## SUPPLEMENTAL MATERIALS

**Lemma 5.** *[15] If $f(\mathbf{x})$ is $\beta$-strongly convex and $\mathbf{x}_*$ denotes the optimal solution to $\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$. For any $\mathbf{x} \in \mathcal{D}$, we have $f(\mathbf{x}) - f(\mathbf{x}_*) \leq 2G_1^2/\beta$.*

*Proof.* From Assumption A1, we have $\|\partial f(\mathbf{x})\|_2 \leq G_1$. Hence

$$f(\mathbf{x}) - f(\mathbf{x}_*) \leq G_1\|\mathbf{x} - \mathbf{x}_*\|_2.$$

Moreover from the strong convexity in $f(\cdot)$ we have

$$f(\mathbf{x}) - f(\mathbf{x}_*) \geq \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}_*\|_2^2.$$

From the two inequalities above, we can easily verify that

$$\|\mathbf{x} - \mathbf{x}_*\|_2 \leq \frac{2G_1}{\beta}, \;\; f(\mathbf{x}) - f(\mathbf{x}_*) \leq \frac{2G_1^2}{\beta}.$$

This completes the proof. □

**Proof of Theorem 2**

The proof of Theorem 2 is based on an important result, as summarized in Lemma 6.

**Lemma 6.** *[20] Assume $\|\mathbf{x}_* - \mathbf{x}_t\|_2 \leq D$ for all $t$. Define $D_T = \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}\|_2^2$ and $\Lambda_T = \sum_{t=1}^T \zeta_t(\mathbf{x})$. We have*

$$\begin{aligned}\Pr&\left(\Lambda_T \leq 4G_1\sqrt{D_T \ln\frac{m}{\epsilon}} + 2G_1 D\ln\frac{m}{\epsilon}\right) \\ &+ \Pr\left(D_T \leq \frac{D^2}{T}\right) \geq 1-\epsilon,\end{aligned}$$

*where $m = \lceil 2\log_2 T\rceil$ and $\sum_{t=1}^T \zeta_t(\mathbf{x}) = \sum_{t=1}^T (\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t))^\top(\mathbf{x} - \mathbf{x}_t)$.*

*Proof of Theorem 2* The proof below follows from techniques used in Lemma 2 and Theorem 1. Since $F(\mathbf{x})$ is $\beta$-strongly convex, we have

$$F(\mathbf{x}_t) - F(\mathbf{x}) \leq (\mathbf{x}_t - \mathbf{x})^\top \nabla F(\mathbf{x}_t) - \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2.$$

Combining the above inequality with the inequality in (8) and taking summation over all $t = 1, \ldots, T$, we have

$$\begin{aligned}\sum_{t=1}^T (F(\mathbf{x}_t) - F(\mathbf{x})) \;\;&\leq\;\; \underbrace{\frac{\|\mathbf{x}_1 - \mathbf{x}\|_2^2}{2\eta} + \eta T(G_1^2 + \lambda^2 G_2^2)}_{BT} \\ &\quad + \sum_{t=1}^T \zeta_t(\mathbf{x}) - \frac{\beta}{2}D_T. \qquad (23)\end{aligned}$$

We substitute the bound in Lemma 6 into the above inequality with $\mathbf{x} = \mathbf{x}^*$. We consider two cases. In the first case, we assume $D_T \leq D^2/T$. As a result, we have

$$\begin{aligned}\sum_{t=1}^T \zeta_t(\mathbf{x}^*) &= \sum_{t=1}^T (\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t))^\top(\mathbf{x}^* - \mathbf{x}_t) \\ &\leq 2G_1\sqrt{TD_T} \leq 2G_1 D,\end{aligned}$$

which together with the inequality in (23) leads to the bound

$$\sum_{t=1}^T (F(\mathbf{x}_t) - F(\mathbf{x}^*)) \leq 2G_1 D + BT.$$

In the second case, we assume

$$\begin{aligned}\sum_{t=1}^T \zeta_t(\mathbf{x}^*) &\leq 4G_1\sqrt{D_T\ln\frac{m}{\epsilon}} + 4G_1\ln\frac{m}{\epsilon} \\ &\leq \frac{\beta}{2}D_T + \left(\frac{8G_1^2}{\beta} + 4G_1\right)\ln\frac{m}{\epsilon},\end{aligned}$$

where the last step uses the fact $2\sqrt{ab} \leq a^2 + b^2$. We thus have

$$\sum_{t=1}^T (F(\mathbf{x}_t) - F(\mathbf{x}^*)) \leq \left(\frac{8G_1^2}{\beta} + 2G_1 D\right)\ln\frac{m}{\epsilon} + BT$$

