Appendix

A Useful Lemmas and Facts

Lemma 7. *[Nesterov, 2004, Theorem 2.1.5]. If f is convex and has L-Lipschitz gradient, then the following inequalities are true*

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{9a}$$

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2L}\big\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\big\|^2 \tag{9b}$$

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L}\big\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\big\|^2. \tag{9c}$$

Note that inequality (9a) does not require the convexity of f.

Lemma 8. *[Nesterov, 2004]. If f is μ-strongly convex and has L-Lipschitz gradient, with $\mathbf{x}^* := \arg\min_{\mathbf{x}} f(\mathbf{x})$, the following inequalities are true*

$$2\mu\big(f(\mathbf{x}) - f(\mathbf{x}^*)\big) \leq \|\nabla f(\mathbf{x})\|^2 \leq 2L\big(f(\mathbf{x}) - f(\mathbf{x}^*)\big) \tag{10a}$$

$$\mu\|\mathbf{x} - \mathbf{x}^*\| \leq \|\nabla f(\mathbf{x})\| \leq L\|\mathbf{x} - \mathbf{x}^*\| \tag{10b}$$

$$\frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}^*\|^2 \tag{10c}$$

$$\big\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \big\rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2. \tag{10d}$$

Proof. By definition $f(\mathbf{x}^*) - f(\mathbf{x}) \geq \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|$, minimizing over $\mathbf{x} - \mathbf{x}^*$ on the RHS results in (10a). Inequality (10b) follows from [Nesterov, 2004, Theorem 2.1.9] and the fact $\nabla f(\mathbf{x}^*) = 0$. Inequality (10c) is from [Nesterov, 2004, Theorem 2.1.7]; and, (10d) is from [Nesterov, 2004, Theorem 2.1.9] □

Proof of Lemma 3:

Proof. If $t_1 \neq t_2$, $N_{t_1:t}$ and $N_{t_2:t}$ are disjoint by definition, since the most recent calculated snapshot gradient can only appear at either t_1 or t_2. Since $\{B_t\}$ are i.i.d., one can find the probability of $N_{t_1:t}$ as

$$\mathbb{P}(N_{t_1:t}) = \begin{cases} \frac{1}{m}\big(1 - \frac{1}{m}\big)^{t-t_1} & \text{if } 1 \leq t_1 \leq t \\ \big(1 - \frac{1}{m}\big)^t & \text{if } t_1 = 0. \end{cases} \tag{11}$$

Hence one can verify that

$$\sum_{t_1=0}^{t}\mathbb{P}(N_{t_1:t}) = \Big(1 - \frac{1}{m}\Big)^t + \sum_{t_1=1}^{t-1}\frac{1}{m}\big(1 - \frac{1}{m}\big)^{t-t_1} + \frac{1}{m}\Big(1 - \frac{1}{m}\Big)^t + \frac{1}{m}\frac{1 - \frac{1}{m} - (1 - \frac{1}{m})^t}{1 - (1 - \frac{1}{m})} + \frac{1}{m} = 1$$

which completes the proof. □

B Technical Proofs in Section 3.1

B.1 Proof of Lemma 4

The following lemmas are needed for the proof.

Lemma 9. *The following equation is true for $t > t_1$*

$$\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | N_{t_1:t}\big] = \sum_{\tau=t_1+1}^{t}\mathbb{E}\big[\|\mathbf{v}_\tau - \mathbf{v}_{\tau-1}\|^2 | N_{t_1:t}\big] - \sum_{\tau=t_1+1}^{t}\mathbb{E}\big[\|\nabla F(\mathbf{x}_\tau) - \nabla F(\mathbf{x}_{\tau-1})\|^2 | N_{t_1:t}\big].$$

Proof. Consider that

$$
\begin{aligned}
&\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] \\
=&\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}) + \nabla F(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1} + \mathbf{v}_{t-1} - \mathbf{v}_t\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] \\
=&\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2 + \mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] + \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2 \\
&+ 2\big\langle \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}), \nabla F(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1} \big\rangle \\
&+ 2\mathbb{E}\big[\big\langle \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1}), \mathbf{v}_{t-1} - \mathbf{v}_t \big\rangle | \mathcal{F}_{t-1}, N_{t_1:t}\big] \\
&+ 2\mathbb{E}\big[\big\langle \nabla F(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}, \mathbf{v}_{t-1} - \mathbf{v}_t \big\rangle | \mathcal{F}_{t-1}, N_{t_1:t}\big] \\
=&\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] - \|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})\|^2 + \|\nabla F(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2 \qquad (12)
\end{aligned}
$$

where the last equation is because $\mathbb{E}[\mathbf{v}_t - \mathbf{v}_{t-1} | \mathcal{F}_{t-1}, N_{t_1:t}] = \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}_{t-1})$. We can expand $\mathbb{E}[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2 | \mathcal{F}_{t-2}, N_{t_1:t}]$ using the same argument. Note that we have $\nabla F(\mathbf{x}_{t_1}) = \mathbf{v}_{t_1}$, which suggests

$$
\mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_1+1}) - \mathbf{v}_{t_1+1}\|^2 | \mathcal{F}_{t_1}, N_{t_1:t}\big] = \mathbb{E}\big[\|\mathbf{v}_{t_1+1} - \mathbf{v}_{t_1}\|^2 | \mathcal{F}_{t_1}, N_{t_1:t}\big] - \|\nabla F(\mathbf{x}_{t_1+1}) - \nabla F(\mathbf{x}_{t_1})\|^2.
$$

Then taking expectation w.r.t. $\mathcal{F}_{t-1}$ and expanding $\mathbb{E}[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{v}_{t-1}\|^2]$ in (12), the proof is completed. $\square$

Proof of Lemma 4: The implication of this Lemma 3 is that *law of total probability* [Gubner, 2006] holds. Specifically, for a random variable C_t that happens in iteration t, the following equation holds

$$
\mathbb{E}\big[C_t\big] = \sum_{t_1=0}^{t} \mathbb{E}\big[C_t | N_{t_1:t}\big] \mathbb{P}\{N_{t_1:t}\}. \qquad (13)
$$

Now we turn to prove Lemma 4. To start with, consider that when $t_1 \neq t$

$$
\begin{aligned}
&\mathbb{E}\big[\|\mathbf{v}_t\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] = \mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1} + \mathbf{v}_{t-1}\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] \\
&= \|\mathbf{v}_{t-1}\|^2 + \mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] + 2\mathbb{E}\big[\langle \mathbf{v}_{t-1}, \mathbf{v}_t - \mathbf{v}_{t-1} \rangle | \mathcal{F}_{t-1}, N_{t_1:t}\big] \\
&\overset{(a)}{=} \|\mathbf{v}_{t-1}\|^2 + \mathbb{E}\Big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 + \frac{2}{\eta}\Big\langle \mathbf{x}_{t-1} - \mathbf{x}_t, \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_{t-1}) \Big\rangle \Big| \mathcal{F}_{t-1}, N_{t_1:t}\Big] \\
&\overset{(b)}{\leq} \|\mathbf{v}_{t-1}\|^2 + \mathbb{E}\Big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 - \frac{2}{\eta L}\|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_{t-1})\|^2 \Big| \mathcal{F}_{t-1}, N_{t_1:t}\Big] \\
&= \|\mathbf{v}_{t-1}\|^2 + \mathbb{E}\Big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 - \frac{2}{\eta L}\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 \Big| \mathcal{F}_{t-1}, N_{t_1:t}\Big] \\
&= \|\mathbf{v}_{t-1}\|^2 + \mathbb{E}\Big[\Big(1 - \frac{2}{\eta L}\Big)\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 \Big| \mathcal{F}_{t-1}, N_{t_1:t}\Big]
\end{aligned}
$$

where (a) follows from (2) and the update $\mathbf{x}_t = \mathbf{x}_{t-1} - \eta \mathbf{v}_{t-1}$; and (b) is the result of (9c). Then by choosing η such that $1 - \frac{2}{\eta L} < 0$, i.e., $\eta < 2/L$, we have

$$
\mathbb{E}\big[\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big] \leq \frac{\eta L}{2 - \eta L}\Big(\|\mathbf{v}_{t-1}\|^2 - \mathbb{E}\big[\|\mathbf{v}_t\|^2 | \mathcal{F}_{t-1}, N_{t_1:t}\big]\Big). \qquad (14)
$$

