Constrained Submodular Minimization Towards Missing Labels and Class Imbalance in Multi-label Learning (Supplementary Material)

Baoyuan Wu
KAUST, Saudi Arabia

Siwei Lyu
SUNY-Albany, NY USA

Bernard Ghanem
KAUST, Saudi Arabia

Here we provide the detailed proofs for the Propositions in the main manuscript. For clarity, here we rewrite the original objective function:

$$\min_{\mathbf{Z}\in\{-1,1\}} \ \mathrm{tr}(\mathbf{Y}_P^\top(\mathbf{Y}-\mathbf{Z})) + \beta\mathrm{tr}(\mathbf{Z}\mathbf{L}_X\mathbf{Z}^\top) + \gamma\mathrm{tr}(\mathbf{Z}^\top\mathbf{L}_C\mathbf{Z}),$$
$$\text{s.t.} \quad \overline{v}_1^l\mathbf{1}_n^\top \le \mathbf{1}_m^\top\mathbf{Z} \le \overline{v}_1^u\mathbf{1}_n^\top, \overline{\boldsymbol{v}}_2^l \le \mathbf{Z}\mathbf{1}_n \le \overline{\boldsymbol{v}}_2^u. \qquad (1)$$

The objective function is denoted as $F(\mathbf{Z})$, and the constraint space is defined as $\Omega_0 = \{\mathbf{Z}|\overline{v}_1^l\mathbf{1}_n^\top \le \mathbf{1}_m^\top\mathbf{Z} \le \overline{v}_1^u\mathbf{1}_n^\top, \overline{\boldsymbol{v}}_2^l \le \mathbf{Z}\mathbf{1}_n \le \overline{\boldsymbol{v}}_2^u\}$, and $\Omega_1 = \{-1,1\}\cap\Omega_0$.

1 Proof to Proposition 1

Before presenting the proof of Proposition 1, we firstly introduce some notations and Lemmas. The matrix variables can be transformed as follows:

$$\begin{aligned}
&\mathbf{z} = \mathrm{vec}(\mathbf{Z}) = [\mathbf{Z}_{11},\ldots,\mathbf{Z}_{m1},\ldots,\mathbf{Z}_{mn}]^\top \in \{-1,+1\}^{mn},\\
&\mathbf{y} = \mathrm{vec}(\mathbf{Y}) = [\mathbf{Y}_{11},\ldots,\mathbf{Y}_{m1},\ldots,\mathbf{Y}_{mn}]^\top \in \{-1,0,+1\}^{mn},\\
&\mathbf{y}_p = \mathrm{vec}(\mathbf{Y}_P) = [\mathbf{P}_{11}\mathbf{Y}_{11},\ldots,\mathbf{P}_{mn}\mathbf{Y}_{mn}]^\top \in \mathrm{R}^{mn\times 1},\\
&\mathbf{W} = \beta\cdot\mathbf{W}_X^\top\otimes\mathbf{I}_m + \gamma\cdot\mathbf{I}_n\otimes\mathbf{W}_C \in \mathrm{R}^{mn\times mn},\\
&\mathbf{L} = \beta\cdot\mathbf{L}_X^\top\otimes\mathbf{I}_m + \gamma\cdot\mathbf{I}_n\otimes\mathbf{L}_C \in \mathrm{R}^{mn\times mn}, \qquad (2)\\
&\mathbf{H}_1 = [\mathbf{1}_m^\top,\mathbf{0}_{(n-1)m}^\top;\mathbf{0}_m^\top,\mathbf{1}_m^\top,\mathbf{0}_{(n-2)m}^\top;\ldots;\mathbf{0}_{(n-1)m}^\top,\mathbf{1}_m^\top]\\
&\quad\in\{0,1\}^{n\times mn},\\
&\mathbf{H}_2 = [\mathbf{I}_m,\ldots,\mathbf{I}_m]\in\{0,1\}^{m\times mn},
\end{aligned}$$

where $\otimes$ denotes the Kronecker product (Zehfuss 1858). Then Problem (1) can be reformulated as follows:

$$\begin{aligned}
&\min_{\mathbf{z}\in\{-1,1\}^{mn}} \sum_i^{mn}\ell(z_i,y_i) + \frac{1}{2}\sum_{i,j}^{mn}\mathbf{W}(i,j)\big[\frac{z_i}{\sqrt{\mathbf{d}_i}} - \frac{z_j}{\sqrt{\mathbf{d}_j}}\big]^2\\
\equiv &\min_{\mathbf{z}\in\{-1,1\}^{mn}} F(\mathbf{z}) = \mathbf{y}_p^\top(\mathbf{y}-\mathbf{z}) + \mathbf{z}^\top\mathbf{L}\mathbf{z}, \qquad (3)\\
&\text{s.t.} \quad [\overline{v}_1^l\mathbf{1}_n;\overline{\boldsymbol{v}}_2^l] \le [\mathbf{H}_1;\mathbf{H}_2]\mathbf{z} \le [\overline{v}_1^u\mathbf{1}_n;\overline{\boldsymbol{v}}_2^u].
\end{aligned}$$

