Fast-and-Light Stochastic ADMM

Appendix

.1 Proof of Proposition 1

Proof. Let $\phi_{i_t} = (\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})) - (\nabla f(x_{t-1}) - \nabla f(\tilde{x}))$. We have

$$
\begin{aligned}
\mathbb{E}\left\|\frac{1}{b}\sum_{i_t\in\mathcal{I}_t}\phi_{i_t}\right\|^2 &= \frac{1}{b^2}\mathbb{E}\sum_{i_t,i_{t'}\in\mathcal{I}_t}\phi_{i_t}^T\phi_{i_{t'}} \\
&= \frac{1}{b^2}\mathbb{E}\sum_{i_t\neq i_{t'}\in\mathcal{I}_t}\phi_{i_t}^T\phi_{i_{t'}} + \frac{1}{b}\mathbb{E}\|\phi_i\|^2 \\
&= \frac{b-1}{bn(n-1)}\sum_{i\neq i'}\phi_i^T\phi_{i'} + \frac{1}{b}\mathbb{E}\|\phi_i\|^2 \\
&= \frac{b-1}{bn(n-1)}\sum_{i,i'}\phi_i^T\phi_{i'} - \frac{b-1}{b(n-1)}\mathbb{E}\|\phi_i\|^2 + \frac{1}{b}\mathbb{E}\|\phi_i\|^2 \\
&= \frac{n-b}{b(n-1)}\mathbb{E}\|\phi_i\|^2, \qquad (1)
\end{aligned}
$$

on using $\frac{1}{n}\sum_i \phi_i = 0$. Hence,

$$
\begin{aligned}
&\mathbb{E}\|\hat{\nabla} f(x_{t-1}) - \nabla f(x_{t-1})\|^2 \\
&= \mathbb{E}\left\|\frac{1}{b}\sum_{i_t\in\mathcal{I}}(\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})) - (\nabla f(x_{t-1}) - \nabla f(\tilde{x}))\right\|^2 \\
&= \frac{n-b}{b(n-1)}\mathbb{E}\|(\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})) - (\nabla f(x_{t-1}) - \nabla f(\tilde{x}))\|^2 \\
&= \frac{n-b}{b(n-1)}\left(\mathbb{E}\|\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})\|^2 - \|\nabla f(x_{t-1}) - \nabla f(\tilde{x})\|^2\right) \\
&\leq \frac{n-b}{b(n-1)}\mathbb{E}\|\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})\|^2 \\
&\leq \frac{2(n-b)}{b(n-1)}\mathbb{E}\|\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(x_*)\|^2 + \frac{2(n-b)}{b(n-1)}\mathbb{E}\|\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x_*)\|^2 \\
&= \frac{2(n-b)}{b(n-1)}\sum_{i=1}^n\frac{1}{n}\|\nabla f_i(x_{t-1}) - \nabla f_i(x_*)\|^2 + \frac{2(n-b)}{b(n-1)}\sum_{i=1}^n\frac{1}{n}\|\nabla f_i(\tilde{x}) - \nabla f_i(x_*)\|^2 \\
&\leq \frac{4L_{\max}(n-b)}{b(n-1)}(f(x_{t-1}) - f(x_*) + f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(x_{t-1} + \tilde{x} - 2x_*)).
\end{aligned}
$$

In the second equality, we use (1). In the third equality, we use $\mathbb{E}\|x_i - \mathbb{E}x_i\|^2 = \mathbb{E}\|x_i\|^2 - \|\mathbb{E}x_i\|^2$. In the second inequality, we use $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. In the last inequality, we employ the following fact [Xiao and Zhang, 2014]: $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(x_*)\|_2^2 \leq 2L_{\max}\left(f(x) - f(x_*) - \nabla f(x_*)^T(x - x_*)\right)$. □

.2 Proof of Theorem 1

First, we introduce the following Lemma.

Lemma 1. $u_* = -\frac{1}{\rho}(A^T)^\dagger\nabla f(x_*)$.

Proof. Consider (4) as a linear system $A^T u = -\frac{1}{\rho}\nabla f(x_*)$ for a random variable u. By [James, 1978], the solutions are given by

$$
U = \left\{u | u = -\frac{1}{\rho}(A^T)^\dagger\nabla f(x_*) + (I - (AA^\dagger)^T)v, v \in \mathbb{R}^l\right\},
$$

and solutions exist iff $(A^\dagger A)^T \nabla f(x_*) = \nabla f(x_*)$. Since u_* exists and $u_* \in U$, then $(A^\dagger A)^T \nabla f(x_*) = \nabla f(x_*)$ holds. Obviously, $u = -\frac{1}{\rho}(A^T)^\dagger \nabla f(x_*) \in U$ with $v = 0$. If A has full row rank, $AA^\dagger = I$ and U has an unique element that $U = \{u | u = -\frac{1}{\rho}(A^T)^\dagger \nabla f(x_*)\}$. Hence, $u_* = -\frac{1}{\rho}(A^T)^\dagger \nabla f(x_*)$. □

