Supplementary Material for "Direct Acceleration of SAGA using Sampled Negative Momentum"

A Proof of Lemma 1

Lemma 1 is technically similar to Lemma 3.4 in [Allen-Zhu, 2017], but since they are not exactly the same, we include a proof here.

$$\begin{aligned}\mathbb{E}_{i_k}\left[\left\|\widetilde{\nabla}_k-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^k)\right\|^2\right]&=\mathbb{E}_{i_k}\left[\left\|\left(\nabla f_{i_k}(y_{i_k}^k)-\nabla f_{i_k}(\phi_{i_k}^k)\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(y_i^k)-\nabla f_i(\phi_i^k)\right)\right\|^2\right]\\&\overset{(a)}{\leq}\mathbb{E}_{i_k}\left[\left\|\nabla f_{i_k}(y_{i_k}^k)-\nabla f_{i_k}(\phi_{i_k}^k)\right\|^2\right]\\&\overset{(b)}{\leq}2L\cdot\mathbb{E}_{i_k}\left[f_{i_k}(\phi_{i_k}^k)-f_{i_k}(y_{i_k}^k)-\left\langle\nabla f_{i_k}(y_{i_k}^k),\phi_{i_k}^k-y_{i_k}^k\right\rangle\right]\\&=2L\left(\frac{1}{n}\sum_{i=1}^{n}\left(f_i(\phi_i^k)-f(y_i^k)\right)-\frac{1}{n}\sum_{i=1}^{n}\left\langle\nabla f_i(y_i^k),\phi_i^k-y_i^k\right\rangle\right),\end{aligned}$$

where (a) follows from $\mathbb{E}\left[\|\zeta-\mathbb{E}\zeta\|^2\right]\leq\mathbb{E}\|\zeta\|^2$ and (b) uses Theorem 2.1.5 in [Nesterov, 2004].

B Proof of Theorem 1

The proof of Theorem 1 combines the ideas in SAGA [Defazio et al., 2014], Katyusha [Allen-Zhu, 2017] and [Zhou et al., 2018].

In order to prove Theorem 1, we need the following useful lemma, which can be regarded as using the 3-point equality of Bregman divergence in the Euclidean norm setting:

Lemma 3. *If two vectors x_{k+1}, $x_k\in\mathbb{R}^d$ satisfy $x_{k+1}=\arg\min_x\{h(x)+\langle\widetilde{\nabla}_k,x\rangle+\frac{1}{2\eta}\|x_k-x\|^2\}$ with a constant vector $\widetilde{\nabla}_k$ and a μ-strongly convex function $h(\cdot)$, then for all $u\in\mathbb{R}^d$, we have*

$$\langle\widetilde{\nabla}_k,x_{k+1}-u\rangle\leq-\frac{1}{2\eta}\|x_{k+1}-x_k\|^2+\frac{1}{2\eta}\|x_k-u\|^2-\frac{1+\eta\mu}{2\eta}\|x_{k+1}-u\|^2+h(u)-h(x_{k+1}).$$

This Lemma is identical to Lemma 3.5 in [Allen-Zhu, 2017], and hence the proof is omitted.

First, we analyze Algorithm 1 at the kth iteration, given that the randomness from previous iterations are fixed.

We start with the convexity of $f_{i_k}(\cdot)$ at $(y_{i_k}^k,x^\star)$. By definition, we have

$$\begin{aligned}f_{i_k}(y_{i_k}^k)-f_{i_k}(x^\star)&\leq\langle\nabla f_{i_k}(y_{i_k}^k),y_{i_k}^k-x^\star\rangle\\&\overset{(\star)}{=}\frac{1-\tau}{\tau}\langle\nabla f_{i_k}(y_{i_k}^k),\phi_{i_k}^k-y_{i_k}^k\rangle+\langle\nabla f_{i_k}(y_{i_k}^k)-\widetilde{\nabla}_k,x_k-x^\star\rangle+\langle\widetilde{\nabla}_k,x_k-x_{k+1}\rangle\\&\quad+\langle\widetilde{\nabla}_k,x_{k+1}-x^\star\rangle,\end{aligned}$$

where $(\star)$ uses the definition of the i_kth entry of "coupled table" that $y_{i_k}^k=\tau x_k+(1-\tau)\phi_{i_k}^k$.

