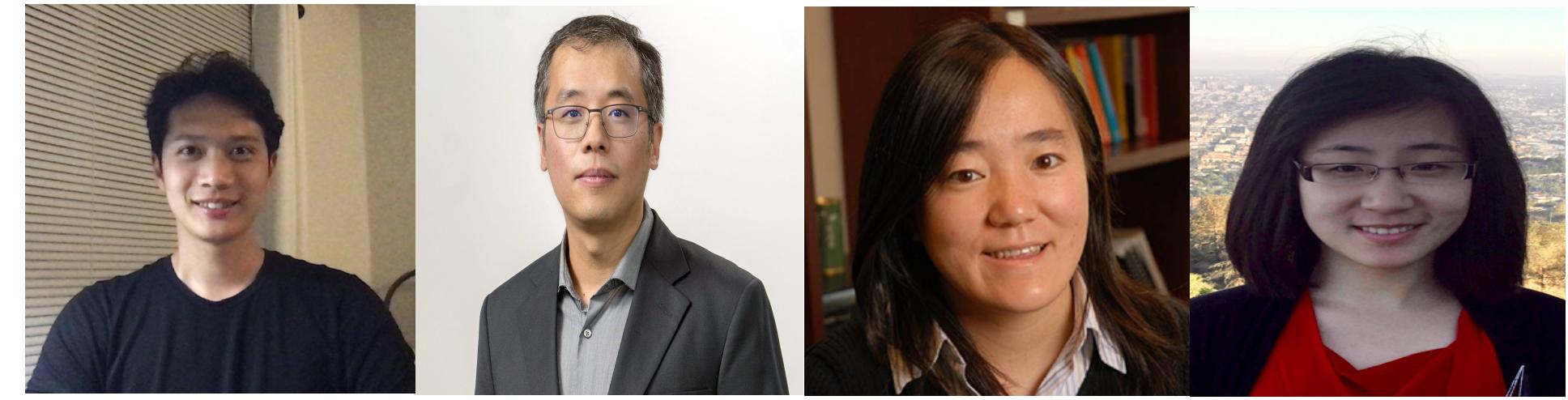


Nonparametric learning with matrix-valued predictors in high dimensions

Chanwoo Lee¹, Lexin Li², Hao Helen Zhang³, and Miaoyan Wang¹

¹University of Wisconsin-Madison ²University of California-Berkeley ³University of Arizona



Problems & Existing methods

Problems: Let $\{(\mathbf{X}_i, y_i) \in \mathbb{R}^{d_1 \times d_2} \times \{-1, 1\}: i = 1, \dots, n\}$ denote an i.i.d. sample from an unknown distribution $\mathcal{X} \times \mathcal{Y}$.

- Classification: How to efficiently classify high-dimensional matrices with limited sample size:

$n \ll d_1 d_2$ = dimension of feature space?

- Regression: How to robustly predict the label probability when little is known about function form of $p(\mathbf{X})$:

