5. Appendix A: Proof for Theorem 2

Recall that the Augmented Lagrangian $\mathcal{L}(W_1, W_2, Y)$ is of the form
\[
(D, W_1) + \langle Y, W_1 - W_2 \rangle + \frac{\rho}{2} \|W_1 - W_2\|^2.
\]

Then let $X = [W_1; W_2]$ be the primal variables and denote
\[
X(Y) := \{X | X = \arg \min_X \mathcal{L}(X, Y)\}
\]
with
\[
\bar{X}^t := \arg\min_{X \in X(Y^t)} \|X - X^t\|,
\]
and let
\[
A X = \left[ \begin{array}{cc}
I & -I \\
D & O
\end{array} \right] \left[ \begin{array}{c}
W_1 \\
W_2
\end{array} \right] = W_1 - W_2 \tag{23}
\]
and
\[
\langle C, X \rangle = \left[ \begin{array}{cc}
D & O
\end{array} \right]^T \left[ \begin{array}{c}
W_1 \\
W_2
\end{array} \right] = (D, W_1) \tag{24}
\]
The Augmented Lagrangian can be re-written as
\[
\mathcal{L}(X, Y) = \langle C, X \rangle + \langle Y, AX \rangle + \frac{\rho}{2} \|AX\|^2. \tag{25}
\]
The dual function is
\[
d(Y) = \min_{X \in \text{Conv}(\mathcal{A}) \times \text{Conv}(\mathcal{G})} \mathcal{L}(X, Y)
\]
and
\[
d^* = \max_Y d(Y)
\]
is the optimal dual function value. Then we measure the sub-optimality of iterates $\{(X^t, Y^t)\}_{t=1}^T$ given by GDMM in terms of dual function difference
\[
\Delta^t_d = d^* - d(Y^t)
\]
and the primal function difference for a given dual iterate $Y^t$:
\[
\Delta^t_p = \mathcal{L}(X^{t+1}, Y^t) - d(Y^t)
\]
yielded by $X^{t+1}$ obtained from AFW steps. Then we have following lemma.

Lemma 1 (Dual Progress). Each iteration of GDMM (Algorithm 1) has
\[
\Delta^t_d - \Delta^{t-1}_d \leq -\eta \langle AX^t, A\bar{X}^t \rangle. \tag{26}
\]

Proof. Let
\[
\Delta^t_d - \Delta^{t-1}_d = (d^* - d(Y^t)) - (d^* - d(Y^{t-1})) = \mathcal{L}(X^{t-1}, Y^{t-1}) - \mathcal{L}(X^t, Y^t) \leq \mathcal{L}(\bar{X}^t, Y^{t-1}) - \mathcal{L}(\bar{X}^t, Y^t) = \langle Y^{t-1} - Y^t, A\bar{X}^t \rangle = -\eta \langle AX^t, A\bar{X}^t \rangle
\]
where the first inequality follows from the optimality of $X^{t-1}$ for the function $\mathcal{L}(X, Y^{t-1})$ defined by $Y^{t-1}$, and the last equality follows from the dual update in GDMM (14).

On the other hand, the following lemma gives an expression on the primal progress that is independent of the algorithm used for minimizing Augmented Lagrangian

Lemma 2 (Primal Progress). Each iteration of GDMM (Algorithm 1) has
\[
\Delta^t_p - \Delta^{t-1}_p \leq \mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) + \eta \|AX^t - A\bar{X}^t\|^2 - \eta \langle AX^t, A\bar{X}^t \rangle
\]

Proof. Let
\[
\Delta^t_p - \Delta^{t-1}_p = \mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) - (d(Y^t) - d(Y^{t-1})) = \mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) + \mathcal{L}(X^t, Y^{t-1}) - \mathcal{L}(X^t, Y^t) + (d(Y^{t-1}) - d(Y^t)) \leq \mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) + \eta \|AX^t - A\bar{X}^t\|^2 - \eta \langle AX^t, A\bar{X}^t \rangle
\]
where the last inequality uses Lemma 1 on $d(Y^{t-1}) - d(Y^t) = \Delta^t_d - \Delta^{t-1}_d$.

By combining results of Lemma 1 and 2, we can obtain a joint progress of the form
\[
\Delta^t_d - \Delta^{t-1}_d + \Delta^t_p - \Delta^{t-1}_p \leq \mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) + \eta \|AX^t - A\bar{X}^t\|^2 - \eta \langle AX^t, A\bar{X}^t \rangle \tag{27}
\]
Note the only term that can be positive in (27) is the second. To guarantee descent of the joint progress, we bound the second term with the primal gap $\mathcal{L}(X^t, Y^t) - d(Y^t)$ given by the following lemma

Lemma 3. \[
\|AX^t - A\bar{X}^t\|^2 \leq \frac{2}{\rho} (\mathcal{L}(X^t, Y^t) - \mathcal{L}(\bar{X}^t, Y^t)) \tag{28}
\]

