Nonparametric approach for binary matrix completion

Miaoyan Wang, Sep 1, 2020

Assumption 1 (GLM tensor)**.** We call the tensor $\mathcal{Y} = [\![y_\omega]\!]$ is exponential family tensor with bounded variance, if the following two assumptions are met.

1. [GLM density] Conditional on canonical parameter tensor $\Theta = [\![\theta_\omega]\!]$, the tensor entries y_ω's are independent of each other, and $y_\omega|\theta_\omega$ follows a generalized linear model (GLM) with density

$$p(y_\omega|\theta_\omega) = c(y_\omega, \phi)\exp\left(\frac{y_\omega\theta_\omega - b(\theta_\omega)}{\phi}\right),$$

where $b(\cdot)$ is a known function depending on the distribution family of y_ω, $\phi > 0$ is the dispersion parameter, and $c(\cdot)$ is a known normalizing function.
2. [Boundedness] The parameter tensor Θ is bounded, i.e, $\|\Theta\|_\infty \leq \alpha$ for some $\alpha > 0$.

Proposition 1 (sub-Gaussian residuals under bounded variance). Let $\mathcal{Y}$ be a GLM data tensor, and $\mathcal{E} = \mathcal{Y} - b'(\Theta)$ be the residual tensor, where $b'(\cdot)$ denotes the first-order derivative. Under Assumption 1, the entries of $\mathcal{E}$ are independent sub-Gaussian entries with parameter (ϕU), where $U = \max_{|\theta|\leq\alpha} b''(\theta) < \infty$ and $\phi > 0$ is the dispersion parameter in the GLM density.

Proof. It is easy to see that the entries of $\mathcal{E} = [\![\varepsilon_\omega]\!]$ are independent conditional on $\Theta = [\![\theta_\omega]\!]$. Furthermore, we show that ε_ω is a sub-Gaussian random variable under the boundedness condition on θ_ω. For notational convinene, we drop the subscript ω, and simply write ε and θ. By the definition of sub-Gaussian random variable, it suffices to show

$$\mathbb{E}\left[\exp(t\varepsilon|\theta)\right] \leq \exp\left(\frac{\phi U t^2}{2}\right), \quad \text{for all } t \in \mathbb{R}.$$

By the definition of GLM density, we have

$$\begin{aligned}
\mathbb{E}[\exp(t\varepsilon|\theta)] &= \int c(y,\phi)\exp\left(\frac{y\theta - b(\theta)}{\phi}\right)\exp\left[t(y - b'(\theta))\right]dy \\
&= \int c(y,\phi)\exp\left(\frac{y(\theta+\phi t) - b(\theta+\phi t) + b(\theta+\phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right)dy \\
&= \exp\left(\frac{b(\theta+\phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) \\
&\leq \exp\left(\frac{\phi U t^2}{2}\right),
\end{aligned}$$

where the last inequality follows from Taylor expansion and the definition of U. Therefore, ε is sub-Gaussian-(ϕU). $\square$

1 Problem

Suppose that we observe a subset of entries from a binary matrix, $\{y_{ij} \in \{-1,1\}\colon (i,j) \in \Omega\}$, where $\Omega \subset [d_1] \times [d_2]$ is the index set of observed entries. How to predict the unobserved entries $\{y_{ij} \in \{-1,1\}\colon (i,j) \in \Omega^c\}$?

$$\begin{bmatrix} -1 & ? & ? & -1 & ? \\ ? & 1 & ? & ? & ? \\ -1 & ? & ? & -1 & ? \\ ? & ? & -1 & ? & 1 \end{bmatrix} \tag{1}$$

2 Earlier solution

First, we perform probability estimation based on parametric models. Assume y_{ij} are independent Bernoulli random variables with success probabilities $P(y_{ij} = 1)$ for all $(i,j) \in [d_1] \times [d_2]$. We model the probability matrix using the GLM logistic model,

$$\mathbb{P}(y_{ij} = 1) = \frac{e^{\theta_{ij}}}{1 + e^{\theta_{ij}}}, \quad \text{where} \quad \Theta = [\![\theta_{ij}]\!] \in \mathbb{R}^{d_1 \times d_2} \text{ is a rank-}r \text{ matrix.}$$

