Supplementary Material of Fast Rate Analysis of Some Stochastic Optimization Algorithms

The following two lemmas are Lemma 3 and Lemma 4 in [1], we present here for completeness.

Lemma A. 1. *Suppose $X_1, ..., X_T$ is a martingale difference sequence with $|X_t| \leq b$. Let*

$$Var_t X_t = Var(X_t | X_1, ..., X_{t-1}).$$

Let $V = \sum_{t=1}^{T} Var_t X_t$ be the sum of conditional variance of $X_t's$. Further, let $\sigma = \sqrt{V}$. Then we have for any $\delta < 1/e$ and $T \geq 3$,

$$Prob(\sum_{t=1}^{T} X_t > max\{2\sigma, 3b\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)}) \leq 4\ln(T)\delta$$

Lemma A. 2. *Suppose $s, r, d, b, \Delta \geq 0$ and we have*

$$s - r \leq \max\{4\sqrt{ds}, 6b\Delta\}\Delta.$$

Then, it follows that

$$s \leq r + 4\sqrt{dr}\Delta + \max\{16d, 6b\}\Delta^2.$$

Regularized Dual Averaging Method

Now we begin the proof of convergence rate of RDA. We define the following conjugate type function used in the proof.

$$V_t(s) = \max_w [\langle s, w - w_0 \rangle - tr(w) - \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2 - \beta_t h(w)]$$

Lemma A. 3. *The function $V_t(\cdot)$ is convex and differentiable. $\nabla V(s_t) = w_{t+1} - w_0$, where $s_t = \sum_{\tau=1}^{t} f'(w_\tau, z_\tau)$. The gradient is Lipschitz continuous with constant $\frac{1}{2\gamma t + \beta_t}$, which is*

$$\|\nabla V_t(s_1) - \nabla V_t(s_2)\|_2 \leq \frac{1}{2\gamma t + \beta_t} \|s_1 - s_2\|_2,$$

Proof. Because $\gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2 + \beta_t h(w)$ is strongly convex with convexity parameter $2\gamma t + \beta_t$. It is a direct result from theorem 1 in [2]. $\square$

A property of $V_t(\cdot)$ with Lipschitz continuous gradient is

$$V_t(s+\delta) \le V_t(s) + \langle \delta, \nabla V_t(s)\rangle + \frac{1}{2(2\gamma t+\beta_t)}\|\delta\|_2^2$$

We refer to [2] for more details.

Proof of Lemma 3.

$$\begin{aligned}
&Reg_t - \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2 \\
&\le \sum_{\tau=1}^{t} \langle f'(w_\tau, z_\tau), w_\tau - w\rangle + \sum_{\tau=1}^{t} r(w_\tau) - tr(w) - \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2 \\
&= \sum_{\tau=1}^{t} \langle f'(w_\tau, z_\tau), w_\tau - w_0\rangle + \sum_{\tau=1}^{t} r(w_\tau) - tr(w) - \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2 \\
&+ \sum_{\tau=1}^{t} \langle f'(w_\tau, z_\tau), w_0 - w\rangle
\end{aligned} \tag{1}$$

where the first inequity holds from the convexity of $f(\cdot, z)$.
Before we bound above terms, we relate $V_{t-1}(-s_t)$ and $V_t(-s_t)$ in the following way

$$\begin{aligned}
&V_{t-1}(-s_t) = \max_w [\langle -s_t, w - w_0\rangle - (t-1)r(w) - \gamma \sum_{\tau=1}^{t-1} \|w_\tau - w\|_2^2 - \beta_{t-1} h(w)] \\
&\ge \langle -s_t, w_{t+1} - w_0\rangle - (t-1)r(w_{t+1}) - \gamma \sum_{\tau=1}^{t-1} \|w_\tau - w_{t+1}\|_2^2 - \beta_{t-1} h(w_{t+1}) \\
&= \langle -s_t, w_{t+1} - w_0\rangle - tr(w_{t+1}) - \gamma \sum_{\tau=1}^{t} \|w_\tau - w_{t+1}\|_2^2 - \beta_t h(w_{t+1}) + r(w_{t+1}) \\
&+ \gamma \|w_t - w_{t+1}\|_2^2 - (\beta_{t-1} - \beta_t) h(w_{t+1})
\end{aligned} \tag{2}$$

