7 Proof Roadmap

The key in proving Theorem 1 and 2 is to establish bounds on the primal-dual progress $\Delta_p^t + \Delta_d^t - \Delta_p^{t-1} - \Delta_d^{t-1}$. As intermediate steps, the two lemmas below bound the dual-progress $\Delta_d^t - \Delta_d^{t-1}$ and the primal-progress $\Delta_p^t - \Delta_p^{t-1}$ with respect to the primal variables $\{\boldsymbol{z}^t\}$ and the optimal primal variables $\{\bar{\boldsymbol{z}}^t\}$ at each iteration.

Lemma 1 (Dual Progress)**.** *The dual progress is upper bounded as*

$$\Delta_d^t - \Delta_d^{t-1} \leq -\eta (M\boldsymbol{z}^t)^T (M\bar{\boldsymbol{z}}^t). \tag{14}$$

Lemma 2 (Primal Progress)**.** *The primal progress is upper bounded as*

$$\begin{aligned}\Delta_p^t - \Delta_p^{t-1} \leq &\mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) \\ &+ \eta \|M\boldsymbol{z}^t\|^2 - \eta \langle M\boldsymbol{z}^t, M\bar{\boldsymbol{z}}^t \rangle\end{aligned}$$

By combining results of Lemma 1 and 2, we obtain an intermediate upper bound on the primal-dual progress:

$$\begin{aligned}&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \\ &\leq \eta \|M\boldsymbol{z}^t - M\bar{\boldsymbol{z}}^t\|^2 - \eta \|M\bar{\boldsymbol{z}}^t\|^2 \\ &+ \mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t)\end{aligned} \tag{15}$$

The following four lemmas provide upper bounds on the three sub-terms in (15), i.e., $\|M\boldsymbol{z}^t - M\bar{\boldsymbol{z}}^t\|^2$, $-\eta\|M\bar{\boldsymbol{z}}^t\|^2$, and $\mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t)$, where the bounds on the last term are algorithm-dependent and therefore are tackled by Lemma 5 and Lemma 19 for Algorithm 1 and Algorithm 2 respectively.

Lemma 3.

$$\|M\boldsymbol{z}^t - M\bar{\boldsymbol{z}}^t\|^2 \leq \frac{2}{\rho}(\mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t)). \tag{16}$$

Lemma 4 (*Hong and Luo 2012*)**.** *There is a constant $\tau > 0$ such that*

$$\Delta_d(\boldsymbol{\mu}) \leq \tau \|M\bar{\boldsymbol{z}}(\boldsymbol{\mu})\|^2. \tag{17}$$

for any $\boldsymbol{\mu}$ in the dual domain and any primal minimizer $\bar{\boldsymbol{z}}(\boldsymbol{\mu})$ satisfying (13).

Lemma 5. *The descent amount of Augmented Lagrangian function produced by one pass of FCFW (in Algorithm 1) has*

$$\begin{aligned}&\mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) \\ &\leq -\frac{m_{\mathcal{M}}}{2|\mathcal{F}|Q}(\mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t))\end{aligned} \tag{18}$$

where $Q = \rho\|M\|^2$.

Lemma 6. *The descent amount of Augmented Lagrangian function produced by iterations of Algorithm 2 has*

$$\begin{aligned}&\mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) \\ &\leq \frac{-m_1}{Q_{max}}(\mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t))\end{aligned} \tag{19}$$

where $Q_{max} = \max_{f\in\mathcal{F}} Q_f$ and

$$m_1 := \frac{1}{\max\{16\theta_1 \Delta\mathcal{L}^0, 2\theta_1(1 + 4L_g^2)/\rho, 6\}} \tag{20}$$

is the generalized strong convexity constant *for function $\mathcal{L}(., \boldsymbol{\mu})$. Here $\Delta\mathcal{L}^0$ is a bound on $\mathcal{L}(\boldsymbol{z}^0, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^0, \boldsymbol{\mu}^t)$, L_g is local Lipschitz-continuous constant of the function $g(\boldsymbol{x}) := \|\boldsymbol{x}\|^2$, and θ_1 is the Hoffman constant depending on the geometry of optimal solution set.*

Now we are ready to prove Theorem 1 and 2.

