Supplementary Materials for
"A Simple Stochastic Variance Reduced Algorithm with Fast Convergence Rates"

Notations. We use $\mathbb{E}_{i_j}$ to denote that the expectation is taken with respect to the jth sample in one epoch, while $\mathbb{E}$ means taking expectation with respect to all randomness in one epoch. We use $O(\cdot)$ to denote computational complexity and $\mathcal{O}(\cdot)$ to denote oracle complexity. x^* refers to the solution to Problem (1) (the proofs for Section 3) or Problem (4) (the proofs for Section 4). $\mathcal{S}$ is the total number of epochs to be executed. Boldface number like **0** refers to a vector with all 0.

Note: In order to give a clean proof, we omit the superscripts for iterates in the same epoch s as x_j instead of x_j^s unless otherwise specified.

A. Useful Lemmas

Lemma 1. (Variance Bound) *Suppose each component function f_i is L-smooth, let $\tilde{\nabla} = \nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1}) + \nabla f(\tilde{x}_{s-1})$, which is the approximate gradient used in* MiG. *Then the following inequality holds:*

$$\mathbb{E}_{i_j}\left[\|\nabla f(y_{j-1}) - \tilde{\nabla}\|^2\right] \leq 2L\big(f(\tilde{x}_{s-1}) - f(y_{j-1}) - \langle \nabla f(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle\big).$$

Proof. This lemma is identical to Lemma 3.4 in (Allen-Zhu, 2017), which provides a tighter upper bound on the gradient estimator variance than those in (Johnson & Zhang, 2013; Xiao & Zhang, 2014). □

Lemma 2. (3-points property) *Assume that z^* is an optimal solution to the following problem,*

$$\min_x \frac{\tau}{2}\|x - z_0\|^2 + \psi(x),$$

where $\tau > 0$, and $\psi(\cdot)$ is a convex function (but possibly non-differentiable). Then for all $z \in \mathbb{R}^d$, there exists a vector $\mathcal{G} \in \partial\psi(z^*)$ with

$$\langle \mathcal{G}, z - z^*\rangle = \frac{\tau}{2}\|z^* - z_0\|^2 - \frac{\tau}{2}\|z - z_0\|^2 + \frac{\tau}{2}\|z - z^*\|^2,$$

where $\partial\psi(z^*)$ denotes the sub-differential of $\psi(\cdot)$ at z^*. If $\psi(\cdot)$ is differentiable, we can simply replace $\mathcal{G} \in \partial\psi(\cdot)$ with $\mathcal{G} = \nabla\psi(\cdot)$.

Proof. By the optimality of z^*, there exists a vector $\mathcal{G} \in \partial\psi(z^*)$ (or $\mathcal{G} = \nabla\psi(z^*)$ for differentiable $\psi(\cdot)$) satisfying

$$\tau(z^* - z_0) + \mathcal{G} = \mathbf{0}.$$

Thus for all $z \in \mathbb{R}^d$,

$$\begin{aligned} 0 &= \langle \tau(z^* - z_0) + \mathcal{G}, z^* - z\rangle \\ &= \tau\langle z^* - z_0, z^* - z\rangle + \langle \mathcal{G}, z^* - z\rangle \\ &\overset{(\star)}{=} \frac{\tau}{2}\|z^* - z_0\|^2 - \frac{\tau}{2}\|z - z_0\|^2 + \frac{\tau}{2}\|z - z^*\|^2 + \langle \mathcal{G}, z^* - z\rangle, \end{aligned}$$

where $(\star)$ uses the fact that $\langle a - b, a - c\rangle = \frac{1}{2}\|a - b\|^2 - \frac{1}{2}\|b - c\|^2 + \frac{1}{2}\|a - c\|^2$. □

Lemma 3. *If two vector $x_j,\, x_{j-1} \in \mathbb{R}^d$ satisfy $x_j = \arg\min_x\{\frac{1}{2\eta}\|x - x_{j-1}\|^2 + \langle\tilde{\nabla}, x\rangle + g(x)\}$ with a constant vector $\tilde{\nabla}$ and a general convex function $g(\cdot)$, then for all $u \in \mathbb{R}^d$, we have*

$$\langle\tilde{\nabla}, x_j - u\rangle \le -\frac{1}{2\eta}\|x_{j-1} - x_j\|^2 + \frac{1}{2\eta}\|x_{j-1} - u\|^2 - \frac{1}{2\eta}\|x_j - u\|^2 + g(u) - g(x_j).$$

Moreover, if $g(\cdot)$ is σ-strongly convex, the above inequality becomes

$$\langle\tilde{\nabla}, x_j - u\rangle \le -\frac{1}{2\eta}\|x_{j-1} - x_j\|^2 + \frac{1}{2\eta}\|x_{j-1} - u\|^2 - \frac{1+\eta\sigma}{2\eta}\|x_j - u\|^2 + g(u) - g(x_j).$$

Proof. Applying Lemma 2 with $z = u$, $z_0 = x_{j-1}$, $z^* = x_j$, $\tau = \frac{1}{\eta}$, $\psi(x) = \langle\tilde{\nabla}, x\rangle + g(x)$, there exists a vector $\mathcal{G} \in \partial g(x_j)$ (or $\mathcal{G} = \nabla g(x_j)$ for differentiable $g(\cdot)$) satisfying

$$\langle\tilde{\nabla}, u - x_j\rangle + \langle\mathcal{G}, u - x_j\rangle = \frac{1}{2\eta}\|x_{j-1} - x_j\|^2 - \frac{1}{2\eta}\|x_{j-1} - u\|^2 + \frac{1}{2\eta}\|x_j - u\|^2.$$

Using the convexity of $g(\cdot)$, we get $g(u) - g(x_j) \ge \langle\mathcal{G}, u - x_j\rangle$ by definition. After rearranging, we conclude that

$$\langle\tilde{\nabla}, x_j - u\rangle \le -\frac{1}{2\eta}\|x_{j-1} - x_j\|^2 + \frac{1}{2\eta}\|x_{j-1} - u\|^2 - \frac{1}{2\eta}\|x_j - u\|^2 + g(u) - g(x_j).$$

If $g(\cdot)$ is further σ-strongly convex, we have $g(u) - g(x_j) \ge \langle\mathcal{G}, u - x_j\rangle + \frac{\sigma}{2}\|x_j - u\|^2$ by (3). Similarly, we can write

$$\langle\tilde{\nabla}, x_j - u\rangle \le -\frac{1}{2\eta}\|x_{j-1} - x_j\|^2 + \frac{1}{2\eta}\|x_{j-1} - u\|^2 - \frac{1+\eta\sigma}{2\eta}\|x_j - u\|^2 + g(u) - g(x_j).$$

□

B. Proofs for Section 3

B.1. Proof of Theorem 1

First, we add the following constraint on the parameters η and θ, which is crucial in the proof of Theorem 1:

$$L\theta + \frac{L\theta}{1-\theta} \le \frac{1}{\eta} \text{ , or equivalently } \eta \le \frac{1-\theta}{L\theta(2-\theta)}. \tag{7}$$

We start with convexity of $f(\cdot)$ at y_{j-1}. By definition, for any vector $u \in \mathbb{R}^d$, we have

$$\begin{aligned} f(y_{j-1}) - f(u) &\le \langle\nabla f(y_{j-1}),\ y_{j-1} - u\rangle \\ &= \langle\nabla f(y_{j-1}),\ y_{j-1} - x_{j-1}\rangle + \langle\nabla f(y_{j-1}),\ x_{j-1} - u\rangle \\ &\overset{(\star)}{=} \frac{1-\theta}{\theta}\langle\nabla f(y_{j-1}),\ \tilde{x}_{s-1} - y_{j-1}\rangle + \langle\nabla f(y_{j-1}),\ x_{j-1} - u\rangle, \end{aligned} \tag{8}$$

where ($\star$) follows from the fact that $y_{j-1} = \theta x_{j-1} + (1-\theta)\tilde{x}_{s-1}$.

Then we further expand $\langle\nabla f(y_{j-1}),\ x_{j-1} - u\rangle$ as

$$\langle\nabla f(y_{j-1}),\ x_{j-1} - u\rangle = \langle\nabla f(y_{j-1}) - \tilde{\nabla},\ x_{j-1} - u\rangle + \langle\tilde{\nabla},\ x_{j-1} - x_j\rangle + \langle\tilde{\nabla},\ x_j - u\rangle. \tag{9}$$

Using L-smooth (2) of $f(\cdot)$ at (y_j, y_{j-1}), we get

$$\begin{aligned} f(y_j) - f(y_{j-1}) &\le \langle\nabla f(y_{j-1}),\ y_j - y_{j-1}\rangle + \frac{L}{2}\|y_j - y_{j-1}\|^2 \\ &\overset{(\star)}{=} \theta\langle\nabla f(y_{j-1}),\ x_j - x_{j-1}\rangle + \frac{L\theta^2}{2}\|x_j - x_{j-1}\|^2 \\ &= \theta\big[\langle\nabla f(y_{j-1}) - \tilde{\nabla},\ x_j - x_{j-1}\rangle + \langle\tilde{\nabla},\ x_j - x_{j-1}\rangle\big] + \frac{L\theta^2}{2}\|x_j - x_{j-1}\|^2, \\ \langle\tilde{\nabla}, x_{j-1} - x_j\rangle &\le \frac{1}{\theta}\big(f(y_{j-1}) - f(y_j)\big) + \langle\nabla f(y_{j-1}) - \tilde{\nabla}, x_j - x_{j-1}\rangle + \frac{L\theta}{2}\|x_j - x_{j-1}\|^2, \end{aligned}$$

where $(\star)$ uses the definition of y_{j-1}.

After plugging in the constraint (7), we have

$$\langle\tilde{\nabla},\ x_{j-1}-x_j\rangle \leq \frac{1}{\theta}\big(f(y_{j-1})-f(y_j)\big)+\langle\nabla f(y_{j-1})-\tilde{\nabla},\ x_j-x_{j-1}\rangle+\frac{1}{2\eta}\|x_j-x_{j-1}\|^2-\frac{L\theta}{2(1-\theta)}\|x_j-x_{j-1}\|^2. \quad (10)$$

Then we are ready to combine (8), (9), (10), as well as Lemma 3 (here $g(x)$ is σ-strongly convex by Assumption 1), which gives

$$\begin{aligned} f(y_{j-1})-f(u) \leq{}& \frac{1-\theta}{\theta}\langle\nabla f(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\langle\nabla f(y_{j-1})-\tilde{\nabla},x_j-u\rangle+\frac{1}{\theta}(f(y_{j-1})-f(y_j)) \\ &-\frac{L\theta}{2(1-\theta)}\|x_j-x_{j-1}\|^2+\frac{1}{2\eta}\|x_{j-1}-u\|^2-\frac{1+\eta\sigma}{2\eta}\|x_j-u\|^2+g(u)-g(x_j). \end{aligned}$$

After taking expectation with respect to the sample i_j, we obtain

$$\begin{aligned} f(y_{j-1})-f(u) \overset{(a)}{\leq}{}& \frac{1-\theta}{\theta}\langle\nabla f(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\mathbb{E}_{i_j}\big[\langle\nabla f(y_{j-1})-\tilde{\nabla},x_j-x_{j-1}\rangle\big]+\frac{1}{\theta}\big(f(y_{j-1})-\mathbb{E}_{i_j}\big[f(y_j)\big]\big) \\ &-\frac{L\theta}{2(1-\theta)}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]+\frac{1}{2\eta}\|x_{j-1}-u\|^2-\frac{1+\eta\sigma}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-u\|^2\big]+g(u)-\mathbb{E}_{i_j}\big[g(x_j)\big] \\ \overset{(b)}{\leq}{}& \frac{1-\theta}{\theta}\langle\nabla f(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\frac{1}{2\beta}\mathbb{E}_{i_j}\big[\|\nabla f(y_{j-1})-\tilde{\nabla}\|^2\big]+\frac{\beta}{2}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big] \\ &+\frac{1}{\theta}\big(f(y_{j-1})-\mathbb{E}_{i_j}\big[f(y_j)\big]\big)-\frac{L\theta}{2(1-\theta)}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]+\frac{1}{2\eta}\|x_{j-1}-u\|^2 \\ &-\frac{1+\eta\sigma}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-u\|^2\big]+g(u)-\mathbb{E}_{i_j}\big[g(x_j)\big], \end{aligned}$$

where (a) holds due to the unbiasedness of the gradient estimator $\mathbb{E}_{i_j}\big[\nabla f(y_{j-1})-\tilde{\nabla}\big]=\mathbf{0}$, and (b) uses the Young's inequality to expand $\mathbb{E}_{i_j}\big[\langle\nabla f(y_{j-1})-\tilde{\nabla},x_j-x_{j-1}\rangle\big]$ with the parameter $\beta>0$.

