Supplementary Materials

A. *Efficient Implementation Procedures of Adaptive Sampling*

Algorithm 2 Updating Tree $\mathcal{T}$

Input: old tree $\mathcal{T}$, new value for i-th leaf π_i'
Compute $\Delta = \pi_i' - \pi_i$
$node = leaf(i)$
while $node.parent \neq$ NULL **do**
  $node = node + \Delta$
  $node = node.parent$
end while
$node = node + \Delta$
Output: new tree $\mathcal{T}'$

Algorithm 3 Sampling Based on Tree $\mathcal{T}$

Input: $\mathcal{T}$, $R \in \left[0, \frac{1}{\alpha_k}\sum_{j=1}^n \pi_j\right]$, $C = \frac{1-\alpha_k}{n\alpha_k}\sum_{j=1}^n \pi_j$
$node = root(\mathcal{T})$
while $node$ is not a leaf **do**
  if $R > node.sum_L + node.left.num \times C$ **then**
    $R = R - node.sum_L - node.left.num \times C$
    $node = node.right$
  else
    $node = node.left$
  end if
end while
Output: $node.ind$

B. *Discussion on the Assumptions*

The following proposition shows that Assumption 3 holds with high probability for large enough N.

Proposition 1. *Suppose that $N \geq 4n\log n/(1-\overline{\alpha})$ and there are K iterations in total, then Assumption 3 holds for all K iterations with probability at least $1 - \frac{\lceil 2K/N \rceil}{n}$.*

Proof. Firstly, in the first $N/2$ iterations, for any $1 \leq j \leq n$, j has been picked with probability at least

$$1 - \left(1 - \min_{1\leq k\leq N/2}\{p_j^k\}\right)^{N/2} \geq 1 - \left(1 - \frac{1-\overline{\alpha}}{n}\right)^{2n\log n/(1-\overline{\alpha})} \geq 1 - \frac{1}{n^2}.$$

Thus, in the first $N/2$ iterates, all indices have been picked at least once with probability at least $1 - \frac{1}{n}$. Furthermore, we know that, for iterations between $(k-1)N/2+1$ and $kN/2$ for each $1 \leq \lceil 2K/N \rceil$, all indices have been picked at least once with probability at least $1 - \frac{\lceil 2K/N \rceil}{n}$. Once this holds, since every N iterations must contain at least one interval $[(k-1)N/2+1, kN/2]$ for some $1 \leq \lceil 2K/N \rceil$, each index has been picked at least once, i.e., Assumption 3 holds. □

C. *Useful Lemmas*

The stochastic gradient at certain iterate $\mathbf{w}$ in SGD-AIS is $\frac{1}{np_i}\nabla f_i(\mathbf{w})$, where i follows the most recently updated distribution $\boldsymbol{p}$. As discussed for (6) and (7), $\boldsymbol{p}$ is a mixture of the sub-optimal distribution and the uniform distribution. To prove our desired result, we introduce an auxiliary distribution $\boldsymbol{p}^{\mathbf{w}}$, which is a mixture of the optimal distribution and the uniform distribution. More specifically,

$$p_i^{\mathbf{w}} = \alpha\frac{\|\nabla f_i(\mathbf{w})\|_2}{\sum_{j=1}^m \|\nabla f_j(\mathbf{w})\|_2} + (1-\alpha)\frac{1}{n},\ \forall i \in [n]. \tag{26}$$

Accordingly, an intermediate stochastic gradient is defined as $\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})$, where $i \sim \boldsymbol{p}^{\mathbf{w}}$. We first prove that the variance of this intermediate stochastic gradient $\mathrm{Var}_{i\sim\boldsymbol{p}^{\mathbf{w}}}\left[\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right]$ is strictly smaller than the variance of uniform distribution $\mathrm{Var}_{i\sim\mathcal{U}}[\nabla f_i(\mathbf{w})]$, which is formally stated as Lemma 1.

