Supplemental Material for "Accelerated Stochastic Block Coordinate Descent with Optimal Sampling"

B. PROOF OF LEMMA A.1

PROOF. Recall Assumption 3.3 that $R(\mathbf{w})$ is block separable. We first define

$$\mathrm{prox}_\eta(\mathbf{w}) = \big[\mathrm{prox}_{\eta,1}(\mathbf{w}_{\mathcal{G}_1})^\top, \ldots, \mathrm{prox}_{\eta,j}(\mathbf{w}_{\mathcal{G}_j})^\top\big]^\top, \tag{B.1}$$

$$\mathbf{g}_{i,\mathcal{G}_j} = \nabla_{\mathcal{G}_j} f_i(\boldsymbol{\phi}_i^{(t)}) - \nabla_{\mathcal{G}_j} f_i(\boldsymbol{\phi}_i^{(t-1)}) + \frac{1}{n}\sum_{k=1}^{n} \nabla_{\mathcal{G}_j} f_k(\boldsymbol{\phi}_k^{(t-1)}), \tag{B.2}$$

$$\boldsymbol{\delta}^{\mathcal{G}_j} = \Big[0, \ldots, 0, \mathrm{prox}_{\eta,j}(\mathbf{w}_{\mathcal{G}_j}^{(t-1)} - \eta \mathbf{g}_{i,\mathcal{G}_j})^\top - \mathrm{prox}_{\eta,j}\big(\mathbf{w}^*_{\mathcal{G}_j} - \eta \nabla_{\mathcal{G}_j} F(\mathbf{w}^*)\big)^\top, 0, \ldots, 0\Big]^\top, \tag{B.3}$$

$$\text{and} \quad \boldsymbol{\delta} = \mathrm{prox}_\eta(\mathbf{w}^{(t-1)} - \eta \mathbf{g}_i) - \mathrm{prox}_\eta\big(\mathbf{w}^* - \eta \nabla F(\mathbf{w}^*)\big). \tag{B.4}$$

Since $R(\mathbf{w})$ is block separable, $\boldsymbol{\delta}^{\mathcal{G}_j}$ and $\boldsymbol{\delta}^{\mathcal{G}_{j'}}$ are orthogonal to each other for all $j \neq j'$, and by (B.3) and (B.4) we have

$$\mathbb{E}_j\big[\|\boldsymbol{\delta}^{\mathcal{G}_j}\|^2\big] = \frac{1}{m}\sum_{j=1}^{m} \|\boldsymbol{\delta}^{\mathcal{G}_j}\|^2 = \frac{\|\boldsymbol{\delta}\|^2}{m}. \tag{B.5}$$

Similarly, for convenience of technical discussions we further define

$$\boldsymbol{\psi}^{\mathcal{G}_j} = \big[0, \ldots, 0, (\mathbf{w}_{\mathcal{G}_j}^{(t-1)} - \mathbf{w}^*_{\mathcal{G}_j})^\top, 0, \ldots, 0\big]^\top \tag{B.6}$$

$$\text{and} \quad \boldsymbol{\psi} = \mathbf{w}^{(t-1)} - \mathbf{w}^*, \tag{B.7}$$

then we are able to obtain their relation:

$$\mathbb{E}_j\big[\|\boldsymbol{\psi}^{\mathcal{G}_j}\|^2\big] = \frac{1}{m}\sum_{j=1}^{m} \|\boldsymbol{\psi}^{\mathcal{G}_j}\|^2 = \frac{\|\boldsymbol{\psi}\|^2}{m}. \tag{B.8}$$

From the definition in (B.2), by exploiting the block separability of $R(\mathbf{w})$, we have

$$\mathbb{E}_j\big[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2\big] = \sum_{k \neq j} \mathbb{E}_k\big[\|\mathbf{w}_{\mathcal{G}_k}^{(t-1)} - \mathbf{w}^*_{\mathcal{G}_k}\|^2\big] + \mathbb{E}_j\Big[\Big\|\mathrm{prox}_{\eta,j}(\mathbf{w}_{\mathcal{G}_j}^{(t-1)} - \eta \mathbf{g}_{i,\mathcal{G}_j}) - \mathrm{prox}_{\eta,j}\big(\mathbf{w}^*_{\mathcal{G}_j} - \eta \nabla_{\mathcal{G}_j} F(\mathbf{w}^*)\big)\Big\|^2\Big].$$

