Nonparametric Trace Regression in High Dimensions via Sign Series Representation

Chanwoo Lee[1], Lexin Li[2], Hao Helen Zhang[3], and Miaoyan Wang*[1]

[1]Department of Statistics, University of Wisconsin–Madison
[2]Division of Biostatistics, University of California–Berkley
[3]Department of Mathematics, University of Arizona

Abstract

Learning of matrix-valued data has recently surged in a range of scientific and business applications. Trace regression is a widely used method to model effects of matrix predictors and has shown great success in matrix learning. However, nearly all existing trace regression solutions rely on two assumptions: (i) a known functional form of the conditional mean, and (ii) a global low-rank structure in the entire range of the regression function, both of which may be violated in practice. In this article, we relax these assumptions by developing a general framework for nonparametric trace regression models via structured sign series representations of high dimensional functions. The new model embraces both linear and nonlinear trace effects, and enjoys rank invariance to order-preserving transformations of the response. In the context of matrix completion, our framework leads to a substantially richer model based on what we coin as the "sign rank" of a matrix. We show that the sign series can be statistically characterized by weighted classification tasks. Based on this connection, we propose a learning reduction approach to learn the regression model via a series of classifiers, and develop a parallelable computation algorithm to implement sign series aggregations. We establish the excess risk bounds, estimation error rates, and sample complexities. Our proposal provides a broad nonparametric paradigm to many important matrix learning problems, including matrix regression, matrix completion, multi-task learning, and compressed sensing. We demonstrate the advantages of our method through simulations and two applications, one on brain connectivity study and the other on high-rank image completion.

1 Introduction

Matrix-valued data are rising ubiquitously in modern data science applications, for instance, brain neuroimaging analysis, integrative genomics, and sensor network localization. Trace regression is one of the most commonly used approaches for modeling matrix data [Fan et al., 2019, Hamidi and

*corresponding author: miaoyan.wang@wisc.edu.

Bayati, 2019]. The model characterizes the relationship between a scalar response Y and a high dimensional matrix predictor $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^{d_1 \times d_2}$ as

$$Y = \langle \boldsymbol{X}, \boldsymbol{B} \rangle + \varepsilon, \text{ with } \boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2} \text{ and } \operatorname{rank}(\boldsymbol{B}) \leq r, \tag{1}$$

where ε is a zero-mean sub-Gaussian noise, and $r \in \mathbb{N}_+$ is the matrix rank typically assumed fixed and much smaller than $\min(d_1, d_2)$. The function $\boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{X}\boldsymbol{B}^T)$ is called the trace effect, where $\operatorname{tr}(\cdot)$ denotes the matrix trace. Over the last decade, the low-rank trace regression (1) has been studied intensively in numerous contexts, including matrix predictor regression, matrix completion, multi-task learning, and compressed sensing.

- **Matrix predictor regression.** Linear trace regression (1) was first proposed to model a matrix-valued predictor [Zhou and Li, 2014, Wang et al., 2014], and was later generalized to model an exponential family response with a known link function [Wang et al., 2017, Fan et al., 2019].

- **Matrix completion.** In addition to the usual regression setting, another application of trace regression (1) is matrix completion, where the goal is to fill in the missing entries of a partially observed matrix [Cai et al., 2016]. Suppose the predictor space $\mathcal{X}$ consists of basis matrices $\boldsymbol{a}_i^T \boldsymbol{b}_j$ in $\mathbb{R}^{d_1 \times d_2}$, with $\boldsymbol{a}_i \in \mathbb{R}^{d_1}$ (respectively, $\boldsymbol{b}_j \in \mathbb{R}^{d_2}$) being the basis vector with 1 at the i-th (respectively, jth) position and 0 elsewhere. Let $\mathbb{P}_{\boldsymbol{X}}$ be a uniform distribution over $\mathcal{X}$. Then model (1) reduces to a matrix completion problem, $Y_{ij} = \langle \boldsymbol{a}_i^T \boldsymbol{b}_j, \boldsymbol{B} \rangle + \varepsilon_{ij} = B_{ij} + \varepsilon_{ij}$, where $Y_{ij}, B_{ij} \in \mathbb{R}$ denotes the (i, j)-th entry of the data matrix $\boldsymbol{Y}$ and the signal matrix $\boldsymbol{B}$, respectively, for $(i, j) \in \Omega \subset \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$ in the observed index set. Moreover, the model becomes a matrix denosing problem [Yang et al., 2016] when the observation set is complete, i.e, $\Omega = \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$.

- **Multi-task learning.** Another application of trace regression is multi-task learning, where the goal is to predict one task response by leveraging the structural similarities among multiple tasks. Here the predictor space $\mathcal{X}$ consists of only matrices that have a single non-zero row. The multi-task problem collects n observations from d_1 different supervised learning tasks. Each task is modeled as a linear regression with an unknown d_2-dimensional parameter $\boldsymbol{b}_i, i = 1, \ldots, d_1$, and the collection of $\boldsymbol{b}_i$ forms the rows of $\boldsymbol{B}$. The model exploits similarities among multiple tasks to predict the response of the i-th task [Caruana, 1997, Fan et al., 2019].

- **Compressed sensing.** Compressed sensing is also a special application of trace regression, where the goal is to recover the structured matrix $\boldsymbol{B}$ from multiple linear combinations of the entry observations. The space $\mathcal{X}$ is the family of measurement matrices given the sampling schemes. For example, Gaussian ensembles use random matrices $\boldsymbol{X}$ with i.i.d. entries from a standard normal distribution [Candes and Plan, 2011], while factorized ensembles use rank-1 matrices $\boldsymbol{X} = \boldsymbol{u}\boldsymbol{v}^T$ for two random vectors $\boldsymbol{u} \in \mathbb{R}^{d_1}, \boldsymbol{v} \in \mathbb{R}^{d_2}$ [Recht et al., 2010].

In this article, we propose and study a nonparametric extension of the trace regression model (1), which encompasses all above matrix learning problems. Particularly, we illustrate our method with two common problems, i.e., matrix predictor regression and matrix completion.

1.1 Inadequacy of low-rank trace regression

The existing trace regression model (1) and its variants rely on two key assumptions: the relationship between $\mathbb{E}(Y|\boldsymbol{X})$ and the trace effect is known a priori through some link function, and the matrix effect is encoded by a global low-rank matrix $\boldsymbol{B}$ in the entire function range. However, despite the popularity of trace regressions, these assumptions are stringent and may often be violated in practice. Next, we use two examples to illustrate the limitations of the classical low-rank trace regression. We present the pitfall in the context of matrix completion, and similar phenomena also occur in general matrix predictor regression.

In the first example, we show the sensitivity of low-rank matrix models to order-preserving transformations. Let $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{V} \in \mathbb{R}^{d\times d}$ be a rank-5 matrix, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times 5}$ consists of i.i.d. standard normal entries and $d = 50$. Now suppose a monotonic transformation $g(b) = (1 + \exp(-cb))^{-1}$ is applied to $\boldsymbol{B}$ entry-wise, and we let $g(\boldsymbol{B})$ be the signal matrix prior to measurements. A small c implies an approximate linearity $b \mapsto -cb$, whereas a large c implies a high nonlinearity $b \mapsto \{0, 1\}$. Fig 1(a) shows that the numerical rank of $g(\boldsymbol{B})$ increases rapidly with c, rendering the classical low-rank model ineffective. In genomic signal processing and other applications, the matrix of interest often undergoes unknown transformation prior to measurements. The sensitivity makes low-rank models less desirable as the global low-rank structure fails to be preserved through monotonic transformations.

In the second example, we show the failure of the classical low-rank model in representing a structured but high-rank effect. We again consider the matrix completion for simplicity, but this time, from a full-rank signal matrix $\boldsymbol{B} \in \mathbb{R}^{d\times d}$, where the (i, j)-th entry is $\log(1+\max(i, j)/d)$ and $d = 10$. Fig 1(b) shows that $\boldsymbol{B}$ is clearly structured, but is of full-rank that $\text{rank}(\boldsymbol{B}) = d$. The classical low-rank model is again ineffective in this case.

These examples reveal the inadequacy of the conventional low-rank trace model (1) in capturing important yet complex matrix effects. This has motivated us to develop a flexible class of nonparametric trace regression for modeling and estimating nonlinear, local, and possibly high-rank effects for high dimensional matrices. We later revisit these two examples in Section 2, and show how those limitations can be overcome using a richer class of matrix models based on a new concept what we coin as the matrix "sign rank".

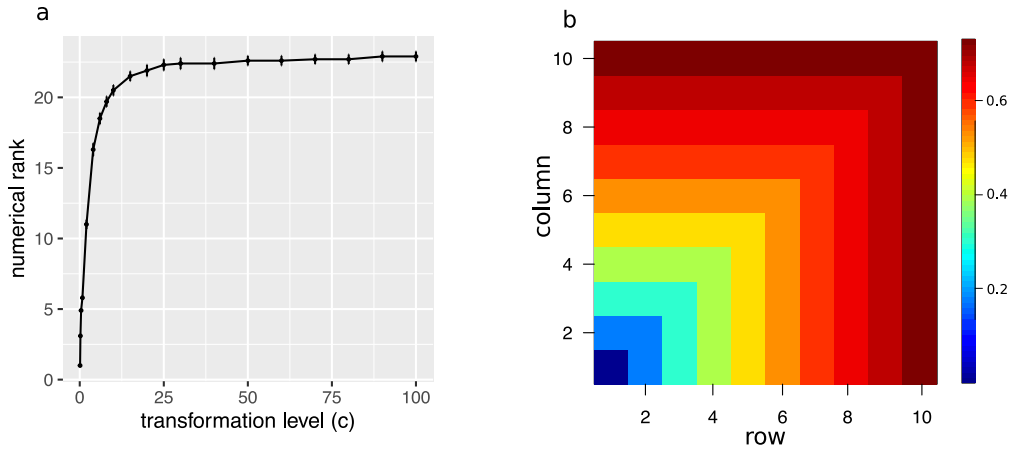

Figure 1: Two examples of high-rank matrix trace models. (a) The numerical rank of the matrix $g(\boldsymbol{B})$ versus c in the transformation, where the numerical rank is defined by $\text{rank}(g(\boldsymbol{B})) = \min\{\text{rank}(\boldsymbol{C}) \colon \|\boldsymbol{C} - g(\boldsymbol{B})\|_F \leq 0.01\|g(\boldsymbol{B})\|_F\}$. The error bar represents standard errors from 10 realizations of $\boldsymbol{B}$. (b) Heatmap of a full-rank matrix $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ with the (i, j)-th entry equal to $\log(1 + \max(i, j))$. In (a), $d = 50$, and in (b), $d = 10$.

1.2 Our proposal and contributions

In this article, we first propose a new notion of low-rank sign representable function, then develop a flexible class of nonparametric trace regression models based on this representation, as well as relevant theory and computational algorithms. Our proposal makes useful contributions on multiple fronts.

First, the proposed work fills a crucial gap between a global parametric model and a local nonparametric model in the literature of matrix modeling. We develop a new nonparametric regression paradigm – structured sign representations – to address the challenges previously difficult or infeasible in trace regressions, especially in the high dimensional regime where $d_1 d_2 \gg n$. Existing literature on matrix regressions almost exclusively focuses on low-rank trace effects in the global scale. However, such a premise often fails, where the rank of global effects may grow with the matrix dimension. By contrast, our proposed model enjoys rank invariance under monotonic transformations, and permits both low-rank and high-rank effects through aggregations of sign representation functions. We show that the low-rank sign functions not only preserve all information for conventional low-rank models, but also provide powerful tools for extracting nonlinear, high-rank trace effects and estimating them accurately. Our framework is flexible and applicable to high-rank matrix learning problems, and it greatly expands the horizon of conventional low-rank matrix models.

Second, we show that the sign function series can be statistically characterized by classification tasks with carefully specified weights. This characterization converts a complex and hard regression problem, *"what is the value of the nonparametric regression function?"* to a series of simpler and easier classification problems, *"does the regression function fall below a threshold?"* Corre-

spondingly, we develop a learning reduction approach to estimate the regression function via a series of classifiers, by leveraging classification solutions from existing state-of-art computational algorithms. Theoretically, we establish the excess risk bounds, estimation error rates, and sample complexities. Particularly, our error bound reveals the well-controlled complexity from sign estimation to regression, where

$$\begin{aligned} \text{sign function error} &\lesssim \underbrace{t_n^{\alpha/(2+\alpha)}}_{\text{classification error}}, \\ \text{regression error} &\lesssim \underbrace{t_n^{\alpha/(2+\alpha)} \log H}_{\text{estimation error inherited from classification}} + \underbrace{\tfrac{1}{H}}_{\text{reduction bias}} + \underbrace{t_n H \log H}_{\text{reduction variance}}, \end{aligned}$$

in which $\alpha \geq 0$ quantifies the smoothness of the nonparametric regression function, $H \in \mathbb{N}_+$ is a resolution parameter that specifies the total number $(2H+1)$ of sign functions to aggregate in our algorithm, $t_n = t_n(d, n) \to 0$ quantifies the convergence rate depending on the specific model, and $d = d_1 = d_2$ for simplicity. In particular, we establish $t_n \asymp n^{-1} \log d$ under a two-way sparse non-parametric trace regression model (see Section 4.1), and $t_n \asymp n^{-1} d$ under a low sign rank non-parametric matrix completion model (see Section 4.2). These results imply that a low sample complexity with respect to the matrix dimension. Note that the sign function estimation reaches a faster $\mathcal{O}(n^{-1})$ rate compared to the $\mathcal{O}(n^{-1/2})$ regression rate when $\alpha = \infty$, which confirms our premise that sign estimation is easier than regression. To our knowledge, these statistical guarantees are among the first for the learning reduction approach in the context of nonparametric matrix regression.

Lastly, we develop an alternating direction method of multipliers (ADMM) algorithm for optimization with a family of large-margin loss functions. From the computational and learning perspectives, the proposed method can be characterized as the **A**ggregation of **S**tructured **SI**gn **S**eries for **T**race regression (**ASSIST**). We show that the **ASSIST** algorithm leverages recent advances in large-margin solvers as well as non-convex optimization for low-rank, two-way sparse matrix learning. As demonstrated in our simulations and real data applications, the **ASSIST** method contributes a new matrix modeling tool of easy interpretability and accurate prediction.

1.3 Related work

Nonparametric learning for matrix data is much more challenging than standard multivariate data. Naively turning a matrix into a vector followed by a classical vector based nonparametric method can destroy rich structural information encoded in the matrix data. Moreover, most nonparametric methods rely on some notion of smoothness in a local neighborhood of the predictors. In the context of matrix regressions, however, the predictor space is huge, rendering the "local smoothness" assumption less practical, which is partially why the topic is barely explored by data with a limited sample size.

Our work is related to but also clearly distinctive from several lines of existing research. The first line is the classical trace regression [Fan et al., 2019, Hamidi and Bayati, 2019]. The key difference is that the existing solutions all adopt a parametric model with a global low-rank structure. By contrast, our method is nonparametric and embraces nonlinear, local, and possibly high-rank effects for high dimensional matrices.

The second line is the recent development of nonparametric methods with matrix-valued or tensor-valued data. In imaging analysis, convolution neural networks (CNNs) have been widely adopted as a nonparametric tool for prediction given matrix-valued images [Goodfellow et al., 2016]. In contrast, our proposal studies not only prediction, but also estimation and interpretability, with the theoretical guarantees. We also numerically compare our method with CNNs. Hao et al. [2019] proposed a sparse additive model with tensor predictors by extending the usual spline basis functions. Zhou et al. [2020] studied tensor predictors and proposed a broadcasting operation to introduce nonlinearity to individual tensor entries. Our nonparametric solution has broader implications than those approaches in estimating local low-rank effects. Our sign series representation of function bridges the gap between regression and classification in high dimensions, and naturally lends the problem to a learning reduction type solution. Moreover, although a matrix can be viewed as a two-dimensional tensor, the problem of nonparametric learning for matrix data itself is more parsimonious and deserves a full investigation. We leave the counterpart problem for nonparametric tensor regression as future research.

The third line is function sign estimation, which is in turn related to classification, or more generally, the level set estimation. The latter problem has a long history in statistics [Tsybakov, 1997] and computational mathematics [Gibou et al., 2018]. Particularly, Wang et al. [2008] proposed a conditional probability estimation method based on support vector machines (SVMs), but their results were restricted to a fixed number of features and vector predictors only. Singh et al. [2009] proposed a tree based method for multiple sets extraction, but their goal was level set estimation instead of function estimation. None of these methods address the regression problem or high dimensional matrix predictors. By contrast, we bridge the problems of level set estimation and nonparametric regression using low-rank sign series representations. Instead of constructing a point-wise function in the domain space, the sign representation partitions the domain space based on the function range. The benefit bears the analogy of Lebesgue versus Riemann integrals in functional analysis, in the sense that the neighborhood is determined by the range space instead of the domain space. The former approach is especially appealing for matrix regressions, where the range space is determined by a simple scalar response, whereas the domain space is huge and high dimensional.

1.4 Notation and organization

We adopt the following notation throughout this article. Let $\mathcal{X} \subset \mathbb{R}^{d_1 \times d_2}$ denote the feature space equipped by some measure $\mathbb{P}_{\boldsymbol{X}}$. For a function $f\colon \mathcal{X} \to \mathbb{R}$, let $\mathrm{sgn} f$ denote its sign function, i.e., $\mathrm{sgn} f(\boldsymbol{X}) = 1$ if $f(\boldsymbol{X}) > 0$ and $\mathrm{sgn} f(\boldsymbol{X}) = -1$ otherwise. Let $\|f\|_1$ denote its L_1 norm, where we define $\|f\|_1 = \mathbb{E}|f(\boldsymbol{X})|$ with the expectation taken with respect to $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$. For a set $A \subset \mathcal{X}$, let $\mathrm{sgn}(\boldsymbol{X} \in A)$ denote the sign function induced by A, i.e., a function taking value 1 on the event $\{\boldsymbol{X} \in A\}$ and -1 otherwise. Let $[n] = \{1, \ldots, n\}$, and $|\cdot|$ denote the cardinality. Let $\|\cdot\|_p$ denote the vector p-norm for $p \geq 0$. For a matrix $\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}$, let $\boldsymbol{B}_i$ denote its i-th row and B_{ij} its (i,j)-th entry. Let $\|\boldsymbol{B}\|_{p,q}$ denote the matrix (p,q)-norm such that $\|\boldsymbol{B}\|_{p,q} = \|\boldsymbol{b}\|_q$, where $\boldsymbol{b} = (\|\boldsymbol{B}_1\|_p, \ldots, \|\boldsymbol{B}_{d_1}\|_p)^T \in \mathbb{R}^{d_1}$ consists of the p-norms for each row of $\boldsymbol{B}$. In particular, let $\|\boldsymbol{B}\|_{1,0} = |\{i \in [d_1]\colon \boldsymbol{B}_i \neq 0\}|$ denote the number of non-zero rows in $\boldsymbol{B}$. Let $\|\boldsymbol{B}\|_F = \sqrt{\langle \boldsymbol{B}, \boldsymbol{B} \rangle}$ denote the matrix Frobenius norm, and $\|\boldsymbol{B}\|_\infty = \max_{(i,j)} |B_{ij}|$ the matrix maximum norm. Denote $a_n \asymp b_n$ if $c_1 \leq \lim_{n\to\infty} a_n/b_n \leq c_2$ for some constants $c_1, c_2 > 0$, and denote $a_n \lesssim b_n$ if $\lim_{n\to\infty} a_n/b_n \leq c$ for some constant $c \geq 0$. Let $\mathcal{O}(\cdot)$ denote the big-O notation, $\tilde{\mathcal{O}}(\cdot)$ the variant that hides the logarithmic factors, and $\mathbb{1}(\cdot)$ the indicator function. Whenever applicable, the basic arithmetic operators are applied to a matrix in an element-wise manner.

The rest of the article is organized as follows. Section 2 presents the low-rank sign representable functions and our nonparametric trace regression model. Section 3 develops the learning reduction approach through weighted classifications, and establishes the corresponding statistical guarantees. Section 4 specializes the general theory to two concrete learning problems, the low-rank sparse matrix predictor regression and the high-rank matrix completion. Section 5 studies the large-margin based estimation and develops an optimization algorithm. Section 6 presents the simulations, and Section 7 two real data applications. Section 8 concludes with a discussion. All technical proofs and additional results are relegated to the Supplementary Appendix.