Consider the objective in the x_t update of Algorithm 1:

$$\left(\frac{1}{b} \sum_{i_t \in \mathcal{I}_t} (\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})) + \nabla f(\tilde{x}) \right)^T x + \frac{\rho}{2} \|Ax + By_t - c + u_{t-1}\|^2 + \frac{\|x - x_{t-1}\|_G^2}{2\eta}.$$

On setting the derivative w.r.t. x at x_t to zero, we have

$$g_t + \frac{1}{\eta} G(x_t - x_{t-1}) = 0, \tag{2}$$

where

$$\begin{aligned} g_t &= v_t + q_t, \\ v_t &= \frac{1}{b} \sum_{i_t \in \mathcal{I}_t} (\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})) + \nabla f(\tilde{x}), \\ q_t &= \rho A^T (Ax_t + By_t - c + u_{t-1}). \end{aligned}$$

Thus, the update can be rewritten as

$$x_t = x_{t-1} - \eta G^{-1} g_t. \tag{3}$$

Let $\alpha_t = \rho(u_t - u_*)$. We first introduce the following Lemmas.

Lemma 2. *For $0 \le \eta \le \frac{1}{L_f}$, we have*

$$f(x) + q_t^T (x - x_t) \ge f(x_t) + g_t^T (x - x_{t-1}) + \frac{\eta}{2} \|g_t\|_{G^{-1}}^2 + (v_t - \nabla f(x_{t-1}))^T (x_t - x).$$

Proof.

$$\begin{aligned} & f(x) + q_t^T (x - x_t) \\ \ge\ & f(x_{t-1}) + \nabla f(x_{t-1})^T (x - x_{t-1}) + q_t^T (x - x_t) \\ \ge\ & f(x_t) - \nabla f(x_{t-1})^T (x_t - x_{t-1}) - \frac{L_f}{2} \|x_t - x_{t-1}\|^2 + \nabla f(x_{t-1})^T (x - x_{t-1}) + q_t^T (x - x_t) \\ \ge\ & f(x_t) - \nabla f(x_{t-1})^T (x_t - x_{t-1}) - \frac{L_f}{2} \|x_t - x_{t-1}\|_G^2 + \nabla f(x_{t-1})^T (x - x_{t-1}) + q_t^T (x - x_t) \\ =\ & f(x_t) - \nabla f(x_{t-1})^T (x_t - x_{t-1}) - \frac{L_f \eta^2}{2} \|g_t\|_{G^{-1}}^2 + \nabla f(x_{t-1})^T (x - x_{t-1}) + q_t^T (x - x_t). \end{aligned}$$

In the first inequality, we use the convexity of f. In the second inequality, we use the smoothness of f at x_{t-1} (Assumption 1). In the last inequality, we use the assumption that $G \succeq I$. In the last equality, we use (3).

Next, consider the sum of inner products on the R.H.S.,

$$\begin{aligned} & -\nabla f(x_{t-1})^T (x_t - x_{t-1}) + \nabla f(x_{t-1})^T (x - x_{t-1}) + q_t^T (x - x_t) \\ =\ & \nabla f(x_{t-1})^T (x - x_t) + (g_t - v_t)^T (x - x_t) \\ =\ & g_t^T (x - x_{t-1} + x_{t-1} - x_t) + (v_t - \nabla f(x_{t-1}))^T (x_t - x) \\ =\ & g_t^T (x - x_{t-1}) + \eta \|g_t\|_{G^{-1}}^2 + (v_t - \nabla f(x_{t-1}))^T (x_t - x). \end{aligned}$$

Combining the results, and with the assumption that $0 \le \eta \le \frac{1}{L_f}$, we obtain

$$\begin{aligned} & f(x) + q_t^T (x - x_t) \\ \ge\ & f(x_t) + g_t^T (x - x_{t-1}) + \frac{\eta}{2} (2 - L_f \eta) \|g_t\|_{G^{-1}}^2 + (v_t - \nabla f(x_{t-1}))^T (x_t - x) \\ \ge\ & f(x_t) + g_t^T (x - x_{t-1}) + \frac{\eta}{2} \|g_t\|_{G^{-1}}^2 + (v_t - \nabla f(x_{t-1}))^T (x_t - x). \end{aligned}$$

□

Lemma 3. $2\eta\mathbb{E}(g(y_t)-g(y_*)-g'(y_*)^T(y_t-y_*)-(B^T\alpha_t)^T(y_*-y_t)) \le \eta\rho\mathbb{E}(\|Ax_{t-1}+By_*-c\|^2-\|Ax_t+By_*-c\|^2+\|u_t-u_{t-1}\|^2)$.