Define the rank-r maximum log-likelihood estimator $\hat{\Theta}^{\text{MLE}} = [\![\hat{\theta}_{ij}^{\text{MLE}}]\!] = \arg\min_{\Theta \in \mathcal{P}(r,\alpha)} L(\Theta)$, where

$$L(\Theta) = -\sum_{(i,j)\in\Omega} \log(e^{y_{ij}\theta_{ij}} + 1), \quad \text{and} \tag{2}$$

$$P(r,\alpha) = \{\Theta \in \mathbb{R}^{d_1 \times d_2} \colon \text{rank}(\Theta) \le r \text{ and } \|\Theta\|_\infty \le \alpha\}.$$

Second, we perform prediction using plug-in estimates,

$$\hat{y}_{ij} = \text{sign } \hat{\theta}_{ij}^{\text{MLE}}, \quad \text{for all } (i,j) \in \Omega^c.$$

3 New proposal

If our goal is to predict the unobserved entries by two labels {-1,1}, there is no need to estimate the probability. We could directly perform the prediction in a nonparametric fashion. This scenario reduces to a special case of our matrix-valued classification problem.

1. Feature space:

$$\begin{aligned} \mathcal{X} &= \{\boldsymbol{X} \in \{0,1\}^{d_1 \times d_2} \big| \text{only one entry of } \boldsymbol{X} \text{ is one, and others are zero}\} \\ &= \{\boldsymbol{e}_i \otimes \boldsymbol{e}_j : (i,j) \in [d_1] \times [d_2]\}. \end{aligned}$$

2. Outcome space: $\mathcal{Y} \in \{0, 1\}$.
3. Uniform marginal distribution $\mathcal{P}(\boldsymbol{X})$ over $\mathcal{X}$. No other joint distribution assumptions on $P(\boldsymbol{X}, y)$;
4. i.i.d. training set: $\{(\boldsymbol{X}_{ij}, y_{ij}) : (i, j) \in \Omega\}$, where $\boldsymbol{X}_{ij} = \boldsymbol{e}_i \otimes \boldsymbol{e}_j \in \{0, 1\}^{d_1 \times d_2}$ is an indicator matrix specifying the observed index, and $y_{ij} \in \{-1, 1\}$ is the observed label at index (i, j). For example, the features in the training sample for problem (1) are

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 0 & \cdots & 1 & 0 \\ 0 & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \cdots, \quad \boldsymbol{X}_7 = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ 0 & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

5. Define the rank-r large-margin estimator $\hat{\Theta}^{\text{margin}} = [\![\hat{\theta}_{ij}^{\text{margin}}]\!] = \arg\min_{\Theta \in \mathcal{P}(r,\alpha)} L(\Theta)$, where

$$L(\Theta) = \sum_{(i,j)\in\Omega} [1 - y_{ij}\langle \boldsymbol{X}_{ij}, \Theta\rangle]_+\,, \text{ and} \tag{3}$$
$$\mathcal{P}(r, \alpha) = \{\Theta \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\Theta) \le r \text{ and } \|\Theta\|_\infty \le \alpha\}.$$