Combing the results of the two cases, we have, with a probability $1 - \epsilon$,

$$\begin{aligned}\sum_{t=1}^T (F(\mathbf{x}_t) - F(\mathbf{x}^*)) &\leq \left(\frac{8G_1^2}{\beta} + 2G_1 D\right)\ln\frac{m}{\epsilon} \\ &\quad + 2G_1 D + BT,\end{aligned}$$

where $C = \left(\frac{8G_1^2}{\beta} + 2G_1 D\right)\ln\frac{m}{\epsilon} + 2G_1 D$. Following the same analysis, we have

$$f(\widetilde{\mathbf{x}}_T) - f(\mathbf{x}_*) \leq \frac{\mu C}{T} + \frac{\mu\|\mathbf{x}_1 - \mathbf{x}_*\|_2^2}{2\eta T} + \mu\eta G^2$$

Let $\Delta_k = f(\mathbf{x}_k^1) - f(\mathbf{x}_*)$. By induction, we have

$$\Delta_{k+1} \leq \frac{\mu C}{T_k} + \frac{\mu\Delta_k}{2\eta_k T_k \beta} + \mu\eta_k G^2$$

Assume $\Delta_k \leq V_k \triangleq \frac{\mu^2 G^2}{2^{k-2}\beta}$, by plugging the values of $\eta_k, T_k$, we have

$$\Delta_{k+1} \leq \frac{V_k}{6} + \frac{V_k}{6} + \frac{V_k}{6} = \frac{V_k}{2} = V_{k+1}$$

where we use $T_1 \geq \max\left(\frac{3C\beta}{\mu G^2}, 9\right)$ and $T_k \geq \max\left(\frac{6\mu c}{V_k}, \frac{18\mu^2 G^2}{V_k \beta}\right)$ and $\eta_k = \frac{V_k}{6\mu G^2} = \frac{2\mu}{2^k(3\beta)}$. This completes the proof of this theorem.

## Proof of Lemma 3

To prove Lemma 3, we derive an inequality similar to Eq. (8); the rest proof of Lemma 3 is similar to that of Lemma 2.

**Corollary 1.** *Given a $\beta$-strongly convex function $\widehat{f}(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, and a sequence $\{\mathbf{x}_t\}$ defined by the update $\mathbf{x}_{t+1} = \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - (\mathbf{x}_t - \eta \mathbf{g}(\mathbf{x}_t))\|_2^2 + \eta g(\mathbf{x})$. Then for any $\mathbf{x}$, we have*

$$\begin{aligned}
& \sum_{t=1}^{T} [f(\mathbf{x}_t) + g(\mathbf{x}_{t+1}) - f(\mathbf{x}) - g(\mathbf{x})] \\
\leq \quad & \frac{\|\mathbf{x} - \mathbf{x}_1\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{g}(\mathbf{x}_t)\|_2^2 + \sum_{t=1}^{T} (\mathbf{x} - \mathbf{x}_t)^\top (\mathbf{g}(\mathbf{x}_t) \\
& -\nabla f(\mathbf{x}_t)) - \frac{\beta}{2}\sum_{t=1}^{T} \|\mathbf{x} - \mathbf{x}_{t+1}\|_2^2.
\end{aligned}$$

Corollary 1 can be proved using techniques similar to the ones in [9] but with extra care on the stochastic gradient. As a consequence we have

$$\begin{aligned}
& \frac{1}{T}\mathrm{E}\left[\sum_{t=1}^{T} \hat{f}(\mathbf{x}_t) - \hat{f}(\mathbf{x})\right] \\
\leq \quad & \frac{\mathrm{E}[\|\mathbf{x} - \mathbf{x}_1\|_2^2]}{2\eta T} + \eta(G_1^2 + \lambda G_2^2) + \frac{g(\mathbf{x}_1) - g(\mathbf{x}_{T+1})}{T}
\end{aligned}$$

