5. Appendix A: Proof for Theorem 2

Recall that the Augmented Lagrangian $\mathcal{L}(W_1, W_2, Y)$ is of the form

$$\langle D, W_1\rangle + \langle Y, W_1 - W_2\rangle + \frac{\rho}{2}\|W_1 - W_2\|^2.$$

Then let $X = [W_1; W_2]$ be the primal variables and denote

$$\mathcal{X}(Y) := \{X | X = arg \min_X \mathcal{L}(X, Y)\}$$

with

$$\bar{X}^t := \underset{\bar{X}\in\mathcal{X}(Y^t)}{argmin} \|\bar{X} - X^t\|,$$

and let

$$\mathbf{A}X = \begin{bmatrix} I & -I \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = W_1 - W_2 \quad (23)$$

and

$$\langle C, X\rangle = \begin{bmatrix} D \\ O \end{bmatrix}^T \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \langle D, W_1\rangle \quad (24)$$

The Augmented Lagrangian can be re-written as

$$\mathcal{L}(X, Y) = \langle C, X\rangle + \langle Y, \mathbf{A}X\rangle + \frac{\rho}{2}\|\mathbf{A}X\|^2. \quad (25)$$

The dual function is

$$d(Y) = \min_{X\in Conv(\mathcal{A})\times Conv(\mathcal{G})} \mathcal{L}(X, Y)$$

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

$$\Delta_d^t = d^* - d(Y^t)$$

and the primal function difference for a given dual iterate Y^t:

$$\Delta_p^t = \mathcal{L}(X^{t+1}, Y^t) - d(Y^t)$$

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*

$$\Delta_d^t - \Delta_d^{t-1} \leq -\eta(\mathbf{A}X^t)^T(\mathbf{A}\bar{X}^t). \quad (26)$$

Proof.

$$\begin{aligned}\Delta_d^t - \Delta_d^{t-1} &= (d^* - d(Y^t)) - (d^* - d(Y^{t-1})) \\ &= \mathcal{L}(\bar{X}^{t-1}, Y^{t-1}) - \mathcal{L}(\bar{X}^t, Y^t) \\ &\leq \mathcal{L}(\bar{X}^t, Y^{t-1}) - \mathcal{L}(\bar{X}^t, Y^t) \\ &= \langle Y^{t-1} - Y^t, \mathbf{A}\bar{X}^t\rangle \\ &= -\eta\langle \mathbf{A}X^t, \mathbf{A}\bar{X}^t\rangle\end{aligned}$$

where the first inequality follows from the optimality of $\bar{X}^{t-1}$ 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*

$$\begin{aligned}\Delta_p^t - \Delta_p^{t-1} \leq& \mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) \\ &+ \eta\|\mathbf{A}X^t - \mathbf{A}\bar{X}^t\|^2 - \eta\langle \mathbf{A}X^t, \mathbf{A}\bar{X}^t\rangle\end{aligned}$$

Proof.

$$\begin{aligned}&\Delta_p^t - \Delta_p^{t-1} \\ =&\mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^{t-1}) - (d(Y^t) - d(Y^{t-1})) \\ =&\mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) + \mathcal{L}(X^t, Y^t) - \mathcal{L}(X^t, Y^{t-1}) \\ &+ (d(Y^{t-1}) - d(Y^t)) \\ \leq&\mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) + \eta\|\mathbf{A}X^t\|^2 - \eta\langle \mathbf{A}X^t, \mathbf{A}\bar{X}^t\rangle\end{aligned}$$

where the last inequality uses Lemma 1 on $d(Y^{t-1}) - d(Y^t) = \Delta_d^t - \Delta_d^{t-1}$. ☐

By combining results of Lemma 1 and 2, we can obtain a joint progress of the form

$$\begin{aligned}&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \\ &\leq \mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) + \eta\|\mathbf{A}X^t - \mathbf{A}\bar{X}^t\|^2 \\ &- \eta\|\mathbf{A}\bar{X}^t\|^2\end{aligned} \quad (27)$$

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.

$$\|\mathbf{A}X^t - \mathbf{A}\bar{X}^t\|^2 \leq \frac{2}{\rho}(\mathcal{L}(X^t, Y^t) - \mathcal{L}(\bar{X}^t, Y^t)) \quad (28)$$

