Appendix

Xi Chen
Machine Learning Department
Carnegie Mellon University
xichen@cs.cmu.edu

Qihang Lin **Javier Peña**
Tepper School of Business
Carnegie Mellon University
{qihangl,jfp}@andrew.cmu.edu

1 Proof of Convergence Rate of ORDA

Theorem 1. *For ORDA, if we require $c \geq 0$ and $c > 0$ when $\mu = 0$, then for any $t \geq 0$:*

$$\phi(x_{t+1}) - \phi(x^*) \leq \theta_t \nu_t \gamma_{t+1} V(x^*, x_0) + \frac{\theta_t \nu_t}{2} \sum_{i=0}^{t} \frac{(\|\Delta_i\|_* + M)^2}{\left(\frac{\mu}{\tau \theta_i} + \frac{\theta_i \gamma_i}{\tau} - \theta_i L\right) \nu_i} + \theta_t \nu_t \sum_{i=0}^{t} \frac{\langle x^* - \widehat{z}_i, \Delta_i \rangle}{\nu_i}, \quad (1)$$

where

$$\widehat{z}_t = \frac{\theta_t \mu}{\mu + \gamma_t \theta_t^2} y_t + \frac{(1 - \theta_t)\mu + \gamma_t \theta_t^2}{\mu + \gamma_t \theta_t^2} z_t, \quad (2)$$

is a convex combination of y_t and z_t and $\widehat{z}_t = z_t$ when $\mu = 0$. Taking the expectation on both sides of Eq.(1):

$$\mathbb{E}\phi(x_{t+1}) - \phi(x^*) \leq \theta_t \nu_t \gamma_{t+1} V(x^*, x_0) + (\sigma^2 + M^2)\theta_t \nu_t \sum_{i=0}^{t} \frac{1}{\left(\frac{\mu}{\tau \theta_i} + \frac{\theta_i \gamma_i}{\tau} - \theta_i L\right) \nu_i}. \quad (3)$$

We first state a basic property for Bregman distance functions in the following Proposition. This proposition generalizes Lemma 1 in [4] by extending one distance function to a sequence of functions.

Proposition 1. *Given any proper lsc convex function $\psi(x)$ and a sequence of $\{z_i\}_{i=0}^{t}$ with each $z_i \in \mathcal{X}$, if $z_+ = \arg\min_{x \in \mathcal{X}} \left\{ \psi(x) + \sum_{i=0}^{t} \eta_i V(x, z_i) \right\}$, where $\{\eta_i \geq 0\}_{i=0}^{t}$ is a sequence of parameters, then $\forall x \in \mathcal{X}$:*

$$\psi(x) + \sum_{i=0}^{t} \eta_i V(x, z_i) \geq \psi(z_+) + \sum_{i=0}^{t} \eta_i V(z_+, z_i) + \left(\sum_{i=0}^{t} \eta_i\right) V(x, z_+). \quad (4)$$

Proof of Proposition 1. For a Bregman distance function $V(x, y)$, let $\nabla_1 V(x, y)$ denote the gradient of $V(\cdot, y)$ at the point x. It is easy to show that:

$$V(x, y) \equiv V(z, y) + \langle \nabla_1 V(z, y), x - z \rangle + V(x, z), \qquad \forall \, x, y, z \in \mathcal{X},$$

which further implies that:

$$\sum_{i=0}^{t} \eta_i V(x, z_i) = \sum_{i=0}^{t} \eta_i V(z_+, z_i) + \sum_{i=0}^{t} \eta_i \langle \nabla_1 V(z_+, z_i), x - z_+ \rangle + \left(\sum_{i=0}^{t} \eta_i\right) V(x, z_+). \quad (5)$$

Since z_+ is the minimizer of the convex function $\psi(x) + \sum_{i=0}^{t} \eta_i V(x, z_i)$, it is known that there exists a subgradient g of ψ at z_+ ($g \in \partial\psi(z_+)$) such that:

$$\left\langle g + \sum_{i=0}^{t} \eta_i \nabla_1 V(z_+, z), x - z_+ \right\rangle \geq 0 \qquad \forall x \in \mathcal{X}. \tag{6}$$

Using the above two relations and the definition of subgradient ($\psi(x) \geq \psi(z_+) + \langle g, x - z_+ \rangle$ for all $x \in \mathcal{X}$), we conclude that:

$$\psi(x) + \sum_{i=0}^{t} \eta_i V(x, z_i)$$

$$\geq \psi(z_+) + \sum_{i=0}^{t} \eta_i V(z_+, z_i) + \left\langle g + \sum_{i=0}^{t} \eta_i \nabla_1 V(z_+, z_i), x - z_+ \right\rangle + \left(\sum_{i=0}^{t} \eta_i \right) V(x, z_+)$$

$$\geq \psi(z_+) + \sum_{i=0}^{t} \eta_i V(z_+, z_i) + \left(\sum_{i=0}^{t} \eta_i \right) V(x, z_+).$$