After substitution with (B.3), (B.4), (B.6), and (B.7), according to (B.5) and (B.8), since

$$\sum_{k \neq j} \mathbb{E}_k\big[\|\boldsymbol{\psi}^{\mathcal{G}_k}\|^2\big] + \mathbb{E}_j\big[\|\boldsymbol{\delta}^{\mathcal{G}_j}\|^2\big] = \frac{(m-1)\|\boldsymbol{\psi}\|^2}{m} + \frac{\|\boldsymbol{\delta}\|^2}{m},$$

by the non-expansiveness of the proximal operator (B.1) [32] and that $\mathbf{w}^*$ is the optimal value in (1.1),

$$\begin{aligned}
&\mathbb{E}_j\big[\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2\big] \\
&= \frac{(m-1)}{m}\|\mathbf{w}^{(t-1)} - \mathbf{w}^*\|^2 + \frac{1}{m}\Big\|\mathrm{prox}_\eta(\mathbf{w}^{(t-1)} - \eta \mathbf{g}_i) - \mathrm{prox}_\eta\big(\mathbf{w}^* - \eta \nabla F(\mathbf{w}^*)\big)\Big\|^2 \\
&\leq \frac{1}{m}\big[(m-1)\|\mathbf{w}^{(t-1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(t-1)} - \eta \mathbf{g}_i - \mathbf{w}^* + \eta \nabla F(\mathbf{w}^*)\|^2\big].
\end{aligned} \tag{B.9}$$

□

C. PROOF OF LEMMA A.2

PROOF. The proof is straightforward using the definition of $\mathbf{g}_i$ in (A.1).

$$\begin{aligned}
\mathbb{E}_i[\mathbf{g}_i] &= \mathbb{E}_i\Big[\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t)}) - \frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)})\Big] + \frac{1}{n}\sum_{k=1}^{n} \nabla f_k(\boldsymbol{\phi}_k^{(t-1)}) \\
&= \sum_{i=1}^{n} \frac{p_i}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \sum_{i=1}^{n} \frac{p_i}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) + \frac{1}{n}\sum_{k=1}^{n} \nabla f_k(\boldsymbol{\phi}_k^{(t-1)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) + \frac{1}{n}\sum_{k=1}^{n} \nabla f_k(\boldsymbol{\phi}_k^{(t-1)}) \\
&= \nabla F(\mathbf{w}^{(t-1)}).
\end{aligned}$$

□

D. PROOF OF LEMMA A.3

PROOF. To prove Lemma A.3, we begin by computing $\mathbb{E}_i[\mathbf{g}_i - \nabla F(\mathbf{w}^*)]$ with $\mathbf{g}_i$ defined in (A.1) and Lemma A.2:

$$\mathbb{E}_i[\mathbf{g}_i - \nabla F(\mathbf{w}^*)] = \nabla F(\mathbf{w}^{(t-1)}) - \nabla F(\mathbf{w}^*). \tag{D.1}$$

By variance decomposition that $\mathbb{E}\big[\, \|\mathbf{x}\|^2 \,\big] = \mathbb{E}\big[\, \|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|^2 \,\big] + \big\|\mathbb{E}[\mathbf{x}]\big\|^2$ for all $\mathbf{x}$, using (D.1),

$$\begin{aligned}
&\mathbb{E}_i\big[\, \|\mathbf{g}_i - \nabla F(\mathbf{w}^*)\|^2 \,\big] \\
&= \mathbb{E}_i\Big[\big\|\mathbf{g}_i - \nabla F(\mathbf{w}^*) - \mathbb{E}_i[\mathbf{g}_i - \nabla F(\mathbf{w}^*)]\big\|^2\Big] + \Big\|\mathbb{E}_i\big[\mathbf{g}_i - \nabla F(\mathbf{w}^*)\big]\Big\|^2 \\
&= \mathbb{E}_i\Bigg[\bigg\|\Big[\frac{1}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) - \nabla F(\mathbf{w}^{(t-1)}) + \nabla F(\mathbf{w}^*)\Big] \\
&\quad - \Big[\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) + \nabla F(\mathbf{w}^*) - \frac{1}{n}\sum_{k=1}^{n} f_k(\boldsymbol{\phi}_k^{(t-1)})\Big]\bigg\|^2\Bigg] + \|\nabla F(\mathbf{w}^{(t-1)}) - \nabla F(\mathbf{w}^*)\|^2.
\end{aligned} \tag{D.2}$$

