Towards optimal stochastic alternating direction method of multipliers: Supplementary material

Samaneh Azadi SAZADI156@GMAIL.COM

UC Berkeley, Berkeley, CA
School of ECE, Shiraz University, Shiraz, Iran

Suvrit Sra SUVRIT@TUEBINGEN.MPG.DE

Carnegie Mellon University, Pittsburgh
Max Planck Institute for Intelligent Systems, Tübingen, Germany

1. The strongly convex case

1.1. Proof of Lemma 1

Lemma 1. *Let f be μ-strongly convex, and let x_{k+1}, y_{k+1} and λ_{k+1} be computed as per Alg. 2. For all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and $w \in \Omega$, it holds for $k \geq 0$ that*

$$\begin{aligned} f(x_k) - f(x) + h(y_{k+1}) - h(y) + \langle w_{k+1} - w,\, F(w_{k+1})\rangle \\ \leq \tfrac{\eta_k}{2}\|g_k\|_2^2 - \tfrac{\mu}{2}\Delta_k + \tfrac{1}{2\eta_k}[\Delta_k - \Delta_{k+1}] + \tfrac{\beta}{2}[\mathcal{A}_k - \mathcal{A}_{k+1}] \\ + \tfrac{1}{2\beta}[\mathcal{L}_k - \mathcal{L}_{k+1}] + \langle \delta_k,\, x_k - x\rangle. \end{aligned} \tag{1}$$

By the strong convexity of f, we have

$$f(x_k) - f(x) \leq \langle f'(x_k),\, x_k - x\rangle - \tfrac{\mu}{2}\|x_k - x\|_2^2. \tag{2}$$

As before, using $\delta_k = f'(x_k) - g_k$; but this time we split the $f'(x_k)$ term differently:

$$\langle f'(x_k),\, x_k - x\rangle = \langle g_k,\, x_{k+1} - x\rangle + \langle \delta_k,\, x_k - x\rangle + \langle g_k,\, x_k - x_{k+1}\rangle. \tag{3}$$

Now for the first part, we just follow the derivation of (Ouyang et al., 2013), before it comes to the critical difference, namely inequality (9). However, for the reader's convenience we include all the details below.

From the optimality condition of Line 2, it follows that

$$\langle g_k + \beta A^T(Ax_{k+1} + By_k - b) - A^T\lambda_k + \eta_k^{-1}(x_{k+1} - x_k),\, x - x_{k+1}\rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

Rearranging this inequality, we obtain

$$\langle g_k,\, x_{k+1} - x\rangle \leq \langle \beta A^T(Ax_{k+1} + By_k - b) - A^T\lambda_k,\, x - x_{k+1}\rangle + \tfrac{1}{\eta_k}\langle x_{k+1} - x_k,\, x - x_{k+1}\rangle,$$

so that a rearrangement similar to (20) yields

$$\begin{aligned} \langle g_k,\, x_{k+1} - x\rangle \leq \langle \beta A^T(Ax_{k+1} + By_k - b) - A^T\lambda_k,\, x - x_{k+1}\rangle \\ + \tfrac{1}{2\eta_k}\left[\|x - x_k\|_2^2 - \|x - x_{k+1}\|_2^2 - \|x_{k+1} - x_k\|_2^2\right]. \end{aligned}$$

Combining this inequality with (2) and (3) we then obtain

$$f(x_k) - f(x) \le \langle \beta A^T(Ax_{k+1} + By_k - b) - A^T\lambda_k,\, x - x_{k+1}\rangle \\ + \tfrac{1}{2\eta_k}\left[\|x - x_k\|_2^2 - \|x - x_{k+1}\|_2^2 - \|x_{k+1} - x_k\|_2^2\right] + \langle \delta_k,\, x_k - x\rangle + \langle g_k,\, x_k - x_{k+1}\rangle - \tfrac{\mu}{2}\|x_k - x\|_2^2.$$

As before, adding and subtracting By_{k+1} allows us to rewrite the AL term as

$$\langle \beta A^T(Ax_{k+1} + By_k - b) - A^T\lambda_k,\, x - x_{k+1}\rangle = \langle -A^T\lambda_{k+1} + \beta A^T B(y_k - y_{k+1}),\, x - x_{k+1}\rangle,$$

which upon manipulations similar to those followed for (22) lead to

$$\langle \beta A^T B(y_k - y_{k+1}),\, x - x_{k+1}\rangle = \beta\langle Ax - Ax_{k+1},\, By_k - By_{k+1}\rangle \\ = \tfrac{\beta}{2}\left[\left(\|Ax + By_k - b\|_2^2 - \|Ax + By_{k+1} - b\|_2^2\right) + \left(\|Ax_{k+1} + By_{k+1} - b\|_2^2 - \|Ax_{k+1} + By_k - b\|_2^2\right)\right] \\ \le \tfrac{\beta}{2}\left(\|Ax + By_k - b\|_2^2 - \|Ax + By_{k+1} - b\|_2^2\right) + \tfrac{1}{2\beta}\|\lambda_{k+1} - \lambda_k\|_2^2.$$