As we will see, the first term on the right side is used to cancel the unwanted inner product term in the variance bound.

By taking expectation with respect to sample i_k and using the unbiasedness that $\mathbb{E}_{i_k}\big[\nabla f_{i_k}(y_{i_k}^k) - \widetilde{\nabla}_k\big] = \mathbf{0}$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} f_i(y_i^k) - f(x^\star) \leq \frac{1-\tau}{\tau n}\sum_{i=1}^{n}\langle \nabla f_i(y_i^k), \phi_i^k - y_i^k\rangle + \mathbb{E}_{i_k}\big[\langle \widetilde{\nabla}_k, x_k - x_{k+1}\rangle\big] + \mathbb{E}_{i_k}\big[\langle \widetilde{\nabla}_k, x_{k+1} - x^\star\rangle\big]. \quad (4)$$

In order to bound $\mathbb{E}_{i_k}\big[\langle \widetilde{\nabla}_k, x_k - x_{k+1}\rangle\big]$, we use the L-smoothness of $f_{I_k}(\cdot)$ at $(\phi_{I_k}^{k+1}, y_{I_k}^k)$, which is

$$f_{I_k}(\phi_{I_k}^{k+1}) - f_{I_k}(y_{I_k}^k) \leq \langle \nabla f_{I_k}(y_{I_k}^k), \phi_{I_k}^{k+1} - y_{I_k}^k\rangle + \frac{L}{2}\|\phi_{I_k}^{k+1} - y_{I_k}^k\|^2.$$

Taking expectation with respect to sample I_k and using our choice of $\phi_{I_k}^{k+1} = \tau x_{k+1} + (1-\tau)\phi_{I_k}^k$ as well as the definition of "coupled table", we conclude that

$$\mathbb{E}_{I_k}\big[f_{I_k}(\phi_{I_k}^{k+1})\big] - \frac{1}{n}\sum_{i=1}^{n} f_i(y_i^k) \leq \tau\Big\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^k), x_{k+1} - x_k\Big\rangle + \frac{L\tau^2}{2}\|x_{k+1} - x_k\|^2,$$

$$\langle \widetilde{\nabla}_k, x_k - x_{k+1}\rangle \leq \frac{1}{\tau n}\sum_{i=1}^{n} f_i(y_i^k) - \frac{1}{\tau}\mathbb{E}_{I_k}\big[f_{I_k}(\phi_{I_k}^{k+1})\big] + \Big\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^k) - \widetilde{\nabla}_k, x_{k+1} - x_k\Big\rangle + \frac{L\tau}{2}\|x_{k+1} - x_k\|^2.$$

Here we see the effect of the independent sample I_k. It decouples the randomness of x_{k+1} and the update position so as to make the above inequalities valid.

Taking expectation with respect to sample i_k, we obtain

$$\begin{aligned}\mathbb{E}_{i_k}\big[\langle \widetilde{\nabla}_k, x_k - x_{k+1}\rangle\big] \leq{}& \frac{1}{\tau n}\sum_{i=1}^{n} f_i(y_i^k) - \frac{1}{\tau}\mathbb{E}_{i_k, I_k}\big[f_{I_k}(\phi_{I_k}^{k+1})\big] + \mathbb{E}_{i_k}\Big[\Big\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^k) - \widetilde{\nabla}_k, x_{k+1} - x_k\Big\rangle\Big] \\ &+ \frac{L\tau}{2}\mathbb{E}_{i_k}\big[\|x_{k+1} - x_k\|^2\big]. \end{aligned} \quad (5)$$

