1. Appendix

1.1. Proof of Lemma 1

It is straight forward to see:

$$\begin{aligned}
\mathbb{E}\bar{H}_{t+1} = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} f_i^{\delta}(\phi_i^{t+1})] &= \mathbb{E}[\frac{1}{n}\big(\sum_{i\in\mathcal{B}} f_i^{\delta}(w^t) + \sum_{i\notin\mathcal{B}} f_i^{\delta}(\phi_i^t)\big)] \\
&= \mathbb{E}[\mathbb{E}[\frac{1}{n}\big(\sum_{i\in\mathcal{B}} f_i^{\delta}(w^t) + \sum_{i\notin\mathcal{B}} f_i^{\delta}(\phi_i^t)\big)|\mathcal{B}]| \quad |\mathcal{B}| = B] \\
&= \frac{1}{n}\big(\frac{\mathbb{E}|\mathcal{B}|}{n}\sum_{i=1}^{n} f_i^{\delta}(w^t) + \frac{n-\mathbb{E}|\mathcal{B}|}{n}\sum_{i=1}^{n} f_i^{\delta}(\phi_i^t)\big) \\
&= \frac{\mathbb{E}|\mathcal{B}|}{n} f^{\delta}(w^t) + \frac{n-\mathbb{E}|\mathcal{B}|}{n}\bar{H}_t
\end{aligned}$$

The second line of equality comes from the rule of total expectation, where the inner expectation is taken over the index set $\mathcal{B}$, and the outer expectation is taken over the set cardinality $|\mathcal{B}|$.

1.2. Proof of Lemma 2

The proof technique is similar to SAGA, as well as a useful inequality (Lemma 4 in (**?**)):

$$\begin{aligned}
f(x) \geq f(y) + \langle f'(y), x-y\rangle + \frac{1}{2(L-\mu)}\|f'(x)-f'(y)\|^2 \\
+ \frac{\mu L}{2(L-\mu)}\|x-y\|^2 - \frac{\mu}{(L-\mu)}\langle f'(x)-f'(y), x-y\rangle.
\end{aligned} \tag{A1}$$

First of all, by the update rule (2):

$$\begin{aligned}
\|w^{t+1}-w^*\|^2 &= \|\mathsf{Prox}_{\gamma g}(w^t-\gamma G(w^t)) - \mathsf{Prox}_{\gamma g}(w^*-\gamma f'(w^*))\|^2 \\
&\leq \|w^t-\gamma G(w^t)-w^*+\gamma f'(w^*)\|^2 \\
&= \|w^t-w^*\|^2 - 2\gamma\langle w^t-w^*, G(w^t)-f'(w^*)\rangle + \gamma^2\|G(w^t)-f'(w^*)\|^2.
\end{aligned} \tag{A2}$$

The inequality follows from non-expansiveness of proximal operator, notice that our stochastic gradient $G(w^t)$ is unbiased, take the expectation to the second term and apply (A1) to each f_i and the average over all i will goes to:

$$\begin{aligned}
-\mathbb{E}[\langle w^t-w^*, G(w^t)-f'(w^*)\rangle] &= -\langle w^t-w^*, f'(w^t)-f'(w^*)\rangle \\
\leq \langle w^t-w^*, f'(w^*)\rangle &+ \frac{L-\mu}{L}[f(w^*)-f(w^t)] - \frac{\mu}{2}\|w^*-w^t\|^2 \\
&-\frac{1}{2Ln}\sum_{i=1}^{n}\|f_i'(w^*)-f_i'(w^t)\|^2 - \frac{\mu}{L}\langle f'(w^*), w^t-w^*\rangle.
\end{aligned} \tag{A3}$$

