7 Appendix A: Analysis for Approximate Low-rank Representation

Theorem 1. *Suppose the low-rank approximation* $\hat{\mathbb{D}}$ *has* $\|\hat{\mathbb{D}} - \mathbb{D}\|_{1,\infty} \leq \epsilon$*, where* $\|A\|_{1,\infty} = \sum_i \max_j |A_{ij}|$*. Then we have*

$$tr(\mathbb{D}^T \hat{W}) \leq tr(\mathbb{D}^T W^*) + 2\epsilon, \tag{25}$$

where $\hat{W}$*,* W^* *are solutions to* (2) *with dissimilarity matrix* $\hat{\mathbb{D}}$ *and* $\mathbb{D}$ *respectively.*

Proof. Since $\hat{W}$ is the optimal solution of (2) with matrix $\hat{\mathbb{D}}$, we have

$$tr(\hat{\mathbb{D}}^T \hat{W}) \leq tr(\hat{\mathbb{D}}^T W^*).$$

Therefore,

$$\begin{aligned}
&tr(\mathbb{D}^T \hat{W}) - tr(\mathbb{D}^T W^*) \leq \\
&\leq tr(\mathbb{D}^T (\hat{W} - W^*)) \\
&= tr(\hat{\mathbb{D}}^T (\hat{W} - W^*)) + tr((\mathbb{D} - \hat{\mathbb{D}})^T (\hat{W} - W^*)) \\
&\leq 0 + tr((\mathbb{D} - \hat{\mathbb{D}})^T (\hat{W} - W^*)) \\
&\leq \sum_i \left(\max_j |\mathbb{D}_{ij} - \hat{\mathbb{D}}_{ij}| \right) \left(\sum_j |\hat{W}_{ij} - W^*_{ij}| \right) \\
&\leq \left(\sum_i \max_j |\mathbb{D}_{ij} - \hat{\mathbb{D}}_{ij}| \right) \left(\max_i \sum_j |\hat{W}_{ij} - W^*_{ij}| \right) \\
&= \|\mathbb{D} - \hat{\mathbb{D}}\|_{1,\infty} * \|\hat{W}_{ij} - W^*_{ij}\|_{max,1} \\
&\leq \|\mathbb{D} - \hat{\mathbb{D}}\|_{1,\infty} \left(\|\hat{W}_{ij}\|_{max,1} + \|W^*_{ij}\|_{max,1} \right) \leq 2\epsilon.
\end{aligned}$$

Note we have used notation $\|A\|_{max,1} = \max_i \sum_j |A_{ij}|$ to distinguish from $\|A\|_{\infty,1} = \sum_j \max_i |A_{ij}|$. □

Theorem 2. *Suppose the low-rank approximation* $\hat{\mathbb{D}}$ *has* $\|\hat{\mathbb{D}} - \mathbb{D}\|_{1,\infty} \leq \epsilon$*, where* $\|A\|_{1,\infty} = \sum_i \max_j |A_{ij}|$*. Then we have*

$$F(\hat{W}) \leq F(W^*) + 2\epsilon, \tag{26}$$

where $F(.)$ *denotes the objective function in* (3)*,* $\hat{W}$*,* W^* *are solutions to* (3) *with dissimilarity matrix* $\hat{\mathbb{D}}$ *and* $\mathbb{D}$ *respectively.*

Proof. Since $\hat{W}$ is the optimal solution of (2) with matrix $\hat{\mathbb{D}}$, we have

$$tr(\hat{\mathbb{D}}^T \hat{W}) + \lambda \|\hat{W}\|_{\infty,1} \leq tr(\hat{\mathbb{D}}^T W^*) + \lambda \|W^*\|_{\infty,1},$$

and thus,

$$\begin{aligned}
&F(\hat{W}) - F(W^*) \leq \\
&\leq tr(D^T (\hat{W} - W^*)) + \|\hat{W}\|_{\infty,1} - \|W^*\|_{\infty,1} \\
&= tr(\hat{\mathbb{D}}^T (\hat{W} - W^*)) + \|\hat{W}\|_{\infty,1} - \|W^*\|_{\infty,1} \\
&+ tr((\mathbb{D} - \hat{\mathbb{D}})^T (\hat{W} - W^*)) \\
&\leq 0 + tr((\mathbb{D} - \hat{\mathbb{D}})^T (\hat{W} - W^*)) \leq 2\epsilon,
\end{aligned}$$

where the final inequality follows from the same reasoning in proof of Theorem 1. □

8 Appendix B: Convergence Analysis of AL-BCD

In this section, we give an analysis that shows global linear convergence of both Algorithm 1 and Algorithm 2. In the first part, we show a linear-type convergence of randomized BCD and Greedy BCD by utilizing the special structure of the AL subproblem. In the second part, we show that by choosing a sufficiently small dual step size η, one can achieve global linear convergence with one step of randomized or Greedy BCD instead of solving each AL subproblem exactly.

