Supplementary Material: Dual Averaging and Proximal Gradient Descent for Online Alternating Direction Multiplier Method

Taiji Suzuki S-TAIJI@STAT.T.U-TOKYO.AC.JP

Department of Mathematical Informatics, The University of Tokyo, Tokyo 113-8656, Japan

In this supplementary material, we give the proofs for the theorems in the main text. We consider more general optimization problem:

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}} \ \frac{1}{T}\sum_{t=1}^{T} f(x,w_t)+\psi(y), \tag{S-1a}$$

$$\text{s.t.} \ \ Ax+By-b=0, \tag{S-1b}$$

where $A \in \mathbb{R}^{l\times m}$, $B \in \mathbb{R}^{l\times d}$ and $b \in \mathbb{R}^{l}$. We solve this problem in an online manner using OPG-ADMM and RDA-ADMM techniques. The algorithm in the main text corresponds to the situation where $B=-I$ and $b=\mathbf{0}$.

Theorems 1, 2, 3 and 4 in the main text correspond to Theorems 7 (and Eq. (S-31)), 8, 9 and 5 in this supplementary material respectively.

A. Convergence rate of OPG-ADMM

Corresponding to the optimization problem (S-1), we consider the following generalized version of OPG-ADMM: Let $x_1=\mathbf{0}$, $\lambda_1=\mathbf{0}$, and $By_1=b$ (we assume there exists y_1 that satisfies this equality for simplicity), and the update rule of the t-th step is given by

$$x_{t+1}=\operatorname*{argmin}_{x\in\mathcal{X}}\left\{g_t^\top x-\lambda_t^\top(Ax+By_t-b)+\frac{\rho}{2}\|Ax+By_t-b\|^2+\frac{1}{2\eta_t}\|x-x_t\|_{G_t}^2\right\}, \tag{S-2a}$$

$$y_{t+1}=\operatorname*{argmin}_{y\in\mathcal{Y}}\left\{\psi(y)-\lambda_t^\top(Ax_{t+1}+By-b)+\frac{\rho}{2}\|Ax_{t+1}+By-b\|^2\right\}, \tag{S-2b}$$

$$\lambda_{t+1}=\lambda_t-\rho(Ax_{t+1}+By_{t+1}-b), \tag{S-2c}$$

where $G_t=\gamma I-\eta_t\rho A^\top A$. Moreover we define

$$\tilde{\lambda}_t=\lambda_t-\rho(Ax_{t+1}+By_t-b). \tag{S-3}$$

This method satisfies the following regret property.

Theorem 4. *For all $x^*\in\mathcal{X}$, $y^*\in\mathcal{Y}$ and $\lambda^*\in\mathbb{R}^l$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}(f(x_t,w_t)+\psi(y_t))-\frac{1}{T}\sum_{t=1}^{T}(f(x^*,w_t)+\psi(y^*))$$
$$+\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}-A^\top\tilde{\lambda}_t\\-B^\top\tilde{\lambda}_t\\Ax_t+By_t-b\end{pmatrix}^\top\begin{pmatrix}x_t-x^*\\y_t-y^*\\\tilde{\lambda}_t-\lambda^*\end{pmatrix}$$

$$
\begin{aligned}
&+ \frac{1}{T}\sum_{t=1}^{T}\frac{\|\lambda_t-\lambda_{t+1}\|^2}{2\rho} + \frac{\|\lambda_{T+1}-\lambda^*\|^2}{2\rho T} \\
\leq &\frac{\|x^*\|_{G_1}^2}{2\eta_1 T} + \frac{1}{T}\sum_{t=2}^{T}\left(\frac{\gamma}{2\eta_t} - \frac{\gamma}{2\eta_{t-1}} - \frac{\sigma}{2}\right)\|x_t - x^*\|^2 + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}\|g_t\|_{G_t^{-1}}^2 \\
&+ \frac{\rho}{2T}\|b - By^*\|^2 + \frac{\|\lambda^*\|^2}{2\rho T} \\
&+ \frac{1}{T}\left(\langle Ax_{T+1}, \lambda^*\rangle + \langle B(y^* - y_{T+1}), \lambda_{T+1} - \lambda^*\rangle\right) - \langle By^* - b, \lambda^*\rangle\right).
\end{aligned}
$$

Proof. Let $f_t(x) := f(x, w_t)$. By the optimality of x_{t+1} and y_t, we have that, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$
\langle g_t - A^\top \tilde{\lambda}_t + G_t(x_{t+1} - x_t)/\eta_t, x - x_{t+1}\rangle \geq 0 \tag{S-4}
$$

and

$$
\langle \nabla\psi(y_t) - B^\top \lambda_t, y - y_t\rangle \geq 0, \tag{S-5}
$$

where we used the relation $\nabla\psi(y_t) - B^\top\lambda_{t-1} + \rho B^\top(Ax_t + By_t - b) = \nabla\psi(y_t) - B^\top\lambda_t$. Using these inequality, for given $x^* \in \mathcal{X}, y^* \in \mathcal{Y}$ and λ^*, we have that

$$
\begin{aligned}
&f_t(x_t) + \psi(y_t) - f_t(x^*) - \psi(y^*) \\
\leq &\langle g_t, x_t - x^*\rangle - \frac{\sigma}{2}\|x_t - x^*\|^2 + \langle \nabla\psi(y_t), y_t - y^*\rangle \\
= &\langle g_t, x_{t+1} - x^*\rangle + \langle \nabla\psi(y_t), y_t - y^*\rangle + \langle g_t, x_t - x_{t+1}\rangle - \frac{\sigma}{2}\|x_t - x^*\|^2 \\
\leq &\langle A^\top\tilde{\lambda}_t - G_t(x_{t+1} - x_t)/\eta_t, x_{t+1} - x^*\rangle + \langle B^\top\lambda_t, y_t - y^*\rangle + \langle g_t, x_t - x_{t+1}\rangle - \frac{\sigma}{2}\|x_t - x^*\|^2 \\
= &\begin{pmatrix} -A^\top\tilde{\lambda}_t \\ -B^\top\tilde{\lambda}_t \end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \end{pmatrix} + \begin{pmatrix} x^* - x_{t+1} \\ y^* - y_t \end{pmatrix}^\top \begin{pmatrix} G_t(x_{t+1} - x_t)/\eta_t \\ -B^\top(\lambda_t - \tilde{\lambda}_t) \end{pmatrix} + \langle A^\top\tilde{\lambda}_t, x_{t+1} - x_t\rangle + \langle g_t, x_t - x_{t+1}\rangle \\
&- \frac{\sigma}{2}\|x_t - x^*\|^2 \\
= &\begin{pmatrix} -A^\top\tilde{\lambda}_t \\ -B^\top\tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix} + \begin{pmatrix} x^* - x_{t+1} \\ y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix}^\top \begin{pmatrix} G_t(x_{t+1} - x_t)/\eta_t \\ -B^\top(\lambda_t - \tilde{\lambda}_t) \\ \frac{\tilde{\lambda}_t - \lambda_t}{\rho} - A(x_t - x_{t+1}), \end{pmatrix} \\
&+ \langle A^\top\tilde{\lambda}_t, x_{t+1} - x_t\rangle + \langle g_t, x_t - x_{t+1}\rangle - \frac{\sigma}{2}\|x_t - x^*\|^2,
\end{aligned}
$$

where, in the last line, we used

$$
Ax_t + By_t - b = -\frac{\tilde{\lambda}_t - \lambda_t}{\rho} + A(x_t - x_{t+1})
$$

by the definition of $\tilde{\lambda}_t$ and λ_t.

