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.

\[
E_{i_k} \left[ \left\| \bar{\nabla}_k - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(y_i^k) \right\|^2 \right] = E_{i_k} \left[ \left\| \left( \nabla f_{i_k}(y_i^k) - \nabla f_{i_k}(\phi_i^k) \right) - \frac{1}{n} \sum_{i=1}^{n} \left( \nabla f_i(y_i^k) - \nabla f_i(\phi_i^k) \right) \right\|^2 \right].
\]

\[
\leq E_{i_k} \left[ \left\| \nabla f_{i_k}(y_i^k) - \nabla f_{i_k}(\phi_i^k) \right\|^2 \right]
\]

\[
\leq 2L \cdot E_{i_k} \left[ f_{i_k}(\phi_i^k) - f_{i_k}(y_i^k) - \langle \nabla f_{i_k}(y_i^k), \phi_i^k - y_i^k \rangle \right]
\]

\[
= 2L \left( \frac{1}{n} \sum_{i=1}^{n} (f_i(\phi_i^k) - f(y_i^k)) - \frac{1}{n} \sum_{i=1}^{n} \langle \nabla f_i(y_i^k), \phi_i^k - y_i^k \rangle \right),
\]

where \((a)\) follows from \(E[\|\zeta - \mathcal{E}\zeta\|^2] \leq E[\|\zeta\|^2]\) 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 \(x_{k+1} = \arg \min_x \{ h(x) + \langle \bar{\nabla}_k, x \rangle + \frac{1}{2\eta} \| x_k - x \|_2^2 \} \) with a constant vector \(\bar{\nabla}_k\) and a \(\mu\)-strongly convex function \(h(\cdot)\), then for all \(u \in \mathbb{R}^d\), we have

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

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 \(k\)th iteration, given that the randomness from previous iterations are fixed. We start with the convexity of \(f_{i_k}(\cdot)\) at \((y_{i_k}^k, x^*)\). By definition, we have

\[
f_{i_k}(y_{i_k}^k) - f_{i_k}(x^*) \leq \langle \nabla f_{i_k}(y_{i_k}^k), y_{i_k}^k - x^* \rangle
\]

\[
\leq \frac{1}{\tau} \langle \nabla f_{i_k}(y_{i_k}^k), \phi_i^k - y_i^k \rangle + \langle \nabla f_{i_k}(y_{i_k}^k) - \bar{\nabla}_k, x_k - x^* \rangle + \langle \bar{\nabla}_k, x_k - x_{k+1} \rangle
\]

\[
- \langle \bar{\nabla}_k, x_{k+1} - x^* \rangle,
\]

where \((\ast)\) uses the definition of the \(i_k\)th entry of “coupled table” that \(y_{i_k}^k = \tau x_k + (1 - \tau)\phi_i^k\).

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 $\mathbb{E}_{i_k}[\nabla f_{i_k}(y^k_{i_k}) - \tilde{\nabla} k] = 0$, we obtain
\[
\frac{1}{n} \sum_{i=1}^{n} f_i(y^k_i) - f(x^*) \leq \frac{1 - \tau}{\tau n} \sum_{i=1}^{n} \langle \nabla f_i(y^k_i), \phi^k_i - y^k_i \rangle + \frac{1}{\tau} \mathbb{E}_{i_k} [\langle \tilde{\nabla} k, x_k - x_{k+1} \rangle] + \mathbb{E}_{i_k} [\langle \tilde{\nabla} k, x_{k+1} - x^* \rangle].
\] (4)

In order to bound $\mathbb{E}_{i_k} [\langle \tilde{\nabla} k, x_k - x_{k+1} \rangle]$, we use the $L$-smoothness of $f_{i_k} (\cdot)$ at $(\phi^{k+1}_{i_k}, y^k_{i_k})$, which is
\[
f_{i_k}(\phi^{k+1}_{i_k}) - f_{i_k}(y^k_{i_k}) \leq \langle \nabla f_{i_k}(y^k_{i_k}), \phi^{k+1}_{i_k} - y^k_{i_k} \rangle + \frac{L}{2} \| \phi^{k+1}_{i_k} - y^k_{i_k} \|^2.
\]