$\square$

To better present the proof of Theorem 1, we denote $G(y_t, \xi_t)$ by $G(y_t)$ and define:

$$\Delta_t := G(y_t) - f'(y_t) = G(y_t, \xi_t) - f'(y_t) \tag{7}$$

We first show some basic properties Δ_t. Let $\xi_{[t]}$ denote the collection of *i.i.d.* random vectors $\{\xi_i\}_{i=0}^{t}$. Since both random vectors y_t and z_t are functions of $\xi_{[t-1]}$ and are independent of $\{\xi_i\}_{i=t}^{N}$, we have that for any $t \geq 1$ and any α, β

$$\mathbb{E}\Delta_t = \mathbb{E}_{\xi_{[t-1]}}[\mathbb{E}_{\xi_t}(\Delta_t|\xi_{[t-1]})] = \mathbb{E}_{\xi_{[t-1]}} 0 = 0; \tag{8}$$

$$\mathbb{E}\|\Delta_t\|_*^2 = \mathbb{E}_{\xi_{[t-1]}}[\mathbb{E}_{\xi_t}(\|\Delta_t\|_*^2|\xi_{[t-1]})] \leq \mathbb{E}_{\xi_{[t-1]}} \sigma^2 = \sigma^2; \tag{9}$$

$$\mathbb{E}\langle \alpha y_t + \beta z_t, \Delta_t \rangle = \mathbb{E}_{\xi_{[t-1]}}[\langle \alpha y_t + \beta z_t, \mathbb{E}_{\xi_t}\Delta_t\rangle|\xi_{[t-1]}] = \mathbb{E}_{\xi_{[t-1]}}[\langle \alpha y_t + \beta z_t, 0\rangle|\xi_{[t-1]}] = 0, \tag{10}$$

Proof of Theorem 1. With our choice of $\theta_t, \nu_t, \gamma_t$, it is easy to show (see [5]) that:

$$\sum_{i=0}^{t} \frac{1}{\nu_i} = \frac{1}{\theta_t \nu_t}, \qquad \frac{1 - \theta_t}{\theta_t \nu_t} = \frac{1}{\theta_{t-1}\nu_{t-1}}, \qquad \theta_t \leq \nu_t. \tag{11}$$

We further define $\frac{1}{\theta_{-1}\nu_{-1}} = 0$. We first bound the objective value $\phi(x_{t+1})$ by:

$$
\begin{aligned}
\phi(x_{t+1}) &= f(x_{t+1}) + h(x_{t+1}) \leq f(y_t) + \langle x_{t+1} - y_t, f'(y_t) \rangle + \frac{L}{2}\|x_{t+1} - y_t\|^2 + M\|x_{t+1} - y_t\| + h(x_{t+1}) \\
&\leq \underbrace{f(y_t) + \langle x_{t+1} - y_t, G(y_t) \rangle + \left(\frac{\mu}{\tau\theta_t^2} + \frac{\gamma_t}{\tau} \right) V(x_{t+1}, y_t) + h(x_{t+1})}_{C_1} \\
&\quad \underbrace{- \frac{1}{2}\left(\frac{\mu}{\tau\theta_t^2} + \frac{\gamma_t}{\tau} - L \right)\|x_{t+1} - y_t\|^2 - \langle x_{t+1} - y_t, \Delta_t \rangle + M\|x_{t+1} - y_t\|}_{C_2}
\end{aligned}
\tag{12}
$$

We bound the terms C_1 and C_2 respectively. Let $\widehat{x}_{t+1}$ be the convex combination of x_t and z_{t+1}:

$$\widehat{x}_{t+1} = (1 - \theta_t)x_t + \theta_t z_{t+1}.$$

Then we have $\widehat{x}_{t+1} - y_t = \theta_t(z_{t+1} - \widehat{z}_t)$, where

$$\widehat{z}_t = \frac{\theta_t \mu}{\mu + \gamma_t \theta_t^2} y_t + \frac{(1 - \theta_t)\mu + \gamma_t \theta_t^2}{\mu + \gamma_t \theta_t^2} z_t,$$

which is a convex combination of y_t and z_t. By the fact that x_{t+1} is the minimizer of C_1 and utilizing the relationship $V(x_{t+1}, y_t) \leq \frac{\tau \|x_{t+1} - y_t\|^2}{2}$ and $\widehat{x}_{t+1} - y_t = \theta_t(z_{t+1} - \widehat{z}_t)$:

$$C_1 \leq f(y_t) + \langle \widehat{x}_{t+1} - y_t, f'(y_t) \rangle + \theta_t \langle z_{t+1} - \widehat{z}_t, \Delta_t \rangle + \left(\frac{\mu + \gamma_t \theta_t^2}{2} \right) \|z_{t+1} - \widehat{z}_t\|^2 + h(\widehat{x}_{t+1}). \quad (13)$$

By the convexity of $\| \cdot \|^2$ and the fact that $\frac{1}{2} \|x - y\|^2 \leq V(x, y)$ for any $x, y \in \mathcal{X}$:

$$\left(\frac{\mu + \gamma_t \theta_t^2}{2} \right) \|z_{t+1} - \widehat{z}_t\|^2 \leq \theta_t \mu V(z_{t+1}, y_t) + \left((1 - \theta_t)\mu + \theta_t^2 \gamma_t \right) V(z_{t+1}, z_t). \quad (14)$$