Note that $\mathbf{W}$ can be understood as the similarity matrix of a large graph $\mathcal{G} = \{\mathcal{V},\mathcal{E}\}$ with mn nodes and $mn_{e_X} + nm_{e_C}$ edges. $\mathcal{G}$ consists of $\mathcal{G}_X$ and $\mathcal{G}_C$: first, treating each entry in $\mathbf{Z}$ as a node, then there are mn nodes; Second, we copy the edges among instances $\mathcal{E}_X$ for each class (building connections between the entries within the same row, for every row of $\mathbf{Z}$); last, we copy the edges among classes $\mathcal{E}_C$ for each instance (building connections between the entries within the same column, for every column of $\mathbf{Z}$). $\mathbf{d}_i \neq \sum_j^{mn}\mathbf{W}(i,j)$: if $(i,j)\in\mathcal{E}_X$, then $\mathbf{d}_i = \mathbf{d}_X(\hat{i})$, with $\hat{i}$ being the instance index (i.e., the column index of $\mathbf{Z}$) corresponding to the node i in $\mathcal{G}$; similarly, if $(i,j)\in\mathcal{E}_C$, then $\mathbf{d}_i = \mathbf{d}_C(\hat{i})$. That means we normalize $\mathbf{W}(i,j)$ by the sum of instance-level neighbors (in the same column) or class-level neighbors (in the same row), rather than the sum of all neighbors. As a result, this problem is a partially normalized graph-cut problem. Interestingly, the formulation in (3) is exactly the same as that of the standard GSSL problem (Zhu 2006). The only difference is that $\mathbf{L}$ is not a normalized Laplacian matrix in (3). Please see Lemma 1.

Lemma 1. $\mathbf{L}$ *matrix satisfies the following conditions:*

1. $\mathbf{L}$ *is not a normalized graph Laplacian matrix;*
2. *The off-diagonal entries of* $\mathbf{L}$ *are non-positive, i.e.,* $\forall i \neq j$, $\mathbf{L}(i,j)\le 0$*;*
3. $\mathbf{L}$ *is positive semi-definite (PSD).*

Proof. 1. It is easy to know the diagonal entries of $\mathbf{L}$ are $\beta\mathbf{L}_X(i,i) + \gamma\mathbf{L}_C(j,j), i = 1,\ldots,n, j = 1,\ldots,m$. As both $\mathbf{L}_X$ and $\mathbf{L}_C$ are normalized Laplacian matrix, then $\mathbf{L}_X(i,i) = \mathbf{L}_C(j,j) = 1$, such that $\mathbf{L}(r,r) = \beta+\gamma, r = 1,\ldots,mn$. Since β and γ are two user-defined parameters, their summation is not always equivalent to 1. Thus $\mathbf{L}$ is not a normalized graph Laplacian matrix.

2. As both $\mathbf{L}_X$ and $\mathbf{L}_C$ are normalized Laplacian matrix, all of their off-diagonal entries are non-positive. According to the definition of Kronecker product, we know the off-diagonal entries of both $\beta\cdot\mathbf{L}_X^\top\otimes\mathbf{I}_m$ and $\gamma\cdot\mathbf{I}_n\otimes\mathbf{L}_C$ are non-positive. Thus $\forall i\neq j$, $\mathbf{L}(i,j)\le 0$ holds.

3. Given two square matrix $\mathbf{S}_1 \in \mathrm{R}^{m\times m}$ and $\mathbf{S}_2 \in \mathrm{R}^{n\times n}$, their eigenvalues are denoted as $\lambda_1,\ldots,\lambda_m$ and $\mu_1,\ldots,\mu_n$. According to the property of Kronecker product, the eigenvalues of $\mathbf{S}_1\otimes\mathbf{S}_2$ are $\lambda_i\mu_j, i = 1,\ldots,m; j = 1,\ldots,n$. In Equation (2), for the first term, $\mathbf{L}_X^\top$ is PSD and I_m is positive definite (PD). Obviously all eigenvalues of $\mathbf{L}_X^\top\otimes\mathbf{I}_m$ are non-zero values, such that $\mathbf{L}_X^\top\otimes\mathbf{I}_m$ is a PSD matrix. Similarly we can obtain that

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

$\mathbf{I}_n \otimes \mathbf{L}_C$ is also PSD. Finally, as $\mathbf{L}$ is the positive weighted linear combination of two PSD matrices, it is easy to conclude that $\mathbf{L}$ is a PSD matrix.

