Supplementary Material: Extracting Certainty from Uncertainty: Transductive Pairwise Classification from Pairwise Similarities

Tianbao Yang[†], **Rong Jin**[‡♮]
[†]The University of Iowa, Iowa City, IA 52242
[‡]Michigan State University, East Lansing, MI 48824
[♮]Alibaba Group, Hangzhou 311121, China
tianbao-yang@uiowa.edu, rongjin@msu.edu

1 Proof of Theorem 1

We first prove Theorem 1 regarding the perfect recovery of the sub-matrix $\widehat{Z}$, which is essentially a Corollary of the following Theorem for matrix completion.

Corollary 1. *(Theorem 1.1 [2]) Let M be an $n_1 \times n_2$ matrix of rank r with singular value decomposition $U\Sigma V^\top$. Without loss of generality, impose the conventions $n_1 \leq n_2$, U is $n_1 \times r$, and V is $n_2 \times r$. Assume that (i) The row and column spaces have coherences bounded above by some positive μ_0, and (ii) the matrix UV has a maximum entry bounded by $\mu_1\sqrt{r/(n1n2)}$ in absolute value for some positive μ_1. Suppose N entries of M are observed with locations sampled uniformly at random denoted by Σ. Then if*

$$N \geq 32\max\{\mu_1^2, \mu_0\}r(n_1+n_2)\beta\log^2(2n_2)$$

for some $\beta > 1$, the minimizer to the problem

$$\min_{X\in\mathbb{R}^{n_1\times n_2}} \|X\|_{tr} \quad s.\ t. \quad X_{i,j} = M_{i,j},\ \forall (i,j)\in\Sigma$$

is unique and equal to M with probability at least $1 - 6\log(n_2)(n_1+n_2)^{2-2\beta} - n_2^{2-2\beta^{1/2}}$

Proof. of Theorem 1 Let m_k denote the number of examples in $\widehat{\mathcal{D}}_m$ that belongs to the k-th class. Since $\widehat{Z} = \sum_{k=1}^r m_k \frac{\widehat{\mathbf{g}}_k}{\sqrt{m_k}}\frac{\widehat{\mathbf{g}}_k^\top}{\sqrt{m_k}}$ is an eigen-decomposition of $\widehat{Z}$, it is easy to verify that the coherence measure μ_0 and μ_1 is given by

$$\mu_0 = \frac{1}{r\min_{1\leq k\leq r} m_k/m}, \quad \mu_1^2 = \frac{1}{r\min_{1\leq k\leq r}(m_k/m)^2}$$

Using the Chernoff bound, we have

$$\Pr\left(m_k < (1-\epsilon)mp_k\right) < \exp\left(-\frac{\epsilon^2 mp_k}{2}\right)$$

By setting $\epsilon = 1/2$, with a probability at least $1 - \sum_{i=1}^r \exp(-mp_i/8)$, we have

$$m_k/m \geq p_k/2, k = 1,\ldots,r$$

and therefore

$$\max\left(\mu_0, \mu_1^2\right) \leq \frac{2}{r\min\limits_{1\leq i\leq r} p_i^2}$$

Then using Corollary 1 with $\beta = 4$, we have, with a probability at least $1 - 6(2m)^{-6}\log m - m^{-2}$, that the solution to (2) in the paper is unique and equal to $\widehat{Z}$ if

$$|\Sigma| \geq 128\mu_1^2 rm \log^2(2m) \tag{1}$$

We complete the proof by using the union bound and $6(2m)^{-6}\log m \leq m^{-2}$. $\square$

2 Proof of Theorem 2

The foundation of the proof of Theorem 2 and of Theorem 3 is the following Corollary that quantifies how large m is in order to ensure the sub-matrix $\widehat{U}_s \in \mathbb{R}^{m\times s}$ has a full column rank.

Corollary 2. *Let U_s be an $n \times s$ matrix with orthonormal columns with a coherence measure μ_s. Let $\widehat{U}_s$ be a $m \times s$ matrix with rows uniformly sampled from the rows of U_s. If $m \geq \frac{2\mu_s}{(1-\epsilon)^2} s \log\left(\frac{s}{\delta}\right)$, then with a probability at least $1-\delta$, the matrix $\widehat{U}_s$ has full column rank and satisfies*

$$\|(\widehat{U}_s^\top)^\dagger\|_2^2 \leq \frac{n}{\epsilon\ell}$$

where $M^\dagger$ denotes the pesudo inverse of a matrix M.

The above Corollary follows immediately from Lemma 1 in [1].

