need

$$K \geq \frac{8(2\sqrt{2}BL_1 - 1)(\sqrt{2}BL_1 - 1)(f(\mathbf{x}_0) - f^\star - \epsilon)d}{AL_0}$$

iterations to reach local ϵ-optimal point.

We summarize the results over all conditions in Table 2.

Conditions[b]	requirement over ρ[a]	Query complexity
$L_1 \leq \frac{1}{\sqrt{2}B}, \lvert\lvert\nabla f(\mathbf{x})\rvert\rvert \geq \frac{AL_0}{1-\sqrt{2}BL_1}$	$\rho \leq \frac{1}{d\sqrt{2\xi\sqrt{d}}}$	$\frac{8d(f(\mathbf{x}_0)-f^\star)}{\epsilon}$
$L_1 \leq \frac{1}{\sqrt{2}B}, \lvert\lvert\nabla f(\mathbf{x})\rvert\rvert \leq \frac{AL_0}{1-\sqrt{2}BL_1}$	$\rho \leq \frac{1}{d}\sqrt{\frac{\epsilon}{2\xi(AL_0+\sqrt{2}BL_1\epsilon)\sqrt{d}}}$	$\frac{8AL_0d(f(\mathbf{x}_0)-f^\star)}{(1-\sqrt{2}BL_1)\epsilon^2}$
$L_1 \geq \frac{1}{\sqrt{2}B}, \lvert\lvert\nabla f(\mathbf{x})\rvert\rvert \leq \frac{AL_0}{\sqrt{2}BL_1-1}$	$\rho \leq \frac{1}{d}\sqrt{\frac{\epsilon}{2\xi(AL_0+\sqrt{2}BL_1\epsilon)\sqrt{d}}}$	$\frac{8AL_0d(f(\mathbf{x}_0)-f^\star)(2\sqrt{2}BL_1-1)}{(\sqrt{2}BL_1-1)\epsilon^2}$
$L_1 \geq \frac{1}{\sqrt{2}B}, \lvert\lvert\nabla f(\mathbf{x})\rvert\rvert \geq \frac{AL_0}{\sqrt{2}BL_1-1}$	$\rho \leq \frac{1}{d}\sqrt{\frac{\epsilon}{2\xi(AL_0+\sqrt{2}BL_1\epsilon)\sqrt{d}}}$	$\frac{8(2\sqrt{2}BL_1-1)(\sqrt{2}BL_1-1)(f(\mathbf{x}_0)-f^\star-\epsilon)d}{AL_0}$

[a] ξ is a constant associated with third-order property of f, detailed in appendix inequality (13).
[b] For forth condition, reaching local ϵ-optimal point instead of ϵ-first-order stationary point, detailed in appendix inequality (17).

Table 2: With dynamic step size strategy, the convergence property of f under relaxed smoothness.

Note that for the specific value of A and B, we have $A = 1+e^c - \frac{e^c-1}{c}$, $B = \frac{e^c-1}{c}$ and $||\mathbf{x}_{k+1} - \mathbf{x}_k|| = ||\alpha\mathbf{s}_k|| = \frac{\gamma_k}{(AL_0+\sqrt{2}BL_1\gamma_k)\sqrt{d}} \leq \frac{c}{L_1} \to c \geq \frac{1}{B\sqrt{2d}} \to e^c \geq 1 + \frac{1}{\sqrt{2d}}$. It is easy to see that such c exists, we can safely consider $A = 1.01, B = 1.01$ for simplicity (under large d) since A and B are expected to be small values. □

C.3 Bound of gradient norm of S2P

Theorem C.2. (Theorem 3.3) Suppose objective function $f(\cdot)$ satisfies Assumption 2. Then the gradient norm $||\nabla f(\mathbf{x}_k)||$ can be bounded in expectation as

$$|\gamma| - \rho d(AL_0 + BL_1||\nabla f(\mathbf{x})||) \leq ||\nabla f(\mathbf{x})|| \leq \sqrt{2}|\gamma| + \sqrt{2}\rho d(AL_0 + BL_1||\nabla f(\mathbf{x})||)$$

where $|\gamma| = \frac{|f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}+\rho\mathbf{s})|}{2\rho}$. Constants $A = 1.01, B = 1.01$ when $\rho \leq \frac{0.001}{2L_1\sqrt{d}}$