We plug Eq.(14) back into RHS of Eq.(13) and substitute $\widehat{x}_{t+1}$ with $(1 - \theta_t)x_t + \theta_t z_{t+1}$. By the convexity of $h(\cdot)$:

$$
\begin{aligned}
C_1 \quad \leq \quad & (1 - \theta_t) \left(f(y_t) + \langle x_t - y_t, f'(y_t) \rangle + h(x_t) \right) \\
& + \theta_t \underbrace{\left(f(y_t) + \langle z_{t+1} - y_t, G(y_t) \rangle + h(z_{t+1}) + \mu V(z_{t+1}, y_t) + \left(\frac{(1 - \theta_t)\mu}{\theta_t} + \gamma_t \theta_t \right) V(z_{t+1}, z_t) \right)}_{C_3} \\
& + \theta_t \langle z_{t+1} - \widehat{z}_t, \Delta_t \rangle + \theta_t \langle y_t - z_{t+1}, \Delta_t \rangle \\
\leq \quad & (1 - \theta_t)\phi(x_t) + C_3 + \theta_t \langle y_t - \widehat{z}_t, \Delta_t \rangle. \quad (15)
\end{aligned}
$$

Now we bound C_3 using Proposition 1. Utilizing the first equality in Eq. (11), we can re-write z_t as

$$z_t = \arg\min_{x \in \mathcal{X}} \left\{ \widetilde{\psi}_t(x) + \sum_{i=0}^{t-1} \frac{\mu}{\nu_i} V(x, y_i) + \gamma_t V(x, x_0) \right\},$$

where

$$\widetilde{\psi}_t(x) := \sum_{i=0}^{t-1} \frac{f(y_i) + \langle x - y_i, G(y_i) \rangle + h(x)}{\nu_i}.$$

Furthermore, we define $\psi_t(x) := \sum_{i=0}^{t-1} \frac{f(y_i) + \langle x - y_i, G(y_i) \rangle + h(x) + \mu V(x, y_i)}{\nu_i}$ and apply Proposition 1 with $x = z_{t+1}$:

$$
\begin{aligned}
\left(\sum_{i=0}^{t-1} \frac{\mu}{\nu_i} + \gamma_t \right) V(z_{t+1}, z_t) \quad \leq \quad & \left(\widetilde{\psi}_t(z_{t+1}) + \sum_{i=0}^{t-1} \frac{\mu}{\nu_i} V(z_{t+1}, y_i) + \gamma_t V(z_{t+1}, x_0) \right) \quad (16) \\
& - \left(\widetilde{\psi}_t(z_t) + \sum_{i=0}^{t-1} \frac{\mu}{\nu_i} V(z_t, y_i) + \gamma_t V(z_t, x_0) \right) \\
= \quad & \psi_t(z_{t+1}) + \gamma_t V(z_{t+1}, x_0) - \psi_t(z_t) - \gamma_t V(z_t, x_0) \quad (17)
\end{aligned}
$$

We can bound the last term in C_3 by Eq.(17). In particular, according to Eq.(11):

$$
\begin{aligned}
\left(\frac{(1 - \theta_t)\mu}{\theta_t} + \gamma_t \theta_t \right) V(z_{t+1}, z_t) \quad \leq \quad & \nu_t \left(\sum_{i=0}^{t-1} \frac{\mu}{\nu_i} + \gamma_t \right) V(z_{t+1}, z_t) \\
\leq \quad & \nu_t \left(\psi_t(z_{t+1}) + \gamma_t V(z_{t+1}, x_0) - \psi_t(z_t) - \gamma_t V(z_t, x_0) \right).
\end{aligned}
$$

With the above inequality, we immediately obtain an upper bound for C_3. Therefore, by the definition of $\psi_t(\cdot)$, we bound the term C_1 by:

$$C_1 \leq (1 - \theta_t)\phi(x_t) + \theta_t \nu_t \left(\psi_{t+1}(z_{t+1}) - \psi_t(z_t) + \gamma_t V(z_{t+1}, x_0) - \gamma_t V(z_t, x_0) \right) + \theta_t \langle y_t - \widehat{z}_t, \Delta_t \rangle. \quad (18)$$

To bound C_2, since the parameter $c > 0$ whenever $\mu = 0$, we always have $\frac{\mu}{\tau \theta_t^2} + \frac{\gamma_t}{\tau} - L > 0$. Using a simple inequality: $-\frac{\alpha}{2}\kappa^2 + \beta\kappa \leq \frac{\beta^2}{2\alpha}$ ($\alpha > 0$), with $\alpha = \frac{\mu}{\tau \theta_t^2} + \frac{\gamma_t}{\tau} - L$, $\beta = \|\Delta_t\|_* + M$ and $\kappa = \|x_{t+1} - y_t\|$, we have:

$$C_2 \leq -\frac{1}{2} \left(\frac{\mu}{\tau \theta_t^2} + \frac{\gamma_t}{\tau} - L \right) \|x_{t+1} - y_t\|^2 + \|x_{t+1} - y_t\| (\|\Delta_t\|_* + M) \leq \frac{(\|\Delta_t\|_* + M)^2}{2 \left(\frac{\mu}{\tau \theta_t^2} + \frac{\gamma_t}{\tau} - L \right)}. \quad (19)$$