8.1 Iteration Complexity of Block Coordinate Descent (BCD)

In this section, we give iteration complexity of inner loop in Algorithm 1 and Algorithm 2 when performed on sub-problem (10). Although (10) is not strongly convex, it has strong convexity when restricted to the subspace $\mathcal{N}^\perp$, where $\mathcal{N}$ is the Nullspace of linear equality $W\mathbf{1} = \mathbf{1}$. This property can be leveraged to prove exponentially fast convergence [31, 33]. In the following, we prove linear convergence of randomized BCD and Greedy BCD for our AL sub-problem (10).

Notice that the AL sub-problem (10) can be expressed as

$$\begin{aligned}
\min_{W,\boldsymbol{\xi}} \quad & \sum_i^N \sum_{j=1}^M \mathbb{D}_{ij} W_{ij} + \sum_{j=1}^M \lambda_j \xi_j + \frac{\rho}{2} \| \sum_{j=1}^M W_j - \boldsymbol{q} \|^2 \\
s.t. \quad & W_{ij} \leq \xi_j \ , \ i \in [N], \\
& W \in [0,1]^{N \times M} \ , \ \boldsymbol{\xi} \in [0,1]^M.
\end{aligned} \tag{27}$$

where $\boldsymbol{q} = \mathbf{1} - \boldsymbol{\alpha}^t / \rho$, and (27) can be further compactly expressed as

$$\min_{x \in \mathcal{X}} \ f(\boldsymbol{x}) = \boldsymbol{b}^T \boldsymbol{x} + g(E\boldsymbol{x}), \tag{28}$$

where $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M]$, $\boldsymbol{x}_j = (W_j, \xi_j)$, $g(E\boldsymbol{x}) = \frac{\rho}{2} \| \sum_{j=1}^M W_j - \boldsymbol{q} \|^2$ and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 ... \times \mathcal{X}_M$ is a

polyhedral set with $\mathcal{X}_j$ defined by the constraints that involves j-th block of variables $\boldsymbol{x}_j = (W_j, \xi_j)$ in (27).

For problem of form (28), [31] shows that the set of optimal solution $\bar{\boldsymbol{x}}$ forms a polyhedron satisfying (i) $E\bar{\boldsymbol{x}} = t^*$, (ii) $\boldsymbol{b}^T\bar{\boldsymbol{x}} = s^*$ and (iii) $\boldsymbol{x} \in \mathcal{X}$. Then we can bound the distance of any point $\bar{\boldsymbol{x}}$ to the optimal polyhedral set by the amount of infeasibility to the three (in)equalities (i)-(iii) using Hoffman's bound introduced as follows.

Lemma 1 (Hoffman's Bound)**.** *Let $\mathcal{S} = \{\boldsymbol{x} \in \mathbb{R}^d \mid A\boldsymbol{x} \leq \boldsymbol{b},\ E\boldsymbol{x} = \boldsymbol{c}\}$ be a polyhedral set. Then for any point $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\boldsymbol{x} - \Pi_{\mathcal{S}}(\boldsymbol{x})\|_2^2 \leq \theta \left\| \begin{array}{c} [A\boldsymbol{x} - \boldsymbol{b}]_+ \\ E\boldsymbol{x} - \boldsymbol{c} \end{array} \right\|_2^2 \tag{29}$$

where $\Pi_{\mathcal{S}}(\boldsymbol{x}) = arg\min_{\boldsymbol{y}\in\mathcal{S}} \|\boldsymbol{y} - \boldsymbol{x}\|$ is the projection of $\boldsymbol{x}$ to the set $\mathcal{S}$, and $\theta > 0$ is a constant depending on the polyhedral set $\mathcal{S}$.

Proof. The Hoffman's bound first appears in [10] and a proof for the ℓ_2-norm's version (29) and the definition of the constant $\theta(\mathcal{S})$ can be found in [31] (lemma 4.3). □

Using Lemma 1, we can bound the distance of any point $\bar{\boldsymbol{x}}$ to the optimal polyhedral set formed by the above conditions (i)-(iii) as

$$\|\boldsymbol{x} - \Pi_{\mathcal{S}}(\boldsymbol{x})\|_2^2 \leq \theta(\|E\boldsymbol{x} - t^*\|^2 + (\boldsymbol{b}^T\boldsymbol{x} - s^*)^2). \tag{30}$$

Note by norm equivalence, we can also have another version of the error bound (30)

$$\left(\sum_{j=1}^{M} \|\boldsymbol{x}_j - \Pi_{\mathcal{S}}(\boldsymbol{x})_j\|_2\right)^2 \leq \theta_1(\|E\boldsymbol{x} - t^*\|^2 + (\boldsymbol{b}^T\boldsymbol{x} - s^*)^2). \tag{31}$$

for some θ_1 satisfying $\theta \leq \theta_1 \leq M\theta$.