Proof.

$$\begin{aligned}
||\nabla f(\mathbf{x})|| &\leq \mathbb{E}[\sqrt{2}|\mathbf{s}^T\nabla f(x)|] = \mathbb{E}[\frac{1}{\sqrt{2}\rho}|2\rho\mathbf{s}^T\nabla f(x)|] &(18)\\
&= \mathbb{E}[\frac{1}{\sqrt{2}\rho}|\big(f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s})\big) - \big(f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s}) - 2\rho\mathbf{s}^T\nabla f(x)\big)|]\\
&\leq \mathbb{E}[\sqrt{2}\frac{|f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s})|}{2\rho} + \frac{1}{\sqrt{2}\rho}|f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s}) - 2\rho\mathbf{s}^T\nabla f(x)\big)|]\\
&= \sqrt{2}|\gamma| + \frac{1}{\sqrt{2}\rho}\mathbb{E}[|f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s}) - 2\rho\mathbf{s}^T\nabla f(x)\big)|]\\
&\leq \sqrt{2}|\gamma| + \frac{1}{\sqrt{2}\rho}\frac{AL_0 + BL_1||\nabla f(\mathbf{x})||}{2}\mathbb{E}[||2\rho\mathbf{s}||^2] &(19)\\
&= \sqrt{2}|\gamma| + \sqrt{2}\rho d(AL_0 + BL_1||\nabla f(\mathbf{x})||).
\end{aligned}$$

Note inequality (18) applies Lemma A.2, inequality (19) applies Lemma A.1. And the same with the following proof.

$$\begin{aligned}
||\nabla f(\mathbf{x})|| &\geq \mathbb{E}[|\mathbf{s}^T\nabla f(x)|] = \mathbb{E}[\frac{1}{2\rho}|2\rho\mathbf{s}^T\nabla f(x)|]\\
&= \mathbb{E}[\frac{1}{2\rho}|\big(f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s})\big) - \big(f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s}) - 2\rho\mathbf{s}^T\nabla f(x)\big)|]\\
&\geq \mathbb{E}[\frac{|f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s})|}{2\rho} - \frac{1}{2\rho}|f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s}) - 2\rho\mathbf{s}^T\nabla f(x)\big)|]\\
&= |\gamma| - \frac{1}{2\rho}\mathbb{E}[|f(\mathbf{x}+\rho\mathbf{s}) - f(\mathbf{x}-\rho\mathbf{s}) - 2\rho\mathbf{s}^T\nabla f(x)\big)|]\\
&\geq |\gamma| - \frac{1}{2\rho}\frac{AL_0 + BL_1||\nabla f(\mathbf{x})||}{2}\mathbb{E}[||2\rho\mathbf{s}||^2]\\
&\geq |\gamma| - \rho d(AL_0 + BL_1||\nabla f(\mathbf{x})||).
\end{aligned}$$

Note that for the specific value of A and B, we have $A = 1 + e^c - \frac{e^c - 1}{c}$, $B = \frac{e^c - 1}{c}$ and $||\mathbf{x}_{k+1} - \mathbf{x}_k|| = ||(\mathbf{x}+\rho\mathbf{s}) - (\mathbf{x}-\rho\mathbf{s})|| = ||2\rho\mathbf{s}|| = 2\rho\sqrt{d} \leq \frac{c}{L_1} \rightarrow c \geq 2\rho L_1\sqrt{d}$. It is easy to see that such c exists, we can safely consider $\rho \leq \frac{1}{2L_1 d}$, then we have $c \geq \frac{1}{\sqrt{d}}$. It is easy to see such c exists, we set $A = 1.01, B = 1.01$ for simplicity. □

D Experiments

D.1 Setup

For experiment over common deep models and datasets, we do grid search for initial learning rate α_0 over list $\{$2e-4, 1e-4, 8e-5, 5e-5, 2e-5, 1e-5$\}$ and for smoothing parameter ρ_0 over list $\{$1e-3, 5e-4, 1e-4, 5e-5, 1e-5$\}$ with all methods. We average the results across 5 random seeds.

Note the selected hyper-parameters directly apply to sign variants. The tunable hyper-parameters are summarized in Table 3.