By upper bounding (4) using (5) and Lemma 3 (with $h(\cdot)$ μ-strongly convex and $u = x^\star$), we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f_i(y_i^k) - f(x^\star) \leq{}& \frac{1-\tau}{\tau n}\sum_{i=1}^{n}\langle \nabla f_i(y_i^k), \phi_i^k - y_i^k\rangle + \frac{1}{\tau n}\sum_{i=1}^{n} f_i(y_i^k) - \frac{1}{\tau}\mathbb{E}_{i_k, I_k}\big[f_{I_k}(\phi_{I_k}^{k+1})\big] \\ &+ \mathbb{E}_{i_k}\Big[\Big\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^k) - \widetilde{\nabla}_k, x_{k+1} - x_k\Big\rangle\Big] + \frac{L\tau}{2}\mathbb{E}_{i_k}\big[\|x_{k+1} - x_k\|^2\big] \\ &- \frac{1}{2\eta}\mathbb{E}_{i_k}\big[\|x_{k+1} - x_k\|^2\big] + \frac{1}{2\eta}\|x_k - x^\star\|^2 - \frac{1+\eta\mu}{2\eta}\mathbb{E}_{i_k}\big[\|x_{k+1} - x^\star\|^2\big] \\ &+ h(x^\star) - \mathbb{E}_{i_k}\big[h(x_{k+1})\big].\end{aligned}$$

Here we add a constraint that $L\tau \leq \frac{1}{\eta} - \frac{L\tau}{1-\tau}$, which is identical to the one used in [Zhou et al., 2018]. Using Young's inequality $\langle a, b\rangle \leq \frac{1}{2\beta}\|a\|^2 + \frac{\beta}{2}\|b\|^2$ to upper bound $\mathbb{E}_{i_k}\Big[\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^k) - \widetilde{\nabla}_k, x_{k+1} - x_k\rangle\Big]$ with $\beta = \frac{L\tau}{1-\tau} > 0$, we can simplify the above inequality as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f_i(y_i^k) - f(x^\star) \leq{}& \frac{1-\tau}{\tau n}\sum_{i=1}^{n}\langle \nabla f_i(y_i^k), \phi_i^k - y_i^k\rangle + \frac{1}{\tau n}\sum_{i=1}^{n} f_i(y_i^k) - \frac{1}{\tau}\mathbb{E}_{i_k, I_k}\big[f_{I_k}(\phi_{I_k}^{k+1})\big] \\ &+ \frac{1-\tau}{2L\tau}\mathbb{E}_{i_k}\Big[\Big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^k) - \widetilde{\nabla}_k\Big\|^2\Big] + \frac{1}{2\eta}\|x_k - x^\star\|^2 - \frac{1+\eta\mu}{2\eta}\mathbb{E}_{i_k}\big[\|x_{k+1} - x^\star\|^2\big] \\ &+ h(x^\star) - \mathbb{E}_{i_k}\big[h(x_{k+1})\big].\end{aligned}$$

By applying Lemma 1 to upper bound the variance term, we see that the additional variance term in the variance bound is canceled by the sampled momentum, which gives

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} f_i(y_i^k) - f(x^\star) \leq \frac{1}{\tau n}\sum_{i=1}^{n} f_i(y_i^k) &- \frac{1}{\tau}\mathbb{E}_{i_k,I_k}\big[f_{I_k}(\phi_{I_k}^{k+1})\big] + \frac{1-\tau}{\tau n}\sum_{i=1}^{n}\big(f_i(\phi_i^k) - f(y_i^k)\big) \\
&+ \frac{1}{2\eta}\|x_k - x^\star\|^2 - \frac{1+\eta\mu}{2\eta}\mathbb{E}_{i_k}\big[\|x_{k+1}-x^\star\|^2\big] + h(x^\star) - \mathbb{E}_{i_k}\big[h(x_{k+1})\big], \\
\frac{1}{\tau}\mathbb{E}_{i_k,I_k}\big[f_{I_k}(\phi_{I_k}^{k+1})\big] - F(x^\star) \leq \frac{1-\tau}{\tau n}\sum_{i=1}^{n} f_i(\phi_i^k) &+ \frac{1}{2\eta}\|x_k - x^\star\|^2 - \frac{1+\eta\mu}{2\eta}\mathbb{E}_{i_k}\big[\|x_{k+1}-x^\star\|^2\big] - \mathbb{E}_{i_k}\big[h(x_{k+1})\big].
\end{aligned} \tag{6}
$$

