

5 Appendix A: Block-Coordinate Frank-Wolfe (BCFW)— Practically Faster-Convergent Variant

Algorithm 4 Block-Coordinate Frank Wolfe (improving upon Algorithm 2)

for $s = 1$ to $|\mathcal{F}|$ **do**
 1. Draw $f \in \mathcal{F}$ uniformly at random.
 2. Find a greedy direction \mathbf{v}_f^+ satisfying (15).
 3. $\mathcal{A}_f^{s+1} = \mathcal{A}_f^s \cup \{\mathbf{v}_f^+\}$.
 4. Solve (25) with active set \mathcal{A}_f^{s+1} .
 5. Maintain $\mathbf{w}_F(\boldsymbol{\alpha})$.
end for

The *Block-Coordinate Frank-Wolfe (BCFW)* (Algorithm 4) differs from *Frank-Wolfe* (Algorithm 2) in that it updates dual variables $\boldsymbol{\alpha}_f$ of each factor sequentially, and the bases \mathbf{v}_f and active sets \mathcal{A}_f are maintained for each factor f separately. For each iteration of BCFW, we find a greedy direction \mathbf{v}_f^+ in the same way (15) as AFW, but for one factor at a time. Then we add \mathbf{v}_f^+ to an active set \mathcal{A}_f maintained for each factor. Since in BCFW we update one factor at a time, we can minimize the following *block active-set subproblem*

$$\mathbf{d}_{\mathcal{A}_f}^* = \underset{\boldsymbol{\alpha}_f + \mathbf{d}_{\mathcal{A}_f} \in \Delta^{|\mathcal{Y}_f|}}{\operatorname{argmin}} \quad \langle \nabla_{\boldsymbol{\alpha}_f} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\lambda}^t), \mathbf{d}_{\mathcal{A}_f} \rangle + \frac{Q_f}{2} \|\mathbf{d}_{\mathcal{A}_f}\|^2 \quad (25)$$

where Q_f is an upper bound on the Hessian of variables in the active set (discussed in section 5.1). The active-set subproblem (25) can be solved via a simplex projection in time $O(|\mathcal{A}_f|)$ [2]. Furthermore, by maintaining $\mathbf{w}_F(\boldsymbol{\alpha})$ after solving each sub-problem (25), we can compute the gradient

$$\nabla_{\boldsymbol{\alpha}_f(\mathbf{y}_f)} \mathcal{L} = \langle \mathbf{w}_F, \boldsymbol{\phi}_f(\mathbf{x}_f, \mathbf{y}_f) \rangle - \delta_f(\mathbf{y}_f, \bar{\mathbf{y}}_f) + \rho_f \sum_{j \in \mathcal{N}(f), \mathbf{y}_j = [\mathbf{y}_f]_j} m_{jf}(\mathbf{y}_j) \quad (26)$$

for $\mathbf{y}_f \in \mathcal{A}_f$ in time $O(|\mathcal{A}_f| \operatorname{nnz}(\boldsymbol{\phi}_f))$, where $\rho_f = -\rho$, $\delta_f = \delta_j$ for $f \in \mathcal{U}$ and $\rho_f = \rho$, $\delta_f = 0$ for $f \notin \mathcal{U}$. Note the size of active set $|\mathcal{A}_f|$ is bounded by the number of GDMM iterations, and in practice $|\mathcal{A}_f|$ converges to a constant much smaller than $|\mathcal{Y}_f|$ for problems of large output domains. Therefore, the bottleneck of the BCFW algorithm lies in the step (15), which as we show in Section 3.2, can be computed in time sublinear to $|\mathcal{Y}_f|$ via an efficient FMO.

5.1 Constant Q_f in Problem (25)

The constant Q_f is an upper bound on the maximum eigenvalue of the Hessian submatrix for variables in the active set \mathcal{A}_f , that is, $\|[\Phi_f \Phi_f^T]_{\mathcal{A}_f}\| + \rho \sum_{j \in \mathcal{N}_f} \| [M_{jf}^T M_{jf}]_{\mathcal{A}_f} \|$, where the notation $[\cdot]_{\mathcal{A}}$ denotes the sub-matrix formed by row and column indexes in \mathcal{A} and $\|\cdot\|$ is the spectral norm of a matrix.

For many types of factors used in practice, $Q_f = O(|\mathcal{A}_f|)$ and is easy to compute in the beginning.

In particular, for unigram factor, we have $M_{jf} = I$ and thus $\| [M_{jf}^T M_{jf}]_{\mathcal{A}} \| = 1$, and for higher-order factor we have $\| [M_{jf}^T M_{jf}]_{\mathcal{A}} \| = |\mathcal{A}|$.