3

By summing up the upper bound for C_1 in Eq.(18) and the bound for C_2 in Eq.(19), we obtain an upper bound for $\phi(x_{t+1})$ according to Eq.(12). Utilizing the second relation in Eq. (11), we build up the following recursive inequality:

$$
\begin{aligned}
\frac{\phi(x_{t+1})}{\theta_t \nu_t} &\leq \frac{\phi(x_t)}{\theta_{t-1}\nu_{t-1}} + \left(\psi_{t+1}(z_{t+1}) - \psi_t(z_t) + \gamma_t V(z_{t+1}, x_0) - \gamma_t V(z_t, x_)\right) \\
&\quad + \frac{(\|\Delta_i\|_* + M)^2}{2\left(\frac{\mu}{\tau\theta_t} + \frac{\theta_t\gamma_t}{\tau} - \theta_t L\right)\nu_t} + \frac{\langle y_t - \widehat{z}_t, \Delta_t\rangle}{\nu_t} \leq \cdots \\
&\leq \frac{\phi(x_0)}{\theta_{-1}\nu_{-1}} + \psi_{t+1}(z_{t+1}) - \psi_0(z_0) + \gamma_t V(z_{t+1}, x_0) - \gamma_t V(z_t, x_) \\
&\quad + \sum_{i=0}^{t} \frac{(\|\Delta_i\|_* + M)^2}{2\left(\frac{\mu}{\tau\theta_i} + \frac{\theta_i\gamma_i}{\tau} - \theta_i L\right)\nu_i} + \sum_{i=0}^{t} \frac{\langle y_i - \widehat{z}_i, \Delta_i\rangle}{\nu_i} \\
&= \psi_{t+1}(z_{t+1}) + \gamma_t V(z_{t+1}, x_0) + \sum_{i=0}^{t} \frac{(\|\Delta_i\|_* + M)^2}{2\left(\frac{\mu}{\tau\theta_i} + \frac{\theta_i\gamma_i}{\tau} - \theta_i L\right)\nu_i} + \sum_{i=0}^{t} \frac{\langle y_i - \widehat{z}_i, \Delta_i\rangle}{\nu_i}, \quad (20)
\end{aligned}
$$

where the last inequality is obtained by the fact that $\frac{1}{\theta_{-1}\nu_{-1}} = 0$, $V(z_0, x_0) = 0$, $\psi_0(z_0) = 0$. Using the fact that $z_{t+1} = \arg\min_{x \in \mathcal{X}}\{\psi_{t+1}(x) + \gamma_{t+1}V(x, x_0)\}$ and $\gamma_t \leq \gamma_{t+1}$, Eq.(20) further implies that:

$$
\begin{aligned}
\frac{\phi(x_{t+1})}{\theta_t \nu_t} &\leq \psi_{t+1}(x^*) + \gamma_{t+1}V(x^*, x_0) + \sum_{i=0}^{t} \frac{(\|\Delta_i\|_* + M)^2}{2\left(\frac{\mu}{\tau\theta_i} + \frac{\theta_i\gamma_i}{\tau} - \theta_i L\right)\nu_i} + \sum_{i=0}^{t} \frac{\langle y_i - \widehat{z}_i, \Delta_i\rangle}{\nu_i} \\
&= \sum_{i=0}^{t} \frac{f(y_i) + \langle x^* - y_i, f'(y_i)\rangle + h(x^*) + \mu V(x^*, y_i)}{\nu_i} + \sum_{i=0}^{t} \frac{\langle x^* - y_i, \Delta_i\rangle}{\nu_i} \\
&\quad + \gamma_{t+1}V(x^*, x_0) + \sum_{i=0}^{t} \frac{(\|\Delta_i\|_* + M)^2}{2\left(\frac{\mu}{\tau\theta_i} + \frac{\theta_i\gamma_i}{\tau} - \theta_i L\right)\nu_i} + \sum_{i=0}^{t} \frac{\langle y_i - \widehat{z}_i, \Delta_i\rangle}{\nu_i} \\
&\leq \sum_{i=0}^{t} \frac{\phi(x^*)}{\nu_i} + \gamma_{t+1}V(x^*, x_0) + \sum_{i=0}^{t} \frac{(\|\Delta_i\|_* + M)^2}{2\left(\frac{\mu}{\tau\theta_i} + \frac{\theta_i\gamma_i}{\tau} - \theta_i L\right)\nu_i} + \sum_{i=0}^{t} \frac{\langle x^* - \widehat{z}_i, \Delta_i\rangle}{\nu_i}. \quad (21)
\end{aligned}
$$