□

Lemma 2. *Let $Q \in \mathrm{R}^{p\times p}$ and $q \in \mathrm{R}^p$, then the quadratic set function $F(A) = q^\top \mathbf{1}_A + \frac{1}{2}\mathbf{1}_A^\top Q \mathbf{1}_A$ is submodular if and only if the off-diagonal entries of Q are non-positive. $\mathbf{1}_A \in \{0,1\}^p$ denotes the indicator vector of the subset A: if $i \in A$, then $\mathbf{1}_A(i) = 1$, otherwise $\mathbf{1}_A(i) = 0$. (See Proposition 6.3 in (Bach 2013)).*

Lemma 3. *The objective function (3) is equivalent to a submodular set function $F : 2^V \to \mathrm{R}$, with $V = \{1, \ldots, mn\}$.*

Proof. Given a subset $A \subset V$, it can be represented by the indicator vector $\mathbf{1}_A \in \{0,1\}^{mn}$. Obviously we know $\mathbf{z} = 2\mathbf{1}_A - 1$. Substitute it into (3), we obtain

$$\mathbf{y}_p^\top(\mathbf{y}-\mathbf{z}) + \mathbf{z}^\top \mathbf{L}\mathbf{z}$$
$$= (-2\mathbf{y}_p - 4\mathbf{L}\mathbf{1}_{mn})^\top \mathbf{1}_A + \frac{1}{2}\mathbf{1}_A^\top (8\mathbf{L})\mathbf{1}_A + \text{const} = F(A).$$

From Lemma 1, we know that the off-diagonal entries of $8\mathbf{L}$ are non-positive. Then according to Lemma 2, we conclude that the objective function (3) is equivalent to a submodular set function. □

Proposition 1. *Problem (1) is equivalent to a constrained submodular minimization (CSM) problem, and it is NP-hard*[1].

Proof. Lemma has demonstrated the objective function (3) is equivalent to a submodular function. And its constraint $[\overline{v}_1^l \mathbf{1}_n; \overline{\boldsymbol{v}}_2^l] \le [\mathbf{H}_1; \mathbf{H}_2]\mathbf{z} \le [\overline{v}_1^u \mathbf{1}_n; \overline{\boldsymbol{v}}_2^u]$ can be seen as the cardinality bounds on local parts of $\mathbf{z}$. Obviously (3) is also a cut function. As demonstrated in (Queyranne and Visitor 2002), if the objective function is a cut function and submodular, then its minimization with cardinality constraint is NP-hard. The local-part cardinality constraint in Problem (3) is tighter than the cardinality constraint of the whole vector $\mathbf{z}$. Thus we conclude that Problem (3) is NP-hard. As Problem (1) is equivalent to Problem (3), it is also NP-hard. □

2 Proof to Proposition 2

Proposition 2. *The Lovasz extension of objective function (1) is formulated as follows:*

$$f(\mathbf{Z}) = -\mathrm{tr}(\mathbf{Y}_P^\top \mathbf{Z}) + \frac{1}{2}\sum_k^m \sum_{i,j}^n \widehat{\mathbf{W}}_X(i,j)|\mathbf{Z}_{ki} - \mathbf{Z}_{kj}|$$
$$+ \frac{1}{2}\sum_k^n \sum_{i,j}^m \widehat{\mathbf{W}}_C(i,j)|\mathbf{Z}_{ik} - \mathbf{Z}_{jk}| + \textit{const}, \quad (4)$$

where $\widehat{\mathbf{W}}_X(i,j) = 2\beta \mathbf{W}_X(i,j)(\mathbf{d}_X(i)\mathbf{d}_X(j))^{-\frac{1}{2}}$ *and* $\widehat{\mathbf{W}}_C(i,j) = 2\gamma \mathbf{W}_C(i,j)(\mathbf{d}_C(i)\mathbf{d}_X(j))^{-\frac{1}{2}}$. *const* $= \mathrm{tr}(\mathbf{Y}_P^\top \mathbf{Y}) + \frac{\beta}{2}\sum_k^m \sum_{i,j}^n [\mathbf{d}_X^{-\frac{1}{2}}(i) - \mathbf{d}_X^{-\frac{1}{2}}(j)]^2 + \frac{\gamma}{2}\sum_k^n \sum_{i,j}^m [\mathbf{d}_C^{-\frac{1}{2}}(i) - \mathbf{d}_C^{-\frac{1}{2}}(j)]^2$. *Note that here* $\mathbf{Z}$ *indicate continuous variables.*