Now we bound $\langle x^* - x_{t+1}, G_t(x_{t+1} - x_t)/\eta_t\rangle$. By Lemma 10, this can be bounded as

$$
\langle x^* - x_{t+1}, G_t(x_{t+1} - x_t)/\eta_t\rangle = -\frac{\|x_{t+1} - x^*\|_{G_t}^2}{2\eta_t} + \frac{\|x_t - x^*\|_{G_t}^2}{2\eta_t} - \frac{\|x_{t+1} - x_t\|_{G_t}^2}{2\eta_t}.
$$

On the other hand,

$$
\langle g_t, x_t - x_{t+1}\rangle \leq \|g_t\|_{G_t^{-1}}\|x_t - x_{t+1}\|_{G_t} \leq \frac{\eta_t}{2}\|g_t\|_{G_t^{-1}}^2 + \frac{1}{2\eta_t}\|x_t - x_{t+1}\|_{G_t}^2.
$$

Combining these two inequalities, we have

$$
\langle x^* - x_{t+1}, G_t(x_{t+1} - x_t)/\eta_t\rangle + \langle g_t, x_t - x_{t+1}\rangle \leq -\frac{\|x_{t+1} - x^*\|_{G_t}^2}{2\eta_t} + \frac{\|x_t - x^*\|_{G_t}^2}{2\eta_t} + \frac{\eta_t}{2}\|g_t\|_{G_t^{-1}}^2.
$$

This and Lemma 11 gives

$$
\begin{aligned}
&\begin{pmatrix} x^* - x_{t+1} \\ y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix}^\top \begin{pmatrix} G_t(x_{t+1} - x_t)/\eta_t \\ -B^\top(\lambda_t - \tilde{\lambda}_t) \\ \frac{\tilde{\lambda}_t - \lambda_t}{\rho} - A(x_t - x_{t+1}), \end{pmatrix} + \langle A^\top \tilde{\lambda}_t, x_{t+1} - x_t \rangle + \langle g_t, x_t - x_{t+1} \rangle, \\
=& \langle x^* - x_{t+1}, G_t(x_{t+1} - x_t)/\eta_t \rangle + \langle g_t, x_t - x_{t+1} \rangle + \langle \lambda^*, A(x_{t+1} - x_t) \rangle \\
&+ \begin{pmatrix} y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix}^\top \begin{pmatrix} -B^\top(\lambda_t - \tilde{\lambda}_t) \\ \frac{\tilde{\lambda}_t - \lambda_t}{\rho}, \end{pmatrix} \\
\le& \frac{\|x_t - x^*\|_{G_t}^2}{2\eta_t} - \frac{\|x_{t+1} - x^*\|_{G_t}^2}{2\eta_t} + \langle \lambda^*, A(x_{t+1} - x_t) \rangle + \frac{\eta_t}{2} \|g_t\|_{G_t^{-1}}^2 \\
&+ \frac{\rho}{2} \|y_t - y^*\|_{B^\top B}^2 - \frac{\rho}{2} \|y_{t+1} - y^*\|_{B^\top B}^2 + \frac{\|\lambda_t - \lambda^*\|^2}{2\rho} - \frac{\|\lambda_{t+1} - \lambda^*\|^2}{2\rho} \\
&+ \langle B(y^* - y_{t+1}), \lambda_{t+1} - \lambda^* \rangle - \langle B(y^* - y_t), \lambda_t - \lambda^* \rangle \qquad \text{(S-6)} \\
&- \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}.
\end{aligned}
$$

Now summing up this bound for $t = 1, \ldots, T$, then we have

$$
\begin{aligned}
&\sum_{t=1}^T (f_t(x_t) + \psi(y_t)) - \sum_{t=1}^T (f_t(x^*) - \psi(y^*)) \\
\le& \sum_{t=1}^T \begin{pmatrix} -A^\top \tilde{\lambda}_t \\ -B^\top \tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix} \\
&+ \frac{\|x_1 - x^*\|_{G_1}^2}{2\eta_1} + \sum_{t=2}^T \left(\frac{\|x_t - x^*\|_{G_t}^2}{2\eta_t} - \frac{\|x_t - x^*\|_{G_{t-1}}^2}{2\eta_{t-1}} \right) + \langle \lambda^*, A(x_{T+1} - x_1) \rangle + \sum_{t=1}^T \frac{\eta_t}{2} \|g_t\|_{G_t^{-1}}^2 \\
&+ \frac{\rho}{2} \|y_1 - y^*\|_{B^\top B}^2 + \frac{\|\lambda_1 - \lambda^*\|^2}{2\rho} - \frac{\|\lambda_{T+1} - \lambda^*\|^2}{2\rho} \\
&+ \langle B(y^* - y_{T+1}), \lambda_{T+1} - \lambda^* \rangle - \langle B(y^* - y_1), \lambda_1 - \lambda^* \rangle \\
&- \sum_{t=1}^T \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho} - \sum_{t=1}^T \frac{\sigma}{2} \|x_t - x^*\|^2.
\end{aligned}
$$

Now since $G_t = \gamma I - \eta_t \rho A^\top A$, we have that

$$
\frac{\|x_t - x^*\|_{G_t}^2}{2\eta_t} - \frac{\|x_t - x^*\|_{G_{t-1}}^2}{2\eta_{t-1}} = \left(\frac{\gamma}{2\eta_t} - \frac{\gamma}{2\eta_{t-1}} \right) \|x_t - x^*\|^2.
$$

This and the initial settings of x_1, y_1, λ_1 give the assertion. □

Here we simplify Theorem 4. First note that by Eq. (S-5)

$$
\langle y^* - y_{T+1}, B^\top(\lambda_{T+1} - \lambda^*) \rangle \le \langle y^* - y_{T+1}, \nabla\psi(y_{T+1}) - B^\top \lambda^* \rangle.
$$

Since the diameters of $\mathcal{X}$ and $\mathcal{Y}$ are bounded by R, $\|g_t\| \le G$, and the subgradient of ψ is bounded by L_ψ, if $\|B^\top \lambda^*\| \le L_\psi$ and B is invertible, there exists a constant K depending on $R, B, \rho, \lambda^*, L_\psi, \eta_1$ such that the bound shown in Theorem 4 can be further bounded as

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

$$+\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}A^\top\tilde{\lambda}_t\\ B^\top\tilde{\lambda}_t\\ Ax_t+By_t-b\end{pmatrix}^\top\begin{pmatrix}x^*-x_t\\ y^*-y_t\\ \tilde{\lambda}_t-\lambda^*\end{pmatrix}$$

$$+\frac{1}{T}\sum_{t=1}^{T}\frac{\|\lambda_t-\lambda_{t+1}\|^2}{2\rho}+\frac{\|\lambda_{T+1}-\lambda^*\|^2}{2\rho T}$$

$$\le\frac{1}{T}\sum_{t=2}^{T}\max\left\{\frac{\gamma}{2\eta_t}-\frac{\gamma}{2\eta_{t-1}}-\frac{\sigma}{2},0\right\}R^2+\frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}G^2+\frac{K}{T} \tag{S-7}$$

$$=:\Xi_T. \tag{S-8}$$

Theorem 5. *Suppose ψ is Lipschitz continuous with a constant L_ψ, i.e., $|\psi(y)-\psi(y')|\le L_\psi\|y-y'\|$ $(\forall y,y'\in\mathcal{Y})$, and B is invertible. We utilize $y'_t:=B^{-1}(b-Ax_t)$ as an estimator of y at the t-th step, and let $\bar{y}'_t:=\frac{1}{t}\sum_{\tau=1}^{t}y'_\tau$. Then, for all $x^*\in\mathcal{X},y^*\in\mathcal{Y}$ such that $Ax^*+By^*-b=0$, there exists a constant K depending on R,A,B,L_ψ,ρ,η_1 such that*

$$\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T,w_T)+\psi(\bar{y}'_T))-(f(x^*,w_T)+\psi(y^*))]+\frac{\rho}{2}\mathrm{E}_{w_{1:T}}\left[\|A\bar{x}_{T+1}+B\bar{y}_{T+1}-b\|^2\right]$$
$$\le\frac{1}{T}\sum_{t=2}^{T}\max\left\{\frac{\gamma}{2\eta_t}-\frac{\gamma}{2\eta_{t-1}}-\frac{\sigma}{2},0\right\}R^2+\frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}G^2+\frac{K}{T}.$$

