Revisiting Sparsity Constraint Under High-Rank Property in Partial Multi-Label Learning

Supplementary Material

7. Proof of Theorem 1

We here prove the theorem 1 as follows.

By definition, the noise matrix $\mathbf{N}$ satisfies $\|\mathbf{N}\|_0 \leq \epsilon$, which implies that it contains at most ϵ non-zero entries. Sparse binary matrices can be approximated as low-rank matrices since their rank is constrained by the number of non-zero entries:

$$\text{rank}(\mathbf{N}) \leq \min(\epsilon, n, l). \tag{19}$$

The rank of the difference of two matrices $\mathbf{Y}$ and $\mathbf{N}$ is bounded by the following rank inequality:

$$\text{rank}(\mathbf{Y}_g) = \text{rank}(\mathbf{Y} - \mathbf{N}) \geq \text{rank}(\mathbf{Y}) - \text{rank}(\mathbf{N}). \tag{20}$$

Since $\mathbf{Y}$ is full rank with $\text{rank}(\mathbf{Y}) = \min(n, l)$ and $\text{rank}(\mathbf{N}) \leq \epsilon$, it follows that:

$$\text{rank}(\mathbf{Y}_g) \geq \min(n, l) - \epsilon. \tag{21}$$

One of our key assumptions is that ϵ is a small value that is smaller than n and l. Indeed, many real-world datasets have the property of $n > \epsilon$, such as the original YeastMF, YeastBP, and YeastCC. Besides, regarding synthetic datasets, the approach described in [21] can also yield extremely sparse noise labels—that is, ϵ is much smaller than n. Moreover, consider more realistic scenarios. For high quality real-world multi-label learning dataset like RCV1-V2, and large scale dataset Amazon-titles with more than 1.3M labels, due to industrial-grade quality control, the noise inevitably introduced during manual labeling but tends to be very sparse, which naturally aligns with our assumption $\epsilon < l$.

Therefore, based on the characteristics of widely-used MLL datasets, we find that many cases satisfy $\epsilon < n, l$. Our use of this assumption is primarily motivated by the desire to establish a sound theoretical analysis within a constrained and analyzable framework.

8. Details of Datasets

We here provide the details of datasets used in our experiments in Table 8.

9. Co-occurrence Induced Low-rank Label Matrix

Co-occurrence is a commonly observed phenomenon in multi-label learning (MLL) settings. Specifically, when a label A always appears together with another label B—i.e., if a sample has label A, it must also have label B—this pattern is referred to as label co-occurrence. Such co-occurrence structures naturally induce a low-rank property in the label matrix, as redundant correlations between labels reduce its effective dimensionality.

To examine the robustness of our method under such conditions, we constructed three variants of the Yeast dataset where the ground-truth label matrix is explicitly made low-rank by introducing artificial co-occurrence relationships among labels. The experimental setup follows the protocol introduced in [20]. We then evaluated three approaches: our proposed method, our method augmented with a low-rank constraint, and NLR [34].

As reported in Table 9, our method consistently and significantly outperforms the other two baselines across all metrics, demonstrating its superior performance and resilience even when the label space exhibits strong low-rank characteristics.

10. Limitations

One limitation of this work is that our method has only been evaluated in the context of stand-alone frameworks, without integration into broader deep learning architectures. While the proposed Schirn formulation is designed to be modular and broadly applicable, its effectiveness in more complex or end-to-end neural architectures remains to be explored. In future work, we plan to extend our investigation by applying Schirn to general-purpose deep learning models to further validate its adaptability and more practical utility.

Algorithm 1 The pseudo code of Schirn

Input: partial multi-label dataset $\mathcal{D}$, hyper-parameters α, β and λ;
Output: The weight matrix $\mathbf{W}$;
Initialize $\mathbf{W} = \mathbf{0}_{d\times l}$, $\mathbf{N} = \mathbf{0}_{n\times l}$, $\mathbf{C} = \mathbf{\Lambda} = \mathbf{1}_{n\times l}$, $\mu = 10^{-4}$, $\mu_{\max} = 10$, $\rho = 1.1$, and the maximum number of iteration $iter = 100$;
for $i = 0$ to iter **do**
    Update $\mathbf{W}$ according to Eq. (9);
    Update $\mathbf{N}$ according to Eq. (13);
    Update $\mathbf{C}$ according to Eq. (16);
    Update $\mathbf{\Lambda}$ according to Eq. (17);
    Update μ according to Eq. (18);
end for
Return Result.