Proof. Let

$$\tilde{\mathcal{L}}(X, Y) = h(X) + g(\mathbf{A}X),$$

where

$$g(\mathbf{A}X) = \frac{\rho}{2}\|\mathbf{A}X\|^2$$

and

$$h(X) = \langle C, X\rangle + \langle Y, \mathbf{A}X\rangle + \boldsymbol{I}_{X\in\mathcal{C}}$$

, where $\boldsymbol{I}_{X\in\mathcal{C}} = 0$ if $X \in \mathcal{C}$ and $\boldsymbol{I}_{X\in\mathcal{C}} = \infty$ otherwise, and

$$\mathcal{C} = \{(W_1, W_2) \mid W_1 \in Conv(\mathcal{A}), W_2 \in Conv(\mathcal{G})\}. \tag{29}$$

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

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

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

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

and

$$\begin{aligned} &g(\mathbf{A}(X^t)) - g(\mathbf{A}(\bar{X}^t)) \\ \geq &\langle \mathbf{A}^T(\nabla g(\mathbf{A}(\bar{X}^t))), X^t - \bar{X}^t \rangle + \frac{\rho}{2}\|\mathbf{A}(X^t)) - \mathbf{A}(\bar{X}^t)\|^2 \end{aligned}$$

The the above two together implies

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

which leads to our conclusion. □

Then to guarantee significant descent of (27) relative to the current sub-optimality, we need to lower bound the magnitude of first term $\mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t)$ and last term $-\eta\|\mathbf{A}\bar{X}^t\|^2$. Note by Danskins theorem, we have

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

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

$$\begin{aligned} d^* - d(Y^t) &\leq \langle \mathbf{A}\bar{X}^t, Y^{t*} - Y^t \rangle \\ &\leq \|\mathbf{A}\bar{X}^t\|\|Y^{t*} - Y^t\| \\ &\leq \|\mathbf{A}\bar{X}^t\| R_Y \end{aligned}$$

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 ℓ_2 norm) of dual iterates $\{Y^t\}_{t=0}^T$ to its projection to the set of maximizer of $d(Y)$. Therefore, the progress (27) can be lower bounded as

$$\begin{aligned} &\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \\ \leq &\mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) \\ &+ \frac{2\eta}{\rho}(\mathcal{L}(X^t, Y^t) - \mathcal{L}(\bar{X}^t, Y^t)) - \frac{\eta}{R_Y^2}\Delta_d^{t2} \end{aligned} \tag{30}$$

The remaining thing to do is show that one good step of Away-Step Frank-Wolfe iterate suffices to give descent amount $\mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t)$ lower bounded by some constant multiple of the primal sub-optimality $\mathcal{L}(X^t, Y^t) - \mathcal{L}(\bar{X}^t, Y^t)$. Then by selecting GDMM step size η 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*

$$\begin{aligned} &\mathcal{L}(X^{t+1}, Y^t) - \mathcal{L}(X^t, Y^t) \\ &\leq -\frac{\kappa}{1+\kappa}(\mathcal{L}(X^t, Y^t) - \mathcal{L}(\bar{X}^t, Y^t)) \end{aligned} \tag{31}$$

where $\kappa := \mu_f/(8C_f^A)$, μ_f is the generalized geometric strong convexity constant for function $\mathcal{L}(.)$ in domain $\mathcal{C}$, and C_f^A is the corresponding smoothnesss constant.

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

$$F(X) = \mathcal{L}(X, Y) = \langle C, X \rangle + f(\mathbf{A}X) \tag{32}$$

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

$$\mu_f \geq \mu(PWidth(\mathcal{C})) \tag{33}$$

where $PWidth(\mathcal{C}) > 0$ is the pyramidal width of polyhedron $\mathcal{C}$ and μ 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

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

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

$$\boldsymbol{v}_{FW}(X) = \begin{bmatrix} \boldsymbol{v}_{FW}^{(1)} \\ \boldsymbol{v}_{FW}^{(2)} \end{bmatrix} \tag{35}$$

and

$$\boldsymbol{v}_A(X) = \begin{bmatrix} \boldsymbol{v}_A^{(1)} \\ \boldsymbol{v}_A^{(2)} \end{bmatrix} \tag{36}$$