Proof. Note that

$$B(y'_t-y_t)=b-Ax_t-By_t=(\lambda_t-\lambda_{t-1})/\rho. \tag{S-9}$$

Thus, noting $y'_1=y_1$ by the initialization, we have

$$B(\bar{y}'_T-\bar{y}_T)=\frac{1}{\rho T}(\lambda_T-\lambda_1). \tag{S-10}$$

Since $b-Ax_t-By_t=(\lambda_t-\lambda_{t-1})/\rho$, the Lagrangian part in the statement of Theorem 4 can be bounded as

$$\begin{pmatrix}A^\top\tilde{\lambda}_t\\ B^\top\tilde{\lambda}_t\\ Ax_t+By_t-b\end{pmatrix}^\top\begin{pmatrix}x^*-x_t\\ y^*-y_t\\ \tilde{\lambda}_t-\lambda^*\end{pmatrix}=\begin{pmatrix}A^\top\lambda^*\\ B^\top\lambda^*\\ Ax^*+By^*-b\end{pmatrix}^\top\begin{pmatrix}x^*-x_t\\ y^*-y_t\\ \tilde{\lambda}_t-\lambda^*\end{pmatrix}$$
$$=\langle\lambda^*,A(x^*-x_t)+B(y^*-y_t)\rangle=\langle\lambda^*,b-Ax_t-By_t\rangle$$
$$=\langle\lambda^*,(\lambda_t-\lambda_{t-1})/\rho\rangle.$$

Therefore

$$\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}A^\top\tilde{\lambda}_t\\ B^\top\tilde{\lambda}_t\\ Ax_t+By_t-b\end{pmatrix}^\top\begin{pmatrix}x^*-x_t\\ y^*-y_t\\ \tilde{\lambda}_t-\lambda^*\end{pmatrix}=\langle\lambda^*,A(x^*-\bar{x}_T)+B(y^*-\bar{y}_T)\rangle$$
$$=\frac{1}{T\rho}\langle\lambda^*,\lambda_T-\lambda_1\rangle, \tag{S-11}$$

where we again used $Ax_1+By_1-b=0$.

Eq. (S-10) gives

$$f(\bar{x}_T,w_T)+\psi(\bar{y}'_T)\le f(\bar{x}_T,w_T)+\psi(\bar{y}_T)+\langle\nabla\psi(\bar{y}'_T),\bar{y}'_T-\bar{y}_T\rangle$$
$$\le f(\bar{x}_T,w_T)+\psi(\bar{y}_T)+\frac{1}{T\rho}\langle\nabla\psi(\bar{y}'_T),B^{-1}(\lambda_T-\lambda_1)\rangle.$$

Thus if we set

$$\lambda^*=B^{-\top}\nabla\psi(\bar{y}'_T),$$

(note that $\|B^\top \lambda^*\| \leq L_\psi$) then we have

$$f(\bar{x}_T, w_T) + \psi(\bar{y}'_T) \leq f(\bar{x}_T, w_T) + \psi(\bar{y}_T) + \frac{1}{T\rho}\langle \lambda^*, \lambda_T - \lambda_1 \rangle.$$

Since $\{w_t\}_{t=1}^T$ is independently identically distributed, Jensen's inequality yields

$$\begin{aligned}
&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T, w_T) + \psi(\bar{y}'_T)) - (f(x^*, w_T) + \psi(y^*))] \\
\leq &\mathrm{E}_{w_{1:T}}[f(\bar{x}_T, w_T) + \psi(\bar{y}_T) - (f(x^*, w_T) + \psi(y^*)) + \frac{1}{T\rho}\langle \lambda^*, \lambda_T - \lambda_1 \rangle] \\
\leq &\mathrm{E}_{w_{1:T}}\left[\frac{1}{T}\sum_{t=1}^T (f(x_t, w_T) + \psi(y_t)) - (f(x^*, w_T) + \psi(y^*)) + \frac{1}{T\rho}\langle \lambda^*, \lambda_T - \lambda_1 \rangle\right] \\
= &\mathrm{E}_{w_{1:T}}\left[\frac{1}{T}\sum_{t=1}^T (f(x_t, w_t) + \psi(y_t)) - (f(x^*, w_t) + \psi(y^*)) \right. \\
&\left. + \frac{1}{T}\sum_{t=1}^T \begin{pmatrix} A^\top \tilde{\lambda}_t \\ B^\top \tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \\ \tilde{\lambda}_t - \lambda^* \end{pmatrix}\right].
\end{aligned} \tag{S-12}$$

By combining this with Theorem 4 (and Eq. (S-8)), we have:

$$\begin{aligned}
&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T, w_T) + \psi(\bar{y}'_T)) - (f(x^*, w_T) + \psi(y^*))] + \mathrm{E}_{w_{1:T}}\left[\frac{1}{T}\sum_{t=1}^T \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}\right] \\
\leq &\Xi_T.
\end{aligned} \tag{S-13}$$

Finally we lower bound the second term of the LHS of Eq. (S-9). Remind that $\lambda_t - \lambda_{t+1} = \rho(Ax_{t+1} + By_{t+1} - b)$ (Eq. (S-10)). Thus

$$\begin{aligned}
\mathrm{E}_{w_{1:T}}\left[\frac{1}{T}\sum_{t=1}^T \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}\right] &= \mathrm{E}_{w_{1:T}}\left[\frac{\rho}{2T}\sum_{t=1}^T \|Ax_{t+1} + By_{t+1} - b\|^2\right] \\
&= \frac{T+1}{T}\mathrm{E}_{w_{1:T}}\left[\frac{\rho}{2(T+1)}\sum_{t=1}^{T+1} \|Ax_t + By_t - b\|^2\right] \quad (\because \text{the definition of } x_1, y_1) \\
&\geq \frac{\rho}{2}\mathrm{E}_{w_{1:T}}\left[\left\|\frac{1}{T+1}\sum_{t=1}^{T+1}(Ax_t + By_t - b)\right\|^2\right] \\
&= \frac{\rho}{2}\mathrm{E}_{w_{1:T}}\left[\|A\bar{x}_{T+1} + B\bar{y}_{T+1} - b\|^2\right].
\end{aligned} \tag{S-14}$$

This gives the assertion. □

Now substituting $\eta_t = \eta_0/\sqrt{t}$, Theorem 5 gives that, for all $\sigma \geq 0$,

$$\begin{aligned}
&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T, w_T) + \psi(\bar{y}'_T)) - (f(x^*, w_T) + \psi(y^*))] \\
\leq &\frac{\gamma}{T}\sum_{t=2}^T \left(\frac{1}{4\eta_0\sqrt{t-1}}\right) R^2 + \frac{1}{T}\sum_{t=1}^T \frac{\eta_0}{2\sqrt{t}}G^2 + \frac{K}{T} \\
\leq &\frac{\gamma}{2\eta_0\sqrt{T}}R^2 + \frac{\eta_0}{\sqrt{T}}G^2 + \frac{K}{T} \\
\leq &\frac{C_2}{\sqrt{T}},
\end{aligned} \tag{S-15}$$

where C_2 is a constant depending on $R, G, A, B, L_\psi, \rho, \eta_0, \gamma$. Moreover if $\sigma > 0$, by letting $\eta_t = \frac{\gamma}{\sigma t}$, we have that

$$\begin{aligned}&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T, w_T) + \psi(\bar{y}'_T)) - (f(x^*, w_T) + \psi(y^*))]\\ \leq& C'_2 \frac{\log(T)}{T},\end{aligned} \tag{S-16}$$

C'_2 is a constant depending on $R, G, A, B, L_\psi, \rho, \eta_0, \gamma, \sigma$.