Proof. We have

$$\begin{aligned}
g(y_t)-g(y_*) &\le g'(y_t)^T(y_t-y_*)\\
&= -(\rho B^T(Ax_{t-1}+By_t-c+u_{t-1}))^T(y_t-y_*)\\
&= -(\rho B^T u_t)^T(y_t-y_*)+(x_{t-1}-x_t)^T\rho A^T B(y_*-y_t)\\
&= -(\rho B^T u_t)^T(y_t-y_*)+\frac{\rho}{2}(\|Ax_{t-1}+By_*-c\|^2-\|Ax_t+By_*-c\|^2)\\
&\quad +\frac{\rho}{2}(\|Ax_t+By_t-c\|^2-\|Ax_{t-1}+By_t-c\|^2)\\
&\le -(\rho B^T u_t)^T(y_t-y_*)+\frac{\rho}{2}(\|Ax_{t-1}+By_*-c\|^2-\|Ax_t+By_*-c\|^2)\\
&\quad +\frac{\rho}{2}\|u_t-u_{t-1}\|^2.
\end{aligned}$$

In the first inequality, we use the convexity of g. In the first equality, we use the optimality condition in the y_t update in Algorithm 1, i.e., $g'(y_t)+\rho B^T(Ax_{t-1}+By_t-c+u_{t-1})=0$. In the second equality, we use the update equation of u_t in Algorithm 1. Result then follows by taking expectation, using the optimality condition in (4), and multiplying by 2η. $\square$

Lemma 4. $2\eta\mathbb{E}(-(Ax_t+By_t-c)^T\alpha_t)=\eta\rho\mathbb{E}(\|u_{t-1}-u_*\|^2-\|u_t-u_*\|^2-\|u_t-u_{t-1}\|^2)$.

Proof. Using the u_t update in Algorithm 1, we obtain

$$\begin{aligned}
-(Ax_t+By_t-c)^T\alpha_t &= \rho(u_{t-1}-u_t)^T(u_t-u_*)\\
&= \frac{\rho}{2}(\|u_{t-1}-u_*\|^2-\|u_t-u_*\|^2-\|u_t-u_{t-1}\|^2).
\end{aligned}$$

Result follows on taking expectation, and multiplying by 2η. $\square$

Proof. (of Theorem 1) Using (2) and x_t in (3), we have

$$\begin{aligned}
\|x_t-x_*\|_G^2 &= \|x_{t-1}-x_*\|_G^2-2\eta(x_{t-1}-x_*)^Tg_t+\eta^2\|g_t\|_{G^{-1}}^2\\
&\le \|x_{t-1}-x_*\|_G^2-2\eta(f(x_t)-f(x_*))\\
&\quad -2\eta(v_t-\nabla f(x_{t-1}))^T(x_t-x_*)+2\eta q_t^T(x_*-x_t),
\end{aligned} \tag{4}$$

where we apply Lemma 2 to obtain the inequality. Now, we bound the term $-2\eta(v_t-\nabla f(x_{t-1}))^T(x_t-x_*)$. Define the convex function

$$\psi_t(x)=\frac{\rho}{2}\|Ax+By_t-c+u_{t-1}\|^2+\frac{1}{2\eta}\|x-x_{t-1}\|_{G-I}^2,$$

and

$$\bar{x}=\mathrm{prox}_{\eta\psi_t}(x_{t-1}-\eta\nabla f(x_{t-1})), \tag{5}$$

where $\mathrm{prox}_{\eta r}(y)=\min_x \eta r(x)+\frac{1}{2}\|x-y\|^2$ is the proximal operator. Note that

$$x_t=\mathrm{prox}_{\eta\psi_t}(x_{t-1}-\eta v_t) \tag{6}$$

since

$$\begin{aligned}
x_t &= \arg\min_x v_t^T x+\frac{\rho}{2}\|Ax+By_t-c+u_{t-1}\|^2+\frac{\|x-x_{t-1}\|_G^2}{2\eta}\\
&= \arg\min_x \eta v_t^T x+\frac{\eta\rho}{2}\|Ax+By_t-c+u_{t-1}\|^2+\frac{\|x-x_{t-1}\|_{G-I}^2}{2}+\frac{\|x-x_{t-1}\|^2}{2}\\
&= \arg\min_x \eta\psi_t(x)+\frac{1}{2}\|x-(x_{t-1}-\eta v_t)\|^2.
\end{aligned}$$