Equipped with those bounds on the minimum distance to an optimal solution, we are ready to show a linear type of convergence for both randomized and greedy BCD algorithms.

Both algorithms work by optimizing a column of variables $\boldsymbol{x}_j = (W_j, \xi_j)$ at a time. In particular, since the constraints in (27) are column-separable, equation (17) is derived as the closed-form solution that minimizes (27) w.r.t. (W_j, ξ_j). The update can be thus denoted as $\boldsymbol{x}^{s+1} - \boldsymbol{x}^s$, where $\boldsymbol{x}^{s+1}$ has $W_j^{s+1} = W_j^s + \boldsymbol{d}_j^*$ as defined in (17), $\xi_j^{s+1} = \|W_j^{s+1}\|_\infty$, and all other variables kept the same as $\boldsymbol{x}^s$.

The minimization of (27) w.r.t. j-th block of coordinates yields the following problem

$$\begin{aligned} \min_{W_j\in[0,1]^N, \xi_j\in[0,1]} \quad & f_j(\boldsymbol{x}_j) = \mathbb{D}_j^T W_j + \lambda_j \xi_j \\ & \quad + \rho(\sum_{k=1}^{M} W_k - \boldsymbol{q})^T W_j + \frac{\rho}{2}\|W_j\|^2 \\ s.t. \quad & W_{ij} \leq \xi_j\ ,\ i \in [N], \end{aligned} \tag{32}$$

whose objective has Lipschitz-continuous gradient with modulus ρ. Therefore, denoting $\Delta\boldsymbol{x}_j = \boldsymbol{x}_j^{s+1} - \boldsymbol{x}_j^s$, we have

$$\begin{aligned} & f_j(\boldsymbol{x}_j^{s+1}) - f_j(\boldsymbol{x}_j^s) \\ & \leq \min_{\Delta\boldsymbol{x}_j}\ h_j(\boldsymbol{x}_j^s + \Delta\boldsymbol{x}_j) + \nabla_j f(\boldsymbol{x}^s)^T\Delta\boldsymbol{x}_j + \frac{\rho}{2}\|\Delta\boldsymbol{x}_j\|^2, \end{aligned} \tag{33}$$

where

$$h_j(\boldsymbol{x}_j) = \begin{cases} 0 & , \boldsymbol{x}_j \in \mathcal{X}_j, \\ \infty & , o.w.. \end{cases} \tag{34}$$

Then the following Theorem gives linear convergence of Randomized BCD (Algorithm 1) by showing that the RHS of (33) has magnitude as large as a constant multiple of suboptimality $f(\boldsymbol{x}^s) - f^*$ in expectation when j is drawn uniformly from $[M]$.

Theorem 7 (Linear Convergence of Randomized BCD)**.** *The iterate $\{\boldsymbol{x}^s\}_{s=1}^{\infty}$ produced by Algorithm 1 has*

$$\mathbb{E}[f(\boldsymbol{x}^{s+1})] - f^* \leq \left(1 - \frac{1}{M\gamma}\right)(f(\boldsymbol{x}^s) - f^*).$$

where f^* is the optimum of (10) and

$$\gamma = \max\left\{16\rho\theta(f^0 - f^*)\ ,\ 2\theta(1 + 4L_g^2)\ ,\ 6\right\}$$

is a constant depending on the initial function difference $f^0 - f^*$, local Lipschitz-continuous constant L_g of the augmented term, and Hoffman constant θ of the optimal (polyhedral) solution set.

Proof. Let $\boldsymbol{x} = \boldsymbol{x}^s$ be s-th iterate and $\boldsymbol{x}^* = \Pi_{\mathcal{S}}(\boldsymbol{x})$ be the projection of $\boldsymbol{x}^s$ on the optimal solution set, and $h(\boldsymbol{x}) = \sum_{j=1}^{M} h_j(\boldsymbol{x}_j)$. Taking expectation w.r.t. j, the descent amount given by (33) has

$$\begin{aligned} & \mathbb{E}[f(\boldsymbol{x}^{s+1})] - f(\boldsymbol{x}^s) \\ & \leq \frac{1}{M}\left(\min_{\Delta\boldsymbol{x}} h(\boldsymbol{x} + \Delta\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x} + \frac{\rho}{2}\|\Delta\boldsymbol{x}\|^2\right) \\ & \leq \frac{1}{M}\left(\min_{\Delta\boldsymbol{x}} h(\boldsymbol{x} + \Delta\boldsymbol{x}) + f(\boldsymbol{x} + \Delta\boldsymbol{x}) - f(\boldsymbol{x}) + \frac{\rho}{2}\|\Delta\boldsymbol{x}\|^2\right) \\ & \leq \frac{1}{M}\left(\min_{\beta\in[0,1]} f(\boldsymbol{x} + \beta(\boldsymbol{x}^* - \boldsymbol{x})) - f(\boldsymbol{x}) + \frac{\rho\beta^2}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2\right) \\ & \leq \frac{1}{M}\left(\min_{\beta\in[0,1]} -\beta(f(\boldsymbol{x}^*) - f(\boldsymbol{x})) + \frac{\rho\beta^2}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2\right) \end{aligned} \tag{35}$$