Taking expectation with respect to sample $I_k$ and using our choice of $\phi^{k+1}_{i_k} = \tau x_{k+1} + (1 - \tau) y^k_{i_k}$ as well as the definition of “coupled table”, we conclude that
\[
\mathbb{E}_{I_k} [f_{i_k}(\phi^{k+1}_{i_k})] - \frac{1}{n} \sum_{i=1}^{n} f_i(y^k_i) \leq \frac{1}{\tau} \mathbb{E}_{I_k} [f_{i_k}(\phi^{k+1}_{i_k})] + \left( \frac{1}{\tau} \sum_{i=1}^{n} \nabla f_i(y^k_i) - \tilde{\nabla} k, x_{k+1} - x_k \right) + \frac{L \tau^2}{2} \| x_{k+1} - x_k \|^2,
\]
\[
\langle \tilde{\nabla} k, x_k - x_{k+1} \rangle \leq \frac{1}{\tau n} \sum_{i=1}^{n} f_i(y^k_i) - \frac{1}{\tau} \mathbb{E}_{I_k} [f_{i_k}(\phi^{k+1}_{i_k})] + \left( \frac{1}{\tau} \sum_{i=1}^{n} \nabla f_i(y^k_i) - \tilde{\nabla} k, x_{k+1} - x_k \right) + \frac{L \tau}{2} \| x_{k+1} - x_k \|^2.
\]

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
\[
\mathbb{E}_{i_k} [\langle \tilde{\nabla} k, x_k - x_{k+1} \rangle] \leq \frac{1}{\tau n} \sum_{i=1}^{n} f_i(y^k_i) - \frac{1}{\tau} \mathbb{E}_{i_k,I_k} [f_{i_k}(\phi^{k+1}_{i_k})] + \mathbb{E}_{i_k} [\langle \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(y^k_i) - \tilde{\nabla} k, x_{k+1} - x_k \rangle]
\]
\[
+ \frac{L \tau}{2} \mathbb{E}_{i_k} [\| x_{k+1} - x_k \|^2].
\] (5)

By upper bounding (4) using (5) and Lemma 3 (with $h(\cdot)$ $\mu$-strongly convex and $u = x^*$), we obtain
\[
\frac{1}{n} \sum_{i=1}^{n} f_i(y^k_i) - f(x^*) \leq \frac{1 - \tau}{\tau n} \sum_{i=1}^{n} \langle \nabla f_i(y^k_i), \phi^k_i - y^k_i \rangle + \frac{1}{\tau} \mathbb{E}_{i_k,I_k} [f_{i_k}(\phi^{k+1}_{i_k})]
\]
\[
+ \mathbb{E}_{i_k} [\langle \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(y^k_i) - \tilde{\nabla} k, x_{k+1} - x_k \rangle] + \frac{L \tau}{2} \mathbb{E}_{i_k} [\| x_{k+1} - x_k \|^2]
\]
\[
- \frac{1}{2 \eta} \mathbb{E}_{i_k} [\| x_{k+1} - x_k \|^2] + \frac{1}{2 \eta} \| x_k - x^* \|^2 - \frac{1 + \eta \mu}{2 \eta} \mathbb{E}_{i_k} [\| x_{k+1} - x^* \|^2]
\]
\[
+ h(x^*) - \mathbb{E}_{i_k} [h(x_{k+1})].
\]

Here we add a constraint that $L \tau \leq \frac{1}{\eta} - \frac{\beta}{L}$, which is identical to the one used in Zhou et al, 2018. Using Young’s inequality $\langle a, b \rangle \leq \frac{\lambda}{2} \| a \|^2 + \frac{\beta}{2} \| b \|^2$ to upper bound $\mathbb{E}_{i_k} [\langle \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(y^k_i) - \tilde{\nabla} k, x_{k+1} - x_k \rangle]$ with $\beta = \frac{L \tau}{1 - \tau} > 0$, we can simplify the above inequality as
\[
\frac{1}{n} \sum_{i=1}^{n} f_i(y^k_i) - f(x^*) \leq \frac{1 - \tau}{\tau n} \sum_{i=1}^{n} \langle \nabla f_i(y^k_i), \phi^k_i - y^k_i \rangle + \frac{1}{\tau} \mathbb{E}_{i_k,I_k} [f_{i_k}(\phi^{k+1}_{i_k})]
\]
\[
+ \frac{1 - \tau}{2 L \tau} \mathbb{E}_{i_k} [\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(y^k_i) - \tilde{\nabla} k \|^2] + \frac{1}{2 \eta} \| x_k - x^* \|^2 - \frac{1 + \eta \mu}{2 \eta} \mathbb{E}_{i_k} [\| x_{k+1} - x^* \|^2]
\]
\[
+ h(x^*) - \mathbb{E}_{i_k} [h(x_{k+1})].
\]
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