Table 8. Characteristics of real-world datasets. Here, n represents the number of samples, d represents the feature dimensions, and l represents the number of labels. Avg. CLs and Avg. GLs represent the average number of candidate labels per sample and that of ground-truth labels per sample, respectively.

Type	Dataset	n	d	l	Avg. CLs	Avg. GLs
Real-world Data Set	Music_emotion	6833	98	11	5.29	2.42
	Music_style	6839	98	10	6.04	1.44
	YeastMF	1408	6139	39	6.54	3.54
	YeastCC	1771	6139	50	9.30	4.30
	YeastBP	3794	6139	217	18.84	8.84
Synthetic Data Set	Scene	2407	294	6	-	1.07
	Birds	645	260	19	-	1.01
	Medical	978	1449	45	-	1.25
	Enron	1702	1001	53	-	3.38
	Chess	1675	585	227	-	2.41
	Philosophy	3971	842	233	-	2.27

Table 9. Comparison of high-rank, low-rank, and NLR settings on Yeast datasets. We report the rank of the label matrix (original and after transformation), average precision (↑), one-error (↓), ranking loss (↓), coverage (↓), and hamming loss (↓), all in percentage format.

Dataset	Condition	rank original	rank new	Avg. Precision ↑	One-error ↓	Ranking Loss ↓	Coverage ↓	Hamming Loss ↓
Yeast-MF	High-rank	36	29	51.2 ± 1.7	48.5 ± 2.8	24.3 ± 1.4	43.4 ± 2.1	11.7 ± 0.5
	Low-rank	–	–	48.8 ± 1.3	50.1 ± 2.2	26.6 ± 1.6	46.9 ± 1.9	11.9 ± 0.6
	NLR	–	–	25.5 ± 1.1	66.0 ± 3.5	46.1 ± 1.7	72.7 ± 2.2	13.1 ± 0.6
Yeast-CC	High-rank	45	37	67.3 ± 2.1	30.6 ± 2.7	15.2 ± 1.7	35.6 ± 2.0	11.4 ± 0.1
	Low-rank	–	–	65.1 ± 2.0	32.4 ± 3.0	16.4 ± 1.7	37.9 ± 1.8	10.9 ± 0.0
	NLR	–	–	24.3 ± 2.2	72.1 ± 2.4	44.1 ± 1.8	76.9 ± 1.8	13.8 ± 0.3
Yeast-BP	High-rank	200	158	44.2 ± 1.0	47.5 ± 1.6	20.1 ± 0.2	48.7 ± 0.7	5.5 ± 0.1
	Low-rank	–	–	41.6 ± 1.0	49.2 ± 2.0	22.2 ± 0.3	53.4 ± 1.2	5.4 ± 0.1
	NLR	–	–	8.8 ± 1.0	88.5 ± 2.1	49.3 ± 4.2	88.8 ± 3.7	6.0 ± 0.1

Table 10. Experimental results on *coverage* (%, lower is better ↓). • denotes the best result among all methods.

