7 Proof Roadmap

The key in proving Theorem 1 and 2 is to establish bounds on the primal-dual progress $\Delta_d^t + \Delta_p^t - \Delta_p^{t-1} - \Delta_d^{t-1}$. 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 $\{z^t\}$ and the optimal primal variables $\{\bar{z}^t\}$ at each iteration.

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

\[
\Delta_d^t - \Delta_d^{t-1} \leq -\eta(Mz^t)^T(M\bar{z}^t). \quad (14)
\]

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

\[
\Delta_p^t - \Delta_p^{t-1} \leq \mathcal{L}(z^{t+1}, \mu^t) - \mathcal{L}(z^t, \mu^t) + \eta\|Mz^t\|^2 - \eta\langle Mz^t, M\bar{z}^t \rangle.
\]

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

\[
\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \leq \eta\|Mz^t - M\bar{z}^t\|^2 - \eta\|M\bar{z}^t\|^2 \quad (15)
\]

The following four lemmas provide upper bounds on the three sub-terms in (15), i.e., $\|Mz^t - M\bar{z}^t\|^2$, $-\eta\|M\bar{z}^t\|^2$, and $\mathcal{L}(z^{t+1}, \mu^t) - \mathcal{L}(z^t, \mu^t)$, 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.**

\[
\|Mz^t - M\bar{z}^t\|^2 \leq \frac{2}{\rho}(\mathcal{L}(z^t, \mu^t) - \mathcal{L}(\bar{z}^t, \mu^t)). \quad (16)
\]

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

\[
\Delta_d(\mu) \leq \tau\|M\bar{z}(\mu)\|^2. \quad (17)
\]

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

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

\[
\mathcal{L}(z^{t+1}, \mu^t) - \mathcal{L}(z^t, \mu^t) \leq -\frac{m_M}{2|\mathcal{F}|Q}(\mathcal{L}(z^t, \mu^t) - \mathcal{L}(\bar{z}^t, \mu^t)) \quad (18)
\]

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

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

\[
\mathcal{L}(z^{t+1}, \mu^t) - \mathcal{L}(z^t, \mu^t) \leq -\frac{m_1}{Q_{\max}}(\mathcal{L}(z^t, \mu^t) - \mathcal{L}(\bar{z}^t, \mu^t)) \quad (19)
\]

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

\[
m_1 := \frac{1}{\max\{16\theta_1 \Delta \mathcal{L}^0, 2\theta_1 (1+4L^2_\mathcal{F})/\rho, 6\}} \quad (20)
\]

is the generalized strong convexity constant for function $\mathcal{L}(\cdot, \mu)$. Here $\mathcal{L}^0$ is a bound on $\mathcal{L}(z^0, \mu^t) - \mathcal{L}(z^0, \mu^t)$, $L_\mathcal{F}$ is local Lipschitz-continuous constant of the function $g(x) := \|x\|^2$, and $\theta_1$ is the Hoffmann constant depending on the geometry of optimal solution set.

Now we are ready to prove Theorem 1 and 2.

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

\[
\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \leq -\frac{\kappa}{1+\kappa}(\mathcal{L}(z^t, \mu^t) - \mathcal{L}(\bar{z}^t, \mu^t)) + \frac{2\eta}{\rho}(\mathcal{L}(z^t, \mu^t) - \mathcal{L}(\bar{z}^t, \mu^t)) - \eta\|M\bar{z}^t\|^2. \quad (21)
\]

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

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

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

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

which then leads to the conclusion.

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

Proof of Lemma 1.
\[
\Delta_{t} - \Delta_{t}^{\dagger} = \mathcal{L}(z^{t-1}, \mu^{t-1}) - \mathcal{L}(z^{t}, \mu^{t}) \\
\leq \mathcal{L}(\hat{z}^{t}, \mu^{t}) - \mathcal{L}(\gamma^{t}, \mu^{t}) \\
= \langle \mu^{t-1}, Mz^{t} \rangle \\
= -\eta \langle Mz^{t}, M\hat{z}^{t} \rangle
\]
where the first inequality follows from the optimality of $z^{t-1}$ for the function $\mathcal{L}(z, \mu^{t-1})$ defined by $\mu^{t-1}$, and the last equality follows from the dual update (9).