Next we bound the last term in (A2):

$$
\begin{aligned}
\mathbb{E}\left\|f'(w^*) - G(w^t)\right\|^2 &= \mathbb{E}\left\|f'(w^*) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(\phi_i^t) - \frac{1}{n}\sum_{i=1}^{n} f_i'(\phi_i^t)\right\|^2 \\
&= \mathbb{E}\Bigg\|\Big[\frac{1}{n}\sum_{i=1}^{n} f_i'(w^t) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\Big] \\
&\quad - \Big[\frac{1}{n}\sum_{i=1}^{n} f_i'(\phi_i^t) - \frac{1}{|\mathcal{B}|} f_i'(\phi_i^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\Big] \\
&\quad + f'(w^*) - \frac{1}{n}\sum_{i=1}^{n} f_i'(w^t)\Bigg\|^2 \qquad \text{(A4)} \\
&\overset{*}{=} \mathbb{E}\Bigg\|\Big[\frac{1}{n}\sum_{i=1}^{n} f_i'(w^t) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\Big] \\
&\quad - \Big[\frac{1}{n}\sum_{i=1}^{n} f_i'(\phi_i^t) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(\phi_i^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\Big]\Bigg\|^2 \\
&\quad + \left\|f'(w^*) - \frac{1}{n}\sum_{i=1}^{n} f_i'(w^t)\right\|^2.
\end{aligned}
$$

In equation $\overset{*}{=}$ we use the property that $\mathbb{E}[X^2] = \mathbb{E}[X - \mathbb{E}[X]]^2 + \mathbb{E}[X]^2$, now use the inequality $\|X+Y\|^2 \le (1+\beta)\|X\|^2 + (1+\beta^{-1})\|Y\|^2$, $\beta > 0$ to the first term:

$$
\begin{aligned}
\mathbb{E}\left\|f'(w^*) - G(w^t)\right\|^2 &\le (1+\beta)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} f_i'(w^t) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\right\|^2 \\
&\quad + (1+\beta^{-1})\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} f_i'(\phi_i^t) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(\phi_i^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\right\|^2 \qquad \text{(A5)} \\
&\quad + \beta\cdot\left\|f'(w^*) - \frac{1}{n}\sum_{i=1}^{n} f_i'(w^t)\right\|^2.
\end{aligned}
$$

Next we bound the first and second terms again by variance decomposition, for simplicity we only take the first term as example:

$$
\begin{aligned}
&\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} f_i'(w^t) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\right\|^2 \\
=&\mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\Big(f_i'(w^*) - f_i'(w^t)\Big)\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}\Big(f_i'(w^*) - f_i'(w^t)\Big)\right\|^2 \\
\overset{(1)}{\le}&\mathbb{E}\Big(\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\|f_i'(w^*) - f_i'(w^t)\|^2\Big) - \left\|\frac{1}{n}\sum_{i=1}^{n}\Big(f_i'(w^*) - f_i'(w^t)\Big)\right\|^2 \qquad \text{(A6)} \\
=&\frac{1}{n}\sum_{i=1}^{n}\|f_i'(w^*) - f_i'(w^t)\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}\Big(f_i'(w^*) - f_i'(w^t)\Big)\right\|^2 \\
\overset{(2)}{\le}&\frac{1}{n}\sum_{i=1}^{n}\|f_i'(w^*) - f_i'(w^t)\|^2,
\end{aligned}
$$

$\overset{(1)}{\leq}$ is by RMS-AM inequality, and in $\overset{(2)}{\leq}$ we drop the negative term. Similarly,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} f_i'(\phi_i^t) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(\phi_i^t) - f'(w^*) + \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} f_i'(w^*)\right\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\|f_i'(w^*) - f_i'(\phi_i^t)\|^2.$$

Plug (A6) into (A5) we get:

$$\begin{aligned}\mathbb{E}\left\|f'(w^*) - G(w^t)\right\|^2 \leq \frac{(1+\beta)}{n}\sum_{i=1}^{n}\|f_i'(w^*) - f_i'(w^t)\|^2 + \frac{(1+\beta^{-1})}{n}\sum_{i=1}^{n}\|f_i'(w^*) - f_i'(\phi_i^t)\|^2 \\ - \beta\|f'(w^t) - f'(w^*)\|^2.\end{aligned} \tag{A7}$$