2 Nonparametric trace regression model

In this section, we present our nonparametric trace regression model. Let $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^{d_1 \times d_2}$ denote the matrix predictor, $Y \in \mathbb{R}$ the scalar response, and $\mathbb{P}_{\boldsymbol{X},Y}$ the joint probability distribution. We consider the model,

$$Y = f(\boldsymbol{X}) + \varepsilon, \tag{2}$$

where $f\colon \mathcal{X} \mapsto \mathbb{R}$ is an unknown regression function of interest, and ε is a mean-zero noise. For a cleaner exposition, we assume the noise is bounded and the range of Y is in $[-1, 1]$; the extension to a sub-Gaussian noise is provided in Section A.2 of the Appendix. In addition, we allow a heterogeneous noise such that ε may depend on $\boldsymbol{X}$. Model (2) therefore incorporates both continuous and binary-valued responses. For instance, we allow the binary regression problem where Y is a $\{0, 1\}$-label from a Bernoulli distribution, in which case, the noise variance depends on the mean,

and f represents the conditional probability, $f(\boldsymbol{X}) = \mathbb{P}(Y = 1|\boldsymbol{X})$. Our goal is to estimate the regression function $f(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$ based on n i.i.d. training samples $(\boldsymbol{X}_i, Y_i)_{i=1,\dots,n}$.

We next introduce the notion of low-rank sign representable function, which is essential to bridge the usual global low-rank trace models to nonparametric local low-rank trace models.

Definition 1 (Rank-r sign representable function)**.** A function $f\colon \mathcal{X} \mapsto [-1, 1]$ is called (r, π)-sign representable, for a given level $\pi \in [-1, 1]$ and a rank $r \in \mathbb{N}_+$, if the function $(f - \pi)$ has the same sign as a rank-r trace function; that is,

$$\operatorname{sgn}(f(\boldsymbol{X}) - \pi) = \operatorname{sgn}(\langle \boldsymbol{X}, \boldsymbol{B}\rangle + b), \quad \text{for all } \boldsymbol{X} \in \mathcal{X}, \tag{3}$$

where $\boldsymbol{B} = \boldsymbol{B}(\pi)$ is a rank-r matrix, and $b = b(\pi)$ is the intercept. A function f is called globally rank-r sign representable, if f is (r, π)-sign representable for all $\pi \in [-1, 1]$. Let $\mathcal{F}_{\text{sgn}}(r)$ denote the rank-r sign representable function family, and let $\Phi(r) = \{\phi\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle + b \;\big|\; \operatorname{rank}(\boldsymbol{B}) \le r, (\boldsymbol{B}, b) \in \mathbb{R}^{d_1 \times d_2} \times \mathbb{R}\}$ denote the rank-r trace function family.

Next, we show that (2) and (3) together form a very general family of models that incorporate most existing matrix regression models, including the low-rank trace regression, single index models, and high-rank matrix completion model.

Example 1 (Generalized trace regression)**.** The linear and generalized trace regression [Zhou and Li, 2014, Wang et al., 2017, Fan et al., 2019] imposes that $f(\boldsymbol{X}) = g(\langle \boldsymbol{X}, \boldsymbol{B}\rangle)$ with a known link function g and a rank-r coefficient matrix $\boldsymbol{B}$. By definition, $\operatorname{sgn}(f(\boldsymbol{X}) - \pi) = \operatorname{sgn}(\langle \boldsymbol{X}, \boldsymbol{B}\rangle - g^{-1}(\pi))$ holds for every π in the function range. Therefore, our model includes the generalized trace regression, i.e, $f \in \mathcal{F}_{\text{sgn}}(r)$. In particular, the usual trace model corresponds to the identity link g. More generally, any monotonic g is allowed as the link function, e.g., the logistic function $g(z) = (1 + \exp(-z))^{-1}$, the arctangent function $g(z) = 1/\pi \arctan(z) + 1/2$, the rectified linear unit (ReLU) function $g(z) = \max(0, z)$, and any inverse cumulative distribution function.

Example 2 (Single index regression model)**.** The monotonic matrix predictor single index model [Balabdaoui et al., 2019, Ganti et al., 2017] assumes a similar form of regression function $f(\boldsymbol{X}) = g(\langle \boldsymbol{X}, \boldsymbol{B}\rangle)$ with a low-rank $\boldsymbol{B}$ and a monotonic g, but the form of g is unknown. By definition, our model family $\mathcal{F}_{\text{sgn}}(r)$ incorporates the single index model and does not require to know g a priori.

Example 3 (Multivariate normal mixture)**.** The prospective model from matrix linear discriminant analysis [Hu et al., 2020] considers a binary response $Y = \{0, 1\}$, and assumes the matrix $\boldsymbol{X}|Y$ follows a Gaussian mixture distribution, $\boldsymbol{X}|\{Y = i\} = \boldsymbol{B}_0 + \boldsymbol{B} \times i + \boldsymbol{E}_i$, $i = 0, 1$, where $\boldsymbol{B}_0$ is an arbitrary baseline matrix, $\boldsymbol{B}$ is a rank-r matrix, and $(\boldsymbol{E}_i)_{i=0,1}$ are two mutually independent noise matrices with i.i.d. standard normal entries. Our model incorporates this model, by noting that $f(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X}) = \text{logistic}(\langle \boldsymbol{B}, \boldsymbol{X}\rangle + b)$ for some $b \in \mathbb{R}$, and thus $f \in \mathcal{F}_{\text{sgn}}(r)$.

Definition 1 leads to another notion, the matrix sign rank, which is important for applying our proposed model for matrix completion as a special nonparametric trace regression. Specifically, for

a given matrix $\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, define its sign rank as:

$$\operatorname{srank}(\boldsymbol{\Theta}) = \min \left\{ \operatorname{rank}(\boldsymbol{\Theta}') \colon \operatorname{sgn}(\boldsymbol{\Theta}') = \operatorname{sgn}(\boldsymbol{\Theta}),\ \boldsymbol{\Theta}' \in \mathbb{R}^{d_1 \times d_2} \right\}.$$

This concept is important in areas such as combinatorics [Cohn and Umans, 2013] and quantum mechanics [De Wolf, 2003], and, to our knowledge, we are the first to exploit this notion for nonparametric learning. To better understand its relation to the proposed nonparametric trace regression, we consider model (3) with the predictor space $\mathcal{X} = \{\boldsymbol{a}_i \boldsymbol{b}_j^T \colon (i,j) \in [d_1] \times [d_2]\}$, and $\boldsymbol{a}_i \in \mathbb{R}^{d_1}, \boldsymbol{b}_j \in \mathbb{R}^{d_2}$ are the basis vectors. For matrix completion, a function f over $\mathcal{X}$ is equivalently represented by a d_1-by-d_2 signal matrix $\boldsymbol{\Theta} = [\![f(\boldsymbol{e}_i \boldsymbol{e}_j^T)]\!]$. Our proposed function family $\mathcal{F}_{\text{sgn}}(r)$ essentially defines a new family of structured matrices with a low sign rank, as shown in the next proposition.

Proposition 1 (Sign-representable function over basis matrices)**.** Consider the predictor space $\mathcal{X} = \{\boldsymbol{a}_i \boldsymbol{b}_j^T \colon (i,j) \in [d_1] \times [d_2]\}$. We represent a bounded function $f \colon \mathcal{X} \to [-1,1]$ by its function values organized as a matrix $\boldsymbol{\Theta} = [\![f(\boldsymbol{a}_i \boldsymbol{b}_j^T)]\!] \in [-1,1]^{d_1 \times d_2}$, for basis vectors $\boldsymbol{a}_i \in \mathbb{R}^{d_1}, \boldsymbol{b}_j \in \mathbb{R}^{d_2}$. If f is rank-r sign representable, then $\max_{\pi \in [-1,1]} \operatorname{srank}(\boldsymbol{\Theta} - \pi) \le r + 1$ (the constant 1 is due to the intercept in (3)). Conversely, if $\max_{\pi \in [-1,1]} \operatorname{srank}(\boldsymbol{\Theta} - \pi) \le r$, then $\boldsymbol{\Theta}$ defines a rank-r sign representable function f.

Define the sign-r representable family for the signal matrix in matrix completion.

$$\mathcal{M}_{\text{sgn}}(r) = \{\boldsymbol{\Theta} \colon \max_{\pi \in [-1,1]} \operatorname{srank}(\boldsymbol{\Theta} - \pi) \le r,\ \|\boldsymbol{\Theta}\|_\infty \le 1\}.$$

The family $\mathcal{M}_{\text{sgn}}(r)$ is a special case of the function family $\mathcal{F}_{\text{sgn}}(r)$ in Definition 1 with $b = 0$ and the predictor space $\mathcal{X} = \{\boldsymbol{a}_i \boldsymbol{b}_j^T \colon (i,j) \in [d_1] \times [d_2]\}$. We next further compare the sign rank with the matrix rank in this setting.

Proposition 2 (Sign-rank vs. matrix rank)**.** Consider the setting in Proposition 1. Then,

(a) $\max_{\pi \in [-1,1]} \operatorname{srank}(\boldsymbol{\Theta} - \pi) \le \operatorname{rank}(\boldsymbol{\Theta}) + 1$.

(b) If $\boldsymbol{\Theta} \in \mathcal{M}_{\text{sgn}}(r)$, then $g(\boldsymbol{\Theta}) / \|g(\boldsymbol{\Theta})\|_\infty \in \mathcal{M}_{\text{sgn}}(r+1)$ for any strictly monotonic function $g \colon \mathbb{R} \to \mathbb{R}$. Here $g(\boldsymbol{\Theta})$ denotes the matrix by applying $g(\cdot)$ to $\boldsymbol{\Theta}$ entry-wise.

(c) For every dimension d, there exists a d-by-d matrix $\boldsymbol{\Theta} \in \mathcal{M}_{\text{sgn}}(2)$ such that $\operatorname{rank}(\boldsymbol{\Theta}) = d$.

Proposition 2 highlights the advantages of using the sign rank in the high dimensional matrix analysis. The first property implies that classical low-rank matrix model is a special case of our low sign rank model. The second property shows that, compared to the matrix rank, the sign rank remains nearly invariant under monotonic transformations, since $\operatorname{srank}(g(\boldsymbol{\Theta})) \le 1 + \operatorname{srank}(\boldsymbol{\Theta})$ for all monotonic functions g. The last property shows that the sign rank can be dramatically smaller than the conventional matrix rank. Therefore, our model $\mathcal{M}_{\text{sgn}}(r)$ is strictly richer than the usual low-rank model.

A key advantage about the sign rank concept is that the low sign rank assumption is more relaxed

and hence more realistic than the classical low matrix rank assumption. We next revisit the high-rank matrix model in Fig 1(a) to show that $\boldsymbol{B}$ is of a high matrix rank but a low sign rank. Meanwhile, we provide some additional examples of low sign rank matrices in Section A.1 of the Appendix, including matrices with repeating patterns [Chan and Airoldi, 2014], banded matrices, and the identity matrix.

Example 4 (Single index model based matrix completion)**.** For the model in Fig 1(a), $g(\boldsymbol{B})$ is a low sign rank matrix because $\text{srank}(g(\boldsymbol{B}) - \pi) \leq 1 + \text{rank}(\boldsymbol{B}) = 6$ for all π in the function range. However, $g(\boldsymbol{B})$ itself is often high-rank as shown in Fig 1(a).

Example 5 (High-rank matrix completion model)**.** For the model in Fig 1(b), the matrix $\boldsymbol{B} = [\![\log(1 + \max(i, j)/d)]\!]$ is full-rank. Remarkably, this high-rank matrix belongs to our sign representable function with rank 2, i.e., $\boldsymbol{B} \in \mathcal{M}_{\text{sgn}}(2)$. This is because $\text{srank}(\boldsymbol{B} - \pi) = \text{srank}(\bar{\boldsymbol{B}})$, where $\bar{\boldsymbol{B}} = [\![\text{sgn}(\max(i, j) - e^{\pi} + 1)]\!]$ is a block matrix with rank at most 2. More generally, matrices of the type $\boldsymbol{B} = [\![g(\max(i, j)/d)]\!]$ belong to $\mathcal{M}_{\text{sgn}}(2r)$, where $g(\cdot)$ is a polynomial of degree r. See Section A.1 of the Appendix.

Our proposed nonparametric matrix regression model $\mathcal{F}_{\text{sgn}}(r)$ therefore implies a new matrix completion model in $\mathcal{M}_{\text{sgn}}(r)$. In next sections, we establish the general theory for $\mathcal{F}_{\text{sgn}}(r)$ first, then specialize the results to the high-rank completion problems in Section 4.2.

3 From classification to regression: a learning reduction approach

In this section, we present a learning reduction approach to estimate f from the model as specified in (2) and (3). Our main crux is to provably convert the regression estimation problem into a series of sign function estimation problems, which are in turn solved by weighted classifications.

More specifically, we dichotomize the response Y_i into a series of binary observations, $\text{sgn}(Y_i - \pi)$, for $\pi \in \mathcal{H} = \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$, where $H \in \mathbb{N}_+$ is a resolution parameter that controls the total number of sign functions to estimate. Then, for each π, we estimate the sign function $\text{sgn}(f - \pi)$ by performing a classification task,

$$\hat{\phi}_\pi = \underset{\phi \in \Phi(r)}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \text{weighted-classification}(\text{sgn}(Y_i - \pi),\ \text{sgn}\phi(\boldsymbol{X}_i)), \tag{4}$$

where $\Phi(r)$ is the collection of rank-r trace functions, and the weighted classification$(\cdot, \cdot)$ denotes a classification objective function with a response-specific weight to each sample point. The weight in the objective function is crucial in our method, and we will detail the form in next section. Our final regression function estimate takes the form,

$$\hat{f} = \frac{1}{2H + 1} \sum_{\pi \in \mathcal{H}} \text{sgn}\hat{\phi}_\pi. \tag{5}$$

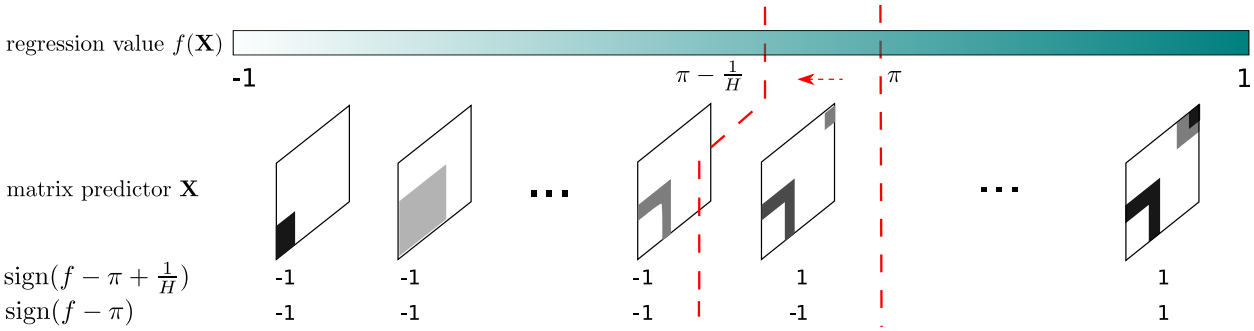

Figure 2: Nonparametric matrix regression via sign function series estimation. We use a series of weighted classifications to estimate the sign functions, then obtain the regression function estimate via sign aggregations. Here, $\boldsymbol{X} \in \mathcal{X}$ denotes matrix-valued predictor, $f\colon \mathcal{X} \to \mathbb{R}$ denotes regression function, and $\text{sgn}(f - \pi) \in \{-1, 1\}$ is the sign function, where $\pi \in \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$ is the series of levels to aggregate in our algorithm.

We comment that the $(2H + 1)$ estimation tasks of the sign functions are fully separable, leading naturally to a parallel type computation. Moreover, the sign functions bridge the problems of level set estimation and Bayes classification, as we will detail in Section 3.2. Fig 2 illustrates our main idea graphically. We refer to our method as the **A**ggegration of **S**tructured **SI**gn **S**eries for **T**race regression, and abbreviate it as **ASSIST**.

Next, we describe the specific form of weighted classification, the uniqueness of the classification optimizer, as well as the accuracy guarantee of the estimator.

3.1 Statistical characterization of sign functions via weighted classification

For a given level $\pi \in [-1, 1]$, define the π-shifted response $\bar{Y}_{\pi,i} = Y_i - \pi$ for $i \in [n]$. We propose a weighted classification objective function in (4) using

$$L(\phi; (\boldsymbol{X}_i, \bar{Y}_{\pi,i})_{i\in[n]}) = \frac{1}{2n}\sum_{i=1}^{n} \underbrace{|\bar{Y}_{\pi,i}|}_{\text{response-specific weight}} \times \underbrace{|\text{sgn}\bar{Y}_{\pi,i} - \text{sgn}\phi(\boldsymbol{X}_i)|}_{\text{classification loss}}, \tag{6}$$

where $\phi \in \Phi(r)$ is the trace function to be optimized, and $|\bar{Y}_{\pi,i}|$ serves as the weight. Such a response-specific weight incorporates the magnitude information of the response into classification, in that the response values that are far away from the target level are penalized more heavily in the objective (6). In the special case of a binary response $Y_i \in \{-1, 1\}$ and target level $\pi = 0$, the objective (6) reduces to the usual classification loss.

Next, define the weighted classification risk,

$$\text{Risk}_\pi(\phi) = \mathbb{E}L(\phi; (\boldsymbol{X}_i, \bar{Y}_{\pi,i})_{i\in[n]}), \tag{7}$$

where the expectation is taken with respect to the joint distribution of $(\boldsymbol{X}_i, Y_i)$ i.i.d. from $\mathbb{P}_{\boldsymbol{X},Y}$.

The next theorem quantifies the global optimum of (7).

Theorem 3.1 (Global optimum of weighted classification risk)**.** *For any given level* $\pi \in [-1, 1]$*, under the model specified in* (2) *and* (3)*, for all functions* $\bar{f}$ *that have the same sign as* $\operatorname{sgn}(f - \pi)$*, it holds that* $\operatorname{Risk}_\pi(\bar{f}) = \inf\{\operatorname{Risk}_\pi(\phi) \colon \phi \in \Phi(r)\}$.

Theorem 3.1 suggests a practical procedure to estimate $\operatorname{sgn}(f - \pi)$ through weighted classifications. The result ensures that the sign function $\operatorname{sgn}(f - \pi)$ minimizes the weighted classification risk. The inverse, however, may not hold true, due to possible multiple global optimizers of $\operatorname{Risk}_\pi(\cdot)$. A simple example is a constant regression $f(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X}) = c$, in which case, every function $\phi \in \Phi(r)$ minimizes $\operatorname{Risk}_\pi(\cdot)$ at the level $\pi = c$. The next section resolves this issue by characterizing the uniqueness of the risk optimizer.

3.2 Identifiability

To establish the statistical guarantee of the minimizer of $\operatorname{Risk}_\pi(\cdot)$, we first address its uniqueness, up to some sign equivalence. It turns out the local behavior of the regression function f around π plays a key role to establish the identifiability of sign function series from weighted classifications.

We introduce some additional notation. We call $S_{\text{bayes}}(\pi) = \{\boldsymbol{X} \in \mathcal{X} \colon f(\boldsymbol{X}) \geq \pi\}$ the Bayes set at level π, and $\partial S_{\text{bayes}}(\pi) = \{\boldsymbol{X} \in \mathcal{X} \colon f(\boldsymbol{X}) = \pi\}$ the level set boundary. Note that there is a one-to-one correspondence between the sign function $\operatorname{sgn}(f - \pi)$ and the Bayes set $S_{\text{bayes}}(\pi)$. We choose to present the results in terms of $S_{\text{bayes}}(\pi)$ for easier comparison with the existing classification literature [Tsybakov, 2004, Singh et al., 2009]. We call a level $\pi \in [0, 1]$ a mass point if the level set boundary $\partial S_{\text{bayes}}(\pi)$ has a non-zero measure under $\mathbb{P}_{\boldsymbol{X}}$. Let $\mathcal{N} = \{\pi \in [-1, 1] \colon \mathbb{P}_{\boldsymbol{X}}[f(\boldsymbol{X}) = \pi] \neq 0\}$ denote the collection of all mass points in f. Assume there exists a constant $c > 0$, independent of the feature space dimension, such that $|\mathcal{N}| \leq c < \infty$. We introduce a notion of smoothness for the cumulative distribution function (CDF) of $f(\boldsymbol{X})$ under measure $\mathbb{P}_{\boldsymbol{X}}$.

Definition 2 (α-smoothness)**.** Suppose $\mathbb{P}_{\boldsymbol{X}}$ is a continuous distribution, and denote the CDF $G(\pi) = \mathbb{P}_{\boldsymbol{X}}[f(\boldsymbol{X}) \leq \pi]$. A function f is called (α, π)-locally smooth, for a given $\pi \notin \mathcal{N}$, if there exist constants $C = C(\pi) > 0$ and $\alpha = \alpha(\pi) \geq 0$, such that

$$\sup_{0 \leq t < \rho(\pi, \mathcal{N})} \frac{G(\pi + t) - G(\pi - t)}{t^\alpha} \leq C, \tag{8}$$

where $\rho(\pi, \mathcal{N}) = \min_{\pi' \in \mathcal{N}} |\pi - \pi'|$ denotes the distance from π to the nearest point in $\mathcal{N}$. We make the convention that $\rho(\pi, \mathcal{N}) = 2$ (which equals the range of $\pi \in [-1, 1]$) when $\mathcal{N}$ is empty, and $\alpha = \infty$ when the numerator in (8) is zero. The largest possible $\alpha = \alpha(\pi)$ in (8) is called the smoothness index at level π. The function f is called α-globally smooth, if (8) holds with a global constant C for all $\pi \in [-1, 1]$ except for a finite number of levels.