Hyper-parameter	Arc.&Dataset	Method		
		GA	AS2P	STP
α_0	ResNet18&CIFAR10	2.0e-5	-	2.0e-4
	ResNet50&CIFAR10	1.0e-5	-	2.0e-4
	ResNet101&CIFAR100	2.0e-5	-	1.0e-4
	ResNet152&CIFAR100	2.0e-5	-	1.0e-4
LR scheduler	All	Cosine decay		
ρ_0	ResNet18&CIFAR10	1e-3	1e-3	-
	ResNet50&CIFAR10	1e-3	5e-4	-
	ResNet101&CIFAR100	5e-4	5e-4	-
	ResNet152&CIFAR100	5e-4	5e-4	-
ρ_{end}	All	-	$\rho_0/10$	-
η_a	All	-	5	-
η_b	ResNet18&CIFAR10	-	5	-
	ResNet50&CIFAR10	-	5	-
	ResNet101&CIFAR100	-	3	-
	ResNet152&CIFAR100	-	5	-
Std Dev(γ_{recent})	All	-	10%	-

Table 3: Summary of hyper-parameters used in experiments over common deep models and datasets. It shows that AS2P has extra hyper-parameters ρ_{end}, η_a, η_b, and Std Dev(γ_{recent}). Basically, those hyper-parameters are unnecessary to tune within above deep models and datasets.

For the experiment over LLM, the six text tasks follow the original settings exactly (Malladi et al., 2023), which randomly samples 1,000 examples and 500 examples for training and validation respectively for each task. We get the results with a fixed random seed. Specifically, for the learning rate and smoothing parameter, we apply the best values mentioned in Malladi et al. (2023) for GA. Then, AS2P directly applies the value of smoothing parameter ρ_0 from GA and only needs to tune one hyper-parameter η_b. For STP method, we search the best α_0 from list $\{$5e-5, 2e-5, 1e-5, 5e-6 ,1e-6, 1e-7$\}$. The details of hyper-parameters are summarized in Table 4, which shows that only η_b is necessary to update among all four extra hyper-parameters ρ_{end}, η_a, η_b, and Std Dev(γ_{recent}) of AS2P compared with experiments about common deep models&datasets.

Hyper-parameter	Task	Method			
		GA	GA constant	AS2P	STP
α_0	SST-2 RTE Copa ReCoRD SQuAD DROP	1e-7	1e-7	-	2e-5
LR scheduler	All	Cosine decay	Constant value	Cosine decay	Cosine decay
ρ_0	All	1e-3	1e-3	1e-3	-
ρ_{end}	All	-	-	$\rho_0/10$	-
η_a	All	-	-	5	-
η_b	All	-	-	50	-
Std Dev(γ_{recent})	All	-	-	10%	-

Table 4: Summary of hyper-parameters used in experiments over LLM. Basically, AS2P needs to tune η_b, and the selected values are robust across varying tasks.

D.2 Additional Experiments

Table version of Figure 1(a) and Figure 2(a). The base of training cost ratio, e.g., {1, 0.8, 0.6, 0.4, 0.2}, normalizes the number of function queries when base method GA reaches {500, 400, 300, 200, 100} epochs with some specific loss values. Then, the training cost ratio aligns with the ratio between the number of function queries of the base method and other methods reaching the same loss values.

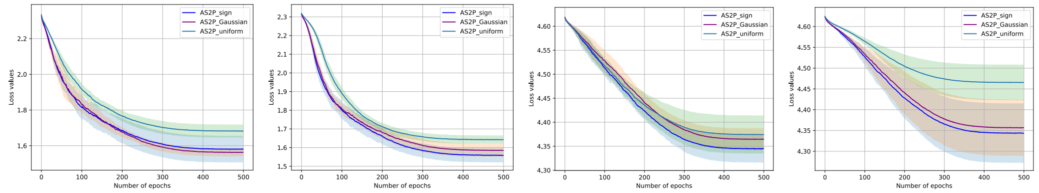
Task	Method	Training cost ratio				
ResNet18&CIFAR10	GA	1	0.80	0.60	0.40	0.20
	STP	1	1	0.80	0.43	0.17
	AS2P	0.56	0.52	0.39	0.24	0.10
ResNet50&CIFAR10	GA	1	0.80	0.60	0.40	0.20
	STP	0.98	0.87	0.54	0.14	0.05
	AS2P	0.40	0.36	0.27	0.14	0.05
ResNet101&CIFAR100	GA	1	0.80	0.60	0.40	0.20
	STP	0.93	0.85	0.61	0.37	0.14
	AS2P	0.39	0.36	0.28	0.17	0.07
ResNet152&CIFAR100	GA	1	0.80	0.60	0.40	0.20
	STP	0.63	0.59	0.45	0.25	0.10
	AS2P	0.32	0.31	0.24	0.14	0.06

Table 5: Training cost ratio of reaching specific loss values under common deep models&datasets.