Combining (A2),(A3),(A7) becomes (5) immediately:

$$\begin{aligned}\|w^t - w^*\|^2 - \mathbb{E}\|w^{t+1} - w^*\|^2 \geq & \gamma\mu\|w^t - w^*\|^2 - (2\gamma^2 - \gamma/L)\mathbb{E}\|f_i'(w^t) - f_i'(w^*)\|^2 \\ & + \gamma^2\|f'(w^t) - f'(w^*)\|^2 + \frac{2\gamma(L-\mu)}{L}f^{\delta}(w^t) - 4\gamma^2 L\bar{H}_t.\end{aligned}$$

1.3. Proof of Theorem 3

It follows directly from Lemma 1 and 2:

$$\begin{aligned}\mathcal{L}_t - \mathbb{E}\mathcal{L}_{t+1} &= c(\bar{H}_t - \mathbb{E}\bar{H}_{t+1}) + (\|w^t - w^*\|^2 - \mathbb{E}\|w^{t+1} - w^*\|^2) \\ &\geq c\Big(\frac{\mathbb{E}|\mathcal{B}|}{n} - \frac{2(1+\beta^{-1})\gamma^2 L}{c}\Big)\bar{H}_t + \gamma\mu\|w^t - w^*\|^2 + (2\mu\beta\gamma^2 + \frac{2\gamma(L-\mu)}{L} - \frac{c\cdot\mathbb{E}|\mathcal{B}|}{n})f^{\delta}(w^t) \\ &\quad + (\frac{\gamma}{L} - (1+\beta)\gamma^2)\mathbb{E}\|f_i'(w^t) - f'(w^*)\|^2 \\ &\overset{?}{\geq} c\Big(\frac{|\mathcal{B}|}{n} - \frac{2(1+\beta^{-1})\gamma^2 L}{c}\Big)\bar{H}_t + \gamma\mu\|w^t - w^*\|^2 \\ &\geq \rho\mathcal{L}_t,\end{aligned} \tag{A8}$$

where $\rho = \min(\frac{|\mathcal{B}|}{n} - \frac{2(1+\beta^{-1})\gamma^2 L}{c}, \gamma\mu)$, the last inequality $\overset{?}{\geq}$ comes with following condition:

$$\begin{aligned}2\mu\beta\gamma^2 + \frac{2\gamma(L-\mu)}{L} - \frac{c|\mathcal{B}|}{n} \geq 0 \\ \frac{\gamma}{L} - (1+\beta)\gamma^2 \geq 0,\end{aligned} \tag{A9}$$

furthermore, to keep our algorithm moving forward, i.e. $\|w^t - w^*\|^2$ decreasing, we should also make sure such condition hold:

$$\frac{|\mathcal{B}|}{n} - \frac{2(1+\beta^{-1})\gamma^2 L}{c} \geq 0. \tag{A10}$$

1.4. Proof of Proposition 1

By plugging $\beta = 2$, $c = \frac{n}{3L\mathbb{E}|\mathcal{B}|}$ into (A9) it is easy to verify both inequalities hold.

1.5. Proof of Proposition 2

In this case we choose $\beta = 1$. From Theorem 3 we know that with a suitable step size γ and c, we have:

$$\mathbb{E}\|w^t - w^*\|^2 \leq \mathbb{E}\mathcal{L}_t \leq (1-\rho)^t\mathcal{L}_0 = (1-\rho)^t\big[\|w^0 - w^*\|^2 + c\bar{H}_0\big].$$