Then, the $-2\eta(v_t-\nabla f(x_{t-1}))^T(x_t-x_*)$ term in (4) becomes

$$\begin{aligned}
&-2\eta(v_t-\nabla f(x_{t-1}))^T(x_t-x_*)\\
&= -2\eta(v_t-\nabla f(x_{t-1}))^T(x_t-\bar{x})-2\eta(v_t-\nabla f(x_{t-1}))^T(\bar{x}-x_*)\\
&\le 2\eta\|v_t-\nabla f(x_{t-1})\|\|x_t-\bar{x}\|-2\eta(v_t-\nabla f(x_{t-1}))^T(\bar{x}-x_*)\\
&\le 2\eta\|v_t-\nabla f(x_{t-1})\|\|(x_{t-1}-\eta v_t)-(x_{t-1}-\eta\nabla f(x_{t-1}))\|\\
&\quad -2\eta(v_t-\nabla f(x_{t-1}))^T(\bar{x}-x_*)\\
&= 2\eta^2\|v_t-\nabla f(x_{t-1})\|^2-2\eta(v_t-\nabla f(x_{t-1}))^T(\bar{x}-x_*),
\end{aligned}$$

where in the first inequality we use the Cauchy-Schwartz inequality. In the second inequality, we use (5), (6) and non-expansiveness of the proximal operator. By combining the above results, we have from (4)

$$\begin{aligned}&\|x_t - x_*\|_G^2 - 2\eta q_t^T(x_* - x_t)\\ &\quad\leq \quad \|x_{t-1} - x_*\|_G^2 - 2\eta(f(x_t) - f(x_*)) + 2\eta^2\|v_t - \nabla f(x_{t-1})\|^2 - 2\eta(v_t - \nabla f(x_{t-1}))^T(\bar{x} - x_*).\end{aligned}$$

Note that $\mathbb{E}v_t = \nabla f(x_{t-1})$. Taking expectation w.r.t. $\mathcal{I}_t$, we obtain

$$\begin{aligned}&\mathbb{E}(\|x_t - x_*\|_G^2 - 2\eta q_t^T(x_* - x_t))\\ &\quad\leq \quad \|x_{t-1} - x_*\|_G^2 - 2\eta(\mathbb{E}f(x_t) - f(x_*)) + 2\eta^2\mathbb{E}\|v_t - \nabla f(x_{t-1})\|^2\\ &\quad\leq \quad \|x_{t-1} - x_*\|_G^2 - 2\eta(\mathbb{E}f(x_t) - f(x_*))\\ &\qquad\quad +8L_{\max}\eta^2\beta(b)(f(x_{t-1}) + f(\tilde{x}) - 2f(x_*) - \nabla f(x_*)^T(x_{t-1} + \tilde{x} - 2x_*)),\end{aligned}$$

where in the second inequality we apply Proposition 1. Taking expectation over $\mathcal{I}_t$ for $t = 1, ..., m$ in the current stage and rearranging terms, we obtain

$$\begin{aligned}&2\eta\mathbb{E}(f(x_t) - f(x_*) - q_t^T(x_* - x_t))\\ &\quad\leq \quad \mathbb{E}\|x_{t-1} - x_*\|_G^2 - \mathbb{E}\|x_t - x_*\|_G^2 + 8L_{\max}\eta^2\beta(b)\mathbb{E}(f(x_{t-1}) - f(x_*) - \nabla f(x_*)^T(x_{t-1} - x_*))\\ &\qquad\quad +8L_{\max}\eta^2\beta(b)(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*)).\end{aligned}$$

By using the optimality condition $\nabla f(x_*) + \rho A^T u_* = 0$, $q_t = \rho A^T u_t$ and $\alpha_t = \rho(u_t - u_*)$, we obtain

$$\begin{aligned}&2\eta\mathbb{E}(f(x_t) - f(x_*) - q_t^T(x_* - x_t))\\ &\quad= \quad 2\eta\mathbb{E}(f(x_t) - f(x_*) - \nabla f(x_*)^T(x_t - x_*) - (\rho A^T u_*)^T(x_t - x_*) - (\rho A^T u_t)^T(x_* - x_t))\\ &\quad= \quad 2\eta\mathbb{E}(f(x_t) - f(x_*) - \nabla f(x_*)^T(x_t - x_*) - (A^T\alpha_t)^T(x_* - x_t)).\end{aligned}$$

Thus, we have

$$\begin{aligned}&2\eta\mathbb{E}(f(x_t) - f(x_*) - \nabla f(x_*)^T(x_t - x_*) - (A^T\alpha_t)^T(x_* - x_t))\\ &\quad\leq \quad \mathbb{E}\|x_{t-1} - x_*\|_G^2 - \mathbb{E}\|x_t - x_*\|_G^2 + 8L_{\max}\eta^2\beta(b)\mathbb{E}(f(x_{t-1}) - f(x_*) - \nabla f(x_*)^T(x_{t-1} - x_*))\\ &\qquad\quad +8L_{\max}\eta^2\beta(b)(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*)). \qquad (7)\end{aligned}$$

Summing from $t = 1, \ldots, m$, and using $2\eta(1 - 4L_{\max}\eta\beta(b)) \leq 2\eta$, and $x_0 = \tilde{x}$, we obtain

$$\begin{aligned}&2\eta(1 - 4L_{\max}\eta\beta(b))\sum_{k=1}^{m}\mathbb{E}(f(x_k) - f(x_*) - \nabla f(x_*)^T(x_k - x_*)) - 2\eta\mathbb{E}\sum_{k=1}^{m}(A^T\alpha_k)^T(x_* - x_k)\\ &\quad\leq \quad \|x_0 - x_*\|_G^2 - \mathbb{E}\|x_m - x_*\|_G^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*))\\ &\quad\leq \quad \|x_0 - x_*\|_G^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*)).\end{aligned}$$