Theorem 6. *Suppose ψ is Lipschitz continuous with a constant L_ψ, i.e., $|\psi(y) - \psi(y')| \leq L_\psi \|y - y'\|$ $(\forall y, y' \in \mathcal{Y})$, and A is invertible. We utilize $x'_t := A^{-1}(b - By_t)$ instead of x_t at the t-th step, and let $\bar{x}'_t := \frac{1}{t}\sum_{\tau=1}^{t} x'_\tau$. Then, for all $x^* \in \mathcal{X}, y^* \in \mathcal{Y}$ such that $Ax^* + By^* - b = 0$, there exists a constant K depending on R, G, A, B, ρ, η_1 such that*

$$\begin{aligned}&\mathrm{E}_{w_{1:T}}[(f(\bar{x}'_T, w_T) + \psi(\bar{y}_T)) - (f(x^*, w_T) + \psi(y^*))]\\ \leq& \frac{1}{T}\sum_{t=2}^{T}\left(\frac{\gamma}{2\eta_t} - \frac{\gamma}{2\eta_{t-1}} - \frac{\sigma}{2}\right)R^2 + \frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}G^2 + \frac{K}{T}.\end{aligned}$$

In particular, for $\eta_t = \eta_0/\sqrt{t}$, the RHS is further bounded as

$$\begin{aligned}&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T, w_T) + \psi(\bar{y}'_T)) - (f(x^*, w_T) + \psi(y^*))]\\ \leq& \frac{C_2}{\sqrt{T}},\end{aligned} \tag{S-17}$$

where C_2 is a constant depending on $R, G, A, B, L_\psi, \rho, \eta_0, \gamma$. Moreover, if $\sigma > 0$, by letting $\eta_t = \gamma/(\sigma t)$, we have that

$$\begin{aligned}&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T, w_T) + \psi(\bar{y}'_T)) - (f(x^*, w_T) + \psi(y^*))]\\ \leq& C'_2 \frac{\log(T)}{T},\end{aligned} \tag{S-18}$$

where C'_2 is a constant depending on $R, G, A, B, L_\psi, \rho, \eta_0, \gamma, \sigma$.

Proof. Eq. (S-10) gives

$$\begin{aligned}f(\bar{x}'_T, w_T) + \psi(\bar{y}_T) &\leq f(\bar{x}'_T, w_T) + \psi(\bar{y}_T) + \langle \nabla_x f(x, w_T)|_{x=\bar{x}'_T}, \bar{x}'_T - \bar{x}_T\rangle\\ &= f(\bar{x}_T, w_T) + \psi(\bar{y}_T) + \langle \nabla_x f(x, w_T)|_{x=\bar{x}'_T}, A^{-1}(b - B\bar{y} - A\bar{x}_T)\rangle\\ &= f(\bar{x}_T, w_T) + \psi(\bar{y}_T) + \frac{1}{T\rho}\langle \nabla_x f(x, w_T)|_{x=\bar{x}'_T}, A^{-1}(\lambda_T - \lambda_1)\rangle.\end{aligned}$$

Thus if we set

$$\lambda^* = A^{-\top}\nabla_x f(x, w_T)|_{x=\bar{x}'_T},$$

(note that $\|A^{-\top}\nabla_x f(x, w_T)|_{x=\bar{x}'_T}\| \leq \|A^{-1}\|G$), then we have

$$f(\bar{x}'_T, w_T) + \psi(\bar{y}_T) \leq f(\bar{x}_T, w_T) + \psi(\bar{y}_T) + \frac{1}{T\rho}\langle \lambda^*, \lambda_T - \lambda_1\rangle.$$

With the same reasoning as Eq. (S-12), we have

$$\begin{aligned}&\mathrm{E}_{w_{1:T}}[(f(\bar{x}'_T, w_T) + \psi(\bar{y}_T)) - (f(x^*, w_T) + \psi(y^*))]\\ \leq& \mathrm{E}_{w_{1:T}}\Bigg[\frac{1}{T}\sum_{t=1}^{T}(f(x_t, w_t) + \psi(y_t)) - (f(x^*, w_t) + \psi(y^*))\\ &+ \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} A^\top\tilde{\lambda}_t \\ B^\top\tilde{\lambda}_t \\ Ax_t + By_t - b\end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \\ \tilde{\lambda}_t - \lambda^*\end{pmatrix}\Bigg].\end{aligned}$$

By combining this with Theorem 4 (and Eq. (S-8)), we have:

$$\begin{aligned}&\mathrm{E}_{w_{1:T}}[(f(\bar{x}'_T, w_T) + \psi(\bar{y}_T)) - (f(x^*, w_T) + \psi(y^*))] \\ \le&\Xi_T. \end{aligned} \tag{S-19}$$

This gives the assertion. □

B. Convergence rate of RDA-ADMM

We define the generalized RDA-ADMM corresponding to (S-1) as follows: Let $x_1 = \mathbf{0}$, $\lambda_1 = \mathbf{0}$, and $By_1 = b - Ax_1$, and the update of the t-th step is given by

$$x_{t+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ \bar{g}_t^\top x - \bar{\lambda}_t^\top Ax + \frac{\rho}{2t}\|Ax\|^2 + \rho(A\bar{x}_t + B\bar{y}_t - b)^\top Ax + \frac{1}{2\eta_t}\|x\|_{G_t}^2 \right\}, \tag{S-20a}$$

$$y_{t+1} = \operatorname*{argmin}_{y\in\mathcal{Y}} \left\{ \psi(y) - \lambda_t^\top (Ax_{t+1} + By - b) + \frac{\rho}{2}\|Ax_{t+1} + By - b\|^2 \right\}, \tag{S-20b}$$

$$\lambda_{t+1} = \lambda_t - \rho(Ax_{t+1} + By_{t+1} - b), \tag{S-20c}$$

where $G_t = \gamma I - \frac{\rho\eta_t}{t}A^\top A$. As in the analysis of OPG-ADMM, we define

$$\tilde{\lambda}_t = \lambda_t - \rho(Ax_{t+1} + By_t - b). \tag{S-21}$$

Moreover we suppose that η_t/t is non-increasing.

Theorem 7. *For all $x^* \in \mathcal{X}$, $y^* \in \mathcal{Y}$ and $\lambda^* \in \mathbb{R}^l$, we have*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T (f(x_t, w_t) + \psi(y_t)) - \frac{1}{T}\sum_{t=1}^T (f(x^*, w_t) + \psi(y^*)) \\
&+ \frac{1}{T}\sum_{t=1}^T \begin{pmatrix} A^\top\tilde{\lambda}_t \\ B^\top\tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \\ \tilde{\lambda}_t - \lambda^* \end{pmatrix} + \frac{1}{T}\sum_{t=1}^T \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho} \\
\le& \frac{1}{T}\sum_{t=2}^T \frac{\eta_{t-1}}{2(t-1)}\|g_t\|_{G_{t-1}^{-1}}^2 + \frac{\eta_1}{T}\|g_1\|_{G_1^{-1}}^2 + \frac{1}{\eta_T}\|x^*\|_{G_T} \\
&+ \frac{\rho}{2T}\|b - By^*\|^2 + \frac{\|\lambda^*\|^2}{2\rho T} \\
&+ \frac{1}{T}\left(\langle y^* - y_{T+1}, B^\top(\lambda_{T+1} - \lambda^*)\rangle - \langle By^* - b, \lambda^*\rangle\right) \\
&+ \frac{\rho}{2T}\|A(x_{T+1} - x^*)\|^2 + \frac{1}{T}\langle\lambda^*, Ax_{T+1}\rangle.
\end{aligned}$$

Proof. First observe that

$$\begin{aligned}
\sum_{t=1}^T (f_t(x_t) + \psi(y_t) - f_t(x^*) - \psi(y^*)) &\le \sum_{t=1}^T (\langle g_t, x_t - x^*\rangle + \langle\nabla\psi(y_t), y_t - y^*\rangle) \\
&= \sum_{t=1}^T (\langle g_t, x_t - x^*\rangle + \langle B^\top\lambda_t, y_t - y^*\rangle) \\
&= \sum_{t=1}^T (\langle g_t, x_t\rangle + \langle\lambda_t, B(y_t - y^*)\rangle) - T\langle\bar{g}_T, x^*\rangle.
\end{aligned} \tag{S-22}$$