Using the convexity of $h(\cdot)$ and that $\phi_{I_k}^{k+1} = \tau x_{k+1} + (1-\tau)\phi_{I_k}^k$, we have

$$
h(\phi_{I_k}^{k+1}) \leq \tau h(x_{k+1}) + (1-\tau)h(\phi_{I_k}^k).
$$

After taking expectation with respect to sample I_k and sample i_k, we obtain

$$
-\mathbb{E}_{i_k}\big[h(x_{k+1})\big] \leq \frac{1-\tau}{\tau n}\sum_{i=1}^{n} h(\phi_i^k) - \frac{1}{\tau}\mathbb{E}_{i_k,I_k}\big[h(\phi_{I_k}^{k+1})\big].
$$

Combining the above inequality with (6) and using the definition that $F_i(\cdot) = f_i(\cdot) + h(\cdot)$, we can write (6) as

$$
\frac{1}{\tau}\mathbb{E}_{i_k,I_k}\big[F_{I_k}(\phi_{I_k}^{k+1}) - F_{I_k}(x^\star)\big] \leq \frac{1-\tau}{\tau}\Big(\frac{1}{n}\sum_{i=1}^{n} F_i(\phi_i^k) - F(x^\star)\Big) + \frac{1}{2\eta}\|x_k - x^\star\|^2 - \frac{1+\eta\mu}{2\eta}\mathbb{E}_{i_k}\big[\|x_{k+1}-x^\star\|^2\big].
$$

Dividing the above inequality by n and adding both sides by $\frac{1}{\tau n}\mathbb{E}_{I_k}\big[\sum_{i\neq I_k}^{n}\big(F_i(\phi_i^k) - F_i(x^\star)\big)\big]$, we obtain

$$
\begin{aligned}
\frac{1}{\tau}\mathbb{E}_{i_k,I_k}\Big[\frac{1}{n}\sum_{i=1}^{n} F_i(\phi_i^{k+1}) - F(x^\star)\Big] &\leq \frac{1-\tau}{\tau n}\Big(\frac{1}{n}\sum_{i=1}^{n}\big(F_i(\phi_i^k) - F_i(x^\star)\big)\Big) + \frac{1}{\tau n}\mathbb{E}_{I_k}\Big[\sum_{i\neq I_k}^{n}\big(F_i(\phi_i^k) - F_i(x^\star)\big)\Big] \\
&\quad + \frac{1}{2\eta n}\|x_k - x^\star\|^2 - \frac{1+\eta\mu}{2\eta n}\mathbb{E}_{i_k}\big[\|x_{k+1}-x^\star\|^2\big] \\
&= \frac{1-\tau}{\tau n}\Big(\frac{1}{n}\sum_{i=1}^{n}\big(F_i(\phi_i^k) - F_i(x^\star)\big)\Big) + \frac{1}{\tau n^2}\sum_{j=1}^{n}\sum_{i\neq j}^{n}\big(F_i(\phi_i^k) - F_i(x^\star)\big) \\
&\quad + \frac{1}{2\eta n}\|x_k - x^\star\|^2 - \frac{1+\eta\mu}{2\eta n}\mathbb{E}_{i_k}\big[\|x_{k+1}-x^\star\|^2\big] \\
&= \frac{1-\frac{\tau}{n}}{\tau}\Big(\frac{1}{n}\sum_{i=1}^{n} F_i(\phi_i^k) - F(x^\star)\Big) + \frac{1}{2\eta n}\|x_k - x^\star\|^2 \\
&\quad - \frac{1+\eta\mu}{2\eta n}\mathbb{E}_{i_k}\big[\|x_{k+1}-x^\star\|^2\big].
\end{aligned} \tag{7}
$$