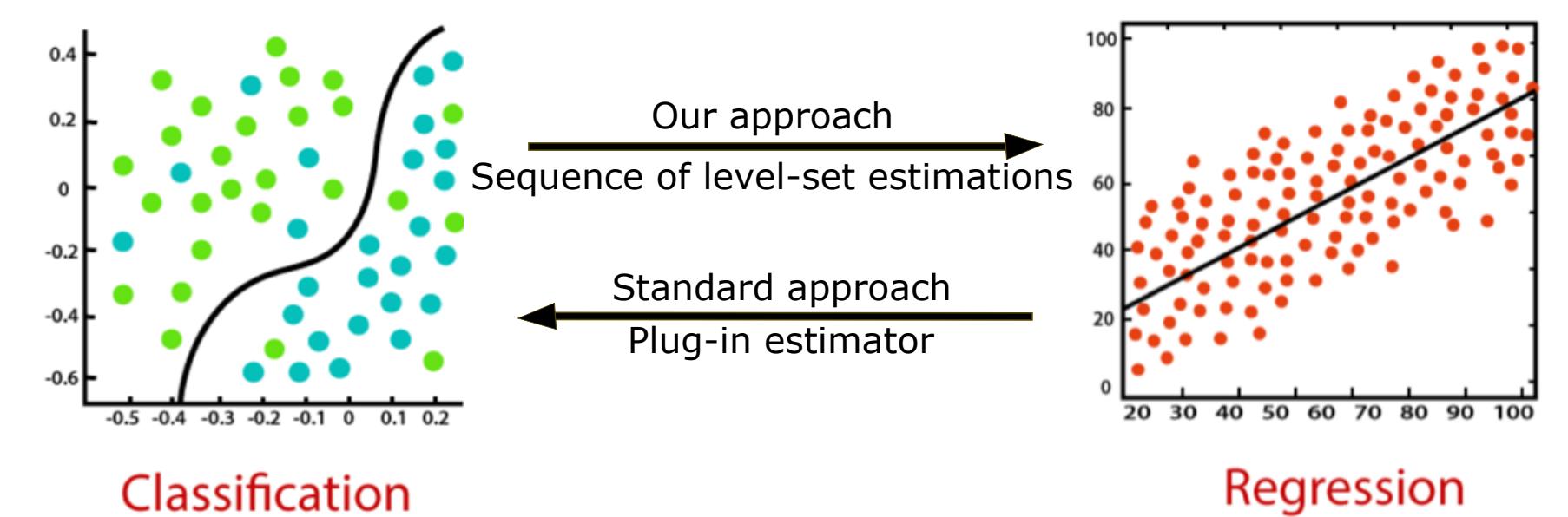
$$p(\mathbf{X}) \stackrel{\text{def}}{=} \mathbb{P}(y = 1 | \mathbf{X})?$$

Existing methods:

- Classification: Decision tree, nearest neighbor, neural network, and support vector machine. However, most methods have focused on vector-valued features.

- Regression: Logistic regression and linear discriminant analysis. However, it is often difficult to justify the assumptions on the function form, especially when the feature space is high-dimensional.

Goal: We propose a nonparametric learning approach with matrix-valued predictors. Unlike classical approaches, our approaches use classification rule to address regression problem.