Proof of Theorem 1. Let $\kappa = m_{\mathcal{M}}/(|\mathcal{F}|Q)$. By lemma 5 and (15), we have

$$\begin{aligned}&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \\ \leq &\frac{-\kappa}{1+\kappa}\left(\mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t)\right) \\ &+ \frac{2\eta}{\rho}(\mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t)) - \eta\|M\bar{\boldsymbol{z}}^t\|^2.\end{aligned} \tag{21}$$

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 $\eta \leq \frac{\kappa\rho}{4(1+\kappa)}$, we have

$$\begin{aligned}&(\Delta_d^t + \Delta_p^t) - (\Delta_d^{t-1} + \Delta_p^{t-1}) \\ \leq &\frac{-\kappa}{2(1+\kappa)}\left(\mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t)\right) - \eta\|M\bar{\boldsymbol{z}}^t\|^2 \\ \leq &\frac{-\kappa}{2(1+\kappa)}\Delta_d^t - \frac{\eta}{\tau}\Delta_d^t \\ \leq &-\min\left(\frac{\kappa}{2(1+\kappa)}, \frac{\eta}{\tau}\right)(\Delta_p^t + \Delta_d^t)\end{aligned}$$

where the second inequality is from Lemma 4. We thus obtain a recursion of the form

$$\Delta_d^t + \Delta_p^t \leq \frac{1}{1 + \min(\frac{\kappa}{2(1+\kappa)}, \frac{\eta}{\tau})}\left(\Delta_d^{t-1} + \Delta_p^{t-1}\right),$$

which then leads to the conclusion. □

The proof of Theorem 2 is the same as above except that the definition of κ is changed to m_1/Q_{max} and Lemma 5 is replaced by Lemma 19.

8 Proof of Lemmas

Proof of Lemma 1.

$$\begin{aligned}\Delta_d^t - \Delta_d^{t-1} &= \mathcal{L}(\bar{\boldsymbol{z}}^{t-1}, \boldsymbol{\mu}^{t-1}) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t)\\ &\le \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^{t-1}) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t)\\ &= \langle \boldsymbol{\mu}^{t-1} - \boldsymbol{\mu}^t, M\bar{\boldsymbol{z}}^t\rangle\\ &= -\eta\langle M\boldsymbol{z}^t, M\bar{\boldsymbol{z}}^t\rangle\end{aligned}$$

where the first inequality follows from the optimality of $\bar{\boldsymbol{z}}^{t-1}$ for the function $\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}^{t-1})$ defined by $\boldsymbol{\mu}^{t-1}$, and the last equality follows from the dual update (9). □

Proof of Lemma 2.

$$\begin{aligned}&\Delta_p^t - \Delta_p^{t-1}\\ =&\mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^{t-1}) - (d(\boldsymbol{\mu}^t) - d(\boldsymbol{\mu}^{t-1}))\\ \le&\mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) + \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^{t-1})\\ &+ (d(\boldsymbol{\mu}^{t-1}) - d(\boldsymbol{\mu}^t))\\ \le&\mathcal{L}(\boldsymbol{z}^{t+1}, \boldsymbol{\mu}^t) - \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) + \eta\|M\boldsymbol{z}^t\|^2 - \eta\langle M\boldsymbol{z}^t, M\bar{\boldsymbol{z}}^t\rangle\end{aligned}$$

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

Proof of Lemma 3. Introduce

$$\tilde{\mathcal{L}}(\boldsymbol{z}, \boldsymbol{\mu}) = h(\boldsymbol{z}) + G(M\boldsymbol{z}),$$

where

$$G(M\boldsymbol{z}) = \frac{\rho}{2}\|M\boldsymbol{z}\|^2,$$

and

$$h(\boldsymbol{z}) = \langle -\boldsymbol{\theta}, \boldsymbol{z}\rangle + \langle \boldsymbol{\mu}, M\boldsymbol{z}\rangle + \boldsymbol{I}_{\boldsymbol{z}\in\mathcal{M}}.$$