Plugging (14) into Lemma 9, we have

$$
\begin{aligned}
\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | \mathcal{F}_{t_1-1}, N_{t_1:t}\big] &\leq \sum_{\tau=t_1+1}^{t} \mathbb{E}\big[\|\mathbf{v}_\tau - \mathbf{v}_{\tau-1}\|^2 | \mathcal{F}_{t_1-1}, N_{t_1:t}\big] \\
&= \frac{\eta L}{2 - \eta L}\mathbb{E}\big[\|\mathbf{v}_{t_1}\|^2 | \mathcal{F}_{t_1-1}, N_{t_1:t}\big] = \frac{\eta L}{2 - \eta L}\|\nabla F(\mathbf{x}_{t_1})\|^2
\end{aligned}
$$

where the last equation is because conditioning on $N_{t_1:t}$, $\mathbf{v}_{t_1} = \nabla F(\mathbf{x}_{t_1})$. Note that when $t_1 = t$, this inequality automatically holds since the LHS equals to 0. Because the randomness of $\nabla F(\mathbf{x}_{t_1})$ is irrelevant to B_{t_1} (thus $N_{t_1:t}$), after taking expectation w.r.t. $\mathcal{F}_{t_1-1}$, we have

$$
\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | N_{t_1:t}\big] \leq \frac{\eta L}{2 - \eta L}\mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_1})\|^2 | N_{t_1:t}\big] = \frac{\eta L}{2 - \eta L}\mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_1})\|^2\big]
$$

which proves the first part of Lemma 4.

For the second part of Lemma 4, by calculating the probability of $N_{t_1:t}$ as in (11), we have

$$\begin{aligned}
\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2\big] &\overset{(c)}{=} \sum_{t_1=0}^{t-1} \mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | N_{t_1:t}\big] \mathbb{P}\{N_{t_1:t}\} \\
&\leq \sum_{t_1=0}^{t-1} \frac{\eta L}{2-\eta L} \mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_1})\|^2\big] \mathbb{P}\{N_{t_1:t}\} \\
&= \frac{\eta L}{2-\eta L}\left[\frac{1}{m}\sum_{\tau=1}^{t-1}\Big(1-\frac{1}{m}\Big)^{t-\tau}\mathbb{E}\big[\|\nabla F(\mathbf{x}_\tau)\|^2\big] + \Big(1-\frac{1}{m}\Big)^{t}\|\nabla F(\mathbf{x}_0)\|^2\right]
\end{aligned}$$

where (c) uses (13), and $\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | N_{t:t}\big] = 0$. The proof is thus completed.

B.2 Proof of Theorem 1

Following Assumption 1, we have

$$\begin{aligned}
F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t) &\leq \big\langle \nabla F(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \big\rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= -\eta\big\langle \nabla F(\mathbf{x}_t), \mathbf{v}_t\big\rangle + \frac{\eta^2 L}{2}\|\mathbf{v}_t\|^2 \\
&= -\frac{\eta}{2}\Big[\|\nabla F(\mathbf{x}_t)\|^2 + \|\mathbf{v}_t\|^2 - \|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2\Big] + \frac{\eta^2 L}{2}\|\mathbf{v}_t\|^2
\end{aligned} \tag{15}$$

where the last equation is because $\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}[\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}-\mathbf{b}\|^2]$. Rearranging the terms, we arrive at

$$\begin{aligned}
\|\nabla F(\mathbf{x}_t)\|^2 &\leq \frac{2\big[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})\big]}{\eta} + \|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 - \big(1-\eta L\big)\|\mathbf{v}_t\|^2 \\
&\leq \frac{2\big[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})\big]}{\eta} + \|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2
\end{aligned}$$

where the last inequality holds since $\eta < 1/L$. Taking expectation and summing over $t = 1, \dots, T$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t)\|^2\Big] &\leq \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}_{T+1})\big]}{\eta} + \sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2\Big] \\
&\overset{(a)}{\leq} \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}_{T+1})\big]}{\eta} + \frac{\eta L}{2-\eta L}\frac{1}{m}\sum_{t=1}^{T}\sum_{\tau=1}^{t-1}\Big(1-\frac{1}{m}\Big)^{t-\tau}\mathbb{E}\big[\|\nabla F(\mathbf{x}_\tau)\|^2\big] \\
&\qquad + \frac{\eta L}{2-\eta L}\sum_{t=1}^{T}\Big(1-\frac{1}{m}\Big)^{t}\|\nabla F(\mathbf{x}_0)\|^2 \\
&\overset{(b)}{\leq} \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}_{T+1})\big]}{\eta} + \frac{\eta L}{2-\eta L}\frac{1}{m}\sum_{t=1}^{T-1}\left[\sum_{\tau=1}^{T-t}\Big(1-\frac{1}{m}\Big)^{\tau}\right]\mathbb{E}\big[\|\nabla F(\mathbf{x}_t)\|^2\big] \\
&\qquad + \frac{m\eta L}{2-\eta L}\|\nabla F(\mathbf{x}_0)\|^2 \\
&\overset{(c)}{\leq} \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}_{T+1})\big]}{\eta} + \frac{\eta L}{2-\eta L}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla F(\mathbf{x}_t)\|^2\big] + \frac{m\eta L}{2-\eta L}\|\nabla F(\mathbf{x}_0)\|^2
\end{aligned}$$

where (a) is the result of Lemma 4; (b) is by changing the order of summation, and $\sum_{t=1}^{T}(1-\frac{1}{m})^t \leq m$; and, (c) is again by $\sum_{\tau=1}^{T-t}(1-\frac{1}{m})^\tau \leq m$. Rearranging the terms and dividing both sides by T, we have

$$\begin{aligned}\Big(1-\frac{\eta L}{2-\eta L}\Big)\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t)\|^2\Big] &\leq \frac{2\big[F(\mathbf{x}_1)-F(\mathbf{x}_{T+1})\big]}{\eta T}+\frac{\eta L}{2-\eta L}\frac{m}{T}\|\nabla F(\mathbf{x}_0)\|^2 \\ &\leq \frac{2\big[F(\mathbf{x}_1)-F(\mathbf{x}^*)\big]}{\eta T}+\frac{\eta L}{2-\eta L}\frac{m}{T}\|\nabla F(\mathbf{x}_0)\|^2. \end{aligned} \tag{16}$$