Multiplying by $\theta_t\nu_t$ on both sides of Eq.(21), we obtain the result in Eq.(1). From the properties of Δ_i in Eq.(8)–(10), we conclude that for all i, $\mathbb{E}\langle x^* - \widehat{z}_i, \Delta_i\rangle = 0$ and $\mathbb{E}(\|\Delta_i\|_* + M)^2 \leq 2\sigma^2 + 2M^2$. By taking the expectation on both sides of Eq.(1) and using the aforementioned properties for Δ_i, we obtain the result in Eq.(3). □

Corollary 1. *For convex $f(x)$ with $\widetilde{\mu} = 0$ (or equivalently $\mu = 0$), by setting $c = \frac{\sqrt{\tau}(\sigma + M)}{2\sqrt{V(x^*, x_0)}}$ and $\Gamma = L$, we obtain:*

$$
\mathbb{E}\phi(x_{N+1}) - \phi(x^*) \leq \frac{4\tau L V(x^*, x_0)}{N^2} + \frac{8(\sigma + M)\sqrt{\tau V(x^*, x_0)}}{\sqrt{N}}. \quad (22)
$$

Proof. When $\mu = 0$, the expected gap in the objective function in Eq.(3) for the last iterate becomes:

$$
\mathbb{E}\phi(x_{N+1}) - \phi(x^*) \leq \theta_N\nu_N\gamma_{N+1}V(x^*, x_0) + (\sigma^2 + M^2)\theta_N\nu_N\sum_{t=0}^{N} \frac{1}{\left(\frac{\gamma_t}{\tau} - L\right)\theta_t\nu_t} \quad (23)
$$

With choice of $\theta_N = \frac{2}{N+2}$, $\nu_N = \frac{2}{N+1}$ and $\gamma_{N+1} = c(N+2)^{3/2} + \tau L$, the first term in Eq.(23) is bounded by:

$$
\theta_N\nu_N\gamma_{N+1}V(x^*, x_0) \leq \frac{4\tau L V(x^*, x_0)}{N^2} + \frac{8c\, V(x^*, x_0)}{\sqrt{N}} \quad (24)
$$

Similarly, the second term in Eq.(23) can be bounded by:

$$
(\sigma^2 + M^2)\theta_N\nu_N\sum_{t=0}^{N} \frac{1}{\left(\frac{\gamma_t}{\tau} - L\right)\theta_t\nu_t} \leq \frac{2\tau(\sigma + M)^2}{c\sqrt{N}} \quad (25)
$$

4

By summing the above two inequalities, we obtain that:

$$\mathbb{E}\phi(x_{N+1}) - \phi(x^*) \leq \frac{4\tau L V(x^*, x_0)}{N^2} + \frac{8c\, V(x^*, x_0)}{\sqrt{N}} + \frac{2\tau(\sigma + M)^2}{c\sqrt{N}} \tag{26}$$

We minimize the RHS of Eq.(26) with respect to c and obtain the convergence rate result in Corollary 1 and the corresponding optimal $c = \frac{\sqrt{\tau}(\sigma+M)}{2\sqrt{V(x^*, x_0)}}$. $\qquad\square$

Corollary 2. *For strongly convex $f(x)$ with $\widetilde{\mu} > 0$, we set $c = 0$ and $\Gamma = L$ and obtain that:*

$$\mathbb{E}\phi(x_{N+1}) - \phi(x^*) \leq \frac{4\tau L V(x^*, x_0)}{N^2} + \frac{4\tau(\sigma^2 + M^2)}{\mu N}. \tag{27}$$

Proof. When $\mu > 0$, we set $c = 0$ and $\gamma_t \equiv \tau L$ and then Eq.(3) becomes:

$$\mathbb{E}\phi(x_{N+1}) - \phi(x^*) \leq \theta_N \nu_N \tau L V(x^*, x_0) + \frac{\tau(\sigma^2 + M^2)}{\mu} \theta_N \nu_N \sum_{t=0}^{N} \frac{\theta_t}{\nu_t} \leq \frac{4\tau L V(x^*, x_0)}{N^2} + \frac{4\tau(\sigma^2 + M^2)}{\mu N}. \tag{28}$$

This gives the result in Eq.(27) in Corollary 1. $\qquad\square$

2 High Probability Bounds for ORDA

Theorem 2. *We assume that (1) $\mathbb{E}\left(\exp\left\{\|G(x, \xi) - f'(x)\|_*^2/\sigma^2\right\}\right) \leq \exp\{1\}$, $\forall x \in \mathcal{X}$ (i.e., "light-tail" assumption) and (2) there exists a constant D such that $\|x^* - \widehat{z}_t\| \leq D$ for all t. By setting $\Gamma = L$ in ORDA, for any iteration t and $\delta \in (0, 1)$, we have, with probability at least $1 - \delta$:*