For the optimal convergence rate, we try to maximize the geometric factor $\rho = \min(\frac{\mathbb{E}|\mathcal{B}|}{n} - \frac{4\gamma^2 L}{c}, \gamma\mu)$. Denote γ_0 as the solution of: $\frac{\mathbb{E}|\mathcal{B}|}{n} - \frac{4\gamma_0^2 L}{c} = \gamma_0\mu$. Notice that $\rho(\gamma) = \gamma\mu$ is increasing with γ when $\gamma \leq \gamma_0 = \frac{c}{8\kappa}\Big(\sqrt{1 + \frac{16\kappa\mathbb{E}|\mathcal{B}|}{cn\mu}} - 1\Big)$

and $\rho(\gamma) = \frac{\mathbb{E}|\mathcal{B}|}{n} - \frac{4\gamma^2 L}{c}$ is decreasing when $\gamma > \gamma_0$. So the optimal step size should be $\gamma = \gamma_0$. However we should also verify that this step size indeed satisfies the condition in (A9). First of all:

$$\gamma_0 = \frac{c}{8\kappa}\Big(\sqrt{1+\frac{16\kappa\mathbb{E}|\mathcal{B}|}{cn\mu}} - 1\Big) \overset{(1)}{\leq} \frac{c}{8\kappa}\sqrt{\frac{16\kappa\mathbb{E}|\mathcal{B}|}{cn\mu}} = \sqrt{\frac{c\mathbb{E}|\mathcal{B}|}{4nL}} \overset{(2)}{\leq} \frac{1}{2L}. \tag{A11}$$

$\overset{(1)}{\leq}$ comes from the fact that $\sqrt{1+x}-1 \leq \sqrt{x}$, $\overset{(2)}{\leq}$ holds by choosing $c = \frac{\tau n}{L\mathbb{E}|\mathcal{B}|}$, where $\tau < 1$ is a small constant. These two inequalities together ensure the upper bound part of (A9). As to the lower bound, we have $\sqrt{1+x}-1 > \sqrt{x}-1$, so:

$$\gamma_0 > \frac{c}{8\kappa}\Big(\sqrt{\frac{16\kappa\mathbb{E}|\mathcal{B}|}{cn\mu}} - 1\Big) \geq \frac{c\mathbb{E}|\mathcal{B}|L}{2n(L-\mu)} \Longrightarrow \tau \leq \Big(\frac{1}{\frac{L}{L-\mu} + \frac{n}{4\kappa\mathbb{E}|\mathcal{B}|}}\Big)^2 < 1.$$

So if we choose τ properly, both sides of (A9) can be satisfied.

1.6. Proof of Corollary 1, 2

Following (7) we take a derivative to $\mathbb{E}|\mathcal{B}|$:

$$\frac{\partial f(\mathbb{E}|\mathcal{B}|)}{\partial \mathbb{E}|\mathcal{B}|} = \frac{(\alpha\mathbb{E}|\mathcal{B}|-1)^2\mathbb{E}|\mathcal{B}|}{\sqrt{1+\alpha^2\mathbb{E}|\mathcal{B}|^2}(\sqrt{1+\alpha^2\mathbb{E}|\mathcal{B}|^2}-1)^2} \geq 0, \tag{A12}$$

where $\alpha = \frac{4\kappa}{\sqrt{\tau n}}$, so there is no optimal batch size, and since we always want to access one data point, i.e. $|\mathcal{B}| \geq 1$ and SAGA style update is optimal.

For Corollary 2, it is easy to see for our algorithm, which choose $|\mathcal{B}| = n$ with probability $p \ll 1$ and $|\mathcal{B}| = 1$ with probability $1-p$, has average batch size $\mathbb{E}|\mathcal{B}| = np + 1 - p \approx np + 1$. For each update, it takes on average time $\tau = n\eta\tau p + (1-p)\tau \approx (1+np\eta)\tau$. If we want to get a ϵ-suboptimal solution, the total iteration will be $N = \frac{\log(\epsilon/\epsilon_0)}{\log(1-\rho)} \propto 1/\rho$, So the running time will be:

$$\begin{aligned} T &\propto \frac{1+np\eta}{\sqrt{\frac{1}{\mathbb{E}|\mathcal{B}|^2} + \frac{16\kappa^2}{\tau n^2}} - \frac{1}{\mathbb{E}|\mathcal{B}|}} \\ &\approx \frac{(\mathbb{E}|\mathcal{B}|^2 - \mathbb{E}|\mathcal{B}|)\eta + \mathbb{E}|\mathcal{B}|}{\sqrt{1+\alpha^2\mathbb{E}|\mathcal{B}|^2} - 1}. \end{aligned} \tag{A13}$$