Applying Lemma 1 to bound the variance term $\mathbb{E}_{i_j}\big[\|\nabla f(y_{j-1})-\tilde{\nabla}\|^2\big]$, we get

$$\begin{aligned} f(y_{j-1})-f(u) \leq{}& \frac{1-\theta}{\theta}\langle\nabla f(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\frac{L}{\beta}\big(f(\tilde{x}_{s-1})-f(y_{j-1})-\langle\nabla f(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle\big) \\ &+\frac{\beta}{2}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]+\frac{1}{\theta}\big(f(y_{j-1})-\mathbb{E}_{i_j}\big[f(y_j)\big]\big)-\frac{L\theta}{2(1-\theta)}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]+\frac{1}{2\eta}\|x_{j-1}-u\|^2 \\ &-\frac{1+\eta\sigma}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-u\|^2\big]+g(u)-\mathbb{E}_{i_j}\big[g(x_j)\big]. \end{aligned}$$

Let $\beta=\frac{L\theta}{1-\theta}>0$, by rearranging the above inequality, we obtain

$$\begin{aligned} 0 \leq{}& \frac{1-\theta}{\theta}f(\tilde{x}_{s-1})-\frac{1}{\theta}\mathbb{E}_{i_j}\big[f(y_j)\big]+F(u)-\mathbb{E}_{i_j}\big[g(x_j)\big]+\frac{1}{2\eta}\|x_{j-1}-u\|^2-\frac{1+\eta\sigma}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-u\|^2\big] \\ \overset{(\star)}{\leq}{}& \frac{1-\theta}{\theta}F(\tilde{x}_{s-1})-\frac{1}{\theta}\mathbb{E}_{i_j}\big[F(y_j)\big]+F(u)+\frac{1}{2\eta}\|x_{j-1}-u\|^2-\frac{1+\eta\sigma}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-u\|^2\big], \\ \frac{1}{\theta}\big(\mathbb{E}_{i_j}\big[F(y_j)\big]-F(u)\big) \leq{}& \frac{1-\theta}{\theta}\big(F(\tilde{x}_{s-1})-F(u)\big)+\frac{1}{2\eta}\|x_{j-1}-u\|^2-\frac{1+\eta\sigma}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-u\|^2\big], \quad (11) \end{aligned}$$

where $(\star)$ follows from the Jensen's inequality and the definition of y_{j-1}, which leads to $-g(x_j)\leq\frac{1-\theta}{\theta}g(\tilde{x}_{s-1})-\frac{1}{\theta}g(y_j)$.

Let $u=x^*$, using our choice of $\omega=1+\eta\sigma$ to sum (11) over $j=1\ldots m$ with increasing weight ω^{j-1}. After taking expectation with respect to all randomness in this epoch, we have

$$\frac{1}{\theta}\sum_{j=0}^{m-1}\omega^j\big(\mathbb{E}\big[F(y_{j+1})\big]-F(x^*)\big)+\frac{\omega^m}{2\eta}\mathbb{E}\big[\|x_m-x^*\|^2\big]\leq\frac{1-\theta}{\theta}\sum_{j=0}^{m-1}\omega^j\big(F(\tilde{x}_{s-1})-F(x^*)\big)+\frac{1}{2\eta}\|x_0-x^*\|^2.$$

Using the Jensen's inequality and $\tilde{x}_s = \theta\big(\sum_{j=0}^{m-1}\omega^j\big)^{-1}\sum_{j=0}^{m-1}\omega^j x_{j+1} + (1-\theta)\tilde{x}_{s-1} = (\sum_{j=0}^{m-1}\omega^j)^{-1}\sum_{j=0}^{m-1}\omega^j y_{j+1}$, we have

$$\frac{1}{\theta}\sum_{j=0}^{m-1}\omega^j\big(\mathbb{E}\big[F(\tilde{x}_s)\big] - F(x^*)\big) + \frac{\omega^m}{2\eta}\mathbb{E}\big[\|x_m - x^*\|^2\big] \le \frac{1-\theta}{\theta}\sum_{j=0}^{m-1}\omega^j\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{1}{2\eta}\|x_0 - x^*\|^2. \quad (12)$$

(I) Consider the first case in Theorem 1 with $\frac{m}{\kappa} \le \frac{3}{4}$, we set $\eta = \sqrt{\frac{1}{3\sigma m L}}$, $\theta = \sqrt{\frac{m}{3\kappa}} \le \frac{1}{2}$, and $m = \Theta(n)$.

First, we evaluate the crucial constraint (7). By substituting in our parameter settings, the constraint becomes

$$L\theta + \frac{L\theta}{1-\theta} \le \frac{1}{\eta} \to \sqrt{\frac{m}{\kappa}} \le \frac{\sqrt{3}}{2}.$$

Thus the constraint is satisfied by meeting the case assumption.

Then we focus on $(1-\theta)\omega^m$, observed that

$$(1-\theta)\omega^m = (1 - \sqrt{\frac{m}{3\kappa}}) \cdot (1 + \sqrt{\frac{1}{3m\kappa}})^m.$$

Let $\zeta = \sqrt{\frac{m}{\kappa}}$, $\zeta \in (0, \frac{\sqrt{3}}{2}]$, we can denote

$$\phi(\zeta) = (1 - \frac{\sqrt{3}}{3}\zeta) \cdot (1 + \frac{\sqrt{3}}{3} \cdot \frac{\zeta}{m})^m$$

as a function of ζ.

By taking derivative with respect to ζ, we find that $\phi(\zeta)$ is monotonically decreasing on $[0, \frac{\sqrt{3}}{2}]$ for any $m > 0$, which means

$$(1-\theta)\omega^m \le \max_{\zeta \in (0, \frac{\sqrt{3}}{2}]} \phi(\zeta) \le \phi(0) = 1.$$

Thus we have $\frac{1}{\theta} \ge \frac{1-\theta}{\theta}\omega^m$. By using this inequality in (12), we get

$$\frac{1-\theta}{\theta}\sum_{j=0}^{m-1}\omega^j\big(\mathbb{E}\big[F(\tilde{x}_s)\big] - F(x^*)\big) + \frac{1}{2\eta}\mathbb{E}\big[\|x_m - x^*\|^2\big]$$
$$\le \omega^{-m} \cdot \Big(\frac{1-\theta}{\theta}\sum_{j=0}^{m-1}\omega^j\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{1}{2\eta}\|x_0 - x^*\|^2\Big).$$

Dividing both sides of the above inequality by $\frac{1-\theta}{\theta}\sum_{j=0}^{m-1}\omega^j$, we get

$$\big(\mathbb{E}\big[F(\tilde{x}_s)\big] - F(x^*)\big) + \frac{\theta}{2\eta(1-\theta)\sum_{j=0}^{m-1}\omega^j}\mathbb{E}\big[\|x_m - x^*\|^2\big]$$
$$\le \omega^{-m} \cdot \Big(\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{\theta}{2\eta(1-\theta)\sum_{j=0}^{m-1}\omega^j}\|x_0 - x^*\|^2\Big).$$

Summing the above inequality over $s = 1 \ldots \mathcal{S}$, we get

$$\big(\mathbb{E}\big[F(\tilde{x}_{\mathcal{S}})\big] - F(x^*)\big) + \frac{\theta}{2\eta(1-\theta)\sum_{j=0}^{m-1}\omega^j}\mathbb{E}\big[\|x_m^{\mathcal{S}} - x^*\|^2\big]$$
$$\le \omega^{-\mathcal{S}m} \cdot \Big(\big(F(\tilde{x}_0) - F(x^*)\big) + \frac{\theta}{2\eta(1-\theta)\sum_{j=0}^{m-1}\omega^j}\|x_0^1 - x^*\|^2\Big).$$

Notice that in order to prevent confusion, we mark iterates with epoch number, such as $x_m^{\mathcal{S}}$ represent the last iterate in epoch $\mathcal{S}$.

Using the fact that $\sum_{j=0}^{m-1} \omega^j \geq m$, we have

$$\big(\mathbb{E}\big[F(\tilde{x}_{\mathcal{S}})\big] - F(x^*)\big) \leq \omega^{-\mathcal{S}m} \cdot \Big(\big(F(\tilde{x}_0) - F(x^*)\big) + \frac{\theta}{2\eta(1-\theta)m}\|x_0^1 - x^*\|^2\Big).$$

Using the σ-strongly convexity of $F(\cdot)$ to bound $\|x_0^1 - x^*\|^2$, which is $\|x_0^1 - x^*\|^2 \leq \frac{2}{\sigma}\big(F(x_0^1) - F(x^*)\big)$, we obtain

$$\mathbb{E}\big[F(\tilde{x}_{\mathcal{S}}) - F(x^*)\big] \leq (1+\eta\sigma)^{-\mathcal{S}m} \cdot \Big(1 + \frac{\theta}{\eta(1-\theta)m\sigma}\Big) \cdot \big(F(\tilde{x}_0) - F(x^*)\big).$$

Note that $\tilde{x}_0 = x_0^1 = x_0$.

By substituting with our parameters setting, we get

$$\begin{aligned}\mathbb{E}\big[F(\tilde{x}_{\mathcal{S}}) - F(x^*)\big] &\overset{(\star)}{\leq} \big(O(1+\sqrt{\frac{1}{3n\kappa}})\big)^{-\mathcal{S}m} \cdot O\Big(1 + 2\theta\sqrt{\frac{\kappa}{n}}\Big) \cdot \big(F(\tilde{x}_0) - F(x^*)\big)\\ &\leq \big(O(1+\sqrt{\frac{1}{3n\kappa}})\big)^{-\mathcal{S}m} \cdot O\big(F(\tilde{x}_0) - F(x^*)\big),\end{aligned}$$

where $(\star)$ holds due to the fact that $\theta \leq \frac{1}{2}$.

The above result implies that the oracle complexity in the case $\frac{m}{\kappa} \leq \frac{3}{4}$ to achieve an ϵ-additive error is $\mathcal{O}\Big(\sqrt{\kappa n}\log\frac{F(\tilde{x}_0)-F(x^*)}{\epsilon}\Big)$.

(II) For another case with $\frac{m}{\kappa} > \frac{3}{4}$, we set $\eta = \frac{2}{3L}$, $\theta = \frac{1}{2}$, and $m = \Theta(n)$.

Again, we evaluate the constraint (7) first. By substituting the parameter setting, the constraint becomes

$$L\theta + \frac{L\theta}{1-\theta} \leq \frac{1}{\eta} \to \eta \leq \frac{2}{3L}.$$

Thus the constraint is satisfied by our parameter choice.