As for the term $\| [\Phi_f \Phi_f^T]_{\mathcal{A}} \|$. In most of applications, Φ_f is a $|\mathcal{Y}_f| \times (|\mathcal{Y}_f|d)$ block-diagonal matrix that duplicates $1 \times d$ feature vector $\boldsymbol{\phi}_f(\mathbf{x}_f)^T$ for $|\mathcal{Y}_f|$ times, for which we have $\| [\Phi_f \Phi_f^T]_{\mathcal{A}} \| = \|\boldsymbol{\phi}_f(\mathbf{x}_f)\|^2$. Note in this case, the quadratic upper bound in (25) is tight for unigram factors.

6 Appendix B: Convergence of GDMM

6.1 Proof of Theorem (1)

Recall that the Augmented Lagrangian $\mathcal{L}(\alpha, \lambda)$ is of the form

$$\mathcal{L}(\alpha, \lambda) := G(\alpha) + \langle \lambda, M\alpha \rangle + \frac{\rho}{2} \|M\alpha\|^2.$$

where M is "the number of consistency constraints" by "the number of variables" matrix and $M\alpha = \mathbf{0}$ encodes all constraints of the form

$$M_{jf}\alpha_f - \alpha_j = [M_{jf} \quad -I_j] \begin{bmatrix} \alpha_f \\ \alpha_j \end{bmatrix} = \mathbf{0}.$$

The function

$$G(\alpha) = \frac{1}{2} \sum_{F \in \mathcal{F}} \|w_F(\alpha_F)\|^2 - \sum_{j \in \mathcal{U}} \delta_j^T \alpha_j$$

can be written in a compact form as

$$\begin{aligned} G(\alpha) &= \frac{1}{2} \|w(\alpha)\|^2 + \delta^T \alpha \\ &= \frac{1}{2} \|\Phi^T \alpha\|^2 + \delta^T \alpha \end{aligned} \tag{27}$$

where Φ is the "number of variables (in α)" by "number of parameters (in w)" design matrix.

Now let α be the "primal variables" and denote

$$\alpha(\lambda) := \{\alpha \mid \alpha = \arg \min_{\alpha} \mathcal{L}(\alpha, \lambda)\} \tag{28}$$

with

$$\bar{\alpha}^t := \arg \min_{\bar{\alpha} \in \alpha(\lambda^t)} \|\bar{\alpha} - \alpha^t\|,$$

and let $\mathcal{M} = \{\alpha \mid \alpha_f \in \Delta^{|\mathcal{Y}_f|}, \forall f \in \mathcal{F}\}$. The dual objective of the augmented problem is

$$d(\lambda) = \min_{\alpha \in \mathcal{M}} \mathcal{L}(\alpha, \lambda)$$

and

$$d^* = \max_{\lambda} d(\lambda)$$

is the optimal dual objective value.

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

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

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

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

yielded by α^{t+1} obtained from one pass of FC-BCFW algorithm on α . Then we have following lemma.

Lemma 1 (Dual Progress). *Each iteration of GDMM (Algorithm 1) has*

$$\Delta_d^t - \Delta_d^{t-1} \leq -\eta (M\alpha^t)^T (M\bar{\alpha}^t). \tag{29}$$

Proof.

$$\begin{aligned} \Delta_d^t - \Delta_d^{t-1} &= d^* - d(\lambda^t) - d^* - d(\lambda^{t-1}) \\ &= \mathcal{L}(\bar{\alpha}^{t-1}, \lambda^{t-1}) - \mathcal{L}(\bar{\alpha}^t, \lambda^t) \\ &\leq \mathcal{L}(\bar{\alpha}^t, \lambda^{t-1}) - \mathcal{L}(\bar{\alpha}^t, \lambda^t) \\ &= \langle \lambda^{t-1} - \lambda^t, M\bar{\alpha}^t \rangle \\ &= -\eta \langle M\alpha^t, M\bar{\alpha}^t \rangle \end{aligned}$$

where the first inequality follows the optimality of $\bar{\alpha}^{t-1}$ for the function $\mathcal{L}(\alpha, \lambda^{t-1})$ defined by λ^{t-1} , and the last equality follows the dual update in GDMM (13). \square

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*

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

Proof.

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

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

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

$$\begin{aligned}\Delta_d^t - \Delta_d^{t-1} + \Delta_p^t - \Delta_p^{t-1} &\leq \mathcal{L}(\alpha^{t+1}, \lambda^t) - \mathcal{L}(\alpha^t, \lambda^t) + \eta \|M\alpha^t - M\bar{\alpha}^t\|^2 - \eta \|M\bar{\alpha}^t\|^2\end{aligned}\tag{30}$$

Note the only positive term in (30) is the second one. To guarantee the descent of joint progress, we upper bound the three terms in (30) with the following lemmas.