Applying the property that $\|\mathbf{x} + \mathbf{y}\|^2 \le (1+\zeta)\,\|\mathbf{x}\|^2 + (1+\zeta^{-1})\,\|\mathbf{y}\|^2$ for all $\mathbf{x}, \mathbf{y}$, and $\zeta > 0$ to (D.2),

$$\begin{aligned}
\mathbb{E}_i\big[\, \|\mathbf{g}_i - \nabla F(\mathbf{w}^*)\|^2 \,\big] &\le \|\nabla F(\mathbf{w}^{(t-1)}) - \nabla F(\mathbf{w}^*)\|^2 \\
&+ (1+\zeta)\mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) - \nabla F(\mathbf{w}^{(t-1)}) + \nabla F(\mathbf{w}^*)\Big\|^2\Big] \\
&+ (1+\zeta^{-1})\mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) + \nabla F(\mathbf{w}^*) - \frac{1}{n}\sum_{k=1}^{n} f_k(\boldsymbol{\phi}_k^{(t-1)})\Big\|^2\Big].
\end{aligned} \tag{D.3}$$

To simplify terms on the right-hand side of (D.3) using variance decomposition, we have

$$\begin{aligned}
&\mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) - \nabla F(\mathbf{w}^{(t-1)}) + \nabla F(\mathbf{w}^*)\Big\|^2\Big] \\
&= \mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) - \mathbb{E}_i\big[\nabla f_i(\mathbf{w}^{(t-1)}) - \nabla f_i(\mathbf{w}^*)\big]\Big\|^2\Big] \\
&= \mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*)\Big\|^2\Big] - \Big\|\mathbb{E}_i\big[\nabla f_i(\mathbf{w}^{(t-1)}) - \nabla f_i(\mathbf{w}^*)\big]\Big\|^2 \\
&= \mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*)\Big\|^2\Big] - \big\|\nabla F(\mathbf{w}^{(t-1)}) - \nabla F(\mathbf{w}^*)\big\|^2,
\end{aligned} \tag{D.4}$$

and we obtain the following inequality by dropping a non-positive term:

$$\begin{aligned}
&\mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) + \nabla F(\mathbf{w}^*) - \frac{1}{n}\sum_{k=1}^{n} f_k(\boldsymbol{\phi}_k^{(t-1)})\Big\|^2\Big] \\
&= \mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*) - \mathbb{E}_i\big[\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \nabla f_i(\mathbf{w}^*)\big]\Big\|^2\Big] \\
&= \mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*)\Big\|^2\Big] - \Big\|\mathbb{E}_i\big[\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \nabla f_i(\mathbf{w}^*)\big]\Big\|^2 \\
&\le \mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*)\Big\|^2\Big].
\end{aligned} \tag{D.5}$$

Plugging (D.4) and (D.5) into (D.3), we complete the proof with

$$\begin{aligned}
\mathbb{E}_i\big[\|\mathbf{g}_i - \nabla F(\mathbf{w}^*)\|^2\big] \le (1+\zeta)\mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\mathbf{w}^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*)\Big\|^2\Big] - \zeta\|\nabla F(\mathbf{w}^{(t-1)}) - \nabla F(\mathbf{w}^*)\|^2 \\
+ (1+\zeta^{-1})\mathbb{E}_i\Big[\Big\|\frac{1}{np_i}\nabla f_i(\boldsymbol{\phi}_i^{(t-1)}) - \frac{1}{np_i}\nabla f_i(\mathbf{w}^*)\Big\|^2\Big].
\end{aligned}$$

□

E. PROOF OF LEMMA A.4

PROOF. For the convenience of this proof, we first define a function

$$h(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\,\|\mathbf{x}\|^2\,. \tag{E.1}$$

Recall that f is strongly convex with the convexity parameter μ and its gradient is Lipschitz continuous with the constant L. By twice differentiating $h(\mathbf{w})$, we obtain that the gradient of h is Lipschitz continuous with the constant $L-\mu$.