Substituting the parameter setting into (12), we get

$$2\sum_{j=0}^{m-1}\omega^j\big(\mathbb{E}\big[F(\tilde{x}_s)\big] - F(x^*)\big) + \frac{3L\omega^m}{4}\mathbb{E}\big[\|x_m - x^*\|^2\big] \leq \sum_{j=0}^{m-1}\omega^j\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{3L}{4}\|x_0 - x^*\|^2.$$

Notice that based on the Bernoulli's inequality, $\omega^m = (1+\frac{2}{3\kappa})^m \geq 1 + \frac{2m}{3\kappa} \geq \frac{3}{2}$, which leads to

$$\begin{aligned}\frac{3}{2}\sum_{j=0}^{m-1}\omega^j\big(\mathbb{E}\big[F(\tilde{x}_s)\big] - F(x^*)\big) + \frac{9L}{8}\mathbb{E}\big[\|x_m - x^*\|^2\big] &\leq \sum_{j=0}^{m-1}\omega^j\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{3L}{4}\|x_0 - x^*\|^2,\\ \sum_{j=0}^{m-1}\omega^j\big(\mathbb{E}\big[F(\tilde{x}_s)\big] - F(x^*)\big) + \frac{3L}{4}\mathbb{E}\big[\|x_m - x^*\|^2\big] &\leq \frac{2}{3}\cdot\big(\sum_{j=0}^{m-1}\omega^j\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{3L}{4}\|x_0 - x^*\|^2\big).\end{aligned}$$

Again, by telescoping the above inequality from $s = 1 \ldots \mathcal{S}$, we get

$$\sum_{j=0}^{m-1}\omega^j\big(\mathbb{E}\big[F(\tilde{x}_{\mathcal{S}})\big] - F(x^*)\big) + \frac{3L}{4}\mathbb{E}\big[\|x_m^{\mathcal{S}} - x^*\|^2\big] \leq \big(\frac{2}{3}\big)^{\mathcal{S}}\cdot\big(\sum_{j=0}^{m-1}\omega^j\big(F(\tilde{x}_0) - F(x^*)\big) + \frac{3L}{4}\|x_0^1 - x^*\|^2\big).$$

Since $\sum_{j=0}^{m-1}\omega^j \geq m$, the above inequality can be rewritten as follows:

$$\begin{aligned}\big(\mathbb{E}\big[F(\tilde{x}_{\mathcal{S}})\big] - F(x^*)\big) &\overset{(\star)}{\leq} \big(\frac{2}{3}\big)^{\mathcal{S}}\cdot\big(1 + \frac{3\kappa}{2m}\big)\cdot\big(F(\tilde{x}_0) - F(x^*)\big)\\ &\leq \big(\frac{2}{3}\big)^{\mathcal{S}}\cdot O\big(F(\tilde{x}_0) - F(x^*)\big),\end{aligned}$$

where $(\star)$ uses the σ-strongly convexity of $F(\cdot)$, that is, $\|x_0^1 - x^*\|^2 \leq \frac{2}{\sigma}\big(F(x_0^1) - F(x^*)\big)$.

This result implies that the oracle complexity in this case is $\mathcal{O}\big(n\log\frac{F(\tilde{x}_0)-F(x^*)}{\epsilon}\big)$.

B.1.1. Detailed comparison between MiG and Katyusha

As mentioned in Section 3.1.1, MiG is a special case of Katyusha. However, it is a non-trivial special case and there are some significant differences from the theoretical perspective between them for the ill-conditioned problems:

The intuition of Katyusha is that Katyusha Momentum is used to further reduce the variance of the iterates so as to make Nesterov's Momentum effective, which can be clearly seen from the parameter choice: the author fixed Katyusha Momentum (for $\tilde{x}$, in the notation of original work) as $\tau_2 = 1/2$, and chose Nesterov's Momentum (for y and z) with $(1/2 - \tau_1, \tau_1)$. This idea is similar to Acc-Prox-SVRG (Nitanda, 2014), which uses sufficiently large mini-batch to make the Nesterov's Momentum effective. Thus, the acceleration comes from Nesterov's Momentum, which is common in previous work (e.g. Catalyst).

In contrast, MiG uses only Katyusha Momentum (negative momentum) to achieve acceleration (i.e., choosing τ_2 in a dynamic way), which is somehow counter-intuitive. To the best of our knowledge, this is the first work to yield an acceleration without using Nesterov's Momentum (which is at the heart of all existing accelerated first-order methods, as stated in (Allen-Zhu, 2017)).

B.2. Proof of Theorem 2

Again, we first impose the following constraint on η and θ:

$$L\theta + \frac{L\theta}{1-\theta} \leq \frac{1}{\eta}\text{ , or equivalently } \eta \leq \frac{1-\theta}{L\theta(2-\theta)}. \tag{13}$$

We start with convexity of $f(\cdot)$ at y_{j-1}. By definition, for any vector $u \in \mathbb{R}^d$, we have

$$\begin{aligned} f(y_{j-1}) - f(u) &\leq \langle \nabla f(y_{j-1}), y_{j-1} - u\rangle \\ &\overset{(\star)}{=} \frac{1-\theta}{\theta}\langle \nabla f(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle + \langle \nabla f(y_{j-1}), x_{j-1} - u\rangle, \end{aligned} \tag{14}$$

where $(\star)$ follows from the fact that $y_{j-1} = \theta x_{j-1} + (1-\theta)\tilde{x}_{s-1}$.

Then we further expand $\langle \nabla f(y_{j-1}), x_{j-1} - u\rangle$ as

$$\langle \nabla f(y_{j-1}), x_{j-1} - u\rangle = \langle \nabla f(y_{j-1}) - \tilde{\nabla}, x_{j-1} - u\rangle + \langle \tilde{\nabla}, x_{j-1} - x_j\rangle + \langle \tilde{\nabla}, x_j - u\rangle. \tag{15}$$

Using L-smooth (2) of $f(\cdot)$ at (y_j, y_{j-1}), we get

$$\begin{aligned} f(y_j) - f(y_{j-1}) &\leq \langle \nabla f(y_{j-1}), y_j - y_{j-1}\rangle + \frac{L}{2}\|y_j - y_{j-1}\|^2 \\ &= \theta\big[\langle \nabla f(y_{j-1}) - \tilde{\nabla}, x_j - x_{j-1}\rangle + \langle \tilde{\nabla}, x_j - x_{j-1}\rangle\big] + \frac{L\theta^2}{2}\|x_j - x_{j-1}\|^2, \\ \langle \tilde{\nabla}, x_{j-1} - x_j\rangle &\leq \frac{1}{\theta}\big(f(y_{j-1}) - f(y_j)\big) + \langle \nabla f(y_{j-1}) - \tilde{\nabla}, x_j - x_{j-1}\rangle + \frac{L\theta}{2}\|x_j - x_{j-1}\|^2. \end{aligned}$$

Using the constraint (13), we have

$$\langle \tilde{\nabla}, x_{j-1} - x_j\rangle \leq \frac{1}{\theta}\big(f(y_{j-1}) - f(y_j)\big) + \langle \nabla f(y_{j-1}) - \tilde{\nabla}, x_j - x_{j-1}\rangle + \frac{1}{2\eta}\|x_j - x_{j-1}\|^2 - \frac{L\theta}{2(1-\theta)}\|x_j - x_{j-1}\|^2. \tag{16}$$

Then we can combine (14), (15), (16), as well as Lemma 3 (with $g(x)$ convex), which leads to

$$\begin{aligned} f(y_{j-1}) - f(u) \leq{}& \frac{1-\theta}{\theta}\langle \nabla f(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle + \langle \nabla f(y_{j-1}) - \tilde{\nabla}, x_j - u\rangle + \frac{1}{\theta}(f(y_{j-1}) - f(y_j)) \\ &- \frac{L\theta}{2(1-\theta)}\|x_j - x_{j-1}\|^2 + \frac{1}{2\eta}\|x_{j-1} - u\|^2 - \frac{1}{2\eta}\|x_j - u\|^2 + g(u) - g(x_j). \end{aligned}$$

After taking expectation with respect to the sample i_j, we get

$$\begin{aligned} f(y_{j-1}) - f(u) \leq & \frac{1-\theta}{\theta}\langle \nabla f(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle + \frac{1}{2\beta}\mathbb{E}_{i_j}\big[\|\nabla f(y_{j-1}) - \tilde{\nabla}\|^2\big] + \frac{\beta}{2}\mathbb{E}_{i_j}\big[\|x_j - x_{j-1}\|^2\big] \\ & + \frac{1}{\theta}\big(f(y_{j-1}) - \mathbb{E}_{i_j}\big[f(y_j)\big]\big) - \frac{L\theta}{2(1-\theta)}\mathbb{E}_{i_j}\big[\|x_j - x_{j-1}\|^2\big] + \frac{1}{2\eta}\|x_{j-1} - u\|^2 \\ & - \frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j - u\|^2\big] + g(u) - \mathbb{E}_{i_j}\big[g(x_j)\big], \end{aligned}$$

which the inequality uses the unbiasedness of the gradient estimator and the Young's inequality with the parameter $\beta > 0$.

Using Lemma 1 to bound the variance term and choosing $\beta = \frac{L\theta}{1-\theta}$, the above inequality becomes

$$\frac{1}{\theta}\mathbb{E}_{i_j}\big[f(y_j)\big] - F(u) \leq \frac{1-\theta}{\theta}f(\tilde{x}_{s-1}) + \frac{1}{2\eta}\|x_{j-1} - u\|^2 - \frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j - u\|^2\big] - \mathbb{E}_{i_j}\big[g(x_j)\big].$$

Let $u = x^*$ and using the fact that $-g(x_j) \leq \frac{1-\theta}{\theta}g(\tilde{x}_{s-1}) - \frac{1}{\theta}g(y_j)$, we get

$$\frac{1}{\theta}\mathbb{E}_{i_j}\big[F(y_j) - F(x^*)\big] \leq \frac{1-\theta}{\theta}\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{1}{2\eta}\|x_{j-1} - x^*\|^2 - \frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j - x^*\|^2\big].$$

Summing the above inequality over $j = 1 \ldots m$ and taking expectation with respect to all randomness in this epoch, we obtain

$$\begin{aligned} \frac{1}{\theta}\mathbb{E}\Big[\frac{1}{m}\sum_{j=1}^{m} F(y_j) - F(x^*)\Big] & \overset{(a)}{\leq} \frac{1-\theta}{\theta}\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{2\theta L}{m}\|x_0 - x^*\|^2 - \frac{2\theta L}{m}\mathbb{E}\big[\|x_m - x^*\|^2\big], \\ \frac{1}{\theta^2}\mathbb{E}\big[F(\tilde{x}_s) - F(x^*)\big] & \overset{(b)}{\leq} \frac{1-\theta}{\theta^2}\big(F(\tilde{x}_{s-1}) - F(x^*)\big) + \frac{2L}{m}\|x_0 - x^*\|^2 - \frac{2L}{m}\mathbb{E}\big[\|x_m - x^*\|^2\big], \end{aligned} \tag{17}$$

where (a) follows from our parameter choice of $\eta = \frac{1}{4L\theta}$, and (b) uses the Jensen's inequality.

At this point, we first check the constraint (13),

$$\eta \leq \frac{1-\theta}{L\theta(2-\theta)} \rightarrow \theta \leq \frac{2}{3},$$

which is satisfied by the parameter setting $\theta = \frac{2}{s+4} \leq \frac{1}{2}$.

Then by evaluating (17) with $s = 1$, we have

$$\frac{1}{\theta_1^2}\mathbb{E}\big[F(\tilde{x}_1) - F(x^*)\big] + \frac{2L}{m}\mathbb{E}\big[\|x_m^1 - x^*\|^2\big] \leq \frac{1-\theta_1}{\theta_1^2}\big(F(\tilde{x}_0) - F(x^*)\big) + \frac{2L}{m}\|x_0^1 - x^*\|^2.$$

Notice that here we mark iterates and θ with epoch number to prevent confusion.

Since $\theta_s = \frac{2}{s+4}$, one can easily verify that $\frac{1-\theta_{s+1}}{\theta_{s+1}^2} \leq \frac{1}{\theta_s^2}$. Thus we can telescope the above inequality from $s = 1 \ldots \mathcal{S}$,

$$\begin{aligned} \frac{1}{\theta_{\mathcal{S}}^2}\mathbb{E}\big[F(\tilde{x}_{\mathcal{S}}) - F(x^*)\big] + \frac{2L}{m}\mathbb{E}\big[\|x_m^{\mathcal{S}} - x^*\|^2\big] & \leq \frac{1-\theta_1}{\theta_1^2}\big(F(\tilde{x}_0) - F(x^*)\big) + \frac{2L}{m}\|x_0^1 - x^*\|^2, \\ \mathbb{E}\big[F(\tilde{x}_{\mathcal{S}}) - F(x^*)\big] & \leq \frac{4(1-\theta_1)}{(\mathcal{S}+4)^2\theta_1^2}\big(F(\tilde{x}_0) - F(x^*)\big) + \frac{8L}{(\mathcal{S}+4)^2 m}\|x_0^1 - x^*\|^2. \end{aligned}$$

In other words, by choosing $m = \Theta(n)$, the total oracle complexity is $\mathcal{O}\big(n\sqrt{\frac{F(\tilde{x}_0) - F(x^*)}{\epsilon}} + \sqrt{\frac{nL\|x_0^1 - x^*\|^2}{\epsilon}}\big)$.