Lemma 3.

$$\|M\alpha^t - M\bar{\alpha}^t\|^2 \leq \frac{2}{\rho} (\mathcal{L}(\alpha^t, \lambda^t) - \mathcal{L}(\bar{\alpha}^t, \lambda^t))\tag{31}$$

Proof. Let

$$\tilde{\mathcal{L}}(\alpha, \lambda) = h(\alpha) + \frac{\rho}{2} \|M\alpha\|^2,$$

where

$$h(\alpha) = G(\alpha) + \langle \lambda, M\alpha \rangle + \mathbf{I}_{\alpha \in \mathcal{M}}.$$

, $\mathbf{I}_{\alpha \in \mathcal{M}} = 0$ if $\alpha \in \mathcal{M}$ and $\mathbf{I}_{\alpha \in \mathcal{M}} = \infty$ otherwise. Note we have $\tilde{\mathcal{L}}(\bar{\alpha}^t, \lambda^t) = \mathcal{L}(\bar{\alpha}^t, \lambda^t)$ and $\tilde{\mathcal{L}}(\alpha^t, \lambda^t) = \mathcal{L}(\alpha^t, \lambda^t)$ due to feasible iterates. Due to the optimality of $\bar{\alpha}^t$, we have

$$\mathbf{0} = \sigma + M^T M\bar{\alpha}^t \in \partial_\alpha \tilde{\mathcal{L}}(\bar{\alpha}^t, \lambda)$$

for some $\sigma \in \partial h(\bar{\alpha}^t)$. And by the convexity of $h(\cdot)$ and the strong convexity of $\frac{\rho}{2} \|\cdot\|^2$, we have

$$h(\alpha^t) - h(\bar{\alpha}^t) \geq \langle \sigma, \alpha^t - \bar{\alpha}^t \rangle$$

and

$$\frac{\rho}{2} \|M\alpha^t\|^2 - \frac{\rho}{2} \|M\bar{\alpha}^t\|^2 \geq \langle M^T M\bar{\alpha}^t, \alpha^t - \bar{\alpha}^t \rangle + \frac{\rho}{2} \|M\alpha^t - M\bar{\alpha}^t\|^2$$

Then the above two together imply

$$\mathcal{L}(\alpha^t, \lambda^t) - \mathcal{L}(\bar{\alpha}^t, \lambda^t) \geq \frac{\rho}{2} \|M(\alpha^t) - M(\bar{\alpha}^t)\|^2$$

which leads to our conclusion. \square

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

$$\Delta_d(\lambda) \leq \tau \|M\bar{\alpha}(\lambda)\|^2.\tag{32}$$

for any λ and any minimizer $\bar{\alpha}(\lambda)$ satisfying (28).

Proof. This is a lemma adapted from [5]. Since our objective (12) satisfies the *assumptions* $A(a)$ — $A(e)$ and $A(g)$ in [5]. Then Lemma 3.1 of [5] guarantees that, as long as $\|\nabla d(\lambda)\|$ is bounded, there is a constant $\tau > 0$ s.t.

$$\Delta_d(\lambda) \leq \tau \|\nabla d(\lambda)\|^2 = \|M\bar{\alpha}(\lambda)\|^2$$

for all λ . Note our problem satisfies the condition of bounded gradient magnitude since

$$\|\nabla d(\lambda)\| = \|M\bar{\alpha}(\lambda)\| \leq \|M\bar{\alpha}(\lambda)\|_1 \leq \|M\|_1 \|\bar{\alpha}(\lambda)\|_1 \leq (\max_f |\mathcal{Y}_f|) |\mathcal{F}|$$

where the last inequality is because $\bar{\alpha}(\lambda)$ lies in a simplex domain. \square

The remaining thing is to show that one pass of AFW (Algorithm 2) or BGCD (Algorithm 3) suffices to give a descent amount $\mathcal{L}(\alpha^{t+1}, \lambda^t) - \mathcal{L}(\alpha^t, \lambda^t)$ lower bounded by some constant multiple of the primal sub-optimality $\mathcal{L}(\alpha^t, \lambda^t) - \mathcal{L}(\bar{\alpha}^t, \lambda^t)$. If it is true, then by selecting a small enough GDM step size η , the RHS of (30) would be negative. For AFW (Algorithm 2), this can be achieved by leveraging recent results from [8], which shows linear convergence of AFW, even for non-strongly convex function of the form (34). We thus have the following lemma.

Lemma 5. *The descent amount of Augmented Lagrangian function produced by one pass of AFW (Algorithm 2) (and FMO parameter ν) has*

$$E[\mathcal{L}(\alpha^{t+1}, \lambda^t)] - \mathcal{L}(\alpha^t, \lambda^t) \leq -\frac{\mu_{\mathcal{M}}}{4(1+\nu)mQ} (\mathcal{L}(\alpha^t, \lambda^t) - \mathcal{L}(\bar{\alpha}^t, \lambda^t)) \quad (33)$$

where $\mu_{\mathcal{M}}$ is the generalized geometric strong convexity constant for function $\mathcal{L}(\alpha)$ in domain \mathcal{M} , Q is the Lipschitz-continuous constant of $\nabla_{\alpha} \mathcal{L}(\alpha)$ and $m = |\mathcal{F}|$.