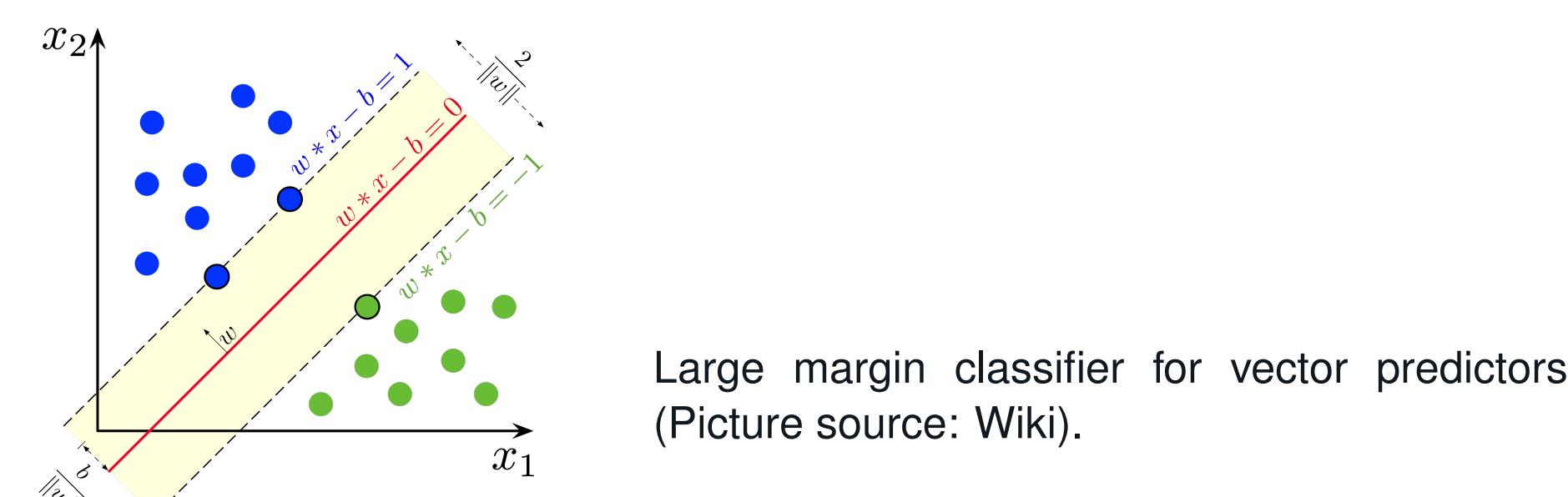
Classification with matrix predictors

- We develop a large-margin classifier for matrix predictors.

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n L(y_i f(\mathbf{X}_i)) + \lambda J(f), \quad (1)$$

- We set $\mathcal{F} = \{f: f(\cdot) = \langle \mathbf{B}, \cdot \rangle \text{ where } \text{rank}(\mathbf{B}) \leq r, \|\mathbf{B}\|_F \leq C\}$, $J(f) = \|\mathbf{B}\|_F^2$, and we choose $L(t)$ to be a large-margin loss, such as hinge loss, logistic loss, etc.

- We also develop nonlinear classifiers for matrix predictors using a new family of matrix-input kernels.



Regression with matrix predictors

- We propose a nonparametric functional estimation method using a sequence of weighted classifiers from (1),

$$\hat{f}_\pi = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \omega_\pi(y_i) L(y_i f(\mathbf{X}_i)) + \lambda J(f),$$

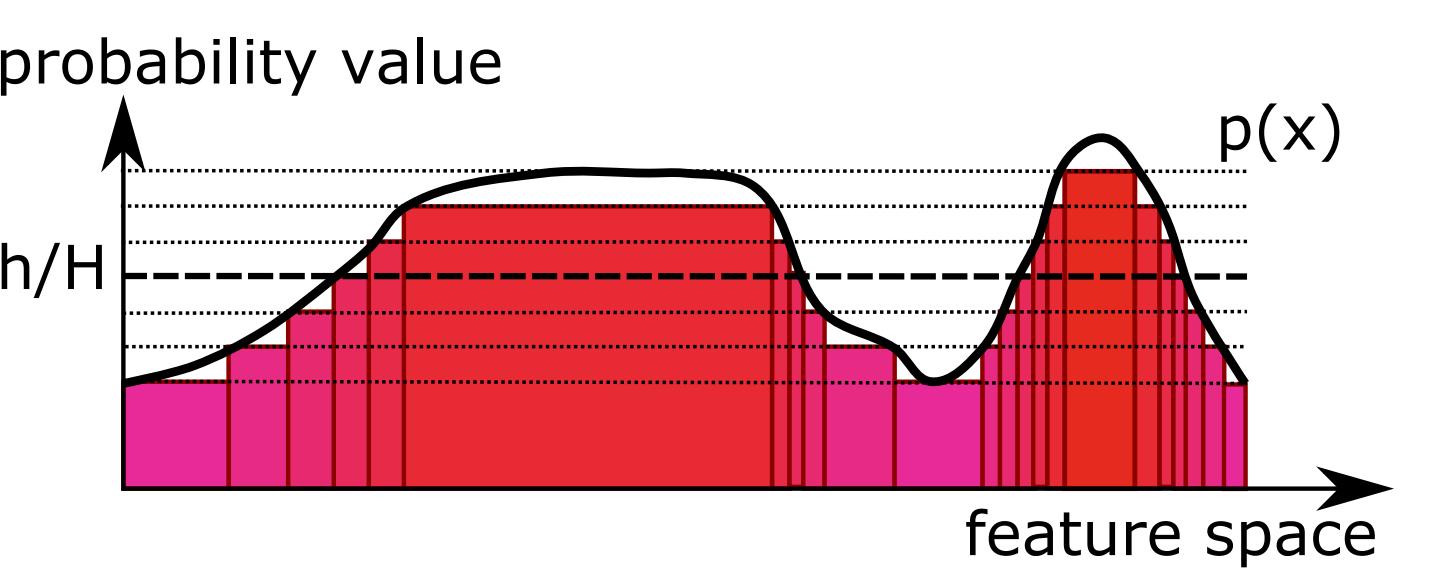
where $\omega_\pi(y) = 1 - \pi$ if $y = 1$ and π if $y = -1$.