Thus, we obtain the following inequality

$$\begin{aligned} &f(x_k) - f(x) + \langle -A^T\lambda_{k+1},\, x_{k+1} - x\rangle \\ &\le \tfrac{1}{2\eta_k}\left[\|x - x_k\|_2^2 - \|x - x_{k+1}\|_2^2 - \|x_{k+1} - x_k\|_2^2\right] \\ &+ \langle \delta_k,\, x_k - x\rangle + \langle g_k,\, x_k - x_{k+1}\rangle - \tfrac{\mu}{2}\|x_k - x\|_2^2 \\ &+ \tfrac{\beta}{2}\left(\|Ax + By_k - b\|_2^2 - \|Ax + By_{k+1} - b\|_2^2\right) + \tfrac{1}{2\beta}\|\lambda_{k+1} - \lambda_k\|_2^2. \end{aligned} \tag{4}$$

To enable cancellation of $\|x_k - x_{k+1}\|_2^2$, we bound $\langle g_k,\, x_k - x_{k+1}\rangle$ using Young's inequality

$$\langle g_k,\, x_k - x_{k+1}\rangle \le \tfrac{\eta_k}{2}\|g_k\|_2^2 + \tfrac{1}{2\eta_k}\|x_k - x_{k+1}\|_2^2. \tag{5}$$

The y terms can be bounded akin to (29) and (30) to obtain

$$\begin{aligned} &h(y_{k+1}) - h(y) \le \langle y_{k+1} - y,\, B^T\lambda_{k+1}\rangle, \\ &\langle \lambda_{k+1} - \lambda,\, Az_{k+1} + By_{k+1} - b\rangle = \tfrac{1}{2\beta}\left[\|\lambda - \lambda_k\|_2^2 - \|\lambda - \lambda_{k+1}\|_2^2 - \|\lambda_{k+1} - \lambda_k\|_2^2\right]. \end{aligned} \tag{6}$$

Adding inequalities (4)–(6) we obtain the overall bound (with Δ_k, $\mathcal{A}_k$, and $\mathcal{L}_k$ redefined with x_k)

$$\begin{aligned} f(x_k) - f(x) + h(y_{k+1}) - h(y) + \langle w_{k+1} - w,\, F(w_{k+1})\rangle &\le \tfrac{\eta_k}{2}\|g_k\|_2^2 - \tfrac{\mu}{2}\Delta_k \\ + \tfrac{1}{2\eta_k}[\Delta_k - \Delta_{k+1}] + \tfrac{\beta}{2}[\mathcal{A}_k - \mathcal{A}_{k+1}] + \tfrac{1}{2\beta}[\mathcal{L}_k - \mathcal{L}_{k+1}] + \langle \delta_k,\, x_k - x\rangle. \end{aligned} \tag{7}$$

1.2. Proof of theorem 2

Theorem 2. *Let f be μ-strongly convex. Let $\eta_k = \frac{2}{\mu(k+2)}$, let x, y_j, λ_j be generated by Alg. 2, and $\bar{x}_k, \bar{y}_k, \bar{\lambda}_k$ computed by (23). Let x^*, y^* be the optimal; then for $k \ge 1$,*

$$\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{x}_k + B\bar{y}_k - b\|_2] \\ \le \frac{2G^2}{\mu(k+1)} + \frac{\beta}{2(k+1)}D_{\mathcal{Y}}^2 + \frac{2\rho^2}{\beta(k+1)}.$$

Up to (7), our analysis has paralleled the one in (Ouyang et al., 2013). But now come two crucial differences: (i) instead of using uniform averages of past iterates we use weighted averages; and instead of stepsizes $\eta_k = 1/(\mu k)$ we use $\eta_k = 2/(\mu(k+1))$. This change is key to making our complexity bound optimal, though under slightly stronger assumptions on $\|B(y_k - y^*)\|_2$ and $\|\lambda_k - \lambda^*\|_2$ than (Ouyang et al., 2013) (namely, boundedness at each iteration k, rather than just at $k = 0$). But this added assumption is the price one has to often pay for acceleration algorithm (Chambolle & Pock, 2011; Goldfarb et al., 2012).

Recall the definitions

$$\bar{x}_k := \tfrac{2}{k(k+1)}\sum_{j=0}^{k-1}(j+1)x_j, \quad \bar{y}_k := \tfrac{2}{k(k+1)}\sum_{j=1}^{k} jy_j, \quad \bar{\lambda}_k := \tfrac{2}{k(k+1)}\sum_{j=1}^{k} j\lambda_j. \tag{8}$$

With definitions (8), as for inequality (41), we obtain

$$\begin{aligned} &f(\bar{x}_k) - f(x) + h(\bar{y}_k) - h(y) + \langle \bar{w}_k - w,\, F(\bar{w}_k)\rangle \\ &\leq \tfrac{2}{k(k+1)} \sum_{j=0}^{k-1} (j+1)[f(x_j) - f(x)] + \tfrac{2}{k(k+1)} \sum\nolimits_{j=1}^{k} j[h(y_j) - h(y) + \langle w_j - w,\, F(w_j)\rangle] \\ &= \tfrac{2}{k(k+1)} \sum_{j=0}^{k-1} (j+1)[f(x_j) - f(x) + h(y_{j+1}) - h(y) + \langle w_{j+1} - w,\, F(w_{j+1})\rangle]. \end{aligned} \tag{9}$$