Fig 3 shows three examples of the CDF with various levels of smoothness. A small value of $\alpha < 1$ indicates the infinite density at level π, or equivalently, when $G(\pi)$ jumps at π. A large value of

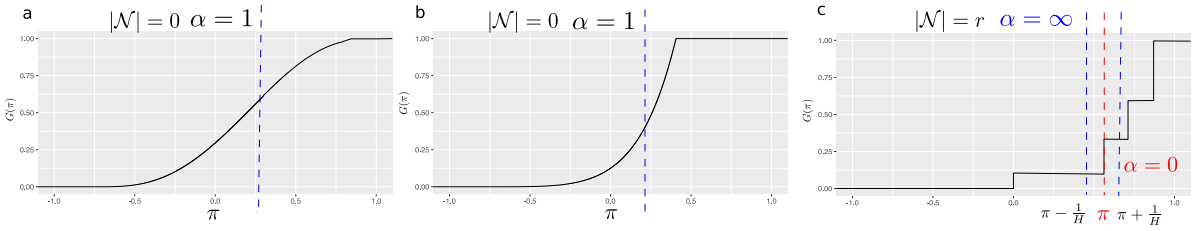

Figure 3: Three examples of CDF, $G(\pi) = \mathbb{P}_{\boldsymbol{X}}(f(\boldsymbol{X}) \leq \pi)$, with local smoothness index α at π depicted in dashed line. (a) and (b). Function $G(\pi)$ $\alpha = 1$ because the $G(\pi)$ has finite sub-derivatives in the range of π; (c). Function $G(\pi)$ with $\alpha = \infty$ at most π (in blue), except for a total number of $|\mathcal{N}| = r$ jump points (in red). Here $|\mathcal{N}|$ denotes the number of jump points.

$\alpha > 1$ corresponds to the case of no point mass around π, or equivalently, when $G(\pi)$ remains flat. An intermediate case is $\alpha = 1$ when $G(\pi)$ has a finite non-zero sub-derivative in the vicinity of π. The global smoothness index is the minimal α over all π's; meanwhile, we allow exceptions for a finite number of levels.

Next, we show that the α-smoothness with $\alpha \neq 0$ implies the uniqueness of $S_{\text{bayes}}(\pi)$ for the optimizer of $\text{Risk}_\pi(\cdot)$. For two sets $S_1, S_2 \in \mathcal{X}$, define the probabilistic set difference,

$$d_\Delta(S_1, S_2) = \mathbb{P}_{\boldsymbol{X}}(S_1 \Delta S_2) = \mathbb{P}_{\boldsymbol{X}}\{\boldsymbol{X} \colon \boldsymbol{X} \in S_1 \setminus S_2 \text{ or } S_2 \setminus S_1\},$$

and the risk difference,

$$d_\pi(S_1, S_2) = \text{Risk}_\pi(\text{sgn}(S_1)) - \text{Risk}_\pi(\text{sgn}(S_2)).$$

Theorem 3.2 (Identifiability)**.** *Suppose f is α-globally smooth over $\mathcal{X}$. Then,*

$$d_\Delta(S, S_{\text{bayes}}(\pi)) \lesssim [d_\pi(S, S_{\text{bayes}}(\pi))]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})} d_\pi(S, S_{\text{bayes}}(\pi)), \tag{9}$$

for all sets $S \in \mathcal{X}$ and all levels $\pi \in [-1, 1]$ except for a finite number of levels.

We make two remarks. First, the bound (9) controls the worst-case perturbation of the classifiers under the measure $\mathbb{P}_{\boldsymbol{X}}$ with respect to the weighted classification risks. When $\alpha \neq 0$, the inequality (9) immediately implies the uniqueness, up to a measure-zero set in $\mathbb{P}_{\boldsymbol{X}}$, of $S_{\text{bayes}}(\pi)$ in minimizing $\text{Risk}_\pi(\cdot)$. Second, our identifiability improves the earlier results for a single level set estimation to multiple level set estimations. Existing work [Singh et al., 2009, Xu et al., 2020] considered only a finite number of π's, and provided only the first term in the bound (9). In contrast, our bound quantifies the full dependence on the level π, and establishes the recovery condition of $S_{\text{bayes}}(\pi)$ uniformly over all possible π's. It turns out both terms in the bound (9) are crucial for our regression function estimation. The first term contributes to the classification error, and the second term contributes to the variance in sign series aggregations.

3.3 Regression risk bound

In this section, we provide the statistical accuracy guarantee for the learning reduction based estimators (4) and (5). Our theory consists of three main ingredients. We first leverage the α-smoothness to provide a sharp rate for $\hat{\phi}_\pi$'s classification risk faster than the usual root-n convergence. The improvement stems from the fact that, under the given assumptions, the variance of the excess classification loss is bounded in terms of its expectation. Because the variance decreases as we approach the optimal $\operatorname{sgn}(f-\pi)$, the risk of $\hat{\phi}_\pi$ converges more quickly to the optimal risk than the simple uniform converge results would suggest. The second step is to convert the risk error into the probability set error by Theorem 3.2. The last step is to aggregate the set error into the final nonparametric function estimation. A careful error analysis reveals the joint contribution from both sign aggregations and variance-bias trade-off.

The next result establishes the estimation accuracy for sign function estimator (4).

Theorem 3.3 (Sign function estimation)**.** *Suppose the regression function $f \in \mathcal{F}_{\operatorname{sgn}}(r)$ is α-globally smooth over $\mathcal{X}$, and let $d_{\max} = \max(d_1, d_2)$. Then, for all $\pi \in [-1,1]$ except for a finite number of levels, with high probability at least $1-\exp(-rd_{\max})$ over the training data $(\boldsymbol{X}_i, Y_i)_{i\in[n]}$, we have,*

$$\|\operatorname{sgn}\hat{\phi}_\pi - \operatorname{sgn}(f-\pi)\|_1 \lesssim \left(\frac{rd_{\max}}{n}\right)^{\frac{\alpha}{2+\alpha}} + \frac{1}{\rho^2(\pi,\mathcal{N})}\left(\frac{rd_{\max}}{n}\right), \tag{10}$$

where the L_1 norm is taken with respect to the measure $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$.

Theorem 3.3 quantifies the statistical convergence of the sign function estimation. For a fixed π, the second term in (10) is absorbed into the first term, leading to the rate $O(n^{-\alpha/(2+\alpha)})$. We find that the sign estimation reaches a fast rate $1/n$ when $\alpha = \infty$, and reaches a slow rate $1/\sqrt{n}$ when the point mass concentrates with $\alpha = 0$. This is consistent with our intuition, because best rate $\alpha = \infty$ corresponds to a clear separation with no point mass at the Bayes set boundary $\partial S_{\text{bayes}}(\pi)$, whereas the worst rate $\alpha = 0$ corresponds to a heavy mass around $\partial S_{\text{bayes}}(\pi)$. Furthermore, the sign function estimation achieves consistency in the high dimensional region as long as $n \gg d_{\max} \to \infty$ and $\alpha \neq 0$. Combining the sign representability of the regression function and the uniform sign estimation accuracy, we obtain our main theoretical result on the nonparametric trace regression.

Theorem 3.4 (Regression function estimation)**.** *Suppose the same conditions in Theorem 3.3 hold. With high probability at least $1-\exp(-rd_{\max})$ over the training data $(\boldsymbol{X}_i, Y_i)_{i\in[n]}$, we have*

$$\|\hat{f} - f\|_1 \lesssim \underbrace{\left(\frac{rd_{\max}\log H}{n}\right)^{\frac{\alpha}{2+\alpha}}}_{\text{estimation error from sign functions}} + \underbrace{\frac{1}{H}}_{\text{reduction bias}} + \underbrace{\left(\frac{rd_{\max}}{n}\right) H\log H}_{\text{reduction variance}}, \tag{11}$$

for any resolution parameter $H \in \mathbb{N}_+$. In particular, setting $H \asymp \left(\frac{n}{rd_{\max}}\right)^{1/2}$ gives

$$\|\hat{f} - f\|_1 \lesssim \left(\frac{rd_{\max}\log n}{n}\right)^{\min\left(\frac{\alpha}{2+\alpha}, \frac{1}{2}\right)}, \tag{12}$$

where the L_1 norm is taken with respect to the measure $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$

Theorem 3.4 establishes the convergence rate of the proposed learning reduction estimator for the nonparametric trace regression. We make three remarks. First, the bound (11) reveals three sources of errors: the estimation error from sign functions, the bias due to sign series representations, and the variance thereof. Recall that H determines the number of sign functions in sign series representations. It controls the bias-variance tradeoff here. Second, the regression is robust to a few off-target classifications, as long as the majorities are accurate. This can also be seen in Fig 3(a) where the classification is nonidentifiable at some mass point (red line). Nevertheless, the regression estimation is still possible because the nearby classifications provide the sign signal (blue lines). This fact shows the benefit of sign aggregations, and also explains the trade-off in choosing H. Intuitively, a larger value of H increases the approximation accuracy, but meanwhile renders the classification harder near the mass points. Third, the final regression error is generally no better than the sign error, as we compare the bounds in (12) with (10). This confirms our premise that classification is easier than regression. On the other hand, our sign representation approach allows us to disentangle the complexity and achieve the theoretical guarantee from classification to regression.

4 Two applications of nonparametric matrix learning

In this section, we apply the general theory in Theorem 3.4 to two specific nonparametric matrix learning problems, the low-rank sparse matrix predictor regression, and the high-rank matrix completion.

4.1 Low-rank sparse matrix predictor regression

The first problem we consider is matrix predictor regression. In addition to the low sign rank structure, we also introduce a two-way sparsity structure. That is, we impose that some rows and columns of $\boldsymbol{B}$ are zeros, where $\boldsymbol{B}$ is as defined in (3). We comment that sparsity is a commonly used structure in matrix data modeling [Zhou and Li, 2014], and it is scientifically plausible in numerous applications [Zhang et al., 2015].

Specifically, we extend the notation $\Phi(r)$ and $\mathcal{F}_{\text{sgn}}(r)$ introduced in Definition 1 to incorporate the sparsity. Let $\Phi(r, s_1, s_2)$ denote the collection of trace functions,

$$\Phi(r, s_1, s_2) = \{\phi\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle + b \,\big|\, \text{rank}(\boldsymbol{B}) \le r, \text{supp}(\boldsymbol{B}) \le (s_1, s_2), (\boldsymbol{B}, b) \in \mathbb{R}^{d_1 \times d_2} \times \mathbb{R}\},$$

where $\text{supp}(\boldsymbol{B})$ denotes the support of $\boldsymbol{B}$, with the sparsity parameters, $s_1 = \|\boldsymbol{B}\|_{1,0} = |\{i \in [d_1]\colon \boldsymbol{B}_i \neq \boldsymbol{0}\}|$, and $s_2 = \|\boldsymbol{B}^T\|_{1,0} = |\{j \in [d_2]\colon \boldsymbol{B}_j^T \neq \boldsymbol{0}\}|$, denoting the number of non-zero rows

and non-zero columns of $\boldsymbol{B}$, respectively. Similarly, let $\mathcal{F}_{\mathrm{sgn}}(r, s_1, s_2)$ denote a family of rank-r, support-(s_1, s_2) sign representable functions based on (3). We have the following result.

Theorem 4.1 (Nonparametric low-rank two-way sparse regression)**.** *Consider the same setup as in Theorem 3.4, except that we replace $\mathcal{F}_{\mathrm{sgn}}(r)$ and $\Phi(r)$ with $\mathcal{F}_{\mathrm{sgn}}(r, s_1, s_2)$ and $\Phi(r, s_1, s_2)$, respectively. Set $H \asymp \left(\frac{n}{r(s_1+s_2)\log d_{\max}}\right)^{1/2}$ in (5). With high probability at least $1 - d_{\max}^{-r(s_1+s_2)}$ over the training data $(\boldsymbol{X}_i, Y_i)_{i\in[n]}$, the estimate (5) is bounded by*

$$\|\hat{f} - f\|_1 \lesssim \left(\frac{r(s_1+s_2)\log d_{\max} \log n}{n}\right)^{\min\left(\frac{\alpha}{2+\alpha}, \frac{1}{2}\right)}. \tag{13}$$

We make two remarks. First, the bound (13) suggests that the estimator remains consistent in the high dimensional regime as $d_{\max}$ and $n \to \infty$, as long as $d_{\max}$ grows sub-exponentially in the sample size n. Such a sample complexity shows the pronounced advantage of the low-rank two-way sparse structural model, by comparing (13) and (12). Second, the two-way sparsity structure facilitates the interpretability, which we further demonstrate through numerical examples in Section 6.2.

4.2 High-rank matrix completion

The second problem we consider is matrix completion. Let $\boldsymbol{Y} \in \mathbb{R}^{d_1 \times d_2}$ be a data matrix generated from the model,

$$\boldsymbol{Y} = \boldsymbol{\Theta} + \boldsymbol{E}, \tag{14}$$

where $\boldsymbol{\Theta} \in \mathcal{M}_{\mathrm{sgn}}(r)$ denotes an unknown signal matrix, and $\boldsymbol{E}$ is an error matrix consisting of zero-mean, independent but not necessarily identically distributed entries. For simplicity, we assume $d_1 = d_2 = d$. Model (14) can be viewed as a special case of model (2), where the predictor space consists of the basis matrices in $\mathbb{R}^{d\times d}$, and the data matrix $\boldsymbol{Y} = [\![Y_{ij}]\!]$ collects the scalar response $Y_{ij} \in \mathbb{R}$. In this case, the problem of regression estimation becomes the estimation of $\boldsymbol{\Theta}$. What is observed is an incomplete data matrix $\boldsymbol{Y}_\Omega$ from (14), where $\Omega \subset [d]^2$ represents the index set of the observed entries. We allow both uniform and non-uniform sampling schemes for Ω. Let $\Pi = \{p_\omega\}$ be an arbitrarily predefined probability distribution over the full index set with $\sum_{\omega\in[d]^2} p_\omega = 1$. Assume the entries ω in Ω are i.i.d. draws with replacement from the full index set following the distribution Π. Denote the sampling rule as $\omega \sim \Pi$, and $\boldsymbol{Y}(\omega)$ the matrix entry indexed by ω.

Now applying our learning reduction approach to the matrix completion problem (14) yields the signal matrix estimate

$$\hat{\boldsymbol{\Theta}} = \frac{1}{2H+1} \sum_{\pi\in\mathcal{H}} \mathrm{sgn}(\hat{\boldsymbol{Z}}_\pi), \tag{15}$$

where, for every $\pi \in \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$, the matrix $\hat{\boldsymbol{Z}}_\pi$ is the solution to the weighted classification

$$\hat{\boldsymbol{Z}}_\pi = \underset{\boldsymbol{Z}:\, \mathrm{rank}(\boldsymbol{Z})\le r}{\arg\min} \sum_{\omega\in\Omega} \underbrace{|\boldsymbol{Y}(\omega) - \pi|}_{\text{weight}} \underbrace{|\mathrm{sgn}(\boldsymbol{Y}(\omega) - \pi) - \mathrm{sgn}(\boldsymbol{Z}(\omega))|}_{\text{classification loss}}.$$

To assess the accuracy of the estimate $\hat{\mathbf{\Theta}} = \hat{\mathbf{\Theta}}_{d\times d}$ in the high dimensional regime $d \to \infty$, we need to put the model in the nonparametric context of Definition 2. We next extend the notion of α-smoothness to a discrete feature space as follows. Let $\Delta s = 1/d^2$ denote a small tolerance, where d^2 represents the number of elements in the feature space. We quantify the distribution of the entries in matrix $\mathbf{\Theta}$ using a pseudo density, i.e., histogram with bin width $2\Delta s$. Specifically, let $G(\pi) = \mathbb{P}_{\omega\sim\Pi}[\mathbf{\Theta}(\omega) \le \pi]$ denote the CDF of $\mathbf{\Theta}(\omega)$ under $\omega \sim \Pi$. We partition $[-1,1] = \mathcal{N} \cup \mathcal{N}^c$, where $\mathcal{N}$ consists of levels whose pseudo density based on $2\Delta s$-bin is asymptotically unbounded; i.e,

$$\mathcal{N} = \left\{ \pi \in [-1,1] \colon \frac{G(\pi+\Delta s) - G(\pi-\Delta s)}{\Delta s} \ge c_1 \right\}, \text{ for some universal constant } c_1 > 0,$$

and $\mathcal{N}^c$ otherwise. Let $|\mathcal{N}|_{\text{cover}}$ be the covering number of $\mathcal{N}$ with $2\Delta s$-bin's; i.e, $|\mathcal{N}|_{\text{cover}} = \text{Leb}(\mathcal{N})/2\Delta s$, where $\text{Leb}(\cdot)$ denotes the Lebesgue measure. The following assumption is a discrete analogy of Definition 2.

Definition 3 (α-smoothness for discrete distribution)**.** Let Π be the sampling distribution over $[d^2]$. We say the signal matrix $\mathbf{\Theta}(\omega)$ is α-globally smooth under $\omega \sim \Pi$, if there exist constants $c_2, c_3 > 0$, such that $|\mathcal{N}|_{\text{cover}} \le c_2$, and for all $\pi \in \mathcal{N}^c$,

$$\sup_{\Delta s \le t < \rho(\pi,\mathcal{N})} \frac{G(\pi+t) - G(\pi-t)}{t^\alpha} \le c_3, \quad \text{with } \rho(\pi,\mathcal{N}) = \min_{\pi'\in\mathcal{N}} |\pi - \pi'| + \Delta s$$

and $\rho(\pi,\mathcal{N})$ denotes the adjusted distance from π to the nearest point in $\mathcal{N}$.

We assess the estimation error of (15) using the mean absolute error (MAE), $\text{MAE}(\hat{\mathbf{\Theta}}, \mathbf{\Theta}) = \mathbb{E}|\hat{\mathbf{\Theta}}(\omega) - \mathbf{\Theta}(\omega)|$, where the expectation is with respect to a future observation $\mathbf{\Theta}(\omega)$ from the distribution G. We have the following result.

Theorem 4.2 (Nonparametric matrix completion)**.** *Consider the matrix model* (14) *with* α*-smooth signal matrix* $\mathbf{\Theta} \in \mathcal{M}_{\text{sgn}}(r)$*. Set* $H \asymp \left(\frac{|\Omega|}{dr}\right)^{1/2}$*. With high probability at least* $1 - \exp(-dr)$ *over* $\boldsymbol{Y}_\Omega$*, the estimate* (15) *satisfies that*

$$\text{MAE}(\hat{\mathbf{\Theta}}, \mathbf{\Theta}) \lesssim \left(\frac{dr \log|\Omega|}{|\Omega|}\right)^{\min(\frac{\alpha}{2+\alpha}, \frac{1}{2})}. \tag{16}$$

We remark that our estimation accuracy (16) applies to both low-rank and high-rank signal matrices. Moreover, the estimation rate depends on the sign complexity $\mathbf{\Theta} \in \mathcal{M}_{\text{sgn}}(r)$, where r can be much smaller than the usual matrix rank as shown in Proposition 2. In fact, our theorem can also be relaxed for a growing $|\mathcal{N}|_{\text{cover}}$ as a function of d, with a slight modification on the setup; see Appendix A.3 for such an extension. We next illustrate Theorem 4.2 with two matrix completion examples and compare with the existing literature.

Example 6 (Stochastic block model based matrix completion)**.** The stochastic block model [Chi et al., 2020] assumes a checkerboard structure under marginal row and column permutations. The signal matrix belongs to our sign representable family $\mathbf{\Theta} \in \mathcal{M}_{\text{sgn}}(r)$, where r is the number of

blocks. Besides, the block matrix is ∞-globally smooth, because $\mathcal{N}$ consists of finitely many $2\Delta s$-bin's covering the block means. Our signal estimate achieves the rate $\tilde{\mathcal{O}}(d^{-1/2})$ when $\alpha = \infty$ with no missingness. This rate agrees with the minimax root-mean-square error (RMSE) rate for stochastic block models with a fixed number of blocks [Gao et al., 2016].