Since $\frac{1}{n}\sum_{i=1}^{n} F_i(\phi_i^k) - F(x^\star)$ may not be positive, we need to involve the following term in our Lyapunov function:

$$
\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n}\langle\nabla F_i(x^\star), \phi_i^{k+1} - x^\star\rangle &= -\frac{1}{n}\langle\nabla F_{I_k}(x^\star), \phi_{I_k}^{k+1} - x^\star\rangle - \frac{1}{n}\sum_{i\neq I_k}^{n}\langle\nabla F_i(x^\star), \phi_i^k - x^\star\rangle \\
&= -\frac{\tau}{n}\langle\nabla F_{I_k}(x^\star), x_{k+1} - x^\star\rangle + \frac{\tau}{n}\langle\nabla F_{I_k}(x^\star), \phi_{I_k}^k - x^\star\rangle \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\langle\nabla F_i(x^\star), \phi_i^k - x^\star\rangle.
\end{aligned}
$$

After taking expectation with respect to sample I_k and i_k, we obtain

$$\mathbb{E}_{i_k,I_k}\Big[-\frac{1}{n}\sum_{i=1}^{n}\langle\nabla F_i(x^\star),\phi_i^{k+1}-x^\star\rangle\Big]=-\Big(1-\frac{\tau}{n}\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\langle\nabla F_i(x^\star),\phi_i^{k}-x^\star\rangle\Big). \tag{8}$$

In order to give a clean proof, we denote $D_k \triangleq \frac{1}{n}\sum_{i=1}^{n}F_i(\phi_i^k)-F(x^\star)-\frac{1}{n}\sum_{i=1}^{n}\langle\nabla F_i(x^\star),\phi_i^k-x^\star\rangle$ and $P_k\triangleq \|x_k-x^\star\|^2$, then by combining (7), (8), we can write the contraction as

$$\frac{1}{\tau}\mathbb{E}_{i_k,I_k}\big[D_{k+1}\big]+\frac{1+\eta\mu}{2\eta n}\mathbb{E}_{i_k}\big[P_{k+1}\big]\le\frac{1-\frac{\tau}{n}}{\tau}D_k+\frac{1}{2\eta n}P_k. \tag{9}$$

Case I: Consider the first case with $\frac{n}{\kappa}\le\frac{3}{4}$, choosing $\eta=\sqrt{\frac{1}{3\mu nL}}$ and $\tau=\frac{n\eta\mu}{1+\eta\mu}=\frac{\sqrt{\frac{n}{3\kappa}}}{1+\sqrt{\frac{1}{3n\kappa}}}<\frac{1}{2}$, we first evaluate the parameter constraint:

$$L\tau\le\frac{1}{\eta}-\frac{L\tau}{1-\tau}\Rightarrow\underbrace{\frac{2-\tau}{1-\tau}}_{<3}\cdot\underbrace{\frac{\sqrt{\frac{n}{3\kappa}}}{1+\sqrt{\frac{1}{3n\kappa}}}}_{\le\sqrt{\frac{n}{3\kappa}}}\le\sqrt{\frac{3n}{\kappa}},$$

which means that the constraint is satisfied by our parameter choices.

Moreover, with this choice of τ, we have

$$\frac{1}{\tau(1+\eta\mu)}=\frac{1-\frac{\tau}{n}}{\tau}=\frac{1}{n\eta\mu}.$$

Thus, the contraction (9) can be written as

$$\frac{1}{n\eta\mu}\mathbb{E}_{i_k,I_k}\big[D_{k+1}\big]+\frac{1}{2\eta n}\mathbb{E}_{i_k}\big[P_{k+1}\big]\le(1+\eta\mu)^{-1}\cdot\Big(\frac{1}{n\eta\mu}D_k+\frac{1}{2\eta n}P_k\Big).$$

After telescoping the above contraction from $k=1\dots K$ and taking expectation with respect to all randomness, we have

$$\frac{1}{n\eta\mu}\mathbb{E}\big[D_{K+1}\big]+\frac{1}{2\eta n}\mathbb{E}\big[P_{K+1}\big]\le(1+\eta\mu)^{-K}\cdot\Big(\frac{1}{n\eta\mu}D_1+\frac{1}{2\eta n}P_1\Big).$$