To this weighted inequality, now apply inequality (7). This yields

$$\begin{aligned} &f(\bar{x}_k) - f(x) + h(\bar{y}_k) - h(y) + \langle \bar{w}_k - w,\, F(\bar{w}_k)\rangle \\ &\leq \tfrac{2}{k(k+1)} \sum_{j=0}^{k-1} (j+1) \Big[\tfrac{\eta_k}{2} \|g_k\|_2^2 - \tfrac{\mu}{2} \Delta_k + \tfrac{1}{2\eta_j} [\Delta_j - \Delta_{j+1}] + \tfrac{\beta}{2} [\mathcal{A}_j - \mathcal{A}_{j+1}] \\ &\qquad + \tfrac{1}{2\beta} [\mathcal{L}_j - \mathcal{L}_{j+1}] + \langle \delta_j,\, x_j - x\rangle \Big]. \end{aligned} \tag{10}$$

Let us bound (10) by considering contributions from the different terms one by one.

$$\begin{aligned} \sum\nolimits_{j=0}^{k-1} \tfrac{\beta(j+1)}{2} [\mathcal{A}_j - \mathcal{A}_{j+1}] &\leq \tfrac{\beta}{2} \sum\nolimits_{j=0}^{k-1} \mathcal{A}_j \\ \sum\nolimits_{j=0}^{k-1} \tfrac{j+1}{2\beta} [\mathcal{L}_j - \mathcal{L}_{j+1}] &\leq \tfrac{1}{2\beta} \sum\nolimits_{j=0}^{k-1} \mathcal{L}_j. \end{aligned} \tag{11}$$

The Δ_j terms require slightly more work. We use $\eta_j = \frac{2}{\mu(j+2)}$ for $j \geq 0$. Thus, we have

$$\begin{aligned} &\sum\nolimits_{j=0}^{k-1} (j+1) \left(\tfrac{1}{2\eta_j} [\Delta_j - \Delta_{j+1}] - \tfrac{\mu}{2} \Delta_j \right) = \sum\nolimits_{j=0}^{k-1} (j+1) \left(\tfrac{\mu(j+2)}{4} \Delta_j - \tfrac{\mu}{2} \Delta_j - \tfrac{\mu(j+2)}{4} \Delta_{j+1} \right) \\ &= \tfrac{\mu}{4} \sum\nolimits_{j=0}^{k-1} (j+1)[j\Delta_j - (j+2)\Delta_{j+1}] \\ &= \tfrac{\mu}{4} \left[1(0 - 2\Delta_1) + 2(\Delta_1 - 3\Delta_2) + \cdots + k((k-1)\Delta_{k-1} - (k+1)\Delta_k) \right] \\ &= \tfrac{\mu}{2} [0 - k(k+1)\Delta_k] \leq 0. \end{aligned} \tag{12}$$

Using (11) and (12) in (10), and taking expectations we obtain

$$\begin{aligned} &\mathbb{E}[f(\bar{x}_k) - f(x) + h(\bar{y}_k) - h(y) + \langle \bar{w}_k - w,\, F(\bar{w}_k)\rangle] \\ &\leq \tfrac{2}{k(k+1)} \sum\nolimits_{j=0}^{k-1} \left(\tfrac{\mathbb{E}[\|g_k\|_2^2]}{\mu} + \tfrac{\beta}{2} \mathcal{A}_j + \tfrac{1}{2\beta} \mathcal{L}_j \right), \end{aligned}$$

where as before $\mathbb{E}[\langle \delta_j,\, x_j - x\rangle] = 0$. Using our assumption $\mathbb{E}[\|g_k\|_2^2] \leq G^2$, and following the same arguments as in (42) we finally obtain

$$\begin{aligned} &\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho \|A\bar{x}_k + B\bar{y}_k - b\|_2] \\ &\leq \frac{2G^2}{\mu(k+1)} + \frac{\beta}{2(k+1)} D_{\mathcal{Y}}^2 + \frac{2\rho^2}{\beta(k+1)}. \end{aligned} \tag{13}$$

2. The smooth case

We begin with a classic result. Let f be a differentiable function with an L-Lipschitz continuous gradient. Then,

$$f(y) \leq f(x) + \langle \nabla f(x),\, y - x\rangle + \tfrac{L}{2} \|y - x\|_2^2. \tag{14}$$

2.1. Proof of Lemma 3

Lemma 3. *Let $x_{k+1}, y_{k+1}, z_{k+1}$ be generated by Alg. 3. For $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $w \in \Omega$, and with $\eta_k = (L + \alpha_k)^{-1}$ the following bound holds for all $k \geq 0$:*

$$
\begin{aligned}
&f(x_{k+1}) + \theta_k[h(y_{k+1}) - h(y)] + \theta_k\langle w_{k+1} - w,\, F(w_{k+1})\rangle \\
&\leq (1-\gamma_k)f(x_k) + \gamma_k f(x) + \tfrac{\gamma_k^2}{2\eta_k}[\Delta_k - \Delta_{k+1}] + \tfrac{1}{2\alpha_k}\|\delta_k\|_2^2 + \gamma_k\langle \delta_k,\, z_k - x\rangle + \tfrac{\beta\gamma_k}{2}[\mathcal{A}_k - \mathcal{A}_{k+1}] + \tfrac{\gamma_k}{2\beta}\left[\mathcal{L}_k - \mathcal{L}_{k+1}\right].
\end{aligned}
$$

In (14) set $y \leftarrow x_{k+1}$ and $x \leftarrow p_k$; then Line 6 of Alg. 3 yields

$$
\begin{aligned}
f(x_{k+1}) &\leq f(p_k) + \langle \nabla f(p_k),\, x_{k+1} - p_k\rangle + \tfrac{L}{2}\|x_{k+1} - p_k\|_2^2 \\
&= f(p_k) + \langle \nabla f(p_k),\, (1-\gamma_k)x_k + \gamma_k z_{k+1} - p_k\rangle + \tfrac{L}{2}\|x_{k+1} - p_k\|_2^2.
\end{aligned} \tag{15}
$$