Data	r	Schirn	NLR	FPML	PML-LRS	PML-NI	P-MAP	P-VLS	PAKS	GLC	PARD
Music_emotion		40.2 ± 0.7 •	42.9 ± 0.7	44.9 ± 0.5	40.7 ± 0.4	40.8 ± 0.3	42.0 ± 1.1	41.0 ± 0.7	40.7 ± 0.4	40.6 ± 0.4	41.4 ± 0.6
Music_style		19.3 ± 0.4 •	19.7 ± 0.7	23.3 ± 0.5	20.7 ± 0.8	19.8 ± 0.8	20.6 ± 0.7	20.7 ± 0.8	19.9 ± 0.8	20.3 ± 0.8	20.6 ± 0.8
YeastMF		35.3 ± 0.8 •	38.2 ± 0.9	49.1 ± 1.7	52.2 ± 1.7	45.0 ± 2.1	54.4 ± 2.2	59.7 ± 1.9	40.2 ± 0.7	52.6 ± 1.9	38.9 ± 3.1
YeastCC		30.6 ± 0.6 •	31.9 ± 0.9	44.0 ± 2.1	54.5 ± 1.0	45.2 ± 1.6	53.4 ± 3.1	61.6 ± 2.2	34.9 ± 0.8	53.0 ± 1.3	47.8 ± 2.0
YeastBP		42.9 ± 1.0 •	44.8 ± 1.1	50.4 ± 1.5	68.2 ± 1.0	60.4 ± 1.3	67.1 ± 1.0	74.3 ± 1.3	48.5 ± 1.1	68.1 ± 1.1	51.5 ± 1.0
Scene	1	7.7 ± 0.3 •	9.3 ± 0.7	19.5 ± 2.1	15.6 ± 0.8	14.5 ± 0.8	10.2 ± 0.6	10.2 ± 0.6	14.1 ± 1.3	15.5 ± 0.8	15.5 ± 1.9
	2	8.1 ± 0.6 •	9.7 ± 1.0	19.9 ± 1.3	20.1 ± 1.1	18.8 ± 1.2	12.0 ± 1.2	11.9 ± 0.7	19.1 ± 1.7	20.0 ± 1.0	20.1 ± 1.3
	3	9.0 ± 0.8 •	10.8 ± 0.9	21.7 ± 2.2	25.9 ± 0.5	25.0 ± 0.6	17.1 ± 0.5	16.0 ± 0.3	26.0 ± 0.9	25.8 ± 0.5	25.9 ± 0.5
Birds	3	26.1 ± 0.5	15.6 ± 3.0	21.6 ± 3.8	16.8 ± 4.1	15.6 ± 3.7 •	20.7 ± 2.4	21.7 ± 1.6	18.9 ± 3.6	15.7 ± 4.1	22.3 ± 1.9
	7	38.8 ± 2.7	19.4 ± 1.4	22.3 ± 3.2	25.3 ± 3.1	22.2 ± 4.0	23.7 ± 2.7	23.9 ± 3.7	20.2 ± 2.9 •	24.6 ± 3.6	23.2 ± 3.8
	11	40.2 ± 4.0	21.8 ± 3.2	24.8 ± 2.1	26.1 ± 1.5	24.5 ± 2.0	28.3 ± 1.8	28.4 ± 3.3	23.1 ± 1.8 •	25.7 ± 1.2	24.5 ± 2.1
Medical	3	3.5 ± 0.8 •	5.0 ± 1.3	6.6 ± 1.6	6.6 ± 0.6	4.0 ± 1.2	11.7 ± 1.8	11.3 ± 2.4	11.5 ± 1.4	6.0 ± 0.8	4.6 ± 1.2
	7	5.1 ± 0.6 •	6.6 ± 1.3	11.8 ± 1.5	16.1 ± 1.2	6.2 ± 1.2	13.2 ± 1.1	13.4 ± 1.4	13.4 ± 0.7	10.6 ± 0.9	6.7 ± 1.4
	11	5.9 ± 1.0 •	7.8 ± 1.5	12.0 ± 2.0	22.1 ± 1.7	7.3 ± 0.7	14.2 ± 1.5	13.5 ± 2.0	14.2 ± 1.8	13.6 ± 1.7	7.8 ± 0.8
Enron	3	24.5 ± 0.6 •	26.6 ± 2.1	31.2 ± 1.7	46.3 ± 2.2	36.7 ± 2.8	30.6 ± 1.1	38.7 ± 1.7	26.7 ± 1.8	47.3 ± 2.3	29.8 ± 2.7
	7	25.7 ± 0.5 •	28.1 ± 2.2	33.0 ± 2.7	50.7 ± 1.6	39.9 ± 2.6	32.6 ± 1.6	40.2 ± 2.8	29.2 ± 1.9	49.8 ± 1.8	34.2 ± 2.7
	11	32.3 ± 1.0	29.3 ± 1.2 •	35.4 ± 1.4	55.0 ± 1.3	42.7 ± 2.5	34.9 ± 1.2	40.4 ± 1.5	30.6 ± 1.9	52.3 ± 1.5	36.7 ± 2.4
Chess	10	24.3 ± 0.9 •	26.7 ± 1.2	29.9 ± 1.6	41.6 ± 1.1	36.2 ± 1.2	34.7 ± 2.1	40.4 ± 3.0	29.2 ± 0.6	43.2 ± 1.2	26.9 ± 0.8
	20	27.6 ± 0.6 •	29.2 ± 1.4	32.0 ± 1.6	46.5 ± 1.2	40.3 ± 0.9	38.3 ± 1.5	41.2 ± 2.7	32.1 ± 1.3	46.7 ± 1.2	31.6 ± 1.0
	30	29.2 ± 0.8 •	31.0 ± 0.9	33.9 ± 1.9	49.4 ± 0.8	43.1 ± 0.8	41.3 ± 1.3	44.0 ± 1.8	35.0 ± 1.3	49.4 ± 0.7	33.8 ± 1.3
Philosophy	10	25.2 ± 0.6 •	27.1 ± 1.0	29.1 ± 1.0	39.8 ± 1.0	36.3 ± 1.3	34.3 ± 0.9	34.4 ± 1.7	30.0 ± 1.6	40.9 ± 0.9	25.5 ± 1.3
	20	29.6 ± 0.4 •	30.3 ± 0.9	32.3 ± 0.8	43.4 ± 0.6	39.8 ± 0.9	36.0 ± 1.5	36.4 ± 1.6	33.5 ± 1.3	44.0 ± 0.6	30.2 ± 1.6
	30	31.8 ± 0.2 •	32.3 ± 1.2	35.1 ± 1.3	45.3 ± 0.8	41.8 ± 0.9	39.3 ± 1.2	38.5 ± 1.7	36.0 ± 1.6	45.8 ± 0.7	32.7 ± 1.6