\[
\frac{1}{n} \sum_{i=1}^{n} f_i(y_i^k) - f(x^*) \leq \frac{1}{\tau n} \sum_{i=1}^{n} f_i(y_i^k) - \frac{1}{\tau} \mathbb{E}_{i_k, I_k} [f_{i_k} (\phi_{i_k}^{k+1})] + \frac{1}{\tau n} \sum_{i=1}^{n} \left( f_i (\phi_i^k) - f(y_i^k) \right) + \frac{1}{\tau n} \left\| x_k - x^* \right\|^2 - \frac{1+\eta \mu}{2\eta} \mathbb{E}_{i_k} \left[ \left\| x_{k+1} - x^* \right\|^2 \right] + h(x^*) - \mathbb{E}_{i_k} [h(x_{k+1})],
\]

\[
\frac{1}{\tau} \mathbb{E}_{i_k, I_k} [f_{i_k} (\phi_{i_k}^{k+1})] - F(x^*) \leq \frac{1-\tau}{\tau n} \sum_{i=1}^{n} f_i (\phi_i^k) + \frac{1}{\tau n} \left\| x_k - x^* \right\|^2 - \frac{1+\eta \mu}{2\eta} \mathbb{E}_{i_k} \left[ \left\| x_{k+1} - x^* \right\|^2 \right] - \mathbb{E}_{i_k} [h(x_{k+1})].
\]

(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

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

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

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

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

\[
\frac{1}{\tau} \mathbb{E}_{i_k, I_k} \left[ \frac{1}{n} \sum_{i=1}^{n} F_i (\phi_i^{k+1}) - F(x^*) \right] \leq \frac{1-\tau}{\tau n} \left( \frac{1}{n} \sum_{i=1}^{n} (F_i (\phi_i^k) - F_i (x^*)) \right) + \frac{1}{\tau n} \mathbb{E}_{i_k} \left[ \sum_{i \neq i_k} (F_i (\phi_i^k) - F_i (x^*)) \right] + \frac{1}{\tau n} \left\| x_k - x^* \right\|^2 - \frac{1+\eta \mu}{2\eta} \mathbb{E}_{i_k} \left[ \left\| x_{k+1} - x^* \right\|^2 \right]
\]

\[
= \frac{1-\tau}{\tau n} \left( \frac{1}{n} \sum_{i=1}^{n} (F_i (\phi_i^k) - F_i (x^*)) \right) + \frac{1}{\tau n^2} \sum_{j=1}^{n} \sum_{i \neq j} (F_j (\phi_j^k) - F_j (x^*)) + \frac{1}{\tau n} \left\| x_k - x^* \right\|^2 - \frac{1+\eta \mu}{2\eta} \mathbb{E}_{i_k} \left[ \left\| x_{k+1} - x^* \right\|^2 \right]
\]

\[
= \frac{1-\tau}{\tau} \left( \frac{1}{n} \sum_{i=1}^{n} F_i (\phi_i^k) - F(x^*) \right) + \frac{1}{2\eta n} \left\| x_k - x^* \right\|^2 - \frac{1+\eta \mu}{2\eta} \mathbb{E}_{i_k} \left[ \left\| x_{k+1} - x^* \right\|^2 \right].
\]

(7)