Task	Method	Training cost ratio			
SST-2	GA cosine decay LR	1	0.75	0.50	0.25
	GA constant LR	0.47	0.50	0.38	0.23
	STP cosine decay LR	0.38	0.38	0.30	0.19
	AS2P cosine decay LR	0.17	0.17	0.15	0.10
RTE	GA cosine decay LR	1	0.75	0.50	0.25
	GA constant LR	0.35	0.35	0.30	0.12
	STP cosine decay LR	0.03	0.03	0.03	0.03
	AS2P cosine decay LR	0.03	0.03	0.03	0.03
Copa	GA cosine decay LR	1	0.75	0.50	0.25
	GA constant LR	0.47	0.47	0.40	0.23
	STP cosine decay LR	0.26	0.26	0.23	0.11
	AS2P cosine decay LR	0.10	0.10	0.07	0.05
ReCoRD	GA cosine decay LR	1	0.75	0.50	0.25
	GA constant LR	0.42	0.42	0.42	0.23
	STP cosine decay LR	0.23	0.23	0.19	0.07
	AS2P cosine decay LR	0.05	0.05	0.05	0.03
SQuAD	GA cosine decay LR	1	0.75	0.50	0.25
	GA constant LR	0.65	0.65	0.53	0.30
	STP cosine decay LR	0.23	0.23	0.23	0.15
	AS2P cosine decay LR	0.15	0.15	0.12	0.07
DROP	GA cosine decay LR	1	0.75	0.50	0.25
	GA constant LR	0.42	0.47	0.40	0.23
	STP cosine decay LR	0.34	0.34	0.30	0.19
	AS2P cosine decay LR	0.15	0.17	0.15	0.10

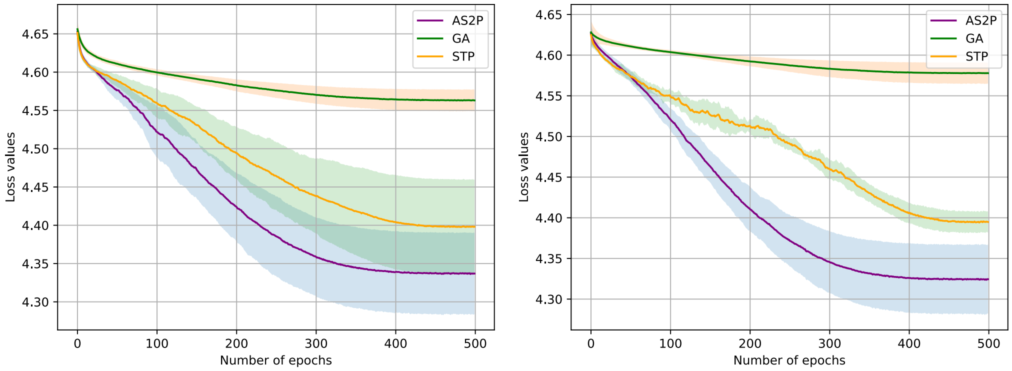
Table 6: Training cost ratio of reaching specific loss values when fully fine-tuning OPT-13B model under various tasks.

Figure 4: Performance comparison between applying different noise distributions such as Normal distribution, Rademacher distribution, and Uniform distribution.