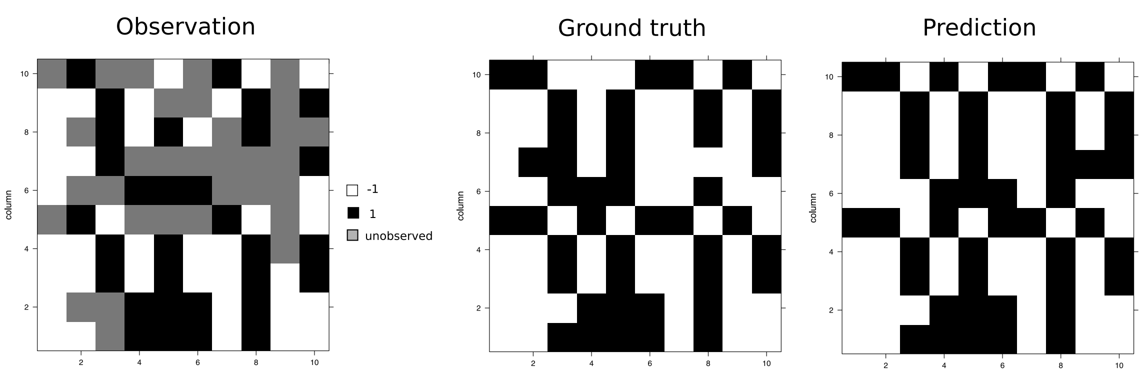
Here, we have omitted the intercept for simplicity.

6. Predict unobserved entries using $\hat{y}_{ij} = \text{sign } \hat{\theta}_{ij}^{\text{margin}}$.
7. Nonparametric probability estimation $\widehat{\mathbb{P}}(y_{ij} = 1 | \boldsymbol{X}_{ij})$ is also possible using a sequence of weighted low-rank classifications (3).

4 Numerical experiments

4.1 Missing data imputation

dimension $d_1 = d_2 = 10$; rank $= 2$; cost $= 1$; observation probability $p = 0.6$.

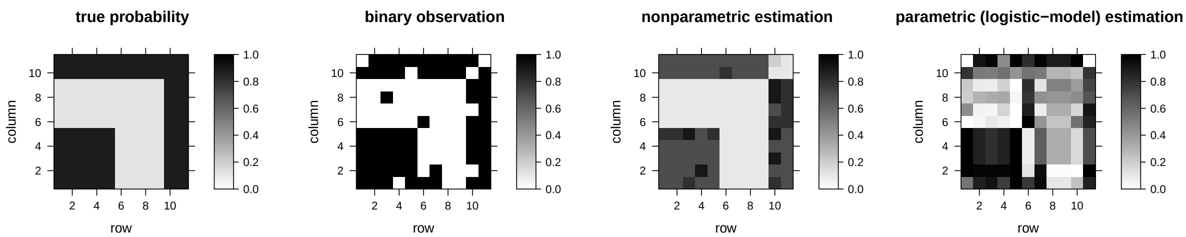


	Unobserved		Observed	
	pred = 1	pred = -1	pred = 1	pred = -1
true = 1	16	3	36	1
true = -1	1	12	1	30

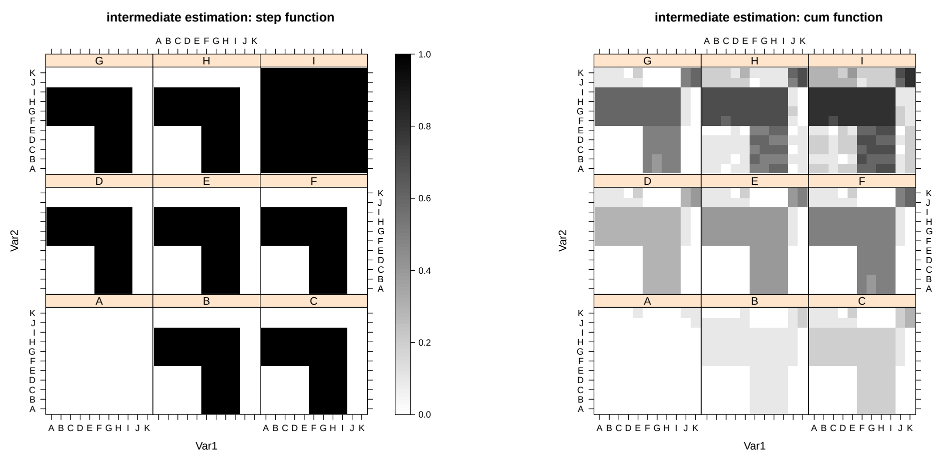
4.2 Probability estimation

dimension $d_1 = d_2 = 11$; cost = 1; observation probability $p = 1$ (no missing data).