where the second and fourth inequality follow from the convexity of $f(\boldsymbol{x})$, and the third inequality follows from the fact that both $\boldsymbol{x}^*$ and $\boldsymbol{x}^s$ are feasible ($h(\boldsymbol{x}^*) = h(\boldsymbol{x}) = 0$). Let $L_g \geq 1$ be a Lipschitz-continuous constant of the augmented term $g(E\boldsymbol{x})$ for $\boldsymbol{x}$ satisfying $\|E\boldsymbol{x}-\boldsymbol{q}\| \leq R_q$ (where L_g is at least ρR_q). Based on the error bound inequality (30), we discuss two cases.

Case 1: $4L_g^2\|E\boldsymbol{x} - t^*\|^2 < (\boldsymbol{b}^T\boldsymbol{x} - s^*)^2$.

In this case, we have

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 &\leq \theta(\|E\boldsymbol{x} - t^*\|^2 + (\boldsymbol{b}^T\boldsymbol{x} - s^*)^2) \\ &\leq \theta(\frac{1}{L_g^2} + 1)(\boldsymbol{b}^T\boldsymbol{x} - s^*)^2 \\ &\leq 2\theta(\boldsymbol{b}^T\boldsymbol{x} - s^*)^2,\end{aligned} \tag{36}$$

and

$$|\boldsymbol{b}^T\boldsymbol{x} - s^*| \geq 2L_g\|E\boldsymbol{x} - t^*\| \geq 2|g(E\boldsymbol{x}) - g(t^*)|$$

by the definition of Lipschitz constant L_g. Note $\boldsymbol{b}^T\boldsymbol{x} - s^*$ is non-negative since otherwise, $f(\boldsymbol{x}) - f^* = g(E\boldsymbol{x}) - g(t^*) + (\boldsymbol{b}^T\boldsymbol{x} - s^*) \leq |g(E\boldsymbol{x}) - g(t^*)| - |\boldsymbol{b}^T\boldsymbol{x} - s^*| \leq -\frac{1}{2}|\boldsymbol{b}^T\boldsymbol{x} - s^*| < 0$, which leads to contradiction. Therefore, we have

$$\begin{aligned}f(\boldsymbol{x}) - f^* &= g(E\boldsymbol{x}) - g(t^*) + (\boldsymbol{b}^T\boldsymbol{x} - s^*) \\ &\geq -|g(E\boldsymbol{x}) - g(t^*)| + (\boldsymbol{b}^T\boldsymbol{x} - s^*) \\ &\geq \frac{1}{2}(\boldsymbol{b}^T\boldsymbol{x} - s^*).\end{aligned} \tag{37}$$

Combining (35), (36) and (37), we have

$$\begin{aligned}&\mathbb{E}[f(\boldsymbol{x}^{s+1})] - f(\boldsymbol{x}^s) \\ &\leq \frac{1}{M}\min_{\beta\in[0,1]} -\frac{\beta}{2}(\boldsymbol{b}^T\boldsymbol{x} - s^*) + \frac{2\rho\theta\beta^2}{2}(\boldsymbol{b}^T\boldsymbol{x} - s^*)^2 \\ &= \begin{cases} -1/(16\rho\theta M) & , 1/(4\rho\theta(\boldsymbol{b}^T\boldsymbol{x} - s^*)) \leq 1 \\ -\frac{1}{4M}(\boldsymbol{b}^T\boldsymbol{x} - s^*) & , o.w.\end{cases}\end{aligned}$$

Furthermore, we have

$$-\frac{1}{16\rho\theta M} \leq -\frac{1}{16\rho\theta M(f^0 - f^*)}\left(f(\boldsymbol{x}) - f^*\right)$$

where $f^0 = f(\boldsymbol{x}^0)$, and

$$-\frac{1}{4M}(\boldsymbol{b}^T\boldsymbol{x} - s^*) \leq -\frac{1}{6M}(f(\boldsymbol{x}) - f^*)$$

since $f(\boldsymbol{x}) - f^* \leq |g(E\boldsymbol{x}) - g(t^*)| + \boldsymbol{b}^T\boldsymbol{x} - s^* \leq \frac{3}{2}(\boldsymbol{b}^T\boldsymbol{x} - s^*)$. In summary, for Case 1 we obtain

$$\mathbb{E}[f(\boldsymbol{x}^{s+1})] - f^* \leq (1 - \frac{1}{M\gamma_1})\left(\mathbb{E}[f(\boldsymbol{x})] - f^*\right) \tag{38}$$

where

$$\gamma_1 = \max\left\{16\rho\theta(f^0 - f^*)\ ,\ 6\right\}. \tag{39}$$

Case 2: $4L_g^2\|E\boldsymbol{x} - t^*\|^2 \geq (\boldsymbol{b}^T\boldsymbol{x} - s^*)^2$.