Lemma 1. *Denote $\mathcal{U}$ as the uniform distribution on $[n]$, and $\boldsymbol{p}^{\mathbf{w}}$ is the distribution defined as (26). If Assumption 2 holds, then for all $\alpha \in [\underline{\alpha}, \overline{\alpha}]$, we have*

$$\mathrm{Var}_{i\sim\mathcal{U}}[\nabla f_i(\mathbf{w})] - \mathrm{Var}_{i\sim\boldsymbol{p}^{\mathbf{w}}}\left[\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right] \geq \alpha\rho G^2. \tag{27}$$

Proof. Since both $\nabla f_i(\mathbf{w})$ and $\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})$ are unbiased estimator of $\nabla F(\mathbf{w})$, we have

$$\mathrm{Var}_{i\sim\mathcal{U}}[\nabla f_i(\mathbf{w})] = \mathbb{E}[\|\nabla f_i(\mathbf{w})\|_2^2] - \|\mathbb{E}[\nabla f_i(\mathbf{w})]\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{w})\|_2^2 - \|\nabla F(\mathbf{w})\|_2^2,$$

and

$$\mathrm{Var}_{i\sim\boldsymbol{p}^{\mathbf{w}}}\left[\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right] = \mathbb{E}\left[\left\|\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right\|_2^2\right] - \left\|\mathbb{E}\left[\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right]\right\|_2^2 = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{p_i^{\mathbf{w}}}\|\nabla f_i(\mathbf{w})\|_2^2 - \|\nabla F(\mathbf{w})\|_2^2.$$

By definition of $p_i^{\mathbf{w}}$ and the fact that $(ax+by)(a/x+b/y) \geq (a+b)^2$ for all x, y, a, $b > 0$, we have

$$\frac{1}{p_i^{\mathbf{w}}} = \frac{1}{\alpha\frac{\|\nabla f_i(\mathbf{w})\|_2}{\sum_{j=1}^{m}\|\nabla f_j(\mathbf{w})\|_2} + (1-\alpha)\frac{1}{n}} \leq \alpha\frac{\sum_{j=1}^{n}\|\nabla f_j(\mathbf{w})\|_2}{\|\nabla f_i(\mathbf{w})\|_2} + (1-\alpha)n,$$

holds for any $\alpha \in [\underline{\alpha}, \overline{\alpha}]$. Therefore,

$$\begin{aligned}
&\mathrm{Var}_{i\sim\mathcal{U}}[\nabla f_i(\mathbf{w})] - \mathrm{Var}_{i\sim\boldsymbol{p}^{\mathbf{w}}}\left[\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right] \\
=&\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{w})\|_2^2 - \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{p_i^{\mathbf{w}}}\|\nabla f_i(\mathbf{w})\|_2^2 \\
\geq&\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{w})\|_2^2 - \frac{1}{n^2}\sum_{i=1}^{n}\left(\alpha\frac{\sum_{j=1}^{n}\|\nabla f_j(\mathbf{w})\|_2}{\|\nabla f_i(\mathbf{w})\|_2} + (1-\alpha)n\right)\|\nabla f_i(\mathbf{w})\|_2^2 \\
=&\frac{\alpha}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{w})\|_2^2 - \frac{\alpha}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\nabla f_i(\mathbf{w})\|_2\|\nabla f_j(\mathbf{w})\|_2 \\
=&\frac{\alpha}{2n^2}\left(\sum_{i=1}^{n}2n\|\nabla f_i(\mathbf{w})\|_2^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}2\|\nabla f_i(\mathbf{w})\|_2\|\nabla f_j(\mathbf{w})\|_2\right) \\
=&\frac{\alpha}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\|\nabla f_i(\mathbf{w})\|_2 - \|\nabla f_j(\mathbf{w})\|_2)^2 \\
\geq&\alpha\rho G^2,
\end{aligned} \tag{28}$$

where the last inequality follows from Assumption 2. ∎

Next, we would like to bound the difference between $\mathrm{Var}_{i\sim\boldsymbol{p}}\left[\frac{1}{np_i}\nabla f_i(\mathbf{w})\right]$ and the intermediate variance $\mathrm{Var}_{i\sim\boldsymbol{p}^{\mathbf{w}}}\left[\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right]$. Lemma 2 plays a key role to achieve this.