- The main idea is to estimate $p(\mathbf{X})$ through two steps of approximations:

$$\begin{aligned} p(\mathbf{X}) &\stackrel{\text{Step 1}}{\approx} \frac{1}{H} \sum_{h \in [H]} \mathbf{1} \left\{ \mathbf{X} : p(\mathbf{X}) \leq \frac{h}{H} \right\} \\ &\stackrel{\text{Step 2}}{\approx} \frac{1}{H} \sum_{h \in [H]} \mathbf{1} \left\{ \mathbf{X} : \underbrace{\text{sign} \left[\hat{f}_{\frac{h}{H}}(\mathbf{X}) \right]}_{\text{estimated decision region from classification}} = -1 \right\}, \end{aligned}$$

where H is a smooth parameter.

- Step 1 is discretization of target function by level sets.



- Step 2 is decision region estimation using a sequence of weighted classifiers.

$$\mathbf{1} \left\{ \mathbf{X} : \underbrace{\text{sign} \left[\hat{f}_\pi(\mathbf{X}) \right]}_{\text{estimated decision region from classification}} = -1 \right\} \xrightarrow{\text{in } p} \mathbf{1} \left\{ \mathbf{X} : \underbrace{\mathbb{P}(Y = 1 | \mathbf{X}) \leq \pi}_{\text{targeted level set}} \right\}.$$

- We provide accuracy guarantees for the above two steps by extending theories in [2] from vectors to high-dimensional matrix predictors.

Algorithms

- We develop an alternating optimization to solve non-convex problem (1).

- We factor the coefficient matrix $\mathbf{B} = \mathbf{U}\mathbf{V}^T$ where $(\mathbf{U}, \mathbf{V}) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$.

Algorithm 1: Classification algorithm with matrix predictors

Input: $(\mathbf{X}_1, y_1), \dots, (\mathbf{X}_n, y_n)$, and prespecified rank r

Initialize: $(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$

Do until converges

 Update \mathbf{U} fixing \mathbf{V} :

$$\mathbf{U} = \arg \min_{\mathbf{U}} \frac{1}{n} \sum_{i=1}^n (1 - \langle \mathbf{U}\mathbf{V}^T, \mathbf{X}_i \rangle)_+ + \lambda \|\mathbf{U}\mathbf{V}^T\|_F^2.$$

 Update \mathbf{V} fixing \mathbf{U} :

$$\mathbf{V} = \arg \min_{\mathbf{V}} \frac{1}{n} \sum_{i=1}^n (1 - \langle \mathbf{U}\mathbf{V}^T, \mathbf{X}_i \rangle)_+ + \lambda \|\mathbf{U}\mathbf{V}^T\|_F^2.$$

Output: $\hat{f}(\mathbf{X}) = \langle \mathbf{U}\mathbf{V}^T, \mathbf{X} \rangle$