Example 7 (Single index model based matrix completion)**.** The single index model based completion [Ganti et al., 2015] admits a signal matrix $\boldsymbol{\Theta} = g(\boldsymbol{B})$, where g is an unknown monotonic function, and $\boldsymbol{B}$ is an unknown low-rank matrix. Note that $\boldsymbol{\Theta}$ itself is often of a high matrix rank as shown in Fig 1(a). Suppose the CDF of $\boldsymbol{\Theta}(\omega)$ has a bounded pseudo density with $\alpha = 1$. Applying Theorem 4.2 yields the estimation error rate $\tilde{\mathcal{O}}(d^{-1/3})$, which is faster compared to the RMSE rate $\tilde{\mathcal{O}}(d^{-1/4})$ obtained earlier [Ganti et al., 2015].

Finally, we obtain the sample complexity of the nonparametric matrix completion, summarized in the next corollary.

Corollary 1 (Sample complexity for nonparametric completion)**.** Suppose the same conditions of Theorem 4.2 hold. When $\alpha \neq 0$, with high probability at least $1 - \exp(-dr)$ over $\mathcal{Y}_\Omega$,

$$\mathrm{MAE}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \to 0, \quad \text{as} \quad \frac{|\Omega|}{dr \log |\Omega|} \to \infty.$$

Corollary 1 improves the earlier work [Yuan and Zhang, 2016, Lee and Wang, 2020] by allowing both low-rank and high-rank signals. Moreover, the sample size requirement depends only on the sign complexity $\tilde{\mathcal{O}}(dr)$, but not the nonparametric complexity α. We also note that $\tilde{\mathcal{O}}(dr)$ roughly matches the degree of freedom of the signals, suggesting the optimality of our sample requirements.

5 Large-margin implementation and ADMM algorithm

In Section 3, we have established the methodology and theory for the nonparametric matrix trace regression under the 0-1 loss, since this is the canonical loss for classification. However, this loss may be difficult to optimize in practice. In this section, we extend it with a continuous large-margin loss, and present the corresponding optimization algorithm. We consider two loss functions: the hinge loss $F(z) = (1 - z)_+$ for support vector machines, and the psi-loss $F(z) = 2\min(1, (1 - z)_+)$ with $z_+ = \max(z, 0)$ [Shen et al., 2003]. These two losses are most commonly used in classification, and both satisfy the linear excess risk bound; see Section 5.4. We focus on the nonparametric low-rank sparse matrix regression problem. With some straightforward modification, the solution applies to matrix completion and other matrix learning problems as well.

5.1 Large-margin learning

Specifically, we generalize the 0-1 loss minimization (6) to the following continuous large-margin loss minimization problem,

$$\hat{\phi}_{\pi,F} = \underset{\phi\in\Phi(r,s_1,s_2)}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} |Y_i - \pi| F(\phi(\boldsymbol{X}_i)\mathrm{sgn}(Y_i - \pi)) + \lambda\|\phi\|_F^2 \right\}, \tag{17}$$

where $F(z)\colon \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is a continuous function of the margin $z = y\phi(\boldsymbol{X})$, $\lambda > 0$ is the penalty parameter, and $\|\phi\|_F$ is the penalty function. We set $\|\phi\|_F = \|\boldsymbol{B}\|_F$, with $\boldsymbol{B}$ being the coefficient matrix associated with $\phi \in \Phi(r, s_1, s_2)$. The use of large-margin loss in (17) allows us to leverage efficient large-margin optimization algorithms, while maintaining desirable statistical properties under mild conditions. The benefit of ridge penalization has been studied [Shen et al., 2003]. We obtain the corresponding regression function estimate as,

$$\hat{f}_F = \frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} \mathrm{sgn}\hat{\phi}_{\pi,F}. \tag{18}$$

5.2 ADMM optimization

We next present an algorithm to solve (17) for a given $\pi \in \mathcal{H}$. We first note that the estimation problem (17) is equivalent to the optimization,

$$\min_{(\boldsymbol{B},b)\colon \mathrm{rank}(\boldsymbol{B})\leq r, \mathrm{supp}(\boldsymbol{B})\leq(s_1,s_2)} \frac{1}{n}\sum_{i=1}^{n} |\bar{Y}_{\pi,i}| F\big([\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle + b]\mathrm{sgn}\bar{Y}_{\pi,i}\big) + \lambda\|\boldsymbol{B}\|_F^2, \tag{19}$$

where we recall $\bar{Y}_{\pi,i} = Y_i - \pi$ is the π-shifted response. The loss function F can be convex, e.g., hinge loss, or non-convex, e.g., psi-loss. Meanwhile, the optimization (19) has a non-convex feasible region because of the low-rank and sparsity constraints.

We propose an alternating direction method of multipliers (ADMM) algorithm to solve (19). We introduce a dual variable and an additional feasibility constraint to perform coordinate descent in the augmented Lagrangian function. The augmented objective of (19) is

$$L(\boldsymbol{B}, b, \boldsymbol{S}, \boldsymbol{\Lambda}, \rho) = \frac{1}{n}\sum_{i=1}^{n} |\bar{Y}_{\pi,i}| F\big([\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle + b]\mathrm{sgn}\bar{Y}_{\pi,i}\big) + \lambda\|\boldsymbol{B}\|_F^2 + \rho\|\boldsymbol{B} - \boldsymbol{S}\|_F^2 + \langle \boldsymbol{\Lambda}, \boldsymbol{B} - \boldsymbol{S}\rangle,$$

where $\boldsymbol{B} \in \mathbb{R}^{d_1\times d_2}$ is the unconstrained primal variable, $\boldsymbol{S} \in \mathbb{R}^{d_1\times d_2}$ is the constrained dual variable satisfying $\mathrm{rank}(\boldsymbol{S}) \leq r$ and $\mathrm{supp}(\boldsymbol{S}) \leq (s_1, s_2)$, $\boldsymbol{\Lambda} \in \mathbb{R}^{d_1\times d_2}$ is the Lagrangian multiplier, and $\rho > 0$ is the step size parameter. Note that in $L(\boldsymbol{B}, b, \boldsymbol{S}, \boldsymbol{\Lambda}, \rho)$, the non-convexity has moved from the first two terms in $\boldsymbol{B}$ to the last two simpler terms in $\boldsymbol{S}$. This separability simplifies the optimization for a wide range of loss functions and constraints.

We next minimize $L(\boldsymbol{B}, b, \boldsymbol{S}, \boldsymbol{\Lambda}, \rho)$ via coordinate descent, by iteratively updating one variable at

a time while holding others fixed. Each update reduces to a simpler problem and can be efficiently solved by standard algorithms.

Specifically, given variables $(\boldsymbol{S}, \boldsymbol{\Lambda}, \rho)$ and $\bar{\boldsymbol{S}} = (2\rho\boldsymbol{S} - \boldsymbol{\Lambda})/[2(\rho + \lambda)]$, the objective with respect to $(\boldsymbol{B}, b)$ is

$$L(\boldsymbol{B}, b|\boldsymbol{S}, \boldsymbol{\Lambda}, \rho) = \frac{1}{n}\sum_{i=1}^{n} |\bar{Y}_{\pi,i}| F\big([\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle + b]\mathrm{sgn}\bar{Y}_{\pi,i}\big) + (\lambda + \rho)\|\boldsymbol{B} - \bar{\boldsymbol{S}}\|_F^2.$$

Optimization with (5.2) is a standard vector based classification problem with a ridge penalty and an offset $\bar{\boldsymbol{S}}$. There are a number of state-of-art algorithms for weighted SVM [Wang et al., 2008] and psi-learning [Shen et al., 2003], which are readily available to solve this problem.

Next, given $(\boldsymbol{B}, b, \boldsymbol{\Lambda}, \rho)$, and $\bar{\boldsymbol{B}} = (2\rho\boldsymbol{B} + \boldsymbol{\Lambda})/(2\rho)$, the objective with respect to $\boldsymbol{S}$ is

$$L(\boldsymbol{S}|\boldsymbol{B}, b, \boldsymbol{\Lambda}, \rho) = \|\boldsymbol{S} - \bar{\boldsymbol{B}}\|_F^2, \quad \text{subject to} \;\; \mathrm{rank}(\boldsymbol{S}) \le r \text{ and } \mathrm{supp}(\boldsymbol{S}) \le (s_1, s_2).$$

This is equivalent to the best sparse low-rank approximation, in the least-square sense, to the matrix $\boldsymbol{B}$. Compared to the original objective (19), the least-square objective is easier to handle. A number of learning algorithms have been designed to solve this problem, e.g., sparse PCA, sparse SVD, and projection pursuit [Ma, 2013]. We adopt the recently developed double projection method, which has a competitive performance in the high dimensional regime [Yang et al., 2016].

Finally, the Lagrangian multiplier $\boldsymbol{\Lambda}$ is updated by $\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + 2\rho(\boldsymbol{B} - \boldsymbol{S})$. Following some common practice in matrix non-convex optimization [Yang et al., 2016], we run the optimization from multiple initializations to locate a final estimate with the lowest objective value. We summarize the above optimization procedure in Algorithm 1.

5.3 Hyperparameter tuning

We briefly describe the hyperparameters in Algorithm 1 and discuss their choices in practice. There are two sets of hyperparameters, one set for model specification, and the other for algorithmic stability. The model hyperparameters are (r, s_1, s_2), which determine the complexity of sign functions. We choose (r, s_1, s_2) via a grid search based on cross-validation regression error. The resolution in grid search depends on the problem size; for instance, in our brain connectivity data example with $d_1 = d_2 = 68$ in Section 7.1, we search for the optimal values of r, s_1, s_2 over $[d]$, with an increment of 5, under the natural constraint $r \le s_1 = s_2$. The algorithm hyperparameters are (H, λ, ρ). For H and λ, their optimal choices are given in Theorems 3.4 and 5.1, respectively. In practice, we default $H = \min(20, \sqrt{n})$, and $\lambda = \min(0.1, n^{-1})$, which perform well in our numerical experiments. For the step size ρ that controls the closeness between the dual and primal variables, we initialize from 1, and increase its value geometrically by 1.1 during the iterations until the relative change in the primal residual $\|\boldsymbol{B} - \bar{\boldsymbol{S}}\|_F$ falls below a threshold [Parikh and Boyd, 2014]. In our numerical

Algorithm 1 Nonparametric low-rank two-way sparse matrix regression via ADMM

Input: data $(\boldsymbol{X}_i, Y_{\pi,i})_{i\in[n]}$, rank r, support (s_1, s_2), ridge parameter λ, resolution parameter H.

for $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$ **do**
  initialize dual variable $\boldsymbol{S}$ randomly, Lagrangian multiplier $\boldsymbol{\Lambda} = \boldsymbol{0}$, step size $\rho = 1$, and $\bar{Y}_{\pi,i}$.
  repeat
    update $(\boldsymbol{B}, b) \leftarrow \arg\min L(\boldsymbol{B}, b|\boldsymbol{S}, \boldsymbol{\Lambda}, \rho)$.
    update $\boldsymbol{S} \leftarrow \arg\min \|\boldsymbol{S} - \frac{1}{2\rho}(2\rho\boldsymbol{B} + \boldsymbol{\Lambda})\|_F^2$ subject to $\mathrm{rank}(\boldsymbol{S}) \le r$ and $\mathrm{supp}(\boldsymbol{S}) \le (s_1, s_2)$.
    update $\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + 2\rho(\boldsymbol{B} - \boldsymbol{S})$.
    update $\rho \leftarrow 1.1\rho$.
  until convergence
  return trace function estimate, $\hat{\phi}_\pi \colon \boldsymbol{X} \mapsto \langle \hat{\boldsymbol{B}}, \boldsymbol{X}\rangle + \hat{b}$.
end for

Output: nonparametric regression function estimate, $\hat{f} = \frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} \mathrm{sgn}\hat{\phi}_\pi$.

analyses, we observe this scheme provides a stable optimization trajectory.

5.4 Large-margin statistical guarantees

We next establish the statistical accuracy for the large-margin estimators under some additional technical assumptions. Let $f_{\mathrm{bayes},\pi} = \mathrm{sgn}(f - \pi)$ denote the ground truth sign function at $\pi \in [-1, 1]$, and let

$$\mathrm{Risk}_\pi(\phi) = \frac{1}{2}\mathbb{E}|Y - \pi|\big|\mathrm{sgn}(Y - \pi) - \mathrm{sgn}\phi(\boldsymbol{X})\big|,$$
$$\mathrm{Risk}_{\pi,F}(\phi) = \mathbb{E}|Y - \pi|F\big(\phi(\boldsymbol{X})\mathrm{sgn}(Y - \pi)\big),$$

denote the 0-1 risk and F-risk, respectively, where F is the surrogate continuous loss, and the expectation is taken with respect to $(\boldsymbol{X}, Y) \sim \mathbb{P}_{\boldsymbol{X},Y}$ following the regression model $f \in \mathcal{F}_{\mathrm{sgn}}(r, s_1, s_2)$. For simplicity, we assume $d_1 = d_2 = d$ and $\|\boldsymbol{X}\|_F \le 1$ with probability 1. We consider the high dimensional regime where both n and d grow, while (r, s_1, s_2) remain fixed. We need the following assumptions.

Assumption 1 (Assumptions on surrogate loss)**.**

(a) (Approximation error) For any given $\pi \in [-1, 1]$, assume there exist a sequence of functions $\phi_\pi^{(n)} \in \Phi(r, s_1, s_2)$, such that $\mathrm{Risk}_{\pi,F}(\phi_\pi^{(n)}) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \le a_n$, for some sequence $a_n \to 0$ as $n \to \infty$. Furthermore, assume $\|\phi_\pi^{(n)}\|_F \le J$ for some constant $J > 0$.

(b) (Common loss) $F(z) = (1 - z)_+$ is hinge loss, or $F(z) = 2\min(1, (1 - z)_+)$ is psi-loss.

Assumption 1(a) quantifies the representation capability of F and $\Phi(r, s1, s2)$. We note that, although the Bayes rule $f_{\mathrm{bayes},\pi}$ also depends on n implicitly through $d = d(n)$, we drop the dependence on n for simpler notation. Assumption 1(b) implies the Fisher consistency bound for

the weighted risk [Scott, 2011],

$$\mathrm{Risk}_\pi(\phi) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi}) \le C[\mathrm{Risk}_{\pi,F}(\phi) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi})], \text{ for all } \pi \in [-1,1] \text{ and all } \phi.$$

where $C = 1$ for the 0-1 or the hinge loss, and $C = 1/2$ for the psi-loss; see Lemma 2 in Appendix. Therefore, it suffices to bound the excess F-risk in order to bound the usual 0-1 risk. Under Assumption 1, we establish the estimation accuracy guarantee for the large-margin estimators (17) and (18).

Theorem 5.1 (Large-margin estimation)**.** *Consider the same setup as in Theorem 4.1, and denote* $t_n = \frac{r(s_1+s_2)\log d}{n}$. *Suppose the surrogate loss* F *satisfies Assumption 1 with* $a_n \lesssim t_n^{(\alpha+1)/(\alpha+2)}$. *Set* $H \asymp t_n^{-1/2}$ *in (18) and* $\lambda \asymp t_n^{(\alpha+1)/(\alpha+2)} + t_n/\rho(\pi,\mathcal{N})$ *in (17). Then, with high probability at least* $1 - \exp(-nt_n)$ *over the training data* $(\boldsymbol{X}_i, Y_i)_{i\in[n]}$, *we have:*

(a) (Sign function estimation). For all $\pi \in [-1,1]$ *except for a finite number of levels,*

$$\|\mathrm{sgn}\hat{\phi}_{\pi,F} - \mathrm{sgn}(f-\pi)\|_1 \lesssim t_n^{\frac{\alpha}{2+\alpha}} + \frac{1}{\rho^2(\pi,\mathcal{N})} t_n.$$

(b) (Regression function estimation).

$$\|\hat{f}_F - f\|_1 \lesssim (t_n \log n)^{\min\left(\frac{1}{2}, \frac{\alpha}{2+\alpha}\right)}.$$

6 Simulations

In this section, we first evaluate the empirical performance of our method **ASSIST** through four experiments, with varying sample size, response type, matrix dimension, and model complexity. We then compare **ASSIST** with some alternative methods.

6.1 Impacts of sample size, matrix dimension, and model complexity

We consider a random matrix predictor $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ with i.i.d. entries sampled from Uniform[0,1], and simulate two types of response, continuous and binary, through

- Continuous regression: $Y = f(\boldsymbol{X}) + \varepsilon$, where $\varepsilon \sim \mathrm{Normal}(0, 0.1^2)$;
- Binary regression: $Y \in \{-1, 1\}$, with $\mathbb{P}(Y = 1|\boldsymbol{X}) = \frac{1}{2}(f(\boldsymbol{X}) + 1)$.

We set the regression function $f(\boldsymbol{X}) = h(z)$, where $h\colon \mathbb{R} \to [-1,1]$ is a non-decreasing function, $z \in \mathbb{R}$ is a nonlinear predictor that $z = (G^{-1} \circ \bar{G})(\langle \boldsymbol{X}, \boldsymbol{B}\rangle)$, $\circ$ denotes function composition, $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ is a fixed rank-r, supp-(s,s) matrix, $\bar{G}\colon \mathbb{R} \to [0,1]$ is the CDF of $\langle \boldsymbol{X}, \boldsymbol{B}\rangle$ induced by $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$ so that $\bar{G}(\langle \boldsymbol{X}, \boldsymbol{B}\rangle) \sim$ Uniform[0,1], and $G\colon \mathbb{R} \to [0,1]$ is the CDF of some reference distribution. This construction yields a highly nonlinear function f. We set the matrix dimension $d = 20, 30, \ldots, 60$, the training sample size $n = 150, 200, \ldots, 400$, and various combinations of (r,s).

In this study, we set $\lambda = 10^{-2}$, $H = 20$, and use the true (r, s) in Algorithm 1, and study parameter tuning in Section 6.2.

The first experiment assesses the impact of the sample size n for the continuous regression. We set $h(z) = [\exp(z) - 1]/[\exp(z) + 1]$, G as the CDF of a standard normal distribution, the matrix dimension $d = 20$, and the model complexity $(r, s) = (2, 2), (2, 3), (5, 5)$. Fig 4(a) summarizes the main model configurations, including the density of $z = z(\boldsymbol{X})$, the function $h = h(z)$, and the resulting density of $f(\boldsymbol{X})$. Fig 4(b) reports the prediction error, $\|\hat{f} - f\|_1$, as the sample n increases. We see that the error decays polynomially with n. We also see that a higher rank r or a higher support s leads to a larger error, as reflected by the upward shift of the curve as (r, s) increases, since it implies a higher model complexity.

The second experiment considers a binary response. Fig 4(c) reports the prediction error $\|\hat{f} - f\|_1$ as the sample size n increases. We see that the error decays polynomially with n. We also note that, in both cases, the matrix predictor has the dimension $20 \times 20 = 400$ whereas n is on the order of hundreds. Nevertheless, our nonparametric method consistently learns the function f well from limited data without specifying a priori the functional form.

The third experiment evaluates the impact of the matrix dimension d. We fix the sample size $n = 200$ and increase d. Fig 4(d) reports the prediction error. We see that the error increases slowly with d, and the growth appears well controlled by the log rate. Note that, in this example, as d increases, the number of effective entries remains unchanged, but the combinatoric complexity increases in the model space. The increasing error is an unavoidable price to pay for not knowing the positions of the s active entries. This example shows the ability of our method to effectively handle a massive number of noisy features.

The fourth experiment investigates the impact of smoothness in regression function. In Section 2, we show that the probabilistic behavior of $f(\boldsymbol{X})$ plays a key role in our learning reduction approach. Here we assess the empirical performance by repeating all the above experiments us-

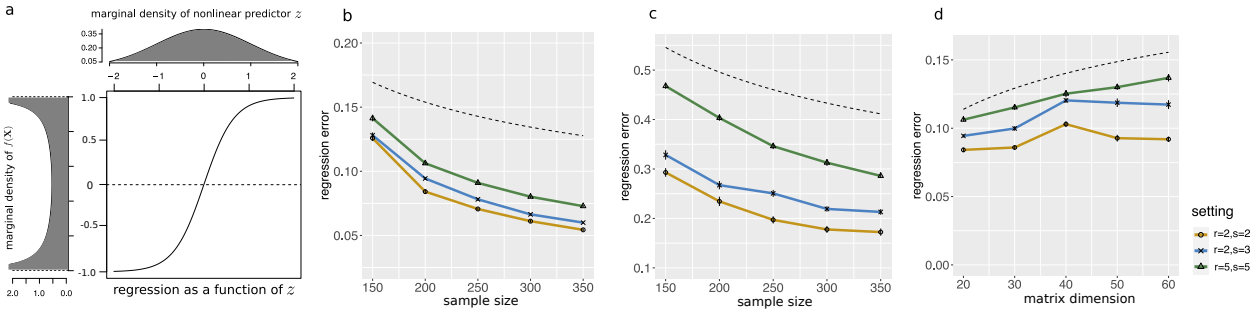

Figure 4: Finite sample performance under a smooth function. (a) simulation setup; (b) prediction error with varying n and $d = 20$ for the continuous response; (c) for the binary response; (d) with varying d and $n = 200$. The dashed lines in panels (b)-(d) represent upper bounds $\mathcal{O}(n^{-1/3})$, $\mathcal{O}(n^{-1/3})$, and $\mathcal{O}(\log d)$, respectively. The results are based on 30 data replications.