$$\phi(x_{t+1}) - \phi(x^*)) \leq \epsilon(t, \delta) \tag{29}$$

with

$$\begin{aligned}
\epsilon(t, \delta) &= \theta_t \nu_t \gamma_{t+1} V(x^*, x_0) + \theta_t \nu_t \sum_{i=0}^{t} \frac{M^2}{\eta_i \nu_i} + \theta_t \left[\sum_{i=0}^{t} \frac{\sigma^2}{\eta_i} + \frac{8\sigma^2 \ln(2/\delta)}{(\frac{\mu+\gamma_0}{\tau} - L)} + 16\sigma^2 \sqrt{\sum_{i=0}^{t} \frac{\ln(2/\delta)}{\eta_i^2}} \right] \\
&\quad + \sqrt{3 \ln \frac{2}{\delta}} \theta_t \nu_t D \sigma \left(\sum_{i=0}^{t} \frac{1}{\nu_i^2} \right)^{1/2},
\end{aligned} \tag{30}$$

where $\eta_i = \left(\frac{\mu}{\tau\theta_i} + \frac{\theta_i \gamma_i}{\tau} - \theta_i L \right)$.

For convex $f(x)$ with $\widetilde{\mu} = 0$, by setting $c = \frac{\sqrt{\tau}(\sigma+M)}{2\sqrt{V(x^*, x_0)}}$ and $\Gamma = L$, we have

$$\begin{aligned}
\epsilon(N, \delta) &= \frac{4\tau L V(x^*, x_0)}{N^2} + \frac{24\sqrt{\tau V(x^*, x_0)}(\sigma + M)}{\sqrt{N}} + \frac{16\ln(2/\delta)\sqrt{\tau V(x^*, x_0)}\sigma}{N} \\
&\quad + \frac{16\sigma\sqrt{\ln(2/\delta)\ln(N+3)V(x^*, x_0)}}{N} + \frac{2\sqrt{\ln(2/\delta)}D\sigma}{\sqrt{N}}.
\end{aligned} \tag{31}$$

For convex $f(x)$ with $\widetilde{\mu} > 0$ (or equivalently $\mu > 0$), by setting $c = 0$ and $\Gamma = L$, , we have

$$\epsilon(N, \delta) = \frac{4\tau L V(x^*, x_0)}{N^2} + \frac{16\tau(\sigma^2 + M^2)\ln(N+2)}{\mu N} + \frac{48\sigma^2 \ln(2/\delta)}{\mu N} + \frac{2\sqrt{\ln(2/\delta)}D\sigma}{\sqrt{N}}. \tag{32}$$

We prove Theorem 2 using the following two lemmas.

Lemma 1 (Lemma 6 in [2]). *Let $\xi_0, \xi_1, \ldots$ be a sequence of i.i.d. random variables and $\varphi_i = \varphi_i(\xi_{[i]})$ be deterministic Borel functions of $\xi_{[i]}$ such that:*

1. $\mathbb{E}(\varphi_i | \xi_{[i-1]}) = 0$;

2. There exists a positive deterministic sequence $\{\sigma_i\}$: $\mathbb{E}\left(\exp\{\varphi_i^2/\sigma_i^2\}|\xi_{[i-1]}\right) \leq \exp\{1\}$.

Then for any $\delta \in (0,1)$, $Prob\left(\sum_{i=0}^{t}\varphi_i \geq \sqrt{3\ln(1/\delta)}(\sum_{i=0}^{t}\sigma_i^2)^{1/2}\right) \leq \delta$.

Lemma 2 (Lemma 5 in [1]). *Under the assumptions in Theorem 2, for any positive and nondecreasing sequence η_i, we have*

$$\sum_{i=0}^{t}\frac{\|\Delta_i\|_*^2}{\eta_i} \geq \sum_{i=0}^{t}\frac{\mathbb{E}\|\Delta_i\|_*^2}{\eta_i} + \max\left\{\frac{8\sigma^2\ln(1/\delta)}{\eta_0}, 16\sigma^2\sqrt{\sum_{i=0}^{t}\frac{\ln(1/\delta)}{\eta_i^2}}\right\}$$

holds with probability at most $\delta \in (0,1)$.

We note that although Lemma 5 in [1] assumes that $\eta_i = \eta\sqrt{i+1}$, its proof and conclusion remain valid for any positive nondecreasing sequence $\{\eta_i\}$.

Proof of Theorem 2. To simply notations, let $\eta_i = \left(\frac{\mu}{\tau\theta_i} + \frac{\theta_i\gamma_i}{\tau} - \theta_i L\right)$. For both convex and strongly convex $f(x)$, according to our setting of parameters, it is easy to verifty that $\{\eta_i\}$ is a positive monotonically increasing sequence. According to Theorem 1:

$$\phi(x_{t+1}) - \phi(x^*) \leq \underbrace{\theta_t\nu_t\gamma_{t+1}V(x^*,x_0)}_{C_1} + \underbrace{\theta_t\nu_t\sum_{i=0}^{t}\frac{M^2}{\eta_i\nu_i}}_{C_1} + \underbrace{\theta_t\nu_t\sum_{i=0}^{t}\frac{\|\Delta_i\|_*^2}{\eta_i\nu_i}}_{C_2} + \underbrace{\theta_t\nu_t\sum_{i=0}^{t}\frac{\langle x^*-\widehat{z}_i,\Delta_i\rangle}{\nu_i}}_{C_3},$$