Application: Probability matrix estimation

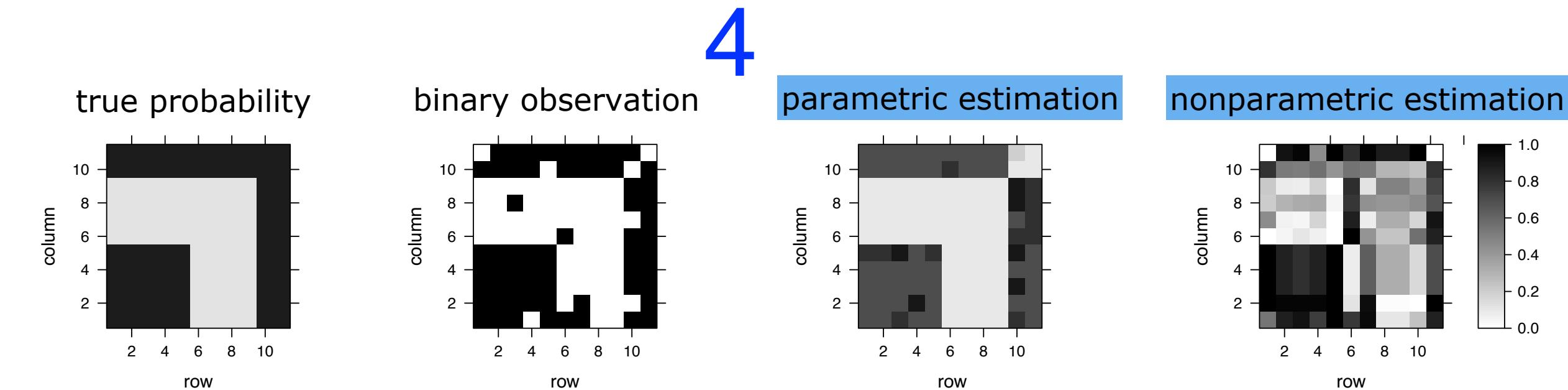
3 lends itself well

- Our method leads itself well to probability matrix estimation problems.
- Goal: Estimate the probability matrix $\mathbf{P} = [p_{ij}] \in [0, 1]$ from binary observations $\mathbf{Y} = [y_{ij}] \in \{0, 1\}$, where $y_{ij} \stackrel{\text{ind.}}{\sim} \text{Ber}(p_{ij})$ for $(i, j) \in [d_1] \times [d_2]$.
- Training set: $\{(\mathbf{X}_{ij}, y_{ij}): (i, j) \in [d_1] \times [d_2]\}$ where

$$[\mathbf{X}_{ij}]_{pq} = \begin{cases} 1 & \text{if } (p, q) = (i, j), \\ 0 & \text{otherwise,} \end{cases}$$

is an indicator matrix with 1 in (i, j) -th position and 0's everywhere.

- We apply our developed methods to estimate $p_{ij} = \mathbb{P}(y_{ij} = 1 | \mathbf{X}_{ij})$.



- Our nonparametric approach provides more robust matrix estimation than parametric approaches [1, 3]

Theoretical results

Theorem 1. Assume that $\{\mathbf{X}_i\}_{i=1}^n$ is a set of i.i.d. Gaussian random matrices with bounded variance. Then with high probability,

$$\mathbb{P}[Y \neq \text{sign}(\hat{f}(\mathbf{X}))] - \mathbb{P}[Y \neq \text{sign}(f^*(\mathbf{X}))] \leq \frac{C \sqrt{r(d_1 + d_2)}}{\sqrt{n}},$$

where f^* is the best predictor in \mathcal{F} .

Theorem 2. Let $\hat{p}: \mathbb{R}^{d_1 \times d_2} \rightarrow [0, 1]$ be the estimated probability function from our method. Under some assumptions on function class \mathcal{F} , penalty parameter λ , and smoothing parameter H , we have

$$\mathbb{E} \|\hat{p} - p\|_1 = \mathcal{O} \left(\left(\frac{\log(n/r(d_1 + d_2))}{(n/r(d_1 + d_2))} \right)^{1/(2-\alpha \wedge 1)} \right),$$

where α is a regularity parameter determined by the true probability function. If $\alpha \leq 1$ and $d_1 = d_2 = d$, we have

$$\mathbb{E} \|\hat{p} - p\|_1 = \mathcal{O} \left(\frac{\log(n/rd)}{(n/rd)} \right).$$

References

- [1] Lee, C. and Wang, M. (2020). Tensor denoising and completion based on ordinal observations. *ICML*.
- [2] Wang, J., Shen, X., and Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167.
- [3] Wang, M. and Li, L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *JMLR*, In press.