By using convexity of f that $f(\frac{1}{m}\sum_{k=1}^m x_k) \leq \frac{1}{m}\sum_{k=1}^m f(x_k)$ and $\tilde{x}_s = \frac{1}{m}\sum_{k=1}^m x_k$, we have

$$\begin{aligned}&2\eta(1 - 4L_{\max}\eta\beta(b))m\mathbb{E}(f(\tilde{x}_s) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_s - x_*)) - 2\eta\mathbb{E}\sum_{k=1}^{m}(A^T\alpha_k)^T(x_* - x_k)\\ &\quad\leq \quad \|x_0 - x_*\|_G^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*))\\ &\quad= \quad \|\tilde{x}_{s-1} - x_*\|_G^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}_{s-1}) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_{s-1} - x_*)),\end{aligned}$$

where in the last inequality, we use that $x_0 = \tilde{x}_{s-1}$. Also we have

$$\begin{aligned}&-(A^T\alpha_t)^T(x_* - x_t) - (B^T\alpha_t)^T(y_* - y_t) - (Ax_t + By_t - c)^T\alpha_t\\ &\quad= \quad -(Ax_* + By_* - c)^T\alpha_t + (Ax_t - Ax_t + By_t - By_t)^T\alpha_t\\ &\quad= \quad 0.\end{aligned}$$

Thus, define $R(x, y) = f(x) - f(x_*) - \nabla f(x_*)^T(x - x_*) + g(y) - g(y_*) - g'(y_*)^T(y - y_*)$. By combining Lemma 3 and

Lemma 4, and $g(\frac{1}{m}\sum_{k=1}^{m} y_k) \leq \frac{1}{m}\sum_{k=1}^{m} g(y_k)$ and $\tilde{y}_s = \frac{1}{m}\sum_{k=1}^{m} y_k$, we obtain

$$\begin{aligned}
&2\eta(1-4L_{\max}\eta\beta(b))m\mathbb{E}R(\tilde{x}_s,\tilde{y}_s)\\
\leq\;& \|\tilde{x}_{s-1}-x_*\|_G^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\
&+\eta\rho\|A\tilde{x}_{s-1}+By_*-c\|^2+\eta\rho\|\tilde{u}_{s-1}-u_*\|^2\\
=\;& \|\tilde{x}_{s-1}-x_*\|_G^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\
&+\eta\rho\|A\tilde{x}_{s-1}-Ax_*\|^2+\eta\rho\|\tilde{u}_{s-1}-u_*\|^2\\
=\;& \|\tilde{x}_{s-1}-x_*\|_{G+\eta\rho A^TA}^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\
&+\eta\rho\|\tilde{u}_{s-1}-u_*\|^2\\
\leq\;& \|G+\eta\rho A^TA\|\|\tilde{x}_{s-1}-x_*\|^2 + 8L_{\max}\eta^2(m+1)\beta(b)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\
&+\eta\rho\|\tilde{u}_{s-1}-u_*\|^2\\
\leq\;& \left(\frac{2\|G+\eta\rho A^TA\|}{\lambda_f}+8L_{\max}\eta^2(m+1)\beta(b)\right)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\
&+\eta\rho\|\tilde{u}_{s-1}-u_*\|^2\\
\leq\;& \left(\frac{2\|G+\eta\rho A^TA\|}{\lambda_f}+8L_{\max}\eta^2(m+1)\beta(b)\right)(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\
&+\left(\frac{2\|G+\eta\rho A^TA\|}{\lambda_f}+8L_{\max}\eta^2(m+1)\beta(b)\right)(g(\tilde{y}_{s-1})-g(y_*)-g'(y_*)^T(\tilde{y}_{s-1}-y_*))\\
&+\eta\rho\|\tilde{u}_{s-1}-u_*\|^2\\
=\;& \left(\frac{2\|G+\eta\rho A^TA\|}{\lambda_f}+8L_{\max}\eta^2(m+1)\beta(b)\right)R(\tilde{x}_{s-1},\tilde{y}_{s-1})+\eta\rho\|\tilde{u}_{s-1}-u_*\|^2.
\end{aligned}$$