Here we define ϱ_t and V_t as follows:

$$\begin{aligned}\varrho_t(g,x) &= \langle g, x\rangle + \langle \bar{\lambda}_t, Ax\rangle - \frac{\rho}{2t}\|Ax\|^2 - \rho(A\bar{x}_t + B\bar{y}_t - b)^\top Ax - \frac{1}{2\eta_t}\|x\|^2_{G_t},\\ V_t(g) &= \max_{x\in\mathcal{X}}\left\{\varrho_t(g,x)\right\}.\end{aligned}$$

Obviously $V_t(-\bar{g}_t) = \varrho_t(-\bar{g}_t, x_{t+1})$ by the update rule of x_t. In particular $\langle x' - x_t, \nabla_x \varrho_{t-1}(-\bar{g}_{t-1}, x)|_{x=x_t}\rangle \le 0$ for all $x' \in \mathcal{X}$. Because $\varrho_{t-1}(-\bar{g}_{t-1}, \cdot)$ is a concave quadratic function, we have

$$\begin{aligned}&\varrho_{t-1}(-\bar{g}_{t-1}, x_{t+1})\\ \le&\varrho_{t-1}(-\bar{g}_{t-1}, x_t) + \langle x_{t+1} - x_t, \nabla_x\varrho_{t-1}(-\bar{g}_{t-1}, x)|_{x=x_t}\rangle + \frac{1}{2}(x_{t+1}-x_t)^\top(\nabla_x^\top\nabla_x\varrho_{t-1}(-\bar{g}_{t-1}, x)|_{x=x_t})(x_{t+1}-x_t)\\ \le&\varrho_{t-1}(-\bar{g}_{t-1}, x_t) + \frac{1}{2}(x_{t+1}-x_t)^\top(\nabla_x^\top\nabla_x\varrho_{t-1}(-\bar{g}_{t-1}, x)|_{x=x_t})(x_{t+1}-x_t).\end{aligned}$$

Here note that $\nabla_x^\top\nabla_x\varrho_{t-1}(-\bar{g}_{t-1}, x)|_{x=x_t} = -\frac{\rho}{t-1}A^\top A - \frac{G_{t-1}}{\eta_{t-1}}$. Therefore

$$\begin{aligned}\varrho_{t-1}(-\bar{g}_{t-1}, x_{t+1}) &\le \varrho_{t-1}(-\bar{g}_{t-1}, x_t) - \frac{\rho}{2(t-1)}\|A(x_{t+1}-x_t)\|^2 - \frac{1}{2\eta_{t-1}}\|x_{t+1}-x_t\|^2_{G_{t-1}}\\ &= V_{t-1}(-\bar{g}_{t-1}) - \frac{\rho}{2(t-1)}\|A(x_{t+1}-x_t)\|^2 - \frac{1}{2\eta_{t-1}}\|x_{t+1}-x_t\|^2_{G_{t-1}}.\end{aligned}$$

Using this inequality, we can compare $V_t(-\bar{g}_t)$ to $V_{t-1}(-\bar{g}_{t-1})$:

$$\begin{aligned}tV_t(-\bar{g}_t) &= t\varrho_t(-\bar{g}_t, x_{t+1})\\ &\le (t-1)\varrho_{t-1}(-\bar{g}_{t-1}, x_{t+1}) + \langle -g_t, x_{t+1}\rangle + \langle\lambda_t, Ax_{t+1}\rangle - \rho\langle Ax_{t+1}, Ax_t + By_t - b\rangle \quad\text{(S-23)}\\ &\le (t-1)V_{t-1}(-\bar{g}_{t-1}) - \frac{\rho}{2}\|A(x_{t+1}-x_t)\|^2 - \frac{t-1}{2\eta_{t-1}}\|x_{t+1}-x_t\|^2_{G_{t-1}}\\ &\quad - \langle g_t, x_t\rangle + \langle g_t, x_t - x_{t+1}\rangle + \langle\lambda_t, Ax_{t+1}\rangle - \rho\langle Ax_{t+1}, Ax_t + By_t - b\rangle. \quad\text{(S-24)}\end{aligned}$$

Since $\langle g_t, x_t - x_{t+1}\rangle \le \|g_t\|_{G_{t-1}^{-1}}\|x_t - x_{t+1}\|_{G_{t-1}} \le \frac{\eta_{t-1}}{2(t-1)}\|g_t\|^2_{G_{t-1}^{-1}} + \frac{t-1}{2\eta_{t-1}}\|x_t - x_{t+1}\|^2_{G_{t-1}}$, the RHS is further bounded by

$$\begin{aligned}tV_t(-\bar{g}_t) \le&(t-1)V_{t-1}(-\bar{g}_{t-1}) - \frac{\rho}{2}\|A(x_{t+1}-x_t)\|^2 + \frac{\eta_{t-1}}{2(t-1)}\|g_t\|^2_{G_t^{-1}}\\ &- \langle g_t, x_t\rangle + \langle\lambda_t, Ax_{t+1}\rangle - \rho\langle Ax_{t+1}, Ax_t + By_t - b\rangle.\end{aligned}$$

Moreover

$$\begin{aligned}V_1(-g_1) =&\langle -g_1, x_2\rangle + \langle\lambda_1, Ax_2\rangle - \frac{\rho}{2}\|Ax_2\|^2 - \rho\langle Ax_1 + By_1 - b, Ax_2\rangle - \frac{1}{2\eta_1}\|x_2\|^2_{G_1}\\ \le&\langle -g_1, x_1\rangle + \langle\lambda_1, Ax_2\rangle - \frac{\rho}{2}\|A(x_1 - x_2)\|^2 - \rho\langle Ax_1 + By_1 - b, Ax_2\rangle\\ &- \frac{1}{2\eta_1}\|x_2 + \eta_1 G_1^{-1}g_1\|^2_{G_1} + \frac{\eta_1}{2}\|g_1\|^2_{G_1^{-1}} + \langle g_1, x_1\rangle - \rho\langle Ax_2, Ax_1\rangle + \frac{\rho}{2}\|Ax_1\|^2.\end{aligned}$$

Summing up $t = 1, \ldots, T$, we have

$$\begin{aligned}TV_T(-\bar{g}_T) \le& -\frac{\rho}{2}\sum_{t=1}^T\|A(x_{t+1}-x_t)\|^2 + \sum_{t=2}^T\frac{\eta_{t-1}}{2(t-1)}\|g_t\|^2_{G_t^{-1}}\\ &- \sum_{t=1}^T\langle g_t, x_t\rangle + \sum_{t=1}^T\langle\lambda_t, Ax_{t+1}\rangle - \sum_{t=1}^T\rho\langle Ax_{t+1}, Ax_t + By_t - b\rangle\\ &+ \frac{\eta_1}{2}\|g_1\|^2_{G_1^{-1}} + \langle g_1, x_1\rangle - \rho\langle Ax_2, Ax_1\rangle + \frac{\rho}{2}\|Ax_1\|^2. \quad\text{(S-25)}\end{aligned}$$