Here

$$\boldsymbol{I}_{\boldsymbol{z}\in\mathcal{M}} = \begin{cases} 0 & \boldsymbol{z}\in\mathcal{M},\\ \infty & \text{otherwise.}\end{cases}$$

As feasibility is strictly enforced during primal updates, we have

$$\tilde{\mathcal{L}}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t) = \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t), \quad \tilde{\mathcal{L}}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) = \mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t). \tag{22}$$

As $\bar{\boldsymbol{z}}^t$ is a critical point of $\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}^t)$, and by definition, $\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}^t) \le \tilde{\mathcal{L}}(\boldsymbol{z}, \boldsymbol{\mu}^t)$, we obtain,

$$0 \in \partial_{\boldsymbol{z}}\tilde{\mathcal{L}}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t) = \partial h(\bar{\boldsymbol{z}}^t) + M^T\nabla G(M\bar{\boldsymbol{z}}^t).$$

Note that $h(\cdot)$ is convex, it follows that

$$h(\boldsymbol{z}^t) - h(\bar{\boldsymbol{z}}^t) \ge \langle \boldsymbol{v}, \boldsymbol{z}^t - \bar{\boldsymbol{z}}^t\rangle, \qquad \forall \boldsymbol{v}\in\partial h(\bar{\boldsymbol{z}}^t). \tag{23}$$

Moreover,

$$\begin{aligned}&G(M(\boldsymbol{z}^t)) - G(M(\bar{\boldsymbol{z}}^t)) \qquad (24)\\ &= \frac{\rho}{2}(\|M\boldsymbol{z}^t\|^2 - \|M\bar{\boldsymbol{z}}^t\|^2)\\ &= \frac{\rho}{2}(\boldsymbol{z}^t - \bar{\boldsymbol{z}}^t)^T M^T M(\boldsymbol{z}^t + \bar{\boldsymbol{z}}^t)\\ &= \rho(\boldsymbol{z}^t - \bar{\boldsymbol{z}}^t)^T M^T M\bar{\boldsymbol{z}}^t + \frac{\rho}{2}(\boldsymbol{z}^t - \bar{\boldsymbol{z}}^t)^T M^T M(\boldsymbol{z}^t - \bar{\boldsymbol{z}}^t)\\ &= \langle M^T\nabla G(M\bar{\boldsymbol{z}}^t), \boldsymbol{z}^t - \bar{\boldsymbol{z}}^t\rangle + \frac{\rho}{2}\|M\boldsymbol{z}^t - M\bar{\boldsymbol{z}}^t\|^2. \qquad (25)\end{aligned}$$

Combing (22), (23), and (25), we arrive at

$$\mathcal{L}(\boldsymbol{z}^t, \boldsymbol{\mu}^t) - \mathcal{L}(\bar{\boldsymbol{z}}^t, \boldsymbol{\mu}^t) \ge \frac{\rho}{2}\|M(\boldsymbol{z}^t) - M(\bar{\boldsymbol{z}}^t)\|^2.$$

□

Proof of Lemma 4. This is a lemma adapted from [22]. Since our primal objective (2) is a linear function with each block of primal variables $\boldsymbol{x}_i$ (or $\boldsymbol{y}_f$) constrained in a simplex domain, it satisfies the *assumptions A(a)—A(e)* and *A(g)* in [22]. Then Lemma 3.1 of [22] guarantees that, as long as $\|\nabla d(\boldsymbol{\mu})\|$ is always bounded, there is a constant $\tau > 0$ s.t.

$$\Delta_d(\boldsymbol{\mu}) \le \tau\|\nabla d(\boldsymbol{\mu})\|^2 = \|M\bar{\boldsymbol{z}}(\boldsymbol{\mu})\|^2$$

for all $\boldsymbol{\mu}$ in the dual domain. Note our problem satisfies the condition of bounded gradient magnitude since

$$\begin{aligned}&\|\nabla d(\boldsymbol{\mu})\| = \|M\bar{\boldsymbol{z}}(\boldsymbol{\mu})\| \le \|M\bar{\boldsymbol{z}}(\boldsymbol{\mu})\|_1\\ &\le \|M\|_1\|\bar{\boldsymbol{z}}(\boldsymbol{\mu})\|_1 \le (\max_f |\mathcal{Y}_f|)(|\mathcal{F}| + |\mathcal{V}|)\end{aligned}$$

where the last inequality is because each block of variables in $\bar{\boldsymbol{z}}(\boldsymbol{\mu})$ lie in a simplex domain. □