In this case, we have

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \leq \theta\left(1 + 4L_g^2\right)\|E\boldsymbol{x} - t^*\|^2, \tag{40}$$

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

$$f(\boldsymbol{x}) - f^* \geq \boldsymbol{b}^T(\boldsymbol{x} - \boldsymbol{x}^*) + \nabla g(t^*)^T E(\boldsymbol{x} - \boldsymbol{x}^*) + \frac{\rho}{2}\|E\boldsymbol{x} - t^*\|^2.$$

Adding inequality $0 = h(\boldsymbol{x}) - h(\boldsymbol{x}^*) \geq \langle \delta^*, \boldsymbol{x} - \boldsymbol{x}^*\rangle$ for some $\delta^* \in \partial h(\boldsymbol{x}^*)$ to the above gives

$$f(\boldsymbol{x}) - f^* \geq \frac{\rho}{2}\|E\boldsymbol{x} - t^*\|^2 \tag{41}$$

since $\delta^* + \boldsymbol{b} + \nabla g(t^*)^T E = \delta^* + \nabla f(\boldsymbol{x}^*) = 0$. Combining (35), (40), and (41), we obtain

$$\begin{aligned}&\mathbb{E}[f(\boldsymbol{x}^{s+1})] - f(\boldsymbol{x}^s) \\ &\leq \frac{1}{M}\left(\min_{\beta\in[0,1]} -\beta(f(\boldsymbol{x}) - f^*) + \frac{\theta(1 + 4L_g^2)\beta^2}{2}\left(f(\boldsymbol{x}) - f^*\right)\right) \\ &= -\frac{1}{2\theta(1 + 4L_g^2)M}\left(f(\boldsymbol{x}) - f^*\right)\end{aligned} \tag{42}$$

Combining results of Case 1 (38) and Case 2 (42), and taking expectation on both sides w.r.t. the history leads to the result. □

A similar linear convergence result can be proved for Greedy BCD on the AL subproblem (10) with approximate column generation criteria (20), as shown in the following theorem.

Theorem 8 (Linear Convergence of Approximate Greedy BCD)**.** *Let $\{\boldsymbol{x}^s\}_{s=1}^{\infty}$ denote iterates produced by Algorithm 2 (without step 2.5) with a fixed $\boldsymbol{\alpha}$ and let f^* be the optimum of* (10)*. Then*

$$\mathbb{E}[f(\boldsymbol{x}^{s+1})] - f^* \leq \left(1 - \frac{1}{m\gamma_2}\right)\left(f(\boldsymbol{x}^s) - f^*\right),$$

where $m = M/R$,

$$\gamma_2 = \max\left\{16\rho\theta_1(f^0 - f^*)\ ,\ 2\theta_1(1 + 4L_g^2)\ ,\ 6\right\}.$$

and θ_1 is the $\ell_{2,1}$-norm version of Hoffman constant of the optimal solution set satisfying $\theta \leq \theta_1 \leq M\theta$.

Proof. Let

$$H_j(\boldsymbol{x}^s) := \min_{\Delta\boldsymbol{x}_j} h_j(\boldsymbol{x}_j^s + \Delta\boldsymbol{x}_j) + \nabla_j f(\boldsymbol{x}^s)^T\Delta\boldsymbol{x}_j + \frac{\rho}{2}\|\Delta\boldsymbol{x}_j\|^2$$