Now we are ready to prove Theorem 2. Let us review the two steps of the proposed algorithm for estimating the ideal matrix $Z = \sum_{k=1}^r \mathbf{g}_k\mathbf{g}_k^\top$. The first step is to recover the sub-matrix $\widehat{Z} = \sum_{k=1}^r \widehat{\mathbf{g}}_k\widehat{\mathbf{g}}_k^\top$ by matrix completion, for which we assume the recovery is perfect due to Theorem 1. Because we assume the column space of Z lies in the subspace spanned by columns of U_s, therefore we can write $\mathbf{g}_k$ and Z as

$$\mathbf{g}_k = U_s\mathbf{a}_k, k \in [r]$$
$$Z = U_s\left(\sum_{k=1}^{\top} \mathbf{a}_k\mathbf{a}_k^\top\right)U_s^\top \tag{2}$$

Thus, the second step is to estimate $\sum_{k=1}^r \mathbf{a}_k\mathbf{a}_k^\top$. The underlying logic is to estimate $\mathbf{a}_k$ by

$$\widehat{\mathbf{a}}_k = \arg\min_{\mathbf{a}\in\mathbb{R}^s} \|\widehat{\mathbf{g}}_k - \widehat{U}_s\mathbf{a}\|_2^2 = (\widehat{U}_s^\top\widehat{U}_s)^\dagger\widehat{U}_s^\top\widehat{\mathbf{g}}_k, \quad k \in [r] \tag{3}$$

Then, the ideal matrix Z can be estimated by

$$\begin{aligned} Z' &= U_s\left(\sum_{k=1}^r \widehat{\mathbf{a}}_k\widehat{\mathbf{a}}_k^\top\right)U_s^\top \\ &= U_s(\widehat{U}_s^\top\widehat{U}_s)^\dagger\widehat{U}_s^\top\left(\sum_{k=1}^r \widehat{\mathbf{g}}_k\widehat{\mathbf{g}}_k^\top\right)\widehat{U}_s(\widehat{U}_s^\top\widehat{U}_s)^\dagger U_s^\top = U_s(\widehat{U}_s^\top\widehat{U}_s)^\dagger\widehat{U}_s^\top\widehat{Z}\widehat{U}_s(\widehat{U}_s^\top\widehat{U}_s)^\dagger U_s^\top \end{aligned} \tag{4}$$

As a result, to prove $Z' = Z$ amounts to showing $\widehat{\mathbf{a}}_k = \mathbf{a}_k, k \in [r]$. To this end, we focus on the optimization problems in (3). Since $\widehat{Z}$ is a perfect recovery of a sub-matrix in Z under the conditions in Theorem 1, it is safe to assume that $\widehat{\mathbf{g}}_k \in \mathbb{R}^m$ is equal to the entries in $\mathbf{g}_k \in \mathbb{R}^n$ that corresponds to the sampled examples. It indicates that $\mathbf{a}_k, k \in [r]$ are solutions to the problems in (3) due to $\mathbf{g}_k = U_s\mathbf{a}_k$. Therefore, in order to show $\widehat{\mathbf{a}}_k = \mathbf{a}_k, k \in [r]$, it is equivalent to show that $\mathbf{a}_k, k \in [r]$ are the unique minimizers of problems (3). It is sufficient to show the optimization problems in (3) are strictly convex, which follows immediately from Corollary 2 since it implies that $\widehat{U}_s^\top\widehat{U}_s$ is a full rank PSD matrix with a high probability. Then using the union bound, we can complete the proof.

3 Proof of Theorem 3

To prove Theorem 3, we first define the following matrix Z_* :

$$Z_* = U_s\left(\sum_{k=1}^r \mathbf{a}_k^*\mathbf{a}_k^{*\top}\right)U_s^\top$$

where

$$\mathbf{a}_k^* = \arg\min_{\mathbf{a}\in\mathbb{R}^s} \|\mathbf{g}_k - U_s\mathbf{a}\|_2^2 = (U_s^\top U_s)^{-1}U_s^\top \mathbf{g}_k = U_s^\top \mathbf{g}_k$$

We introduce a matrix $E \in \{0,1\}^{n\times m}$ with columns selected from the identity matrix corresponding to the indices of $\widehat{\mathcal{D}}_m$ in $\mathcal{D}_n$. Then, we can write $\widehat{U}_s = E^\top U_s$, $\widehat{\mathbf{g}}_k = E^\top \mathbf{g}_k$, and have the solution to (3) written as

$$\begin{aligned}\widehat{\mathbf{a}}_k &= (\widehat{U}_s^\top \widehat{U}_s)^{-1}\widehat{U}_s^\top \widehat{\mathbf{g}}_k = ([E^\top U_s]^\top E^\top U_s)^{-1}[E^\top U_s]^\top E^\top \mathbf{g}_k \\ &= (U_s^\top EE^\top U_s)^{-1}U_s^\top EE^\top \mathbf{g}_k\end{aligned}$$

where we use inverse in place of pseudo inverse because we assume $\widehat{U}_s^\top \widehat{U}_s$ is a full rank matrix. To proceed, we write $\mathbf{g}_k$ as

$$\mathbf{g}_k = \mathbf{g}_k^\perp + \mathbf{g}_k^\parallel$$

where $\mathbf{g}_k^\parallel = U_s U_s^\top \mathbf{g}_k$ is the projection of $\mathbf{g}_k$ into the subspace spanned by $\mathbf{u}_1,\ldots,\mathbf{u}_s$ and $\mathbf{g}_k^\perp = \mathbf{g}_k - \mathbf{g}_k^\parallel$. Then, we have

$$\begin{aligned}\widehat{\mathbf{a}}_k &= (U_s^\top EE^\top U_s)^{-1}U_s^\top EE^\top U_s U_s^\top \mathbf{g}_i^\parallel + (U_s^\top EE^\top U_s)^{-1}U_s^\top EE^\top \mathbf{g}_k^\perp \\ &= \mathbf{a}_k^* + (U_s^\top EE^\top U_s)^{-1}U_s^\top EE^\top \mathbf{g}_k^\perp \\ &= \mathbf{a}_k^* + (U_s^\top E)^\dagger E^\top \mathbf{g}_k^\perp = \mathbf{a}_k^* + (\widehat{U}_s^\top)^\dagger E^\top \mathbf{g}_k^\perp\end{aligned}$$