Proof. We firstly prove that the Lovasz extension of the quadratic term $\mathbf{z}^\top \mathbf{L}\mathbf{z}$ of (3) is

$$\hat{f}(\mathbf{z}) = \frac{1}{2}\sum_{i,j}^{mn} \mathbf{W}_{ij}\big[2(\mathbf{d}_i\mathbf{d}_j)^{-\frac{1}{2}}|z_i - z_j| + (\mathbf{d}_i^{-\frac{1}{2}} - \mathbf{d}_j^{-\frac{1}{2}})^2\big]. \quad (5)$$

Define a set function $F : 2^V \to \mathrm{R}$ corresponding to the quadratic term in (3), we have

$$\hat{F}(A) = \frac{1}{2}\sum_{i,j}^{mn} \mathbf{W}_{ij}\big[\frac{z_i}{\sqrt{\mathbf{d}_i}} - \frac{z_j}{\sqrt{\mathbf{d}_j}}\big]^2, \quad (6)$$

where $A \subset V$, and $\mathbf{z}$ can be seen as the sign vector of A: when $i \in A$, then $z_i = 1$, otherwise $z_i = -1$. Note that the value of the null set is

$$\hat{F}(\emptyset) = \frac{1}{2}\sum_{i,j}^{mn} \mathbf{W}_{ij}\big[\frac{1}{\sqrt{\mathbf{d}_i}} - \frac{1}{\sqrt{\mathbf{d}_j}}\big]^2, \quad (7)$$

which is a constant of $\mathbf{z}$. To facilitate the following proof, we define a modified set function as follows:

$$\overline{F}(A) = \hat{F}(A) - \hat{F}(\emptyset) = \frac{1}{2}\sum_{i,j}^{mn} \overline{\mathbf{W}}_{ij}(z_i - z_j)^2 = \mathbf{z}^\top \overline{\mathbf{L}}\mathbf{z}, \quad (8)$$

where $\overline{\mathbf{W}}_{ij} = \frac{\mathbf{W}_{ij}}{\sqrt{\mathbf{d}_i\mathbf{d}_j}}$, and $\overline{\mathbf{L}} = \overline{\mathbf{D}} - \overline{\mathbf{W}}$ denotes the corresponding unnormalized Laplacian matrix. We have $\overline{F}(\emptyset) = \overline{F}(V) = 0$. Obviously it is a standard cut function. Thus, as demonstrated in Section 6.2 of (Bach 2013), the Lovasz extension of $\overline{F}$ is

$$\overline{f}(\mathbf{z}) = \sum_{i,j}^{mn} \overline{\mathbf{W}}_{ij}|z_i - z_j|. \quad (9)$$

Then adding the constant term $F(\emptyset)$, we obtain $\hat{f}(\mathbf{z}) = \overline{f}(\mathbf{z}) + F(\emptyset)$, i.e., Equation (5). Besides, the Lovasz extension of the linear term $\mathbf{y}_p^\top(\mathbf{y}-\mathbf{z})$ is still the same form. Thus the Lovasz extension of the objective function (3) is

$$f(\mathbf{z}) = \mathbf{y}_p^\top(\mathbf{y}-\mathbf{z}) + \hat{f}(\mathbf{z}). \quad (10)$$

Finally, utilizing the transformations between variables, it is easy to obtain (4). □

3 Proof to Proposition 3

Utilizing Proposition 2, we obtain the following optimization problem:

$$\min_{\mathbf{Z}\in\Omega_2} f(\mathbf{Z}), \quad (11)$$

where $\Omega_2 = [-1,1]^{m\times n} \cap \Omega_0$.

Lemma 4. *Let F be a submodular function and f its Lovasz extension; then we have*

$$\min_{A\subset V} F(A) = \min_{\mathbf{a}\in\{0,1\}^p} f(\mathbf{a}) = \min_{\mathbf{a}\in[0,1]^p} f(\mathbf{a}).$$

(See Proposition 3.7 in (Bach 2013)).

[1]Due to the space limit, the proofs of all propositions in this manuscript will be presented in **supplementary material**.