Finally, since $\mathbf{v}_0=\nabla F(\mathbf{x}_0)$, we have

$$\begin{aligned} F(\mathbf{x}_1)-F(\mathbf{x}_0) &\leq \big\langle \nabla F(\mathbf{x}_0), \mathbf{x}_1-\mathbf{x}_0\big\rangle + \frac{L}{2}\|\mathbf{x}_1-\mathbf{x}_0\|^2 \\ &= -\eta\|\nabla F(\mathbf{x}_0)\|^2 + \frac{\eta^2 L}{2}\|\nabla F(\mathbf{x}_0)\|^2 \leq 0 \end{aligned} \tag{17}$$

where the last inequality follows from $\eta < 1/L$. Hence we have $F(\mathbf{x}_1) \leq F(\mathbf{x}_0)$, which is applied to (16) to have

$$\Big(1-\frac{\eta L}{2-\eta L}\Big)\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t)\|^2\Big] \leq \frac{2\big[F(\mathbf{x}_0)-F(\mathbf{x}^*)\big]}{\eta T}+\frac{\eta L}{2-\eta L}\frac{m}{T}\|\nabla F(\mathbf{x}_0)\|^2.$$

Now if we choose $\eta < 1/L$ such that $1-\frac{\eta L}{2-\eta L} \geq C_\eta$ with C_η being a positive constant, then we have

$$\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t)\|^2\Big] = \mathcal{O}\bigg(\frac{F(\mathbf{x}_0)-F(\mathbf{x}^*)}{\eta T C_\eta}+\frac{m\eta L\|\nabla F(\mathbf{x}_0)\|^2}{T C_\eta}\bigg).$$

B.3 Proof of Corollaries 1 and 2

From Theorem 1, it is clear that upon choosing $\eta=\mathcal{O}(1/L)$, we have $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big]=\mathcal{O}(m/T)$. This means that $T=\mathcal{O}(m/\epsilon)$ iterations are needed to guarantee $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big]=\epsilon$.

Per iteration requires $\frac{n}{m}+2(1-\frac{1}{m})$ IFO calls in expectation. And n IFO calls are required when computing $\mathbf{v}_0$.

Combining these facts together, we have that $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big]=\mathcal{O}(\sqrt{n}/T)$ if $m=\Theta(\sqrt{n})$. And the IFO complexity is $n+\big[\frac{n}{m}+2(1-\frac{1}{m})\big]T=\mathcal{O}(n+n/\epsilon)$.

Similarly, if $m=\Theta(n)$, we have $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big]=\mathcal{O}(n/T)$. And the IFO complexity in this case becomes $\mathcal{O}(n+n/\epsilon)$.

B.4 Proof of Corollary 3

From Theorem 1, it is clear that with a large m, choosing $\eta=\mathcal{O}(1/\sqrt{m}L)$ leads to $C_\eta \geq 0.5$. Thus we have $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big]=\mathcal{O}(\sqrt{m}/T)$. This translates to the need of $T=\mathcal{O}(\sqrt{m}/\epsilon)$ iterations to guarantee $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big]=\epsilon$.

Choosing $m=\Theta(n)$, we have $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big]=\mathcal{O}(\sqrt{n}/T)$. And the number of IFO calls is $n+\big[\frac{n}{m}+2(1-\frac{1}{m})\big]T=\mathcal{O}(n+\sqrt{n}/\epsilon)$.

C Technical Proofs in Section 3.2

Using the Bernoulli random variable B_t introduced in (4), L2S (Alg. 2) can be rewritten in an equivalent form as Alg. 4.

Algorithm 4 L2S Equivalent Form

Initialize: $\mathbf{x}_0, \eta, m, T$
Compute $\mathbf{v}_0 = \nabla F(\mathbf{x}_0)$
$\mathbf{x}_1 = \mathbf{x}_0 - \eta \mathbf{v}_0$
for $t = 1, 2, \dots, T$ **do**
    Randomly generate B_t: $B_t = 1$ w.p. $\frac{1}{m}$, and $B_t = 0$ w.p. $1 - \frac{1}{m}$
    if $B_t = 1$ **then**,
        $\mathbf{v}_t = \nabla F(\mathbf{x}_t)$
    else
        $\mathbf{v}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_{t-1}) + \mathbf{v}_{t-1}$
    end if
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$
end for
Output: randomly chosen from $\{\mathbf{x}_t\}_{t=1}^T$

Recall that a known $N_{t_1:t}$ is equivalent to $B_{t_1} = 1, B_{t_1+1} = 0, \cdots, B_t = 0$. Now we are ready to prove Lemma 5.

C.1 Proof of Lemma 5

It can be seen that Lemma 9 still holds for nonconvex problems, thus we have

$$\begin{aligned}
\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | N_{t_1:t}\big] &\leq \sum_{\tau=t_1+1}^{t} \mathbb{E}\big[\|\mathbf{v}_\tau - \mathbf{v}_{\tau-1}\|^2 | N_{t_1:t}\big] \\
&= \sum_{\tau=t_1+1}^{t} \mathbb{E}\big[\|\nabla f_{i_\tau}(\mathbf{x}_\tau) - \nabla f_{i_\tau}(\mathbf{x}_{\tau-1})\|^2 | N_{t_1:t}\big] \\
&\leq \eta^2 L^2 \sum_{\tau=t_1+1}^{t} \mathbb{E}\big[\|\mathbf{v}_{\tau-1}\|^2 | N_{t_1:t}\big] = \eta^2 L^2 \sum_{\tau=t_1}^{t-1} \mathbb{E}\big[\|\mathbf{v}_\tau\|^2 | N_{t_1:t}\big] \qquad (18)
\end{aligned}$$

where the last inequality follows from Assumption 1 and $\mathbf{x}_\tau = \mathbf{x}_{\tau-1} - \eta \mathbf{v}_{\tau-1}$. The first part of this lemma is thus proved. Next, we have

$$\begin{aligned}
&\mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2\big] \overset{(a)}{=} \sum_{t_1=0}^{t-1} \mathbb{E}\big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | N_{t_1:t}\big] \mathbb{P}\big\{N_{t_1:t}\big\} \\
&\overset{(b)}{\leq} \eta^2 L^2 \sum_{t_1=0}^{t-1} \sum_{\tau=t_1}^{t-1} \mathbb{E}\big[\|\mathbf{v}_\tau\|^2 | N_{t_1:t}\big] \mathbb{P}\big\{N_{t_1:t}\big\} \overset{(c)}{=} \eta^2 L^2 \sum_{\tau=0}^{t-1} \bigg[\sum_{t_1=0}^{\tau} \mathbb{E}\big[\|\mathbf{v}_\tau\|^2 | N_{t_1:t}\big] \mathbb{P}\big\{N_{t_1:t}\big\}\bigg] \\
&\overset{(d)}{=} \eta^2 L^2 \sum_{\tau=0}^{t-1} \bigg[\mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big] - \sum_{t_1=\tau+1}^{t} \mathbb{E}\big[\|\mathbf{v}_\tau\|^2 | N_{t_1:t}\big] \mathbb{P}\big\{N_{t_1:t}\big\}\bigg] \\
&\overset{(e)}{=} \eta^2 L^2 \sum_{\tau=0}^{t-1} \bigg[\mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big] - \sum_{t_1=\tau+1}^{t} \mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big] \mathbb{P}\big\{N_{t_1:t}\big\}\bigg] \\
&= \eta^2 L^2 \sum_{\tau=0}^{t-1} \bigg[\sum_{t_1=0}^{\tau} \mathbb{P}\big\{N_{t_1:t}\big\}\bigg] \mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big] = \eta^2 L^2 \sum_{\tau=0}^{t-1} \Big(1 - \frac{1}{m}\Big)^{t-\tau} \mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big]
\end{aligned}$$