Note that $D_1=F(x_1)-F(x^\star)$ and $\mathbb{E}\big[D_{K+1}\big]\ge 0$ based on convexity. After substituting the parameter choices, we have

$$\mathbb{E}\big[\|x_{K+1}-x^\star\|^2\big]\le\Big(1+\sqrt{\frac{1}{3n\kappa}}\Big)^{-K}\cdot\Big(\frac{2}{\mu}\big(F(x_1)-F(x^\star)\big)+\|x_1-x^\star\|^2\Big).$$

Case II: Consider another case with $\frac{n}{\kappa}>\frac{3}{4}$, choosing $\eta=\frac{1}{2\mu n}$, $\tau=\frac{n\eta\mu}{1+\eta\mu}=\frac{\frac{1}{2}}{1+\frac{1}{2n}}<\frac{1}{2}$. Again, we first evaluate the constraint:

$$L\tau\le\frac{1}{\eta}-\frac{L\tau}{1-\tau}\Rightarrow\tau\cdot\underbrace{\frac{2-\tau}{1-\tau}}_{<3}<\frac{3}{2}<\frac{2n}{\kappa}.$$

Then by rewriting the contraction (9), telescoping from $k=1\dots K$ and taking expectation with respect to all randomness, we obtain

$$2\mathbb{E}\big[D_{K+1}\big]+\frac{1}{2\eta n}\mathbb{E}\big[P_{K+1}\big]\le(1+\eta\mu)^{-K}\cdot\Big(2D_1+\frac{1}{2\eta n}P_1\Big).$$

By substituting the parameter choices, we have

$$\mathbb{E}\big[\|x_{K+1}-x^\star\|^2\big]\le\Big(1+\frac{1}{2n}\Big)^{-K}\cdot\Big(\frac{2}{\mu}\big(F(x_1)-F(x^\star)\big)+\|x_1-x^\star\|^2\Big).$$

C About the Lyapunov functions for SAGA and SVRG

The Lyapunov functions used to prove the convergence of SAGA (and SSNM) and SVRG (and its variants):

$$\text{SAGA:}\ \ \frac{1}{n}\sum_{i=1}^{n}F_i(\phi_i)-F(x^\star)-\frac{1}{n}\sum_{i=1}^{n}\langle\nabla F_i(x^\star),\phi_i-x^\star\rangle+c_1\|x-x^\star\|^2 \tag{10}$$

$$\text{SVRG:}\ \ F(\tilde{x})-F(x^\star)+c_2\|x-x^\star\|^2, \tag{11}$$

where c_1 and c_2 are constants. Thus, the convergence of SAGA (and SSNM) is built with respect to $\|x-x^\star\|^2$ and that of SVRG (and its variants) is built with respect to $F(\tilde{x})-F(x^\star)$. If $h\equiv 0$ in Problem (1), $F(x)-F(x^\star)$ and $\|x-x^\star\|^2$ only have a constant difference. However, when $h\not\equiv 0$, we only have $(\mu/2)\|x-x^\star\|^2\leq F(x)-F(x^\star)$. For SAGA (and SSNM), this subtle difference prevents us from using techniques that involve restart (e.g., AdaptSmooth, APPA, Catalyst). In the case where $h\equiv 0$, we can use them but an additional $\log(L/\mu)$ factor will appear in the rate. This difference somehow explains why the SVRG-like variance reduction technique is more favorable in theory than that of SAGA.

D Experimental setup in Section 6

All the algorithms were implemented in C++ and executed through a MATLAB interface for fair comparison. We ran experiments on an HP Z440 machine with a single Intel Xeon E5-1630v4 with 3.70GHz cores, 16GB RAM, Ubuntu 16.04 LTS with GCC 4.9.0, MATLAB R2017b.