For simplicity we denote $B = \mathbb{E}|\mathcal{B}|$. By taking the partial derivative and set it to zero $\partial T/\partial B = 0$ can solve the best batch size:

$$\big((2B|-1)\eta+1\big)(\sqrt{1+\alpha^2B^2}-1) = \big((B^2-B)\eta+B\big)\frac{\alpha^2B}{\sqrt{1+\alpha^2B^2}}, \tag{A14}$$

after solving the above equation we get:

$$B = \left(\frac{1}{\eta}-1\right)\left(\frac{\xi-1}{2-\xi}\right), \quad \xi = \frac{\alpha^2B^2}{1+\alpha^2B^2-\sqrt{1+\alpha^2B^2}}. \tag{A15}$$

By showing the second order derivative $\partial^2 T/\partial B^2 \geq 0$ it's easy to verify that this solution is actually a global minimum.

1.7. Proof of Lemma 4

We begin with non-expansiveness of proximal operation:

$$\begin{aligned} \|w^{t+1}-w^*\|^2 &= \|\mathsf{Prox}_{\gamma g}(w^t - \gamma G(w^t)) - \mathsf{Prox}_{\gamma g}(w^* - \gamma f'(w^*))\|^2 \\ &\leq \|w^t - \gamma G(w^t) - w^* + \gamma f'(w^*)\|^2 \\ &= \|w^t - w^*\|^2 - 2\gamma\langle w^t - w^*, G(w^t) - f'(w^*)\rangle + \gamma^2\|G(w^t) - f'(w^*)\|^2, \end{aligned} \tag{A16}$$

where $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$ By taking expectation on each side and notice $G(w^t)$ is a unbiased estimation of $f'(w^t)$:

$$\mathbb{E}\|w^{t+1}-w^*\|^2 = \|w^t-w^*\|^2 - 2\gamma\langle w^t-w^*, f'(w^t) - f'(w^*)\rangle + \gamma^2\mathbb{E}\|G(w^t) - f'(w^*)\|^2, \tag{A17}$$

and then apply the following bounds for strongly convex function f:

$$\begin{aligned}\langle w^t - w^*, f'(w^t) - f'(w^*)\rangle &\geq \mu\|w^t - w^*\|^2 \\ \langle w^t - w^*, f'(w^t) - f'(w^*)\rangle &\geq \frac{1}{L}\|f'(w^t) - f'(w^*)\|^2,\end{aligned} \tag{A18}$$

so the inner product term have a composite upper bound:

$$-2\gamma\langle w^t - w^*, f'(w^t) - f'(w^*)\rangle \leq -\gamma(\mu\|w^t - w^*\|^2 + \frac{1}{L}\|f'(w^t) - f'(w^*)\|^2) \tag{A19}$$

on the other hand, we can bound $\mathbb{E}\|G(w^t) - f'(w^*)\|^2$ as (A5) but we only need to care about one sample in a batch case, since we are comparing SAGA with SVRG update style:

$$\mathbb{E}\|G(w^t) - f'(w^*)\|^2 \leq 2\mathbb{E}\|f_i'(\phi_i^t) - f_i'(w^*)\|^2 + 2\mathbb{E}\|f_i'(w^t) - f_i'(w^*)\|^2 - \|f'(w^t) - f'(w^*)\|^2. \tag{A20}$$

Remember we have proved above formula in (A7), for $\mathbb{E}\|f_i'(w^t) - f_i'(w^*)\|^2$ we have:

$$\begin{aligned}\mathbb{E}\|f_i'(w^t) - f_i'(w^*)\|^2 &\leq \frac{2L}{n}\sum_{i=1}^{n} f_i(w^t) - f_i(w^*) - f_i'(w^*)^\intercal(w^t - w^*) \\ &= 2L(f(w^t) - f(w^*) - f'(w^*)^\intercal(w^t - w^*)).\end{aligned} \tag{A21}$$