where (a) is by Lemma 3 (or law of total probability) and $\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 | N_{t:t}] = 0$; (b) is obtained by plugging (18) in; (c) is established by changing the order of summation; (d) is again by Lemma 3 (or law of total probability); and (e) is because of the independence of $\mathbf{v}_\tau$ and $N_{t_1:t}$ when $t_1 > \tau$, that is, $\mathbb{E}[\|\mathbf{v}_\tau\|^2 | N_{t_1:t}] = \mathbb{E}[\|\mathbf{v}_\tau\|^2 | B_{t_1} = 1, B_{t_1+1} = 0, \dots, B_t = 0] = \mathbb{E}[\|\mathbf{v}_\tau\|^2]$. To be more precise, given $t_1 > \tau$, the randomness of $\mathbf{v}_\tau$ comes from $B_1, B_2, \dots B_\tau$ and $i_1, i_2, \cdots, i_\tau$, thus is independent with $B_{t_1}, B_{t_1+1}, \dots, B_t$.

C.2 Proof of Theorem 2

Following the same steps of (15) in Theorem 1, we have

$$\|\nabla F(\mathbf{x}_t)\|^2 \leq \frac{2\big[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})\big]}{\eta} + \|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2 - \big(1 - \eta L\big)\|\mathbf{v}_t\|^2.$$

Taking expectation and summing over t, we have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t)\|^2\Big] &\leq \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}^*)\big]}{\eta} + \sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t) - \mathbf{v}_t\|^2\Big] - \big(1 - \eta L\big)\sum_{t=1}^{T}\mathbb{E}\big[\|\mathbf{v}_t\|^2\big] \\
&\overset{(a)}{\leq} \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}^*)\big]}{\eta} + \eta^2 L^2 \sum_{t=1}^{T}\sum_{\tau=0}^{t-1}\left(1 - \frac{1}{m}\right)^{t-\tau}\mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big] - \big(1 - \eta L\big)\sum_{t=1}^{T}\mathbb{E}\big[\|\mathbf{v}_t\|^2\big] \\
&\overset{(b)}{\leq} \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}^*)\big]}{\eta} + \eta^2 L^2 \sum_{t=1}^{T}\sum_{\tau=0}^{t-1}\left(1 - \frac{1}{m}\right)^{t-\tau}\mathbb{E}\big[\|\mathbf{v}_\tau\|^2\big] - \big(1 - \eta L\big)\sum_{t=1}^{T-1}\mathbb{E}\big[\|\mathbf{v}_t\|^2\big] \\
&\overset{(c)}{\leq} \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}^*)\big]}{\eta} + m\eta^2 L^2 \sum_{t=0}^{T-1}\mathbb{E}\big[\|\mathbf{v}_t\|^2\big] - \big(1 - \eta L\big)\sum_{t=1}^{T-1}\mathbb{E}\big[\|\mathbf{v}_t\|^2\big] \\
&= \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}^*)\big]}{\eta} + m\eta^2 L^2\|\mathbf{v}_0\|^2 + \big(m\eta^2 L^2 + \eta L - 1\big)\sum_{t=1}^{T-1}\mathbb{E}\big[\|\mathbf{v}_t\|^2\big]
\end{aligned} \tag{19}$$

where (a) is by Lemma 5; (b) holds when $1 - \eta L \geq 0$; and (c) is by exchanging the order of summation and $\sum_{t=1}^{T-1}(1 - \frac{1}{m})^t \leq m$. Upon choosing η such that $m\eta^2 L^2 + \eta L - 1 \leq 0$, i.e., $\eta \in (0, \frac{\sqrt{4m+1}-1}{2mL}] = \mathcal{O}\big(\frac{1}{L\sqrt{m}}\big)$, we can eliminate the last term in (19). Plugging m in and dividing both sides by T, we arrive at

$$\begin{aligned}
\mathbb{E}\Big[\|\nabla F(\mathbf{x}_a)\|^2\Big] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla F(\mathbf{x}_t)\|^2\Big] &\leq \frac{2\big[F(\mathbf{x}_1) - F(\mathbf{x}^*)\big]}{\eta T} + \frac{m\eta^2 L^2}{T}\|\nabla F(\mathbf{x}_0)\|^2 \\
&\overset{(d)}{\leq} \frac{2\big[F(\mathbf{x}_0) - F(\mathbf{x}^*)\big]}{\eta T} + \frac{m\eta^2 L^2}{T}\|\nabla F(\mathbf{x}_0)\|^2 \\
&= \mathcal{O}\bigg(\frac{L\sqrt{m}\big[F(\mathbf{x}_0) - F(\mathbf{x}^*)\big]}{T} + \frac{\|\nabla F(\mathbf{x}_0)\|^2}{T}\bigg)
\end{aligned}$$

where (d) is because $F(\mathbf{x}_0) \geq F(\mathbf{x}_1)$ when $\eta \leq 2/L$, which we have already seen from (17). The proof is thus completed.

C.3 Proof of Corollary 5

From Theorem 2, choosing $\eta = \mathcal{O}(1/L\sqrt{m})$, we have $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big] = \mathcal{O}(\sqrt{m}/T)$. This means that $T = \mathcal{O}(\sqrt{m}/\epsilon)$ iterations are required to ensure $\mathbb{E}\big[\|\nabla F(\mathbf{x}_a)\|^2\big] = \epsilon$.

Per iteration it takes in expectation $\frac{n}{m} + 2(1 - \frac{1}{m})$ IFO calls. And n IFO calls are required for computing $\mathbf{v}_0$

Hence choosing $m = \Theta(n)$, the IFO complexity is $n + \big[\frac{n}{m} + 2(1 - \frac{1}{m})\big]T = \mathcal{O}(n + \sqrt{n}/\epsilon)$.

D Technical Proofs in Section 3.3

D.1 Proof of Lemma 6

We borrow the following lemmas from [Nguyen et al., 2017] and summarize them below.

Lemma 10. *[Nguyen et al., 2017, Theorem 1a] Suppose that Assumptions 1 - 3 hold. Choosing step size $\eta \leq 2/L$ in SARAH (Alg. 1), then for a particular inner loop s and any $t \geq 1$, we have*

$$\mathbb{E}\big[\|\mathbf{v}_t^s\|^2\big] \leq \bigg[1 - \bigg(\frac{2}{\eta L} - 1\bigg)\mu^2\eta^2\bigg]^t\,\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big].$$

Lemma 11. *[Nguyen et al., 2017, Theorem 1b] Suppose that Assumptions 1 and 4 hold. Choosing step size $\eta < 2/(\mu+L)$ in SARAH (Alg. 1), then for a particular inner loop s and any $t \geq 1$, we have*

$$\mathbb{E}\big[\|\mathbf{v}_t^s\|^2\big] \leq \bigg[1 - \frac{2\mu L\eta}{\mu+L}\bigg]^t \mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big].$$

Now we are ready to prove Lemma 6.