Using this inequality, we observe that

$$
\begin{aligned}
& -\langle T\bar{g}_T, x^*\rangle \\
=& T\varrho_T(-\bar{g}_T, x^*) - \langle T\bar{\lambda}_T, Ax^*\rangle + \frac{\rho}{2}\|Ax^*\|^2 + T\rho\langle Ax^*, A\bar{x}_T + B\bar{y}_T - b\rangle + \frac{T}{\eta_T}\|x^*\|^2_{G_T} \\
\le& TV_T(-\bar{g}_T) - \langle T\bar{\lambda}_T, Ax^*\rangle + \frac{\rho}{2}\|Ax^*\|^2 + T\rho\langle Ax^*, A\bar{x}_T + B\bar{y}_T - b\rangle + \frac{T}{\eta_T}\|x^*\|^2_{G_T} \\
\le& -\frac{\rho}{2}\sum_{t=1}^{T}\|A(x_{t+1}-x_t)\|^2 - \sum_{t=2}^{T}\frac{\eta_{t-1}}{2(t-1)}\|g_t\|^2_{G_t^{-1}} \\
& -\sum_{t=1}^{T}\langle g_t, x_t\rangle + \sum_{t=1}^{T}\langle \lambda_t, Ax_{t+1}\rangle - \sum_{t=1}^{T}\rho\langle Ax_{t+1}, Ax_t + By_t - b\rangle \\
& + \frac{\eta_1}{2}\|g_1\|^2_{G_1^{-1}} + \langle g_1, x_1\rangle - \rho\langle Ax_2, Ax_1\rangle + \frac{\rho}{2}\|Ax_1\|^2 \\
& - \langle T\bar{\lambda}_T, Ax^*\rangle + \frac{\rho}{2}\|Ax^*\|^2 + T\rho\langle Ax^*, A\bar{x}_T + B\bar{y}_T - b\rangle + \frac{T}{\eta_T}\|x^*\|^2_{G_T} \quad (\because \text{Eq. (S-25)}) \\
=& -\frac{\rho}{2}\sum_{t=1}^{T}\|A(x_{t+1}-x_t)\|^2 + \sum_{t=2}^{T}\frac{\eta_{t-1}}{2(t-1)}\|g_t\|^2_{G_t^{-1}} \\
& -\sum_{t=1}^{T}\langle g_t, x_t\rangle + \sum_{t=1}^{T}\langle \lambda_t - \rho(Ax_t + By_t - b), A(x_{t+1}-x^*)\rangle + \frac{\rho}{2}\|Ax^*\|^2 + \frac{T}{\eta_T}\|x^*\|^2_{G_T} \\
& + \frac{\eta_1}{2}\|g_1\|^2_{G_1^{-1}} + \langle g_1, x_1\rangle - \rho\langle Ax_2, Ax_1\rangle + \frac{\rho}{2}\|Ax_1\|^2.
\end{aligned}
$$

Substituting this inequality into the RHS of Eq. (S-22), we obtain

$$
\begin{aligned}
& \sum_{t=1}^{T}(\langle g_t, x_t\rangle + \langle \lambda_t, B(y_t - y^*)\rangle) - T\langle \bar{g}_T, x^*\rangle \\
\le& \sum_{t=1}^{T}\langle \lambda_t, B(y_t - y^*)\rangle + \sum_{t=1}^{T}\langle \lambda_t - \rho(Ax_t + By_t - b), A(x_{t+1}-x^*)\rangle \\
& - \frac{\rho}{2}\sum_{t=1}^{T}\|A(x_{t+1}-x_t)\|^2 + \sum_{t=2}^{T}\frac{\eta_{t-1}}{2(t-1)}\|g_t\|^2_{G_t^{-1}} \\
& + \frac{\rho}{2}\|Ax^*\|^2 + \frac{T}{\eta_T}\|x^*\|^2_{G_T} \\
& + \frac{\eta_1}{2}\|g_1\|^2_{G_1^{-1}} + \langle g_1, x_1\rangle - \rho\langle Ax_2, Ax_1\rangle + \frac{\rho}{2}\|Ax_1\|^2.
\end{aligned} \tag{S-26}
$$

From now on, we bound the first two terms of the RHS: (i) $\sum_{t=1}^{T}\langle \lambda_t, B(y_t - y^*)\rangle$ and (ii) $\sum_{t=1}^{T}\langle \lambda_t - \rho(Ax_t + By_t - b), A(x_{t+1}-x^*)\rangle$.

(i) *Evaluating* $\sum_{t=1}^{T}\langle \lambda_t, B(y_t - y^*)\rangle$. We have

$$
\langle \lambda_t, B(y_t - y^*)\rangle = \langle \tilde{\lambda}_t, B(y_t - y^*)\rangle + \langle B(y_t - y^*), \lambda_t - \tilde{\lambda}_t\rangle. \tag{S-27}
$$

(ii) *Evaluating* $\sum_{t=1}^{T}\langle \lambda_t - \rho(Ax_t + By_t - b), A(x_{t+1}-x^*)\rangle$. By the definition of $\tilde{\lambda}_t$, we have

$$
\begin{aligned}
& \langle \lambda_t - \rho(Ax_t + By_t - b), A(x_{t+1}-x^*)\rangle = \langle \tilde{\lambda}_t + \rho A(x_{t+1}-x_t), A(x_{t+1}-x^*)\rangle \\
=& \langle \tilde{\lambda}_t, A(x_t - x^*)\rangle + \langle \tilde{\lambda}_t - \lambda^*, A(x_{t+1}-x_t)\rangle + \langle \lambda^*, A(x_{t+1}-x_t)\rangle \\
& + \rho\langle A(x_{t+1}-x_t), A(x_{t+1}-x^*)\rangle
\end{aligned}
$$

$$\leq \langle \tilde{\lambda}_t, A(x_t - x^*) \rangle + \langle \tilde{\lambda}_t - \lambda^*, A(x_{t+1} - x_t) \rangle + \langle \lambda^*, A(x_{t+1} - x_t) \rangle$$
$$+ \frac{\rho}{2}\|A(x_{t+1} - x^*)\|^2 - \frac{\rho}{2}\|A(x_t - x^*)\|^2 + \frac{\rho}{2}\|A(x_{t+1} - x_t)\|^2 \quad (\because \text{Lemma 10}). \tag{S-28}$$

Here by substituting the relation $A(x_{t+1} - x_t) = \frac{\tilde{\lambda}_t - \lambda_t}{\rho} - (Ax_t + By_t - b)$ to the second term in the RHS, we have

$$\langle \lambda_t - \rho(Ax_t + By_t - b), A(x_{t+1} - x^*) \rangle$$
$$= \langle \tilde{\lambda}_t, A(x_t - x^*) \rangle + \left\langle \tilde{\lambda}_t - \lambda^*, \frac{\tilde{\lambda}_t - \lambda_t}{\rho} - (Ax_t + By_t - b) \right\rangle + \langle \lambda^*, A(x_{t+1} - x_t) \rangle$$
$$+ \frac{\rho}{2}\|A(x_{t+1} - x^*)\|^2 - \frac{\rho}{2}\|A(x_t - x^*)\|^2 + \frac{\rho}{2}\|A(x_{t+1} - x_t)\|^2.$$

Substituting Eqs. (S-27),(S-28) into the RHS of Eq. (S-30), we have that

$$\sum_{t=1}^{T} (\langle g_t, x_t \rangle + \langle \lambda_t, B(y_t - y^*) \rangle) - T\langle \bar{g}_T, x^* \rangle$$
$$\leq \sum_{t=1}^{T} \begin{pmatrix} -A^\top \tilde{\lambda}_t \\ -B^\top \tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix} + \sum_{t=1}^{T} \begin{pmatrix} y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix}^\top \begin{pmatrix} -B^\top(\lambda_t - \tilde{\lambda}_t) \\ \frac{\tilde{\lambda}_t - \lambda_t}{\rho}, \end{pmatrix}$$
$$+ \langle \lambda^*, A(x_{T+1} - x_1) \rangle + \frac{\rho}{2}\|A(x_{T+1} - x^*)\|^2 - \frac{\rho}{2}\|A(x_1 - x^*)\|^2 + \sum_{t=2}^{T} \frac{\eta_{t-1}}{2(t-1)} \|g_t\|^2_{G_t^{-1}}$$
$$+ \frac{\rho}{2}\|Ax^*\|^2 + \frac{T}{\eta_T}\|x^*\|^2_{G_T} + \frac{\eta_1}{2}\|g_1\|^2_{G_1^{-1}} + \langle g_1, x_1 \rangle - \rho\langle Ax_2, Ax_1 \rangle + \frac{\rho}{2}\|Ax_1\|^2. \tag{S-29}$$