## Proof of Lemma 4

The lemma is a corollary of results in [6] for general convex optimization. In particular, if we consider the stochastic composite optimization

$$F(\mathbf{x}) = \phi(\mathbf{x}) + g(\mathbf{x})$$

where $g(\mathbf{x})$ is a simple function such that its proximal mapping can be easily solved and $\phi(\mathbf{x})$ is only accessible through a stochastic oracle that returns a stochastic subgradient $\mathbf{g}(\mathbf{x})$. To state the convergence of ORDA for general convex problems, [6] makes the following assumptions: (i) $\mathrm{E}[\|\mathbf{g}(\mathbf{x}) - \mathrm{E}\mathbf{g}(\mathbf{x})\|_2^2] \leq \sigma^2$ and (ii)

$$\phi(\mathbf{y}) - \phi(\mathbf{x}) - (\mathbf{y} - \mathbf{x})^\top \partial\phi(\mathbf{x}) \leq M\|\mathbf{y} - \mathbf{x}\|_2$$

When $\|\partial\phi(\mathbf{x})\|_2 \leq G$, the first inequality holds $\sigma = G$ and the second inequality holds with $M = 2G$. Applying to the augmented objective

$$F(\mathbf{x}) = f(\mathbf{x}) + \lambda[c(\mathbf{x})]_+ + g(\mathbf{x})$$

We note that $\sigma = G_1$ and $M = 2(G_1 + \lambda G_2)$. Follow the inequality (26) in the appendix of [6], we obtain that

$$\mathrm{E}[F(\mathbf{x}_{T+2}) - F(\mathbf{x}_*)] \leq \frac{4\|\mathbf{x}_1 - \mathbf{x}_*\|_2^2}{\eta\sqrt{T}} + \frac{2\eta(\sigma + M)^2}{\sqrt{T}}$$

by using the Euclidean distance $V(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ and their notation $\tau = 1$, and noting that $\eta$ is the inverse of their notation $c$. Then the second inequality is Lemma 4 can be proved similarly as for Lemma 2.

## Proof of Theorem 3

*Proof.* Recall $\mu = \rho/(\rho - G_1/\lambda)$ and $G = 3G_1 + 2\lambda G_2$. Let $V_k = \left(\mu^2 G^2\right) / \left(2^{k-2}\beta\right)$. By the values of $\eta_k$ and $T_k$ we have

$$T_k = 2^{k+3} = \frac{32\mu^2 G^2}{V_k \beta}, \eta_k = \frac{\mu}{2^{(k-1)/2}\beta} = \frac{V_k\sqrt{T_k}}{8\mu G^2}.$$

Define $\Delta_k = \hat{f}(\mathbf{x}_1^k) - \hat{f}(\mathbf{x}_*)$. We first prove the inequality

$$\mathrm{E}[\Delta_k] \leq V_k$$

by induction. It is true for $k = 1$ because of Lemma 5, $\mu > 1$ and $G^2 > G_1^2$. Now assume it is true for $k$ and we prove it for $k+1$. For a random variable $X$ measurable with respect to the randomness up to epoch $k + 1$. Let $\mathrm{E}_k[X]$ denote the expectation conditioned on all the randomness up to epoch $k$. Following Lemma 2, we have

$$\mathrm{E}_k[\Delta_{k+1}] \leq \mu\left[\frac{2\eta_k G^2}{\sqrt{T_k}} + \frac{\mathrm{E}[4\|\mathbf{x}_1^k - \mathbf{x}_*\|_2^2]}{\eta_k\sqrt{T_k}}\right] \tag{24}$$