Lemma 2. *Consider the k-th iteration. Denote $\tau_j = \max\{k' : k' \leq k, i_{k'} = j\}$ for all $j \in [n]$. Let $\alpha_k \in (\underline{\alpha}, \overline{\alpha})$ be in Algorithm 1. $\boldsymbol{p}$ is the most recently updated probability distribution in Algorithm 1, i.e.,*

$$p_i = \alpha_k\left(\frac{\|\nabla f_i(\mathbf{w}_{\tau_i})\|_2}{\sum_{j=1}^{n}\|\nabla f_j(\mathbf{w}_{\tau_j})\|_2}\right) + (1-\alpha_k)\frac{1}{n}. \tag{29}$$

$p_i^{\mathbf{w}_k}$ is defined as the right hand side of equation (26), i.e.,

$$p_i^{\mathbf{w}_k} = \alpha_k\left(\frac{\|\nabla f_i(\mathbf{w}_k)\|_2}{\sum_{j=1}^{n}\|\nabla f_j(\mathbf{w}_k)\|_2}\right) + (1-\alpha_k)\frac{1}{n}. \tag{30}$$

By (16) and Assumption 3, as well as $\eta := \max\{\eta_k : k \in \mathbb{N}\} \leq (1-\overline{\alpha})\delta/NL$, we have

$$\sum_{i=1}^{n} |p_i - p_i^{\mathbf{w}_k}| \leq \frac{2\overline{\alpha}L\eta m}{(1-\overline{\alpha})\delta - L\eta m}. \tag{31}$$

Proof. For $j \in [n]$, we first consider the difference of the following gradient norms,

$$\begin{aligned}
|\|\nabla f_j(\mathbf{w}_{\tau_j})\|_2 - \|\nabla f_j(\mathbf{w}_k)\|_2| &\leq \|\nabla f_j(\mathbf{w}_{\tau_j}) - \nabla f_j(\mathbf{w}_k)\|_2 \\
&\leq L\|\mathbf{w}_{\tau_j} - \mathbf{w}_k\|_2 \\
&= L\left\| \sum_{\kappa=\tau_j}^{k} \eta_\kappa \frac{1}{np_{i_\kappa}} \nabla f_{i_\kappa}(\mathbf{w}_\kappa) \right\|_2 \\
&\leq L\eta \sum_{\kappa=\tau_j}^{k} \left\| \frac{1}{np_{i_\kappa}} f_{i_\kappa}(\mathbf{w}_\kappa) \right\|_2 \\
&\leq L\eta \sum_{\kappa=\tau_j}^{k} \frac{G}{1-\overline{\alpha}} \\
&= \frac{GL\eta}{1-\overline{\alpha}}(k - \tau_j + 1) \\
&\leq \frac{GL\eta N}{1-\overline{\alpha}}.
\end{aligned} \tag{32}$$

The fourth inequality in (32) is bacause (29) implies $p_{i_\kappa} > (1-\overline{\alpha})/n$, and (16) implies $\|\nabla f_{i_\kappa}(\mathbf{w}_\kappa)\|_2 \leq G$. The last inequality in (32) is because Assumption 3 implies that $k + 1 - \tau_j \leq N$. (32) further implies that, for all $j \in [n]$

$$\|\nabla f_j(\mathbf{w}_k)\|_2 - \frac{GL\eta N}{1-\overline{\alpha}} \leq \|\nabla f_j(\mathbf{w}_{\tau_j})\|_2 \leq \|\nabla f_j(\mathbf{w}_k)\|_2 + \frac{GL\eta m N}{1-\overline{\alpha}}. \tag{33}$$