Case 1: Assumptions 1 – 3 hold. Following Assumption 1, we have

$$F(\mathbf{x}_{t+1}^s) - F(\mathbf{x}_t^s) \leq -\frac{\eta}{2}\Big[\|\nabla F(\mathbf{x}_t^s)\|^2 + \|\mathbf{v}_t^s\|^2 - \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2\Big] + \frac{(\eta)^2 L}{2}\|\mathbf{v}_t^s\|^2. \tag{20}$$

The derivation is exactly the same as (15), so we do not repeat it here. Rearranging the terms and dividing both sides with $\eta/2$, we have

$$\begin{aligned}
\|\nabla F(\mathbf{x}_t^s)\|^2 &\leq \frac{2\big[F(\mathbf{x}_t^s) - F(\mathbf{x}_{t+1}^s)\big]}{\eta} + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2 - \big(1-\eta L\big)\|\mathbf{v}_t^s\|^2 \\
&\overset{(a)}{\leq} \frac{2\big\langle \nabla F(\mathbf{x}_t^s), \mathbf{x}_t^s - \mathbf{x}_{t+1}^s\big\rangle}{\eta} + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2 - \big(1-\eta L\big)\|\mathbf{v}_t^s\|^2 \\
&\overset{(b)}{\leq} \frac{2}{\eta}\bigg[\frac{\delta\|\nabla F(\mathbf{x}_t^s)\|^2}{2} + \frac{\|\mathbf{x}_t^s - \mathbf{x}_{t+1}^s\|^2}{2\delta}\bigg] + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2 - \big(1-\eta L\big)\|\mathbf{v}_t^s\|^2
\end{aligned}$$

where (a) follows from the convexity of F; (b) uses Young's inequality with $\delta > 0$ to be specified later. Since $\mathbf{x}_{t+1}^s = \mathbf{x}_t^s - \eta\mathbf{v}_t^s$, rearranging the terms we have

$$\bigg(1 - \frac{\delta}{\eta}\bigg)\|\nabla F(\mathbf{x}_t^s)\|^2 \leq \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2 - \bigg(1 - \eta L - \frac{\eta}{\delta}\bigg)\|\mathbf{v}_t^s\|^2.$$

Choosing $\delta = 0.5\eta$, we have

$$\frac{1}{2}\|\nabla F(\mathbf{x}_t^s)\|^2 \leq \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2 + \big(1+\eta L\big)\|\mathbf{v}_t^s\|^2. \tag{21}$$

Then, taking expectation w.r.t. $\mathcal{F}_{t-1}$, applying Lemma 1 to $\mathbb{E}[\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2]$ and Lemma 10 to $\mathbb{E}[\|\mathbf{v}_t^s\|^2]$, with $t = m$ we have

$$\frac{1}{2}\mathbb{E}\big[\|\nabla F(\mathbf{x}_m^s)\|^2\big] \leq \frac{\eta L}{2-\eta L}\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2 + \big(1+\eta L\big)\bigg[1 - \bigg(\frac{2}{\eta L} - 1\bigg)\mu^2\eta^2\bigg]^m \mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big].$$

Multiplying both sides by 2 completes the proof.

Case 2: Assumptions 1 and 4 hold. Using exactly same arguments as Case 1 we can arrive at (21). Now applying Lemma 11, we have

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\big[\|\nabla F(\mathbf{x}_m^s)\|^2\big] &\leq \frac{\eta L}{2-\eta L}\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2 + \big(1+\eta L\big)\Big(1 - \frac{2\mu L\eta}{\mu+L}\Big)^m \mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big] \\
&= \frac{\eta L}{2-\eta L}\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2 + \big(1+\eta L\big)\Big(1 - \frac{2L\eta}{1+\kappa}\Big)^m \mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big].
\end{aligned}$$

Multiplying both sides by 2 completes the proof.

D.2 Proof of Theorem 3

We will only analyze case 1 where Assumptions 1 – 3 hold. The other case where Assumptions 1 and 4 are true can be analyzed in the same manner.

For analysis, let sequence $\{0, t_1, t_2, \ldots, t_N\}$, be the iteration indices where $B_{t_i} = 1$ (0 is automatically contained since at the beginning of L2S-SC, $\mathbf{v}_0$ is calculated). For a given sequence $\{0, t_1, t_2, \ldots, t_S\}$, it can be seen that due to the step back in Line 7 of Alg. 3, $\mathbf{x}_{t_i-1}$ plays the role of the starting point of an inner loop of SARAH; while $\mathbf{x}_{t_{i+1}-1}$ is analogous to $\mathbf{x}_m^s$ of SARAH's inner loop. Define $\mathbf{x}_{-1} = \mathbf{x}_0$ and

$$\lambda_{i+1} := \left\{ \frac{2\eta L}{2-\eta L} + \big(2+2\eta L\big)\left[1-\left(\frac{2}{\eta L}-1\right)\mu^2\eta^2\right]^{t_{i+1}-t_i} \right\}. \tag{22}$$

Using similar arguments of Lemma 6, when $\eta \leq 2/(3L)$, it is guaranteed to have

$$\begin{aligned}
\mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_S-1})\|^2 \big| \{0,t_1,t_2,\dots,t_S\}\big] &\leq \lambda_S \mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_{S-1}})\|^2 \big| \{0,t_1,t_2,\dots,t_S\}\big] \\
&= \lambda_S \mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_{S-1}-1})\|^2 \big| \{0,t_1,t_2,\dots,t_S\}\big] \\
&\leq \lambda_S\lambda_{S-1}\dots\lambda_1 \|\nabla F(\mathbf{x}_0)\|^2.
\end{aligned} \tag{23}$$

For convenience, let us define

$$\theta := 1 - \left(\frac{2}{\eta L}-1\right)\mu^2\eta^2.$$

Note that choosing η properly we can have $\theta < 1$. Now it can be seen that

$$\mathbb{E}[\theta^{t_{i+1}-t_i}|t_i] \leq \sum_{j=1}^{\infty} \frac{1}{m}\left(1-\frac{1}{m}\right)^{j-1}\theta^j \leq \frac{1}{m-1}\frac{\theta(1-\frac{1}{m})}{1-\theta(1-\frac{1}{m})}.$$

Note that this inequality is irrelevant with t_i. Thus if we further take expectation w.r.t. t_i, we arrive at

$$\mathbb{E}[\theta^{t_{i+1}-t_i}] \leq \frac{1}{m-1}\frac{\theta(1-\frac{1}{m})}{1-\theta(1-\frac{1}{m})}. \tag{24}$$

Plugging (24) into (22) we have

$$\mathbb{E}[\lambda_i] \leq \frac{2\eta L}{2-\eta L} + \frac{2+2\eta L}{m-1}\frac{\theta(1-\frac{1}{m})}{1-\theta(1-\frac{1}{m})} := \lambda, \forall i.$$

Note that the randomness of λ_{i+1} comes from $t_{i+1}-t_i$, which is the length of the interval between the calculation of two snapshot gradient. Since $\mathbb{P}\{t_{i+1}-t_i = u, t_{i+2}-t_{i+1} = v\} = \mathbb{P}\{t_{i+1}-t_i = u\}\mathbb{P}\{t_{i+2}-t_{i+1} = v\}$ for positive integers u and v, it can be seen $\{t_{i+1}-t_i\}$ are mutually independent, which further leads to the mutual independence of $\lambda_1, \lambda_2, \dots, \lambda_S$. Therefore, taking expectation w.r.t. $\{0, t_1, t_2, \dots, t_S\}$ on both sides of (23), we have

$$\mathbb{E}\big[\|\nabla F(\mathbf{x}_{t_S-1})\|^2\big] = \mathbb{E}[\lambda_S\lambda_{S-1}\dots\lambda_1]\|\nabla F(\mathbf{x}_0)\|^2 \leq \lambda^S \|\nabla F(\mathbf{x}_0)\|^2$$

which completes the proof.