Similarly, for $\|f'(w^t) - f'(w^*)\|^2$ we recall f is a μ-strongly convex function:

$$\|f'(w^t) - f'(w^*)\|^2 \geq 2\mu(f(w^t) - f(w^*) - f'(w^*)^\intercal(w^t - w^*)). \tag{A22}$$

Add those inequalities together:

$$\mathbb{E}\|w^{t+1} - w^*\|^2 \leq (1 - \gamma\mu)\|w^t - w^*\|^2 + (4L\gamma^2 - \frac{2\mu\gamma}{L} - 2\mu\gamma^2)f^\delta(w^t) + 2\gamma^2\mathbb{E}\|f_i'(\phi_i^t) - f_i'(w^*)\|^2. \tag{A23}$$

1.8. Proof of Lemma 5

Since we know the distribution of random variable τ, also denote t_s as the index of the latest gradient snapshot so for SVRG/SAGA++ $t_s = kT$ where k is the number of outer iteration and T is the length of inner iteration, for SAGA $t_s = 0$ so in either method we have $t_s \geq 0$ then by conditional expectation relationship:

$$\begin{aligned}\mathbb{E}[\|\alpha_i - f_i'(w^*)\|^2|\mathcal{F}_0] &\overset{(1)}{=} \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[\|\alpha_k - f_k'(w^*)\|^2|\mathcal{F}_{t_s}] \\ &\overset{(2)}{=} \frac{1}{n}\sum_{k=1}^{n}\sum_{l=t_s}^{t} p_l\|f_k'(w^l) - f_k'(w^*)\|^2 \\ &= \sum_{l=t_s}^{t} p_l \frac{1}{n}\sum_{k=1}^{n}\|f_k'(w^l) - f_k'(w^*)\|^2 \\ &\leq 2L\sum_{l=t_s}^{t} p_l(f(w^l) - f(w^*) - f'(w^*)(w^l - w^*)),\end{aligned} \tag{A24}$$

$\overset{(1)}{=}$ is taken over the choices of i, while $\overset{(2)}{=}$ is taken over the random variable τ in $\alpha_k = f_k'(w^\tau)$. Because the regularization function $g(w)$ is convex, and from optimal condition we know: $-f'(w^*) \in \partial g(w^*)$, we have:

$$\begin{aligned}f(w^l) - f(w^*) - f'(w^*)(w^l - w^*) &= f(w^l) - f(w^*) + v^l(w^l - w^*) \\ &\leq f(w^l) - f(w^*) + g(w^l) - g(w^*) \\ &= F(w^l) - F(w^*),\end{aligned} \tag{A25}$$

where $v^l \in \partial g(w^l)$. Finally we have $\mathbb{E}[\|\alpha_i - f_i'(w^*)\|^2|\mathcal{F}_0] \leq 2L\sum_{l=t_s}^{t} p_l(F(w^l) - F(w^*))$.

1.9. Proof of Proposition 3

Recall the quadratic upper bound of L-Lipschitz function:

$$f(w^t - \gamma G(w^t)) \leq f(w^t) - \gamma \nabla f^{\intercal}(w^t) G(w^t) + \frac{L\gamma^2}{2}\|G(w^t)\|^2. \tag{A26}$$

By taking the expectation,

$$\begin{aligned}\mathbb{E}[f(w^t - \gamma G(w^t))|\mathcal{F}_t] &\leq f(w^t) - \gamma\|f(w^t)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}[\|G(w^t)\|^2|\mathcal{F}_t] \\ &\leq f(w^t) - (\gamma - \frac{L\gamma^2}{2})\|\nabla f(w^t)\|^2 + \frac{L\gamma^2}{2}\mathsf{Var}[G(w^t)].\end{aligned} \tag{A27}$$