Table 11. Experimental results on *hamming loss* (%, the smaller the better ↓). • denotes the best result achieved among all methods.

Data	r	Schirn	NLR	FPML	PML-LRS	PML-NI	P-MAP	P-VLS	PAKS	GLC	PARD
Music_emotion		20.2 ± 0.3 •	22.2 ± 0.4	23.3 ± 0.2	25.6 ± 0.5	21.6 ± 0.2	22.7 ± 0.4	21.2 ± 0.5	21.5 ± 0.4	21.5 ± 0.2	22.0 ± 0.2
Music_style		11.3 ± 0.2 •	12.0 ± 0.3	12.4 ± 0.3	16.1 ± 3.2	11.4 ± 0.2	11.7 ± 0.2	11.9 ± 0.2	11.6 ± 0.2	11.5 ± 0.3	14.4 ± 0.2
YeastMF		9.3 ± 0.1	10.6 ± 0.7	11.7 ± 0.5	11.7 ± 0.3	10.7 ± 0.2	13.3 ± 0.1	11.5 ± 0.5	10.2 ± 0.2	11.3 ± 0.4	9.0 ± 0.2 •
YeastCC		8.3 ± 0.2	8.9 ± 0.9	9.0 ± 0.3	10.8 ± 0.2	9.4 ± 0.3	11.3 ± 0.4	9.5 ± 0.4	7.7 ± 0.3 •	10.2 ± 0.3	8.3 ± 0.2
YeastBP		3.9 ± 0.1 •	4.6 ± 0.2	4.3 ± 0.1	5.2 ± 0.1	4.8 ± 0.0	5.3 ± 0.2	4.0 ± 0.1	4.2 ± 0.1	5.1 ± 0.1	3.9 ± 0.1 •
Scene	1	10.8 ± 0.3 •	16.3 ± 2.5	20.3 ± 1.1	28.8 ± 0.4	15.2 ± 0.8	11.5 ± 0.7	12.8 ± 0.7	14.4 ± 1.2	15.9 ± 0.9	15.7 ± 1.1
	2	11.7 ± 0.8 •	18.6 ± 1.3	20.7 ± 0.6	29.5 ± 1.6	18.7 ± 0.9	13.1 ± 0.7	14.7 ± 0.5	18.5 ± 0.9	19.5 ± 0.7	17.9 ± 0.2
	3	16.2 ± 0.5 •	21.5 ± 2.1	19.8 ± 0.9	29.3 ± 1.3	23.1 ± 0.4	17.3 ± 0.8	20.4 ± 0.9	23.1 ± 0.7	23.6 ± 0.5	17.9 ± 0.2
Birds	3	8.6 ± 0.6 •	10.1 ± 0.6	11.7 ± 1.7	11.5 ± 1.7	10.1 ± 0.6	14.6 ± 1.9	18.8 ± 2.3	11.0 ± 0.5	10.3 ± 0.6	10.9 ± 2.0
	7	9.5 ± 0.4 •	13.3 ± 1.0	12.2 ± 0.8	12.9 ± 4.5	12.9 ± 0.7	17.6 ± 0.9	9.7 ± 0.7	11.4 ± 0.6	13.2 ± 0.4	11.0 ± 2.4