Proof. Note the Augmented Lagrangian is of the form

$$H(\alpha) := \mathcal{L}(\alpha, \lambda^t) = g(B\alpha) + \langle b, \alpha \rangle \quad (34)$$

where

$$B := \begin{bmatrix} \Phi^T \\ M \end{bmatrix}, \quad b := \delta + M^T \lambda^t$$

and function $g\left(\begin{bmatrix} w \\ v \end{bmatrix}\right) = \frac{1}{2}\|w\|^2 + \frac{\rho}{2}\|v\|^2 + \text{const.}$ is strongly convex with parameter $\bar{\rho} = \min(1, \rho)$. Without loss of generality, assume $\rho \leq 1$ and thus $\bar{\rho} = \rho$. Since we are minimizing the function subject to a convex, polyhedral domain \mathcal{M} , by Theorem 10 of [8], we have the *generalized geometrical strong convexity* constant $\mu_{\mathcal{M}}$ of the form

$$\mu_{\mathcal{M}} := \mu(PWidth(\mathcal{M}))^2 \quad (35)$$

where $PWidth(\mathcal{M}) > 0$ is the pyramidal width of the simplex domain \mathcal{M} and μ is the *generalized strong convexity* constant of function (34) (defined by Lemma 9 of [8]). By definition of the geometric strong convexity constant, we have

$$H(\alpha^t) - H^* \leq \frac{g_t^2}{2\mu_{\mathcal{M}}} \quad (36)$$

from (23) in [8], where $g_t := \langle -\nabla H(\alpha^t), v^F - v^A \rangle$, and v^F is the Frank-Wolfe direction

$$v^F := \arg \min_{v \in \mathcal{M}} \langle \nabla H(\alpha^t), v \rangle,$$

v^A is the away direction

$$v^A := \arg \max_{v \in \mathcal{A}^t} \langle \nabla H(\alpha^t), v \rangle$$

Then let $m = |\mathcal{F}|$ be the number of factors. The FMO returns $v_f^+ = v_f^F$ with probability at least $\frac{1}{\nu}$, and suppose we set v_f^+ to α_f^t whenever $\langle \nabla_{\alpha_f} H, v_f^+ - \alpha_f^t \rangle \not\leq 0$. We have $\langle \nabla H, d_F \rangle \leq \frac{1}{\nu} \langle \nabla H, v^F - \alpha^t \rangle$ and thus

$$(1 + \frac{1}{\nu}) \langle \nabla H, d^t \rangle \leq \frac{1}{\nu} \langle \nabla H, v^F - \alpha^t \rangle + \frac{1}{\nu} \langle \nabla H, \alpha^t - v^A \rangle$$

and $\langle \nabla H, \mathbf{d}^t \rangle \leq -\frac{1}{1+\nu} g_t$. Therefore, for any $\forall \gamma \in [0, 1]$,

$$E[H(\boldsymbol{\alpha}^{t+1})] - H(\boldsymbol{\alpha}^t) \leq -\gamma \frac{g_t}{1+\nu} + \frac{Q}{2} \|\gamma(\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^t)\|^2 \leq -\gamma \frac{g_t}{1+\nu} + \frac{2mQ}{2} \gamma^2 \quad (37)$$

where Q is an upper bound on the spectral norm of Hessian $\|\nabla^2 H(\boldsymbol{\alpha})\|$ and $2m$ is the square of the radius of domain \mathcal{M} . Now we need to consider two cases. When the greedy direction \mathbf{d}_F in (17) is chosen, we have $\gamma^* = \min(\frac{g_s}{mQ_{\max}}, 1)$, which gives us

$$E[H(\boldsymbol{\alpha}^{t+1})] - H(\boldsymbol{\alpha}^t) \leq -\frac{g_t^2}{4(1+\nu)mQ}. \quad (38)$$

While in case \mathbf{d}_A in (17) is chosen, we have $\gamma^* = \min(\frac{g_s}{mQ_{\max}}, c_{v^-})$. When $\gamma^* = c_{v^-}$, a basis \mathbf{v}^- is removed from the active set and this is called a *drop step* [8] and it is hard to show sufficient descent in this case. Nevertheless, we can ignore those drop steps since the number of them is at most half of the iterates. For a non-drop step t , with the error bound (36), we have

$$E[H(\boldsymbol{\alpha}^{t+1})] - H(\boldsymbol{\alpha}^t) \leq -\frac{\mu_{\mathcal{M}}(H(\boldsymbol{\alpha}^t) - H^*)}{4(1+\nu)mQ}. \quad (39)$$