Proof. Let
\[
\hat{\mathcal{L}}(X, Y) = h(X) + g(AX),
\]
where
\[
g(AX) = \frac{\rho}{2} \|AX\|^2
\]
and
\[
h(X) = \langle C, X \rangle + \langle Y, AX \rangle + I_{X \in C}
\]
where $I_{X \in C} = 0$ if $X \in C$ and $I_{X \notin C} = \infty$ otherwise, and

$$C = \{(W_1, W_2) | W_1 \in Conv(A), W_2 \in Conv(G)\}. \quad (29)$$

Note we have $\hat{L}(\hat{X}^t, Y^t) = L(\hat{X}^t, Y^t)$, $\hat{L}(X^t, Y^t) = L(X^t, Y^t)$ due to feasible iterates. According to the definition of $d(Y)$, we know that

$$0 \in \partial_X \hat{L}(\hat{X}^t, Y) = \partial h(\hat{X}^t) + A^T \nabla g(A(\hat{X}^t))$$

And by the convexity of $h(\cdot)$ and the strong convexity of $g(\cdot)$, we have

$$h(X^t) - h(\hat{X}^t) \geq \langle \partial h(\hat{X}^t), X^t - \hat{X}^t \rangle$$

and

$$g(A(X^t)) - g(A(\hat{X}^t)) \geq \langle A^T \nabla g(A(\hat{X}^t)), X^t - \hat{X}^t \rangle + \frac{\rho}{2} \|A(X^t) - A(\hat{X}^t)\|^2$$

The above two together implies

$$L(X^t, Y^t) - L(\hat{X}^t, Y^t) \geq \frac{\rho}{2} \|A(X^t) - A(\hat{X}^t)\|^2$$

which leads to our conclusion. \qed

Then to guarantee significant descent of (27) relative to the current sub-optimality, we need to lower bound the magnitude of first term $L(X^{t+1}, Y^t) - L(X^t, Y^t)$ and last term $-\eta \|A X^t\|^2$. Note by Danskin's theorem, we have

$$\nabla d(Y^t) = A \hat{X}^t$$

and we have the following lower bound on $\|A \hat{X}^t\|$ by the concavity of $d(Y)$

$$d^* - d(Y^t) \leq \langle A \hat{X}^t, Y^{t*} - Y^t \rangle \leq \|A \hat{X}^t\| \|Y^{t*} - Y^t\| \leq \|A \hat{X}^t\| R_Y$$

where $Y^{t*}$ is the maximizer of $d(Y)$ that is closest to $Y^t$ and $R_Y$ is an upper bound on the distance (in $\ell_2$ norm) of dual iterates $\{Y^t\}_{t=0}$ to its projection to the set of maximizer of $d(Y)$. Therefore, the progress (27) can be lower bounded as

$$\Delta^t_d - \Delta^{t-1}_d + \Delta^t_p - \Delta^{t-1}_p \leq L(X^{t+1}, Y^t) - L(X^t, Y^t) \leq \frac{2\eta}{\rho} \|A(X^t) - A(\hat{X}^t)\|^2 \quad (30)$$

The remaining thing to do is show that one good step of Away-Step Frank-Wolfe iterate suffices to give descent amount $L(X^{t+1}, Y^t) - L(X^t, Y^t)$ lower bounded by some constant multiple of the primal sub-optimality $L(X^t, Y^t) - L(\hat{X}^t, Y^t)$. Then by selecting GDMM step size $\eta$ small enough, the RHS of (30) leads to a positive descent amount. Note this can be achieved by leveraging recent result from (Lacoste-Julien & Jaggi, 2015), who shows a linear-type convergence of AFW, even for non-strongly convex function of form (25). We thus provide the following lemma.

**Lemma 4.** The AFW (Algorithm 2) performed on $X = (W_1, W_2)$ gives descent amount

$$L(X^{t+1}, Y^t) - L(X^t, Y^t) \leq -\frac{\kappa}{1 + \kappa} (L(X^t, Y^t) - L(\hat{X}^t, Y^t)) \quad (31)$$

where $\kappa := \mu_f/(8C^A_f)$, $\mu_f$ is the generalized geometric strong convexity constant for function $L(\cdot)$ in domain $C$, and $C^A_f$ is the corresponding smoothness constant.

**Proof.** Note the AL (25) is of the form

$$F(X) = L(X, Y) = \langle C, X \rangle + f(AX) \quad (32)$$

where $f(AX) = \frac{2}{\rho} \|AX + Y/\rho\|^2 + const.$ is a $\rho$-strongly convex function w.r.t. $AX$, and we are minimizing the function subject to a polyhedral domain $C$ (defined at (29)). Therefore, by Theorem 10 of (Lacoste-Julien & Jaggi, 2015), we have the generalized geometrical strong convexity constant $\mu_f$ for function $L(\cdot)$ in domain $C$ that has

$$\mu_f \geq \mu(PWidh(C)) \quad (33)$$

where $PWidh(C) > 0$ is the pyramidal width of polyhedron $C$ and $\mu$ is the generalized strong convexity constant of function (32) defined in Lemma 9 of (Lacoste-Julien & Jaggi, 2015). By definition of the geometric strong convexity constant, we have