Lines 3 and 6 show that $x_{k+1} - p_k = \gamma_k(z_{k+1} - z_k)$. Writing $f(p_k) = (1-\gamma_k)f(p_k) + \gamma_k f(p_k)$ in (15) we obtain

$$
\begin{aligned}
f(x_{k+1}) &\leq (1-\gamma_k)[f(p_k) + \langle \nabla f(p_k),\, x_k - p_k\rangle] \\
&\quad + \gamma_k[f(p_k) + \langle \nabla f(p_k),\, z_{k+1} - p_k\rangle + \tfrac{L\gamma_k}{2}\|z_{k+1} - z_k\|_2^2].
\end{aligned} \tag{16}
$$

Since f is convex, $(1-\gamma_k)[f(p_k) + \langle \nabla f(p_k),\, x_k - p_k\rangle] \leq (1-\gamma_k)f(x_k)$. Let us now bound the terms in the second line of (16). Denoting the error in the gradient by $\delta_k := \nabla f(p_k) - g_k$ (recall from Line 4 that $E[g_k] = \nabla f(p_k)$), we obtain

$$
\langle \nabla f(p_k),\, z_{k+1} - p_k\rangle = \langle \nabla f(p_k),\, x - p_k\rangle + \langle g_k,\, z_{k+1} - x\rangle + \langle \delta_k,\, z_{k+1} - x\rangle. \tag{17}
$$

To tackle the second term, notice that z_{k+1} is computed via Line 5, and thus satisfies the optimality condition

$$
\langle g_k + A^T(\beta\theta_k(Az_{k+1} + By_k - b) - \theta_k\lambda_k) + \gamma_k\eta_k^{-1}(z_{k+1} - z_k),\, x - z_{k+1}\rangle \geq 0, \quad \forall x \in \mathcal{X}.
$$

Rearranging this inequality we bound the $\langle g_k,\, z_{k+1} - x\rangle$ term as

$$
\langle g_k,\, z_{k+1} - x\rangle \leq \beta\theta_k\langle A^T(Az_{k+1} + By_k - b) - \theta_k A^T\lambda_k,\, x - z_{k+1}\rangle + \tfrac{\gamma_k}{\eta_k}\langle z_{k+1} - z_k,\, x - z_{k+1}\rangle. \tag{18}
$$

To enable telescoping, we apply the following identity

$$
\langle a - b,\, c - a\rangle = \tfrac{1}{2}\left(\|c - b\|_2^2 - \|c - a\|_2^2 - \|a - b\|_2^2\right), \tag{19}
$$

to the last term in (18). The leads us to replace (18) by

$$
\langle g_k,\, z_{k+1} - x\rangle \leq \theta_k\langle \beta A^T(Az_{k+1} + By_k - b) - A^T\lambda_k,\, x - z_{k+1}\rangle + \tfrac{\gamma_k}{2\eta_k}\left[\|x - z_k\|_2^2 - \|x - z_{k+1}\|_2^2 - \|z_{k+1} - z_k\|_2^2\right]. \tag{20}
$$

Now we work on the first term on the rhs of (20). Add and subtract By_{k+1} from it and use Line 8 to obtain

$$
\theta_k\langle \beta A^T(Az_{k+1} + By_k - b) - A^T\lambda_k,\, x - z_{k+1}\rangle = \theta_k\langle -A^T\lambda_{k+1} + \beta A^T B(y_k - y_{k+1})),\, x - z_{k+1}\rangle. \tag{21}
$$

Combining equations (21) and (17) with inequality (20), and plugging into the second part of inequality (16) we obtain

$$
\begin{aligned}
&f(p_k) + \langle \nabla f(p_k),\, z_{k+1} - p_k\rangle + \tfrac{L\gamma_k}{2}\|z_{k+1} - z_k\|_2^2 \\
&\leq f(p_k) + \langle \nabla f(p_k),\, x - p_k\rangle + \langle \delta_k,\, z_{k+1} - x\rangle + \tfrac{\gamma_k}{2\eta_k}\left[\Delta_k - \Delta_{k+1} - \|z_{k+1} - z_k\|_2^2\right] + \tfrac{L\gamma_k}{2}\|z_{k+1} - z_k\|_2^2 \\
&+ \theta_k\langle -A^T\lambda_{k+1} + \beta A^T B(y_k - y_{k+1}),\, x - z_{k+1}\rangle,
\end{aligned} \tag{22}
$$

where we wrote $\Delta_k := \|x - z_k\|_2^2$ to simplify notation.

Let us further simplify (22) by first bounding a part of the AL term; following (Ouyang et al., 2013) we have

$$
\begin{aligned}
&\theta_k\langle \beta A^T B(y_k - y_{k+1}),\, x - z_{k+1}\rangle = \theta_k\beta\langle Ax - Az_{k+1},\, By_k - By_{k+1}\rangle \\
&= \tfrac{\beta\theta_k}{2}\left[\left(\|Ax + By_k - b\|_2^2 - \|Ax + By_{k+1} - b\|_2^2\right) + \left(\|Az_{k+1} + By_{k+1} - b\|_2^2 - \|Az_{k+1} + By_k - b\|_2^2\right)\right] \\
&\leq \tfrac{\beta\theta_k}{2}\left(\|Ax + By_k - b\|_2^2 - \|Ax + By_{k+1} - b\|_2^2\right) + \tfrac{\theta_k}{2\beta}\|\lambda_{k+1} - \lambda_k\|_2^2,
\end{aligned} \tag{23}
$$

where the last inequality follows by discarding the negative term and using Line 8.