Finally we bound the second term of the RHS using Lemma 11, we obtain

$$\sum_{t=1}^{T} (\langle g_t, x_t \rangle + \langle \lambda_t, B(y_t - y^*) \rangle) - T\langle \bar{g}_T, x^* \rangle$$
$$\leq \sum_{t=1}^{T} \begin{pmatrix} -A^\top \tilde{\lambda}_t \\ -B^\top \tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^\top \begin{pmatrix} x^* - x_t \\ y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix}$$
$$+ \frac{\rho}{2}\|y_1 - y^*\|^2_{B^\top B} + \frac{\|\lambda_1 - \lambda^*\|^2}{2\rho}$$
$$+ \langle B(y^* - y_{T+1}), \lambda_{T+1} - \lambda^* \rangle - \langle B(y^* - y_1), \lambda_1 - \lambda^* \rangle - \sum_{t=1}^{T} \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}$$
$$+ \langle \lambda^*, A(x_{T+1} - x_1) \rangle$$
$$+ \frac{\rho}{2}\|A(x_{T+1} - x^*)\|^2 - \frac{\rho}{2}\|A(x_1 - x^*)\|^2 + \sum_{t=2}^{T} \frac{\eta_{t-1}}{2(t-1)} \|g_t\|^2_{G_t^{-1}}$$
$$+ \frac{\rho}{2}\|Ax^*\|^2 + \frac{T}{\eta_T}\|x^*\|^2_{G_T} + \frac{\eta_1}{2}\|g_1\|^2_{G_1^{-1}} + \langle g_1, x_1 \rangle - \rho\langle Ax_2, Ax_1 \rangle + \frac{\rho}{2}\|Ax_1\|^2. \tag{S-30}$$

This, Eq. (S-22) and the initial settings of x_1, y_1, λ_1 give the assertion. □

Here we simplify Theorem 7. First note that by Eq. (S-5)

$$\langle y^* - y_{T+1}, B^\top(\lambda_{T+1} - \lambda^*) \rangle \leq \langle y^* - y_{T+1}, \nabla\psi(y_{T+1}) - B^\top\lambda^* \rangle.$$

Since the diameters of $\mathcal{X}$ and $\mathcal{Y}$ are bounded by R, $\|g_t\| \le G$, and the subgradient of ψ is bounded by L_ψ, if B is invertible, then there exists a constant K depending on $R, G, L_\psi, \rho, A, B, \eta_1, \lambda^*$ such that the bound shown in Theorem 7 can be further bounded as

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(f(x_t,w_t)+\psi(y_t)) - \frac{1}{T}\sum_{t=1}^{T}(f(x^*,w_t)+\psi(y^*)) \\
&- \langle \lambda^*, A\bar{x}_T + B\bar{y}_T - (Ax^*+By^*)\rangle + \frac{1}{T}\sum_{t=1}^{T}\frac{\|\lambda_t-\lambda_{t+1}\|^2}{2\rho} \\
\le& \frac{1}{T}\sum_{t=2}^{T}\frac{\eta_{t-1}}{2(t-1)}G^2 + \frac{\gamma}{\eta_T}R^2 + \frac{K}{T}, \qquad \text{(S-31)}
\end{aligned}$$

where we used Eq. (S-11).

Theorem 8. *Suppose ψ is Lipschitz continuous with a constant L_ψ, i.e., $|\psi(y)-\psi(y')| \le L_\psi\|y-y'\|$ $(\forall y, y' \in \mathcal{Y})$, and B is invertible. We utilize $y'_t := B^{-1}(b - Ax_t)$ as an estimator of y at the t-th step, and let $\bar{y}'_t := \frac{1}{t}\sum_{\tau=1}^{t} y'_\tau$. Then, for all $x^* \in \mathcal{X}, y^* \in \mathcal{Y}$ such that $Ax^* + By^* - b = 0$, there exists a constant K depending on $R, G, L_\psi, \rho, A, B, \eta_1$ such that*

$$\begin{aligned}
&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T,w_T)+\psi(\bar{y}'_T)) - (f(x^*,w_T)+\psi(y^*))] \\
\le& \frac{1}{T}\sum_{t=2}^{T}\frac{\eta_{t-1}}{2(t-1)}G^2 + \frac{\gamma}{\eta_T}R^2 + \frac{K}{T}.
\end{aligned}$$

Proof. Using $\eta_t = \eta_0\sqrt{t}$ and Theorem 7 instead of $\eta_t = \eta_0/\sqrt{t}$ and Theorem 4 respectively, the same proof as Theorem 5 yields the assertion. □

Substituting $\eta_t = \eta_0\sqrt{t}$, the bound in Theorem 8 can be simplified as

$$\begin{aligned}
&\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T,w_T)+\psi(\bar{y}'_T)) - (f(x^*,w_T)+\psi(y^*))] \\
\le& \frac{1}{T}\sum_{t=2}^{T}\frac{\eta_0\sqrt{t-1}}{2(t-1)}G^2 + \frac{\gamma}{\eta_0\sqrt{T}}R^2 + \frac{K}{T} \\
\le& \frac{\eta_0}{\sqrt{T}}G^2 + \frac{\gamma}{\eta_0\sqrt{T}}R^2 + \frac{K}{T} \\
\le& \frac{C_2}{\sqrt{T}},
\end{aligned}$$

where C_2 is a constant depending on $R, G, L_\psi, \rho, \eta_0, A, B, G, \gamma$.

Theorem 9. *Suppose ψ is Lipschitz continuous with a constant L_ψ, i.e., $|\psi(y)-\psi(y')| \le L_\psi\|y-y'\|$ $(\forall y, y' \in \mathcal{Y})$, and A is invertible. We utilize $x'_t := A^{-1}(b - By_t)$ instead of x_t at the t-th step, and let $\bar{x}'_t := \frac{1}{t}\sum_{\tau=1}^{t} x'_\tau$. Then, for all $x^* \in \mathcal{X}, y^* \in \mathcal{Y}$ such that $Ax^* + By^* - b = 0$, there exists a constant K depending on R, G, A, B, ρ, η_1 such that*

$$\begin{aligned}
&\mathrm{E}_{w_{1:T}}[(f(\bar{x}'_T,w_T)+\psi(\bar{y}_T)) - (f(x^*,w_T)+\psi(y^*))] \\
\le& \frac{1}{T}\sum_{t=2}^{T}\frac{\eta_{t-1}}{2(t-1)}G^2 + \frac{\gamma}{\eta_T}R^2 + \frac{K}{T}.
\end{aligned}$$

In particular, for $\eta_t = \eta_0\sqrt{t}$, the RHS is further bounded as

$$\mathrm{E}_{w_{1:T}}[(f(\bar{x}_T,w_T)+\psi(\bar{y}'_T)) - (f(x^*,w_T)+\psi(y^*))] \le \frac{C_2}{\sqrt{T}}, \qquad \text{(S-32)}$$

where C_2 is a constant depending on $R, G, A, B, L_\psi, \rho, \eta_0, \gamma$.