□

The above Lemma shows a significant progress made by the AFW algorithm. In the following, we provide a similar Lemma for minimizing AL subproblem with the Block-Greedy Coordinate Descent (BGCD) (Algorithm 3). Note that for problem of the form (34), the optimal solution is profiled by a polyhedral set $\mathcal{S} := \{\boldsymbol{\alpha} \mid B\boldsymbol{\alpha} = \mathbf{t}^*, \mathbf{b}^T \boldsymbol{\alpha} = s^*, \boldsymbol{\alpha} \in \mathcal{M}\}$. Therefore, let $\bar{\boldsymbol{\alpha}} := \Pi_{\mathcal{S}}(\boldsymbol{\alpha})$. We can bound the distance of any feasible point $\boldsymbol{\alpha} \in \mathcal{M}$ to its projection $\Pi_{\mathcal{S}}(\boldsymbol{\alpha})$ on \mathcal{S} using the Hoffman's inequality [4]

$$\|\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{2,1}^2 = \sum_{i=1}^n \left(\sum_{f \in \mathcal{F}_i} \|\bar{\boldsymbol{\alpha}}_f - \boldsymbol{\alpha}_f\|_2 \right)^2 \leq \theta_1 \left(\|B\boldsymbol{\alpha} - \mathbf{t}^*\|^2 + \|\mathbf{b}^T \boldsymbol{\alpha} - s^*\|^2 \right) \quad (40)$$

where θ_1 is a constant depending on the set \mathcal{S} . Then we can establish the following Lemma using the error bound (40).

Lemma 6. *The descent amount of Augmented Lagrangian function given by one pass of BGCD (Algorithm 3) with FMO multiplicative-approximation parameter ν has*

$$E[\mathcal{L}(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\lambda}^t)] - \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\lambda}^t) \leq \frac{-1}{1 + \nu Q_{\max}/\mu_1} (\mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\lambda}^t) - \mathcal{L}(\bar{\boldsymbol{\alpha}}^t, \boldsymbol{\lambda}^t)) \quad (41)$$

where

$$\mu_1 := \frac{1}{\max\{16\theta_1 \Delta \mathcal{L}^0, 2\theta_1(1 + 4L_g^2)\}}.$$

is the generalized strong convexity constant for function $\mathcal{L}(\boldsymbol{\alpha})$ with feasible domain \mathcal{M} , $\Delta \mathcal{L}^0$ is a bound on $\mathcal{L}(\boldsymbol{\alpha}^0) - \mathcal{L}(\bar{\boldsymbol{\alpha}}^0)$, L_g is the local Lipschitz-continuous constant of $g(\cdot)$ and $Q_{\max} = \max_{f \in \mathcal{F}} Q_f$.

Proof. For each iteration s of Algorithm 3, let i be the chosen sample and suppose that out of ν partitions the one containing greedy factor satisfying (18) is chosen. We have

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}^{s+1}) - \mathcal{L}(\boldsymbol{\alpha}^s) &\leq \min_{\boldsymbol{\alpha}_{f^*}^s + \mathbf{d}_{f^*} \in \Delta^{|\mathcal{Y}_{f^*}|}} \langle \nabla_{\boldsymbol{\alpha}_{f^*}} \mathcal{L}, \mathbf{d}_{f^*} \rangle + \frac{Q_{\max}}{2} \|\mathbf{d}_{f^*}\|^2 \\ &= \min_{\boldsymbol{\alpha}_f^s + \mathbf{d}_f \in \Delta^{|\mathcal{Y}_f|}} + \sum_{f \in \mathcal{F}_i} \langle \nabla_{\boldsymbol{\alpha}_f} \mathcal{L}, \mathbf{d}_f \rangle + \frac{Q_{\max}}{2} \left(\sum_{f \in \mathcal{F}_i} \|\mathbf{d}_f\| \right)^2 \end{aligned} \quad (42)$$

where the second equality follows from the optimality of f^* w.r.t. (18). Then consider i being uniformly sampled from $[n]$, and consider the probability that the partition containing greedy factor