Next we use the stepsize $\eta_k = 1/(L+\alpha_k)$ to rewrite the last term on the second line of (22) as

$$\begin{aligned}&\tfrac{\gamma_k}{2\eta_k}\left[\Delta_k - \Delta_{k+1} - \|z_{k+1}-z_k\|_2^2\right] + \tfrac{L\gamma_k}{2}\|z_{k+1}-z_k\|_2^2\\ &= -\tfrac{\gamma_k\alpha_k}{2}\|z_{k+1}-z_k\|_2^2 + \tfrac{\gamma_k}{2\eta_k}\left[\Delta_k - \Delta_{k+1}\right],\end{aligned} \tag{24}$$

It remains to bound the stochastic error $\langle \delta_k,\, z_{k+1}-x\rangle$; here we use Young's inequality to write

$$\langle \delta_k,\, z_{k+1}-x\rangle = \langle \delta_k,\, z_{k+1}-z_k\rangle + \langle \delta_k,\, z_k - x\rangle \le \tfrac{1}{2\gamma_k\alpha_k}\|\delta_k\|_2^2 + \tfrac{\gamma_k\alpha_k}{2}\|z_{k+1}-z_k\|_2^2 + \langle \delta_k,\, z_k - w\rangle. \tag{25}$$

Adding equation (24) and inequality (25), after cancellation we obtain the inequality

$$\begin{aligned}&\tfrac{\gamma_k}{2\eta_k}\left[\Delta_k - \Delta_{k+1} - \|z_{k+1}-z_k\|_2^2\right] + \tfrac{L\gamma_k}{2}\|z_{k+1}-z_k\|_2^2 + \langle \delta_k,\, z_{k+1}-x\rangle\\ &\le \quad \tfrac{1}{2\gamma_k\alpha_k}\|\delta_k\|_2^2 + \langle \delta_k,\, z_k - x\rangle + \tfrac{\gamma_k}{2\eta_k}\left[\Delta_k - \Delta_{k+1}\right].\end{aligned} \tag{26}$$

Using (26) and (23) in inequality (22) and noting that $f(p_k) + \langle \nabla f(p_k),\, x - p_k\rangle \le f(x)$ yields

$$\begin{aligned}&f(p_k) + \langle \nabla f(p_k),\, z_{k+1}-p_k\rangle + \tfrac{L\gamma_k}{2}\|z_{k+1}-z_k\|_2^2\\ &\quad\le f(x) + \tfrac{1}{2\gamma_k\alpha_k}\|\delta_k\|_2^2 + \langle \delta_k,\, z_k - x\rangle + \tfrac{\gamma_k}{2\eta_k}\left[\Delta_k - \Delta_{k+1}\right] + \theta_k\langle z_{k+1}-x,\, A^T\lambda_{k+1}\rangle\\ &\quad+ \tfrac{\beta\theta_k}{2}\left(\|Ax+By_k-b\|_2^2 - \|Ax+By_{k+1}-b\|_2^2\right) + \tfrac{\theta_k}{2\beta}\|\lambda_{k+1}-\lambda_k\|_2^2.\end{aligned} \tag{27}$$

Multiplying (27) by γ_k and plugging it back into (16), we therefore obtain the inequality

$$\begin{aligned}f(x_{k+1}) \le{}& (1-\gamma_k)f(x_k) + \gamma_k f(x) + \tfrac{1}{2\alpha_k}\|\delta_k\|_2^2 + \gamma_k\langle \delta_k,\, z_k - x\rangle + \tfrac{\gamma_k^2}{2\eta_k}[\Delta_k - \Delta_{k+1}]\\ &+ \gamma_k\theta_k\Big\{\tfrac{\beta}{2}\left(\|Ax+By_k-b\|_2^2 - \|Ax+By_{k+1}-b\|_2^2\right) + \tfrac{1}{2\beta}\|\lambda_{k+1}-\lambda_k\|_2^2 + \langle z_{k+1}-x,\, A^T\lambda_{k+1}\rangle\Big\}.\end{aligned} \tag{28}$$

Now let us obtain a bound on the y terms. From convexity of $h(y)$, we have $h(y_{k+1}) \le h(y) + \langle h'(y_{k+1}),\, y_{k+1}-y\rangle$. Therefore, upon using Lines 7 and 8 it follows that

$$h(y_{k+1}) - h(y) \le \theta_k\langle y_{k+1}-y,\, B^T\lambda_{k+1}\rangle. \tag{29}$$

Line 6 also yields

$$\langle \lambda_{k+1}-\lambda,\, Az_{k+1}+By_{k+1}-b\rangle = \tfrac{1}{2\beta}\left[\|\lambda-\lambda_k\|_2^2 - \|\lambda-\lambda_{k+1}\|_2^2 - \|\lambda_{k+1}-\lambda_k\|_2^2\right]. \tag{30}$$