By the property of f that is convex and has a Lipschitz continuous gradient: $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \right\|^2 / (2L)$ for all $\mathbf{x}$ and $\mathbf{y}$ [30] (Theorem 2.1.5), we have

$$h(\mathbf{x}) \geq h(\mathbf{y}) + \langle \nabla h(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{1}{2(L-\mu)} \left\| \nabla h(\mathbf{x}) - \nabla h(\mathbf{y}) \right\|^2.$$

By substitution of $h(\mathbf{x})$ according to (E.1),

$$\begin{aligned} f(\mathbf{x}) - \frac{\mu}{2}\left\|\mathbf{x}\right\|^2 &\geq f(\mathbf{y}) - \frac{\mu}{2}\left\|\mathbf{y}\right\|^2 + \langle \nabla f(\mathbf{y}) - \mu\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle \\ &+ \frac{1}{2(L-\mu)}\left[\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|^2 + \mu^2\left\|\mathbf{y}-\mathbf{x}\right\|^2 + 2\mu\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{y}-\mathbf{x}\rangle\right]. \end{aligned}$$

Re-arranging terms gives the following relation:

$$\begin{aligned} \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle &\leq f(\mathbf{x}) - f(\mathbf{y}) - \frac{1}{2(L-\mu)}\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|^2 - \frac{\mu}{L-\mu}\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{y}-\mathbf{x}\rangle \\ &- \Big(\frac{\mu}{2}\left\|\mathbf{x}\right\|^2 - \frac{\mu}{2}\left\|\mathbf{y}\right\|^2 - \mu\langle \mathbf{y}, \mathbf{x}-\mathbf{y}\rangle\Big) - \frac{\mu^2}{2(L-\mu)}\left\|\mathbf{y}-\mathbf{x}\right\|^2. \end{aligned} \tag{E.2}$$

After simplifying terms on the right-hand side of (E.2) by

$$\begin{aligned} &\frac{\mu}{2}\left\|\mathbf{x}\right\|^2 - \frac{\mu}{2}\left\|\mathbf{y}\right\|^2 - \mu\langle \mathbf{y}, \mathbf{x}-\mathbf{y}\rangle \\ &= \frac{\mu}{2}\left\|\mathbf{x}\right\|^2 - \frac{\mu}{2}\left\|\mathbf{y}\right\|^2 - \mu\langle \mathbf{x}, \mathbf{y}\rangle + \mu\left\|\mathbf{y}\right\|^2 \\ &= \frac{\mu}{2}\left\|\mathbf{y}-\mathbf{x}\right\|^2, \end{aligned}$$

we are able to obtain the conclusion of Lemma A.4:

$$\langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \leq f(\mathbf{x}) - f(\mathbf{y}) - \frac{1}{2(L-\mu)}\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|^2 - \frac{\mu}{L-\mu}\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{y}-\mathbf{x}\rangle - \frac{L\mu}{2(L-\mu)}\left\|\mathbf{y}-\mathbf{x}\right\|^2.$$

□

F. PROOF OF LEMMA A.5

PROOF. Recall that in Algorithm 1, at each iteration one component function f_i is sampled at probability p_i from n functions. Thus,

$$\mathbb{E}_i[f_i(\boldsymbol{\phi}_i^{(t)})] = p_i f_i(\boldsymbol{\phi}_i^{(t)}) + (1-p_i) f_i(\boldsymbol{\phi}_i^{(t-1)}). \tag{F.1}$$

Plugging (F.1) and $\boldsymbol{\phi}_i^{(t)} = \mathbf{w}^{(t-1)}$ into $\mathbb{E}_i[n^{-1}\sum_{i=1}^n L_i(np_i)^{-1} f_i(\boldsymbol{\phi}_i^{(t)})]$, we obtain

$$\begin{aligned} &\mathbb{E}_i\Big[\frac{1}{n}\sum_{i=1}^n \frac{L_i}{np_i} f_i(\boldsymbol{\phi}_i^{(t)})\Big] \\ &= \frac{1}{n}\sum_{i=1}^n p_i \frac{L_i}{np_i} f_i(\mathbf{w}^{(t-1)}) + \frac{1}{n}\sum_{i=1}^n (1-p_i)\frac{L_i}{np_i} f_i(\boldsymbol{\phi}_i^{(t-1)}) \\ &= \frac{1}{n}\sum_{i=1}^n \frac{L_i}{n} f_i(\mathbf{w}^{(t-1)}) + \frac{1}{n}\sum_{i=1}^n \frac{(1-p_i)L_i}{np_i} f_i(\boldsymbol{\phi}_i^{(t-1)}). \end{aligned}$$

□