Proof of Lemma 5. Recall that the Augmented Lagrangian $\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu})$ is of the form

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}) = \langle -\boldsymbol{\theta} + M^T\boldsymbol{\mu}, \boldsymbol{z}\rangle + G(M\boldsymbol{z})\ , \forall i\in\mathcal{V} \tag{26}$$

where M is the matrix that encodes all constraints of the form

$$M_{if}\boldsymbol{z}_f - \boldsymbol{z}_i = \begin{bmatrix} M_{if} & -I_i\end{bmatrix}\begin{bmatrix}\boldsymbol{z}_f\\ \boldsymbol{z}_i\end{bmatrix} = \boldsymbol{0}.$$

and function $G(\boldsymbol{w}) = \frac{\rho}{2}\|\boldsymbol{w}\|^2$ is strongly convex with parameter ρ. Let

$$H(\boldsymbol{z}) := \mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}). \tag{27}$$

Since we are minimizing the function subject to a convex, polyhedral domain $\mathcal{M}$, by Theorem 10 of [23], we have the *generalized geometrical strong convexity* constant $m_{\mathcal{M}}$ of the form

$$m_{\mathcal{M}} := m(PWidth(\mathcal{M}))^2 \tag{28}$$

where $PWidth(\mathcal{M}) > 0$ is the pyramidal width of the simplex domain $\mathcal{M}$ and m is the *generalized strong convexity* constant of function (26) (defined by Lemma 9 of [23]). By definition of the geometric strong convexity constant, we have

$$H(\boldsymbol{z}) - H^* \le \frac{g_{FW}^2}{2m_{\mathcal{M}}} \tag{29}$$

from (23) in [23], where $g_{FW} := \langle \nabla H(\boldsymbol{z}), \boldsymbol{v}_{FW} - \boldsymbol{v}_A\rangle$. $\boldsymbol{v}_{FW}$ is the greedy Frank-Wolfe (FW) direction

$$\boldsymbol{v}_{FW} := arg\min_{\boldsymbol{v}\in\mathcal{M}}\langle \nabla H(\boldsymbol{z}), \boldsymbol{v}\rangle \tag{30}$$

and $\boldsymbol{v}_A$ is the away direction

$$\boldsymbol{v}_A := arg \max_{\boldsymbol{v}\in\mathcal{M}} \langle \nabla \tilde{H}(\boldsymbol{z}), \boldsymbol{v}\rangle \tag{31}$$

where

$$\nabla_k \tilde{H}(\boldsymbol{z}) = \begin{cases} \nabla_k H(\boldsymbol{z}), & z_k \neq 0 \\ -\infty, & o.w. \end{cases}$$

Then let $m = |\mathcal{F}|$ be the number of factors. For each inner iteration s of the Fully-Corrective FW, by minimizing subproblem (5) w.r.t. an active set that contains the FW direction and also the away direction (by the definition (31)), we have, for any $\forall \gamma \in [0,1]$,

$$H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \le \gamma g^t_{FW} + mQ\gamma^2. \tag{32}$$

Suppose the minimizer of (32) $\gamma^* = -\frac{g^t_{FW}}{2mQ}$ has $\gamma^* < 1$, we have

$$H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \le -\frac{g^{t2}_{FW}}{4mQ} \tag{33}$$

Otherwise, let $\gamma^* = 1$, we have

$$\begin{aligned} &H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \\ &\le g^t_{FW} + mQ \le \frac{g^t_{FW}}{2} < -\frac{g^{t2}_{FW}}{2mQ} \le -\frac{g^{t2}_{FW}}{4mQ}, \end{aligned}$$

where the second inequality holds since $-\frac{g^t_{FW}}{2Qm} \ge 1$.

Combining with the error bound (29), we have

$$H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \le -\frac{m_{\mathcal{M}}(H(\boldsymbol{z}^t) - H^*)}{2mQ}. \tag{34}$$

□

Proof of Lemma 19.