	11	12.2 ± 1.0 •	15.5 ± 1.4	13.4 ± 2.2	15.4 ± 3.3	13.5 ± 0.4	17.2 ± 0.8	13.3 ± 0.9	12.5 ± 0.7	13.5 ± 0.5	13.6 ± 1.6
Medical	3	1.0 ± 0.1 •	1.2 ± 0.2	2.1 ± 0.1	3.9 ± 0.1	1.3 ± 0.1	3.0 ± 0.2	2.1 ± 0.3	2.8 ± 0.1	1.9 ± 0.1	2.6 ± 0.1
	7	1.2 ± 0.2 •	1.3 ± 0.1 •	2.1 ± 0.1	3.8 ± 0.1	1.7 ± 0.1	3.4 ± 0.3	2.4 ± 0.1	2.9 ± 0.2	3.3 ± 0.3	2.6 ± 0.1
	11	1.8 ± 0.1 •	2.3 ± 0.4	2.1 ± 0.1	3.8 ± 0.1	2.2 ± 0.3	4.5 ± 0.7	2.3 ± 0.3	2.8 ± 0.2	4.3 ± 0.4	2.5 ± 0.1
Enron	3	4.7 ± 0.2 •	5.4 ± 0.5	5.7 ± 0.1	11.1 ± 1.4	5.4 ± 0.2	5.5 ± 0.2	6.3 ± 0.2	5.3 ± 0.1	6.7 ± 0.2	5.7 ± 0.1
	7	4.8 ± 0.2 •	5.4 ± 0.6	5.7 ± 0.1	12.0 ± 1.3	6.0 ± 0.2	5.9 ± 0.3	6.4 ± 0.2	5.3 ± 0.1	7.7 ± 0.2	5.7 ± 0.2
	11	4.9 ± 0.2 •	5.6 ± 0.4	5.8 ± 0.2	10.8 ± 1.6	6.7 ± 0.2	5.9 ± 0.2	6.4 ± 0.2	5.3 ± 0.2	8.6 ± 0.2	5.7 ± 0.2
Chess	10	1.0 ± 0.0 •	1.1 ± 0.0	1.1 ± 0.1	9.1 ± 1.7	1.4 ± 0.0	1.4 ± 0.1	1.1 ± 0.0	1.1 ± 0.0	1.7 ± 0.0	1.2 ± 0.0
	20	1.1 ± 0.1 •	1.2 ± 0.0	1.1 ± 0.1 •	6.9 ± 1.3	1.7 ± 0.0	1.5 ± 0.0	1.1 ± 0.0 •	1.2 ± 0.0	2.0 ± 0.0	2.0 ± 0.1
	30	1.1 ± 0.1 •	1.2 ± 0.0	1.2 ± 0.1	4.7 ± 1.3	1.8 ± 0.0	1.6 ± 0.0	1.1 ± 0.0 •	1.2 ± 0.1	2.2 ± 0.0	3.6 ± 0.1
Philosophy	10	1.0 ± 0.0 •	1.1 ± 0.0	1.0 ± 0.0 •	7.2 ± 0.5	1.3 ± 0.0	1.3 ± 0.0	1.0 ± 0.0 •	1.0 ± 0.0 •	1.5 ± 0.0	1.0 ± 0.0 •
	20	1.0 ± 0.0 •	1.1 ± 0.0	1.1 ± 0.0	6.3 ± 0.5	1.6 ± 0.0	1.3 ± 0.1	1.0 ± 0.0 •	1.2 ± 0.0	1.8 ± 0.0	1.0 ± 0.0 •
	30	1.0 ± 0.0 •	1.1 ± 0.0	1.1 ± 0.0	5.0 ± 0.4	1.8 ± 0.0	1.4 ± 0.1	1.1 ± 0.0	1.3 ± 0.0	2.0 ± 0.0	1.0 ± 0.0 •