C. Proofs for Section 4

Lemma 4 (Sparse Variance Bound 1)**.** *If Assumption 3 holds, for any $y, \tilde{x} \in \mathbb{R}^d$ and the sample i_j, denote $\tilde{\nabla} = \nabla f_{i_j}(y) - \nabla f_{i_j}(\tilde{x}) + D_{i_j}\nabla F(\tilde{x})$, where D_{i_j} is defined in Section 4.1, we can bound the variance $\mathbb{E}_{i_j}\big[\|\tilde{\nabla}\|^2\big]$ as*

$$\mathbb{E}_{i_j}\big[\|\tilde{\nabla}\|^2\big] \leq 4L\big(F(y) - F(x^*)\big) + 4L\big(F(\tilde{x}) - F(x^*)\big).$$

Proof. From Lemma 10 in (Mania et al., 2017), we have

$$\mathbb{E}_{i_j}\left[\|\tilde{\nabla}\|^2\right] \leq 2\mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(y) - \nabla f_{i_j}(x^*)\|^2\right] + 2\mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(\tilde{x}) - \nabla f_{i_j}(x^*)\|^2\right], \tag{18}$$

which provides an upper bound for the variance of the sparse stochastic variance reduced gradient estimator.

From Theorem 2.1.5 in (Nesterov, 2004), we have

$$\|\nabla f_{i_j}(y) - \nabla f_{i_j}(x^*)\|^2 \leq 2L\big(f_{i_j}(y) - f_{i_j}(x^*) - \langle \nabla f_{i_j}(x^*), y - x^*\rangle\big).$$

Taking expectation with the sample i_j, the above inequality becomes

$$\mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(y) - \nabla f_{i_j}(x^*)\|^2\right] \leq 2L\big(F(y) - F(x^*)\big).$$

Similarly, we have

$$\mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(\tilde{x}) - \nabla f_{i_j}(x^*)\|^2\right] \leq 2L\big(F(\tilde{x}) - F(x^*)\big).$$

Substituting the above inequalities into (18) yields the desired result. □

Lemma 5 (Sparse Variance Bound 2)**.** *If Assumption 3 holds, using same notations as in Algorithm 3, we can bound the variance as*

$$\begin{aligned}\mathbb{E}_{i_j}\left[\|\nabla F(y_{j-1}) - \tilde{\nabla}_S\|^2\right] \leq{}& 2L\big(F(\tilde{x}_{s-1}) - F(y_{j-1}) - \langle \nabla F(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle\big)\\ &+ 2L(D_m^2 - D_m)\cdot\big(F(\tilde{x}_{s-1}) - F(x^*)\big),\end{aligned}$$

where $D_m = \max_{k=1\ldots d}\frac{1}{p_k}$, i.e., the inverse probability of the most sparse coordinates.

Proof. First, we expand the variance term as

$$\begin{aligned}&\mathbb{E}_{i_j}\left[\|\nabla F(y_{j-1}) - \tilde{\nabla}_S\|^2\right]\\ ={}& \mathbb{E}_{i_j}\left[\left\|\big(\nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1})\big) - \big(\nabla F(y_{j-1}) - D_{i_j}\nabla F(\tilde{x}_{s-1})\big)\right\|^2\right]\\ ={}& \mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1})\|^2\right] - 2\mathbb{E}_{i_j}\left[\langle \nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1}), \nabla F(y_{j-1}) - D_{i_j}\nabla F(\tilde{x}_{s-1})\rangle\right]\\ &+ \mathbb{E}_{i_j}\left[\|\nabla F(y_{j-1}) - D_{i_j}\nabla F(\tilde{x}_{s-1})\|^2\right].\end{aligned} \tag{19}$$

Then we focus on the last two terms in (19),

$$\begin{aligned}&\mathbb{E}_{i_j}\left[\|\nabla F(y_{j-1}) - D_{i_j}\nabla F(\tilde{x}_{s-1})\|^2\right]\\ ={}& \|\nabla F(y_{j-1})\|^2 - 2\mathbb{E}_{i_j}\left[\langle \nabla F(y_{j-1}), D_{i_j}\nabla F(\tilde{x}_{s-1})\rangle\right] + \mathbb{E}_{i_j}\left[\|D_{i_j}\nabla F(\tilde{x}_{s-1})\|^2\right]\\ \overset{(\star)}{=}{}& \|\nabla F(y_{j-1})\|^2 - 2\langle \nabla F(y_{j-1}), \nabla F(\tilde{x}_{s-1})\rangle + \langle \nabla F(\tilde{x}_{s-1}), D\nabla F(\tilde{x}_{s-1})\rangle,\end{aligned} \tag{20}$$

where $(\star)$ uses the fact that $P_{i_j} = P_{i_j}\cdot P_{i_j}$ and the unbiased sparse estimator $\mathbb{E}_{i_j}\big[D_{i_j}\nabla F(x)\big] = \nabla F(x)$.

$$\begin{aligned}&\mathbb{E}_{i_j}\left[\langle \nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1}), \nabla F(y_{j-1}) - D_{i_j}\nabla F(\tilde{x}_{s-1})\rangle\right]\\ ={}& \|\nabla F(y_{j-1})\|^2 - \langle \nabla F(y_{j-1}), \nabla F(\tilde{x}_{s-1})\rangle - \mathbb{E}_{i_j}\left[\langle \nabla f_{i_j}(y_{j-1}), D_{i_j}\nabla F(\tilde{x}_{s-1})\rangle\right]\\ &+ \mathbb{E}_{i_j}\left[\langle \nabla f_{i_j}(\tilde{x}_{s-1}), D_{i_j}\nabla F(\tilde{x}_{s-1})\rangle\right]\\ \overset{(\star)}{=}{}& \|\nabla F(y_{j-1})\|^2 - \langle \nabla F(y_{j-1}), \nabla F(\tilde{x}_{s-1})\rangle - \langle \nabla F(y_{j-1}), D\nabla F(\tilde{x}_{s-1})\rangle + \langle \nabla F(\tilde{x}_{s-1}), D\nabla F(\tilde{x}_{s-1})\rangle,\end{aligned} \tag{21}$$

where $(\star)$ follows from $\langle \nabla f_{i_j}(y_{j-1}), D_{i_j}\nabla F(\tilde{x}_{s-1})\rangle = \langle \nabla f_{i_j}(y_{j-1}), D\nabla F(\tilde{x}_{s-1})\rangle$, since $\nabla f_{i_j}(y_{j-1})$ and D_{i_j} are both supported on the sample i_j.

By substituting (20) and (21) into (19), we get

$$\begin{aligned}&\mathbb{E}_{i_j}\left[\|\nabla F(y_{j-1}) - \tilde{\nabla}_S\|^2\right]\\ ={}& \mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1})\|^2\right] - \|\nabla F(y_{j-1})\|^2 + 2\langle \nabla F(y_{j-1}), D\nabla F(\tilde{x}_{s-1})\rangle - \|D\nabla F(\tilde{x}_{s-1})\|^2\\ &+ \langle (D - I)\nabla F(\tilde{x}_{s-1}), D\nabla F(\tilde{x}_{s-1})\rangle\\ ={}& \mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1})\|^2\right] - \|\nabla F(y_{j-1}) - D\nabla F(\tilde{x}_{s-1})\|^2 + \langle (D - I)\nabla F(\tilde{x}_{s-1}), D\nabla F(\tilde{x}_{s-1})\rangle\\ \leq{}& \mathbb{E}_{i_j}\left[\|\nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1})\|^2\right] + \langle (D - I)\nabla F(\tilde{x}_{s-1}), D\nabla F(\tilde{x}_{s-1})\rangle.\end{aligned} \tag{22}$$

From Theorem 2.1.5 in (Nesterov, 2004), we have

$$\|\nabla f_{i_j}(y_{j-1}) - \nabla f_{i_j}(\tilde{x}_{s-1})\|^2 \leq 2L\big(f_{i_j}(\tilde{x}_{s-1}) - f_{i_j}(y_{j-1}) - \langle \nabla f_{i_j}(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle\big). \tag{23}$$

By substituting (23) into (22), we obtain

$$\begin{aligned}&\mathbb{E}_{i_j}\big[\|\nabla F(y_{j-1}) - \tilde{\nabla}_S\|^2\big]\\ &\leq 2L\big(F(\tilde{x}_{s-1}) - F(y_{j-1}) - \langle \nabla F(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle\big) + \langle (D-I)\nabla F(\tilde{x}_{s-1}), D\nabla F(\tilde{x}_{s-1})\rangle.\end{aligned}$$

Since $D_m = \max_{k=1\ldots d} \frac{1}{p_k}$, we can bound the last term as follows:

$$\begin{aligned}&\langle (D-I)\nabla F(\tilde{x}_{s-1}), D\nabla F(\tilde{x}_{s-1})\rangle\\ &= \sum_{k=1}^{d} (\frac{1}{p_k^2} - \frac{1}{p_k})[\nabla F(\tilde{x}_{s-1})]_k^2\\ &\leq (D_m^2 - D_m)\|\nabla F(\tilde{x}_{s-1}) - \nabla F(x^*)\|^2\\ &\overset{(\star)}{\leq} 2L(D_m^2 - D_m)\cdot\big(F(\tilde{x}_{s-1}) - F(x^*)\big),\end{aligned}$$

where $(\star)$ follows from applying Theorem 2.1.5 in (Nesterov, 2004). □

C.1. Proof of Theorem 3

For the purpose of analyzing in the asynchronous setting, we need to analyze the convergence of MiG starting with iterate difference.

We start with the iterate difference between y_j and x^*. By expanding iterate difference and taking expectation with respect to the sample i_j, we get

$$\begin{aligned}\mathbb{E}_{i_j}\big[\|y_j - x^*\|^2\big] &\overset{(a)}{=} \mathbb{E}_{i_j}\big[\|\theta(x_{j-1} - \eta\cdot\tilde{\nabla}_S) + (1-\theta)\tilde{x}_{s-1} - x^*\|^2\big]\\ &= \mathbb{E}_{i_j}\big[\|y_{j-1} - \eta\theta\cdot\tilde{\nabla}_S - x^*\|^2\big]\\ &\overset{(b)}{=} \|y_{j-1} - x^*\|^2 - 2\eta\theta\langle \nabla F(y_{j-1}), y_{j-1} - x^*\rangle + \eta^2\theta^2\mathbb{E}_{i_j}\big[\|\tilde{\nabla}_S\|^2\big],\end{aligned} \tag{24}$$

where (a) uses the definition of y, and (b) follows from the unbiasedness of the sparse gradient estimator $\mathbb{E}_{i_j}\big[\tilde{\nabla}_S\big] = \nabla F(y_{j-1})$.