(a) Under pre-trained ResNet18&CIFAR10 (b) Under pre-trained ResNet50&CIFAR10 (c) Under pre-trained ResNet101&CIFAR100 (d) Under pre-trained ResNet152&CIFAR100

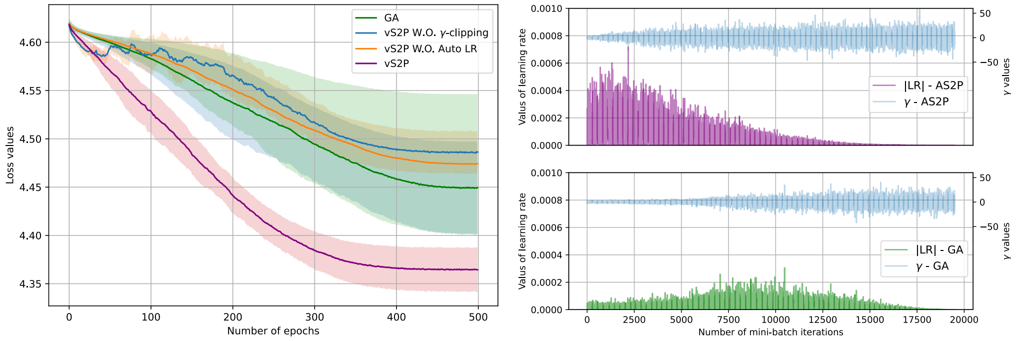
Figure 5: Convergence rate of pre-trained ResNet18&CIFAR100 and pre-trained ResNet50&CIFAR100.



(a) pre-trained ResNet18&CIFAR100

(b) pre-trained ResNet50&CIFAR100

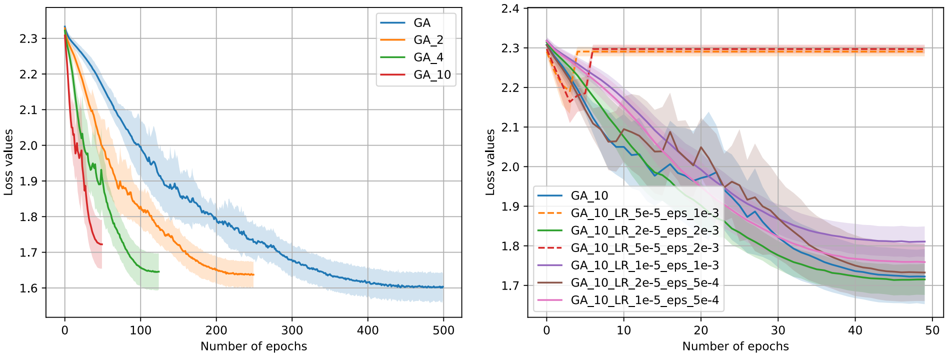
Figure 6: Verification of effectiveness of proposed method under pre-trained ResNet101&CIFAR100. Left-side figure demonstrated the convergence rate of AS2P without (W.O.) automatic learning rate and without progressive γ-clipping. Right-side two figures demonstrate the dynamics of learning rate and γ;



(a) Convergence rate comparison

(b) Learning rate comparison

Figure 7: Under ResNet18 and CIFAR10, the performance of GA with the different number of symmetric perturbations for each update. The left-side figure shows performance under the varying number of symmetric random perturbation per update where the number of function query for each setting are the same. The right-side figure demonstrates that under varying training settings, the convergence of GA with 10 symmetric random perturbations for gradient approximation per update. Basically, we can conclude that one symmetric symmetric random perturbation per update converges to smaller loss values under the same number of function queries.



(a) Same initial LR and LR scheduler

(b) Varying α and ρ under GA_10.

Figure 8: Performance comparison under VGG11 and CIFAR10. The left-side figure demonstrates the dynamics of training loss; The right-side figure demonstrates the training cost ratio of reaching the same specific loss values. The proposed method AS2P converges faster than other baseline methods and nearly requires 0.5× number of queries to reach the same specific loss values. Note that the hyper-parameters directly follow the setting of ReSNet18&CIFAR10 in Table 3.