Since $\Delta_k = f(\mathbf{x}_1^k) - f(\mathbf{x}_*) \geq \beta\|\mathbf{x}_1^k - \mathbf{x}_*\|_2^2/2$ by the strong convexity, we have

$$\begin{aligned}
\mathrm{E}[\Delta_{k+1}] \quad & \leq \quad \mu\left[\frac{2\eta_k G^2}{\sqrt{T_k}} + \frac{\mathrm{E}[8\Delta_k]}{\eta_k\sqrt{T_k}\beta}\right] \\
& = \quad \frac{2\eta_k \mu G^2}{\sqrt{T_k}} + \frac{V_k\mu}{\eta_k\sqrt{T_k}\beta} = \frac{V_k}{4} + \frac{V_k}{4} = \frac{V_k}{2}
\end{aligned}$$

where we use the fact $\eta_k/\sqrt{T_k} = V_k/(8\mu G^2)$ and $T_k = 32\mu^2 G^2/(V_k\beta)$. Thus, we get

$$\mathrm{E}[f(\mathbf{x}_1^{k^\dagger+1})] - f(\mathbf{x}_*) = \mathrm{E}[\Delta_{k^\dagger+1}] \leq V_{k^\dagger+1} = \frac{\mu^2 G^2}{2^{k^\dagger-1}\beta}$$

Note that the total number of epochs satisfies

$$\sum_{k=1}^{k^\dagger}(T_k + 1) = 16(2^{k^\dagger} - 1) + k^\dagger \leq T$$

By some reformulations, we complete the proof of this theorem. □

## Proof of Lemma 6

The proof of Lemma 6 is based on *the Bernstein Inequality for Martingales* [4]. We present its main result below for completeness.

**Theorem 4.** *[Bernstein Inequality for Martingales] Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence with respect to the filtration $\mathcal{F} = (\mathcal{F}_i)_{1 \leq i \leq n}$ and with $\|X_i\| \leq K$. Let*

$$S_i = \sum_{j=1}^{i} X_j$$

*be the associated martingale. Denote the sum of the conditional variances by*

$$\Sigma_n^2 = \sum_{t=1}^{n} \mathrm{E}\left[X_t^2 | \mathcal{F}_{t-1}\right],$$

*Then for all constants t, $\nu > 0$,*

$$\Pr\left[\max_{i=1,\ldots,n} S_i > t \text{ and } \Sigma_n^2 \leq \nu\right] \leq \exp\left(-\frac{t^2}{2(\nu + Kt/3)}\right),$$

*and therefore,*

$$\Pr\left[\max_{i=1,\ldots,n} S_i > \sqrt{2\nu t} + \frac{\sqrt{2}}{3} Kt \text{ and } \Sigma_n^2 \leq \nu\right] \leq e^{-t}.$$

*Proof of Lemma 6.* Define martingale difference $X_t = (\mathbf{x} - \mathbf{x}_t)^\top (\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t))$ and martingale $\Lambda_T = \sum_{t=1}^{T} X_t$. Define the conditional variance $\Sigma_T^2$ as

$$\Sigma_T^2 = \sum_{t=1}^{T} \mathrm{E}_{\xi_t}\left[X_t^2\right] \leq 4G_1^2 \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{x}\|_2^2 = 4G_1^2 D_T.$$

Define $K = 2G_1 D$. Thus, $\|X_t\|_2 \leq K$. We have

$$\begin{aligned}
&\Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \sqrt{2}K\tau/3\right) \\
&= \Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 D_T\right) \\
&= \Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 D_T,\right. \\
&\quad \left. D_T \leq \frac{D^2}{T}\right) + \sum_{i=1}^{m} \Pr\left(\Lambda_T \geq 2\sqrt{4G_1^2 D_T \tau} + \right. \\
&\quad \left. \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 D_T, \frac{D^2}{T} 2^{i-1} < D_T \leq \frac{D^2}{T} 2^i\right) \\
&\leq \Pr\left(D_T \leq \frac{D^2}{T}\right) + \sum_{i=1}^{m} \Pr\left(\Lambda_T \geq \sqrt{2 \times 4G_1^2 \frac{D^2}{T} 2^i \tau}\right. \\
&\quad \left. + \sqrt{2}K\tau/3, \Sigma_T^2 \leq 4G_1^2 \frac{D^2}{T} 2^i\right) \\
&\leq \Pr\left(D_T \leq \frac{D^2}{T}\right) + me^{-\tau},
\end{aligned}$$

where we use the fact $\|\mathbf{x}_t - \mathbf{x}\|_2^2 \leq D^2$ for all $t$ and $m = \lceil 2 \log_2 T \rceil$, and the last step follows the Bernstein inequality for martingales. We complete the proof by setting $\tau = \ln(m/\epsilon)$.