In the first equality, we use that $Ax_*+By_*=c$. In the last inequality, we use the convexity of g so that $g(\tilde{y}_{s-1})-g(y_*)-g'(y_*)^T(\tilde{y}_{s-1}-y_*)$ is non-negative. We now turn to bound $\|\tilde{u}_{s-1}-u_*\|^2$. Since we assume that A has full row rank, by Lemma 1, we have $u_*=-\frac{1}{\rho}(A^T)^\dagger\nabla f(x_*)$. By using the update rule $\tilde{u}_{s-1}=-\frac{1}{\rho}(A^T)^\dagger\nabla f(\tilde{x}_{s-1})$, we obtain

$$\begin{aligned}
\|\tilde{u}_{s-1}-u_*\|^2 &= \frac{1}{\rho^2}\|\nabla f(\tilde{x}_{s-1})-\nabla f(x_*)\|^2_{A^\dagger(A^\dagger)^T}\\
&\leq \frac{2L_f\|A^\dagger(A^\dagger)^T\|}{\rho^2}(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*))\\
&= \frac{2L_f}{\rho^2\sigma_{\min}(AA^T)}(f(\tilde{x}_{s-1})-f(x_*)-\nabla f(x_*)^T(\tilde{x}_{s-1}-x_*)).
\end{aligned}$$

Thus, combining the results, we have

$$\begin{aligned}
&2\eta(1-4L_{\max}\eta\beta(b))m\mathbb{E}R(\tilde{x}_s,\tilde{y}_s)\\
\leq\;& \left(\frac{2\|G+\eta\rho A^TA\|}{\lambda_f}+8L_{\max}\eta^2(m+1)\beta(b)+\frac{2L_f\eta}{\rho\sigma_{\min}(AA^T)}\right)R(\tilde{x}_{s-1},\tilde{y}_{s-1}).
\end{aligned}$$

Let $\kappa=\frac{\|G+\eta\rho A^TA\|}{\lambda_f\eta(1-4L_{\max}\eta\beta(b))m}+\frac{4L_{\max}\eta\beta(b)(m+1)}{(1-4L_{\max}\eta\beta(b))m}+\frac{L_f}{\rho(1-4L_{\max}\eta\beta(b))\sigma_{\min}(AA^T)m}$, we have

$$\mathbb{E}R(\tilde{x}_s,\tilde{y}_s) \leq \kappa R(\tilde{x}_{s-1},\tilde{y}_{s-1}).$$

Thus, we obtain

$$\mathbb{E}R(\tilde{x}_s,\tilde{y}_s)\leq\kappa^s R(\tilde{x}_0,\tilde{y}_0)$$

which completes the proof. □

.3 Proof of Theorem 2

Firstly, we introduce a variant of Lemma 4

Lemma 5. *For any $\alpha=\rho u$, $2\eta\mathbb{E}(-(Ax_t+By_t-c)^T(\alpha_t-\alpha))=\eta\rho\mathbb{E}(\|u_{t-1}-u_*-u\|^2-\|u_t-u_*-u\|^2-\|u_t-u_{t-1}\|^2)$.*

Proof. By using $Ax_t + By_t - c = u_t - u_{t-1}$, we obtain

$$\begin{aligned} -(Ax_t + By_t - c)^T(\alpha_t - \alpha) &= \rho(u_{t-1} - u_t)^T(u_t - u_* - u) \\ &= \frac{\rho}{2}(\|u_{t-1} - u_* - u\|^2 - \|u_t - u_* - u\|^2 - \|u_t - u_{t-1}\|^2). \end{aligned}$$

Result follows on taking expectation, and multiplying by 2η. $\square$

Proof. (of Theorem 2) Recall (7),

$$\begin{aligned} &2\eta\mathbb{E}(f(x_t) - f(x_*) - \nabla f(x_*)^T(x_t - x_*) - (A^T\alpha_t)^T(x_* - x_t)) \\ \leq\ & \mathbb{E}\|x_{t-1} - x_*\|_G^2 - \mathbb{E}\|x_t - x_*\|_G^2 + 8L_{\max}\eta^2\beta(b)\mathbb{E}(f(x_{t-1}) - f(x_*) - \nabla f(x_*)^T(x_{t-1} - x_*)) \\ & +8L_{\max}\eta^2\beta(b)(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*)). \end{aligned}$$

By summing over $t = 1, \ldots, m$, we obtain

$$\begin{aligned} & 2\eta(1 - 4L_{\max}\eta\beta(b))\sum_{k=1}^m \mathbb{E}(f(x_k) - f(x_*) - \nabla f(x_*)^T(x_k - x_*)) - 2\eta\mathbb{E}\sum_{k=1}^m (A^T\alpha_k)^T(x_* - x_k) \\ \leq\ & 8L_{\max}\eta^2\beta(b)(f(x_0) - f(x_*) - \nabla f(x_*)^T(x_0 - x_*)) + \|x_0 - x_*\|_G^2 \\ & -\mathbb{E}\left(8L_{\max}\eta^2\beta(b)(f(x_m) - f(x_*) - \nabla f(x_*)^T(x_m - x_*)) + \|x_m - x_*\|_G^2\right) \\ & +8L_{\max}\eta^2 m\beta(b)(f(\tilde{x}) - f(x_*) - \nabla f(x_*)^T(\tilde{x} - x_*)). \end{aligned}$$