Define $\widehat{A} = (\widehat{\mathbf{a}}_1,\ldots,\widehat{\mathbf{a}}_r) \in \mathbb{R}^{s\times r}$ and $A_* = (\mathbf{a}_1^*,\ldots,\mathbf{a}_r^*) \in \mathbb{R}^{s\times r}$. Then we have

$$\begin{aligned}\|\widehat{A} - A_*\|_F &= \sqrt{\sum_{k=1}^r \|\widehat{\mathbf{a}}_k - \mathbf{a}_k^*\|^2} \le \sqrt{\|(\widehat{U}_s^\top)^\dagger\|_2 \sum_{k=1}^r \|\mathbf{g}_k^\perp\|^2} \\ &= \sqrt{\|(\widehat{U}_s^\top)^\dagger\|_2 \sum_{k=1}^r \left\|\mathbf{g}_k - \mathbf{g}_k^\parallel\right\|_2^2} = \sqrt{\|(\widehat{U}_s^\top)^\dagger\|_2 tr\left(\sum_{k=1}^r \left(\mathbf{g}_k - \mathbf{g}_k^\parallel\right)\left(\mathbf{g}_k - \mathbf{g}_k^\parallel\right)^\top\right)}\end{aligned}$$

Note that we can also write $Z_* = \sum_{k=1}^r \mathbf{g}_k^\parallel {\mathbf{g}_k^\parallel}^\top$, then we have

$$Z - Z_* = \sum_{k=1}^r (\mathbf{g}_k - \mathbf{g}_k^\parallel)(\mathbf{g}_k - \mathbf{g}_k^\parallel)^\top + (\mathbf{g}_k - \mathbf{g}_k^\parallel){\mathbf{g}_k^\parallel}^\top + \mathbf{g}_k^\parallel(\mathbf{g}_k - \mathbf{g}_k^\parallel)^\top$$

Due to that $\mathbf{g}_k - \mathbf{g}_k^\parallel$ is perpendicular to $\mathbf{g}_k^\parallel$, we have

$$tr\left(\sum_{k=1}^r \left(\mathbf{g}_k - \mathbf{g}_k^\parallel\right)\left(\mathbf{g}_k - \mathbf{g}_k^\parallel\right)^\top\right) = tr(Z - Z_*)$$

As a result,

$$\|\widehat{A} - A_*\|_F \le \sqrt{\|(\widehat{U}_s^\top)^\dagger\|_2 tr(Z - Z^*)}$$

Then we can bound $\|Z' - Z_*\|_F$ by

$$\begin{aligned}\|Z' - Z_*\|_F &= \left\|\sum_{i=1}^s U_s(\widehat{\mathbf{a}}_i\widehat{\mathbf{a}}_i^\top - \mathbf{a}_i^*[\mathbf{a}_i^*]^\top)U_s^\top\right\|_F = \|\widehat{A}\widehat{A}^\top - A_*A_*^\top\|_F \\ &\le 2\|\widehat{A} - A_*\|_F\|A_*\|_F + \|\widehat{A} - A_*\|_F^2 \\ &\le 2\sqrt{\|(\widehat{U}_s^\top)^\dagger\|_2 tr(Z - Z_*)tr(Z_*)} + \|(\widehat{U}_s^\top)^\dagger\|_2 tr(Z - Z_*)\end{aligned}$$

where the last step follows from the fact $\|A_*\|_F^2 \le tr(Z_*)$. We can further bound $tr(Z - Z_*)$ as follows:

$$\begin{aligned}tr(Z - Z_*) &= tr(Z - U_sU_s^\top ZU_sU_s^\top) = tr(Z(I - U_sU_s^\top)) = tr\left(\sum_{k=1}^r (I - P_{U_s})\mathbf{g}_k\mathbf{g}_k^\top\right) \\ &= \sum_{k=1}^r \|(I - P_{U_s})\mathbf{g}_k\|_2^2 = \varepsilon\end{aligned}$$

We complete the proof by using $\|Z' - Z\|_F \le \|Z' - Z_*\|_F + \|Z_* - Z\|_F$, $\|Z - Z_*\|_F \le tr(Z - Z_*)$, $tr(Z_*) \le tr(Z) = n$, and the result in Corollary 2.

References

[1] A. Gittens. The spectral norm errors of the naive nystrom extension. *CoRR*, abs/1110.5305, 2011.

[2] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.