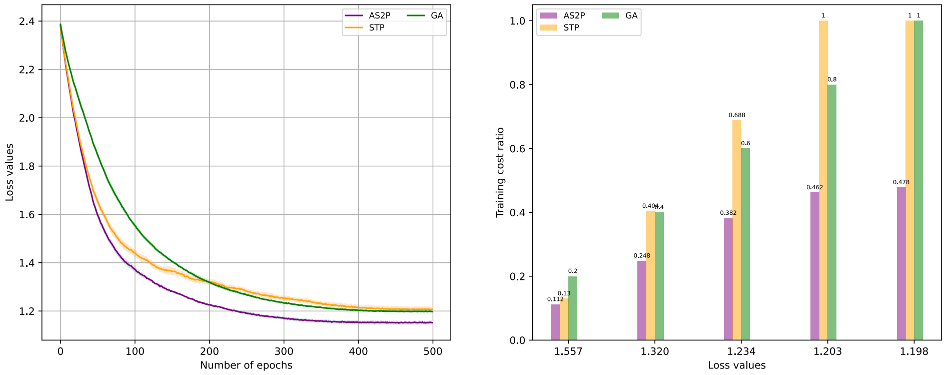
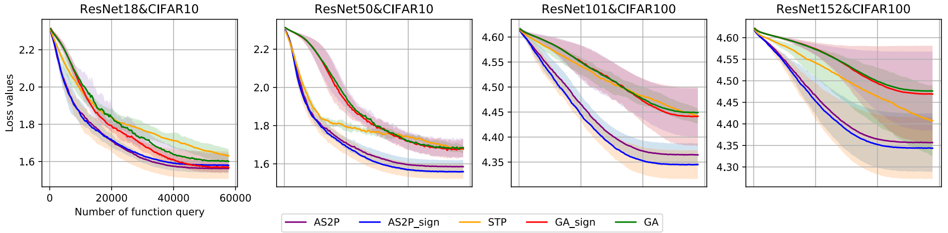
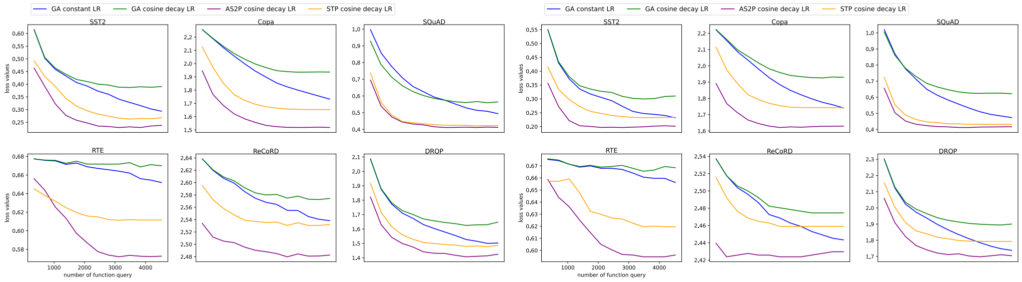


Figure 9: Performance comparison with various baselines under common deep models&datasets where the x-axis is the number of function queries. This figure is adopted from Figure 1.





(b) Training loss

(c) Evaluation loss

Figure 10: Performance comparison with full fine-tuning OPT-13B model where the x-axis is the number of function queries. This figure is adopted from Figure 2.

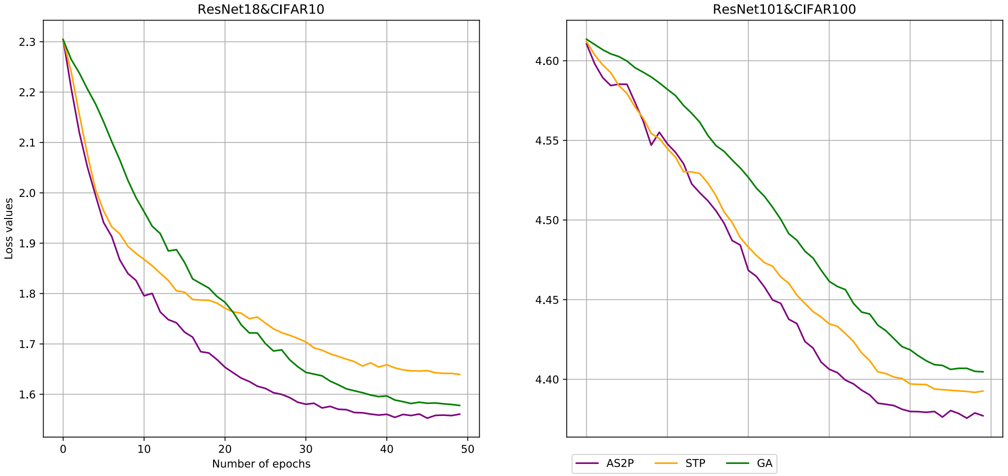


Figure 11: Corresponding validation performance of Figure 1(b) under setting ResNet18&CIFAR10 and ResNet101&CIFAR100. Using one seed only.