Thus,

$$\frac{\|\nabla f_i(\mathbf{w}_k)\|_2 - \frac{GL\eta N}{1-\overline{\alpha}}}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 + \frac{GL\eta Nn}{1-\overline{\alpha}}} \leq \frac{\|\nabla f_i(\mathbf{w}_{\tau_i})\|_2}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_{\tau_j})\|_2} \leq \frac{\|\nabla f_i(\mathbf{w}_k)\|_2 + \frac{GL\eta N}{1-\overline{\alpha}}}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 - \frac{GL\eta Nn}{1-\overline{\alpha}}}, \tag{34}$$

where the second inequality is ensured to be positive by (16) and $\eta < (1-\overline{\alpha})\delta/NL$. (34) implies that at least one of the following two inequalities hold, i.e.,

$$\begin{aligned}
|p_i - p_i^{\mathbf{w}_k}| &\leq \overline{\alpha} \left| \frac{\|\nabla f_i(\mathbf{w}_k)\|_2 + \frac{GL\eta N}{1-\overline{\alpha}}}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 - \frac{GL\eta Nn}{1-\overline{\alpha}}} - \frac{\|\nabla f_i(\mathbf{w}_k)\|_2}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2} \right| \\
&= \overline{\alpha} \frac{\frac{GL\eta N}{1-\overline{\alpha}} \sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 + \frac{GL\eta Nn}{1-\overline{\alpha}} \|\nabla f_i(\mathbf{w}_k)\|_2}{\left(\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 - \frac{GL\eta Nn}{1-\overline{\alpha}}\right) \sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2} \\
&:= A,
\end{aligned} \tag{35}$$

or

$$\begin{aligned}
|p_i - p_i^{\mathbf{w}_k}| &\leq \overline{\alpha} \left| \frac{\|\nabla f_i(\mathbf{w}_k)\|_2 - \frac{GL\eta N}{1-\overline{\alpha}}}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 + \frac{GL\eta N}{1-\overline{\alpha}}} - \frac{\|\nabla f_i(\mathbf{w}_k)\|_2}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2} \right| \\
&= \overline{\alpha} \frac{\frac{GL\eta N}{1-\overline{\alpha}} \sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 + \frac{GL\eta Nn}{1-\overline{\alpha}} \|\nabla f_i(\mathbf{w}_k)\|_2}{\left(\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 + \frac{GL\eta Nn}{1-\overline{\alpha}}\right) \sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2} \\
&:= B.
\end{aligned} \tag{36}$$

It is obvious that $A \geq B$, thus inequality (35) always holds. Taking $i = 1, 2, \dots, n$ in (35) and summing the n inequalities, this yields

$$\begin{aligned}
\sum_{i=1}^{n} |p_i - p_i^{\mathbf{w}_k}| &\leq \overline{\alpha} \frac{\frac{GL\eta Nn}{1-\overline{\alpha}} \sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 + \frac{GL\eta Nn}{1-\overline{\alpha}} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{w}_k)\|_2}{\left(\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 - \frac{GL\eta Nn}{1-\overline{\alpha}}\right) \sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2} \\
&= \overline{\alpha} \frac{\frac{GL\eta Nn}{1-\overline{\alpha}}}{\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2 - \frac{GL\eta Nn}{1-\overline{\alpha}}} \\
&= \frac{\overline{\alpha} GL\eta Nn}{(1-\overline{\alpha})\left(\sum_{j=1}^{n} \|\nabla f_j(\mathbf{w}_k)\|_2\right) - GL\eta Nn} \\
&\leq \frac{\overline{\alpha} GL\eta Nn}{(1-\overline{\alpha}) n\delta G - GL\eta Nn} \\
&= \frac{\overline{\alpha} L\eta N}{(1-\overline{\alpha})\delta - L\eta N}
\end{aligned} \tag{37}$$

where the second inequality comes from (16). Note that $\eta < \frac{(1-\overline{\alpha})\delta}{NL}$, thus the upper bound in (37) is positive. □