Table 12. Ablation study of Schirn in terms of *coverage* and *hamming loss* on synthetic datasets (in %). The settings for each synthetic dataset are $r = 1$ for Scene, $r = 3$ for Birds, Medical, and Enron, and $r = 10$ for Chess and Philosophy.

High Rank	Sparsity	Low Rank	**coverage ↓** Scene	Birds	Medical	Enron	Chess	Philosophy	**hamming loss ↓** Scene	Birds	Medical	Enron	Chess	Philosophy
×	✓	×	9.7 ± 0.8	32.4 ± 2.2	5.1 ± 1.1	27.7 ± 1.2	26.4 ± 1.3	27.5 ± 0.9	11.2 ± 0.2	9.2 ± 0.5	1.1 ± 0.1	4.9 ± 0.2	1.1 ± 0.0	1.0 ± 0.0
✓	×	×	29.2 ± 2.2	46.3 ± 3.3	8.6 ± 1.3	60.5 ± 2.0	63.9 ± 5.4	62.4 ± 2.5	16.8 ± 0.3	9.5 ± 0.6	1.1 ± 0.1	5.2 ± 0.1	1.1 ± 0.0	1.0 ± 0.0
×	✓	✓	9.0 ± 0.7	34.7 ± 2.8	5.1 ± 1.0	27.8 ± 1.4	26.2 ± 1.8	27.3 ± 0.3	11.3 ± 0.5	9.7 ± 0.5	1.1 ± 0.2	5.0 ± 0.2	1.1 ± 0.0	1.0 ± 0.0
✓	✓	×	7.7 ± 0.3	26.1 ± 0.5	3.5 ± 0.8	24.5 ± 0.6	24.3 ± 0.9	25.2 ± 0.6	10.8 ± 0.3	8.6 ± 0.6	1.0 ± 0.1	4.7 ± 0.2	1.0 ± 0.0	1.0 ± 0.0

Table 13. Ablation study of Schirn in terms of *one-error* on synthetic data sets. The settings for each synthetic dataset are $r = 1$ for Scene, $r = 3$ for Birds, Medical and Enron, and $r = 10$ for Chess and Philosophy.

High Rank	Sparsity	Low Rank	**one-error** ↓ Scene	Birds	Medical	Enron	Chess	Philosophy
×	✓	×	.282 ± .014	.508 ± .071	.138 ± .022	.226 ± .022	.463 ± .015	.484 ± .006
✓	×	×	.602 ± .020	.550 ± .033	.164 ± .025	.326 ± .021	.992 ± .007	.840 ± .023
×	✓	✓	.278 ± .018	.502 ± .070	.130 ± .026	.238 ± .017	.468 ± .032	.474 ± .014
✓	✓	×	.234 ± .011	.410 ± .013	.115 ± .019	.212 ± .018	.431 ± .016	.462 ± .017