Notice the summation of first four terms is $V_t(-s_t)$.
When $t > 1$, since β_t is an increasing sequence , we have

$$V_t(-s_t) + r(w_{t+1}) + \gamma \|w_t - w_{t+1}\|_2^2 \le V_{t-1}(-s_t)$$

We then upper bound $V_{t-1}(-s_t)$

$$\begin{aligned}
V_{t-1}(-s_t) &= V_{t-1}(-s_{t-1} - f'(w_t, z_t)) \\
&\le V_{t-1}(-s_{t-1}) - \langle \nabla V_{t-1}(-s_{t-1}), f'(w_t, z_t)\rangle + \frac{1}{2(2\gamma(t-1)+\beta_{t-1})}\|f'(w_t, z_t)\|_2^2,
\end{aligned} \tag{3}$$

where the inequality holds from the property of Lipschitz continuous of $\nabla V_t(\cdot)$.

Now we have

$$\begin{aligned} V_t(-s_t) - V_{t-1}(-s_{t-1}) \leq & -\langle \nabla V_{t-1}(s_{t-1}), f'(w_t, z_t)\rangle + \frac{1}{2(2\gamma(t-1)+\beta_{t-1})}\|f'(w_t,z_t)\|_2^2 \\ & - r(w_{t+1}) - \gamma\|w_t - w_{t+1}\|_2^2 \\ \leq & -\langle w_t - w_0, f'(w_t, z_t)\rangle + \frac{1}{2(2\gamma(t-1)+\beta_{t-1})}\|f'(w_t,z_t)\|_2^2 \\ & - r(w_{t+1}), \end{aligned} \tag{4}$$

where the second inequality uses the fact $\nabla V(s_t) = w_{t+1} - w_0$ from Lemma A.3
.
When $t = 1$, we have

$$V_1(-s_1) - 0 \leq -\langle w_1 - w_0, f'(w_1, z_1)\rangle + \frac{\|f'(w_1, z_1)\|_2^2}{2\beta_0} - r(w_2) + (\beta_0 - \beta_1)h(w_2). \tag{5}$$

Sum both sides of $V_\tau(s_\tau)$ from $\tau = 1$ to t, we have
$V_t(-s_t) \leq -\sum_{\tau=1}^t \langle w_\tau - w_0, f'(w_\tau, z_\tau)\rangle + \sum_{\tau=1}^t \frac{\|f'(w_\tau,z_\tau)\|_2^2}{2(2\gamma(\tau-1)+\beta_{\tau-1})} - \sum_{\tau=2}^{t+1} r(w_\tau) +$ $(\beta_0 - \beta_1)h(w_2)$.
We then bound $Reg_t - \gamma\sum_{\tau=1}^t \|w_\tau - w\|_2^2$ for all $w \in \mathcal{F}_D$ using above result,

$$\begin{aligned} Reg_t - \gamma\sum_{\tau=1}^t \|w_\tau - w\|_2^2 \leq & \sum_{\tau=1}^t r(w_\tau) + \sum_{\tau=1}^t \langle f'(w_\tau, z_\tau), w_\tau - w_0\rangle + \max_{w\in\mathcal{F}_D} [\langle -s_t, w - w_0\rangle - tr(w) \\ & - \gamma\sum_{\tau=1}^t \|w_\tau - w\|_2^2] \\ \leq & \sum_{\tau=1}^t r(w_\tau) + \sum_{\tau=1}^t \langle f'(w_\tau, z_\tau), w_\tau - w_0\rangle + V_t(-s_t) + \beta_t D^2 \\ \leq & \, r(w_1) - r(w_{t+1}) + (\beta_0 - \beta_1)h(w_2) + \sum_{\tau=1}^t \frac{\|f'(w_\tau, z_\tau)\|_2^2}{2(2\gamma(\tau-1)+\beta_{\tau-1})} \\ & + \beta_t D^2, \end{aligned} \tag{6}$$

where the second inequality holds from the fact that

$$\max_{w\in\mathcal{F}_D} [\langle s_t, w - w_0\rangle - tr(w) - \gamma\sum_{\tau=1}^t \|w_\tau - w\|_2^2] \leq V_t(-s_t) + \beta_t D^2.$$