D. ***Proof of Theorem 1***

Proof. We first consider the following bound

$$\begin{aligned}
\left| \mathrm{Var}_{i\sim \boldsymbol{p}} \left[\frac{1}{np_i} \nabla f_i(\mathbf{w})\right] - \mathrm{Var}_{i\sim \boldsymbol{p}^{\mathbf{w}}} \left[\frac{1}{np_i^{\mathbf{w}}} \nabla f_i(\mathbf{w})\right] \right| &= \left| \frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{1}{p_i} - \frac{1}{p_i^{\mathbf{w}}}\right) \|\nabla f_i(\mathbf{w})\|_2^2 \right| \\
&\leq \frac{G^2}{n^2} \sum_{i=1}^{n} \left| \frac{1}{p_i} - \frac{1}{p_i^{\mathbf{w}}} \right| \\
&= \frac{G^2}{n^2} \sum_{i=1}^{n} \frac{|p_i^{\mathbf{w}} - p_i|}{p_i p_i^{\mathbf{w}}} \\
&\leq \frac{G^2}{n^2} \left(\frac{n}{1-\overline{\alpha}}\right)^2 \sum_{i=1}^{n} |p_i^{\mathbf{w}} - p_i| \\
&\leq \frac{G^2}{(1-\overline{\alpha})^2} \frac{\overline{\alpha} L\eta N}{(1-\overline{\alpha})\delta - L\eta N} \\
&= \frac{\overline{\alpha} G^2 L\eta N}{(1-\overline{\alpha})^3\delta - (1-\overline{\alpha})^2 L\eta N},
\end{aligned} \tag{38}$$

where the last inequality follows from Lemma 2, and the final obtained bound in is positive since $\eta < \frac{(1-\overline{\alpha})^3\delta\rho}{(1-\overline{\alpha})^2 NL\rho + NL} < \frac{(1-\overline{\alpha})\delta}{NL}$. Therefore,

$$\begin{aligned}
\mathrm{Var}_{i\sim \boldsymbol{p}} \left[\frac{1}{np_i} \nabla f_i(\mathbf{w})\right] &\leq \mathrm{Var}_{i\sim \boldsymbol{p}^{\mathbf{w}}} \left[\frac{1}{np_i^{\mathbf{w}}} \nabla f_i(\mathbf{w})\right] + \frac{\overline{\alpha} G^2 L\eta N}{(1-\overline{\alpha})^3\delta - (1-\overline{\alpha})^2 L\eta N} \\
&\leq \mathrm{Var}_{i\sim \mathcal{U}}[\nabla f_i(\mathbf{w})] - \underline{\alpha}\rho G^2 + \frac{\overline{\alpha} G^2 L\eta N}{(1-\overline{\alpha})^3\delta - (1-\overline{\alpha})^2 L\eta N} \\
&= \mathrm{Var}_{i\sim \mathcal{U}}[\nabla f_i(\mathbf{w})] - \left(\underline{\alpha}\rho - \frac{\overline{\alpha} L\eta N}{(1-\overline{\alpha})^3\delta - (1-\overline{\alpha})^2 L\eta N}\right) G^2 \\
&= \mathrm{Var}_{i\sim \mathcal{U}}[\nabla f_i(\mathbf{w})] - \gamma G^2,
\end{aligned} \tag{39}$$

where the second inequality results from Lemma 1. In addition, $\gamma = \underline{\alpha}\rho - \frac{\overline{\alpha} L\eta N}{(1-\overline{\alpha})^3\delta - (1-\overline{\alpha})^2 L\eta N} > 0$ since $\eta < \frac{(1-\overline{\alpha})^3\delta\rho}{(1-\overline{\alpha})^2 NL\rho + NL}$, and $\gamma < 1$ since $\underline{\alpha}, \rho < 1$ □

E. ***Proofs of Theorems 2 & 3***

Prepared with the above two lemmas, we can finally connect our desired variances $\mathrm{Var}_{i\sim \boldsymbol{p}}\left[\frac{1}{np_i}\nabla f_i(\mathbf{w})\right]$ and $\mathrm{Var}_{i\sim \mathcal{U}}[\nabla f_i(\mathbf{w})]$ by bridging over the intermediate variance $\mathrm{Var}_{i\sim \boldsymbol{p}^{\mathbf{w}}}\left[\frac{1}{np_i^{\mathbf{w}}}\nabla f_i(\mathbf{w})\right]$.