Firstly, we analyze the last term C_3 using Lemma 1. Let $\varphi_i(\xi_{[i]}) := \frac{\langle x^*-\widehat{z}_i,\Delta_i\rangle}{\nu_i}$ and hence $C_3 = \theta_t\nu_t\sum_{i=0}^{t}\varphi_i$. It is easy to verify that $\mathbb{E}(\varphi_i|\xi_{[i-1]}) = 0$ and there exists a sequence $\sigma_i = \frac{D\sigma}{\nu_i}$ such that:

$$\mathbb{E}(\exp\{\varphi_i^2/\sigma_i^2\}|\xi_{[i-1]}) \equiv \mathbb{E}\left(\exp\left\{\left(\frac{\langle x^*-\widehat{z}_i,\Delta_i\rangle}{\nu_i}\right)^2/\frac{D^2\sigma^2}{\nu_i^2}\right\}\right) \leq \mathbb{E}\left(\exp\left\{\frac{\|x^*-\widehat{z}_i\|^2\|\Delta_i\|_*^2}{D^2\sigma^2}\right\}\right) \leq \exp\{1\},$$

where the last inequality holds because $\|x^*-\widehat{z}_t\| \leq D$ and our "light-tail" assumption. By Lemma 1, we conclude that for any $\delta \in (0,1)$,

$$\Pr\left(C_3 \geq \underbrace{\sqrt{3\ln\frac{2}{\delta}}\theta_t\nu_tD\sigma(\sum_{i=0}^{t}\frac{1}{\nu_i^2})^{1/2}}_{D_3}\right) \leq \frac{\delta}{2}. \tag{33}$$

Secondly, we bound the term C_2 using Lemma 2. Since ν_i is decreasing in i, we have

$$C_2 = \theta_t\nu_t\sum_{i=0}^{t}\frac{\|\Delta_i\|_*^2}{\eta_i\nu_i} \leq \theta_t\nu_t\sum_{i=0}^{t}\frac{\|\Delta_i\|_*^2}{\eta_i\nu_t} = \theta_t\sum_{i=0}^{t}\frac{\|\Delta_i\|_*^2}{\eta_i}. \tag{34}$$

Since η_i is increasing in i when $\Gamma = L$, we can directly apply Lemma 2 as follows:

$$\Pr\left(C_2 \geq \theta_t\underbrace{\left[\sum_{i=0}^{t}\frac{\sigma^2}{\eta_i} + \frac{8\sigma^2\ln(2/\delta)}{(\frac{\mu+\gamma_0}{\tau}-L)} + 16\sigma^2\sqrt{\sum_{i=0}^{t}\frac{\ln(2/\delta)}{\eta_i^2}}\right]}_{D_2}\right)$$

$$\leq \Pr\left(\theta_t\sum_{i=0}^{t}\frac{\|\Delta_i\|_*^2}{\eta_i} \geq \theta_t\left[\sum_{i=0}^{t}\frac{\mathbb{E}\|\Delta_i\|_*^2}{\eta_i} + \max\left\{\frac{8\sigma^2\ln(2/\delta)}{(\frac{\mu+\gamma_0}{\tau}-L)}, 16\sigma^2\sqrt{\sum_{i=0}^{t}\frac{\ln(2/\delta)}{\eta_i^2}}\right\}\right]\right)$$

$$\leq \frac{\delta}{2}$$

where the first inequality is from Eq. (34), $a + b \geq \max\{a,b\}$ and the fact $\mathbb{E}\|\Delta_i\|_*^2 \leq \sigma^2\ln\left(\mathbb{E}\exp\left(\frac{\|\Delta_i\|_*^2}{\sigma^2}\right)\right) \leq \sigma^2\ln(e) = \sigma^2$ and the second inequality is due to Lemma 2.

Combining Eq.(35) and Eq. (33), by the union bound:

$$\Pr\left(\phi(x_{t+1}) - \phi(x^*) \geq C_1 + D_2 + D_3\right) \leq \Pr\left(C_1 + C_2 + C_3 \geq C_1 + D_2 + D_3\right)$$

$$\leq \quad \Pr\left(C_2 \geq D_2\right) + \Pr\left(C_3 \geq D_3\right) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta, \tag{35}$$

we immediately obtain Eq.(30). The bounds in Eq. (31) and Eq. (32) can be derived by plugging all the parameters into Eq. (30).

$\square$

3   Proof of Convergence Rate for Multi-stage ORDA

Theorem 3. *If we run multi-stage ORDA for K stages with $K = \log \frac{\mathcal{V}_0}{\epsilon}$ for any given ϵ, we have $\mathbb{E}(\phi(\widetilde{x}_K)) - \phi(x^*) \leq \epsilon$ and the total number of iterations is upper bounded by:*

$$N = \sum_{k=1}^{K} N_k \leq 4\sqrt{\frac{L}{\mu}} \log \frac{\mathcal{V}_0}{\epsilon} + \frac{1024(\sigma^2 + M^2)}{\mu\epsilon}. \tag{36}$$

To prove theorem 3, we first state a corollary of Theorem 1.