Goal: estimate probability matrix $P \in [0,1]^{d_1 \times d_2}$ from binary observations $\boldsymbol{Y} = \{0,1\}^{d_1 \times d_2}$.



Intermediate steps.



Define a sequence of cumulative probability matrices $\frac{1}{10}\sum_{h\leq 1} F_h$ (Panel A), $\frac{1}{10}\sum_{h\leq 2} F_h$ (Panel B), $\ldots$, $\frac{1}{10}\sum_{h\leq 9} F_h$ (Panel I), where $F_h = \mathbb{1}(P \leq \frac{h}{10}) \in \{0,1\}^{d_1\times d_2}$ is the indicator function.

For each matrix entry (i,j), estimate the probability using estimated cumulative probability

$$\hat{P}(i,j) = \frac{1}{10} \operatorname*{arg\,max}_{H\in[10]} \sum_{h\leq H} \widehat{F_h}(i,j), \quad \text{for} \quad (i,j) \in [d_1]\times[d_2],$$

where $\widehat{F_h}(i,j)$ is the predicted class indicator for (i,j)-th entry, based on weighted classifiers.

5 Theory

Definition 1 (Misclassification error)**.** Let $\boldsymbol{Y} = [\![y_{ij}]\!]$, $\boldsymbol{Z} = [\![z_{ij}]\!] \in \{0,1\}^{d_1 \times d_2}$ be two binary matrices. We define the misclassification error (MCE),

$$\text{MCE}(\boldsymbol{Y}, \boldsymbol{Z}) = \frac{1}{d_1 d_2} \sum_{(i,j)\in[d_1]\times[d_2]} \mathbb{1}\{y_{ij} \neq z_{ij}\}.$$

Theorem 5.1 (Generalization error bounds)**.** *Consider a binary target matrix* $\boldsymbol{Y} = [\![y_{ij}]\!] \in \{-1,1\}^{d_1 \times d_2}$ *whose entires are independent realizations from some unknown distributions* $Ber(p_{ij})$*, for all* $(i,j) \in [d_1] \times [d_2]$*. Suppose that we observe a subset of entries,* $\mathbf{Y}_\Omega := \{y_{ij}\}_{(i,j)\in\Omega}$*, where* $\Omega \subset [d_1] \times [d_2]$ *is a random set with* $|\Omega|$ *entries, and each entry in* Ω *is an i.i.d. drawn uniformly from* $[d_1] \times [d_2]$*. Let* $\hat{\Theta} = [\![\hat{\theta}_{ij}]\!] \in \mathcal{P}(r,\alpha)$ *denote any estimator based on the observations* $\mathbf{Y}_\Omega$*, where* r *is the rank bound and* α *is the infinity norm bound. Then, with probability at least* $1-\delta$ *over* $\boldsymbol{Y}$ *and the sample selection* Ω*, the following bound holds uniformly for all* $\hat{\Theta} \in \mathcal{P}(r,\alpha)$*,*

$$\underbrace{\text{MCE}(\boldsymbol{Y}, \text{sign}\ \hat{\Theta})}_{\textit{misclassification error in targeted matrix}} \leq \underbrace{\frac{L}{|\Omega|} \sum_{(i,j)\in\Omega} \text{surrogate-loss}(y_{ij}\hat{\theta}_{ij})}_{\textit{surrogate loss in sample}} + C_1 \alpha \sqrt{\frac{(d_1+d_2)r}{|\Omega|}} + C_2 \sqrt{\frac{\log(3/\delta)}{2|\Omega|}},$$

where where $C_1, C_2 > 0$ *are two universal constants, and* $L > 0$ *is the Lipschitz constant of the surrogate loss,*

$$L = \begin{cases} 1, & \textit{for hinge loss } S(t) = (1-t)_+, \\ \frac{1}{\log 2,} & \textit{for logistic loss } S(t) = \log_2(e^t + 1), \end{cases}$$