Proposition 3. *$F(\mathbf{Z})$ and $f(\mathbf{Z})$ satisfy the following conditions:*

$$\min_{\mathbf{Z}\in\Omega_1} f(\mathbf{Z}) \geq \min_{\mathbf{Z}\in\Omega_1} F(\mathbf{Z}) \geq \min_{\mathbf{Z}\in[-1,1]} f(\mathbf{Z}) = \min_{\mathbf{Z}\in\{-1,1\}} F(\mathbf{Z}),$$
$$\min_{\mathbf{Z}\in\Omega_1} f(\mathbf{Z}) \geq \min_{\mathbf{Z}\in\Omega_2} f(\mathbf{Z}) \geq \min_{\mathbf{Z}\in[-1,1]} f(\mathbf{Z}) = \min_{\mathbf{Z}\in\{-1,1\}} F(\mathbf{Z}).$$

Proof. According to Proposition 2 and Lemma 4, as well as a simple transformation from $[-1, 1]$ to $[0, 1]$, it is easy to obtain

$$\min_{\mathbf{Z}\in[-1,1]} f(\mathbf{Z}) = \min_{\mathbf{Z}\in\{-1,1\}} F(\mathbf{Z}). \tag{12}$$

□

As $\Omega_1 \subset \{-1, 1\}$, we have

$$\min_{\mathbf{Z}\in\Omega_1} F(\mathbf{Z}) \geq \min_{\mathbf{Z}\in\{-1,1\}} F(\mathbf{Z}) = \min_{\mathbf{Z}\in[-1,1]} f(\mathbf{Z}).$$

As $\Omega_2 \subset [-1, 1]$, we have

$$\min_{\mathbf{Z}\in\Omega_2} f(\mathbf{Z}) \geq \min_{\mathbf{Z}\in[-1,1]} f(\mathbf{Z}) = \min_{\mathbf{Z}\in\{-1,1\}} F(\mathbf{Z}),$$

As $\Omega_1 \subset \Omega_2$, we have

$$\min_{\mathbf{Z}\in\Omega_1} f(\mathbf{Z}) \geq \min_{\mathbf{Z}\in\Omega_2} f(\mathbf{Z}).$$

Then the only remaining proof is

$$\min_{\mathbf{Z}\in\Omega_1} f(\mathbf{Z}) \geq \min_{\mathbf{Z}\in\Omega_1} F(\mathbf{Z}). \tag{13}$$

We firstly prove the following inequality $f(\mathbf{Z}) = f(\mathbf{z}) \geq F(\mathbf{Z}) = F(\mathbf{z})$ holds for all $\mathbf{z} \in [-1, 1]$, i.e.,

$$\begin{aligned} &f(\mathbf{z}) - F(\mathbf{z}) = \hat{f}(\mathbf{z}) - \mathbf{z}^\top L\mathbf{z} \qquad (14)\\ \equiv &\frac{1}{2}\sum_{i,j}^{mn} \mathbf{W}_{ij}\big[2t_it_j|z_i - z_j| + (t_i - t_j)^2 - (t_iz_i - t_jz_j)^2\big] \geq 0. \end{aligned}$$

We use $t_i = \mathbf{d}_i^{-\frac{1}{2}} > 0$ and $t_j = \mathbf{d}_j^{-\frac{1}{2}} > 0$ for clarity. When $z_i > z_j$, we have

$$\begin{aligned} &2t_it_j|z_i - z_j| + (t_i - t_j)^2 - (t_iz_i - t_jz_j)^2 \qquad (15)\\ =&2t_it_j(z_i - z_j) + (t_i - t_j)^2 - (t_iz_i - t_jz_j)^2\\ =&2t_it_j\big[\sqrt{(1 - z_i^2)(1 - z_j^2)} - (1 - z_i)(1 + z_j)\big]\\ &+ \left[t_i\sqrt{1 - z_i^2} - t_j\sqrt{1 - z_j^2}\right]^2\\ \geq&2t_it_j\big[\sqrt{(1 - z_i^2)(1 - z_j^2)} - (1 - z_i)(1 + z_j)\big]\\ \geq&2t_it_j\big[\sqrt{(1 - z_i)^2(1 + z_j)^2} - (1 - z_i)(1 + z_j)\big] \geq 0. \end{aligned}$$

It is easy to prove that when $z_i \leq z_j$, the inequality (15) still holds. Thus, considering $\mathbf{W}_{ij} \geq 0$, it is easy to know the inequality (14) holds $\forall \mathbf{z} \in [-1, +1]$. Furthermore, as $\Omega_1 \subset [-1, 1]$, the last inequality (13) is proved. Thus all proofs are finished.

References

Bach, F. 2013. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning* 228.

Queyranne, M., and Visitor, I. 2002. An introduction to submodular functions and optimization.

Zehfuss, G. 1858. Über eine gewisse determinante. *Zeitschrift für Mathematik und Physik* 3:298–301.

Zhu, X. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison* 2:3.