Proof of Theorem 2. For all $k \in \mathbb{N}$, conditioning on $\mathbf{w}_k$, along with (41), we have

$$\mathbb{E}_{i\sim \boldsymbol{p}}[F(\mathbf{w}_{k+1})] - F(\mathbf{w}_k) \le -2\eta\sigma(F(\mathbf{w}_k) - F^*) + \frac{\eta^2 L}{2}(1-\gamma)G^2.$$

Subtracting F^* from both sides, taking total expectation, and rearranging, this yields

$$\mathbb{E}[F(\mathbf{w}_{k+1}) - F^*] \le (1-2\eta\sigma)\mathbb{E}[F(\mathbf{w}_k) - F^*] + \frac{\eta^2 L}{2}(1-\gamma)G^2.$$

Applying this inequality repeatedly through iteration $k \in \mathbb{N}$ to get

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_k) - F^*] \le&(1-2\eta\sigma)^{k-1}(F(\mathbf{w}_1) - F^*) + \frac{\eta^2 L}{2}(1-\gamma)G^2\sum_{l=1}^{k}(1-2\eta\sigma)^{l-1} \\ \le&(1-2\eta\sigma)^{k-1}(F(\mathbf{w}_1) - F^*) + \frac{\eta L}{4\sigma}(1-\gamma)G^2 \\ \xrightarrow{k\to\infty}&\frac{\eta L(1-\gamma)G^2}{4\sigma},\end{aligned} \tag{40}$$

where the last limit comes from $1-2\eta\sigma < 1$, which is implied by

$$\eta < \frac{(1-\overline{\alpha})^3\underline{\alpha}\delta\rho}{(1-\overline{\alpha})^2\underline{\alpha}NL\rho + \overline{\alpha}NL} < \frac{1}{2L} < \frac{1}{2\sigma},$$

since $\delta, \rho \le 1$. □

Proof of Theorem 3. By (23) and the definition of η_k, the following inequality holds for all $k \in \mathbb{N}$,

$$\eta_k < \frac{(1-\overline{\alpha})^3\underline{\alpha}\delta\rho}{(1-\overline{\alpha})^2\underline{\alpha}NL\rho + \overline{\alpha}NL}.$$

For all $k \in \mathbb{N}$, conditioning on $\mathbf{w}_k$, we have

$$\begin{aligned}\mathbb{E}_{i\sim \boldsymbol{p}}[F(\mathbf{w}_{k+1})] - F(\mathbf{w}_k) \le& -\eta_k \mathbb{E}_{i\sim \boldsymbol{p}}\left[\left\langle \frac{1}{np_i}\nabla f_i(\mathbf{w}), \nabla F(\mathbf{w}_k)\right\rangle\right] + \frac{\eta_k^2 L}{2}\mathbb{E}_{i\sim \boldsymbol{p}}\left[\left\|\frac{1}{np_i}\nabla f_i(\mathbf{w})\right\|_2^2\right] \\ \le& -\eta_k\|\nabla F(\mathbf{w}_k)\|_F^2 + \frac{\eta_k^2 L}{2}\left(\mathbb{E}_{i\sim\mathcal{U}}\left[\|\nabla f_i(\mathbf{w})\|_2^2\right] - \gamma G^2\right) \\ \le& -\eta_k\|\nabla F(\mathbf{w}_k)\|_F^2 + \frac{\eta_k^2 L}{2}(1-\gamma)G^2 \\ \le& -2\eta_k\sigma(F(\mathbf{w}_k) - F^*) + \frac{\eta_k^2 L}{2}(1-\gamma)G^2,\end{aligned} \tag{41}$$

where $\boldsymbol{p}$ denotes the most recently updated sampling distribution in SGD-AIS. In (41), the first inequality is implied by the L-smoothness of F, the second inequality follows from Theorem 1, the third inequality is due to (16), and the last inequality are come from strong convexity of F. Subtracting F^* from both sides, taking total expectation, and rearranging, this yields

$$\mathbb{E}[F(\mathbf{w}_{k+1}) - F^*] \le (1-2\eta_k\sigma)\mathbb{E}[F(\mathbf{w}_k) - F^*] + \frac{\eta_k^2 L}{2}(1-\gamma)G^2.$$