In particular, the generalization error of $\hat{\Theta}$ *converges to zero as long as the sample size* $|\Omega| \geq \tilde{\mathcal{O}}(d_{\max} r)$*.*

Corollary 1 (Large-margin estimator)**.** *Consider the same set-up as in Theorem 5.1. Let* $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ *be the optimal estimator in* $\mathcal{P}(r,\alpha)$ *that minimizes the MCE, i.e.,*

$$\Theta^* = \underset{\Theta \in \mathcal{P}(r,\alpha)}{\arg\min}\ \text{MCE}(\boldsymbol{Y}, \text{sign}\ \Theta).$$

Then, for the constrained MLE defined in (2) *and large-margin estimator defined in* (3)*, we have*

$$\text{MCE}(\boldsymbol{Y}, \text{sign}\ \hat{\Theta}^{\text{margin}}) - \text{MCE}(\boldsymbol{Y}, \text{sign}\ \Theta^*) \leq C_1 \alpha \sqrt{\frac{(d_1+d_2)r}{|\Omega|}} + C_2 \sqrt{\frac{\log(1/\delta)}{2|\Omega|}},$$

$$\text{MCE}(\boldsymbol{Y}, \text{sign}\ \hat{\Theta}^{\text{MLE}}) - \text{MCE}(\boldsymbol{Y}, \text{sign}\ \Theta^*) \leq C_1 \alpha \sqrt{\frac{(d_1+d_2)r}{|\Omega|}} + C_2 \sqrt{\frac{\log(1/\delta)}{2|\Omega|}},$$

with probability at least $1-\delta$ *over* $\boldsymbol{Y}$ *and the sample selection* Ω*.*

Remark 1 (Approximation error)**.** What is the total estimation error of sign $\hat{\Theta}$ from the bayes rule? Two sources of error: generalization error + approximation error. The approximation error reaches zero when the "bayes rule" binary matrix is included in the set of candidate sign matrices. Namely, there exists a low-rank, entrywise bounded matrix $\Theta^* \in \mathcal{P}(r,\alpha)$ such that

$$\Theta^* \overset{\text{equal in sign}}{=} [\![p_{ij} - 0.5]\!], \quad \text{or equivalently,} \quad \text{MCE}(\Theta^*, \underbrace{\text{sign}(p_{ij}-0.5)}_{\text{"bayes rule" binary matrix}}) = 0.$$

Remark 2. Given a bayes rule binary matrix, how can we tell whether it is the sign matrix for some low-rank matrix in $\mathbb{P}(r,\alpha)$? For matrix completion problem, the sample size $|\Omega|$ is always smaller than the feature dimension d_1d_2. What does "decision boundary" mean when the feature space is discrete?

Remark 3. If full rank, then $\theta_{ss'}$ = intercept = sample average = $\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} y_{ij}^{\text{test}}$ for all $(s,s') \in \Omega^c$. Non-vanishing MCE unless $|\Omega| \approx d_1d_2$.

Remark 4. MCE vs. MSE. sharpness compared to earlier paper?

Remark 5 (Nonlinear extension)**.**

6 Proofs

Proof of Theorem 5.1. Because the desired conclusion is a uniform bound over all $\hat{\Theta} \in \mathcal{P}(r,\alpha)$, we write Θ in place of $\hat{\Theta}$ for notational convenience. One should note that Θ is a random variable depending on the realizations of the training set $\{y_{ij}\}_{(i,j)\in\Omega}$.