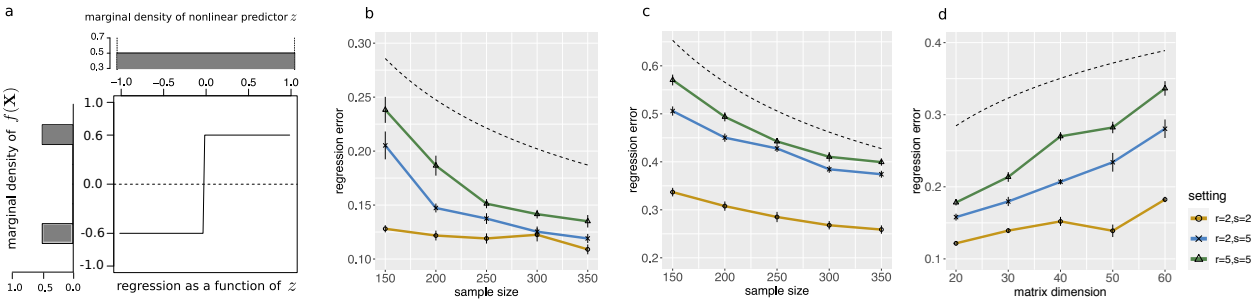

Figure 5: Finite sample performance under a non-smooth function. The setup is similar as Fig 4. The dashed lines in panels (b)-(d) represent upper bounds $\mathcal{O}(n^{-1/2})$, $\mathcal{O}(n^{-1/2})$, and $\mathcal{O}(\log d)$, respectively.

ing a model configuration with $z = z(\boldsymbol{X}) \sim \text{Uniform}[-1, 1]$, $h(z) = -0.6 + 1.2\mathbb{1}(z > 0)$, and $(r, s) = c(2, 2), (2, 5), (5, 5)$. This case falls on the other end of the spectrum in contrast to the infinity smooth function in Fig 4(a). That is, $f(\boldsymbol{X})$ now concentrates at two mass points $\pi = \pm 0.6$. This makes the π-sign function estimation challenging around $\pi = \pm 0.6$ because of the non-identifiability. Fig 5 reports the new model configurations and the corresponding results. Interestingly, we find that our method still maintains a good performance. Such a robustness may be explained by the fact that we aggregate in total $2H + 1$ sign functions, each of which incurs at most $1/(2H + 1)$ error to the regression function estimation. Therefore, our function estimate is robust against some off-target sign estimates, as long as the majority are accurate. This observation is consistent with the consistency result established in Section 3.

6.2 Comparison with alternative methods

Next, we compare our method with several popular alternative solutions. In this comparison, we adopt the simulation setup as in Relión et al. [2019], but add more challenging matrix effects. Particularly, in this setup, the response is binary, and the predictor is a symmetric matrix that encodes a network. In this article, we have been targeting a general matrix predictor, which is directly applicable to a symmetric matrix, though we do not focus on symmetry. Moreover, as we show in Section A.4 of the Appendix, the data generating model falls into our general family of nonparametric trace regression when there is no noise, but no longer so when there is noise. Therefore, we also investigate the performance of our method under model misspecification when including the noise.

More specifically, we simulate from a latent variable model $(\boldsymbol{X}, Y)|\pi$, where we generate π i.i.d. from Uniform[0,1], and conditional on π, we generate $Y \sim \text{Bernoulli}(\pi)$, and

$$\boldsymbol{X} = [\![\boldsymbol{X}_{ij}]\!], \ \boldsymbol{X}_{ij} \overset{\text{indep.}}{\sim} \text{Normal}\left(g_{ij}(\pi)\mathbb{1}(\text{edge } (i, j) \text{ is active}), \sigma^2\right),$$

where the edge connectivity strength, denoted by $g_{ij}(\pi)$, varies depending on the location of $(i, j) \in$

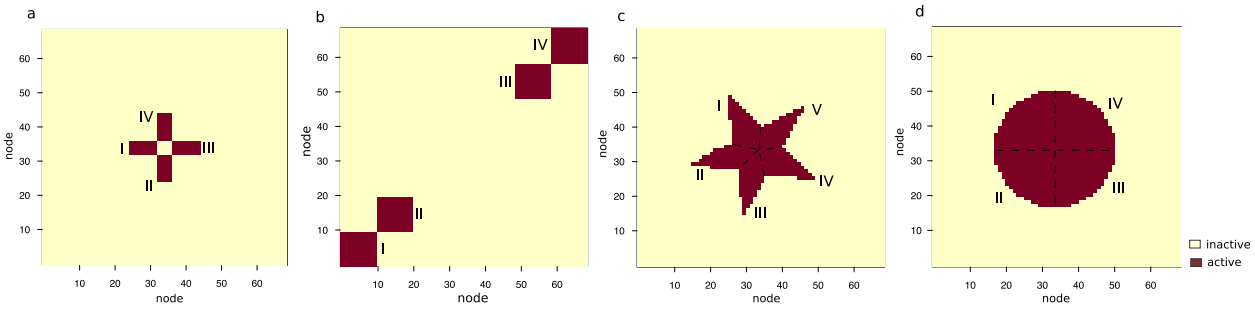

Figure 6: Four activation patterns in simulations. The active region is divided into four or five subregions, denoted by I, II, ..., V, each of which has its own edge connectivity signal $g_{pq}(\pi)$.

$[d]^2$, and the mean response π. Fig 6 shows the activation pattern we consider that specifies the locations of the active edges. The active region is further divided into several subregions, each of which has its own signal function $g_{ij}(\cdot)\colon [0,1] \to \mathbb{R}$. The function form of $g_{ij}(\cdot)$ is randomly drawn from a pre-specified library consisting of common polynomial, log, and trigonometric functions. We set $d = 68$, the training sample size $n = 160$, and the testing size 80. In the noiseless case $\sigma = 0$ in (20), the cross and block patterns are low-rank with $r = 3$ and 5, respectively, whereas the star and circle patterns are nearly full-rank, with a numerical rank $r \approx 30$ on the supported submatrix.

We compare the following four estimation methods.

- Unstructured logistic regression for vector predictors (**LogisticV**, [Zou and Hastie, 2005]). This method vectorizes the matrix predictor into a high dimensional vector, then employs a logistic loss with an elastic net penalty.
- Generalized trace regression for matrix predictors (**LogisticM**, [Relión et al., 2019]). This method fits a parametric trace regression model with a logistic link and a symmetric matrix predictor. It imposes a group lasso penalty to encourage two-way sparsity.
- Convolutional Neural Network (**CNN**) with two hidden layers implemented in Keras [Chollet and Allaire, 2018]. We apply 64 filters with 3×3 convolutional kernels to the matrix-valued predictor, followed by a pooling layer with size 5×5. The resulting features are fed to a fully connected layer of neural network with ReLU activation.
- **A**ggegration of **S**tructured **SI**gn **S**eries for **T**race regression (**ASSIST**), our method.

Among these methods, **LogisticV** serves as a baseline to assess the gain of modeling a matrix predictor over a vector predictor, **LogisticM** is a parametric model, whereas **CNN** and **ASSIST** are nonparametric solutions for matrix predictors. We feed each method with the binary response and the network adjacency matrix as the predictor after randomly permuting the node indices. Because **LogisticM** only supports a symmetric matrix predictor, we provide it with $(\boldsymbol{X} + \boldsymbol{X}^T)/2$ as the input. We use the default parameters of **LogisticM**, and select the tuning parameters of **LogisticV**, **CNN**, and our method **ASSIST**, including the rank r and sparsity parameters (r, s), by

5-fold cross validation.

Fig 7 reports both the prediction error $\|\hat{f} - f\|_1$ and the misclassification error at $\pi = 1/2$ of the four methods evaluated on the testing data. For prediction, we see that **ASSIST** consistently outperforms the alternatives, and the improvement is substantial. For example, the relative reduction using **ASSIST** over the next best approach, **LogisticM**, is over 20% for patterns (a) and (d), and over 15% for patterns (b) and (c). These results clearly demonstrate the benefit of our nonparametric approach. Moreover, we find that neither **LogisticV** nor **CNN** has a satisfactory prediction. A possible explanation is that **LogisticV** takes the vectorized matrix as the input and therefore loses the two-way pairing information. Meanwhile, **CNN** assumes spacial ordering within row and column indices. Although local similarity is important for the usual imaging analysis, the row and column indices take no particular order for a network. Actually, adjacency matrices after row or column permutation represent the same network, and thus the index-invariant methods, such as **LogisticM** and **ASSIST**, perform better. For classification, we also see that our method overall performs the best. The only exception is the circle pattern where **CNN** has a slightly lower classification error. This is perhaps due to the fact that the circle is nearly full rank and thus favors a more complicated model. Interestingly, we also find that the advantage of our method is more substantial in regression prediction than in classification, since classification is easier than regression. Moreover, with model noise included, our method still performs well even though the true model does not exactly follow our model specification.

Finally, to illustrate its capability of producing an estimate of high interpretability, Fig 8 reports the output of **ASSIST** based on the moving average of the feature weights $(\hat{\boldsymbol{B}}_\pi)_{\pi \in \Pi}$. It is observed that the identified activation region agrees well with the truth. We also investigate the relationship between the edge connectivity for individual i and the estimated mean response $\hat{\pi}_i$ for $i = 1, \ldots, n$. The trajectory accurately resembles the ground truth function in each subregion, demonstrating that our method is able to recover the pattern in the matrix predictors $\boldsymbol{X}_i$ against $\hat{\pi}_i$ on a continuous spectrum.

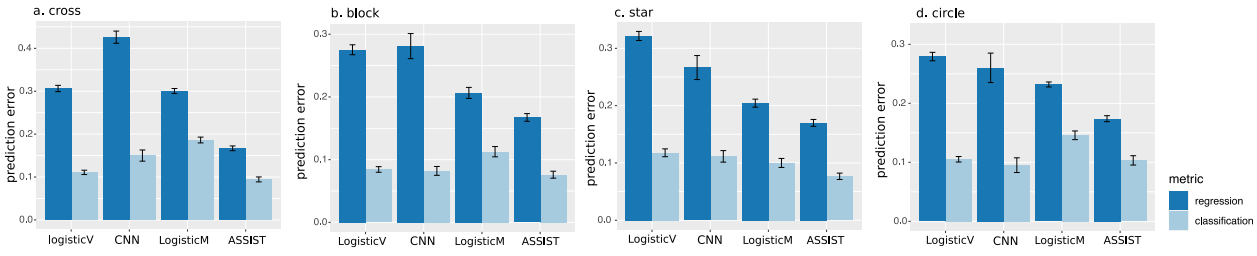

Figure 7: Performance comparison of various methods under four different activation patterns. Reported are the prediction error $\|\hat{f} - f\|_1$, denoted by "regression", and the misclassification error at $\pi = 1/2$, denoted by "classification". The results are based on 30 data replications.

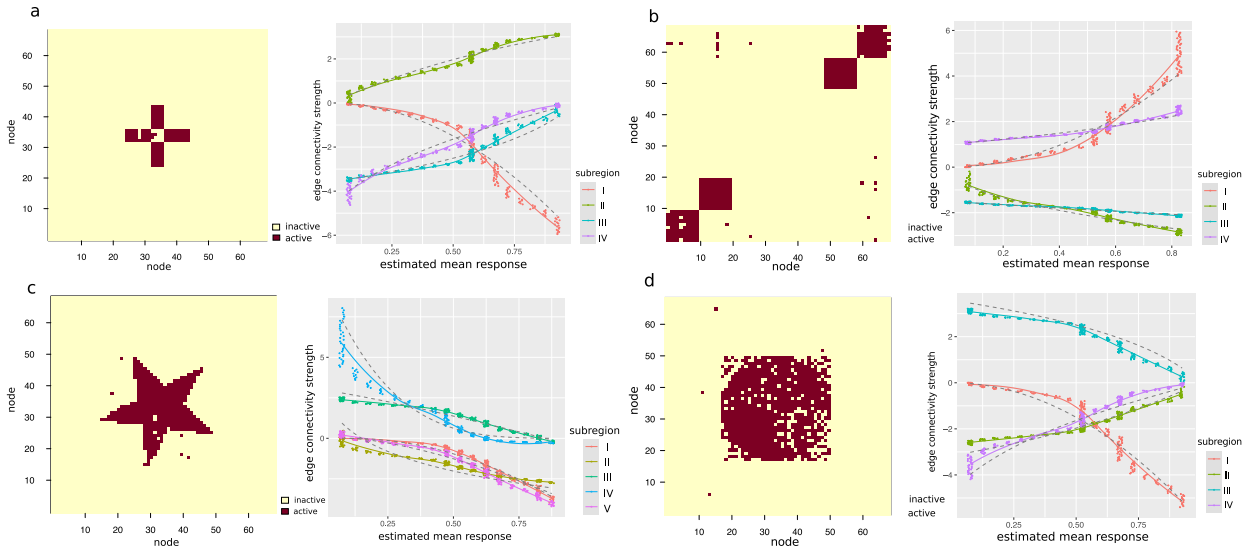

Figure 8: Example output returned by **ASSIST** based on the moving average of the feature weights, and the scatter plot of the edge connectivity strength, averaged by each subregion, versus the estimated mean response. The dashed curve shows the true function.

7 Real data applications

We present two real data applications, in parallel to the two matrix learning tasks studied in Section 4. The first task is the binary-valued trait prediction based on brain connectivity matrix regression, and the second is the continuous-valued matrix completion for imaging analysis.

7.1 Brain connectivity analysis

The first example is a brain connectivity data analysis, which aims to understand the relation between brain connectivity network and cognitive performance. The data is obtained from the Human Connectome Project (HCP) [Van Essen et al., 2013], and consists of $n = 212$ healthy subjects. For each subject, a binary connectivity network is extracted, with nodes corresponding to $d = 68$ brain regions-of-interest following the Desikan atlas [Desikan et al., 2006], and links corresponding to the structural connectivity evaluated by diffusion tensor imaging [Zhang et al., 2018]. The outcome is the dichotomized version of a visuospatial processing test score, corresponding to a high or low performance score [Wang et al., 2019]. We adjust age and gender as additional covariates in our analysis. We note that, although our model focuses on a matrix predictor, it is straightforward to incorporate additional vector-valued covariates. We use a random 60-20-20 split of the data for training, validation, and testing.

We compare our method with the same alternatives as in Section 6.2. Table 1(a) shows that our method achieves the highest accuracy, measured by the area under receiver operating characteristic (AUC). Moreover, as common in the high dimensional setting, we see the model with a good cross-

Table 1: Brain connectivity analysis. (a) Comparison of prediction accuracy measured by AUC, with standard errors over 5-fold cross validation in the parentheses. For **CNN**, there is no report for node selection. (b) Top edges selected by the method **ASSIST-p**. The letters "r" and "l" in node names indicate the right and left hemisphere, respectively. The p-value is calculated from the two-sample test of edge connection strength between two individual groups.

a

Method	AUC	% of Active Nodes
ASSIST-p	**0.73 (0.03)**	88.2
ASSIST	**0.77 (0.04)**	97.3
LogisticM	0.72 (0.02)	100.0
LogisticV	0.68 (0.01)	89.7
CNN	0.67 (0.03)	-

b

Rank	Node	Node	p-value
1	r-inferiortemporal	r-middletemporal	0.01
2	r-parstriangularis	r-supramarginal	3e-5
3	l-posteriorcingulate	r-precentral	0.01
4	l-caudalmiddlefronta	l-isthmuscingulate	2e-5
5	l-lateralorbitofrontal	r-parstriangularis	1e-4

validation accuracy tends to include a large number of noise variables. A useful heuristic called the "one-standard-error rule", suggested by Hastie et al. [2015b], selects the most parsimonious model with cross-validation accuracy within one standard error of the best. We apply this rule and report the results as **ASSIST-p**. It is remarkable to see that **ASSIST-p** results in 12% reduction of active nodes but still achieves a comparable accuracy to the best one. Table 1(b) lists the top brain links identified by our method. The edges are ranked by their maximal values in the feature weights $(\hat{\boldsymbol{B}}_\pi)_{\pi\in\mathcal{H}}$ via moving averaging. We find that the top edges involve connections between frontal and occipital regions in the right hemisphere. This is consistent with recent findings of dysfunction in right posterior regions for deficits in visuospatial processing [Wang et al., 2019]. Fig 9(a) shows the top selected edges overlaid on a brain template. Moreover, we find the relationship between the edge connection strength and the mean response to be nonlinear. Fig 9(b) plots the edge connectivity strength versus the estimated mean response. We see that the connection between r-parstriangularis and r-supramarginal grows slowly when the mean response is small but fast when it is large. In contrary, the connection between r-posteriorcingulate and r-precentral grows fast initially, then

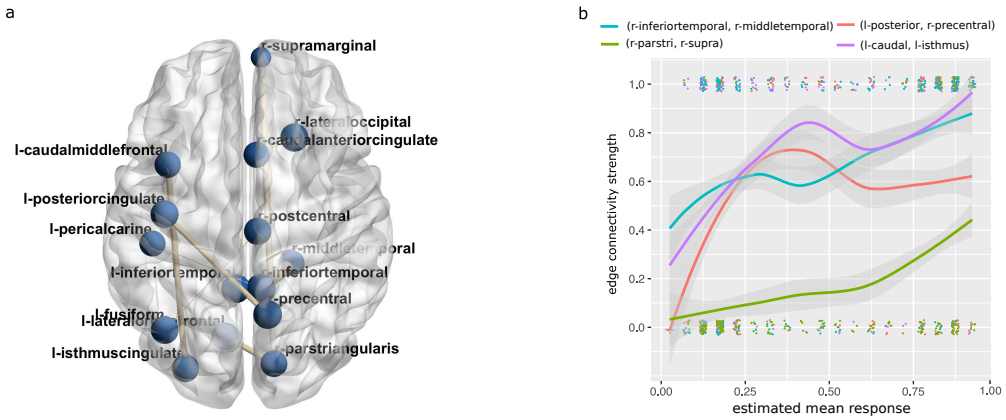

Figure 9: Brain connectivity analysis. (a) Top edges overlaid on a brain template. (b) Edge connectivity strength versus estimated mean response. Colored curves represent the moving averages of connectivity strengths, gray bands represent one standard error, and jitter points represent the raw connectivity values (0 or 1).

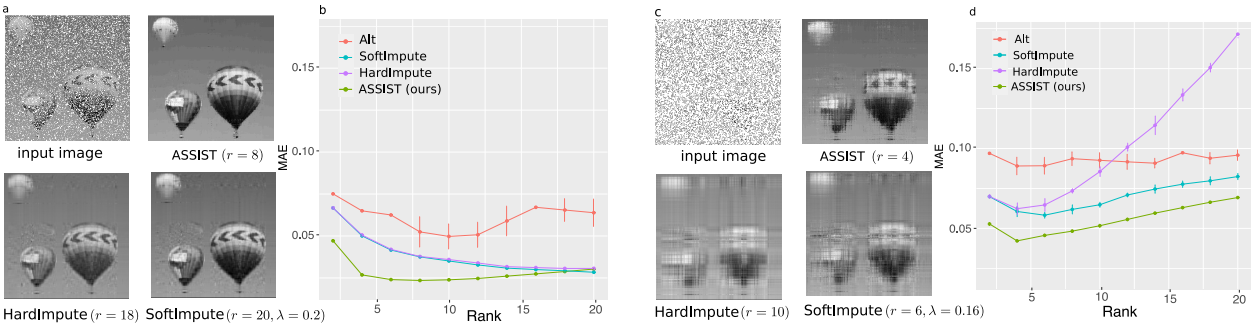

Figure 10: Matrix completion analysis. (a)-(b) correspond to the 40% missing rate, and (c)-(d) the 80% missing rate. Error bars represent the standard error over 5-fold cross-validation. Numbers in the parentheses represent the selected tuning parameters for each method. In (a) and (c), we omit the worst method **ALT** for space consideration.

reaches a plateau as the mean response increases. Such patterns suggest heterogeneous changes in brain connectivity with respect to the visuospatial processing capability.

7.2 Imaging matrix completion

The second application is an imaging matrix completion, where the goal is to recover and restore the partially observed gray-scaled hot air balloon image. This image is a standard benchmark in computer vision, and is organized as a 217-by-217 matrix, whose entries represent pixel values in $[0, 1]$. We randomly mask a subset of entries and perform matrix completion based on the observed entries.