Subtracting F^* from both sides, taking the expectation and rearranging, this yields

$$\mathbb{E}[F(\mathbf{w}_{k+1}) - F^*] \le (1-2\eta_k\sigma)\mathbb{E}[F(\mathbf{w}_k) - F^*] + \frac{\eta_k^2 L}{2}(1-\gamma)G^2. \tag{42}$$

Then we prove $\mathbb{E}[F(\mathbf{w}_k) - F^*] \le \nu/(\xi + k)$ by induction. Firstly, the definition of ν ensures that it holds for $k = 1$. Then, assume it holds for some $k > 1$, it follows from (42) that

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_{k+1}) - F^*] \le&(1-\frac{2\sigma\beta}{\xi+k})\frac{\nu}{\xi+k} + \frac{\beta^2 L(1-\gamma)G^2}{2(\xi+k)^2} \\ =&\frac{\xi+k-1}{(\xi+k)^2}\nu - \frac{2(2\sigma\beta-1)\nu - \beta^2 L(1-\gamma)G^2}{2(\xi+k)^2} \\ \le&\frac{\nu}{\xi+k+1}.\end{aligned} \tag{43}$$

The last inequality holds because of $(\xi+k-1)(\xi+k+1) < (\xi+k)^2$ and the definition of ν. □

F. *Supplementary Convergence Analysis*

Theorem 1 holds without requiring the convexity of the objective function $F(\mathbf{w})$, thus we can get the convergence results of SGD-AIS for the nonconvex cases, which are formally stated as the following two theorems.

Theorem 4. *Under Assumptions 1-3, suppose that the objective function $F(\mathbf{w})$ is a L-smooth function, and the SGD-AIS is run with a fixed stepsize, $\eta_k = \eta$ for all $k \in \mathbb{N}$, satisfying*

$$0 < \eta < \frac{(1-\overline{\alpha})^3\underline{\alpha}\delta\rho}{(1-\overline{\alpha})^2\underline{\alpha}NL\rho + \overline{\alpha}NL}.$$

Then, the average-squared gradient of F corresponding to the iterates satisfy

$$\begin{aligned}\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{w}_k)\|_2^2\right] \leq & \frac{\eta L}{2}(1-\gamma)G^2 + \frac{F(\mathbf{w}_k) - F_{\inf}}{K\eta} \\ \xrightarrow{K\to\infty} & \frac{\eta L}{2}(1-\gamma)G^2.\end{aligned} \tag{44}$$

Proof. Taking the total expectation of (41) yields

$$\mathbb{E}[F(\mathbf{w}_{k+1})] - \mathbb{E}[F(\mathbf{w}_k)] \leq -\eta\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|_F^2] + \frac{\eta^2 L}{2}(1-\gamma)G^2.$$

Summing both sides of this inequality for $1 \leq k \leq K$ and dividing by K gives

$$\frac{\mathbb{E}[F(\mathbf{w}_{K+1})] - F(\mathbf{w}_1)}{K} \leq -\eta\sum_{k=1}^{K}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|_F^2] + \frac{\eta^2 L}{2}(1-\gamma)G^2.$$

To get (44), we only need to use the inequality $\mathbb{E}[F(\mathbf{w}_{K+1})] \geq F_{\inf}$. □

Theorem 5. *Under Assumptions 1-3, suppose that the objective function $F(\mathbf{w})$ is L-smooth, and SGD-AIS is run with a diminishing stepsize sequence that satisfies, for all $k \in \mathbb{N}$,*

$$0 < \eta_k < \frac{(1-\overline{\alpha})^3\underline{\alpha}\delta\rho}{(1-\overline{\alpha})^2\underline{\alpha}NL\rho + \overline{\alpha}NL}, \tag{45}$$

and

$$A_K = \sum_{k=1}^{K}\eta_k = \infty, \quad \textit{and} \quad B_K = \sum_{k=1}^{K}\eta_k^2 < \infty. \tag{46}$$

Then, the average-squared gradient of F corresponding to the SGD iterates satisfy

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K}\eta_k\|\nabla F(\mathbf{w}_k)\|_2^2\right] \\ \leq & \frac{L(1-\gamma)G^2 B_K}{A_K} + \frac{2\left(F(\mathbf{w}_k) - F_{\inf}\right)}{A_K} \xrightarrow{K\to\infty} 0.\end{aligned} \tag{47}$$