and

$$j^* = \underset{j}{argmin}\ H_j(\boldsymbol{x}^s).$$

Suppose the Greedy BCD chooses column j^* to update. We have

$$
\begin{aligned}
&f(\boldsymbol{x}^{s+1}) - f(\boldsymbol{x}^s) \le H_{j^*}(\boldsymbol{x}^s) \\
&= \min_{\Delta\boldsymbol{x}:\boldsymbol{x}+\Delta\boldsymbol{x}\in\mathcal{X}} \nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x} + \frac{\rho}{2}(\sum_{j=1}^{M}\|\Delta\boldsymbol{x}_j\|_2)^2 \\
&\le \min_{\Delta\boldsymbol{x}:\boldsymbol{x}+\Delta\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}+\Delta\boldsymbol{x}) - f(\boldsymbol{x}) + \frac{\rho}{2}(\sum_{j=1}^{M}\|\Delta\boldsymbol{x}_j\|_2)^2 \\
&\le \min_{\beta\in[0,1]} f(\boldsymbol{x}+\beta\Delta\boldsymbol{x}^*) - f(\boldsymbol{x}) + \frac{\rho\beta^2}{2}(\sum_{j=1}^{M}\|\Delta\boldsymbol{x}_j^*\|_2)^2 \\
&\le \min_{\beta\in[0,1]} -\beta(f(\boldsymbol{x}^*) - f(\boldsymbol{x})) + \frac{\rho\beta^2}{2}(\sum_{j=1}^{M}\|\Delta\boldsymbol{x}_j^*\|_2)^2
\end{aligned}
$$

where $\Delta\boldsymbol{x}^* = \boldsymbol{x}^* - \boldsymbol{x}$. The first equality is from the fact that linear objective subject to $\ell_{2,1}$-norm constraint has solution on the corner point, which corresponds to $\Delta\boldsymbol{x}$ with $\Delta\boldsymbol{x}_j = 0, \forall j \neq j^*$.

Now consider the approximate greedy column generation based on (20). Note that the approximate score given by (20) is always an underestimate of the true score, and if a column j is picked for generating reference vector $\boldsymbol{q}_j$ the computed score for column j is always exact, and therefore the approximate column generation returns a column that is at least as good as the best one used as reference vectors. By drawing R reference vectors, there is R/M probability column j^* is picked. As a result, the iterates produced by approximate column generation have

$$
\begin{aligned}
&f(\boldsymbol{x}^{s+1}) - f(\boldsymbol{x}^s) \le \frac{1}{m}H_{j^*}(\boldsymbol{x}^s) \\
&\le \frac{1}{m}\left(\min_{\beta\in[0,1]} -\beta(f(\boldsymbol{x}^*) - f(\boldsymbol{x})) + \frac{\rho\beta^2}{2}(\sum_{j=1}^{M}\|\Delta\boldsymbol{x}_j^*\|_2)^2\right),
\end{aligned}
$$

where $m = M/R$. Then following the same reasoning as in Theorem 7 with (30) replaced by the $\ell_{2,1}$-norm version of the Hoffman bound (31) gives the result. □

8.2 Iteration Complexity of AL-BCD

In this section, we show global linear convergence of both randomized AL-BCD (Algorithm 1) and AL-BCD with column generation (Algorithm 2).

Recall that $\mathcal{L}(W;\boldsymbol{\alpha})$ is the Augmented Lagrangian function. The dual objective of the exemplar clustering problem is

$$d(\boldsymbol{\alpha}) := \min_{W\ge 0} \mathcal{L}(W,\boldsymbol{\alpha})$$

and

$$d^* := \max_{\boldsymbol{\alpha}} d(\boldsymbol{\alpha})$$

is the optimal dual objective.

Then we measure the sub-optimality of iterates $\{(W^t,\boldsymbol{\alpha}^t)\}_{t=1}^T$ given by our algorithms in terms of the dual function difference

$$\Delta_d^t = d^* - d(\boldsymbol{\alpha}^t)$$

and the primal function difference:

$$\Delta_p^t = \mathcal{L}(W^{t+1},\boldsymbol{\alpha}^t) - d(\boldsymbol{\alpha}^t).$$

It is clear that any $(W^{t+1},\boldsymbol{\alpha}^t)$ satisfying $\Delta_d^t = \Delta_p^t = 0$ is an optimal solution to our original problem (3).

Lemma 2 (Dual Progress)**.** *Each dual update* (11) *leads to*

$$\Delta_d^t - \Delta_d^{t-1} \le -\eta(W^t\mathbf{1}_M - \mathbf{1}_N)^T(\bar{W}^t\mathbf{1}_M - \mathbf{1}_N), \quad (43)$$

where $\bar{W}^t$ is the projection of W^t to the optimal solution set of AL function

$$\operatorname*{argmin}_W \mathcal{L}(W,\boldsymbol{\alpha}^t).$$

Proof.

$$
\begin{aligned}
\Delta_d^t - \Delta_d^{t-1} &= d^* - d(\boldsymbol{\alpha}^t) - d^* - d(\boldsymbol{\alpha}^{t-1}) \\
&= \mathcal{L}(\bar{W}^{t-1},\boldsymbol{\alpha}^{t-1}) - \mathcal{L}(\bar{W}^t,\boldsymbol{\alpha}^t) \\
&\le \mathcal{L}(\bar{W}^t,\boldsymbol{\alpha}^{t-1}) - \mathcal{L}(\bar{W}^t,\boldsymbol{\alpha}^t) \\
&= \langle\boldsymbol{\alpha}^{t-1} - \boldsymbol{\alpha}^t, \bar{W}^t\mathbf{1}_M - \mathbf{1}_N\rangle \\
&= -\eta\langle W^t\mathbf{1}_M - \mathbf{1}_N, \bar{W}^t\mathbf{1}_M - \mathbf{1}_N\rangle
\end{aligned}
$$

where the first inequality follows from the optimality of $\bar{W}^{t-1}$ for the function $\mathcal{L}(W,\boldsymbol{\alpha}^{t-1})$, and the last equality follows from the dual update (11). □