D.3 When to Use An n-dependent Step Size in Convex Problems

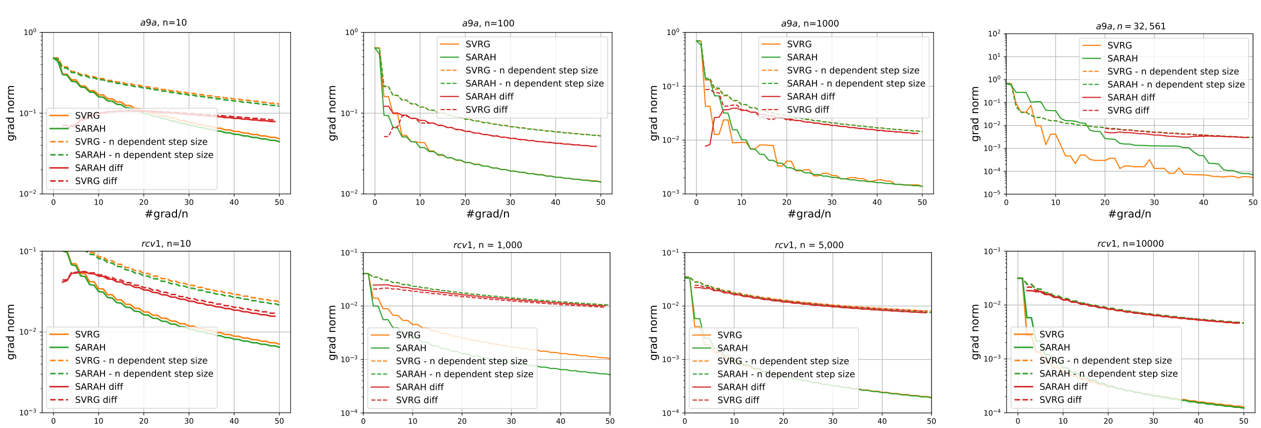

Figure 4: Performances of n-dependent step size and n-independent step size under on subsample datasets *rcv1* and *a9a*.

We perform SVRG and SARAH with n-dependent/independent step sizes to solve logistic regression problems on subsampled *rcv1* and *a9a*. The results can be found in Fig. 4. It can be seen that n-independent step sizes perform better than those of n-dependent step sizes in all the tests. In addition, as n increases, i) the gradient norm of solutions obtained via n-dependent step sizes becomes smaller; and ii) the performance gap between n-dependent and n-independent step sizes reduces. These observations suggest n-dependent step sizes can reveal their merits when n is extremely large (at least it should be larger than the size of *a9a*, which is $n = 32561$).

E Boosting the Practical Merits of SARAH

Algorithm 5 D2S

```
Initialize: x̃_0, η, m, S
for s = 1, 2, ..., S do
    x_0^s = x̃^{s-1}
    v_0^s = ∇F(x_0^s)
    x_1^s = x_0^s - η v_0^s
    for t = 1, 2, ..., m do
        Sample i_t according to p_t^s in (26)
        Compute v_t^s via (27)
        x_{t+1}^s = x_t^s - η v_t^s
    end for
    x̃^s uniformly rnd. chosen from {x_t^s}_{t=0}^m
end for
Output: x̃^S
```

Assumption 5. *Each* $f_i : \mathbb{R}^d \to \mathbb{R}$ *has* L_i*-Lipchitz gradient, and* F *has* L_F*-Lipchitz gradient; that is,* $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L_i \|\mathbf{x} - \mathbf{y}\|$, *and* $\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L_F \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

This section presents a simple yet effective variant of SARAH to enable a larger step size. The improvement stems from making use of the data dependent L_i in Assumption 5. The resultant algorithm that we term **D**ata **D**ependent **S**ARAH (D2S) is summarized in Alg. 5. For simplicity D2S is developed based on SARAH, but it generalizes to L2S as well.

Intuitively, each f_i provides a distinct gradient to be used in the updates. The insight here is that if one could quantify the "importance" of f_i (or the gradient it provides), those more important ones should be used more frequently. Formally, our idea is to draw i_t of outer loop s according to a probability mass vector $\mathbf{p}_t^s \in \Delta_n$, where $\Delta_n := \{\mathbf{p} \in \mathbb{R}_+^n | \langle \mathbf{1}, \mathbf{p} \rangle = 1\}$. With $\mathbf{p}_t^s = 1/n$, D2S boils down to SARAH.

Ideally, finding $\mathbf{p}_t^s$ should rely on the estimation error as optimality crietrion. Specifically, we wish to minimize $\mathbb{E}[\|\mathbf{v}_t^s - \nabla F(\mathbf{x}_t^s)\|^2 | \mathcal{F}_{t-1}]$ in Lemma 1. Writing the expectation explicitly, the problem can be posed as

$$\min_{\mathbf{p}_t^s \in \Delta_n} \frac{1}{n^2} \sum_{i \in [n]} \frac{\|\nabla f_i(\mathbf{x}_t^s) - \nabla f_i(\mathbf{x}_{t-1}^s)\|^2}{p_{t,i}^s} \quad \Rightarrow \quad (p_{t,i}^s)^* = \frac{\|\nabla f_i(\mathbf{x}_t^s) - \nabla f_i(\mathbf{x}_{t-1}^s)\|}{\sum_{j \in [n]} \|\nabla f_j(\mathbf{x}_t^s) - \nabla f_j(\mathbf{x}_{t-1}^s)\|} \tag{25}$$

where the $(p_{t,i}^s)^*$ denotes the optimal solution. Though finding out $\mathbf{p}_t^s$ via (25) is optimal, it is intractable to implement because $\nabla f_i(\mathbf{x}_{t-1}^s)$ and $\nabla f_i(\mathbf{x}_t^s)$ for all $i \in [n]$ must be computed, which is even more expensive than computing $\nabla F(\mathbf{x}_t^s)$ itself. However, (25) implies that a larger probability should be assigned to those $\{f_i\}$ whose gradients on $\mathbf{x}_t^s$ and $\mathbf{x}_{t-1}^s$ change drastically. The intuition behind this observation is that a more abrupt change of the gradient suggests a larger residual to be optimized. Thus, $\|\nabla f_i(\mathbf{x}_t^s) - \nabla f_i(\mathbf{x}_{t-1}^s)\|^2$ in (25) can be approximated by its upper bound $L_i^2 \|\mathbf{x}_t^s - \mathbf{x}_{t-1}^s\|^2$, which inaccurately captures gradient changes. The resultant problem and its optimal solution are

$$\min_{\mathbf{p}_t^s \in \Delta_n} \frac{1}{n^2} \sum_{i \in [n]} \frac{L_i^2 \|\mathbf{x}_t^s - \mathbf{x}_{t-1}^s\|^2}{p_{t,i}^s} \quad \Rightarrow \quad (p_{t,i}^s)^* = \frac{L_i}{\sum_{j \in [n]} L_j}, \forall t, \forall s. \tag{26}$$

Choosing $\mathbf{p}_t^s$ according to (26) is computationally attractive not only because it eliminates the need to compute gradients, but also because L_i is usually cheap to obtain in practice (at least for linear and logistic regression losses). Knowing