f^* is chosen, the expected descent amount is

$$\begin{aligned}
& E[\mathcal{L}(\alpha^{s+1})] - \mathcal{L}(\alpha^s) \\
& \leq \frac{1}{n\nu} \left(\min_{\alpha_f^s + d_f \in \Delta^{\mathcal{Y}_f}} \sum_{f \in \mathcal{F}} \langle \nabla_{\alpha_f} \mathcal{L}, d_f \rangle + \frac{Q_{\max}}{2} \sum_{i=1}^n \left(\sum_{f \in \mathcal{F}_i} \|d_f\| \right)^2 \right) \\
& \leq \frac{1}{n\nu} \left(\min_{\alpha_f^s + d_f \in \Delta^{\mathcal{Y}_f}} \mathcal{L}(\alpha^s + d) - \mathcal{L}(\alpha^s) + \frac{Q_{\max}}{2} \sum_{i=1}^n \left(\sum_{f \in \mathcal{F}_i} \|d_f\| \right)^2 \right) \\
& \leq \frac{1}{n\nu} \left(\min_{\beta \in [0,1]} \mathcal{L}(\alpha^s + \beta(\bar{\alpha}^s - \alpha^s)) - \mathcal{L}(\alpha^s) + \frac{Q_{\max}\beta^2}{2} \sum_{i=1}^n \left(\sum_{f \in \mathcal{F}_i} \|\bar{\alpha}_f^s - \alpha_f^s\| \right)^2 \right) \\
& \leq \frac{1}{n\nu} \left(\min_{\beta \in [0,1]} \beta(\mathcal{L}(\bar{\alpha}^s) - \mathcal{L}(\alpha^s)) + \frac{Q_{\max}\beta^2}{2} \|\bar{\alpha}^s - \alpha^s\|_{2,1}^2 \right)
\end{aligned} \tag{43}$$

where $\bar{\alpha}^s = \Pi_{\mathcal{S}}(\alpha^s)$ is the projection of $\bar{\alpha}^s$ 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 we discuss two cases in the following.

Case 1: $4L_g^2 \|B\alpha^s - t^*\|^2 < (b^T \alpha^s - s^*)^2$.

In this case, by the Hoffman inequality (40), we have

$$\begin{aligned}
\|\alpha^s - \bar{\alpha}^s\|_{2,1}^2 & \leq \theta_1 (\|B\bar{\alpha}^s - t^*\|^2 + (b^T \alpha^s - s^*)^2) \\
& \leq \theta_1 \left(\frac{1}{4L_g^2} + 1 \right) (b^T \alpha^s - s^*)^2 \\
& \leq 2\theta_1 (b^T \alpha^s - s^*)^2,
\end{aligned} \tag{44}$$

since $\frac{1}{4L_g^2} \leq 1$. Then

$$|b^T \alpha^s - s^*| \geq 2L_g \|B\alpha^s - t^*\| \geq 2|g(B\alpha^s) - g(t^*)|$$

by the definition of Lipschitz constant L_g . Note that $b^T \alpha^s - s^*$ is non-negative since otherwise we have contradiction $\mathcal{L}(\alpha^s) - \mathcal{L}^* = g(B\alpha^s) - g(t^*) + (b^T \alpha^s - s^*) \leq |g(B\alpha^s) - g(t^*)| - |b^T \alpha^s - s^*| \leq -\frac{1}{2}|b^T \alpha^s - s^*| < 0$. Therefore, we have

$$\begin{aligned}
\mathcal{L}(\alpha^s) - \mathcal{L}^* & = g(B\alpha^s) - g(t^*) + (b^T \alpha^s - s^*) \\
& \geq -|g(B\alpha^s) - g(t^*)| + (b^T \alpha^s - s^*) \\
& \geq \frac{1}{2}(b^T \alpha^s - s^*).
\end{aligned} \tag{45}$$

Combining (43), (44) and (45), we have

$$\begin{aligned}
& \mathbb{E}[\mathcal{L}(\alpha^{s+1})] - \mathcal{L}(\alpha^s) \\
& \leq \frac{1}{n\nu} \left(\min_{\beta \in [0,1]} -\frac{\beta}{2} (b^T \alpha^s - s^*) + \frac{2\theta_1 Q_{\max} \beta^2}{2} (b^T \alpha^s - s^*)^2 \right) \\
& = \begin{cases} -1/(16\theta_1 Q_{\max} n\nu) & , 1/(4\theta_1 Q_{\max} (b^T \alpha^s - s^*)) \leq 1 \\ -\frac{1}{4n\nu} (b^T \alpha^s - s^*) & , o.w. \end{cases}
\end{aligned}$$

Furthermore, we have

$$-\frac{1}{16Q_{\max}\theta_1 n\nu} \leq -\frac{1}{16Q_{\max}\theta_1 n\nu(\mathcal{L}^0 - \mathcal{L}^*)} (\mathcal{L}(\alpha^s) - \mathcal{L}^*) \tag{46}$$

where $\mathcal{L}^0 = \mathcal{L}(\alpha^0)$, and

$$-\frac{1}{4n\nu} (b^T \alpha^s - s^*) \leq -\frac{1}{6n\nu} (\mathcal{L}(\alpha^s) - \mathcal{L}^*) \tag{47}$$

since $\mathcal{L}(\alpha^s) - \mathcal{L}^* \leq |g(B\alpha^s) - g(t^*)| + \mathbf{b}^T \alpha^s - s^* \leq \frac{3}{2}(\mathbf{b}^T \alpha^s - s^*)$. Since the bound (46) is much smaller than (47). For Case 1, we obtain

$$\mathbb{E}[\mathcal{L}(\alpha^{s+1})] - \mathcal{L}^s \leq -\frac{\mu_1}{n\nu Q_{\max}} (\mathcal{L}(\alpha^s) - \mathcal{L}^*) \quad (48)$$

where

$$\mu_1 = \frac{1}{16\theta(\mathcal{L}^0 - \mathcal{L}^*)}. \quad (49)$$

Case 2: $4L_g^2 \|B\alpha^s - t^*\|^2 \geq (\mathbf{b}^T \alpha^s - s^*)^2$.