By using convexity of f, and $\hat{x}_s = x_m$, $\tilde{x}_s = \frac{1}{m}\sum_{k=1}^m x_k$ and $\tilde{x} = \tilde{x}_{s-1}$, and taking expectation over whole history, we have

$$\begin{aligned} & 2\eta(1 - 4L_{\max}\eta\beta(b))m\mathbb{E}(f(\tilde{x}_s) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_s - x_*)) - 2\eta\mathbb{E}\sum_{k=1}^m (A^T\alpha_k)^T(x_* - x_k) \\ \leq\ & \mathbb{E}\left(8L_{\max}\eta^2\beta(b)(f(\hat{x}_{s-1}) - f(x_*) - \nabla f(x_*)^T(\hat{x}_{s-1} - x_*)) + \|\hat{x}_{s-1} - x_*\|_G^2\right) \\ & -\mathbb{E}\left(8L_{\max}\eta^2\beta(b)(f(\hat{x}_s) - f(x_*) - \nabla f(x_*)^T(\hat{x}_s - x_*)) + \|\hat{x}_s - x_*\|_G^2\right) \\ & +\mathbb{E}\left(8L_{\max}\eta^2 m\beta(b)(f(\tilde{x}_{s-1}) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_{s-1} - x_*))\right). \end{aligned}$$

Define sequence $T_k = \|\hat{x}_k - x_*\|_G^2 + 8L_{\max}\eta^2\beta(b)(f(\hat{x}_k) - f(x_*) - \nabla f(x_*)^T(\hat{x}_k - x_*)) + 8L_{\max}\eta^2 m\beta(b)(f(\tilde{x}_k) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_k - x_*))$. By subtracting $8L_{\max}\eta^2 m\beta(b)(f(\tilde{x}_s) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_s - x_*))$ from both sides, we have

$$\begin{aligned} & 2\eta(1 - 8L_{\max}\eta\beta(b))m\mathbb{E}(f(\tilde{x}_s) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_s - x_*)) - 2\eta\sum_{k=1}^m (A^T\alpha_k)^T(x_* - x_k) \\ \leq\ & \mathbb{E}\left(8L_{\max}\eta^2\beta(b)(f(\hat{x}_{s-1}) - f(x_*) - \nabla f(x_*)^T(\hat{x}_{s-1} - x_*)) + \|\hat{x}_{s-1} - x_*\|_G^2\right) \\ & -\mathbb{E}\left(8L_{\max}\eta^2\beta(b)(f(\hat{x}_s) - f(x_*) - \nabla f(x_*)^T(\hat{x}_s - x_*)) + \|\hat{x}_s - x_*\|_G^2\right) \\ & +\mathbb{E}\left(8L_{\max}\eta^2 m\beta(b)(f(\tilde{x}_{s-1}) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_{s-1} - x_*))\right) \\ & -\mathbb{E}\left(8L_{\max}\eta^2 m\beta(b)(f(\tilde{x}_s) - f(x_*) - \nabla f(x_*)^T(\tilde{x}_s - x_*))\right) \\ =\ & T_{s-1} - T_s. \end{aligned}$$

Also, we have

$$\begin{aligned} & -(A^T\alpha_t)^T(x_* - x_t) - (B^T\alpha_t)^T(y_* - y_t) - (Ax_t + By_t - c)^T(\alpha_t - \alpha) \\ =\ & -(Ax_* + By_* - c)^T\alpha_t + (Ax_t - Ax_t + By_t - By_t)^T\alpha_t + (Ax_t + By_t - c)^T\alpha \\ =\ & (Ax_t + By_t - c)^T\alpha. \end{aligned}$$

By combining Lemma 3 and Lemma 5, and $\tilde{y} = \frac{1}{m}\sum_{k=1}^m y_k$, $\hat{y} = y_m$, $\hat{u} = u_m$, $2\eta(1 - 8L_{\max}\eta\beta(b)) \leq 2\eta$, $\hat{x}_0 = \tilde{x}_0$, and summing over all stages, we have