$$F(X) - F^* \leq \frac{g^2}{2\mu_f} \quad (34)$$

from (28) in (Lacoste-Julien & Jaggi, 2015), where $g_X = \langle \nabla F(X), v_{FW}(X) - v_A(X) \rangle$ for any FW atom $v_{FW}(X)$ and away atom $v_A(X)$ at point $X$. Note, since the convex polyhedron $C$ is separable w.r.t. $W_1, W_2$, we have

$$v_{FW}(X) = \begin{bmatrix} v_{FW}^{(1)} \\ v_{FW}^{(2)} \end{bmatrix} \quad (35)$$

and

$$v_A(X) = \begin{bmatrix} v_A^{(1)} \\ v_A^{(2)} \end{bmatrix} \quad (36)$$

Then consider the progress given by a non-drop ("good")
step at iterate \( s \) of the AFW. We have
\[
F(X^{s+1}) - F(X^s) \leq -\frac{\gamma}{2} g_s + \frac{C_f}{2} \gamma^2
\]
\[
\leq -\frac{g_s^2}{16C_f^2} \leq -\frac{\mu_f F(X^s) - F^*}{8C_f^2}
\]
assuming \( \gamma^* = g_s/(2C) < 1 \), where \( g_s = \langle -\nabla F, v_{FW}(X^s) - v_A(X^s) \rangle \), \( C_f \) is the curvature constant of \( F(X) \) on domain \( \mathcal{C} \) (eq. (26) in (Lacoste-Julien & Jaggi, 2015)). The first inequality follows from the fact that AFW chooses the smaller one between \( \langle \nabla F, d_{FW} \rangle \) and \( \langle \nabla F, d_A \rangle \) as the descent direction. The second inequality is given by minimizing RHS w.r.t. \( \gamma \in [0, 1] \). And the third inequality is from (34). In case \( \gamma^* = g_s/(2C) > 1 \), we have \( \gamma = 1 \) and
\[
F(X^{s+1}) - F(X^s) \leq -\frac{\gamma}{2} g_s + \frac{C_f}{2} \gamma^2
\]
\[
\leq -\frac{g_s}{4} \leq -(F(X^s) - F^*)/4
\]
\[
\leq -\frac{\mu_f F(X^s) - F^*}{8C_f^2}
\]
which leads to the same result.

Then let \( \kappa = \mu_f/(8C_f^2) \). We have
\[
F(X^{t+1}) - F(X^t) \leq F(X^{s+1}) - F(X^s)
\]
\[
\leq -\kappa (F(X^s) - F^*)
\]
\[
\leq -\kappa (F(X^{t+1}) - F^*)
\]
where the first inequality is due to \( F(X^t) \geq F(X^s) \) (since AFW is an descent algorithm). Through rearrangement we have
\[
F(X^{t+1}) - F^* \leq \frac{1}{1 + \kappa} (F(X^t) - F^*)
\]
which then leads to the conclusion. \( \square \)

Now we provide proof of the main theorem 2 as follows.

**Proof.** By lemma 4 and (30), we have
\[
\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1}
\]
\[
\leq \frac{-\kappa}{1 + \kappa} (L(X^t, Y^t) - L(\bar{X}^t, Y^t))
\]
\[
+ \frac{2\eta}{\rho} (L(X^t, Y^t) - L(\bar{X}^t, Y^t)) - \frac{\eta}{R_Y^2} \Delta_d^t.
\]
(39)

Then by choosing \( \eta < \frac{\kappa \rho}{2(1 + \kappa)} \), we have guaranteed descent on \( \Delta_p^t + \Delta_d^t \) for each GDMM iteration. By choosing \( \eta \leq \frac{\kappa \rho}{2(1 + \kappa)} \), we have
\[
(\Delta_p^t + \Delta_d^t) - (\Delta_p^{t-1} + \Delta_d^{t-1})
\]
\[
\leq \frac{-\kappa}{2(1 + \kappa)} (L(X^t, Y^t) - L(\bar{X}^t, Y^t)) - \frac{\eta}{R_Y^2} \Delta_d^t
\]
\[
\leq \frac{-\kappa}{2(1 + \kappa)} \Delta_p^t - \frac{\kappa \rho}{4(1 + \kappa)R_Y^2} \Delta_d^t
\]
\[
\leq \frac{-\kappa}{2(1 + \kappa)} (\Delta_p^0 + \Delta_d^0) \Delta_p^t - \frac{\kappa \rho}{4(1 + \kappa)R_Y^2} \Delta_d^t
\]
\[
\leq \left( -\frac{\kappa}{4(1 + \kappa)} \min\left( \frac{1}{\Delta_p^0 + \Delta_d^0}, \frac{\rho}{2R_Y^2} \right) \right) (\Delta_p^t + \Delta_d^t)^2
\]
where the third inequality is by non-increasing of \( \{\Delta_p^t + \Delta_d^t\}_{t=1}^\infty \). Then recursion of the form \( \Delta^t - \Delta^{t-1} \leq c\Delta^{t/2} \) leads to the conclusion. \( \square \)