Proof of Lemma 2.
\[
\Delta_{p}^{t} - \Delta_{p}^{t-1} = \mathcal{L}(z^{t+1}, \mu^{t}) - \mathcal{L}(z^{t}, \mu^{t}) - (d(\mu^{t}) - d(\mu^{t-1})) \\
\leq \mathcal{L}(z^{t+1}, \mu^{t}) - \mathcal{L}(z^{t}, \mu^{t}) + \mathcal{L}(z^{t+1}, \mu^{t}) - \mathcal{L}(z^{t}, \mu^{t-1}) \\
+ (d(\mu^{t-1}) - d(\mu^{t})) \\
\leq \mathcal{L}(z^{t+1}, \mu^{t}) - \mathcal{L}(z^{t}, \mu^{t}) + \eta \langle Mz^{t}, M\gamma^{t} \rangle
\]
where the last inequality uses Lemma 1 on $d(\mu^{t-1}) - d(\mu^{t}) = \Delta_{p}^{t} - \Delta_{p}^{t-1}$.

Proof of Lemma 3. Introduce
\[
\tilde{\mathcal{L}}(z, \mu) = h(z) + G(Mz),
\]
where
\[
G(Mz) = \frac{\rho}{2} \|Mz\|^2,
\]
and
\[
h(z) = \langle -\theta, z \rangle + \langle \mu, Mz \rangle + I_{z \in M}.
\]
Here
\[
I_{z \in M} = \begin{cases} 
0 & z \in M, \\
\infty & \text{otherwise}.
\end{cases}
\]
As feasibility is strictly enforced during primal updates, we have
\[
\tilde{\mathcal{L}}(z^{t}, \mu^{t}) = \mathcal{L}(z^{t}, \mu^{t}) - \mathcal{L}(z^{t}, \mu^{t}) = \mathcal{L}(z^{t}, \mu^{t}).
\]
As $z^{t}$ is a critical point of $\mathcal{L}(z, \mu^{t})$, and by definition, $\mathcal{L}(z, \mu^{t}) \leq \tilde{\mathcal{L}}(z, \mu^{t})$, we obtain,
\[
0 \in \partial_{z} \tilde{\mathcal{L}}(z^{t}, \mu^{t}) = \partial h(z^{t}) + M^{T} \nabla G(Mz^{t}).
\]
Note that $h(\cdot)$ is convex, it follows that
\[
h(z^{t}) - h(z^{t}) \geq \langle v, z^{t} - z^{t} \rangle, \quad \forall v \in \partial h(z^{t}).
\]
Moreover,
\[
G(Mz^{t}) - G(Mz^{t}) \\
= \frac{\rho}{2} \|Mz^{t}\|^2 - \|M\hat{z}^{t}\|^2 \\
= \frac{\rho}{2} (z^{t} - \hat{z}^{t})^{T} M^{T} M (z^{t} - \hat{z}^{t}) \\
= \rho (z^{t} - \hat{z}^{t})^{T} M^{T} M z^{t} + \rho (z^{t} - \hat{z}^{t})^{T} M^{T} M (z^{t} - \hat{z}^{t}) \\
= \langle M^{T} \nabla G(Mz^{t}), z^{t} - \hat{z}^{t} \rangle + \frac{\rho}{2} \|Mz^{t} - M\hat{z}^{t}\|^2.
\]
Combing (22), (23), and (25), we arrive at
\[
\mathcal{L}(z^{t}, \mu^{t}) - \mathcal{L}(z^{t}, \mu^{t}) \geq \frac{\rho}{2} \|M(z^{t}) - M(\hat{z}^{t})\|^2.
\]