Then consider the progress given by a non-drop ("good") step at iterate s of the AFW. We have

$$F(X^{s+1})-F(X^s)\le -\frac{\gamma}{2}g_s+\frac{C_f^A}{2}\gamma^2 \le -\frac{g_s^2}{16C_f^A} \le -\frac{\mu_f(F(X^s)-F^*)}{8C_f^A} \tag{37}$$

assuming $\gamma^*=g_s/(2C)<1$, where $g_s=\langle -\nabla F,\boldsymbol{v}_{FW}(X^s)-\boldsymbol{v}_A(X^s)\rangle$, C_f^A 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 $\langle\nabla F,\boldsymbol{d}_{FW}\rangle$ and $\langle\nabla F,\boldsymbol{d}_A\rangle$ 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

$$F(X^{s+1})-F(X^s)\le -\frac{\gamma}{2}g_s+\frac{C_f^A}{2}\gamma^2 \le -g_s/4\le -(F(X^s)-F^*)/4 \le -\frac{\mu_f(F(X^s)-F^*)}{8C_f^A} \tag{38}$$

which leads to the same result.

Then let $\kappa=\mu_f/(8C_f^A)$. We have

$$\begin{aligned}F(X^{t+1})-F(X^t)&\le F(X^{s+1})-F(X^s)\\&\le -\kappa(F(X^s)-F^*)\\&\le -\kappa(F(X^{t+1})-F^*)\end{aligned}$$

where the first inequality is due to $F(X^t)\ge F(X^s)$ (since AFW is an descent algorithm). Through rearrangement we have

$$F(X^{t+1})-F^*\le\frac{1}{1+\kappa}(F(X^t)-F^*)$$

which then leads to the conclusion. □

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

Proof. By lemma 4 and (30), we have

$$\begin{aligned}&\Delta_d^t-\Delta_d^{t-1}+\Delta_p^t-\Delta_p^{t-1}\\ \le&\frac{-\kappa}{1+\kappa}\left(\mathcal{L}(X^t,Y^t)-\mathcal{L}(\bar{X}^t,Y^t)\right)\\&+\frac{2\eta}{\rho}(\mathcal{L}(X^t,Y^t)-\mathcal{L}(\bar{X}^t,Y^t))-\frac{\eta}{R_Y^2}\Delta_d^t.\end{aligned} \tag{39}$$

Then by choosing $\eta<\frac{\kappa\rho}{2(1+\kappa)}$, we have guaranteed descent on $\Delta_p+\Delta_d$ for each GDMM iteration. By choosing $\eta\le\frac{\kappa\rho}{4(1+\kappa)}$, we have

$$\begin{aligned}&(\Delta_d^t+\Delta_p^t)-(\Delta_d^{t-1}+\Delta_p^{t-1})\\ \le&\frac{-\kappa}{2(1+\kappa)}\left(\mathcal{L}(X^t,Y^t)-\mathcal{L}(\bar{X}^t,Y^t)\right)-\frac{\eta}{R_Y^2}\Delta_d^{t\,2}\\ \le&\frac{-\kappa}{2(1+\kappa)}\Delta_p^t-\frac{\kappa\rho}{4(1+\kappa)R_Y^2}\Delta_d^{t\,2}\\ \le&\frac{-\kappa}{2(1+\kappa)(\Delta_p^0+\Delta_d^0)}\Delta_p^{t\,2}-\frac{\kappa\rho}{4(1+\kappa)R_Y^2}\Delta_d^{t\,2}\\ \le&-\left(\frac{\kappa}{4(1+\kappa)}\min\left(\frac{1}{\Delta_p^0+\Delta_d^0},\frac{\rho}{2R_Y^2}\right)\right)(\Delta_p^t+\Delta_d^t)^2\end{aligned}$$

where the third inequality is by non-increasing of $\{\Delta_p^t+\Delta_d^t\}_{t=1}^{\infty}$. Then recursion of the form $\Delta^t-\Delta^{t-1}\le c\Delta^{t2}$ leads to the conclusion. □