For problem of the form (13), the optimal solution is profiled by the polyhedral set $\mathcal{S} := \{\boldsymbol{z} \mid M\boldsymbol{z} = \boldsymbol{t}^*,\ \boldsymbol{\Delta}^T\boldsymbol{z} = s^*,\ \boldsymbol{z} \in \mathcal{M}\}$ for some $\boldsymbol{t}^*$, s^*. Denoting $\bar{\boldsymbol{z}} := \Pi_{\mathcal{S}}(\boldsymbol{z})$, we can bound the distance of any feasible point $\boldsymbol{z}$ to its projection $\Pi_{\mathcal{S}}(\boldsymbol{z})$ to set $\mathcal{S}$ by

$$\begin{aligned} &\|\bar{\boldsymbol{z}} - \boldsymbol{z}\|^2_{2,1} = (\sum_{f\in\mathcal{F}} \|\bar{\boldsymbol{z}}_f - \boldsymbol{z}_f\|_2)^2 \\ &\le \theta_1 \left(\|M\boldsymbol{z} - \boldsymbol{t}^*\|^2 + \|\boldsymbol{\Delta}^T\boldsymbol{z} - s^*\|^2\right) \end{aligned} \tag{35}$$

where θ_1 is a constant depending on the set $\mathcal{S}$, using the Hoffman's inequality [37].

Then for each iteration t of the Algorithm 2, consider the descent amount produced by the update w.r.t. the selected factor satisfying (11). We have

$$\begin{aligned} &H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \\ &\le \min_{\boldsymbol{z}^t_{f^*} + \boldsymbol{d}_{f^*}\in\Delta_{f^*}} \langle \nabla_{\boldsymbol{z}_{f^*}} H, \boldsymbol{d}_{f^*}\rangle + \frac{Q_{\max}}{2}\|\boldsymbol{d}_{f^*}\|^2 \\ &= \min_{\boldsymbol{z}^t + \boldsymbol{d}\in\mathcal{M}} \sum_{f\in\mathcal{F}} \langle \nabla_{\boldsymbol{z}_f} H, \boldsymbol{d}_f\rangle + \frac{Q_{\max}}{2}\left(\sum_{f\in\mathcal{F}} \|\boldsymbol{d}_f\|\right)^2 \end{aligned} \tag{36}$$

where the second equality is from the definition (11) of f^*.

Then we have

$$\begin{aligned} &H(\boldsymbol{z}^{t+1})] - H(\boldsymbol{z}^t) \\ &\le \min_{\boldsymbol{z}^t + \boldsymbol{d}\in\mathcal{M}} \left(\sum_{f\in\mathcal{F}} \langle \nabla_{\boldsymbol{z}_f} H, \boldsymbol{d}_f\rangle + \frac{Q_{\max}}{2}\left(\sum_{f\in\mathcal{F}} \|\boldsymbol{d}_f\|\right)^2\right) \\ &\le \min_{\boldsymbol{z}^t + \boldsymbol{d}\in\mathcal{M}} H(\boldsymbol{z}^t + \boldsymbol{d}) - H(\boldsymbol{z}^t) + \frac{Q_{\max}}{2}\left(\sum_{f\in\mathcal{F}} \|\boldsymbol{d}_f\|\right)^2 \\ &\le \min_{\beta\in[0,1]} H(\boldsymbol{z}^t + \beta(\bar{\boldsymbol{z}}^t - \boldsymbol{z}^t)) - H(\boldsymbol{z}^t) \\ &\quad + \frac{Q_{\max}\beta^2}{2}\left(\sum_{f\in\mathcal{F}} \|\bar{\boldsymbol{z}}^t_f - \boldsymbol{z}^t_f\|\right)^2 \\ &\le \min_{\beta\in[0,1]} \beta(H(\bar{\boldsymbol{z}}^t) - H(\boldsymbol{z}^t)) + \frac{Q_{\max}\beta^2}{2}\|\bar{\boldsymbol{z}}^t - \boldsymbol{z}^t\|^2_{2,1} \end{aligned} \tag{37}$$

where $\bar{\boldsymbol{z}}^t = \Pi_{\mathcal{S}}(\boldsymbol{z}^t)$ is the projection of $\boldsymbol{z}^t$ to the optimal solution set $\mathcal{S}$. The second and last inequality is due to convexity, and the third inequality is due to a confinement of optimization domain. Then let L_g be the local Lipschitz-continuous constant of function $G(M\boldsymbol{z}) = \frac{\rho}{2}\|M\boldsymbol{z}\|^2$ in the bounded domain of $M\boldsymbol{z}$. We discuss two cases in the following.