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 $x_{i}$ (or $y_{j}$) constrained in a simplex domain, it satisfies the assumptions $A(a) - A(c)$ and $A(g)$ in [22]. Then Lemma 3.1 of [22] guarantees that, as long as $\|d(\mu)\|$ is always bounded, there is a constant $\tau > 0$ s.t.
\[
\Delta_{d}(\mu) \leq \tau \|\nabla d(\mu)\|^2 = \|M\hat{z}(\mu)\|^2
\]
for all $\mu$ in the dual domain. Note our problem satisfies the condition of bounded gradient magnitude since
\[
\|\nabla d(\mu)\| = \|M\hat{z}(\mu)\| \leq \|M\hat{z}(\mu)\|_{1} \leq \|M\|_{1} \|\hat{z}(\mu)\|_{1} \leq (\max_{i} |Y_{i}|)(|F_{i}| + |V|)
\]
where the last inequality is because each block of variables in $\hat{z}(\mu)$ lie in a simplex domain.

Proof of Lemma 5. Recall that the Augmented Lagrangian $\mathcal{L}(z, \mu)$ is of the form
\[
\mathcal{L}(z, \mu) = \langle -\theta + M^{T}, \mu, z \rangle + G(Mz), \forall \mu \in V
\]
where $M$ is the matrix that encodes all constraints of the form
\[
M_{if} z_{f} - z_{i} = \begin{bmatrix} M_{if} & -I_{i} \end{bmatrix} \begin{bmatrix} z_{f} \\
\hat{z}_{i} \end{bmatrix} = 0.
\]
and function $G(w) = \frac{\rho}{2} \|w\|^2$ is strongly convex with parameter $\rho$. Let
\[
H(z) := \mathcal{L}(z, \mu).
\]
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
\[
m_{\mathcal{M}} := m_{PW\text{idth}(\mathcal{M})}^{2}
\]
where $PW\text{idth}(\mathcal{M}) > 0$ 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
\[
H(z) - H^{\star} \leq \frac{g_{FW}^{2}}{2m_{\mathcal{M}}}
\]
from (23) in [23], where $g_{FW} := \langle \nabla H(z), v_{FW} - v_{A} \rangle$. $v_{FW}$ is the greedy Frank-Wolfe (FW) direction
\[
v_{FW} := \arg\min_{v \in \mathcal{M}} \langle \nabla H(z), v \rangle
\]
and \( v_A \) is the away direction
\[
v_A := \arg \max_{v \in \mathcal{M}} \langle \nabla \tilde{H}(z), v \rangle
\] (31)
where
\[
\nabla \tilde{H}(z) = \begin{cases} \nabla H(z), & z_k \neq 0 \\ -\infty, & \text{o.w.} \end{cases}
\]

Then let \( m = |\mathcal{F}| \) be the number of factors. For each inner iteration \( t \) 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 \( \forall \gamma \in [0, 1], \)
\[
H(z^{t+1}) - H(z^t) \leq \gamma g_{FW} + mQ \gamma^2.
\] (32)

Suppose the minimizer of (32) \( \gamma^* = -\frac{g_{FW}}{2mQ} \) has \( \gamma^* < 1 \), we have
\[
H(z^{t+1}) - H(z^t) \leq -\frac{g_{FW}^2}{4mQ}.
\] (33)

Otherwise, let \( \gamma^* = 1 \), we have
\[
H(z^{t+1}) - H(z^t) \leq g_{FW} + mQ \leq \frac{g_{FW}}{2} \leq \frac{g_{FW}^2}{4mQ}
\]
where the second inequality holds since \( \frac{g_{FW}}{2mQ} \geq 1 \).

Combining with the error bound (29), we have
\[
H(z^{t+1}) - H(z^t) \leq -\frac{mM(H(z^t) - H^*)}{2mQ}
\] (34)

Proof of Lemma 19.