We are optimizing the following binary problem with $a_i\in\mathbb{R}^d$, $b_i\in\{-1,+1\}$, $i=1\ldots m$:

$$\ell_2\text{-Logistic Regression:}\ \ \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp\left(-b_i a_i^T x\right)\right)+\frac{\lambda}{2}\|x\|^2,$$

where λ is the regularization parameter and all the datasets used were normalized before the experiments.

The parameter settings used in the experiments:

- SAGA. We set the learning rate as $\frac{1}{2(\mu n+L)}$, which is analyzed theoretically in [Defazio et al., 2014].
- SSNM. We used the same settings as suggested in Algorithm 1, which are $\eta=\sqrt{\frac{1}{3\mu n L}}$ and $\tau=\frac{n\eta\mu}{1+\eta\mu}$.
- Katyusha. As suggested by the author, we fixed $\tau_2=\frac{1}{2}$, set $\eta=\frac{1}{3\tau_1 L}$ and chose $\tau_1=\sqrt{\frac{m}{3\kappa}}$ [Allen-Zhu, 2017] (In the notations of the original work).
- MiG. We set $\eta=\frac{1}{3\theta L}$ and chose $\theta=\sqrt{\frac{m}{3\kappa}}$ as analyzed in [Zhou et al., 2018].

E An empirical comparison with Point-SAGA

Here we report an experiment comparing the performance of SAGA, Point-SAGA and SSNM with respect to iteration counter. The detailed experimental setting is given in Section 6 in the main paper. Since Point-SAGA requires the exact proximal operator of each $F_i(\cdot)$ in theory, we focus on training ridge regression in this section:

$$\textit{Ridge Regression:}\ \ \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(a_i^T x+b_i)^2+\frac{\lambda}{2}\|x\|^2.$$

Note that the proximal operator of each $F_i(\cdot)=\frac{1}{2}(a_i^T x+b_i)^2+\frac{\lambda}{2}\|x\|^2$ can be efficiently computed as mentioned in [Defazio, 2016].

A memory issue of Point-SAGA: In fact, when we involve an $\ell 2$-regularizer in each $F_i(\cdot)$ [11], we cannot use the trick of representing a gradient by a scalar since the update equation of the new table entry g_j^{k+1} (in original notations) contains a term that correlates to the weight x_k, which leads to an $O(nd)$ memory complexity. A possible solution is to separate the proximal computations for the component functions and the regularizer, but it does not fit in the analysis of Point-SAGA.

[11] An $\ell 2$-regularizer is always the source of strong convexity for real world problems.

We used the same parameter settings for SAGA and SSNM as in Section 6 in the main paper. For Point-SAGA, we chose the learning rate γ suggested by the original work [Defazio, 2016],

$$\gamma = \frac{\sqrt{(n-1)^2 + 4n\frac{L}{\mu}}}{2Ln} - \frac{1-\frac{1}{n}}{2L}.$$

The result is shown in Figure 3. As we can see, the convergence rates of Point-SAGA and SSNM are quite similar and consistently faster than SAGA. Although Point-SAGA is shown to be slightly faster than SSNM in this experiment, considering the general objective assumption and the memory issue of Point-SAGA mentioned above, SSNM is a more favorable accelerated variant of SAGA than Point-SAGA in practice. Interestingly, both accelerated variants are more unstable than SAGA in this experiment.

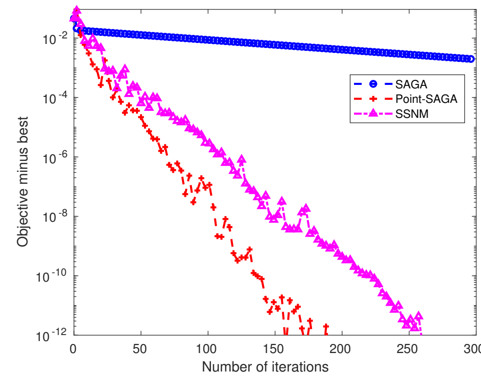

Figure 3: Comparison of SAGA, Point-SAGA and SSNM for solving ridge regression on `covtype` with $\lambda = 10^{-8}$.