Since $\arg\min_w h(w) = \arg\min_w r(w)$ and $w_1 = \arg\min h(w)$, $r(w_1) - r(w_{t+1}) \leq 0$. We set $\beta_0 = \beta_1 = \gamma$ and $\beta_t = \gamma(1 + \ln t)$, then we have

$$Reg_T - \gamma\sum_{t=1}^T \|w_t - w\|_2^2 \leq \gamma D^2(1+\ln(T)) + \sum_{t=1}^T \frac{L^2}{2\gamma(2t-1+\ln t)} \leq (C_1\gamma D^2 + \frac{C_2 L^2}{\gamma})(1+\ln T).$$

□

Proof of Theorem 2. Since we have already known $Reg_T - \gamma\sum_{t=1}^T \|w_t - w\|_2^2 \le C_1 \ln T + C_2$, using similar steps in the proof of Theorem 1, we have,

$$\frac{1}{2}Diff_T \le \sum_{t=1}^T \xi_t + C_1 \ln T + C_2.$$

Then we apply Lemma 2, Lemma A.1 and Lemma A.2 to get the result.

□

OPG-ADMM

The following Lemma is extracted from Theorem 4 in the appendix of [3].

Lemma A. 4. *Let $\{x_t\}_{t=1}^T$, $\{y_t\}_{t=1}^T$ and $\{\lambda_t\}_{t=1}^T$ be the sequence generated by the algorithm. For all $\hat{x} \in \mathcal{X}$, $\hat{y} \in \mathcal{Y}$ and $\hat{\lambda} \in R^l$ and f is weakly convex, we have*

$$\sum_{t=1}^T (f(x_t, z_t) + \psi(y_t)) - \sum_{t=1}^T (f(\hat{x}, z_t) + \psi(\hat{y})) + \sum_{t=1}^T \begin{pmatrix} -A^T\tilde{\lambda}_t \\ -B^T\tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^T \begin{pmatrix} x_t - \hat{x} \\ y_t - \hat{y} \\ \tilde{\lambda}_t - \hat{\lambda} \end{pmatrix}$$
$$+ \sum_{t=1}^T \frac{\|\lambda_t - \lambda_{t+1}\|_2^2}{2\rho} + \frac{\|\lambda_{T+1} - \hat{\lambda}\|_2^2}{2\rho}$$
$$\le \frac{\|\hat{x}\|_{G_1}^2}{2\eta_1} + \sum_{t=2}^T (\frac{\gamma}{2\eta_t} - \frac{\gamma}{2\eta_{t-1}})\|x_t - \hat{x}\|_2^2 + \sum_{t=1}^T \frac{\eta_t}{2}\|g_t\|_{G_t^{-1}}^2 + \frac{\rho}{2}\|b - B\hat{y}\|_2^2$$
$$+ \frac{\|\hat{\lambda}\|_2^2}{2\rho} + \langle Ax_{T+1}, \hat{\lambda}\rangle + \langle B(\hat{y} - y_{T+1}), \lambda_{T+1} - \hat{\lambda}\rangle - \langle B\hat{y} - b, \hat{\lambda}\rangle, \tag{7}$$

where g_t denotes $f'(x_t, z_t)$ for short.