For problem of the form (13), the optimal solution is profiled by the polyhedral set \( S := \{ z \mid Mz = t^*, \Delta^Tz = s^*, z \in \mathcal{M} \} \) for some \( t^*, s^* \). Denoting \( \bar{z} := \Pi_S(z) \), we can bound the distance of any feasible point \( z \) to its projection \( \Pi_S(z) \) to set \( S \) by
\[
\|z - \bar{z}\|_2^2 = (\sum_{j \in \mathcal{F}} \|z_j - \bar{z}_j\|)^2 \leq \theta_1 (\|Mz - t^*\|^2 + \|\Delta^Tz - s^*\|^2)
\] (35)
where \( \theta_1 \) is a constant depending on the set \( S \), 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
\[
H(z^{t+1}) - H(z^t) \leq \min_{z^t + d \in S} \langle \nabla z_t, H, d \rangle + \frac{Q_{\max}}{2} \|d\|_2^2
\] (36)

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

Then we have
\[
H(z^{t+1}) - H(z^t) \leq \min_{z^t + d \in \mathcal{M}} \left( \sum_{j \in \mathcal{F}} \langle \nabla z_j, H_j, d_j \rangle + \frac{Q_{\max}}{2} \|d_j\|_2^2 \right)
\]
\[
\leq \min_{z^t + d \in \mathcal{M}} \left( H(z^t + d) - H(z^t) + \frac{Q_{\max}}{2} \|d\|_2^2 \right)
\]
\[
\leq \min_{\beta \in [0, 1]} \left( H(z^t + \beta(z^t - z^*)) - H(z^t) \right)
\]
\[
+ \frac{Q_{\max}}{2} \|z^t - z^*\|_2^2
\]
\[
\leq \min_{\beta \in [0, 1]} \beta (H(z^t) - H(z^t)) + \frac{Q_{\max}}{2} \|z^t - z^*\|_2^2
\] (37)

where \( \bar{z} := \Pi_S(z^t) \) is the projection of \( z^t \) to the optimal solution set \( S \). 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 \( G(Mz) = \frac{\theta}{2} \|Mz\|^2 \) in the bounded domain of \( Mz \). We discuss two cases in the following.

Case 1: \( 4L_g^2 \|Mz^t - t^*\|^2 < (\Delta^Tz^t - s^*)^2 \).

In this case, we have
\[
\|z^t - z^*\|_2^2 \leq \theta_1 (\|Mz^t - t^*\|^2 + (\Delta^Tz^t - s^*)^2)
\]
\[
\leq \theta_1 \left( \frac{1}{L_g^2} + 1 \right) (\Delta^Tz^t - s^*)^2
\]
\[
\leq 2\theta_1 (\Delta^Tz^t - s^*)^2,
\] (38)

and
\[
|\Delta^Tz^t - s^*| \geq 2L_g \|Mz^t - t^*\| \geq 2|G(Mz^t) - G(t^*)|
\]

by the definition of Lipschitz constant \( L_g \). Note \( \Delta^Tz^t - s^* \) is non-negative since otherwise, \( H(z^t) - H^* = G(Mz^t) - G(t^*) + (\Delta^Tz^t - s^*) \leq |G(Mz^t) - G(t^*)| - |\Delta^Tz^t - s^*| < -\frac{\theta}{4} |\Delta^Tz^t - s^*| < 0 \), which leads to contradiction. Therefore, we have
\[
H(z^t) - H^*
\]
\[
= G(Mz^t) - G(t^*) + (\Delta^Tz^t - s^*)
\]
\[
\geq -G(Mz^t) - G(t^*) + (\Delta^Tz^t - s^*)
\]
\[
\geq \frac{1}{2} (\Delta^Tz^t - s^*).
\] (39)
Combining (37), (38) and (39), we have
\[
H(z^{t+1}) - H(z^t) \\
\leq \min_{\beta \in [0,1]} -\frac{\beta}{2} (\Delta^T z^t - s^*)^2 + \frac{2Q_{\max} \theta_1 \beta^2}{2} (\Delta^T z^t - s^*)^2 \\
= \begin{cases} \\
-1/(16Q_{\max} \theta_1), & 1/(4\theta_1(\Delta^T z^t - s^*)) \leq 1 \\
-\frac{1}{4}(\Delta^T z^t - s^*), & \text{o.w.} \\
\end{cases}
\]
Furthermore, we have
\[
\frac{1}{4}(\Delta^T z^t - s^*) \leq \frac{1}{6}(H(z^t) - H^*)
\]
since \( H(z^t) - H^* \leq |G(M z^t) - G(t^*)| + \Delta^T z^t - s^* \leq \frac{1}{2}(\Delta^T z^t - s^*). \) In summary, for Case 1 we obtain
\[
H(z^{t+1}) - H^* \leq (1 - \frac{m_0}{Q_{\max}}) (H(z^t) - H^*)
\]
where
\[
m_0 = \frac{1}{\max \{16\theta_1(H^0 - H^*), 6\}}.
\]