Case 1: $4L_g^2\|M\boldsymbol{z}^t - \boldsymbol{t}^*\|^2 < (\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*)^2$.

In this case, we have

$$\begin{aligned} \|\boldsymbol{z}^t - \bar{\boldsymbol{z}}^t\|^2_{2,1} &\le \theta_1(\|M\boldsymbol{z}^t - \boldsymbol{t}^*\|^2 + (\boldsymbol{\Delta}^T\boldsymbol{z}^s - s^*)^2) \\ &\le \theta_1(\frac{1}{L_g^2} + 1)(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*)^2 \\ &\le 2\theta_1(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*)^2, \end{aligned} \tag{38}$$

and

$$|\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*| \ge 2L_g\|M\boldsymbol{z}^t - \boldsymbol{t}^*\| \ge 2|G(M\boldsymbol{z}^t) - G(\boldsymbol{t}^*)|$$

by the definition of Lipschitz constant L_g. Note $\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*$ is non-negative since otherwise, $H(\boldsymbol{z}^t) - H^* = G(M\boldsymbol{z}^t) - G(\boldsymbol{t}^*) + (\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*) \le |G(M\boldsymbol{z}^t) - G(\boldsymbol{t}^*)| - |\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*| \le -\frac{1}{2}|\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*| < 0$, which leads to contradiction. Therefore, we have

$$\begin{aligned} &H(\boldsymbol{z}^t) - H^* \\ &= G(M\boldsymbol{z}^t) - G(\boldsymbol{t}^*) + (\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*) \\ &\ge -|G(M\boldsymbol{z}^t) - G(\boldsymbol{t}^*)| + (\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*) \\ &\ge \frac{1}{2}(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*). \end{aligned} \tag{39}$$

Combining (37), (38) and (39), we have

$$
\begin{aligned}
&H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \\
&\leq \min_{\beta\in[0,1]} -\frac{\beta}{2}(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*) + \frac{2Q_{max}\theta_1\beta^2}{2}(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*)^2 \\
&= \begin{cases} -1/(16Q_{\max}\theta_1) & , 1/(4\rho\theta_1(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*)) \leq 1 \\ -\frac{1}{4}(\boldsymbol{\Delta}^T\boldsymbol{\alpha}^s - s^*) & , o.w. \end{cases}
\end{aligned}
$$

Furthermore, we have

$$
-\frac{1}{16Q_{max}\theta_1} \leq -\frac{1}{16Q_{max}\theta_1(H^0 - H^*)}\left(H(\boldsymbol{z}^t) - H^*\right)
$$

where $H^0 = H(\boldsymbol{z}^0)$, and

$$
-\frac{1}{4}(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*) \leq -\frac{1}{6}(H(\boldsymbol{z}^t) - H^*)
$$

since $H(\boldsymbol{z}^t) - H^* \leq |G(M\boldsymbol{z}^t) - G(\boldsymbol{t}^*)| + \boldsymbol{\Delta}^T\boldsymbol{z}^t - s^* \leq \frac{3}{2}(\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*)$. In summary, for Case 1 we obtain

$$
H(\boldsymbol{z}^{t+1})] - H^* \leq (1 - \frac{m_0}{Q_{max}})\left(H(\boldsymbol{z}^t) - H^*\right) \quad (40)
$$

where

$$
m_0 = \frac{1}{\max\{16\theta_1(H^0 - H^*)\ ,\ 6\}}. \quad (41)
$$

Case 2: $4L_g^2\|M\boldsymbol{z}^t - \boldsymbol{t}^*\|^2 \geq (\boldsymbol{\Delta}^T\boldsymbol{z}^t - s^*)^2$.