Define the function class $\mathcal{F} = \{f(\boldsymbol{X})\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \Theta\rangle \;\big|\; \Theta \in \mathcal{P}(r,M)\}$. Given the features in the training set, $\{\boldsymbol{X}_{ij} = \boldsymbol{e}_i \otimes \boldsymbol{e}_j : (i,j) \in \Omega\}$, we consider the empirical Rademacher complexity of $\mathcal{F}$ conditional on Ω and the training set. Let $\{\xi_{ij}\}$ a set of i.i.d. Rademacher random variables with equal probability on ± 1, then

$$\mathcal{R}_\Omega(\mathcal{F}) = \frac{2}{|\Omega|}\mathbb{E}_{\xi_{ij}}\left\{\sup_{\Theta\in\mathcal{P}(r,\alpha)} \sum_{(i,j)\in\Omega} \xi_{ij}\Theta_{ij}\right\}. \tag{4}$$

Note that $\Theta \in \mathcal{P}(r,M)$ implies that $\|\Theta\|_{\max} \leq \sqrt{r}\alpha$, where $\|\Theta\|_{\max} = \min_{\Theta=\boldsymbol{U}^T\boldsymbol{V}}\{\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}\}$ denotes the matrix max-qnorm. Therefore, the inequality (4) is upper bounded,

$$\begin{aligned}\frac{2}{|\Omega|}\mathbb{E}_{\xi_{ij}}\left\{\sup_{\Theta\in\mathcal{P}(r,M)} \sum_{(i,j)\in\Omega} \xi_{ij}\Theta_{ij}\right\} &\leq \frac{2}{|\Omega|}\mathbb{E}_{\xi_{ij}}\left\{\sup_{\|\Theta\|_{\max}\leq\sqrt{r}\alpha} \sum_{(i,j)\in\Omega} \xi_{ij}\Theta_{ij}\right\}\\ &\leq c\alpha\sqrt{\frac{r(d_1+d_2)}{|\Omega|}},\end{aligned}$$

where the last inequality follows from Ghadermarzy et al. [2019, Lemma 31].

Using the generalization error inequality in the earlier notes, we have that, with probability at least $1-\delta$ over the sample selection Ω and training data $\{y_{ij}^{\text{train}}\}_{(i,j)\in\Omega}$, the following bound holds uniformly over $\Theta = [\![\theta_{ij}]\!] \in \mathcal{P}(r, M)$,

$$\mathbb{P}\left[y^{\text{test}} \neq \text{sign}\ \langle \boldsymbol{X}^{\text{test}}, \Theta\rangle\right] \leq \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} \text{hinge-loss}(y_{ij}^{\text{train}}, \theta_{ij}) + C_1\alpha\sqrt{\frac{r(d_1+d_2)}{|\Omega|}} + \sqrt{\frac{\log(1/\delta)}{2|\Omega|}}, \quad (5)$$

for some constant $C_1 > 0$, where the probability at the left hand side is taken with respect to $(\boldsymbol{X}^{\text{test}}, y^{\text{test}}) \in \{\boldsymbol{e}_s \otimes \boldsymbol{e}'_s \colon (s,s') \in [d_1]\times[d_2]\} \times \{0,1\}$, independent of training data $\{y_{ij}^{\text{train}}\}_{(i,j)\in\Omega}$.

Now, the i.i.d. uniform sampling assumption implies the mutual independence of the events $\mathbb{1}\{y_{ss'} \neq \text{sign}\ \theta_{ss'}\}$ and marginal uniform distribution $\mathbb{P}(\boldsymbol{X}^{\text{test}} = \boldsymbol{e}_s \otimes \boldsymbol{e}_{s'}) = \frac{1}{d_1 d_2}$ for all $(s,s') \in [d_1]\times[d_2]$. By properties of conditional expectation and concentration inequality, we have, for any $\alpha > 0$,

$$\begin{aligned}\mathbb{P}\left[y^{\text{test}} \neq \text{sign}\ \langle \boldsymbol{X}^{\text{test}}, \Theta\rangle\right] &= \frac{1}{d_1 d_2}\sum_{(s,s')\in[d_1]\times[d_2]} \mathbb{E}_{y_{ss'}^{\text{test}}}\mathbb{1}\left\{y_{ss'}^{\text{test}} \neq \text{sign}\ \theta_{ss'}\right\} \\ &\geq \frac{1}{d_1 d_2}\sum_{(s,s')\in[d_1]\times[d_2]} \mathbb{1}\left\{y_{ss'}^{\text{test}} \neq \text{sign}\ \theta_{ss'}\right\} - C\alpha\sqrt{\frac{1}{d_1 d_2}}, \end{aligned} \quad (6)$$