**Case 2:** \( 4L_g^2 \| M z^t - t^* \|^2 \geq (\Delta^T z^t - s^*)^2. \)
In this case, we have
\[
\| \tilde{z}^t - z^t \|^2 \leq \theta_1 (1 + 4L_g^2) \| M z^t - t^* \|^2,
\]
and by strong convexity of \( G(z) \),
\[
H(z^t) - H^* \geq \\
\Delta^T (z^t - z^*) + \nabla G(t^*)^T M(z^t - z^*), + \frac{\rho}{2} \| M(z^t - t^*) \|^2.
\]
Now let \( h(\alpha) \) be a function that takes value 0 when \( z \) is feasible and takes value \( \infty \) otherwise. Adding inequality \( 0 = h(z^t) - h(\tilde{z}^t) \geq \langle \sigma^*, z^t - \tilde{z}^t \rangle \) for some \( \sigma^* \in \partial h(\tilde{z}^t) \) to the above gives
\[
H(z^t) - H^* \geq \frac{\rho}{2} \| M z^t - t^* \|^2
\]
since \( \sigma^* + \Delta + \nabla G(t^*)^T M = \sigma^* + \nabla H(z^t) = 0 \). Combining (37), (42), and (43), we obtain
\[
H(z^{t+1}) - H(z^t) \\
\leq \min_{\beta \in [0,1]} -\beta(H(z^t) - H^*) + \theta_1 (1 + 4L_g^2)Q_{\max} \beta^2 \\
= -\frac{\rho}{2\theta_1 (1 + 4L_g^2)Q_{\max}} (H(z^t) - H^*)
\]
Combining results of Case 1 (40) and Case 2 (44), we have
\[
H(z^{t+1}) - H(z^t) \leq -\frac{m_1}{Q_{\max}} (H(z^t) - H^*),
\]
where
\[
m_1 = \frac{1}{\max \{16\theta_1 \Delta^2 \| z \|^2, 2\theta_1 (1 + 4L_g^2)/\rho, 6\}}
\]
This leads to the conclusion.

### 9 Active set size statistics for all experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th></th>
<th>[F]</th>
<th>[\mathcal{A}_f]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiLabel</td>
<td></td>
<td>7544670</td>
<td>6128.2</td>
</tr>
<tr>
<td>Dataset</td>
<td>[\mathcal{Y}_f]</td>
<td>[\mathcal{E}_f,</td>
<td>[\mathcal{A}_f]</td>
</tr>
<tr>
<td>Segmentation</td>
<td></td>
<td>441</td>
<td>4.9</td>
</tr>
<tr>
<td>ImageAlignment</td>
<td></td>
<td>6889</td>
<td>2.4</td>
</tr>
<tr>
<td>Protein</td>
<td></td>
<td>163216</td>
<td>12.7</td>
</tr>
<tr>
<td>GraphMatching</td>
<td></td>
<td>1069156</td>
<td>1.66</td>
</tr>
</tbody>
</table>

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