Recall that $w^T := [z; y; \lambda]$ and $[F(w)]^T := [-A^T\lambda; -B^T\lambda; Az+By-b]$. Thus, combining (28)–(30) we obtain

$$\begin{aligned}&f(x_{k+1}) + \gamma_k[h(y_{k+1}) - h(y)] + \gamma_k\theta_k\langle w_{k+1}-w,\, F(w_{k+1})\rangle\\ &\qquad\le (1-\gamma_k)f(x_k) + \gamma_k f(x) + \tfrac{1}{2\alpha_k}\|\delta_k\|_2^2 + \gamma_k\langle \delta_k,\, z_k - x\rangle + \tfrac{\gamma_k^2}{2\eta_k}[\Delta_k - \Delta_{k+1}]\\ &\qquad+ \tfrac{\beta\gamma_k\theta_k}{2}[\mathcal{A}_k - \mathcal{A}_{k+1}] + \tfrac{\gamma_k\theta_k}{2\beta}\left[\mathcal{L}_k - \mathcal{L}_{k+1}\right].\end{aligned} \tag{31}$$

2.2. Proof of key bound: Lemma 4

Lemma 4. *Using the notation of Lemma 3, for* $(1-\gamma_{k+1})/\gamma_{k+1}^2 \le 1/\gamma_k^2$, *inequality* (31) *yields*

$$\begin{aligned}&\tfrac{1}{\gamma_k^2}(f(x_{k+1}) - f(x)) + \sum_{j=1}^k \tfrac{1}{\gamma_j}\left[h(y_{j+1}) - h(y)\right] + \tfrac{\theta_j}{\gamma_j}\left[\langle w_{j+1}-w,\, F(w_{j+1})\rangle\right]\\ &\le \tfrac{L+\alpha_k}{2}R^2 + \tfrac{\beta}{2}\sum_{j=1}^k \mathcal{A}_j + \tfrac{1}{2\beta}\sum_{j=1}^k \mathcal{L}_j + \sum_{j=1}^k \tfrac{1}{\gamma_j^2\alpha_j}\|\delta_j\|_2^2 + \tfrac{1}{\gamma_j}\langle \delta_j,\, z_j - x\rangle.\end{aligned} \tag{32}$$

Proof. The proof below builds on the nice analysis of (Tseng, 2008), adapted to our stochastic setting. We start by subtracting $f(x)$ from both sides of (31) and dividing by γ_k^2. This yields

$$\begin{aligned}
&\tfrac{1}{\gamma_k^2}[f(x_{k+1}) - f(x)] + \tfrac{1}{\gamma_k}\left[h(y_{k+1}) - h(y)\right] + \tfrac{\theta_k}{\gamma_k}\left[\langle w_{k+1} - w,\, F(w_{k+1})\rangle\right] \\
&\quad\leq \tfrac{1-\gamma_k}{\gamma_k^2}[f(x_k) - f(x)] + \tfrac{1}{2\alpha_k\gamma_k^2}\|\delta_k\|_2^2 + \tfrac{1}{\gamma_k}\langle \delta_k,\, z_k - x\rangle + \tfrac{1}{2\eta_k}[\Delta_k - \Delta_{k+1}] \\
&\quad+\ \tfrac{\beta\theta_k}{2\gamma_k}\left[\mathcal{A}_k - \mathcal{A}_{k+1}\right] + \tfrac{\theta_k}{2\beta\gamma_k}[\mathcal{L}_k - \mathcal{L}_{k+1}].
\end{aligned} \tag{33}$$

Since $\gamma_k = \frac{2}{k+1}$, it follows that $\frac{1-\gamma_{k+1}}{\gamma_{k+1}^2} \leq \frac{1}{\gamma_k^2}$. Assuming $f(x_{k+1}) \geq f(x)$ we then obtain

$$\tfrac{1-\gamma_{k+1}}{\gamma_{k+1}^2}[f(x_{k+1}) - f(x)] \leq \tfrac{1}{\gamma_k^2}[f(x_{k+1}) - f(x)]. \tag{34}$$

Further assuming that $f(x_k) \geq f(x)$ (for all k), inequality (34) allows us to unroll (32) to obtain the bound

$$\begin{aligned}
&\tfrac{1-\gamma_{k+1}}{\gamma_{k+1}^2}[f(x_{k+1}) - f(x)] + \sum_{j=1}^{k} \tfrac{1}{\gamma_j}\left[h(y_{j+1}) - h(y)\right] + \tfrac{\theta_j}{\gamma_j}\left[\langle w_{j+1} - w,\, F(w_{j+1})\rangle\right] \\
&\leq \tfrac{1-\gamma_1}{\gamma_1^2}[f(x_1) - f(x)] + \sum_{j=1}^{k}\left\{\tfrac{1}{2\alpha_j\gamma_j^2}\|\delta_j\|_2^2 + \tfrac{1}{\gamma_j}\langle \delta_j,\, z_j - x\rangle + \tfrac{1}{2\eta_j}[\Delta_j - \Delta_{j+1}]\right\} \\
&\quad+ \sum_{j=1}^{k}\left\{\tfrac{\beta\theta_j}{2\gamma_j}\left[\mathcal{A}_j - \mathcal{A}_{j+1}\right] + \tfrac{\theta_j}{2\beta\gamma_j}[\mathcal{L}_j - \mathcal{L}_{j+1}]\right\}.
\end{aligned} \tag{35}$$