Using σ-strongly convex, we get a bound for $-\langle \nabla F(y_{j-1}), y_{j-1} - x^*\rangle$ as

$$\langle \nabla F(y_{j-1}), y_{j-1} - x^*\rangle \geq F(y_{j-1}) - F(x^*) + \frac{\sigma}{2}\|y_{j-1} - x^*\|^2. \tag{25}$$

Using Lemma 4, we have the following variance bound:

$$\mathbb{E}_{i_j}\big[\|\tilde{\nabla}_S\|^2\big] \leq 4L\big(F(y_{j-1}) - F(x^*)\big) + 4L\big(F(\tilde{x}_{s-1}) - F(x^*)\big). \tag{26}$$

By combining (24), (25) and (26), we get

$$\begin{aligned}\mathbb{E}_{i_j}\big[\|y_j - x^*\|^2\big] &\leq \|y_{j-1} - x^*\|^2 - \eta\theta\sigma\|y_{j-1} - x^*\|^2 - 2\eta\theta\big(F(y_{j-1}) - F(x^*)\big) + \eta^2\theta^2\mathbb{E}_{i_j}\big[\|\tilde{\nabla}_S\|^2\big]\\ &\leq (1-\eta\theta\sigma)\cdot\|y_{j-1} - x^*\|^2 + (4L\eta^2\theta^2 - 2\eta\theta)\cdot\big(F(y_{j-1}) - F(x^*)\big)\\ &\quad + 4L\eta^2\theta^2\big(F(\tilde{x}_{s-1}) - F(x^*)\big),\end{aligned}$$

$$\begin{aligned}&(2\eta\theta - 4L\eta^2\theta^2)\cdot\big(F(y_{j-1}) - F(x^*)\big)\\ &\leq \|y_{j-1} - x^*\|^2 - \mathbb{E}_{i_j}\big[\|y_j - x^*\|^2\big] + 4L\eta^2\theta^2\big(F(\tilde{x}_{s-1}) - F(x^*)\big).\end{aligned}$$

Summing the above inequality over $j = 1\ldots m$ and taking expectation with respect to all randomness in this epoch, we get

$$\begin{aligned}&(2\eta\theta - 4L\eta^2\theta^2)\cdot\mathbb{E}\big[\sum_{j=1}^{m}\big(F(y_{j-1}) - F(x^*)\big)\big]\\ &\leq \|y_0 - x^*\|^2 - \mathbb{E}\big[\|y_m - x^*\|^2\big] + 4L\eta^2\theta^2 m\big(F(\tilde{x}_{s-1}) - F(x^*)\big).\end{aligned}$$

Using Jensen’s inequality and $\tilde{x}_s = \frac{1}{m}\sum_{j=0}^{m-1} y_j$, we obtain

$$\begin{aligned}(2\eta\theta - 4L\eta^2\theta^2)m \cdot \mathbb{E}\big[\big(F(\tilde{x}_s) - F(x^*)\big)\big] &\le \|y_0 - x^*\|^2 + 4L\eta^2\theta^2 m\big(F(\tilde{x}_{s-1}) - F(x^*)\big),\\ \mathbb{E}\big[\big(F(\tilde{x}_s) - F(x^*)\big)\big] &\overset{(\star)}{\le} \frac{\frac{2}{\sigma} + 4L\eta^2\theta^2 m}{(2\eta\theta - 4L\eta^2\theta^2)m} \cdot \big(F(\tilde{x}_{s-1}) - F(x^*)\big),\end{aligned}$$

where $(\star)$ follows from the fact $x_0 = \tilde{x}_{s-1} = y_0$ and σ-strongly convexity of $F(\cdot)$.

By choosing $m = 25\kappa$, $\eta = \frac{1}{L}$, $\theta = \frac{1}{10}$, the above inequality becomes

$$\mathbb{E}\big[\big(F(\tilde{x}_s) - F(x^*)\big)\big] \le 0.75 \cdot \big(F(\tilde{x}_{s-1}) - F(x^*)\big),$$

which means that the total oracle complexity is $\mathcal{O}\big((n+\kappa)\log\frac{F(\tilde{x}_0) - F(x^*)}{\epsilon}\big)$.

C.2. Proof of Theorem 4

Here we consider improving the convergence rate for the Serial Sparse MiG, referencing to Algorithm 3 with Option II. Our analysis is based on function difference similar to the proofs in Appendix B, which make the parameter θ effective.

Similarly, we first add the following constraint:

$$L\theta + \frac{L\theta}{1-\theta} \le \frac{1}{\eta}. \tag{27}$$

We start with the convexity of $F(\cdot)$ at y_{j-1} and x^*,

$$\begin{aligned}F(y_{j-1}) - F(x^*) &\le \langle \nabla F(y_{j-1}), y_{j-1} - x^*\rangle\\ &= \frac{1-\theta}{\theta}\langle \nabla F(y_{j-1}), \tilde{x}_{s-1} - y_{j-1}\rangle + \langle \nabla F(y_{j-1}), x_{j-1} - x^*\rangle.\end{aligned} \tag{28}$$

Then we further expand $\langle \nabla F(y_{j-1}), x_{j-1} - x^*\rangle$ as

$$\langle \nabla F(y_{j-1}), x_{j-1} - x^*\rangle = \langle \nabla F(y_{j-1}) - \tilde{\nabla}_S, x_{j-1} - x^*\rangle + \langle \tilde{\nabla}_S, x_{j-1} - x_j\rangle + \langle \tilde{\nabla}_S, x_j - x^*\rangle. \tag{29}$$

Using L-smooth (2) of $F(\cdot)$ at (y_j, y_{j-1}), we can bound $\langle \tilde{\nabla}_S, x_{j-1} - x_j\rangle$ as

$$\begin{aligned}F(y_j) - F(y_{j-1}) &\le \langle \nabla F(y_{j-1}), y_j - y_{j-1}\rangle + \frac{L}{2}\|y_j - y_{j-1}\|^2\\ &\overset{(\star)}{=} \theta\big[\langle \nabla F(y_{j-1}) - \tilde{\nabla}_S, x_j - x_{j-1}\rangle + \langle \tilde{\nabla}_S, x_j - x_{j-1}\rangle\big] + \frac{L\theta^2}{2}\|x_j - x_{j-1}\|^2,\\ \langle \tilde{\nabla}_S, x_{j-1} - x_j\rangle &\le \frac{1}{\theta}\big(F(y_{j-1}) - F(y_j)\big) + \langle \nabla F(y_{j-1}) - \tilde{\nabla}_S, x_j - x_{j-1}\rangle + \frac{L\theta}{2}\|x_j - x_{j-1}\|^2,\end{aligned}$$

where $(\star)$ uses the definition of y.

Using the constraint (27), we have

$$\begin{aligned}\langle \tilde{\nabla}_S, x_{j-1} - x_j\rangle \le{}& \frac{1}{\theta}\big(F(y_{j-1}) - F(y_j)\big) + \langle \nabla F(y_{j-1}) - \tilde{\nabla}_S, x_j - x_{j-1}\rangle + \frac{1}{2\eta}\|x_j - x_{j-1}\|^2\\ &- \frac{L\theta}{2(1-\theta)}\|x_j - x_{j-1}\|^2.\end{aligned} \tag{30}$$

Since we can write an equivalent update as $x_j = \arg\min_x\{\frac{1}{2}\|x - x_{j-1}\|^2 + \eta\langle \tilde{\nabla}_S, x\rangle\}$, by applying Lemma 2 with $z = x^*$, $z_0 = x_{j-1}$, $z^* = x_j$, $\tau = \frac{1}{\eta}$, $\psi(x) = \langle \tilde{\nabla}_S, x\rangle$, we have

$$\langle \tilde{\nabla}_S, x_j - x^*\rangle = -\frac{1}{2\eta}\|x_{j-1} - x_j\|^2 + \frac{1}{2\eta}\|x_{j-1} - x^*\|^2 - \frac{1}{2\eta}\|x_j - x^*\|^2. \tag{31}$$

Combining (28), (29), (30) and (31), we get

$$\begin{aligned}F(y_{j-1})-F(x^*)\leq&\frac{1-\theta}{\theta}\langle\nabla F(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\frac{1}{\theta}\big(F(y_{j-1})-F(y_j)\big)+\langle\nabla F(y_{j-1})-\tilde{\nabla}_S,x_j-x^*\rangle\\&-\frac{L\theta}{2(1-\theta)}\|x_j-x_{j-1}\|^2+\frac{1}{2\eta}\|x_{j-1}-x^*\|^2-\frac{1}{2\eta}\|x_j-x^*\|^2.\end{aligned}$$

Taking expectation with respect to the sample i_j, we obtain

$$\begin{aligned}F(y_{j-1})-F(x^*)\overset{(a)}{\leq}&\frac{1-\theta}{\theta}\langle\nabla F(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\frac{1}{\theta}\big(F(y_{j-1})-\mathbb{E}_{i_j}\big[F(y_j)\big]\big)-\frac{L\theta}{2(1-\theta)}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]\\&+\mathbb{E}_{i_j}\big[\langle\nabla F(y_{j-1})-\tilde{\nabla}_S,x_j-x_{j-1}\rangle\big]+\frac{1}{2\eta}\|x_{j-1}-x^*\|^2-\frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-x^*\|^2\big]\\\overset{(b)}{\leq}&\frac{1-\theta}{\theta}\langle\nabla F(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\frac{1}{\theta}\big(F(y_{j-1})-\mathbb{E}_{i_j}\big[F(y_j)\big]\big)-\frac{L\theta}{2(1-\theta)}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]\\&+\frac{1}{2\beta}\mathbb{E}_{i_j}\big[\|\nabla F(y_{j-1})-\tilde{\nabla}_S\|^2\big]+\frac{\beta}{2}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]+\frac{1}{2\eta}\|x_{j-1}-x^*\|^2-\frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-x^*\|^2\big],\end{aligned}$$

where (a) uses the unbiasedness of the sparse gradient estimator $\mathbb{E}_{i_j}\big[\tilde{\nabla}_S\big]=\nabla F(y_{j-1})$, and (b) uses the Young's inequality with the parameter $\beta>0$.

Using Lemma 5 to bound the variance term, we get

$$\begin{aligned}F(y_{j-1})-F(x^*)\leq&\frac{1-\theta}{\theta}\langle\nabla F(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle+\frac{1}{\theta}\big(F(y_{j-1})-\mathbb{E}_{i_j}\big[F(y_j)\big]\big)-\frac{L\theta}{2(1-\theta)}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]\\&+\frac{L}{\beta}\big(F(\tilde{x}_{s-1})-F(y_{j-1})-\langle\nabla F(y_{j-1}),\tilde{x}_{s-1}-y_{j-1}\rangle\big)+\frac{L(D_m^2-D_m)}{\beta}\big(F(\tilde{x}_{s-1})-F(x^*)\big)\\&+\frac{\beta}{2}\mathbb{E}_{i_j}\big[\|x_j-x_{j-1}\|^2\big]+\frac{1}{2\eta}\|x_{j-1}-x^*\|^2-\frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-x^*\|^2\big].\end{aligned}$$

By choosing $\beta=\frac{L\theta}{1-\theta}>0$, the above inequality becomes

$$\begin{aligned}F(y_{j-1})-F(x^*)\leq&\frac{1}{\theta}\big(F(y_{j-1})-\mathbb{E}_{i_j}\big[F(y_j)\big]\big)+\frac{1-\theta}{\theta}\big(F(\tilde{x}_{s-1})-F(y_{j-1})\big)\\&+\frac{1-\theta}{\theta}(D_m^2-D_m)\cdot\big(F(\tilde{x}_{s-1})-F(x^*)\big)+\frac{1}{2\eta}\|x_{j-1}-x^*\|^2-\frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-x^*\|^2\big],\\\frac{1}{\theta}\big(\mathbb{E}_{i_j}\big[F(y_j)\big]-F(x^*)\big)\leq&\frac{1-\theta}{\theta}\big(F(\tilde{x}_{s-1})-F(x^*)\big)+\frac{1-\theta}{\theta}(D_m^2-D_m)\cdot\big(F(\tilde{x}_{s-1})-F(x^*)\big)\\&+\frac{1}{2\eta}\|x_{j-1}-x^*\|^2-\frac{1}{2\eta}\mathbb{E}_{i_j}\big[\|x_j-x^*\|^2\big].\end{aligned}$$

Summing the above inequality over $j=1\ldots m$ and taking expectation with respect to all randomness in this epoch, we get

$$\begin{aligned}&\frac{1}{\theta}\Big(\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\big[F(y_j)\big]-F(x^*)\Big)\\\leq&\frac{1-\theta}{\theta}(1+D_m^2-D_m)\cdot\big(F(\tilde{x}_{s-1})-F(x^*)\big)+\frac{1}{2\eta m}\Big(\|x_0-x^*\|^2-\mathbb{E}\big[\|x_m-x^*\|^2\big]\Big).\end{aligned}$$

Using Jensen's inequality and $\tilde{x}_s=\frac{1}{m}\sum_{j=1}^{m}y_j$ in Option II, we have

$$\begin{aligned}&\frac{1}{\theta}\mathbb{E}\big[F(\tilde{x}_s)-F(x^*)\big]\\\leq&\frac{1-\theta}{\theta}\Big(1+D_m^2-D_m\Big)\cdot\big(F(\tilde{x}_{s-1})-F(x^*)\big)+\frac{1}{2\eta m}\Big(\|x_0-x^*\|^2-\mathbb{E}\big[\|x_m-x^*\|^2\big]\Big).\end{aligned}$$