In this case, we have

$$\|\alpha^s - \bar{\alpha}^s\|_{2,1}^2 \leq \theta_1 (1 + 4L_g^2) \|B\alpha^s - t^*\|^2, \quad (50)$$

and by strong convexity of $g(\cdot)$,

$$\mathcal{L}(\alpha^s) - \mathcal{L}^* \geq \mathbf{b}^T (\alpha^s - \alpha^*) + \nabla g(t^*)^T B(\alpha^s - \bar{\alpha}^s) + \frac{\rho}{2} \|B\alpha^s - t^*\|^2.$$

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

$$\mathcal{L}(\alpha^s) - \mathcal{L}^* \geq \frac{\rho}{2} \|B\alpha^s - t^*\|^2 \quad (51)$$

because $\sigma^* + \mathbf{b} + \nabla g(t^*)^T B = \sigma^* + \nabla \mathcal{L}(\bar{\alpha}^s) = 0$ for some $\sigma^* \in \partial h(\bar{\alpha}^s)$. Combining (43), (50), and (51), we obtain

$$\begin{aligned} & \mathbb{E}[\mathcal{L}(\alpha^{s+1})] - \mathcal{L}(\alpha^s) \\ & \leq \frac{1}{n\nu} \left(\min_{\beta \in [0,1]} -\beta(\mathcal{L}(\alpha^s) - \mathcal{L}^*) + \frac{\theta_1(1 + 4L_g^2)Q_{\max}\beta^2}{\rho} (\mathcal{L}(\alpha^s) - \mathcal{L}^*) \right) \\ & \leq -\frac{\rho}{n\nu\theta_1(1 + 4L_g^2)Q_{\max}} (\mathcal{L}(\alpha^s) - \mathcal{L}^*) \end{aligned} \quad (52)$$

Combining results of Case 1 (48) and Case 2 (52), and taking expectation on both sides w.r.t. the history, we have

$$E[\mathcal{L}(\alpha^{s+1})] - \mathcal{L}(\alpha^s) \leq -\frac{\mu_1}{Q_{\max}n\nu} (\mathcal{L}(\alpha^s) - \mathcal{L}^*). \quad (53)$$

where

$$\mu_1 := \min\left\{\frac{1}{16\theta(\Delta\mathcal{L}^0)}, \frac{\rho}{\theta_1(1 + 4L_g^2)}\right\}.$$

Taking summation of (53) over iterates $s = 1 \dots n$, we have

$$\begin{aligned} E[\mathcal{L}(\alpha^{t+1})] - \mathcal{L}(\alpha^t) & \leq -\frac{\mu_1}{Q_{\max}n\nu} \left(\sum_{s=1}^n \mathcal{L}(\alpha^s) - \mathcal{L}^* \right) \\ & \leq -\frac{\mu_1}{Q_{\max}\nu} (\mathcal{L}(\alpha^{t+1}) - \mathcal{L}^*). \end{aligned} \quad (54)$$

Rearranging terms gives the conclusion. \square

Now we provide proof of Theorem 1 as follows.

Proof. Let $\kappa = 4(1 + \nu)mQ/\mu_{\mathcal{M}}$. By lemma 3, 5, 4 and (30), we have

$$\begin{aligned} & \Delta_d^t - \Delta_d^{t-1} + E[\Delta_p^t] - \Delta_p^{t-1} \\ & \leq \frac{-1}{1 + \kappa} (\mathcal{L}(\alpha^t, \lambda^t) - \mathcal{L}(\bar{\alpha}^t, \lambda^t)) + \frac{2\eta}{\rho} (\mathcal{L}(\alpha^t, \lambda^t) - \mathcal{L}(\bar{\alpha}^t, \lambda^t)) - \frac{\eta}{\tau} \Delta_d^t. \end{aligned} \quad (55)$$

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

$$\begin{aligned} & (\Delta_d^t + E[\Delta_p^t]) - (\Delta_d^{t-1} + \Delta_p^{t-1}) \\ & \leq \frac{-1}{2(1+\kappa)} (\mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\lambda}^t) - \mathcal{L}(\bar{\boldsymbol{\alpha}}^t, \boldsymbol{\lambda}^t)) - \frac{\eta}{\tau} \Delta_d^t \\ & \leq -\min\left(\frac{1}{2(1+\kappa)}, \frac{\eta}{\tau}\right) (\Delta_p^t + \Delta_d^t) \end{aligned}$$

which then leads to the conclusion. \square

The proof for Theorem 2 follows the same line of above reasoning with step (55) replaced by application of Lemma 6 instead of Lemma 5.