$L = \max_{i\in[n]} L_i$ is critical for SARAH [Nguyen et al., 2017]; hence, finding $\mathbf{p}_t^s$ only introduces negligible overhead compared to SARAH. Accounting for $\mathbf{p}_t^s$, the gradient estimator $\mathbf{v}_t^s$ is also modified to an importance sampling based one to compensate for those less frequently sampled $\{f_i\}$

$$\mathbf{v}_t^s = \frac{\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\mathbf{x}_{t-1}^s)}{np_{t,i_t}^s} + \mathbf{v}_{t-1}^s. \tag{27}$$

Note that $\mathbf{v}_t^s$ is still biased, since $\mathbb{E}[\mathbf{v}_t^s|\mathcal{F}_{t-1}] = \nabla F(\mathbf{x}_t^s) - \nabla F(\mathbf{x}_{t-1}^s) + \mathbf{v}_{t-1}^s \neq \nabla F(\mathbf{x}_t^s)$. As asserted next, with $\mathbf{p}_t^s$ as in (26) and $\mathbf{v}_t^s$ computed via (27), D2S indeed improves SARAH's convergence rate.

Theorem 4. *If Assumptions 5, 2, and 3 hold, upon choosing $\eta < 1/\bar{L}$ and a large enough m such that $\sigma_m := \frac{1}{\mu\eta(m+1)} + \frac{\eta\bar{L}}{2-\eta\bar{L}} < 1$, D2S convergences linearly; that is,*

$$\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}_s)\|^2\big] \leq (\sigma_m)^s \|\nabla F(\tilde{\mathbf{x}}_0)\|^2, \forall s.$$

Compared with SARAH's linear convergence rate $\tilde{\sigma}_m = \frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2-\eta L}$ [Nguyen et al., 2017], the improvement on the convergence constant σ_m is twofold: i) if η and m are chosen the same in D2S and SARAH, it always holds that $\sigma_m \leq \tilde{\sigma}_m$, which implies D2S converges faster than SARAH; and ii) the step size can be chosen more aggressively with $\eta < 1/\bar{L}$, while the standard SARAH step size has to be less than $1/L$. The improvements are further corroborated in terms of the number of IFO calls, especially for ERM problems that are ill-conditioned.

Corollary 7. *If Assumptions 5, 2, and 3 hold, to find $\tilde{\mathbf{x}}^s$ such that $\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^s)\|^2\big] \leq \epsilon$, D2S requires $\mathcal{O}\big((n+\bar{\kappa})\ln(1/\epsilon)\big)$ IFO calls, where $\bar{\kappa} := \bar{L}/\mu$.*

E.1 Optimal Solution of (25)

The optimal solution of (25) can be directly obtained from the partial Lagrangian

$$\mathcal{L}(\mathbf{p}_t^s, \lambda) = \frac{1}{n^2}\sum_{i\in[n]} \frac{\|\nabla f_i(\mathbf{x}_t^s) - \nabla f_i(\mathbf{x}_{t-1}^s)\|^2}{p_{t,i}^s} + \lambda \sum_{i\in[n]} p_{t,i}^s - \lambda.$$

Taking derivative w.r.t. $\mathbf{p}_t^s$ and set it to $\mathbf{0}$, we have

$$p_{t,i}^s = \frac{\|\nabla f_i(\mathbf{x}_t^s) - \nabla f_i(\mathbf{x}_{t-1}^s)\|}{\sqrt{\lambda} n}.$$

Note that if $\lambda > 0$, it automatically satisfies $p_{t,i}^s \geq 0$. Then let $\sum_{i\in[n]} p_{t,i}^s = 1$, it is not hard to find the value of λ and obtain (25). The solution of (26) can be derived in a similar manner.

E.2 Proof of Theorem 4

The proof generalizes the original proof of SARAH for strongly convex problems [Nguyen et al., 2017, Theorem 2]. Notice that the importance sampling based gradient estimator enables the fact $\mathbb{E}_{i_t}\big[\mathbf{v}_t^s|\mathcal{F}_{t-1}\big] = \nabla F(\mathbf{x}_t^s) - \nabla F(\mathbf{x}_{t-1}^s) + \mathbf{v}_{t-1}^s$. By exploring this fact, it is not hard to see that the following lemmas hold. The proof has almost the same steps as those in [Nguyen et al., 2017], except for the expectation now is w.r.t. a nonuniform distribution $\mathbf{p}_t^s$.

Lemma 12. *[Nguyen et al., 2017, Lemma 1] In any outer loop s, if $\eta \leq 1/L_F$, we have*

$$\sum_{t=0}^{m} \mathbb{E}\big[\|\nabla F(\mathbf{x}_t^s)\|^2\big] \leq \frac{2}{\eta}\mathbb{E}\big[F(\mathbf{x}_0^s) - F(\mathbf{x}^*)\big] + \sum_{t=0}^{m}\mathbb{E}\big[\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|\big].$$

Lemma 13. *The following equation is true*

$$\mathbb{E}\big[\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2\big] = \sum_{\tau=1}^{t}\mathbb{E}\big[\|\mathbf{v}_\tau^s - \mathbf{v}_{\tau-1}^s\|^2\big] - \sum_{\tau=1}^{t}\mathbb{E}\big[\|\nabla F(\mathbf{x}_\tau^s) - \nabla F(\mathbf{x}_{\tau-1}^s)\|^2\big].$$

Lemma 14. *In any outer loop s, if η is chosen to satisfy $1 - \frac{2}{\eta\bar{L}} < 0$, we have*

$$\mathbb{E}\big[\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2|\mathcal{F}_{t-1}\big] \leq \frac{\eta\bar{L}}{2-\eta\bar{L}}\Big(\|\mathbf{v}_{t-1}^s\|^2 - \mathbb{E}\big[\|\mathbf{v}_t^s\|^2|\mathcal{F}_{t-1}\big]\Big), \forall t \geq 1.$$

Proof. Consider that for any $t \geq 1$

$$\begin{aligned}
&\mathbb{E}_{i_t}\big[\|\mathbf{v}_t^s\|^2|\mathcal{F}_{t-1}\big] = \mathbb{E}_{i_t}\big[\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s + \mathbf{v}_{t-1}^s\|^2|\mathcal{F}_{t-1}\big] \\
&= \|\mathbf{v}_{t-1}^s\|^2 + \mathbb{E}\big[\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2|\mathcal{F}_{t-1}\big] + 2\mathbb{E}\big[\langle \mathbf{v}_{t-1}^s, \mathbf{v}_t^s - \mathbf{v}_{t-1}^s\rangle|\mathcal{F}_{t-1}\big] \\
&\overset{(a)}{=} \|\mathbf{v}_{t-1}^s\|^2 + \mathbb{E}\Big[\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2 + \frac{2}{\eta}\Big\langle \mathbf{x}_{t-1}^s - \mathbf{x}_t^s, \frac{\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\mathbf{x}_{t-1}^s)}{np_{t,i_t}^s}\Big\rangle\Big|\mathcal{F}_{t-1}\Big] \\
&\overset{(b)}{\leq} \|\mathbf{v}_{t-1}^s\|^2 + \mathbb{E}\Big[\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2 - \frac{2}{\eta L_{i_t} np_{t,i_t}^s}\|\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\mathbf{x}_{t-1}^s)\|^2\Big|\mathcal{F}_{t-1}\Big] \\
&\overset{(c)}{=} \|\mathbf{v}_{t-1}^s\|^2 + \mathbb{E}\Big[\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2 - \frac{2np_{t,i_t}^s}{\eta L_{i_t}}\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2\Big|\mathcal{F}_{t-1}\Big] \\
&\overset{(d)}{=} \|\mathbf{v}_{t-1}^s\|^2 + \mathbb{E}\Big[\Big(1 - \frac{2}{\eta \bar{L}}\Big)\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2\Big|\mathcal{F}_{t-1}\Big]
\end{aligned}$$