Corollary 3. *For strongly convex $f(x)$, by setting $c = 0$ and $\Gamma = \Lambda + L$ in ORDA, we obtain that:*

$$\mathbb{E}\phi(x_{N+1}) - \phi(x^*) \leq \frac{4\tau(\Lambda + L)V(x^*, x_0)}{N^2} + \frac{(N+3)(\sigma^2 + M^2)}{\Lambda}. \tag{37}$$

The proof technique follows the proof in [3]. The main idea is to show that $\mathbb{E}(\phi(\widetilde{x}_k)) - \phi(x^*) \leq \mathcal{V}_0 2^{-k}$, where $\widetilde{x}_k$ is the solution from the k-th stage.

Proof. We show by induction that

$$\mathbb{E}(\phi(\widetilde{x}_k)) - \phi(x^*) \leq \mathcal{V}_0 2^{-k}. \tag{38}$$

By the definition of $\mathcal{V}_0$ ($\mathcal{V}_0 > \phi(\widetilde{x}_0) - \phi(x^*)$), this inequality holds for $k = 0$.

Assuming Eq.(38) holds for the $(k-1)$-th stage, by the strong convexity of $f(x)$, we have

$$\mathbb{E}[V(x^*, \widetilde{x}_{k-1})] \leq \mathbb{E}\left[\frac{\tau}{2}\|\widetilde{x}_{k-1} - x^*\|^2\right] \leq \mathbb{E}\left[\frac{\tau}{\mu}(\phi(\widetilde{x}_{k-1}) - \phi(x^*))\right] \leq \frac{\mathcal{V}_0 2^{-(k-1)}}{\mu}$$

According to Corollary 3 and the setting of N_k and Γ_k, we have

$$
\begin{aligned}
\mathbb{E}[\phi(\widetilde{x}_k) - \phi(x^*)] \quad &\leq \quad \frac{4\tau(\Lambda_k + L)\mathbb{E}V(x^*, \widetilde{x}_{k-1})}{N_k^2} + \frac{(N_k + 3)(\sigma^2 + M^2)}{\Lambda_k} \\
&\leq \quad \frac{4\tau L \mathcal{V}_0 2^{-(k-1)}}{\mu N_k^2} + \frac{4\tau \Lambda_k \mathcal{V}_0 2^{-(k-1)}}{\mu N_k^2} + \frac{4N_k(\sigma^2 + M^2)}{\Lambda_k} \\
&\leq \quad \frac{4\tau L \mathcal{V}_0 2^{-(k-1)}}{\mu N_k^2} + \frac{8\sqrt{(\sigma^2 + M^2)\tau \mathcal{V}_0 2^{-(k-1)}}}{\sqrt{\mu}N_k} \\
&\leq \quad \frac{\mathcal{V}_0 2^{-k}}{2} + \frac{\mathcal{V}_0 2^{-k}}{2} = \mathcal{V}_0 2^{-k}.
\end{aligned}
$$

Therefore, we prove that $\mathbb{E}[\phi(\widetilde{x}_k) - \phi(x^*)] \leq \mathcal{V}_0 2^{-k}$ for $k \geq 1$.

After running K stages of multi-stage ORDA with $K = \log_2\left(\frac{\mathcal{V}_0}{\epsilon}\right)$, we have $\mathbb{E}[\phi(\widetilde{x}_k) - \phi(x^*)] \leq \mathcal{V}_0 2^{-K} = \epsilon$. The total number of iterations from these K stages is upper bounded by:

$$
\begin{aligned}
\sum_{k=1}^{K} N_k \quad &\leq \quad \sum_{k=1}^{K} \max\left\{ 4\sqrt{\frac{\tau L}{\mu}}, \frac{2^{k+9}\tau(\sigma^2 + M^2)}{\mu \mathcal{V}_0} \right\} \\
&\leq \quad \sum_{k=1}^{K} \left[4\sqrt{\frac{\tau L}{\mu}} + \frac{2^{k+9}\tau(\sigma^2 + M^2)}{\mu \mathcal{V}_0} \right] \\
&= \quad 4\sqrt{\frac{\tau L}{\mu}} K + \frac{1024\tau(\sigma^2 + M^2)(2^K - 1)}{\mu \mathcal{V}_0} \\
&\leq \quad 4\sqrt{\frac{\tau L}{\mu}} \log_2\left(\frac{\mathcal{V}_0}{\epsilon}\right) + \frac{1024\tau(\sigma^2 + M^2)}{\mu \epsilon}
\end{aligned}
$$

$\square$

References

[1] J. Duchi, P. L. Bartlett, and M. Wainwright. Randomized smoothing for stochastic optimization. arXiv:1103.4296v1, 2011.

[2] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2010.

[3] G. Lan and S. Ghadimi. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, part ii: shrinking procedures and optimal algorithms. Technical report, University of Florida, 2010.

[4] G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $o(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 2009.

[5] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization (Submitted)*, 2008.