where the last statement holds with probability at least $1-\exp(-\alpha^2)$ over the test matrix $\boldsymbol{Y}^{\text{test}} = [\![y_{ss'}^{\text{test}}]\!] \in \{0,1\}^{d_1\times d_2}$.

Combining (5) and (6) with $\alpha = \sqrt{\log(1/\delta)}$ yields the uniform bound for all $\Theta \in \mathcal{P}(r,\alpha)$,

$$\text{MCE}(\boldsymbol{Y}^{\text{test}}, \text{sign}\ \Theta) \leq \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} \text{hinge-loss}(y_{ij}^{\text{train}}, \theta_{ij}) + C_1\alpha\sqrt{\frac{r(d_1+d_2)}{|\Omega|}} + C_2\sqrt{\frac{\log(1/\delta)}{2|\Omega|}}, \quad (7)$$

with probability at least $1-2\alpha$ taken jointly over the test data $\boldsymbol{Y}^{\text{test}}$, sample selection Ω, and training data $\{y_{ij}^{\text{train}}\}_{(i,j)\in\Omega}$. Note that in the bound (7), the test data $\boldsymbol{Y}^{\text{test}}$ at the left hand side and training data $\{y_{ij}^{\text{train}}\}_{(i,j)\in\Omega}$ at the right hand side are independent of each other.

We write the target binary matrix $\boldsymbol{Y} = [\![y_{ij}]\!] \in \{0,1\}^{d_1\times d_2}$ as

$$y_{ij} = \begin{cases} y_{ij}^{\text{train}}, & (i,j)\in\Omega, \\ y_{ij}^{\text{test}}, & (i,j)\in\Omega^c. \end{cases}$$

Then, the classification error satisfies

$$\text{MCE}(\boldsymbol{Y}, \text{sign}\ \Theta) = \frac{1}{d_1 d_2}\left\{\sum_{(s,s')\in\Omega^c} \mathbb{1}\left\{y_{ss'}^{\text{test}} \neq \text{sign}\ \theta_{ss'}\right\} + \sum_{(s,s')\in\Omega} \mathbb{1}\left\{y_{ss'}^{\text{test}} \neq \text{sign}\ \theta_{ss'}\right\}\right\}$$

$$
\begin{aligned}
&+ \frac{1}{d_1 d_2} \left\{ \sum_{(s,s')\in\Omega} \mathbb{1}\left\{ y_{ss'}^{\text{train}} \neq \text{sign } \theta_{ss'} \right\} - \sum_{(s,s')\in\Omega} \mathbb{1}\left\{ y_{ss'}^{\text{test}} \neq \text{sign } \theta_{ss'} \right\} \right\} \\
\leq & \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \text{hinge-loss}(y_{ij}^{\text{train}}, \theta_{ij}) + C_1 \alpha \sqrt{\frac{r(d_1+d_2)}{|\Omega|}} + C_2' \sqrt{\frac{\log(1/\delta)}{2|\Omega|}},
\end{aligned}
$$

with probability at least $1-3\delta$, where the last line follows from (7) and the concentration inequality for $\sum_{(s,s')\in\Omega} \left[\mathbb{1}\left\{ y_{ss'}^{\text{train}} \neq \text{sign } \theta_{ss'} \right\} - \mathbb{1}\left\{ y_{ss'}^{\text{test}} \neq \text{sign } \theta_{ss'} \right\} \right]$. $\square$

References

Navid Ghadermarzy, Yaniv Plan, and Özgür Yilmaz. Near-optimal sample complexity for convex tensor completion. *Information and Inference: A Journal of the IMA*, 8(3):577–619, 2019.