where (a) follows from (27) and the update $\mathbf{x}_t^s = \mathbf{x}_{t-1}^s - \eta \mathbf{v}_t^s$; (b) is the result of (9c); (c) is by the definition of $\mathbf{v}_t^s$; and (d) is by plugging (26) in. By choosing η such that $1 - \frac{2}{\eta \bar{L}} < 0$, we have

$$\mathbb{E}\big[\|\mathbf{v}_t^s - \mathbf{v}_{t-1}^s\|^2|\mathcal{F}_{t-1}\big] \leq \frac{\eta \bar{L}}{2 - \eta \bar{L}}\Big(\|\mathbf{v}_{t-1}^s\|^2 - \mathbb{E}\big[\|\mathbf{v}_t^s\|^2|\mathcal{F}_{t-1}\big]\Big)$$

which concludes the proof. ∎

Proof of Theorem 4: Using Lemmas 13 and 14 we have

$$\begin{aligned}
\mathbb{E}\big[\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^2\big] &= \sum_{\tau=1}^{t}\mathbb{E}\big[\|\mathbf{v}_\tau^s - \mathbf{v}_{\tau-1}^s\|^2\big] - \sum_{\tau=1}^{t}\mathbb{E}\big[\|\nabla F(\mathbf{x}_\tau^s) - \nabla F(\mathbf{x}_{\tau-1}^s)\|^2\big] \\
&\leq \frac{\eta \bar{L}}{2 - \eta \bar{L}}\mathbb{E}\big[\|\mathbf{v}_0^s\|^2\big]. \qquad (28)
\end{aligned}$$

If we further let $\eta \leq 1/L_F$, plugging (28) into Lemma 12, we have

$$\sum_{t=0}^{m}\mathbb{E}\big[\|\nabla F(\mathbf{x}_t^s)\|^2\big] \leq \frac{2}{\eta}\mathbb{E}\big[F(\mathbf{x}_0^s) - F(\mathbf{x}^*)\big] + \frac{(m+1)\eta \bar{L}}{2 - \eta \bar{L}}\mathbb{E}\big[\|\mathbf{v}_0^s\|^2\big].$$

Since $\tilde{\mathbf{x}}^s$ is uniformly randomized chosen from $\{\mathbf{x}_t^s\}_{t=0}^m$, by exploiting the fact $\mathbf{v}_0^s = \nabla F(\tilde{\mathbf{x}}^{s-1})$ and $\mathbf{x}_0^s = \tilde{\mathbf{x}}^{s-1}$, we have that

$$\begin{aligned}
\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^s)\|^2\big] &\leq \frac{2}{\eta(m+1)}\mathbb{E}\big[F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*)\big] + \frac{\eta \bar{L}}{2 - \eta \bar{L}}\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big] \\
&\leq \Big(\frac{2}{\mu\eta(m+1)} + \frac{\eta \bar{L}}{2 - \eta \bar{L}}\Big)\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big] \qquad (29)
\end{aligned}$$

where the last inequality follows from (10a). Unrolling $\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^{s-1})\|^2\big]$ in (29), Theorem 4 can be proved.

E.3 Proof of Corollary 7

The proof is modified from [Nguyen et al., 2017, Corollary 3]. By choosing $\eta = 0.5/(\bar{L})$ and $m = 4.5\bar{\kappa}$, we have σ_m in Theorem 4 bounded by

$$\sigma_m = \frac{1}{\frac{1}{2\bar{\kappa}}(4.5\bar{\kappa} + 1)} + \frac{0.5}{1.5} < \frac{7}{9}.$$

Then by Theorem 4, by choosing S as

$$S \geq \frac{\ln\big(\|\nabla F(\tilde{\mathbf{x}}^0)\|^2/\epsilon\big)}{\ln(9/7)} \geq \log_{7/9}(\|\nabla F\big(\tilde{\mathbf{x}}^0\big)\|^2/\epsilon)$$

we have $\mathbb{E}\big[\|\nabla F(\tilde{\mathbf{x}}^S)\|^2\big] \leq (\sigma_m)^2\|\nabla F\big(\tilde{\mathbf{x}}^0\big)\|^2 \leq \epsilon$. Thus the number of IFO calls is

$$(n + 2m)S = \mathcal{O}\big((n + \bar{\kappa})\ln(1/\epsilon)\big).$$

Table 1: A summary of datasets used in numerical tests

Dataset	d	n (train)	density	n (test)	L	λ
a9a	123	$32,561$	11.28%	$16,281$	3.4672	0.0005
rcv1	$47,236$	$20,242$	0.157%	$677,399$	0.25	0.0001
w7a	300	$24,692$	3.89%	$25,057$	2.917	0.005

F Numerical Experiments

Experiments for (strongly) convex cases are performed using python 3.7 on an Intel i7-4790CPU @3.60 GHz (32 GB RAM) desktop. The details of the used datasets are summarized in Table 1. The smoothness parameter L_i can be calculated via $L_i = \|\mathbf{a}_i\|^2/4$ by checking the Hessian matrix.

L2S. Since we are considering the convex case, we set $\lambda = 0$ in (8). SVRG, SARAH and SGD are chosen as benchmarks, where SGD is modified with step size $\eta_k = 1/\big(\bar{L}(k+1)\big)$ on the k-th epoch. For both SARAH and SVRG, the length of inner loop is chosen as $m = n$. For a fair comparison, we use the same m for L2S [cf. (3)]. The step sizes of SARAH and SVRG are selected from $\{0.01/\bar{L}, 0.1/\bar{L}, 0.2/\bar{L}, 0.3/\bar{L}, 0.4/\bar{L}, 0.5/\bar{L}, 0.6/\bar{L}, 0.7/\bar{L}, 0.8/\bar{L}, 0.9/\bar{L}, 0.95/\bar{L}\}$ and those with best performances are reported. Note that the SVRG theory only effects when $\eta < 0.25/\bar{L}$. The step size of L2S is the same as that of SARAH for fairness.

L2S-SC. The parameters are chosen in the same manner as the test of L2S.

L2S for on Nononvex Problems We perform classification on MNIST dataset using a $784 \times 128 \times 10$ feedforward neural network through Pytorch. The activation function used in hidden layer is sigmoid. SGD, SVRG, and SARAH are adopted as benchmarks. In all tested algorithms the batch sizes are $b = 32$. The step size of SGD is $\mathcal{O}(\sqrt{b}/(k+1))$, where k is the index of epoch; the step size is chosen as $b/(Ln^{2/3})$ for SVRG [Reddi et al., 2016a]; and the step sizes are $\sqrt{b}/(2\sqrt{n}L)$ for SARAH [Nguyen et al., 2019] and L2S. The inner loop lengths are selected to be $m = n/b$ for SVRG and SARAH, while the same m is used for L2S.