In order to give a clean proof, we set $\tilde{D}_s \triangleq F(\tilde{x}_s) - F(x^*)$, $P_0^s \triangleq \|x_0^s - x^*\|^2$ and the terms that related to D_m as $\zeta = D_m^2 - D_m$. Then the inequality becomes

$$\frac{1}{\theta}\mathbb{E}\big[\tilde{D}_s\big] \le \frac{1-\theta}{\theta}(1+\zeta)\cdot\tilde{D}_{s-1} + \frac{1}{2\eta m}\mathbb{E}\big[P_0^s - P_m^s\big],$$

$$\frac{\theta+\theta\zeta-\zeta}{\theta}\cdot\mathbb{E}\big[\tilde{D}_s\big] \le \frac{1-\theta}{\theta}(1+\zeta)\cdot\mathbb{E}\big[\tilde{D}_{s-1}-\tilde{D}_s\big] + \frac{1}{2\eta m}\mathbb{E}\big[P_0^s - P_m^s\big].$$

Suppose we run $\mathcal{S}$ epochs before a restart. By summing the above inequality over $s = 1\ldots\mathcal{S}$, we get

$$\frac{\theta+\theta\zeta-\zeta}{\theta}\cdot\sum_{s=1}^{\mathcal{S}}\mathbb{E}\big[\tilde{D}_s\big] \le \frac{1-\theta}{\theta}(1+\zeta)\cdot\mathbb{E}\big[\tilde{D}_0-\tilde{D}_{\mathcal{S}}\big] + \frac{1}{2\eta m}\mathbb{E}\big[P_0^1 - P_m^{\mathcal{S}}\big].$$

Choosing the initial vector for next $\mathcal{S}$ epochs as $x_0^{new} = \frac{1}{\mathcal{S}}\sum_{s=1}^{\mathcal{S}}\tilde{x}_s$, we have

$$\begin{aligned}\mathcal{S}\cdot\frac{\theta+\theta\zeta-\zeta}{\theta}\cdot\mathbb{E}\big[\tilde{D}_0^{new}\big] &\le \frac{1-\theta}{\theta}(1+\zeta)\cdot\mathbb{E}\big[\tilde{D}_0-\tilde{D}_{\mathcal{S}}\big] + \frac{1}{2\eta m}\mathbb{E}\big[P_0^1 - P_m^{\mathcal{S}}\big]\\ &\overset{(\star)}{\le} \Big(\frac{(1-\theta)(1+\zeta)}{\theta} + \frac{1}{\eta m\sigma}\Big)\cdot\tilde{D}_0,\\ \mathbb{E}\big[\tilde{D}_0^{new}\big] &\le \frac{(1-\theta)\cdot(1+\zeta)+\frac{\theta}{\eta m\sigma}}{\mathcal{S}\cdot(\theta+\zeta\theta-\zeta)}\tilde{D}_0,\end{aligned}$$

where $(\star)$ uses the σ-strongly convexity of $F(\cdot)$ and $\tilde{x}_0 = x_0^1$, that is, $P_0^1 \le \frac{2}{\sigma}\tilde{D}_0$.

Setting $\mathcal{S} = \Big\lceil 2\cdot\frac{(1-\theta)\cdot(1+\zeta)+\frac{\theta}{\eta m\sigma}}{\theta+\zeta\theta-\zeta}\Big\rceil$, we have that $\tilde{D}_0$ decreases by a factor of at least $\frac{1}{2}$ every $\mathcal{S}$ rounds of epochs. So in order to achieve an ϵ-additive error, we need totally $O(\log\frac{\tilde{D}_0}{\epsilon})$ restarts of the above procedure.

C.2.1. Parameter setting for two cases

(I) Consider the first case with $\frac{m}{\kappa} \le \frac{3}{4}$, by choosing identical parameters settings $\eta = \sqrt{\frac{1}{3\sigma mL}}$, $\theta = \sqrt{\frac{m}{3\kappa}} \le \frac{1}{2}$ and $m = \Theta(n)$ as in Section 3 (so the constraint (27) is satisfied), we have

$$\mathcal{S} = \bigg\lceil 2\cdot\frac{2-\big(\theta+\zeta\cdot(\theta-1)\big)}{\theta+\zeta\cdot(\theta-1)}\bigg\rceil.$$

Imposing an additional constraint on the sparse variance: $\zeta \le \sqrt{\frac{m}{4\kappa}}$, we have $\mathcal{S} = O(\sqrt{\frac{\kappa}{n}})$, which means that the total oracle complexity is

$$\mathcal{O}\bigg(\mathcal{S}\cdot O\bigg(\log\frac{\tilde{D}_0}{\epsilon}\bigg)\cdot(m+n)\bigg) = \mathcal{O}\bigg(\sqrt{\kappa n}\log\frac{F(\tilde{x}_0)-F(x^*)}{\epsilon}\bigg)$$

(II) Consider another case with $\frac{m}{\kappa} > \frac{3}{4}$ and $\zeta \le C_\zeta$, let $\hat{\theta} \triangleq \frac{C_\zeta+0.5}{C_\zeta+1}$, by choosing $\theta = \frac{\zeta+0.5}{\zeta+1} \in [0.5, \hat{\theta}]$, $\eta = \frac{1-\hat{\theta}}{2m\sigma\hat{\theta}} \le \frac{1-\theta}{L\theta(2-\theta)}$ (the constraint (27) is satisfied) and $m = \Theta(n)$, we have $\mathcal{S} = O(1)$ (correlated with C_ζ), the total oracle complexity $\mathcal{O}\big(n\log\frac{F(\tilde{x}_0)-F(x^*)}{\epsilon}\big)$.

C.2.2. Discussion about Sparse Variance Bound (Lemma 4 and Lemma 5)

In the dense update case (with $D_m = 1$), we see that Lemma 5 degenerates to Lemma 1, this bound is much tighter than the bound in Lemma 4 (Lemma 4 is identical for both the sparse and dense cases).

However, in the sparse update case, Lemma 5 highly correlates with dataset sparsity ($\propto D_m^2$), which could be much looser than Lemma 4 in some extreme cases (imagine a dataset with some dimensions contain only one entry among the n samples, so $D_m = n$). Unfortunately, our MiG algorithm (as well as Katyusha) relies on canceling the additional variance term to yield a tight correlation inside one iteration:

$$F(\tilde{x}_{s-1}) - F(y_{j-1}) - \boxed{\langle\nabla F(y_{j-1}), \tilde{x}_{s-1}-y_{j-1}\rangle} \to \text{canceled by coupling term } (1-\theta)\tilde{x}_s$$

If D_m is as large as n, the oracle complexity in Theorem 4 could be even worse than that in Theorem 3.

It is still an open problem whether we can have a tighter variance bound in the sparse update setting that is uncorrelated with D_m.

C.3. Proof of Theorem 5

Here we analyze Algorithm 4 based on the “perturbed iterate analysis” framework (Mania et al., 2017).

To begin with, we need to specify the iterates labeling order, which is crucial in our asynchronous analysis.

Choice of labeling order. There are “Before Read” (Mania et al., 2017) and “After Read” (Leblond et al., 2017) labeling schemes proposed in recent years which are reasonable in asynchronous analysis. Among these two schemes, “Before Read” requires considering the updates from “future”, which leads to a complex analysis. “After Read” enjoys a simpler analysis but requires changing the order of sampling step to ensure uniform distributed samples[16]. In order to give a clean proof, we adopt the “After Read” labeling scheme and make the following assumptions:

Assumption C.1. *The labeling order increases after the step (14) in Algorithm 4 finished, so the future perturbation is not included in the effect of asynchrony in the current step.*

Assumption C.2. *We explicitly assume uniform distributed samples and the independence of the sample i_j with $\hat{x}_{j-1}$.*

In other words, we are analyzing the following procedure:

1. Inconsistent read the iterate $\hat{x}_{j-1}$.
2. Increase iterates counter j and sample a random index i_j.
3. Compute an update $-\eta \cdot \tilde{\nabla}(\hat{y}_{j-1})$.
4. Atomic write the update to shared memory coordinately.

From (Leblond et al., 2017), we can model the effect of asynchrony as follows:

$$\hat{x}_j - x_j = \eta \sum_{k=(j-1-\tau)_+}^{j-2} \mathcal{S}_k^j \tilde{\nabla}(\hat{y}_k), \tag{32}$$

where $\mathcal{S}_k^j$ is a diagonal matrix with entries in $\{0, +1\}$. This definition models the coordinate perturbation from the past updates. Here τ represents the maximum number of overlaps between concurrent threads (Mania et al., 2017). We further denote $\Delta = \max_{k=1\ldots d} p_k$ following (Leblond et al., 2017), which provides a measure of sparsity.

Then we start our analysis with the iterate difference between “fake” y_j and x^*. By expanding the iterate difference and taking expectation with respect to the sample i_j, we get

$$\begin{aligned}\mathbb{E}_{i_j}\left[\|y_j - x^*\|^2\right] &= \mathbb{E}_{i_j}\left[\|\theta(x_{j-1} - \eta \cdot \tilde{\nabla}(\hat{y}_{j-1})) + (1-\theta)\tilde{x}_{s-1} - x^*\|^2\right] \\ &= \mathbb{E}_{i_j}\left[\|y_{j-1} - \eta\theta \cdot \tilde{\nabla}(\hat{y}_{j-1}) - x^*\|^2\right] \\ &\overset{(\star)}{=} \|y_{j-1} - x^*\|^2 - 2\eta\theta\langle \nabla F(\hat{y}_{j-1}), \hat{y}_{j-1} - x^*\rangle + \eta^2\theta^2\mathbb{E}_{i_j}\left[\|\tilde{\nabla}(\hat{y}_{j-1})\|^2\right] \\ &\quad + 2\eta\theta\mathbb{E}_{i_j}\left[\langle \tilde{\nabla}(\hat{y}_{j-1}), \hat{y}_{j-1} - y_{j-1}\rangle\right],\end{aligned} \tag{33}$$

where $(\star)$ uses the unbiasedness $\mathbb{E}_{i_j}\left[\tilde{\nabla}(\hat{y}_{j-1})\right] = \nabla F(\hat{y}_{j-1})$ and the independence Assumption C.2.

Using Lemma 4, we get the variance bound

$$\mathbb{E}_{i_j}\left[\|\tilde{\nabla}(\hat{y}_{j-1})\|^2\right] \leq 4L\big(F(\hat{y}_{j-1}) - F(x^*)\big) + 4L\big(F(\tilde{x}_{s-1}) - F(x^*)\big). \tag{34}$$

[16] So there are always two versions (analyzed, implemented) of algorithms in the works with “After Read” scheme (Leblond et al., 2017; Pedregosa et al., 2017)

Using the σ-strongly convex of $F(\cdot)$, we can bound $-\langle \nabla F(\hat{y}_{j-1}), \hat{y}_{j-1} - x^*\rangle$ as follows:

$$\begin{aligned}\langle \nabla F(\hat{y}_{j-1}), \hat{y}_{j-1} - x^*\rangle &\geq F(\hat{y}_{j-1}) - F(x^*) + \frac{\sigma}{2}\|\hat{y}_{j-1} - x^*\|^2 \\ &\overset{(\star)}{\geq} F(\hat{y}_{j-1}) - F(x^*) + \frac{\sigma}{4}\|y_{j-1} - x^*\|^2 - \frac{\sigma}{2}\|\hat{y}_{j-1} - y_{j-1}\|^2, \end{aligned} \tag{35}$$

where $(\star)$ uses the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

Combining (33), (34) and (35), we get

$$\begin{aligned}\mathbb{E}_{i_j}\big[\|y_j - x^*\|^2\big] \leq (1 - \frac{\eta\theta\sigma}{2})\|y_{j-1} - x^*\|^2 + \eta\theta\sigma\|\hat{y}_{j-1} - y_{j-1}\|^2 + 2\eta\theta\mathbb{E}_{i_j}\big[\langle\tilde{\nabla}(\hat{y}_{j-1}), \hat{y}_{j-1} - y_{j-1}\rangle\big] \\ + (4L\eta^2\theta^2 - 2\eta\theta)\big(F(\hat{y}_{j-1}) - F(x^*)\big) + 4L\eta^2\theta^2\big(F(\tilde{x}_{s-1}) - F(x^*)\big).\end{aligned} \tag{36}$$