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

Lemma 3 (Primal Progress)**.** *Each iteration of Algorithm 1 and Algorithm 2 satisfies*

$$
\begin{aligned}
\Delta_p^t - \Delta_p^{t-1} \le& \mathcal{L}(W^{t+1},\boldsymbol{\alpha}^t) - \mathcal{L}(W^t,\boldsymbol{\alpha}^t) \\
&+ \eta\|W^t\mathbf{1}_M - \mathbf{1}_N\|^2 \\
&- \eta\langle W^t\mathbf{1}_M - \mathbf{1}_N, \bar{W}^t\mathbf{1}_M - \mathbf{1}_N\rangle
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\Delta_p^t - \Delta_p^{t-1} \\
=&\mathcal{L}(W^{t+1},\boldsymbol{\alpha}^t) - \mathcal{L}(W^t,\boldsymbol{\alpha}^{t-1}) - (d(\boldsymbol{\alpha}^t) - d(\boldsymbol{\alpha}^{t-1})) \\
\le&\mathcal{L}(W^{t+1},\boldsymbol{\alpha}^t) - \mathcal{L}(W^t,\boldsymbol{\alpha}^t) + \mathcal{L}(W^t,\boldsymbol{\alpha}^t) - \mathcal{L}(W^t,\boldsymbol{\alpha}^{t-1}) \\
&+ (d(\boldsymbol{\alpha}^{t-1}) - d(\boldsymbol{\alpha}^t)) \\
\le&\mathcal{L}(W^{t+1},\boldsymbol{\alpha}^t) - \mathcal{L}(W^t,\boldsymbol{\alpha}^t) + \eta\|W^t\mathbf{1}_M - \mathbf{1}_N\|^2 \\
&- \eta\langle W^t\mathbf{1}_M - \mathbf{1}_N, \bar{W}^t\mathbf{1}_M - \mathbf{1}_N\rangle
\end{aligned}
$$

where the last inequality uses Lemma 2 on $d(\boldsymbol{\alpha}^{t-1}) - d(\boldsymbol{\alpha}^t) = \Delta_d^t - \Delta_d^{t-1}$. □

By combining results of Lemma 2 and 3, we can obtain a joint progress of the form

$$\begin{aligned}&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1}\\ &\leq \mathcal{L}(W^{t+1}, \boldsymbol{\alpha}^t) - \mathcal{L}(W^t, \boldsymbol{\alpha}^t) + \eta\|W^t\mathbf{1}_M - \bar{W}^t\mathbf{1}_M\|^2\\ &- \eta\|\bar{W}^t\mathbf{1}_M - \mathbf{1}_N\|^2\end{aligned} \tag{44}$$

Note the only positive term in (44) is the term $\eta\|W^t\mathbf{1}_M - \bar{W}^t\mathbf{1}_M\|^2$. To guarantee descent of the joint measure of suboptimality $\Delta_d^t + \Delta_p^t$, we bound the second term of (44) with the primal gap $\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - d(\boldsymbol{\alpha}^t)$ given by the following lemma.

Lemma 4.

$$\|W^t\mathbf{1}_M - \bar{W}^t\mathbf{1}_M\|^2 \leq \frac{2}{\rho}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t)) \tag{45}$$

Proof. Let

$$\tilde{\mathcal{L}}(W, \boldsymbol{\alpha}) = h(W) + g(W\mathbf{1}_M),$$

where

$$g(W\mathbf{1}_M) = \frac{\rho}{2}\|W\mathbf{1}_M - \mathbf{1}_N\|^2$$

and

$$h(W) = tr(\mathbb{D}^T W) + \lambda\|W\|_{\infty,1} + \boldsymbol{\alpha}^T(W\mathbf{1}_M - \mathbf{1}_N),$$

, if $W \geq 0$ and $h(W) = \infty$ if W is infeasible. According to the definition of $d(\boldsymbol{\alpha})$, we know that

$$0 \in \partial_W \tilde{\mathcal{L}}(\bar{W}^t, \boldsymbol{\alpha}) = \partial h(\bar{W}^t) + \nabla_W g(W\mathbf{1}_M).$$