We compare our method with three alternatives: a soft imputation method based on matrix nuclear norm regularization (**SoftImpute**) [Hastie et al., 2015a], a hard imputation method with ridge regression (**HardImpute**) [Mazumder et al., 2010], and a hard imputation based on alternating SVD (**ALT**) [Rennie and Srebro, 2005]. We evaluate the recovery accuracy by MAE on the unobserved entries, and we tune all the parameters based on 5-fold cross-validation.

We investigate missing percentages at 40% and 80%, and vary the rank $r = 2, 4, \ldots, 20$. Fig 10 reports the performances of the four methods. We see clearly that our method achieves the best image recovery, with the smallest MAE. Besides, the advantage of our method compared to the alternative solutions is more clear when the missing percentage increases.

8 Discussion

We have developed a nonparametric trace regression model for studying the relationship between a scalar response and a high dimensional matrix predictor. We propose a learning reduction approach, **ASSIST**, using the structured sign function series, which bridges between regression and

classification. We establish the theoretical bounds, which concern the fundamental statistical errors, are independent of specific algorithms, and serve as a benchmark on how well any algorithmic procedure could perform. Our numerical results demonstrate the competitive performance of the proposed method.

Our work unlocks several possible future directions. One is nonparametric modeling of other nonconventional predictors, such as tensors, functions, and manifold data. Other directions include multi-task learning and compressed sensing. Moreover, our learning reduction approach can be coupled with more sophisticated classifiers, such as neural networks, decision trees, and boosting, for sign function estimation. Finally, the theoretical guarantees we obtain are for the global optimum. How to characterize the behavior of the actual minimizer, or relatedly, the computational error for non-convex matrix based regression remains challenging and open. All these questions are warranted for future research.

Acknowledgements

The research was supported in part by NSF DMS-1915978, NSF DMS-2023239, Wisconsin Alumni Research Foundation (to M. Wang), NIH R01 AG061303 (to L. Li), and NSF CCF-1740858 (to H. Zhang)

References

Fadoua Balabdaoui, Cécile Durot, and Hanna Jankowski. Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B):3276–3310, 2019.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Tianxi Cai, T. Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514):621–633, 2016.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Stanley Chan and Edoardo Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216, 2014.

Eric C Chi, Brian J Gaines, Will Wei Sun, Hua Zhou, and Jian Yang. Provable convex co-clustering of tensors. *Journal of Machine Learning Research*, 21(214):1–58, 2020.

François Chollet and Joseph J Allaire. *Deep Learning mit R und Keras: Das Praxis-Handbuch von den Entwicklern von Keras und RStudio*. MITP-Verlags GmbH & Co. KG, 2018.

Henry Cohn and Christopher Umans. Fast matrix multiplication using coherent configurations. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1074–1087. SIAM, 2013.

Ronald De Wolf. Nondeterministic quantum query and communication complexities. *SIAM Journal on Computing*, 32(3):681–699, 2003.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202, 2019.

Ravi Ganti, Nikhil Rao, Laura Balzano, Rebecca Willett, and Robert Nowak. On learning high dimensional structured single index models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1898–1904, 2017.

Ravi Sastry Ganti, Laura Balzano, and Rebecca Willett. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, volume 28, pages 1873–1881, 2015.

Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *Journal of Machine Learning Research*, 17(1):5602–5630, 2016.

Frederic Gibou, Ronald Fedkiw, and Stanley Osher. A review of level-set methods and some recent applications. *Journal of Computational Physics*, 353:82–109, 2018.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1(2). MIT press Cambridge, 2016.

Nima Hamidi and Mohsen Bayati. On low-rank trace regression under general sampling distribution. *arXiv preprint arXiv:1904.08576*, 2019.

Botao Hao, Boxiang Wang, Pengyuan Wang, Jingfei Zhang, Jian Yang, and Will Wei Sun. Sparse tensor additive regression. *arXiv preprint arXiv:1904.00479*, 2019.

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1): 3367–3402, 2015a.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015b.

Wei Hu, Weining Shen, Hua Zhou, and Dehan Kong. Matrix linear discriminant analysis. *Technometrics*, 62(2):196–205, 2020.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

Chanwoo Lee and Miaoyan Wang. Tensor denoising and completion based on ordinal observations. In *International Conference on Machine Learning*, pages 5778–5788, 2020.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Jesús D Arroyo Relión, Daniel Kessler, Elizaveta Levina, and Stephan F Taylor. Network classification with applications to brain connectomics. *The Annals of Applied Statistics*, 13(3):1648–1677, 2019.

Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719, 2005.

Clayton Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *International Conference on Machine Learning*, 2011.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615, 1994.

Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Aarti Singh, Clayton Scott, and Robert Nowak. Adaptive Hausdorff estimation of density level sets. *The Annals of Statistics*, 37(5B):2760–2782, 2009.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Alexandre B Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, and Wu-Minn HCP Consortium. The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Junhui Wang, Xiaotong Shen, and Yufeng Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, 2008.

Lu Wang, Zhengwu Zhang, and David Dunson. Common and individual structure of brain networks. *The Annals of Applied Statistics*, 13(1):85–112, 2019.

Xiao Wang, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112 (519):1156–1168, 2017.

Zi Wang, Edward Curry, and Giovanni Montana. Network-guided regression for detecting associations between dna methylation and gene expression. *Bioinformatics*, 30(19):2693–2701, 2014.

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning*, 2020.

Dan Yang, Zongming Ma, and Andreas Buja. Rate optimal denoising of simultaneously sparse and low rank matrices. *Journal of Machine Learning Research*, 17(92):1–27, 2016.

Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

Tingting Zhang, Jingwei Wu, Fan Li, Brian Caffo, and Dana Boatman-Reich. A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *Journal of the American Statistical Association*, 110:93–106, 2015.

Zhengwu Zhang, Maxime Descoteaux, Jingwen Zhang, Gabriel Girard, Maxime Chamberland, David Dunson, Anuj Srivastava, and Hongtu Zhu. Mapping population-based structural connectomes. *NeuroImage*, 172:130–145, 2018.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483, 2014.

Ya Zhou, Raymond KW Wong, and Kejun He. Broadcasted nonparametric tensor regression. *arXiv preprint arXiv:2008.12927*, 2020.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix for "Nonparametric Trace Regression in High Dimensions via Sign Series Representation"

A Additional theoretical results

A.1 Sign rank and matrix rank

In the main paper, we have provided several examples with high matrix rank but low sign rank. This section provides more examples and their proofs.

Example 8 (Max graphon)**.** Suppose the matrix $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$ takes the form

$$\mathbf{\Theta}(i,j) = \log\left(1 + \frac{1}{d}\max(i,j)\right), \text{ for all } (i,j) \in [d]^2.$$

Then

$$\operatorname{rank}(\mathbf{\Theta}) = d, \quad \text{and} \quad \operatorname{srank}(\mathbf{\Theta} - \pi) \le 2 \text{ for all } \pi \in \mathbb{R}.$$

Proof. The full-rankness of $\mathbf{\Theta}$ is verified from elementary row operations as follows

$$\begin{pmatrix} (\mathbf{\Theta}_2 - \mathbf{\Theta}_1)/(\log(1+\frac{2}{d}) - \log(1+\frac{1}{d})) \\ (\mathbf{\Theta}_3 - \mathbf{\Theta}_2)/(\log(1+\frac{3}{d}) - \log(1+\frac{2}{d})) \\ \vdots \\ (\mathbf{\Theta}_d - \mathbf{\Theta}_{d-1})/(\log(1+\frac{d}{d}) - \log(1+\frac{d-1}{d})) \\ \mathbf{\Theta}_d/\log(1+\frac{d}{d}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ddots & \ddots & 0 \\ 1 & 1 & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

where $\mathbf{\Theta}_i$ denotes the i-th row of $\mathbf{\Theta}$. Now it suffices to show $\operatorname{srank}(\mathbf{\Theta} - \pi) \le 2$ for π in the feasible range $(\log(1+\frac{1}{d}),\ \log 2)$. In this case, there exists an index $i^* \in \{2,\dots,d\}$, such that $\log(1+\frac{i^*-1}{d}) < \pi \le \log(1+\frac{i^*}{d})$. By definition, the sign matrix $\operatorname{sgn}(\mathbf{\Theta} - \pi)$ takes the form

$$\operatorname{sgn}(\mathbf{\Theta}(i,j) - \pi) = \begin{cases} -1, & \text{both } i \text{ and } j \text{ are smaller than } i^*; \\ 1, & \text{otherwise.} \end{cases}$$

Therefore, the matrix $\operatorname{sgn}(\mathbf{\Theta} - \pi)$ is a rank-2 block matrix, which implies $\operatorname{srank}(\mathbf{\Theta} - \pi) = 2$. □

In fact, Example 8 is a special case of the following proposition.

Proposition 3 (Min/Max graphon)**.** Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuous function such that $g(z) = 0$ has at most $r \ge 1$ distinct real roots. For given numbers $x_i, y_j \in [0,1]$ all $(i,j) \in [d]^2$, define a matrix $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$ with entries

$$\mathbf{\Theta}(i,j) = g(\max(x_i, y_j)), \quad (i,j) \in [d]^2. \tag{21}$$

Then, the sign rank of $\boldsymbol{\Theta}$ satisfies

$$\operatorname{srank}(\boldsymbol{\Theta}) \le 2r.$$

The same conclusion holds if we use min in place of max in (21).

Proof. Without loss of generality, assume $x_1 \le \cdots \le x_d$ and $y_1 \le \cdots \le y_d$. Based on the construction of $\boldsymbol{\Theta}$, the reordering does not change the rank of $\boldsymbol{\Theta}$. Let $z_1 < \cdots < z_r$ be the r distinct real roots for the equation $g(z) = 0$. We separate the proof for two cases, $r = 1$ and $r \ge 2$.

- When $r = 1$. The continuity of $g(\cdot)$ implies that the function $g(z)$ has at most one sign change point. Based on the similar argument as in Example 8, the matrix $\operatorname{sgn}(\boldsymbol{\Theta})$ is a rank-2 block matrix; i.e.,

  $$\operatorname{sgn}(\boldsymbol{\Theta}) = 1 - 2\boldsymbol{a} \otimes \boldsymbol{b} \quad \text{or} \quad \operatorname{sgn}(\boldsymbol{\Theta}) = 2\boldsymbol{a} \otimes \boldsymbol{b} - 1,$$

  where $\boldsymbol{a}, \boldsymbol{b}$ are binary vectors defined by

  $$\boldsymbol{a} = (\underbrace{1, \ldots, 1,}_{\text{positions for which } x_i < z_1} 0, \ldots, 0)^T, \quad \boldsymbol{b} = (\underbrace{1, \ldots, 1,}_{\text{positions for which } y_j < z_1} 0, \ldots, 0)^T.$$

  Therefore, $\operatorname{srank}(\boldsymbol{\Theta}) \le \operatorname{rank}(\operatorname{sgn}(\boldsymbol{\Theta})) = 2$.

- When $r \ge 2$. By continuity, the function $g(z)$ is non-zero and remains an unchanged sign in each of the intervals (z_s, z_{s+1}), for $1 \le s \le r-1$. Define the index set

  $$\mathcal{I} = \{s \in \mathbb{N}_+ \colon \text{the interval } (z_s, z_{s+1}) \text{ in which } g(z) < 0\}.$$

  We now prove that the sign matrix $\operatorname{sgn}(\boldsymbol{\Theta})$ has rank bounded by $2r-1$. To see this, consider the matrix indices for which $\operatorname{sgn}(\boldsymbol{\Theta}) = -1$,

  $$\begin{aligned} \{(i,j) \colon \boldsymbol{\Theta}(i,j) < 0\} &= \{(i,j) \colon g(\max(x_i, y_j)) < 0\} \\ &= \cup_{s \in \mathcal{I}} \{(i,j) \colon \max(x_i, y_j) \in (z_s, z_{s+1})\} \\ &= \cup_{s \in \mathcal{I}} \Big(\{(i,j) \colon x_i < z_{s+1}, y_j < z_{s+1}\} \cap \{(i,j) \colon x_i \le z_s, y_j \le z_{s+1}\}^c \Big) \end{aligned} \tag{22}$$

  The equation (22) is equivalent to

  $$\mathbb{1}(\boldsymbol{\Theta}(i,j) < 0) = \sum_{s \in \mathcal{I}} \left(\mathbb{1}(x_i < z_{s+1}) \mathbb{1}(y_j < z_{s+1}) - \mathbb{1}(x_i \le z_s) \mathbb{1}(y_j \le z_s) \right), \tag{23}$$

  for all $(i,j) \in [d]^2$, where $\mathbb{1}(\cdot) \in \{0,1\}$ denotes the indicator function. The equation (23) implies the low-rank representation of $\operatorname{sgn}(\boldsymbol{\Theta})$,

  $$\operatorname{sgn}(\boldsymbol{\Theta}) = 1 - 2 \sum_{s \in \mathcal{I}} \left(\boldsymbol{a}_{s+1} \otimes \boldsymbol{b}_{s+1} - \bar{\boldsymbol{a}}_s \otimes \bar{\boldsymbol{b}}_s \right), \tag{24}$$

where $\boldsymbol{a}_{s+1}, \bar{\boldsymbol{a}}_s$ are binary vectors defined by

$$\boldsymbol{a}_{s+1} = (\underbrace{1, \ldots, 1,}_{\text{positions for which } x_i < z_{s+1}} 0, \ldots, 0)^T, \quad \text{and} \quad \bar{\boldsymbol{a}}_s = (\underbrace{1, \ldots, 1,}_{\text{positions for which } x_i \le z_s} 0, \ldots, 0)^T,$$

and $\boldsymbol{b}_{s+1}, \bar{\boldsymbol{b}}_s$ are binary vectors defined similarly by using y_j in place of x_i. Therefore, by (24) and the assumption $|\mathcal{I}| \le r-1$, we conclude that

$$\operatorname{srank}(\boldsymbol{\Theta}) \le 1 + 2(r-1) = 2r - 1.$$

Combining two cases yields that $\operatorname{srank}(\boldsymbol{\Theta}) \le 2r$ for any $r \ge 1$. □

Example 9 (Banded matrices)**.** Let $\boldsymbol{a} = (1, 2, \ldots, d)^T$ be a d-dimensional vector, and define a d-by-d banded matrix $\boldsymbol{M} = |\boldsymbol{a} \otimes \boldsymbol{1} - \boldsymbol{1} \otimes \boldsymbol{a}|$. Then

$$\operatorname{rank}(\boldsymbol{M}) = d, \quad \text{and} \quad \operatorname{srank}(\boldsymbol{M} - \pi) \le 3, \quad \text{for all } \pi \in \mathbb{R}.$$

Proof. Note that $\boldsymbol{M}$ is a banded matrix with entries

$$\boldsymbol{M}(i,j) = |i-j|, \quad \text{for all } (i,j) \in [d]^2.$$

Elementary row operation shows that $\boldsymbol{M}$ is full rank as follows,

$$\begin{pmatrix} (\boldsymbol{M}_1 + \boldsymbol{M}_d)/(d-1) \\ \boldsymbol{M}_1 - \boldsymbol{M}_2 \\ \boldsymbol{M}_2 - \boldsymbol{M}_3 \\ \vdots \\ \boldsymbol{M}_{d-1} - \boldsymbol{M}_d \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ -1 & 1 & 1 & \cdots & 1 & 1 \\ -1 & -1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & -1 & 1 \end{pmatrix}.$$

We now show $\operatorname{srank}(\boldsymbol{M} - \pi) \le 3$ by construction. Define two vectors $\boldsymbol{b} = (2^{-1}, 2^{-2}, \ldots, 2^{-d})^T \in \mathbb{R}^d$ and $\operatorname{rev}(\boldsymbol{b}) = (2^{-d}, \ldots, 2^{-1})^T \in \mathbb{R}^d$. We construct the following matrix

$$\boldsymbol{A} = \boldsymbol{b} \otimes \operatorname{rev}(\boldsymbol{b}) + \operatorname{rev}(\boldsymbol{b}) \otimes \boldsymbol{b}. \tag{25}$$

The matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is banded with entries

$$\boldsymbol{A}(i,j) = \boldsymbol{A}(j,i) = \boldsymbol{A}(d-i, d-j) = \boldsymbol{A}(d-j, d-i) = 2^{-d-1}\left(2^{j-i} + 2^{i-j}\right), \text{ for all } (i,j) \in [d]^2.$$

Furthermore, the entry value $\boldsymbol{A}(i,j)$ decreases with respect to $|i-j|$; i.e.,

$$\boldsymbol{A}(i,j) \ge \boldsymbol{A}(i',j'), \quad \text{for all } |i-j| \ge |i'-j'|. \tag{26}$$

Notice that for a given $\pi \in \mathbb{R}$, there exists $\pi' \in \mathbb{R}$ such that $\operatorname{sgn}(\boldsymbol{A} - \pi') = \operatorname{sgn}(\boldsymbol{M} - \pi)$. This is because both $\boldsymbol{A}$ and $\boldsymbol{M}$ are banded matrices satisfying monotonicity (26). By definition (60), $\boldsymbol{A}$ is

a rank-2 matrix. Henceforce, $\operatorname{srank}(\boldsymbol{M} - \pi) = \operatorname{srank}(\boldsymbol{A} - \pi') \leq 3$. □

Example 10 (Identity matrices)**.** Let $\boldsymbol{I}$ be a d-by-d identity matrix. Then

$$\operatorname{rank}(\boldsymbol{I}) = d, \quad \text{and} \quad \operatorname{srank}(\boldsymbol{I} - \pi) \leq 3 \text{ for all } \pi \in \mathbb{R}.$$

Proof. Depending on the value of π, the sign matrix $\operatorname{sgn}(\boldsymbol{I} - \pi)$ falls into one of the two cases:

(a) $\operatorname{sgn}(\boldsymbol{I} - \pi)$ is a matrix of all 1, or of all -1;

(b) $\operatorname{sgn}(\boldsymbol{I} - \pi) = 2\boldsymbol{I} - \mathbf{1}_d \otimes \mathbf{1}_d$.

The first case is trivial, so it suffices to show $\operatorname{srank}(\boldsymbol{I} - \pi) \leq 3$ in the second case. Based on Example 9, the rank-2 matrix $\boldsymbol{A}$ in (60) satisfies

$$\boldsymbol{A}(i,j) \begin{cases} = 2^{-d}, & i = j, \\ \geq 2^{-d} + 2^{-d-2}, & i \neq j. \end{cases}$$

Therefore, $\operatorname{sgn}\left(2^{-d} + 2^{-d-3} - \boldsymbol{A}\right) = 2\boldsymbol{I} - \mathbf{1}_d \otimes \mathbf{1}_d$. We conclude that $\operatorname{srank}(\boldsymbol{I} - \pi) \leq \operatorname{rank}(2^{-d} + 2^{-d-3} - \boldsymbol{A}) = 3$. □

A.2 Extension to sub-Gaussian noise

In the main paper, we have assumed the bounded noise (and thus bounded response) in the regression model. Here we extend the results to unbounded response with sub-Gaussian noise. For notational simplicity, we state the results for the matrix completion problem with $d_1 = d_2 = d$. The results extend similarly to general nonparamatrix matrix regression; we omit the elaboration but only state the difference in the remark.

Consider the signal plus noise model on matrix $\boldsymbol{Y} \in \mathbb{R}^{d \times d}$,

$$\boldsymbol{Y} = \boldsymbol{\Theta} + \boldsymbol{E},$$

where $\boldsymbol{E}$ consists of zero-mean, independent noise entries, and $\boldsymbol{\Theta} \in \mathcal{M}_{\text{sgn}}(r)$ is an α-smooth matrix. Theoretical results in Section 4 of the main paper are based on bounded observation $\|\boldsymbol{Y}\|_\infty \leq 1$. Here, we extend the results to unbounded observation with the following assumption.

Assumption 2 (Sub-Gaussian noise)**.**

1. There exists a constant $\beta > 0$, independent of matrix dimension, such that $\|\boldsymbol{\Theta}\|_\infty \leq \beta$. Without loss of generality, we set $\beta = 1$.
2. The noise entries $\boldsymbol{E}(\omega)$ are independent zero-mean sub-Gaussian random variables with variance proxy $\sigma^2 > 0$; i.e, $\mathbb{P}(|\boldsymbol{E}(\omega)| \geq B) \leq 2e^{-B^2/2\sigma^2}$ for all $B > 0$.

We say that an event A occurs "with high probability" if $\mathbb{P}(A)$ tends to 1 as the dimension $d \to \infty$. The following result show that the sub-Gaussian noise incurs an additional $\log d$ factor compared to the bounded case.