From Lemma 1 in (Leblond et al., 2017), we borrow the notations $C_1 = 1 + \sqrt{\Delta}\tau$, $C_2 = \sqrt{\Delta} + \eta\theta\sigma C_1$ and bound the asynchronous variance terms $\|\hat{y}_{j-1} - y_{j-1}\|^2$, $\mathbb{E}_{i_j}\big[\langle\tilde{\nabla}(\hat{y}_{j-1}), \hat{y}_{j-1} - y_{j-1}\rangle\big]$ using (32) as

$$\mathbb{E}_{i_j}\big[\langle\tilde{\nabla}(\hat{y}_{j-1}), \hat{y}_{j-1} - y_{j-1}\rangle\big] \leq \frac{\eta\theta\sqrt{\Delta}}{2}\sum_{k=(j-1-\tau)_+}^{j-2}\|\tilde{\nabla}(\hat{y}_k)\|^2 + \frac{\eta\theta\sqrt{\Delta}\tau}{2}\mathbb{E}_{i_j}\big[\|\tilde{\nabla}(\hat{y}_{j-1})\|^2\big], \tag{37}$$

$$\|\hat{y}_{j-1} - y_{j-1}\|^2 \leq \eta^2\theta^2(1 + \sqrt{\Delta}\tau)\sum_{k=(j-1-\tau)_+}^{j-2}\|\tilde{\nabla}(\hat{y}_k)\|^2. \tag{38}$$

Upper bounding the asynchronous terms in (36) using (37) and (38), we get

$$\begin{aligned}\mathbb{E}_{i_j}\big[\|y_j - x^*\|^2\big] \leq (1 - \frac{\eta\theta\sigma}{2})\|y_{j-1} - x^*\|^2 + \eta^2\theta^2(\sqrt{\Delta} + \eta\theta\sigma(1 + \sqrt{\Delta}\tau))\sum_{k=(j-1-\tau)_+}^{j-2}\|\tilde{\nabla}(\hat{y}_k)\|^2 \\ + (4L\eta^2\theta^2(1+\sqrt{\Delta}\tau) - 2\eta\theta)\big(F(\hat{y}_{j-1}) - F(x^*)\big) + 4L\eta^2\theta^2(1+\sqrt{\Delta}\tau)\big(F(\tilde{x}_{s-1}) - F(x^*)\big).\end{aligned}$$

Defining $a_j \triangleq \|y_j - x^*\|^2$, $\hat{D}_{j-1} = F(\hat{y}_{j-1}) - F(x^*)$, $\tilde{D}_{s-1} = F(\tilde{x}_{s-1}) - F(x^*)$ for a clean proof and rearranging, we obtain

$$\begin{aligned}\mathbb{E}_{i_j}\big[a_j\big] &\leq (1 - \frac{\eta\theta\sigma}{2})a_{j-1} + \eta^2\theta^2 C_2\sum_{k=(j-1-\tau)_+}^{j-2}\|\tilde{\nabla}(\hat{y}_k)\|^2 + (4L\eta^2\theta^2 C_1 - 2\eta\theta)\hat{D}_{j-1} + 4L\eta^2\theta^2 C_1\tilde{D}_{s-1}, \\ (2\eta\theta - 4L\eta^2\theta^2 C_1)\hat{D}_{j-1} &\overset{(\star)}{\leq} (a_{j-1} - \mathbb{E}_{i_j}\big[a_j\big]) + \eta^2\theta^2 C_2\sum_{k=(j-1-\tau)_+}^{j-2}\|\tilde{\nabla}(\hat{y}_k)\|^2 + 4L\eta^2\theta^2 C_1\tilde{D}_{s-1},\end{aligned} \tag{39}$$

where $(\star)$ uses the fact that $1 - \frac{\eta\theta\sigma}{2} \leq 1$.

Summing (39) over $j = 1 \ldots m$ and taking expectation with all randomness in this epoch, we get

$$(2\eta\theta - 4L\eta^2\theta^2 C_1)\sum_{j=1}^{m}\mathbb{E}\big[\hat{D}_{j-1}\big] \leq (a_0 - \mathbb{E}\big[a_m\big]) + \eta^2\theta^2 C_2\sum_{j=1}^{m}\sum_{k=(j-1-\tau)_+}^{j-2}\mathbb{E}\big[\|\tilde{\nabla}(\hat{y}_k)\|^2\big] + 4L\eta^2\theta^2 C_1 m\tilde{D}_{s-1}. \tag{40}$$

Then we focus on upper bounding the second term on the right side of (40),

$$\sum_{j=1}^{m}\sum_{k=(j-1-\tau)_+}^{j-2}\mathbb{E}\big[\|\tilde{\nabla}(\hat{y}_k)\|^2\big] \leq \tau\sum_{j=1}^{m-1}\mathbb{E}\big[\|\tilde{\nabla}(\hat{y}_{j-1})\|^2\big] \leq \tau\sum_{j=1}^{m}\mathbb{E}\big[\|\tilde{\nabla}(\hat{y}_{j-1})\|^2\big] \overset{(\star)}{\leq} 4L\tau\big(\sum_{j=1}^{m}\mathbb{E}\big[\hat{D}_{j-1}\big] + m\tilde{D}_{s-1}\big),$$

where $(\star)$ uses the variance bound (34).

Substituting the above inequality into (40), we get

$$(2\eta\theta - 4L\eta^2\theta^2 C_1 - 4L\eta^2\theta^2 C_2\tau)\sum_{j=1}^{m}\mathbb{E}\big[\hat{D}_{j-1}\big] \le a_0 + (4L\eta^2\theta^2 C_1 m + 4L\eta^2\theta^2 C_2\tau m)\tilde{D}_{s-1},$$

$$\tilde{D}_s \overset{(\star)}{\le} \frac{\frac{2}{\sigma} + 4L\eta^2\theta^2 C_1 m + 4L\eta^2\theta^2 C_2\tau m}{(2\eta\theta - 4L\eta^2\theta^2 C_1 - 4L\eta^2\theta^2 C_2\tau)m}\cdot\tilde{D}_{s-1},$$

where $(\star)$ uses the σ-strongly convex of $F(\cdot)$ and $\tilde{x}_0 = x_0 = y_0$, $\tilde{x}_s = \frac{1}{m}\sum_{j=0}^{m-1}\hat{y}_j$.

By choosing $m = 60\kappa$, $\eta = \frac{1}{5L}$, $\theta = \frac{1}{6}$, we get

$$\tilde{D}_s \le \frac{2 + \frac{4}{15}(C_1 + C_2\tau)}{4 - \frac{4}{15}(C_1 + C_2\tau)}\cdot\tilde{D}_{s-1}.$$

In order to ensure linear speed up, τ needs to satisfy the following constraint:

$$\rho \triangleq \frac{2 + \frac{4}{15}(C_1 + C_2\tau)}{4 - \frac{4}{15}(C_1 + C_2\tau)} \le 1.$$

By simply setting $\tau \le \min\{\frac{5}{4\sqrt{\Delta}}, 2\kappa, \sqrt{\frac{2\kappa}{\sqrt{\Delta}}}\}$, the above constraint is satisfied with $\rho \le 0.979$, which implies that the total oracle complexity is $\mathcal{O}\big((n+\kappa)\log\frac{F(\tilde{x}_0)-F(x^*)}{\epsilon}\big)$.

D. Experimental Setup

All our algorithms were implemented in C++ and parameters were passed through MATLAB interface for fair comparison[17]. Detailed settings are divided into the following cases.

D.1. In Serial Dense Case

In this case, we ran experiments on the HP Z440 machine with 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 problems with $a_i \in \mathbb{R}^d$, $b_i \in \{-1, +1\}$, $i = 1 \ldots m$:

$$\begin{aligned}\text{Logistic Regression: } & \frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp\left(-b_i a_i^T x\right)\right) + \frac{\lambda}{2}\|x\|^2,\\ \text{Ridge Regression: } & \frac{1}{n}\sum_{i=1}^{n}(a_i^T x + b_i)^2 + \frac{\lambda}{2}\|x\|^2,\\ \text{LASSO: } & \frac{1}{n}\sum_{i=1}^{n}(a_i^T x + b_i)^2 + \lambda\|x\|_1,\end{aligned} \tag{41}$$

where λ is the regularization parameter.

We used datasets from LibSVM website[18], including a9a (32,561 samples, 123 features), covtype.binary (581,012 samples, 54 features), w8a (49,749 samples, 300 features), ijcnn1 (49,990 samples, 22 features). We added one dimension as bias to all the datasets and then normalized all data vectors to 1 for the ease of experimental setup.

We mainly compared MiG with the following state-of-the-art algorithms:

- SVRG. For theoretical evaluation, we set the learning rate as $\frac{1}{4L}$, which is a reasonable learning rate for SVRG in theory. Otherwise in the strongly convex case, we tuned the learning rate. For non-smooth regularizers (e.g., LASSO), we ran Prox-SVRG (Xiao & Zhang, 2014) instead.

[17] The code of our method can be downloaded from the anonymous link: https://www.dropbox.com/s/1a5v1gvioqbjtv3/Async_Sparse_MiG.zip?dl=0.

[18] https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/

- SAGA. For theoretical evaluation, we set the learning rate as $\frac{1}{2(\sigma n+L)}$ following (Defazio et al., 2014). Otherwise (include the non-strongly convex case), we tuned the learning rate.

- Acc-Prox-SVRG. This algorithm is quite unstable if mini-batch size is set to 1, we set same learning rate as in SVRG and tuned the momentum parameter β (Nitanda, 2014).

- Catalyst on SVRG. We set the same learning rate as in SVRG and carefully tuned the parameters α_0 and κ (Lin et al., 2015).

- Katyusha. As suggested by the author, we fixed $\tau_2 = \frac{1}{2}$ (sometimes we tuned τ_2 for a better performance), set $\eta = \frac{1}{3\tau_1 L}$ and tuned only τ_1 (Allen-Zhu, 2017). For theoretical evaluation, we chose $\tau_1 = \sqrt{\frac{m}{3\kappa}}$. For Katyusha$^{\text{ns}}$, we used $\tau_1 = \frac{2}{s+4}$, $\alpha = \frac{1}{a\tau_1 L}$ and tune a.

- MiG. Similarly, we set $\eta = \frac{1}{3\theta L}$ and tuned only θ. For theoretical evaluation, we chose $\theta = \sqrt{\frac{m}{3\kappa}}$. For MiG$^{\text{NSC}}$, we chose $\theta = \frac{2}{s+4}$, $\eta = \frac{1}{aL\theta}$ and tuned a.

More theoretical evaluation results for $\ell 2$-logistic regression and ridge regression problems are shown in Figures 1 and 2, respectively, where the regularization parameter was set to some relatively small values, e.g., 10^{-6}, 10^{-7}, and 10^{-8}.

More practical evaluation results (with parameter tuning for all the algorithms and for relatively large λ) are shown in Figures 3 and 4.

We also compared the performance of MiG$^{\text{NSC}}$ with that of all the algorithms mentioned above, as shown in Figures 5 and 6, where the results are for the non-strongly convex logistic regression (i.e., $\lambda = 0$) and LASSO, respectively.

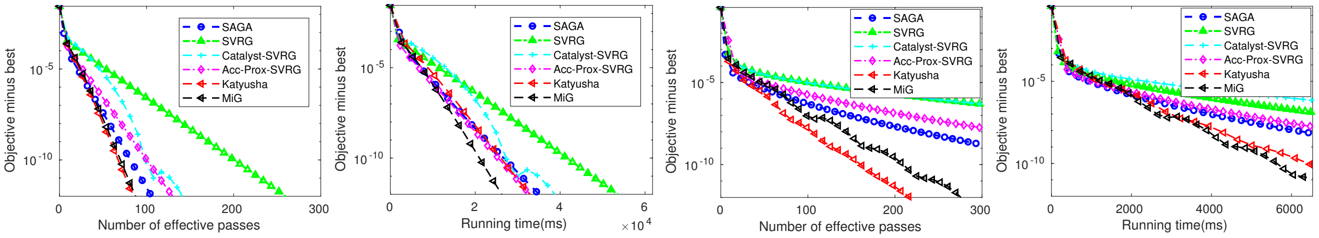


Figure 1. Theoretical evaluations of MiG and other state-of-the-art algorithms for solving $\ell 2$-logistic regression on covtype ($\lambda = 10^{-7}$, the first two figures) and a9a ($\lambda = 10^{-7}$, the last two figures).