Proof. Using $\eta_t = \eta_0\sqrt{t}$ and Theorem 7 instead of $\eta_t = \eta_0/\sqrt{t}$ and Theorem 4 respectively, the same proof as Theorem 6 yields the assertion. □

C. Auxiliary Lemmas

Lemma 10. *For all symmetric matrix H, we have*

$$(a-b)^\top H(c-b) = \frac{1}{2}\|a-b\|_H^2 - \frac{1}{2}\|a-c\|_H^2 + \frac{1}{2}\|c-b\|_H^2. \tag{S-33}$$

Proof.

$$\begin{aligned}(a-b)^\top H(c-b) &= \left(a - \frac{c+b}{2} + \frac{c+b}{2} - b\right)^\top H(c-b)\\ &= \left(\frac{a-c}{2} + \frac{a-b}{2}\right)^\top H(c-b) + \left(\frac{c-b}{2}\right)^\top H(c-b)\\ &= \left(\frac{a-c}{2} + \frac{a-b}{2}\right)^\top H\{(a-b)-(a-c)\} + \left(\frac{c-b}{2}\right)^\top H(c-b)\\ &= \frac{1}{2}\|a-b\|_H^2 - \frac{1}{2}\|a-c\|_H^2 + \frac{1}{2}\|c-b\|_H^2.\end{aligned}$$

□

Lemma 11. *Under the update rule* (S-2) *or* (S-20)*, we have the following bound:*

$$\begin{aligned}&\begin{pmatrix} y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix}^\top \begin{pmatrix} -B^\top(\lambda_t - \tilde{\lambda}_t) \\ \frac{\tilde{\lambda}_t - \lambda_t}{\rho} \end{pmatrix}\\ \leq &\frac{\rho}{2}\|y_t - y^*\|_{B^\top B}^2 - \frac{\rho}{2}\|y_{t+1} - y^*\|_{B^\top B}^2 + \frac{\|\lambda_t - \lambda^*\|^2}{2\rho} - \frac{\|\lambda_{t+1} - \lambda^*\|^2}{2\rho}\\ &+ \langle B(y^* - y_{t+1}), \lambda_{t+1} - \lambda^*\rangle - \langle B(y^* - y_t), \lambda_t - \lambda^*\rangle\\ &- \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}.\end{aligned} \tag{S-34}$$

Proof. Since $\tilde{\lambda}_t = \lambda_{t+1} - \rho B(y_t - y_{t+1})$, we have

$$\begin{aligned}&\langle y^* - y_t, -B^\top(\lambda_t - \tilde{\lambda}_t)\rangle\\ =&\langle y^* - y_t, -B^\top(\lambda_t - \lambda_{t+1})\rangle + \langle y^* - y_t, \rho B^\top B(y_{t+1} - y_t)\rangle\\ \leq&\langle y^*, B^\top(\lambda_{t+1} - \lambda_t)\rangle - \langle y_t, B^\top(\lambda_{t+1} - \lambda_t)\rangle\\ &+ \frac{\rho}{2}\|y_t - y^*\|_{B^\top B}^2 - \frac{\rho}{2}\|y_{t+1} - y^*\|_{B^\top B}^2 + \frac{\rho}{2}\|y_{t+1} - y_t\|_{B^\top B}^2,\end{aligned} \tag{S-35}$$

where we used Lemma 10 in the last inequality.

On the other hand, by Lemma 10, we have

$$\begin{aligned}&\langle \lambda^* - \tilde{\lambda}_t, (\tilde{\lambda}_t - \lambda_t)/\rho\rangle\\ =&\underbrace{-\frac{\|\tilde{\lambda}_t - \lambda^*\|^2}{2\rho}}_{\text{(i)}} + \frac{\|\lambda_t - \lambda^*\|^2}{2\rho} \underbrace{-\frac{\|\tilde{\lambda}_t - \lambda_t\|^2}{2\rho}}_{\text{(ii)}}.\end{aligned} \tag{S-36}$$

The first term (i) in Eq. (S-36) can be evaluated as

$$\begin{aligned}-\frac{\|\tilde{\lambda}_t - \lambda^*\|^2}{2\rho} &= -\frac{\|\tilde{\lambda}_t - \lambda_{t+1}\|^2}{2\rho} - \frac{\langle \tilde{\lambda}_t - \lambda_{t+1}, \lambda_{t+1} - \lambda^*\rangle}{\rho} - \frac{\|\lambda_{t+1} - \lambda^*\|^2}{2\rho}\\ &= -\frac{\|\tilde{\lambda}_t - \lambda_{t+1}\|^2}{2\rho} - \frac{\langle \rho B(y_{t+1} - y_t), \lambda_{t+1} - \lambda^*\rangle}{\rho} - \frac{\|\lambda_{t+1} - \lambda^*\|^2}{2\rho}.\end{aligned} \tag{S-37}$$

Since $\tilde{\lambda}_t - \lambda_{t+1} = \rho B(y_{t+1} - y_t)$, the third term (ii) in Eq. (S-36) can be evaluated as

$$-\frac{\|\tilde{\lambda}_t - \lambda_t\|^2}{2\rho} = -\frac{\|\tilde{\lambda}_t - \lambda_{t+1}\|^2}{2\rho} - \frac{1}{\rho}\langle \tilde{\lambda}_t - \lambda_{t+1}, \lambda_{t+1} - \lambda_t\rangle - \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}$$
$$= -\frac{\rho\|B(y_{t+1} - y_t)\|^2}{2} - \frac{1}{\rho}\langle \rho B(y_{t+1} - y_t), \lambda_{t+1} - \lambda_t\rangle - \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}. \tag{S-38}$$

Here note that, by Eq. (S-5), we have

$$\langle B(y_{t+1} - y_t), \lambda_{t+1} - \lambda_t\rangle \geq \langle y_{t+1} - y_t, \nabla\psi(y_{t+1}) - \nabla\psi(y_t)\rangle \geq 0,$$

by the monotonicity of subgradients. Therefore the RHS of Eq. (S-38) is further bounded as

$$-\frac{\|\tilde{\lambda}_t - \lambda_t\|^2}{2\rho} \leq -\frac{\rho\|y_{t+1} - y_t\|^2_{B^\top B}}{2} - \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}. \tag{S-39}$$

Combining Eqs. (S-36), (S-37), (S-39), then

$$\begin{aligned}
&\langle \lambda^* - \tilde{\lambda}_t, (\tilde{\lambda}_t - \lambda_t)/\rho\rangle \\
\leq& -\frac{\rho}{2}\|y_{t+1} - y_t\|^2_{B^\top B} - \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho} + \frac{\|\lambda_t - \lambda^*\|^2}{2\rho} \\
&- \frac{\|\tilde{\lambda}_t - \lambda_{t+1}\|^2}{2\rho} - \langle B(y_{t+1} - y_t), \lambda_{t+1} - \lambda^*\rangle - \frac{\|\lambda_{t+1} - \lambda^*\|^2}{2\rho}
\end{aligned} \tag{S-40}$$

Finally we combine Eqs. (S-40) and (S-35) so that we obtain

$$\begin{aligned}
&\begin{pmatrix} y^* - y_t \\ \lambda^* - \tilde{\lambda}_t \end{pmatrix}^\top \begin{pmatrix} -B^\top(\lambda_t - \tilde{\lambda}_t) \\ \frac{\tilde{\lambda}_t - \lambda_t}{\rho} \end{pmatrix} \\
\leq& \langle y^*, B^\top(\lambda_{t+1} - \lambda_t)\rangle - \langle y_t, B^\top(\lambda_{t+1} - \lambda_t)\rangle \\
&+ \frac{\rho}{2}\|y_t - y^*\|^2_{B^\top B} - \frac{\rho}{2}\|y_{t+1} - y^*\|^2_{B^\top B} + \frac{\rho}{2}\|y_{t+1} - y_t\|^2_{B^\top B}, \\
&- \frac{\rho}{2}\|y_{t+1} - y_t\|^2_{B^\top B} - \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho} + \frac{\|\lambda_t - \lambda^*\|^2}{2\rho} \\
&- \frac{\|\tilde{\lambda}_t - \lambda_{t+1}\|^2}{2\rho} - \langle B(y_{t+1} - y_t), \lambda_{t+1} - \lambda^*\rangle - \frac{\|\lambda_{t+1} - \lambda^*\|^2}{2\rho} \\
\leq& \frac{\rho}{2}\|y_t - y^*\|^2_{B^\top B} - \frac{\rho}{2}\|y_{t+1} - y^*\|^2_{B^\top B} + \frac{\|\lambda_t - \lambda^*\|^2}{2\rho} - \frac{\|\lambda_{t+1} - \lambda^*\|^2}{2\rho} \\
&+ \langle B(y^* - y_{t+1}), \lambda_{t+1} - \lambda^*\rangle - \langle B(y^* - y_t), \lambda_t - \lambda^*\rangle \\
&- \frac{\|\lambda_t - \lambda_{t+1}\|^2}{2\rho}.
\end{aligned}$$

This gives the assertion. □