Proof. Similarly, taking the total expectation of (41) yields

$$\mathbb{E}[F(\mathbf{w}_{k+1})] - \mathbb{E}[F(\mathbf{w}_k)] \leq -\eta_k\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|_F^2] + \frac{\eta_k^2 L}{2}(1-\gamma)G^2.$$

Summing both sides of this inequality for $1 \leq k \leq K$ and dividing by A_K gives

$$\frac{\mathbb{E}[F(\mathbf{w}_{K+1})] - F(\mathbf{w}_1)}{A_K} \leq -\mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K}\eta_k\|\nabla F(\mathbf{w}_k)\|_F^2\right] + \frac{\eta^2 L(1-\gamma)G^2 B_K}{2A_K}.$$

Use the inequality $\mathbb{E}[F(\mathbf{w}_{K+1})] \geq F_{\inf}$, we can easily get the first inequality of (47), while the limitation holds because of (46). □

G. CNN Architecture Used in the Experiments (printed in PyTorch format)

```
Net(
(conv1): Conv2d(3, 6, kernel-size=(5, 5), stride=(1, 1))
(pool): MaxPool2d(kernel-size=2, stride=2, padding=0, dilation=1, ceil-mode=False)
(conv2): Conv2d(6, 16, kernel-size=(5, 5), stride=(1, 1))
(fc1): Linear(in-features=400, out-features=120, bias=True)
(fc2): Linear(in-features=120, out-features=84, bias=True)
(fc3): Linear(in-features=84, out-features=10, bias=True)
)
```

H. Dataset Sizes and Algorithmic Parameters

In our experiments, we adopt diminishing stepsizes $\eta_k = \frac{\beta}{\xi+k}$ for SGD-based algorithms and constant stepsize η for SGDm/ADAM-based algorithms. The sizes of the real datasets and specific choices of the parameters are given in the following tables.

TABLE III: Sizes of Datasets

	a2a	ijcnn1	w8a	gisette
n	2265	49990	49749	6000
d	123	22	300	5000

TABLE IV: Parameters of SGD-based Algorithms for Logistic Regression

	a2a	w8a	ijcnn1	gisette
Stepsize Parameter β	1100	200	100	100
Stepsize Parameter ξ	7000	100000	6000	20000
Regularization Parameter λ	0.01	0.01	0.01	0.01

TABLE V: Parameters of SGD-based Algorithms for SVM

	a2a	w8a	ijcnn1	gisette
Stepsize Parameter β	300	100	1100	50
Stepsize Parameter ξ	7000	100000	6000	500000
Regularization Parameter λ	0.01	0.01	0.01	0.01

TABLE VI: Parameters of SGDm-based Algorithms for SVM

	a2a	w8a	ijcnn1	gisette
Constant Stepsize η	0.001	0.0002	0.0001	0.0001
Regularization Parameter λ	0.01	0.01	0.01	0.01

TABLE VII: Parameters of ADAM-based Algorithms for SVM

	a2a	w8a	ijcnn1	gisette
Constant Stepsize η	0.005	0.0005	0.0005	0.0008
Regularization Parameter λ	0.01	0.01	0.01	0.01

TABLE VIII: Parameters of SGDm-based Algorithms for Neural Networks

	MLP (MINIST)	LeNet (MINIST)	CNN (Cifar-10)
Mini-batch Size	8	16	16
Stepsize η	0.001	0.001	0.001
Learning Rate Decay (per 100 steps) ρ	0.999	0.995	0.99
Regularization Parameter λ	0.01	0.01	0.01

TABLE IX: Parameters of ADAM-based Algorithms for Neural Networks

	MLP (MINIST)	LeNet (MINIST)	CNN (Cifar-10)
Mini-batch Size	8	16	16
Stepsize η	0.00003	0.001	0.001
Learning Rate Decay (per 100 steps) ρ	0.999	0.995	0.99
Regularization Parameter λ	0.01	0.01	0.01