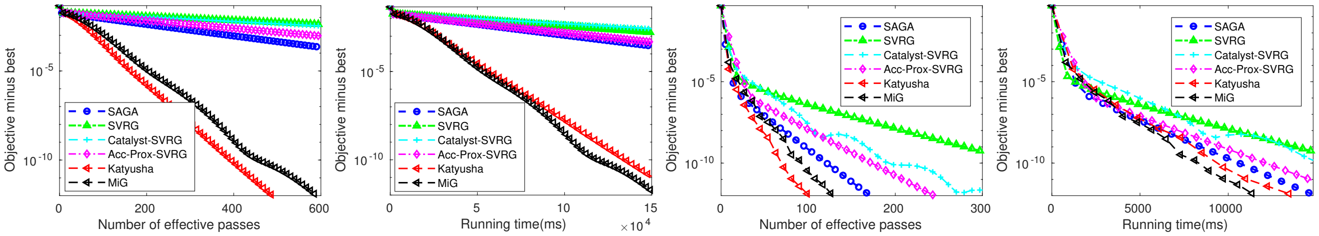
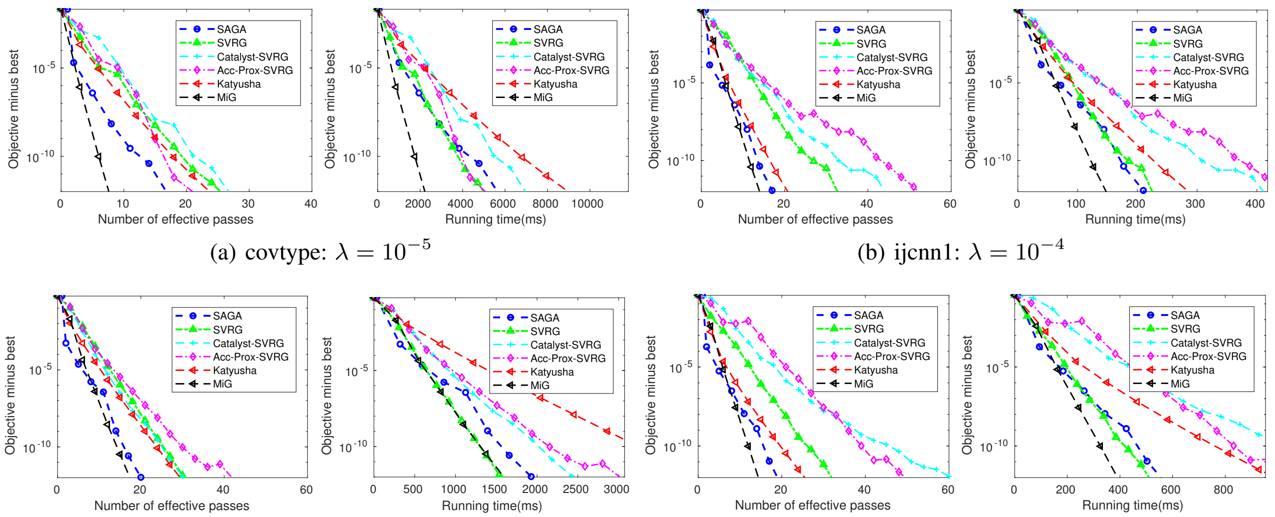


Figure 2. Theoretical evaluations of MiG and other state-of-the-art algorithms for solving ridge regression on covtype ($\lambda = 10^{-8}$, the first two figures) and w8a ($\lambda = 10^{-6}$, the last two figures).



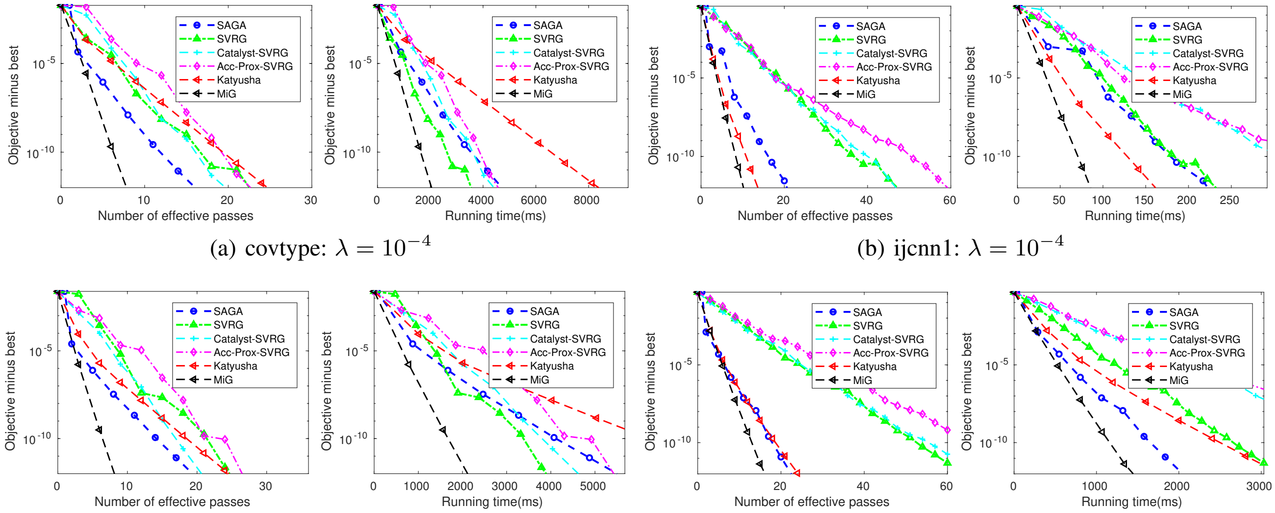
(a) covtype: $\lambda = 10^{-5}$

(b) ijcnn1: $\lambda = 10^{-4}$

(c) w8a: $\lambda = 10^{-4}$

(d) a9a: $\lambda = 10^{-4}$

Figure 3. Practical evaluations of MiG and other state-of-the-art algorithms for solving $\ell 2$-logistic regression on covtype (a), ijcnn1 (b), w8a (c), and a9a (d).



(a) covtype: $\lambda = 10^{-4}$

(b) ijcnn1: $\lambda = 10^{-4}$

(c) covtype: $\lambda = 10^{-5}$

(d) w8a: $\lambda = 10^{-4}$

Figure 4. Practical evaluations of MiG and other state-of-the-art algorithms for solving ridge regression on covtype (a), ijcnn1 (b), covtype (c), and w8a (d).

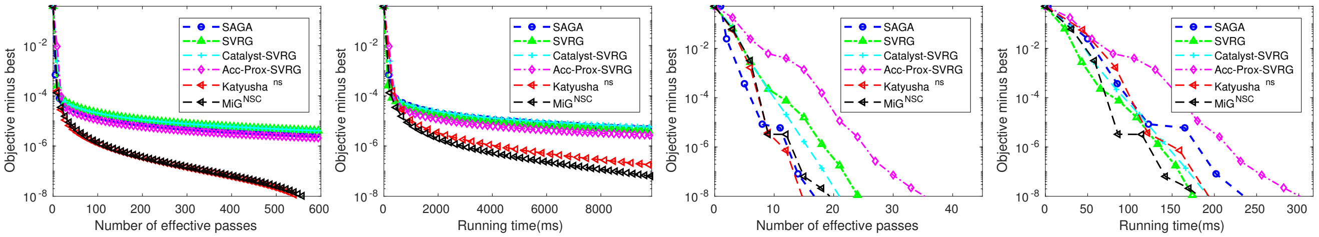


Figure 5. Comparison of MiG$^{\text{NSC}}$ and other state-of-the-art algorithms for solving non-strongly convex logistic regression without regularizer (i.e., $\lambda = 0$) on a9a (the first two figures) and ijcnn1 (the last two figures).

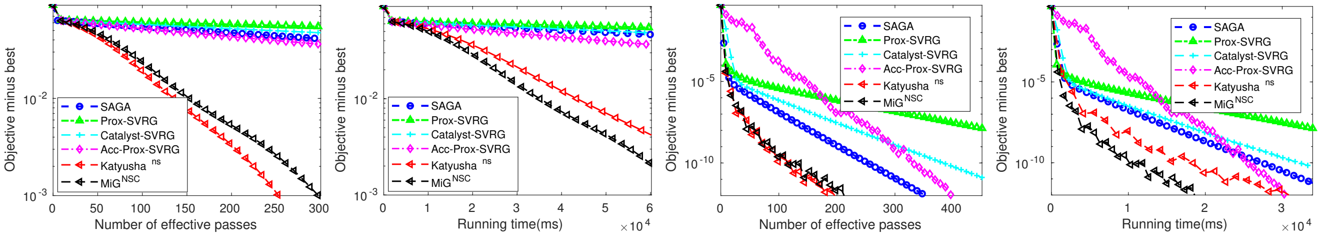


Figure 6. Comparison of MiG$^{\text{NSC}}$ and other state-of-the-art algorithms for solving the non-strongly convex problem (LASSO) on covtype ($\lambda = 10^{-6}$, the first two figures) and w8a ($\lambda = 10^{-7}$, the last two figures).

D.2. In Asynchronous Sparse Case

Experiments in this setting were running on a server with 4 Intel Xeon E7-4820 v3 with 10 cores each 1.90GHz, 512 GB RAM, CentOS 7.4.1708 with GCC 4.8.5, MATLAB R2017a. Multi-threads experiments were running on certain CPU cores using numactl. We are optimizing the following smooth and strongly convex problem:

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

where λ is the regularization parameter. Notice that the $\ell 2$ regularizer is dense with respect to the sample i, so we used the following sparse and unbiased stochastic gradient as in (Leblond et al., 2017):

$$\nabla_i(x) = \nabla f_i(x) + \lambda D_i x.$$

We tuned learning rates for ASAGA and KroMagnon, learning rate and θ for MiG.

D.2.1. Parameter Tuning Criteria

For the relatively small RCV1 dataset, we carefully tuned the learning rate for ASAGA and KroMagnon to achieve best performance. For MiG, following theoretical intuition in the dense case, we first chose a large learning rate and then carefully tuned θ to achieve best performance (thus the effect of θ can be regard as stabilize iterates and allow for a larger learning rate).

For the KDD2010 dataset, we first made 3 subsets with 19,000 samples, 190,000 samples and 1,900,000 samples (corresponds to $\lambda = 10^{-7}$, 10^{-8} and 10^{-9}). Then we tuned parameters for them following the above criteria and finally used relatively "safe" parameter settings for the entire dataset.

D.2.2. On the Effectiveness of Our Acceleration Trick

In this part, we evaluated the effect of our acceleration trick in MiG. One can easily verify that when θ is set to 1, the coupling term is neglected and the algorithm is quite similar to SVRG (different on the choice of the snapshot point). Based on the theoretical analysis in Section 3 and Section 4.1, we claim that the effect of θ is to stabilize the iterates to adopt a larger learning rate. Thus, we designed an experiment to justify this claim.

We compared the best-tuned performance of MiG, MiG with $\theta = 1$, and KroMagnon on RCV1 on a server with the 16 threads, as shown in Figure 7.

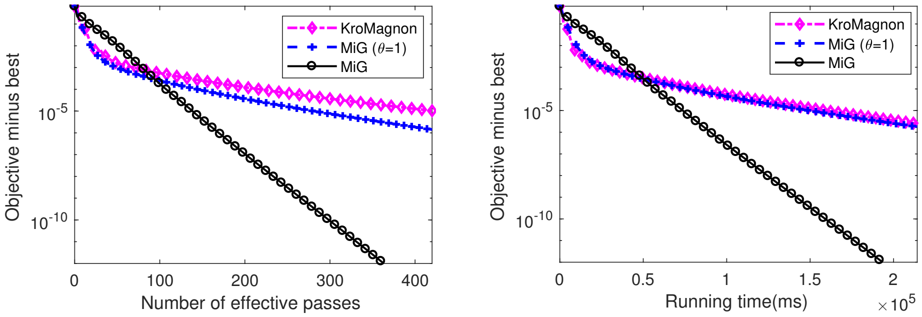


Figure 7. Evaluation of the effectiveness of our acceleration trick of our algorithm for solving ℓ2-logistic regression with $\lambda = 10^{-9}$.

Based on the parameter choice in this experiment, MiG with $\theta = 1$ cannot use the same large learning rate as MiG does. When the suitable parameter θ is chosen, MiG outperforms both algorithms as desired, which further indicates the effectiveness of our acceleration trick.