On the other hand, for μ-strongly convex f, we have:

$$\|\nabla f(w^t)\|^2 \geq 2\mu(f(w^t) - f^*), \tag{A28}$$

so if $\mathsf{Var}[G(w^t)]$ also converges to zero at the order of $f^{\text{sub}}(w^t) = f(w^t) - f^*$ then γ can keep to a small constant rather than damping like SGD. In fact (**?**)(Corollary 3) already proved it for SVRG, here we prove a similar result for SAGA style update:

$$\begin{aligned}\mathsf{Var}[G(w^t)|\mathcal{F}_s] &= \mathbb{E}\Big[\Big\|\nabla f_{i_k}(w^t) - \nabla f_{i_k}(\phi^t_{i_k}) - \frac{1}{n}\sum_{j=1}^n \big(\nabla f_j(w^t) - \nabla f_j(\phi^t_j)\big)\Big\|^2\Big|\mathcal{F}_s\Big] \\ &= \mathbb{E}\Big[\Big\|\nabla f_{i_k}(w^t) - \nabla f_{i_k}(\phi^t_{i_k})\Big\|^2\Big|\mathcal{F}_s\Big] - \Big\|\frac{1}{n}\sum_{j=1}^n \big(\nabla f_j(w^t) - \nabla f_j(\phi^t_j)\big)\Big\|^2 \\ &\leq \mathbb{E}\Big[\mathbb{E}\Big[\Big\|\nabla f_{i_k}(w^t) - \nabla f_{i_k}(\phi^t_{i_k})\Big\|^2\Big|\mathcal{F}_t\Big|\mathcal{F}_s\Big]\Big] \\ &= \frac{2}{n}\sum_{j=1}^n \mathbb{E}\big[\big\|\nabla f_j(w^t) - \nabla f_j(w^*)\big\|^2|\mathcal{F}_s\big] + \frac{2}{n}\sum_{j=1}^n \mathbb{E}\big[\big\|\nabla f_j(\phi^t_j) - \nabla f_j(w^*)\big\|^2|\mathcal{F}_s\big] \\ &\leq 4L\big(\mathbb{E}[f(w^t)|\mathcal{F}_s] - f(w^*)\big) + \frac{4L}{n}\sum_{j=1}^n\sum_{\tau=s}^t p_\tau\big(\mathbb{E}[f_j(w_\tau)|\mathcal{F}_s] - f_j(w^*)\big) \\ &= 4L\big(\mathbb{E}[f(w^t)|\mathcal{F}_s] - f(w^*)\big) + 4L\sum_{\tau=s}^t p_\tau\big(\mathbb{E}[f(w_\tau)|\mathcal{F}_s] - f(w^*)\big),\end{aligned} \tag{A29}$$

here $\{\mathcal{F}_t\}_{t\geq 0}$ is the filtered probability space, $t - T \leq s \leq t$ (recall T is the length of inner iteration) is the latest available full gradient time stamp, p_τ is the probability distribution of stored gradient discussed in (10). Since $t-s$ is upper bounded (this is true for SVRG/SAGA++, as to SAGA, the expectation is $n\log n$ by "*Coupon collection problem*"), together with linear convergence, we know the second term is close to the first term up to a constant.

1.10. Proof of Theorem 6

First of all, we have the following recursive formula:

$$\begin{aligned}P_g(x,\eta,c,n) &= \mathsf{Prox}_g(P_g(x,\eta,c,n-1) - c) \\ &= \begin{cases} P(x,\eta,c,n-1) - c - \eta, & \text{if } P(x,\eta,c,n-1) \geq c + \eta \\ 0, & \text{if } c - \eta \leq P(x,\eta,c,n-1) \leq c + \eta \\ P(x,\eta,c,n-1) - c + \eta, & \text{if } P(x,\eta,c,n-1) \leq c - \eta \end{cases}.\end{aligned} \tag{A30}$$

Because c can be either positive or negative but η is always positive, we consider about following cases:

- ($c < -\eta$) In this case $0 > c + \eta > c - \eta$, if:

1. $x \geq c + \eta$, then $P(x, \eta, c, n) = x - n(c + \eta)$;
2. $x < c + \eta$, then suppose $x = q(c - \eta) + \epsilon$, $q \in \mathbb{N}$, $\epsilon \in [c - \eta, c + \eta]$, if $q \geq n$ then $P(x, \eta, c, n) = x - n(c - \eta)$; else $P(x, \eta, c, q) = \epsilon$, $P(x, \eta, c, q + 1) = 0$, $P(x, \eta, c, n) = -(n - q - 1)(c + \eta)$.

- $(c > \eta)$ In this case $0 < c - \eta < c + \eta$ which is symmetric to previous case, if:
  1. $x \leq c - \eta$, then $P(x, \eta, c, n) = x - n(c - \eta)$;
  2. $x > c - \eta$, then suppose $x = q(c + \eta) + \epsilon$, $q \in \mathbb{N}$, $\epsilon \in [c - \eta, c + \eta]$, if $q \geq n$ then $P(x, \eta, c, n) = x - n(c - \eta)$; else $P(x, \eta, c, q) = \epsilon$, $P(x, \eta, c, q + 1) = 0$, $P(x, \eta, c, n) = -(n - q - 1)(c - \eta)$.
- $(-\eta \leq c \leq \eta)$ finally, $c - \eta \leq 0 \leq c + \eta$, if:
  1. $x \geq n(c + \eta)$, then $P(x, \eta, c, n) = x - n(c + \eta)$;
  2. $x \leq n(c - \eta)$, then $P(x, \eta, c, n) = x + n(c - \eta)$;
  3. otherwise, $\lfloor \frac{x}{c+\eta} \rfloor < n$ or $\lfloor \frac{-x}{-c+\eta} \rfloor < n$ then we know it will eventually be zero: $P(x, \eta, c, n) = 0$.

Clearly this is a piecewise linear function with tangent either 1 or 0.

1.11. ℓ_2 Logistic Regression Experiment

In this supplemental experiment, we conduct the ℓ_2 logistic regression experiment, formulated as follows

$$w^* = \arg\min_w \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(y_i x_i^\intercal w)\right) + \frac{\lambda}{2} \|w\|_2^2. \tag{A31}$$

The datasets and settings are the same as ℓ_1 experiment discussed in the main text. The experiment result is exhibited in Figure 1.

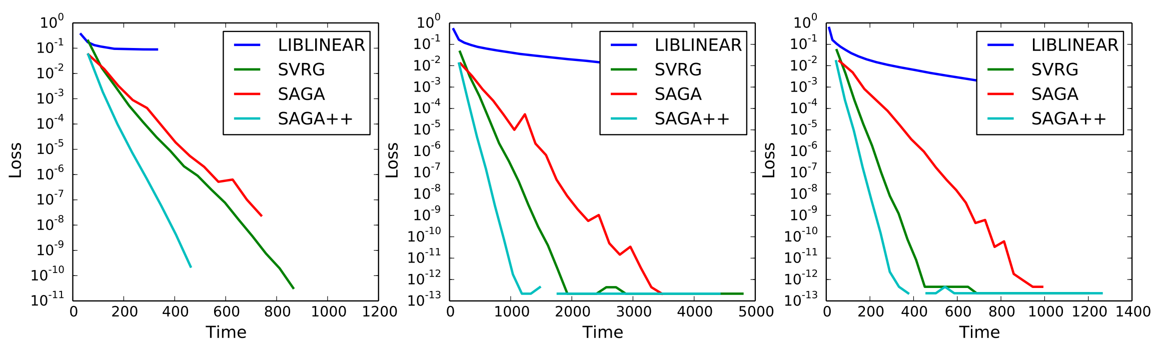

Figure 1. Running time comparison among different data ($\lambda = 1.0 \times 10^{-7}$ for all data).