Theorem A.1 (Extension of Theorem 4.2 to sub-Gaussian noise)**.** *Consider the same conditions of Theorem 4.2. Under Assumption 2, with high probability over training data $\boldsymbol{Y}_\Omega$, we have*

(a) (Sign matrix estimation). For all $\pi \in [-1, 1]$ except for a finite number of levels,

$$\mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi, \mathrm{sgn}(\boldsymbol{\Theta} - \pi)) \lesssim \left(\frac{r\sigma^2 d \log d}{|\Omega|}\right)^{\frac{\alpha+1}{\alpha+2}} + \frac{1}{\rho(\pi, \mathcal{N})}\left(\frac{r\sigma^2 d \log d}{|\Omega|}\right). \tag{27}$$

(b) (Signal matrix estimation) Set $H \asymp \left(\frac{|\Omega|}{r\sigma^2 d \log d}\right)^{1/2}$. We have

$$\mathrm{MAE}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \lesssim \mathcal{O}\left\{\left(\frac{r\sigma^2 d \log d \log |\Omega|}{|\Omega|}\right)^{\min(\frac{\alpha}{\alpha+2}, \frac{1}{2})}\right\}.$$

The proof is provided in Section B.6.

Remark 1 (Extending to general non-parametric matrix regression)**.** We have used matrix completion as an example to show the extension to unbounded noise; similar result applies to general matrix regression. For matrix nonparametrix regression (Theorem 3.4 of the main paper), the extension of bounded noise to sub-Gaussian noise incurs an additional $\log n$ factor, where n is the sample size. The techniques of handling sub-Gaussian noise is identical to the above extension, and is thus omitted in the paper.

A.3 Extension to unbounded number of mass points

Theorem 4.2 of our main paper assumes the bounded $|\mathcal{N}|_{\text{cover}} < c < \infty$ for some constant $c > 0$, where $|\mathcal{N}|_{\text{cover}}$ is defined as the covering number of $\mathcal{N}$ with $2\Delta s$-bin's. Recall that $\mathcal{N}$ corresponds to regions of jumps greater than $\Delta s = 1/d^2$ in the CDF $G(\pi) = \mathbb{P}_{\omega\sim\Pi}(\Theta(\omega) \le \pi)$. This setup gives a cleaner exposition of our results but may be restricted in some cases. For example, the high-rank matrices in Example 5 and Figure 1(b) are excluded, because $\alpha = \infty$ and $|\mathcal{N}|_{\text{cover}} = d$ in this setup. Fortunately, our framework still applies to this family of matrices with a little amendment.

We now extend the setup to allow for more general structured matrices including those in Example 5. Redefine $\Delta s = 1/d$. Correspondingly, redefine the smoothness index α and the set $\mathcal{N}$ for the psudo density of $\boldsymbol{\Theta}(\omega)$ with new bin width $2\Delta s$. Let $|\mathcal{N}|_{\text{cover}}$ be the covering number of $\mathcal{N}$ with new $2\Delta s$-bin's. Under this new setup, the signal matrix in Example 5 has $|\mathcal{N}|_{\text{cover}} = 0$ and $\alpha < \infty$. Following the same line as in Theorem 4.2 and use the fact that $\Delta s \lesssim t_d$, we obtain that

$$\mathrm{MAE}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \lesssim (t_d \log H)^{\alpha/(\alpha+2)} + \frac{1}{H} + t_d H \log H, \quad \text{with } t_d = \frac{dr}{|\Omega|}.$$

Therefore, setting $H \asymp t_d^{-1/2}$ yields the error bound

$$\mathrm{MAE}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \leq \mathcal{O}\left\{ \left(\frac{dr \log |\Omega|}{|\Omega|} \right)^{\min(\frac{\alpha}{2+\alpha},\ \frac{1}{2})} \right\}. \tag{28}$$

The result (28) applies to cases when the signal matrices belong to $\mathcal{M}_{\mathrm{sgn}}(r)$ and have at most d distinct entries with repetition patterns.

A.4 Connection to structured matrix model with functional coefficients

In Section 6.2 of the main paper, we simulate data $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ from latent variable model $(\boldsymbol{X}, Y)|\pi$ based on the following scheme,

$$\pi \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1] \overset{\text{conditional on } \pi}{\longrightarrow} \begin{cases} Y \sim \mathrm{Ber}(\pi),\ Y \perp \boldsymbol{X}|\pi, \\ \boldsymbol{X} = [\![\boldsymbol{X}_{pq}]\!],\ \text{where } \boldsymbol{X}_{pq} \overset{\text{indep.}}{\sim} \mathcal{N}(g_{pq}(\pi)\mathbb{1}(\text{edge } (p,q) \text{ is active}), \sigma^2). \end{cases}$$

Notice that, for any given π, $\boldsymbol{X}$ is a rank-r, (s_1, s_2) matrix as shown in Fig 6 of the main paper.

Here we provide justification to this simulation. We will show that the, in the absence of noise $\sigma = 0$, the conditional expectation $\mathbb{E}(Y|\boldsymbol{X}) = f(\boldsymbol{X})$ from the above simulation falls into the low-rank sign-representable function family of our interest.

Specifically, we consider a structured matrix model with functional coefficients

$$\boldsymbol{X}_\pi \overset{\text{def}}{=} \boldsymbol{B}_0 + \sum_{s=1}^{r} g_s(\pi)\boldsymbol{B}_s + \sigma \boldsymbol{E}, \quad Y_\pi \sim \mathrm{Ber}(\pi), \quad \boldsymbol{X}_\pi \perp Y_\pi|\pi, \tag{29}$$

where $\pi \in [0,1]$ is drawn from $\mathrm{Unif}[0,1]$; $\boldsymbol{E}$ is a noise matrix consisting of i.i.d. entries in $N(0,1)$; σ is the noise level; $\boldsymbol{B}_0$ is an arbitrary baseline matrix; $(\boldsymbol{B}_s)_{s=1}^r$ is a set of rank-1 matrices in $\{0,1\}^{d_1 \times d_2}$ that satisfy three conditions:

1. non-overlapping supports, i.e., $\langle \boldsymbol{B}_s, \boldsymbol{B}_{s'} \rangle = 0$ for all $s \neq s'$
2. bounded total support, i.e., $\sum_{s \in [r]} \mathrm{supp}(\boldsymbol{B}_s) \leq (s_1, s_2)$;
3. At least one of the functions $(g_s)_{s=1}^r$ is strictly monotonic with respect to π for all $s \in [r]$.

Proposition 4 (Connection to structured matrix model with functional coefficients)**.** Let $\mathbb{P}_{\boldsymbol{X},Y}$ denote the joint distribution induced by $(\boldsymbol{X}_\pi, Y_\pi)_{\pi \in [0,1]}$ drawn from from (29). In the noiseless case $\sigma = 0$, let $f(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$ denote the regression function based on $\mathbb{P}_{\boldsymbol{X},Y}$. Then $f \in \mathcal{F}_{\mathrm{sgn}}(r, s_1, s_2)$.

Proof. We restrict ourselves to the noiseless case with $\sigma = 0$ in (29). Let

$$\mathcal{X} = \{\boldsymbol{X}_\pi \colon \boldsymbol{X}_\pi \text{ has structure specified in (29) for } \pi \in [0,1]\}$$

denote the predictor space. The mapping between π and $\boldsymbol{X} \in \mathcal{X}$ is one-to-one based on the construction of $\boldsymbol{X}_\pi$. We use $\Pi\colon [0,1] \to \mathcal{X}$ to denote the mapping and Π^{-1} the inverse. Based on the property 3, without loss of generality, assume g_1 is a strictly increasing function.

For any given $\pi \in [0,1]$, we have

$$\mathbb{E}_{Y|\pi}[Y|\pi] = \pi = \Pi^{-1}(\boldsymbol{X}).$$

This implies the regression function $f = \Pi^{-1}$. To show $f \in \mathcal{F}_{\text{sgn}}(r, s_1, s_2)$, it suffices to show $\Pi^{-1} \in \mathcal{F}_{\text{sgn}}(r, s_1, s_2)$. For any given $\pi' \in [0,1]$, write

$$\begin{aligned}\{\boldsymbol{X} \in \mathcal{X}\colon \operatorname{sgn}(\Pi^{-1} - \pi') = 1\} &= \{\boldsymbol{X} \in \mathcal{X}\colon \Pi^{-1}(\boldsymbol{X}) \geq \pi'\} \\ &= \{\boldsymbol{X} \in \mathcal{X}\colon g_1(\Pi^{-1}(\boldsymbol{X})) \geq g_1(\pi')\} \\ &= \left\{\boldsymbol{X} \in \mathcal{X}\colon \langle \boldsymbol{X}, \boldsymbol{B}_1\rangle \geq g_1(\pi')\langle \boldsymbol{B}_1, \boldsymbol{B}_1\rangle + \langle \boldsymbol{B}_0, \boldsymbol{B}_1\rangle\right\},\end{aligned}$$

where the second line uses the fact that g_1 is strictly increasing.

Therefore, the sign function $\operatorname{sgn}(\Pi^{-1} - \pi')$ can be expressed as the sign of trace function,

$$\operatorname{sgn}(\Pi^{-1} - \pi') = \operatorname{sgn}(\underbrace{\langle \boldsymbol{X}, \boldsymbol{B}_1\rangle}_{\text{trace}} - \underbrace{g_1(\pi')\langle \boldsymbol{B}_1, \boldsymbol{B}_1\rangle - \langle \boldsymbol{B}_0, \boldsymbol{B}_1\rangle}_{\text{intercept}}), \quad \text{for all } \boldsymbol{X} \in \mathcal{X},$$

where $\boldsymbol{B}_1$ is a rank-1, supp-(s_1, s_2) matrix coefficient. The proof is complete. □

Remark 2. The above result shows the connection of our method to joint matrix model (29) $(\boldsymbol{X}_\pi, Y_\pi)_{\pi\in[0,1]}$. We should point out, despite of the seeming similarity, a fundamental challenge arises in our setting when the latent index π is unobserved. Our sign aggregation approach essentially learns the right ordering of $\boldsymbol{X}_\pi$ against the index $\pi \in [0,1]$ (see Figure 2 of the main paper), thereby facilitating the estimation of regression function f.

A.5 Adjusting for intercept and additional covariates

In the main paper, we estimate the trace function $\hat{\phi}_{\pi,F}\colon \boldsymbol{X} \mapsto \langle \hat{\boldsymbol{B}}, \boldsymbol{X}\rangle + \hat{b}$ using optimization

$$\begin{aligned}(\hat{\boldsymbol{B}}, \hat{b}) = \operatorname*{arg\,min}_{(\boldsymbol{B}, b)} &\left\{\frac{1}{n}\sum_{i=1}^{n} |\bar{Y}_{\pi,i}| F\big([\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle + b]\operatorname{sgn}\bar{Y}_{\pi,i}\big) + \lambda\|\boldsymbol{B}\|_F^2\right\}, \\ \text{subject to } \operatorname{rank}(\boldsymbol{B}) \leq r,\ &\operatorname{supp}(\boldsymbol{B}) \leq (s_1, s_2).\end{aligned} \tag{30}$$

The optimizer may not be unique; however, the following lemma shows that we can always choose an optimizer $(\hat{\boldsymbol{B}}, \hat{b})$ with bounded intercept without loss of generality.

Lemma 1 (bounded intercept)**.** *Consider 0-1 loss, hinge loss, or phi-loss. Let $(\boldsymbol{X}_i, Y_i)_{i\in[n]}$ be an arbitrary sample with $\|\boldsymbol{X}_i\|_F \leq 1$. Then, there exists a global optimizer $(\boldsymbol{B}_{\text{opt}}, \boldsymbol{b}_{\text{opt}})$ of (30) such that $|\boldsymbol{b}_{\text{opt}}| \leq \|\boldsymbol{B}_{\text{opt}}\|_F + 1$.*

Therefore, in this appendix, we will always assume the trace function family has the additional structure as in Lemma 1, i.e,

$$\Phi(r, s_1, s_2) := \{\phi\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle + b \;\big|\; \operatorname{rank}(\boldsymbol{B}) \le r,\ \operatorname{supp}(\boldsymbol{B}) \le (s_1, s_2),\ |b| \le \|\boldsymbol{B}\|_F + 1\}.$$

For ease of notation, we still use $\Phi(r, s_1, s_2)$ to denote this constrained trace function family.

Proof of Lemma 1. We show that there always exists a global optimizer $(\boldsymbol{B}_{\text{opt}}, b_{\text{opt}})$ of (30) such that

$$\min_{i\in[n]} |\langle \boldsymbol{X}_i, \boldsymbol{B}_{\text{opt}}\rangle + b_{\text{opt}}| \le 1. \tag{31}$$

Let $(\hat{\boldsymbol{B}}, \hat{b})$ be an arbitrary global optimizer of (30). Write $\hat{\phi}(\boldsymbol{X}_i) = \langle \boldsymbol{X}_i, \hat{\boldsymbol{B}}\rangle + \hat{b}$, and $\bar{Y}_i = \bar{Y}_{\pi,i}$ for all $i \in [n]$. If $(\hat{\boldsymbol{B}}, \hat{b})$ satisfies (31), then we keep this $(\hat{\boldsymbol{B}}, \hat{b})$. Otherwise, we aim to construct another global optimizer that satisfies (31). Without loss of generality, assume that $(\hat{\boldsymbol{B}}, \hat{b})$ does not satisfy (31). The construction is divided into two cases based on loss functions.

Case 1: F is 0-1 loss or psi-loss.

Denote

$$i^* = \arg\min_{i\in[n]} |\hat{\phi}(\boldsymbol{X}_i)|, \quad \text{and} \quad m := \min_{i\in[n]} |\hat{\phi}(\boldsymbol{X}_i)| = |\hat{\phi}(\boldsymbol{X}_{i^*})| > 1.$$

We construct a shifted trace function,

$$\phi^*\colon \boldsymbol{X} \mapsto \hat{\phi}(\boldsymbol{X}) - (m-1)\operatorname{sgn}\hat{\phi}(\boldsymbol{X}_{i^*}) = \underbrace{\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle}_{\text{trace}} + \underbrace{\hat{b} - (m-1)\operatorname{sgn}\hat{\phi}(\boldsymbol{X}_{i^*})}_{\text{new intercept} =: \hat{b}^*}.$$

The assumption $m > 1$ implies that, for each $i \in [n]$, $\hat{\phi}(\boldsymbol{X}_i)\operatorname{sgn}\bar{Y}_i$ is either $\ge m > 1$ or $\le -m < -1$. By the definition of ϕ^* and loss function F, we have

$$F\left(\phi^*(\boldsymbol{X}_i)\operatorname{sgn}\bar{Y}_i\right) = \begin{cases} F\big(\hat{\phi}(\boldsymbol{X}_i)\operatorname{sgn}\bar{Y}_i\big), & \text{if } \hat{\phi}(\boldsymbol{X}_i)\operatorname{sgn}\bar{Y}_i \ge m > 1, \\ F\big(\hat{\phi}(\boldsymbol{X}_i)\operatorname{sgn}\bar{Y}_i\big), & \text{if } \hat{\phi}(\boldsymbol{X}_i)\operatorname{sgn}\bar{Y}_i \le m < -1. \end{cases}$$

Therefore, $(\hat{\boldsymbol{B}}, \hat{b}^*)$ is also an optimizer of (30). Notice that $|\phi^*(\boldsymbol{X}_{i^*})| = |\langle \boldsymbol{X}_{i^*}, \hat{\boldsymbol{B}}\rangle + \hat{b}^*| = 1$. Hence, we have found a global optimizer that satisfies (31).

Case 2: F is hinge loss.

We construct $\hat{\phi}^*$ based on misclassified sample points. Denote

$$\mathcal{I}_+ = \{i \in [n]\colon \operatorname{sgn}\hat{\phi}(\boldsymbol{X}_i) = -1 \text{ and } \operatorname{sgn}\bar{Y}_i = 1\}, \quad \mathcal{I}_- = \{i \in [n]\colon \operatorname{sgn}\hat{\phi}(\boldsymbol{X}_i) = 1 \text{ and } \operatorname{sgn}\bar{Y}_i = -1\}.$$

If $\mathcal{I}_+ = \mathcal{I}_- = \emptyset$, then we construct a shifted trace function $\hat{\phi}^*$ as in Case 1. Straightforward calculation shows that the resulting $(\hat{\boldsymbol{B}}, \hat{b}^*)$ satisfies (31). Now, suppose at least one of $\mathcal{I}_+, \mathcal{I}_-$ is

nonempty. Define

$$L_+ = \sum_{i\in\mathcal{I}_+} |\bar{Y}_i|, \qquad \text{and} \qquad L_- = \sum_{i\in\mathcal{I}_-} |\bar{Y}_i|,$$

where we make the conversion that the sum $\sum |\bar{Y}_i|$ is $-\infty$ if the index set is empty. Define

$$i^* = \begin{cases} \arg\min_{i\in\mathcal{I}_+} |\hat{\phi}(\boldsymbol{X}_i)|, & \text{if } L_+ \geq L_-, \\ \arg\min_{i\in\mathcal{I}_-} |\hat{\phi}(\boldsymbol{X}_i)|, & \text{otherwise,} \end{cases} \qquad \text{and} \qquad m := |\hat{\phi}(\boldsymbol{X}_{i^*})| > 1,$$

Notice that the construction of i^* ensures $(L_+ - L_-)\text{sgn}\hat{\phi}(\boldsymbol{X}_{i^*}) = -|L_+ - L_-|$. We construct a shifted trace function

$$\phi^* \colon \boldsymbol{X} \mapsto \hat{\phi}(\boldsymbol{X}) - (m-1)\text{sgn}\phi(\boldsymbol{X}_{i^*}) = \underbrace{\langle \boldsymbol{X}, \hat{\boldsymbol{B}} \rangle}_{\text{trace}} + \underbrace{\hat{b} - (m-1)\text{sgn}\hat{\phi}(\boldsymbol{X}_{i^*})}_{\text{intercept}}. \tag{32}$$

By construction,

$$F\left(\phi^*(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i\right) = \begin{cases} F\left(\hat{\phi}(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i\right) = 0, & \text{if } \hat{\phi}(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i \geq m > 1, \\ F\left(\hat{\phi}(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i\right) + (m-1)\text{sgn}\hat{\phi}(\boldsymbol{X}_{i^*}), & \text{if } i \in \mathcal{I}_+, \\ F\left(\hat{\phi}(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i\right) - (m-1)\text{sgn}\hat{\phi}(\boldsymbol{X}_{i^*}), & \text{if } i \in \mathcal{I}_-. \end{cases}$$

Therefore ϕ^* defined in (32) is a global optimizer of (30), since

$$\sum_{i\in[n]} |Y_i| F(\hat{\phi}^*(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i) = \sum_{i\in[n]} |Y_i| F(\hat{\phi}(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i) - (m-1)|L_+ - L_-| \leq \sum_{i\in[n]} |Y_i| F(\hat{\phi}(\boldsymbol{X}_i)\text{sgn}\bar{Y}_i).$$

Notice that $|\phi^*(\boldsymbol{X}_{i^*})| = 1$. Hence, we have found a global optimizer that satisfies (31).

Finally, the property (31) implies that

$$|b_{\text{opt}}| \leq 1 + \max_{i\in[n]} |\langle \boldsymbol{X}_i, \boldsymbol{B}_{\text{opt}} \rangle| \leq 1 + \|\boldsymbol{B}_{\text{opt}}\|_F.$$

□

Our Algorithm 1 in the main paper can be extended to a mixture of matrix-valued predictors and usual vector-valued predictors. Specifically, we consider classifiers of the type $f(\boldsymbol{X}) = \langle \boldsymbol{X}, \boldsymbol{B} \rangle + \boldsymbol{W}^T\boldsymbol{C}$, where $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ represents the matrix-valued predictor of our interest, $\boldsymbol{W} \in \mathbb{R}^p$ represents the additional covariate including intercept, and $\boldsymbol{C} \in \mathbb{R}^p$ is the unconstrained coefficient parameter. In our neuroimaging analysis (see Section 7.1 of the main paper), we have used $\boldsymbol{W}$ to capture covariates such as age, gender, etc, in the prediction model. Our algorithm is amenable to this case. The only change is the primal update in the algorithm (Line 4 in Algorithm 1 of main paper). The decision variables now consist of $(\boldsymbol{B}, \boldsymbol{C})$ and we solve them simultaneously. Because both $\boldsymbol{B}$ and $\boldsymbol{C}$ are unconstrained decision variables, the algorithm lends itself well to this context.