In this case, we have

$$
\|\bar{\boldsymbol{z}}^t - \boldsymbol{z}^t\|^2 \leq \theta_1\left(1 + 4L_g^2\right)\|M\boldsymbol{z}^t - \boldsymbol{t}^*\|^2, \quad (42)
$$

and by strong convexity of $G(.)$,

$$
\begin{aligned}
&H(\boldsymbol{z}^t) - H^* \geq \\
&\boldsymbol{\Delta}^T(\boldsymbol{z}^t - \boldsymbol{z}^*) + \nabla G(t^*)^T M(\bar{\boldsymbol{z}}^t - \boldsymbol{z}^t) + \frac{\rho}{2}\|M\boldsymbol{z}^t - \boldsymbol{t}^*\|^2.
\end{aligned}
$$

Now let $h(\boldsymbol{\alpha})$ be a function that takes value 0 when $\boldsymbol{z}$ is feasible and takes value ∞ otherwise. Adding inequality $0 = h(\boldsymbol{z}^t) - h(\bar{\boldsymbol{z}}^t) \geq \langle \boldsymbol{\sigma}^*, \boldsymbol{z}^t - \bar{\boldsymbol{z}}^t\rangle$ for some $\boldsymbol{\sigma}^* \in \partial h(\bar{\boldsymbol{z}}^t)$ to the above gives

$$
H(\boldsymbol{z}^t) - H^* \geq \frac{\rho}{2}\|M\boldsymbol{z}^t - \boldsymbol{t}^*\|^2 \quad (43)
$$

since $\boldsymbol{\sigma}^* + \boldsymbol{\Delta} + \nabla G(\boldsymbol{t}^*)^T M = \boldsymbol{\sigma}^* + \nabla H(\boldsymbol{z}^t) = 0$. Combining (37), (42), and (43), we obtain

$$
\begin{aligned}
&H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \\
&\leq \min_{\beta\in[0,1]} -\beta(H(\boldsymbol{z}^t) - H^*) + \frac{\theta_1(1 + 4L_g^2)Q_{max}\beta^2}{2\rho}\left(H(\boldsymbol{z}^t) - H^*\right) \\
&= -\frac{\rho}{2\theta_1(1 + 4L_g^2)Q_{max}}\left(H(\boldsymbol{z}^t) - H^*\right)
\end{aligned}
\quad (44)
$$

Combining results of Case 1 (40) and Case 2 (44), we have

$$
H(\boldsymbol{z}^{t+1}) - H(\boldsymbol{z}^t) \leq -\frac{m_1}{Q_{\max}}(H(\boldsymbol{z}^t) - H^*), \quad (45)
$$

where

$$
m_1 = \frac{1}{\max\{16\theta_1\Delta\mathcal{L}^0, 2\theta_1(1 + 4L_g^2)/\rho, 6\}}
$$

This leads to the conclusion.

□

9 Active set size statistics for all experiments

Dataset	$\|\mathcal{F}\|$	$\mathbb{E}_t\|\mathcal{A}_{\mathcal{F}}^t\|$
MultiLabel	7544670	6128.2
Dataset	$\|\mathcal{Y}_f\|$	$\mathbb{E}_{t,f}\|\mathcal{A}_f^t\|$
Segmentation	441	4.9
ImageAlignment	6889	2.4
Protein	163216	12.7
GraphMatching	1069156	1.66

Table 3: Run time statistics for GDMM active set. For multilabel dataset, we use Algorithm 2, thus $|\mathcal{F}|$ and $\mathbb{E}_t|\mathcal{A}_{\mathcal{F}}^t|$ are compared, where $\mathbb{E}_t|\mathcal{A}_{\mathcal{F}}^t|$ is the expected size of $\mathcal{A}_{\mathcal{F}}$ over all iterations. For other datasets, we use Algorithm 1, thus $|\mathcal{Y}_f|$ and $\mathbb{E}_{t,f}|\mathcal{A}_f^t|$ are compared, the latter is the expected size of $\mathcal{A}_f$ over all iterations and bigram factors.