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

$$h(W^t) - h(\bar{W}^t) \geq \langle \delta^*, W^t - \bar{W}^t\rangle$$

for any $\delta^* \in \partial h(\bar{W}^t)$. and

$$\begin{aligned}&g(W^t\mathbf{1}_M) - g(\bar{W}^t\mathbf{1}_M))\\ \geq&\langle \nabla_W g(W^t\mathbf{1}_M), W^t - \bar{W}^t\rangle + \frac{\rho}{2}\|W^t\mathbf{1}_M - \bar{W}^t\mathbf{1}_M\|^2\end{aligned}$$

The the above two together implies

$$\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t) \geq \frac{\rho}{2}\|W^t\mathbf{1}_M - \bar{W}^t\mathbf{1}_M\|^2$$

which leads to our conclusion. □

Now the joint progress can be written as

$$\begin{aligned}&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1}\\ &\leq \mathcal{L}(W^{t+1}, \boldsymbol{\alpha}^t) - \mathcal{L}(W^t, \boldsymbol{\alpha}^t)\\ &+ \frac{2\eta}{\rho}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t))\\ &- \eta\|\bar{W}^t\mathbf{1}_M - \mathbf{1}_N\|^2,\end{aligned} \tag{46}$$

which indicates that as long as the primal update leads to a descent of AL function $\mathcal{L}(W^{t+1}, \boldsymbol{\alpha}^t) - \mathcal{L}(W^t, \boldsymbol{\alpha}^t)$ proportional to the current primal suboptimality $\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t)$, one can choose a small enough constant step size η to guarantee descent of the joint suboptimality $\Delta_d^t + \Delta_p^t$.

Note the term $\bar{W}^t\mathbf{1}_M - \mathbf{1}_N$ is actually gradient of dual objective $\nabla d(\boldsymbol{\alpha}^t)$, and since problem (3) satisfies the Assumption A(a)-A(e) in [11], the error bound

$$\Delta_d(\boldsymbol{\alpha}) \leq \tau\|\nabla d(\boldsymbol{\alpha})\|^2 \tag{47}$$

in Lemma 3.1 of [11] applies to our dual objective $d(\boldsymbol{\alpha})$ with compact domain $\boldsymbol{\alpha} \in R(\boldsymbol{\alpha}^0)$, where $\tau > 0$ is a constant that depends on geometry of solution set S.

Now we are ready to give the convergence results for Randomized ALBCD and Greedy ALBCD.

Theorem 9 (Linear Convergence of Randomized ALBCD)**.** *The iterates $\{(W^t, \boldsymbol{\alpha}^t)\}_{t=1}^{\infty}$ produced by Algorithm 1 has*

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

for any $0 < \eta \leq \rho/4M\gamma$, where $\tau > 0$ is a constant depending on the geometry of optimal solution set.

Proof. By (46), (47) and Theorem 7, we have

$$\begin{aligned}&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1}\\ &\leq -\frac{1}{M\gamma}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t))\\ &+ \frac{2\eta}{\rho}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t))\\ &- \frac{\eta}{\tau}\Delta_d^t.\end{aligned}$$

Setting $\eta \leq \rho/4M\gamma$, we have

$$\begin{aligned}&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1}\\ &\leq -\frac{1}{2M\gamma}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t)) - \frac{\eta}{\tau}\Delta_d^t\\ &\leq -\frac{1}{2M\gamma}\Delta_p^t - \frac{\eta}{\tau}\Delta_d^t,\end{aligned}$$

which leads to the result. □

Theorem 10 (Linear Convergence of Greedy ALBCD)**.** *The iterates $\{(W^t, \boldsymbol{\alpha}^t)\}_{t=1}^{\infty}$ produced by Algorithm 2 has*

$$\Delta_d^t + \Delta_p^t \leq \frac{1}{1 + \min(\frac{1}{2m\gamma_2}, \frac{\eta}{\tau})}\left(\Delta_d^{t-1} + \Delta_p^{t-1}\right),$$

for any $0 < \eta \leq \rho/4m\gamma_2$.

Proof. By (46), (47) and Theorem 8, we have

$$\begin{aligned}
&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \\
&\leq -\frac{1}{m\gamma_2}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t)) \\
&+ \frac{2\eta}{\rho}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t)) \\
&- \frac{\eta}{\tau}\Delta_d^t.
\end{aligned}$$

Setting $\eta \leq \rho/4m\gamma_2$, we have

$$\begin{aligned}
&\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} \\
&\leq -\frac{1}{2m\gamma_2}(\mathcal{L}(W^t, \boldsymbol{\alpha}^t) - \mathcal{L}(\bar{W}^t, \boldsymbol{\alpha}^t)) - \frac{\eta}{\tau}\Delta_d^t \\
&\leq -\frac{1}{2m\gamma_2}\Delta_p^t - \frac{\eta}{\tau}\Delta_d^t,
\end{aligned}$$

which leads to the result. □