$$
\begin{aligned}
& 2\eta(1-8L_{\max}\eta\beta(b))m\sum_{k=1}^{s}\mathbb{E}R(\tilde{x}_k,\tilde{y}_k)+2\eta m\sum_{k=1}^{s}\mathbb{E}(A\tilde{x}_k+B\tilde{y}_k-c)^T\alpha \\
\leq\ & 8L_{\max}\eta^2\beta(b)(f(\hat{x}_0)-f(x_*)-\nabla f(x_*)^T(\hat{x}_0-x_*))+\|\hat{x}_0-x_*\|_G^2 \\
& +8L_{\max}\eta^2\beta(b)m(f(\tilde{x}_0)-f(x_*)-\nabla f(x_*)^T(\tilde{x}_0-x_*))+\eta\rho\|A\hat{x}_0+By_*-c\|^2 \\
& +\eta\rho\|\hat{u}_0-u_*-u\|^2 \\
=\ & 8L_{\max}\eta^2\beta(b)(f(\hat{x}_0)-f(x_*)-\nabla f(x_*)^T(\hat{x}_0-x_*))+\|\hat{x}_0-x_*\|_G^2 \\
& +8L_{\max}\eta^2\beta(b)m(f(\hat{x}_0)-f(x_*)-\nabla f(x_*)^T(\hat{x}_0-x_*))+\eta\rho\|A\hat{x}_0-Ax_*\|^2 \\
& +\eta\rho\|\hat{u}_0-u_*-u\|^2 \\
=\ & 8L_{\max}\eta^2\beta(b)(m+1)(f(\hat{x}_0)-f(x_*)-\nabla f(x_*)^T(\hat{x}_0-x_*))+\|\hat{x}_0-x_*\|_{G+\eta\rho A^TA}^2 \\
& +\eta\rho\|\hat{u}_0-u_*-u\|^2.
\end{aligned}
$$

With convexity of f and g, and $\bar{x}=\frac{1}{s}\sum_{k=1}^{s}\tilde{x}_k$, $\bar{y}=\frac{1}{s}\sum_{k=1}^{s}\tilde{y}_k$, and set $\alpha=\zeta\frac{A\bar{x}+b\bar{y}-c}{\|A\bar{x}+b\bar{y}-c\|}$ with any $\zeta>0$, we have

$$
\begin{aligned}
& \mathbb{E}\left(R(\bar{x},\bar{y})+\zeta\|A\bar{x}+b\bar{y}-c\|\right) \\
\leq\ & \frac{4L_{\max}\eta\beta(b)(m+1)}{(1-8L_{\max}\eta\beta(b))ms}(f(\hat{x}_0)-f(x_*)-\nabla f(x_*)^T(\hat{x}_0-x_*)) \\
& +\frac{1}{2\eta(1-8L_{\max}\eta\beta(b))ms}\|\hat{x}_0-x_*\|_{G+\eta\rho A^TA}^2+\frac{\rho}{2(1-8L_{\max}\eta\beta(b))ms}\|\hat{u}_0-u_*-u\|^2 \\
\leq\ & \frac{4L_{\max}\eta\beta(b)(m+1)}{(1-8L_{\max}\eta\beta(b))ms}(f(\hat{x}_0)-f(x_*)-\nabla f(x_*)^T(\hat{x}_0-x_*)) \\
& +\frac{1}{2\eta(1-8L_{\max}\eta\beta(b))ms}\|\hat{x}_0-x_*\|_{G+\eta\rho A^TA}^2+\frac{\rho}{(1-8L_{\max}\eta\beta(b))ms}(\|\hat{u}_0-u_*\|^2+\|u\|^2) \\
=\ & \frac{4L_{\max}\eta\beta(b)(m+1)}{(1-8L_{\max}\eta\beta(b))ms}(f(\hat{x}_0)-f(x_*)-\nabla f(x_*)^T(\hat{x}_0-x_*)) \\
& +\frac{1}{2\eta(1-8L_{\max}\eta\beta(b))ms}\|\hat{x}_0-x_*\|_{G+\eta\rho A^TA}^2+\frac{\rho}{(1-8L_{\max}\eta\beta(b))ms}(\|\hat{u}_0-u_*\|^2+\frac{\zeta^2}{\rho^2}).
\end{aligned}
$$

□

.4 Proof of Corollary 1

Theorem 1 and Markov's inequality imply

$$\mathrm{Prob}(R(\tilde{x}_s,\tilde{y}_s)\geq\epsilon)\leq\frac{\mathbb{E}R(\tilde{x}_s,\tilde{y}_s)}{\epsilon}\leq\frac{\kappa^sR(\tilde{x}_0,\tilde{y}_0)}{\epsilon}.$$

Result follows on setting $\frac{\kappa^sR(\tilde{x}_0,\tilde{y}_0)}{\epsilon}\leq\delta$ and taking logarithm on both sides.

.5 Proof of Proposition 3

With $G=\gamma I-\eta\rho A^TA$ and $\gamma=\gamma_{\min}$, we have

$$\kappa=\frac{\eta\rho\|A^TA\|+1}{\lambda_f\eta(1-4L_{\max}\eta\beta(b))m}+\frac{4L_{\max}\eta\beta(b)(m+1)}{(1-4L_{\max}\eta\beta(b))m}+\frac{L_f}{\rho(1-4L_{\max}\eta\beta(b))\sigma_{\min}(AA^T)m}.$$

It can be shown that κ is convex w.r.t. $\rho>0$. Hence, by simple differentiation, choosing $\rho=\rho_*$, minimizes κ.

References

[James, 1978] M James. The generalised inverse. *The Mathematical Gazette*, pages 109–114, 1978.

[Xiao and Zhang, 2014] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4), 2014.