Proof of Lemma 4. We subtract $\sum_{t=1}^T \frac{\beta}{4}\|x_t - \hat{x}\|_2^2$ at both side of Lemma A.4. Notice $\langle \hat{y} - y_{T+1}, B^T(\lambda_{T+1} - \hat{\lambda})\rangle \le \langle \hat{y} - y_{T+1}, \nabla\psi(y_{T+1}) - B^T\hat{\lambda}\rangle$ using the optimality of y_{t+1} in the algorithm, i.e., $\langle \nabla\psi(y_t) - B^T\lambda_t, y - y_t\rangle \ge 0$. So this term can also be bounded if $\hat{\lambda}$ is bounded, in particular we choose $\hat{\lambda} = 0$.
Notice $G_t \succeq I$ in the algorithm by choosing γ, ρ, η_t. Similar to the proof of Lemma 1, the term $\sum_{t=2}^T(\frac{\gamma}{2\eta_t} - \frac{\gamma}{2\eta_{t-1}})\|x_t - \hat{x}\|_2^2 + \sum_{t=1}^T \frac{\eta_t}{2}\|g_t\|_{G_t^{-1}}^2 - \sum_{t=1}^T \frac{\beta}{4}\|x_t - \hat{x}\|_2^2$ is bounded by $C_1 \ln T + C_2$, if we choose $\eta_t = \frac{2\gamma}{\beta t}$.
We choose $\hat{\lambda} = 0$ to simplify the left hand side. For all $\hat{x}, \hat{y}$ such that $A\hat{x} + B\hat{y} = b$, we have

$$
\begin{aligned}
\sum_{t=1}^{T}\begin{pmatrix}-A^T\tilde{\lambda}_t\\ -B^T\tilde{\lambda}_t\\ Ax_t+By_t-b\end{pmatrix}^T\begin{pmatrix}x_t-\hat{x}\\ y_t-\hat{y}\\ \tilde{\lambda}_t-\hat{\lambda}\end{pmatrix} &= \sum_{t=1}^{T}\begin{pmatrix}A^T\hat{\lambda}\\ B^T\hat{\lambda}\\ A\hat{x}+B\hat{y}-b\end{pmatrix}^T\begin{pmatrix}\hat{x}-x^t\\ \hat{y}-y_t\\ \tilde{\lambda}_t-\hat{\lambda}\end{pmatrix}\\
&= \sum_{t=1}^{T}\langle\hat{\lambda}, A(\hat{x}-x_t)+B(\hat{y}-y_t)\rangle\\
&= \sum_{t=1}^{T}\langle\hat{\lambda}, b-Ax_t-By_t\rangle\\
&= \sum_{t=1}^{T}\langle\hat{\lambda}, (\lambda_t-\lambda_{t-1})\rangle\\
&= \langle\frac{1}{\rho}\hat{\lambda}, \lambda_T-\lambda_1\rangle,
\end{aligned}
\tag{8}
$$

where last two equality hold from the fact that $b - Ax_t - By_t = \frac{\lambda_t-\lambda_{t-1}}{\rho}$ and $Ax_1 + By_1 - b = 0$.
We set $\hat{\lambda} = 0$, so the third term on the left side in Lemma 4 is 0.

Also notice $\frac{\|\hat{x}\|_{G_1}^2}{2\eta_1}$ and $\frac{\rho}{2}\|b - B\hat{y}\|_2^2$ are bounded under our assumption. Thus the RHS of the Lemma 4 is bounded by $C_1 \ln T + C_2$ when $\hat{\lambda} = 0$. □

Similar to the previous proof, we define

$$
Diff = \sum_{t=1}^{T}(F(x_t) + \psi(y_t)) - \sum_{t=1}^{T}(F(x^*) + \psi(y^*))
$$

and

$$
Reg = \sum_{t=1}^{T}(f(x_t, z_t) + \psi(y_t)) - \sum_{t=1}^{T}(f(x^*, z_t) + \psi(y^*)),
$$

where $F(x) = Ef(x, z), G(x, y) = F(x) + \psi(y)$. Remind that $Ax_t + By_t - b \neq 0$ in general, thus we use $y'_t = B^{-1}(b - Ax_t)$ as an estimator of y at the t-th step.