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

\[
-\frac{1}{n} \sum_{i=1}^{n} \langle \nabla F_i (x^*), \phi_i^{k+1} - x^* \rangle = -\frac{1}{n} \langle \nabla F_{i_k} (x^*), \phi_{i_k}^{k+1} - x^* \rangle - \frac{1}{n} \sum_{i \neq i_k} \langle \nabla F_i (x^*), \phi_i^k - x^* \rangle
\]

\[
= -\frac{\tau}{n} \langle \nabla F_{i_k} (x^*), x_{k+1} - x^* \rangle + \frac{\tau}{n} \langle \nabla F_{i_k} (x^*), \phi_{i_k}^k - x^* \rangle - \frac{1}{n} \sum_{i=1}^{n} \langle \nabla F_i (x^*), \phi_i^k - x^* \rangle.
\]
After taking expectation with respect to sample $I_k$ and $i_k$, we obtain

$$\mathbb{E}_{i_k, I_k} \left[ - \frac{1}{n} \sum_{i=1}^{n} (\nabla F_i(x^*), \phi_{i}^{k+1} - x^*) \right] = - \left( 1 - \frac{\tau}{n} \right) \left( \frac{1}{n} \sum_{i=1}^{n} (\nabla F_i(x^*), \phi_{i}^{k} - x^*) \right). \quad (8)$$

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

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

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

$$L \tau \leq \frac{1}{\tau} - \frac{L \tau}{1 - \tau} \Rightarrow \frac{2 - \tau}{1 - \tau} \frac{\sqrt{3n}}{1 + \sqrt{3n \kappa}} \leq \sqrt{\frac{3n}{\kappa}},$$

which means that the constraint is satisfied by our parameter choices. Moreover, with this choice of $\tau$, we have

$$\frac{1}{\tau} = \frac{1}{n \eta \mu}.$$

Thus, the contraction (9) can be written as

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

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

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

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

$$\mathbb{E} [\|x_{K+1} - x^*\|^2] \leq \left( 1 + \sqrt{\frac{1}{3n \kappa}} \right)^{-K} \cdot \left( \frac{2}{\mu} (F(x_1) - F(x^*)) + \|x_1 - x^*\|^2 \right).$$

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

$$L \tau \leq \frac{1}{\tau} - \frac{L \tau}{1 - \tau} \Rightarrow \frac{2 - \tau}{1 - \tau} < \frac{2}{\kappa}.$$ 

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

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

By substituting the parameter choices, we have

$$\mathbb{E} [\|x_{K+1} - x^*\|^2] \leq \left( 1 + \frac{1}{2 \eta n} \right)^{-K} \cdot \left( \frac{2}{\mu} (F(x_1) - F(x^*)) + \|x_1 - x^*\|^2 \right).$$
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):

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

\[
\text{SVRG: } F(\bar{x}) - F(x^*) + c_2 \|x - x^*\|^2,
\]

where \(c_1\) and \(c_2\) are constants. Thus, the convergence of SAGA (and SSNM) is built with respect to \(\|x - x^*\|^2\) and that of SVRG (and its variants) is built with respect to \(F(\bar{x}) - F(x^*)\). If \(h \equiv 0\) in Problem (1), \(F(x) - F(x^*)\) and \(\|x - x^*\|^2\) only have a constant difference. However, when \(h \neq 0\), we only have \((\mu/2)\|x - x^*\|^2 \leq F(x) - F(x^*)\). 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 \(a_i \in \mathbb{R}^d, b_i \in \{-1, +1\}, i = 1 \ldots m\):

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

where \(\lambda\) 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 \(\frac{1}{2(\mu + L)}\), which is analyzed theoretically in [Defazio et al., 2014].
- **SSNM.** We used the same settings as suggested in Algorithm 1 in the original work.
- **Katyusha.** As suggested by the author, we fixed \(\tau_2 = \frac{1}{2}\), set \(\eta = \frac{1}{3 \tau_1 L}\), and chose \(\tau_1 = \sqrt{\frac{\mu}{3 \eta L}}\) [Allen-Zhu, 2017].
- **MiG.** We set \(\eta = \frac{1}{3 \sigma L}\) and chose \(\theta = \sqrt{\frac{\mu}{5 \sigma L}}\) 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:

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

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 \(F_i(\cdot)\) 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.

\[11\text{An } \ell_2\text{-regularizer is always the source of strong convexity for real world problems.}\]
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 $\gamma$ suggested by the original work [Defazio, 2016],

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

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 $\lambda = 10^{-8}$.  