But $\gamma_1 = 1$, so the first term disappears. Combining (33) with the recursive relation (35) we then obtain[1]

$$\begin{aligned}
&\tfrac{1}{\gamma_k^2}[f(x_{k+1}) - f(x)] + \sum_{j=1}^{k} \tfrac{1}{\gamma_j}\left[h(y_{j+1}) - h(y)\right] + \tfrac{\theta_j}{\gamma_j}\left[\langle w_{j+1} - w,\, F(w_{j+1})\rangle\right] \\
&\leq \sum_{j=1}^{k}\left\{\tfrac{1}{2\alpha_j\gamma_j^2}\|\delta_j\|_2^2 + \tfrac{1}{\gamma_j}\langle \delta_j,\, z_j - x\rangle + \tfrac{1}{2\eta_j}[\Delta_j - \Delta_{j+1}] + \tfrac{\beta\theta_j}{2\gamma_j}\left[\mathcal{A}_j - \mathcal{A}_{j+1}\right] + \tfrac{\theta_j}{2\beta\gamma_j}[\mathcal{L}_j - \mathcal{L}_{j+1}]\right\}.
\end{aligned} \tag{36}$$

Let us bound the different parts now. First, consider the part with $\Delta_j - \Delta_{j+1}$. From our assumption, we have $\Delta_j = \|z_j - x\|_2^2 \leq R^2$ for all $z_j \in \mathcal{X}$. Setting $\eta_j = \frac{1}{L+\alpha_j}$ we get

$$\sum_{j=1}^{k}\left(\tfrac{1}{2\eta_j}[\Delta_j - \Delta_{j+1}]\right) \leq \tfrac{\Delta_1}{2\eta_1} + \sum_{j=2}^{k}\tfrac{\Delta_j}{2}\left(\tfrac{1}{\eta_j} - \tfrac{1}{\eta_{j-1}}\right) \leq \tfrac{L+\alpha_k}{2}R^2.$$

Similarly processing the $A_j = \mathcal{A}_j - \mathcal{A}_{j+1}$ terms, and using $0 \leq \frac{\theta_j}{\gamma_j} - \frac{\theta_{j-1}}{\gamma_{j-1}} \leq 1$ for $j \geq 2$, we obtain

$$\sum_{j=1}^{k}\tfrac{\beta\theta_j}{2\gamma_j}[\mathcal{A}_j - \mathcal{A}_{j+1}] \leq \tfrac{\beta\theta_1\mathcal{A}_1}{2\gamma_1} + \sum_{j=2}^{k}\tfrac{\beta\mathcal{A}_j}{2}\left(\tfrac{\theta_j}{\gamma_j} - \tfrac{\theta_{j-1}}{\gamma_{j-1}}\right) \leq \tfrac{\beta}{2}\sum_{j=1}^{k}\mathcal{A}_j.$$

where we set $\theta_1 = \gamma_1 = 1$ for simplicity. The same manipulation yields

$$\sum_{j=1}^{k}\tfrac{\theta_j}{2\beta\gamma_j}[\mathcal{L}_j - \mathcal{L}_{j+1}] \leq \tfrac{1}{2\beta}\sum_{j=1}^{k}\mathcal{L}_j.$$

Putting these bounds back into (36) we obtain the lemma. □

2.3. Proof of Theorem 5

Recall that we define the averaged solution vectors

$$\bar{x}_k := x_{k+1}, \quad \bar{y}_k := \sum_{j=1}^{k}\nu_j y_{j+1}, \quad \bar{z}_k := \sum_{j=1}^{k}\nu_j z_{j+1}, \quad \bar{\lambda}_k := \sum_{j=1}^{k}\nu_j\lambda_j. \tag{37}$$

Theorem 5. *Let $\bar{x}_k$, $\bar{z}_k$, $\bar{y}_k$, and $\bar{\lambda}_k$ be as defined in (37). Then, for $\theta_k = 1$, for $k \geq 0$,*

$$\tfrac{1}{\gamma_k^2}\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2] \leq \tfrac{LR^2}{2} + \tfrac{\beta D_{\mathcal{Y}}^2}{2} + \tfrac{\rho^2}{2\beta} + \tfrac{\alpha_k R^2}{2} + \sigma^2\sum_{j=1}^{k}\tfrac{1}{\gamma_j^2\alpha_j}.$$

[1] The key to realizing this is to notice that we set up a recursion $c_{k+1}r_{k+1} \leq c_k r_k + \theta_k$. Unroll this; and then later as per (33) use $d_k r_{k+1} \leq c_k r_k + \theta_k$ and apply the recursive bound to $c_k r_k$ to obtain (36).

Proof. Multiplying both sides of (32) by γ_k^2, invoking the assumption $\|\delta_j\|_2^2 \leq \sigma^2$, and taking expectations, we obtain

$$\begin{aligned}&\mathbb{E}[f(x_{k+1}) - f(x)] + \sum_{j=1}^{k} \tfrac{\gamma_k^2}{\gamma_j}\mathbb{E}\left[h(y_{j+1}) - h(y) + \langle w_{j+1} - w,\, F(w_{j+1})\rangle\right] \\ &\leq \gamma_k^2 \mathbb{E}\left[\tfrac{L+\alpha_k}{2}R^2 + \tfrac{\beta}{2}\sum_{j=1}^{k}\mathcal{A}_j + \tfrac{1}{2\beta}\sum_{j=1}^{k}\mathcal{L}_j + \sum_{j=1}^{k}\tfrac{1}{\gamma_j^2\alpha_j}\|\delta_j\|_2^2 + \tfrac{1}{\gamma_j}\langle \delta_j,\, z_j - x\rangle\right].\end{aligned} \tag{38}$$