Proof of Theorem 3. Similar to the previous proof in OPG, we define

$$
\begin{aligned}
\xi_t &= F(x_t) + \psi(y_t) - F(x^*) - r(y^*) - (f(x_t, z_t) + r(y_t) - f(x^*, z_t) - \psi(y^*))\\
&= F(x_t) - F(x^*) - (f(x_t, z_t) - f(x^*, z_t)).
\end{aligned}
\tag{9}
$$

ξ_t is a martingale difference, since x_t just depends on the data from time step $1, ..., t-1$, $E_{t-1}f(x^*, z_t) = F(x^*)$, $E_{t-1}f(x_t, z_t) = F(x_t)$. Using Lemma 2, $Var_{t-1}\xi_t = E_{t-1}\xi_t^2 \leq L^2\|x_t - x^*\|_2^2$.
Next we relate $Diff$ to $\sum_{t=1}^{T} Var_{t-1}\xi_t$.

$$
\begin{aligned}
Diff &\geq \sum_{t=1}^{T} \langle \nabla F(x^*), x_t - x^* \rangle + \frac{\beta}{2} \|x_t - x^*\|_2^2 + \langle \nabla \psi(y^*), y_t - y^* \rangle \\
&= \sum_{t=1}^{T} [\langle \nabla F(x^*), x_t - x^* \rangle + \langle \nabla \psi(y^*), y'_t - y^* \rangle + \langle \nabla \psi(y^*), y_t - y'_t \rangle + \frac{\beta}{2} \|x_t - x^*\|_2^2],
\end{aligned} \tag{10}
$$

where the first inequality holds from the convexity of F and ψ.
Recall that $y'_t = B^{-1}(b - Ax_t)$ and $Ax^* + By^* - b = 0$, so $\langle \nabla F(x^*), x_t - x^* \rangle + \langle \nabla \psi(y^*), y'_t - y^* \rangle \geq 0$ using the optimality of (x^*, y^*). Thus we have the following relation.

$$
Diff + \frac{1}{\rho} \langle B^{-T} \nabla \psi(y^*), \lambda_T - \lambda_1 \rangle = Diff + \sum_{t=1}^{T} \langle \nabla \psi(y^*), y'_t - y_t \rangle \geq \sum_{t=1}^{T} \frac{\beta}{2} \|x_t - x^*\|_2^2, \tag{11}
$$

where the first equality holds from the fact that $B(y'_t - y_t) = b - Ax_t - By_t = \frac{\lambda_t - \lambda_{t-1}}{\rho}$ and $Ax_1 + By_1 - b = 0$.
We denote $\frac{1}{\rho} \langle B^{-T} \nabla \psi(y^*), \lambda_T - \lambda_1 \rangle$ as N_T, and discuss two conditions.
When $N_T \leq 0$, we have $Diff \geq \sum_{t=1}^{T} \frac{\beta}{2} \|x_t - x^*\|_2^2$.

When $N_T \geq 0$, we need a upper bound of N_T.

$$
\begin{aligned}
N_T &= \frac{1}{\rho} \langle B^{-T} \nabla \psi(y^*), \lambda_T - \lambda_1 \rangle \\
&\leq \frac{3}{2\rho} \|B^{-T} \nabla \psi(y^*)\|_2^2 + \frac{1}{6\rho} \|\lambda_T - \lambda_1\|_2^2 \\
&= \frac{3}{2\rho} \|B^{-T} \nabla \psi(y^*)\|_2^2 + \frac{1}{6\rho} \|\lambda_T - \lambda_{T+1} + \lambda_{T+1} - \lambda_1\|_2^2 \\
&\leq \frac{3}{2\rho} \|B^{-T} \nabla \psi(y^*)\|_2^2 + \frac{1}{2\rho} (\|\lambda_T - \lambda_{T+1}\|_2^2 + \|\lambda_{T+1}\|_2^2 + \|\lambda_1\|_2^2),
\end{aligned} \tag{12}
$$

where the first and second inequalities holds from the Cauchy-Schwarz inequality. Notice $\frac{3}{2\rho} \|B^{-T} \nabla \psi(y^*)\|_2^2$ can be bounded by our assumption.
Remind that instead of evaluating $F(\bar{x}_T) + \psi(\bar{y}_T) - F(x^*) - \psi(y^*)$, our aim is to bound $F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*)$.