7 Appendix C: Implementation details of FMO

7.1 C-1: Indicator Factor

Here we assume $\delta(y_j, \bar{y}_j)$ is constant for $\forall y_j \neq \bar{y}_j$ as in the case of Hamming error. Then we find maximizers of the 4 cases as following

- (i) Visit \mathbf{y}_f in descending order of $v(\cdot)$ to find the first $\mathbf{y}_f: m_{if}(y_i) = 0, m_{jf}(y_j) = 0$.
- (ii) $\forall y_j: m_{jf}(y_j) \neq 0$, visit y_i in descending order of $v(\cdot)$ to find the first $y_i: m_{if}(y_i) = 0$.
- (iii) $\forall y_i: m_{if}(y_i) \neq 0$, visit y_j in descending order of $v(\cdot)$ to find the first $y_j: m_{jf}(y_j) = 0$.
- (iv) Evaluate (26) for $\forall (y_i, y_j): m_{if}(y_i) \neq 0, m_{jf}(y_j) \neq 0$.

Then \mathbf{y}_f^* is returned as label ($\neq \bar{\mathbf{y}}_f$) of maximum gradient (26) among the 4 cases. One can verify the above procedure considers all labels that have potential to be \mathbf{y}_f^* . The complexities for (ii)-(iv) are bounded by $O(\text{nnz}(\mathbf{m}_{if})\text{nnz}(\mathbf{m}_{jf}))$, where $\text{nnz}(\mathbf{m}_{jf}) \leq |\hat{\mathcal{A}}_j^t|$. When BCFW adopts sampling without replacement, we have $|\hat{\mathcal{A}}_f^t| \leq t$. In practice, as t keeps increasing, $|\hat{\mathcal{A}}_f^t|$ converges to a constant that depends on the optimal $\text{nnz}(\boldsymbol{\alpha}_f^*)$. Note $\text{nnz}(\boldsymbol{\alpha}_f^*)$ is equivalent to the number of labels \mathbf{y}_f that attains the maximum of hinge loss (8), which is small in general as long as there are few labels with larger responses than the others.

Define $\mathcal{Y}_{NZ} = \{\mathbf{y}_f | m_{if}(y_i) \neq 0 \vee m_{jf}(y_j) \neq 0\}$ as the set of labels with messages from one of the variables involved, and $\mathcal{Y}_{Inc} = \{\mathbf{y}_f | \mathbf{y}_f \in \mathcal{Y}_{NZ} \wedge v(\mathbf{y}_f, \mathbf{x}_f) > v(\mathbf{y}'_f, \mathbf{x}_f), \forall \mathbf{y}'_f \notin \mathcal{Y}_{NZ}\}$ as the subset being inconsistently ranked at the top in the multimap. The complexity of step (i) is $O(|\mathcal{Y}_{Inc}|)$, where

$$|\mathcal{Y}_{Inc}| \leq \max(|\mathcal{Y}_i| |\hat{\mathcal{A}}_i^t|, |\mathcal{Y}_j| |\hat{\mathcal{A}}_j^t|), \quad (56)$$

which is sublinear to the size of factor domain $|\mathcal{Y}_f| = |\mathcal{Y}_i| |\mathcal{Y}_j|$. Although the bound (56) is already sublinear to $|\mathcal{Y}_f|$, it is a *very loose bound*. In our experiments, we observed the average number of elements being visited at stage (i) is no more than 5 for problems of $|\mathcal{Y}_f|$ up to 10^7 , presumably because the inconsistency between factors is small in real applications.

7.2 C-2: Binary-Variable Interaction Factor

Similar to Appendix C-1, we're trying to find active factors with largest gradient. Here is the procedure.

- (i) Visit \mathbf{y}_f in descending order of $v(\cdot)$ to find the first $\mathbf{y}_f: i \notin \mathcal{A}, j \notin \mathcal{A}$.
- (ii) $\forall j: j \notin \mathcal{A}$, visit i in descending order of $v(\cdot)$ to find the first $i: i \notin \mathcal{A}$.
- (iii) $\forall i: i \notin \mathcal{A}$, visit j in descending order of $v(\cdot)$ to find the first $j: j \notin \mathcal{A}$.
- (iv) Compute gradient for $\forall (i, j): i \in \mathcal{A}, j \in \mathcal{A}$.

A similar reasoning as C-1 applies here for complexity analysis.