Now invoke independence of δ_k and z_k and unbiasedness of the stochastic gradients to conclude $\mathbb{E}[\langle \delta_k,\, z_k - x\rangle] = 0$. But to proceed from (38) to the final result, some work is needed. Recall the weighted averages (37). Set $\nu_j = \frac{2(j+1)}{(k+1)(k+2)}$ as the weights for $\bar{y}_k$ (note that $\sum_{j=1}^{k}\nu_j = 1$). With this definition, using convexity of h we have

$$h(\bar{y}_k) - h(y) \leq \tfrac{2}{(k+1)(k+2)}\sum_{j=1}^{k}(j+1)[h(y_{j+1}) - h(y)] < \gamma_k^2\sum_{j=1}^{k}\tfrac{1}{\gamma_j}[h(y_{j+1}) - h(y)]. \tag{39}$$

Since F is skew-symmetric and monotone and the duality gap is always nonnegative, using $\zeta_j = 2\gamma_k^2\theta_j/\gamma_j$, we obtain

$$\langle \bar{w}_k - w,\, F(w)\rangle = \sum_{j=1}^{k}\nu_j\langle w_{j+1} - w,\, F(w)\rangle \leq \sum_{j=1}^{k}\tfrac{\gamma_k^2}{\gamma_j}\langle w_{j+1} - w,\, F(w_{j+1})\rangle. \tag{40}$$

Combining the above results, we obtain the inequality

$$\begin{aligned}&f(\bar{x}_k) + h(\bar{y}_k) - f(x) - h(y) + \langle \bar{w}_k - w,\, F(w)\rangle \\ &\leq f(x_{k+1}) - f(x) + \sum_{j=1}^{k}\tfrac{\gamma_k^2}{\gamma_j}\left[h(y_{j+1}) - h(y) + \langle w_{j+1} - w,\, F(w_{j+1})\rangle\right].\end{aligned} \tag{41}$$

Before we can translate (41) to obtain the theorem, we need to control the dual variable λ. For that, we follow (Ouyang et al., 2013) and note that key inequality (38) holds for $\lambda \in \mathbb{R}^n$, so it also holds within the norm-ball $\mathcal{B} = \{\lambda \mid \|\lambda\|_2 \leq \rho\}$. This ball allows us to then find the worst case value that the left-hand-side of (41) may attain at optimal $x = x^*$ and $y = y^*$. Since $Ax^* + By^* = b$ (feasibility), it then follows that

$$\begin{aligned}&\sup_{\lambda\in\mathcal{B}}\{f(\bar{x}_k) + h(\bar{y}_k) - f(x) - h(y) + \langle \bar{w}_k - w,\, F(w)\rangle\} \\ &= \sup_{\lambda\in\mathcal{B}}\{f(\bar{x}_k) - f(x) + h(\bar{y}_k) - h(y) + \langle \bar{\lambda}_k,\, Ax^* + By^* - b\rangle - \langle \lambda,\, A\bar{z}_k + B\bar{y}_k - b\rangle\} \\ &= f(\bar{x}_k) - f(x) + h(\bar{y}_k) - h(y) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2.\end{aligned} \tag{42}$$

Using (42) in conjunction with (41) and (38) then yields

$$\mathbb{E}[f(\bar{x}_k) - f(x) + h(\bar{y}_k) - h(y) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2] \leq \gamma_k^2\mathbb{E}\left[\tfrac{L+\alpha_k}{2}R^2 + \sum_{j=1}^{k}\left(\tfrac{\beta}{2}\mathcal{A}_j + \tfrac{1}{2\beta}\mathcal{L}_j\right) + \sum_{j=1}^{k}\tfrac{1}{\gamma_j^2\alpha_j}\|\delta_j\|_2^2\right].$$

Using $z = z^*$, $y = y^*$, $x^* = z^*$, and $\lambda_j \in \mathcal{B}$, we see that

$$\mathcal{A}_j = \|Az^* + By_j - b\|_2^2 = \|B(y_j - y^*)\|_2^2 \leq D_{\mathcal{Y}}^2, \qquad \mathbb{E}\big[\max_{\lambda\in\mathcal{B}}\mathcal{L}_j\big] \leq 2\rho^2.$$

With these bounds we obtain

$$\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2] \leq \gamma_k^2\tfrac{LR^2}{2} + \tfrac{\gamma_k^2\beta}{2}D_{\mathcal{Y}}^2 + \tfrac{\gamma_k^2}{\beta}\rho^2 + \tfrac{\gamma_k^2\alpha_k R^2}{2} + \gamma_k^2\sigma^2\sum_{j=1}^{k}\tfrac{1}{\gamma_j^2\alpha_j}.$$

Now set $\alpha_j = c^{-1}\sigma(j+1)^{3/2}$ (for a tunable constant c) and $\gamma_j = 2/(j+1)$ we finally obtain

$$\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2] \leq \tfrac{2LR^2}{(k+1)^2} + \tfrac{2\beta D_{\mathcal{Y}}^2}{k+1} + \tfrac{2\rho^2}{\beta(k+1)} + \tfrac{2\sigma(c^{-1}+c)}{\sqrt{k+1}}. \tag{43}$$

References

Chambolle, A. and Pock, T. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

Goldfarb, D., Ma, S., and Scheinberg, K. Fast alternating linearization methods for minimizing the sum of two convex functions. *Math. Prog. Ser. A*, 2012.

Ouyang, Hua, He, Niao, Tran, Long, and Gray, Alexander G. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 80–88, 2013.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *Unpublished*, 2008.