$$
\begin{aligned}
T(F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*)) &\leq T(F(\bar{x}_T) + \psi(\bar{y}_T) - F(x^*) - \psi(y^*)) \\
&\quad + T \langle \nabla \psi(\bar{y}'_T), \bar{y}'_T - \bar{y}_T \rangle \\
&\leq Diff + T \langle B^{-T} \nabla \psi(\bar{y}'_T), B(\bar{y}'_T - \bar{y}_T) \rangle \\
&= Diff + \frac{1}{\rho} \langle B^{-T} \nabla \psi(\bar{y}'_T), \lambda_T - \lambda_1 \rangle,
\end{aligned} \tag{13}
$$

where the first inequality holds from the convexity of ψ, the second inequality uses the convexity of F and ψ, and the last equality holds from the fact $B(y'_t - y_t) = b - Ax_t - By_t = \frac{\lambda_t - \lambda_{t-1}}{\rho}$ and $Ax_1 + By_1 - b = 0$.
We also need to consider two cases, i.e., $\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle$ is negative or not.
If it is negative, $T(F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*)) \leq Diff$.
If it is not negative, we need to bound it

$$\begin{aligned}
&Diff + \frac{1}{\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle \\
&\leq Diff + \frac{1}{2\rho}(6\|B^{-T}\nabla\psi(\bar{y}'_T)\|_2^2 + \frac{1}{6}\|\lambda_T - \lambda_1\|_2^2) \\
&\leq Diff + \frac{3}{\rho}\|B^{-T}\nabla\psi(\bar{y}'_T)\|_2^2 + \frac{1}{4\rho}(\|\lambda_T - \lambda_{T+1}\|_2^2 + \|\lambda_{T+1}\|_2^2 + \|\lambda_1\|_2^2).
\end{aligned} \tag{14}$$

Notice $\frac{3}{\rho}\|B^{-T}\nabla\psi(\bar{y}'_T)\|_2^2$ can be bounded by our assumption.
Totally, we need to consider four cases.

Case 1 $N_T \leq 0, \langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle \leq 0$.
In this case, $Diff \geq \sum_{t=1}^T \frac{\beta}{2}\|x_t - x^*\|_2^2$.
Using the similar technique in the proof of Theorem 1, we have following condition with probability at least $1 - 4\delta\ln T$.
$\frac{1}{2}Diff - (Reg - \frac{\beta}{4}\sum_{t=1}^T \|x_t - x^*\|_2^2) \leq \xi_t \leq \max\{2\sqrt{\frac{2L^2}{\beta}(Diff)}, 6B\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)}$, which implies

$$\frac{1}{2}Diff - (Reg - \frac{\beta}{4}\sum_{t=1}^T \|x_t - x^*\|_2^2) \leq \max\{2\sqrt{\frac{2L^2}{\beta}(Diff)}, 6B\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)}. \tag{15}$$

Notice $Reg - \frac{\beta}{4}\sum_{t=1}^T \|x_t - x^*\|_2^2$ is bounded by $C_1 \ln T + C_2$ in Lemma 4 with $\hat{\lambda} = 0$ ($\sum_{t=1}^T \frac{\|\lambda_t - \lambda_{t+1}\|_2^2}{2\rho} + \frac{\|\lambda_{T+1}\|_2^2}{2\rho}$ is a positive term). Following similar steps in the proof of Theorem 1, we solve this inequality using Lemma A.2. Then we get $Diff \leq C_3 \ln T + C_4$ with high probability. In this case $T(F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*)) \leq Diff$, thus $F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*) \leq O(\frac{\ln T}{T})$ with high probability.
Case 2 $N_T \geq 0, \langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle \leq 0$.
We have following relation by (11) with probability at least $1 - 4\delta\ln T$.

$$\begin{aligned}
&\frac{1}{2}(Diff + N_T) - (Reg - \frac{\beta}{4}\sum_{t=1}^T \|x_t - x^*\|_2^2 + \frac{N_T}{2}) \\
&\leq \max\{2\sqrt{\frac{2L^2}{\beta}(Diff + N_T)}, 6B\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)}.
\end{aligned} \tag{16}$$

Notice $Reg - \frac{\beta}{4}\sum_{t=1}^T \|x_t - x^*\|_2^2 + \frac{N_T}{2}$ is bounded by $C_1 \ln T + C_2$, using the Lemma 4 and (12) with $\hat{\lambda} = 0$. Solve above inequality using Lemma A.2, we have

$Diff + N_T \leq C_3 \ln T + C_4$ with high probability which implies $Diff \leq O(\ln T)$. Since $T(F(\bar{x_T}) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*)) \leq Diff$, we have $F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*) \leq O(\frac{\ln T}{T})$ with high probability.
Case 3 $N_T \leq 0, \langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle \geq 0$.
We have following relation with probability at least $1 - 4\delta \ln T$.

$$\begin{aligned}
&\frac{1}{2}(Diff + \frac{1}{\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle) - (Reg - \frac{\beta}{4}\sum_{t=1}^{T}\|x_t - x^*\|_2^2 + \frac{1}{2\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle) \\
&\leq \max\{2\sqrt{\frac{2L^2}{\beta}(Diff)}, 6B\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)} \\
&\leq \max\{2\sqrt{\frac{2L^2}{\beta}(Diff + \frac{1}{\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle)}, 6B\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)}.
\end{aligned} \tag{17}$$

$Reg - \frac{\beta}{4}\sum_{t=1}^{T}\|x_t - x^*\|_2^2 + \frac{1}{2\rho}\langle \nabla B^{-T}\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle$ is bounded by $C_1 \ln T + C_2$ by Lemma 4 and (14) with $\hat{\lambda} = 0$. We get $Diff + \frac{1}{\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle \leq C_3 \ln T + C_4$ with high probability. Thus we have $T(F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*)) \leq C_3 \ln T + C_4$ by (13).
Case 4 $N_T \geq 0, \langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle \geq 0$
We have following relation with probability at least $1 - 4\delta \ln T$.

$$\begin{aligned}
&\frac{1}{2}(Diff + N_T + \frac{1}{\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle) - (Reg - \frac{\beta}{4}\sum_{t=1}^{T}\|x_t - x^*\|_2^2 + \frac{N_T}{2} \\
&+ \frac{1}{2\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle) \\
&\leq \max\{2\sqrt{\frac{2L^2}{\beta}(Diff + N_T)}, 6B\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)} \\
&\leq \max\{2\sqrt{\frac{2L^2}{\beta}(Diff + N_T + \frac{1}{\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle)}, 6B\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)}.
\end{aligned} \tag{18}$$

Notice $Reg - \frac{\beta}{4}\sum_{t=1}^{T}\|x_t - x^*\|_2^2 + \frac{N_T}{2} + \frac{1}{2\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle$ is bounded by $C_1 \ln T + C_2$, using Lemma 4, (12) and (14) with $\hat{\lambda} = 0$. Solve the inequality, we get $Diff + N_T + \frac{1}{\rho}\langle B^{-T}\nabla\psi(\bar{y}'_T), \lambda_T - \lambda_1\rangle \leq C_3 \ln T + C_4$ with high probability. Thus $T(F(\bar{x}_T) + \psi(\bar{y}'_T) - F(x^*) - \psi(y^*)) \leq C_3 \ln T + C_4$ with high probability by (13) and the fact that $N_T \geq 0$.
In all cases, we have $G(\bar{x}_T, \bar{y}'_T) - G(x^*, y^*) \leq O(\frac{\ln T}{T})$ with high probability, thus we finish our proof. □

References

[1] Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

[2] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[3] Taiji Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning*, pages 392–400, 2013.