B Proofs

B.1 Main notation

Notation	Definition
$(\boldsymbol{X}, Y)$	matrix predictor and univariate response
$(\boldsymbol{X}_i, Y_i)_{i=1}^n$	a sample of size n
$\mathcal{X}$	predictor space
$\bar{Y}_\pi = Y - \pi$	shifted response
$f\colon \boldsymbol{X} \mapsto \mathbb{E}(Y\|\boldsymbol{X})$	ground truth regression function
$\hat{f}\colon \boldsymbol{X} \mapsto \mathbb{R}$	estimated regression function
$f_{\text{bayes},\pi} = \text{sgn}(f - \pi)$	Bayes classifier at level π
$S_{\text{bayes}}(\pi) = \{\boldsymbol{X} \in \mathcal{X}\colon f(\boldsymbol{X}) \geq \pi\}$	Indicator set corresponding to $f_{\text{bayes},\pi}$
r	matrix rank
(s_1, s_2)	support parameter
$\mathcal{F}_{\text{sgn}}(r)$	set of r-sign representable functions
$\Phi(r, s_1, s_2)$	rank-r, supp-(s_1, s_2) trace functions
$\Phi(r)$	family of rank-r trace functions
$\boldsymbol{B}$	rank-r, supp-(s_1, s_2) matrix in trace function
α	smoothness index of $G(\pi)$
$\mathcal{N}$	set of mass points associated with CDF $G(\pi) = \mathbb{P}_{\boldsymbol{X}}[f(\boldsymbol{X}) \leq \pi]$
$\rho(\pi, \mathcal{N})$	distance from π to nearest point in $\mathcal{N}$
H	resolution parameter in sign aggregation
ϕ	an arbitrary classifier function from $\mathcal{X}$ to $\mathbb{R}$
$S_\phi = \{\boldsymbol{X} \in \mathcal{X}\colon \phi(\boldsymbol{X}) \geq 0\}$	Indicator set corresponding to ϕ
F	surrogate large-margin loss function from $\mathbb{R}$ to $\mathbb{R}_{\geq 0}$
$\hat{\phi}_{\pi,F}$	estimated classifier function based on regularized empirical F-risk
$\ell_{\pi,F}$	weighted F-loss function, i.e., $\ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)) = \|\bar{Y}_\pi\| F(\phi(\boldsymbol{X})\text{sgn}\bar{Y}_\pi)$
Risk_π	weighted 0-1 risk
$\text{Risk}_{\pi,F}$	weighted surrogate F-risk
$\widehat{\text{Risk}}_{\pi,F}$	empirical weighted F-risk, $\widehat{\text{Risk}}_\pi$ is when F is the 0-1 risk
S, S_1, S_2	subsets in $\mathcal{X}$
$d_\Delta(S_1, S_2)$	probability set difference, equal to $\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{X} \in \mathcal{X}\colon \boldsymbol{X} \in S_1/S_2 \text{ or } S_2/S_1)$
$d_\pi(S_1, S_2)$	risk difference, equal to $\text{Risk}_\pi(\text{sgn}S_1) - \text{Risk}_\pi(\text{sgn}S_2)$
$\boldsymbol{Y}$	data matrix with complete observation
$\Omega \subset [d_1] \times [d_2]$	index set of observations
$\boldsymbol{Y}_\Omega$	data matrix with incomplete observation
$\mathcal{M}_{\text{sgn}}(r)$	family of rank-r sign representable matrices
$\boldsymbol{\Theta} \in \mathcal{M}_{\text{sgn}}(r)$	signal matrix in matrix completion problem
$\boldsymbol{E}$	noise matrix
$\boldsymbol{Z}$	an arbitrary matrix

B.2 Proof of Theorem 3.1

Proof. Fix $\pi \in [-1, 1]$. For any arbitrary function $\phi \in \Phi(r)$, we evaluate the excess risk between $\operatorname{sgn}(f-\pi)$ and $\operatorname{sgn}\phi$,

$$\begin{aligned}
&\operatorname{Risk}_\pi(\operatorname{sgn}\phi) - \operatorname{Risk}_\pi(\operatorname{sgn}(f-\pi)) \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{X}} \underbrace{\mathbb{E}_{Y|\boldsymbol{X}}\left\{|Y-\pi|\left[|\operatorname{sgn}(Y-\pi) - \operatorname{sgn}\phi| - |\operatorname{sgn}(Y-\pi) - \operatorname{sgn}(f-\pi)|\right]\right\}}_{\stackrel{\text{def}}{=} I}.
\end{aligned} \tag{33}$$

Here, $I = I(\boldsymbol{X})$ is a function of $\boldsymbol{X}$, and its expression can be simplified as

$$\begin{aligned}
I &= \mathbb{E}_{Y|\boldsymbol{X}}\left[(Y-\pi)(\operatorname{sgn}(f-\pi) - \operatorname{sgn}\phi)\mathbb{1}(Y \geq \pi) + (\pi - Y)(\operatorname{sgn}\phi - \operatorname{sgn}(f-\pi))\mathbb{1}(Y < \pi)\right] \\
&= \mathbb{E}_{Y|\boldsymbol{X}}\left[(\operatorname{sgn}(f-\pi) - \operatorname{sgn}\phi)(Y-\pi)\right] \\
&= \left[\operatorname{sgn}(f-\pi) - \operatorname{sgn}\phi\right]\left[f-\pi\right] \\
&= |\operatorname{sgn}(f-\pi) - \operatorname{sgn}\phi||f-\pi|,
\end{aligned} \tag{34}$$

where the third line uses the fact $\mathbb{E}_{Y|\boldsymbol{X}} Y = f(\boldsymbol{X})$. Combining (34) with (33), we conclude that, for all $\phi \in \Phi(r)$,

$$\operatorname{Risk}_\pi(\operatorname{sgn}\phi) - \operatorname{Risk}_\pi(\operatorname{sgn}(f-\pi)) = \frac{1}{2}\mathbb{E}_{\boldsymbol{X}}|\operatorname{sgn}(f-\pi) - \operatorname{sgn}\phi||f-\pi| \geq 0,$$

where the last line equals to zero when $\operatorname{sgn}\phi = \operatorname{sgn}(f-\pi)$ or $f \equiv \pi$ is a constant function. Note that $(f-\pi)$ is r-sign representable by assumption. Therefore,

$$\operatorname{Risk}_\pi(\operatorname{sgn}(f-\pi)) = \inf\{\operatorname{Risk}_\pi(\operatorname{sgn}\phi)\colon \phi \in \Phi(r)\}.$$

Based on the definition of 0-1 classification loss, the $\operatorname{Risk}_\pi(\cdot)$ relies only on the sign of the argument function. Therefore, for all functions $\bar{f}\colon \mathcal{X} \to \mathbb{R}$ that have the same sign as $\operatorname{sgn}(f-\pi)$, we have

$$\operatorname{Risk}_\pi(\bar{f}) = \inf\{\operatorname{Risk}_\pi(\operatorname{sgn}\phi)\colon \phi \in \Phi(r)\} = \inf\{\operatorname{Risk}_\pi(\phi)\colon \phi \in \Phi(r)\}.$$

□

B.3 Proof of Theorem 3.2

Proof. Fix $\pi \in [-1, 1]$. For ease of notation, we drop the dependence of π in $S_{\text{bayes}}(\pi)$ and simply write S_{bayes}. Based on (34) in the proof of Theorem 3.1, we have

$$\begin{aligned}
d_\pi(S, S_{\text{bayes}}) &\stackrel{\text{def}}{=} \operatorname{Risk}_\pi(\operatorname{sgn}(S)) - \operatorname{Risk}_\pi(\operatorname{sgn}(S_{\text{bayes}})) \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{X}}\left(|\operatorname{sgn}(S) - \operatorname{sgn}(S_{\text{bayes}})|\,|\pi - f|\right)
\end{aligned}$$

$$= \int_{\boldsymbol{X} \in S \Delta S_{\text{bayes}}} |f(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}}. \tag{35}$$

We divide the proof into two cases: $\alpha > 0$ and $\alpha = \infty$.

Case 1: $\alpha > 0$.

Consider an arbitrary set $S \subset \mathbb{R}^{d_1 \times d_2}$. Let t be an arbitrary number in the interval $[0, 1]$, and define the set $A = \{\boldsymbol{X} \in \mathcal{X} \colon |f(\boldsymbol{X}) - \pi| > t\}$.

$$\begin{aligned}
\int_{\boldsymbol{X} \in S \Delta S_{\text{bayes}}} |f(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}} &\geq t \left[\mathbb{P}_{\boldsymbol{X}}((S \Delta S_{\text{bayes}}) \cap A)\right] \\
&\geq t \left(\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) - \mathbb{P}_{\boldsymbol{X}}(A^c)\right) \\
&\geq t \left(\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) - C t^{\alpha}\right), \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}),
\end{aligned}$$

where the last inequality is from α-globally smoothness condition. Combining the above inequality with the identity (35) yields

$$d_{\pi}(S, S_{\text{bayes}}) \geq t \left(\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) - C t^{\alpha}\right), \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{36}$$

We maximize the lower bound of (36) with respect to t, and obtain the optimal t_{opt},

$$t_{\text{opt}} = \begin{cases} \rho(\pi, \mathcal{N}), & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) > C(1+\alpha)\rho^{\alpha}(\pi, \mathcal{N}), \\ \left[\frac{1}{2C(1+\alpha)} \mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right)\right]^{1/\alpha}, & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) \leq C(1+\alpha)\rho^{\alpha}(\pi, \mathcal{N}). \end{cases}$$

The corresponding lower bound of the inequality (36) becomes

$$d_{\pi}(S, S_{\text{bayes}}) \geq \begin{cases} c_1 \rho(\pi, \mathcal{N}) \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}), & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) > C(1+\alpha)\rho^{\alpha}(\pi, \mathcal{N}), \\ c_2 \left[\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right)\right]^{\frac{1+\alpha}{\alpha}}, & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) \leq C(1+\alpha)\rho^{\alpha}(\pi, \mathcal{N}), \end{cases}$$

where $c_1, c_2 > 0$ are two constants independent of S. Combining both cases gives

$$d_{\Delta}(S, S_{\text{bayes}}) \stackrel{\text{def}}{=} \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}) \lesssim \left[d_{\pi}(S, S_{\text{bayes}})\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})} d_{\pi}(S, S_{\text{bayes}}), \tag{37}$$

where we have absorbed the constants into the relationship $\lesssim$.

Case 2: $\alpha = \infty$.

The inequality (36) now becomes

$$d_{\pi}(S, S_{\text{bayes}}) \geq t \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}) = t d_{\Delta}(S, S_{\text{bayes}}), \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{38}$$

The conclusion (37) follows by taking $t = \frac{\rho(\pi, \mathcal{N})}{2}$ in the inequality (38).

□

Remark 3 (Bounding L_1 distance by classification risk)**.** The bound controls the L_1 distance to $f_{\text{bayes},\pi} = \text{sgn}(f - \pi)$ using the classification excess risk to $\text{Risk}_\pi(f_{\text{bayes},\pi})$. The result applies uniformly to $\pi \in [-1, 1]$ if f is globally-α smooth; i.e., the bound

$$\|\text{sgn}\phi - f_{\text{bayes},\pi}\|_1 \lesssim [\text{Risk}_\pi(\phi) - \text{Risk}_\pi(f_{\text{bayes},\pi})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})} [\text{Risk}_\pi(\phi) - \text{Risk}_\pi(f_{\text{bayes},\pi})]$$

holds for all functions $\phi\colon \mathcal{X} \to \mathbb{R}$ and for all $\pi \in [-1, 1]$ except for a finite number of points. In fact, the similar inequality holds by replacing the 0-1 risk to hinge risk or T-truncated hinge risk. Specifically, the following bound holds for all functions $\phi\colon \mathcal{X} \to \mathbb{R}$ and all $\pi \in [-1, 1]$ except for a finite number of points.

- For hinge loss $F(z) = (1 - z)_+$,

  $$\|\phi - f_{\text{bayes},\pi}\|_1 \lesssim [\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})} [\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})] .$$

- For T-truncated hinge loss $F(z) = \min((1 - z)_+, T)$ with $T \geq 2$,

  $$\|\phi^T - f_{\text{bayes},\pi}\|_1 \lesssim [\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})} [\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})] ,$$

  where ϕ^T is a truncation of function ϕ; see formal definition in (44).

See Lemma 5 for proofs.

B.4 Proofs of Theorem 3.3 and Part (a) in Theorems 5.1

We provide a unified framework that incorporates Theorem 3.3, Part (a) in Theorems 5.1 in the main paper. In addition, part of the proof in Theorem 4.1 is given with the same framework. For any given $\pi \in [-1, 1]$, write $\bar{Y}_\pi = Y - \pi$, and let $\ell_{\pi,F}(\phi; (\boldsymbol{X}, Y))$ denote the weighted F-loss

$$\ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)) \stackrel{\text{def}}{=} |\bar{Y}_\pi| F\big(\phi(\boldsymbol{X})\text{sgn}(Y - \pi)\big),$$

where the loss function F could be either standard 0-1 loss $F(z) = \mathbb{1}(z > 0)$ or surrogate loss satisfying Assumption 1. Assume $\mathbb{P}(\|\boldsymbol{X}\|_F \leq 1) = 1$. Consider the large-margin estimate

$$\hat{\phi}_{\pi,F} = \underset{\phi \in \Phi(r,s_1,s_2)}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{\pi,F}(\phi; (\boldsymbol{X}_i, Y_i)) + \lambda \|\phi\|_F^2 \right\},$$

where the trace function family

$$\Phi(r, s_1, s_2) = \{\phi\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B} \rangle + b \,\big|\text{rank}(\boldsymbol{B}) \leq r,\ \text{supp}(\boldsymbol{B}) \leq (s_1, s_2), |b| \leq \|\boldsymbol{B}\|_F + 1\}$$

is the search domain. Notice that we have imposed the additional constraint $|b| \leq \|\boldsymbol{B}\|_F + 1$ without altering the estimation; see Section A.5.

The following theorem states the accuracy for sign function estimate $\operatorname{sgn}\hat{\phi}_{\pi,F}\colon \mathcal{X} \to \{-1, 1\}$.

Theorem B.1 (Sign estimation)**.** *Fix $\pi \notin \mathcal{N}$. Suppose the regression function $f \in \mathcal{F}_{\text{sgn}}(r, s_1, s_2)$ is (π, α)-smooth over $\mathcal{X}$. Then, with high probability at least $1 - \exp(-nt_n)$ over training data $(\boldsymbol{X}_i, Y_i)_{i\in[n]}$, the estimate (30) satisfies*

$$\|\operatorname{sgn}\hat{\phi}_{\pi,F} - f_{\text{bayes},\pi}\|_1 \lesssim t_n^{\alpha/(2+\alpha)} + \frac{1}{\rho^2(\pi, \mathcal{N})} t_n, \tag{39}$$

under the following three specifications:

(a) (Theorem 3.3) 0-1 loss $F(z) = \mathbb{1}(z > 0)$, no penalization $\lambda = 0$, $(s_1, s_2) = (d_1, d_2)$, and $t_n = \frac{1}{n} r d_{\max}$;

(b) (Theorem 4.1) 0-1 loss $F(z) = \mathbb{1}(z > 0)$, no penalization $\lambda = 0$, constant (s_1, s_2), and $t_n = \frac{1}{n} r(s_1 + s_2) \log d_{\max}$;

(c) (Theorem 5.1) Surrogate loss satisfying Assumption 1, constant (s_1, s_2), $t_n = \frac{1}{n} r(s_1+s_2) \log d_{\max}$, penalization $\lambda \asymp t_n^{(\alpha+1)/(\alpha+2)} + t_n/\rho(\pi, \mathcal{N})$, approximation error $a_n^{(\alpha+1)/(\alpha+2)} \le t_n$.

Here, the constants suppressed in the $\lesssim$ of (39) are independent of π.

Remark 4 (One-sided tail)**.** Inspection of the proof shows that the conclusion (39) holds for all $t \ge t_n$. That is, for all $t \ge t_n$, with high probability at least $1 - \exp(-nt)$, we have

$$\|\operatorname{sgn}\hat{\phi}_{\pi,F} - f_{\text{bayes},\pi}\|_1 \lesssim t^{\alpha/(2+\alpha)} + \frac{1}{\rho^2(\pi, \mathcal{N})} t.$$

Remark 5 (Ridge penalization)**.** The estimation under 0-1 loss requires no penalization, because only the sign, but not the magnitude, of ϕ affects the 0-1 risk. One can constrain $\|\phi\|_F = 1$ in the empirical 0-1 risk minimization without altering the solution. In contrast, the surrogate loss such as as hinge loss is scale-sensitive, rending the possible unboundedness of ϕ. We impose penalization to control the magnitude of the $\|\phi\|_F$ and thus the local complexity. The resulting estimation enjoys the fast convergence as in sieve estimate [Shen and Wong, 1994] under well tuned λ.

We provide the proof after introducing two main lemmas. There are two key ingredients in the proof. The first step is to quantify the convergence of $\hat{\phi}_{\pi,F}$'s excess F-risk using Lemmas 2 and 3. The second step is to relate the excess F-risk to excess 0-1 risk using Lemma 2, and then establish the sign function accuracy using Theorem 3.2.

Recall that $\hat{\phi}_{\pi,F}$ is the minimizer of empirical F-risk. To quantify the $\hat{\phi}_{\pi,F}$'s excess F-risk, we notice that

$$\begin{aligned}
&\operatorname{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \inf_{\text{all } \phi} \operatorname{Risk}_{\pi,F}(\phi) \\
=&\underbrace{\operatorname{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \inf_{\phi\in\Phi(r,s_1,s_2)} \operatorname{Risk}_{\pi,F}(\phi_\pi^*)}_{\text{estimation error}} + \underbrace{\inf_{\phi\in\Phi(r,s_1,s_2)} \operatorname{Risk}_{\pi,F}(\phi) - \inf_{\text{all } \phi} \operatorname{Risk}_{\pi,F}(\phi)}_{\text{approximation error}},
\end{aligned}$$

The simplest way to bound $\hat{\phi}_{\pi,F}$'s excess risk is to use a uniform convergence of excess risk over classifiers $\Phi(r, s_1, s_2)$; however, this approach ignores the local complexity around $\hat{\phi}_{\pi,F}$ and yields a suboptimal rate. Here we adopt the local iterative techniques of Wang et al. [2008, Theorem 3] to obtain a better rate. The improvement stems from the fact that, under considered assumptions, the variance of the excess loss is bounded in terms of its expectation. Because the variance decreases as we approach the optimal ϕ^*_π, the risk of the empirical minimizer converges more quickly to the optimal risk than the simple uniform converge results would suggest.

The following result summarizes the key properties of four common losses: 0-1 loss, hinge loss, T-truncated hinge loss, and psi-loss. Here, the T-truncated hinge loss is defined as $F(z) = \min((1 - z)_+, T)$ for a given $T \geq 2$. We will use T-truncated hinge loss to facilitate the proofs of Lemma 3 and Theorem B.1.

Lemma 2 (Conversion inequalities)**.** *Suppose the regression function f is (π, α)-smooth, and denote $f_{\text{bayes},\pi} = \text{sgn}(f - \pi)$ for $\pi \in [-1, 1]$. Let F be 0-1 loss, hinge loss, T-truncated hinge loss, or psi-loss. Then, the following three properties hold for all $\pi \in [-1, 1]$.*

(a) Optimality: $\inf_{\text{all } \phi} \text{Risk}_{\pi,F}(\phi) = \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})$.

(b) Excess risk bound: for all classifers $\phi \colon \mathcal{X} \to \mathbb{R}$,

$$\text{Risk}_\pi(\phi) - \text{Risk}_\pi(f_{\text{bayes},\pi}) \leq C \left[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right], \tag{40}$$

where $C = 1$ for 0-1, hinge loss or T-truncated loss, and $C = 1/2$ for psi-loss.

(c) Variance-to-mean relationship: Suppose F is 0-1 loss, T-truncated loss, or psi-loss. Then, for all classifiers $\phi \colon \mathcal{X} \to \mathbb{R}$,

$$\begin{aligned} &\text{Var}\left[\ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(f_{\text{bayes},\pi}; (\boldsymbol{X}, Y))\right] \\ \lesssim\ &\left[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right]^{\alpha/(1+\alpha)} + \frac{1}{\rho(\pi, \mathcal{N})} \left[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right]. \end{aligned} \tag{41}$$

Remark 6. The property (c) holds for bounded loss functions only, i.e, excluding hinge loss.

Below we establish the estimation convergence rate for $\hat{\phi}_{\pi,F}$'s excess F-risk. The variance-to-mean relationship in Lemma 2 plays a key role in determining the convergence rate based on Shen and Wong [1994, Theorem 3]; also see Theorem C.2 in Section C. Our proof of Lemma 3 adopts the local iterative techniques from Wang et al. [2008, Theorem 3]. Similar techniques have been used in Bartlett et al. [2006, Theorem 4] for similar estimate but without ridge penalization.

Lemma 3 (Classification risk error)**.** *Consider the set-up as in Theorem B.1. Then, with high probability and t_n specified in Theorem B.1, the following holds for all $\pi \notin \mathcal{N}$.*

(a) If F is 0-1 loss or psi-loss, then

$$\text{Risk}_\pi(\hat{\phi}_{\pi,F}) - \text{Risk}_\pi(f_{\text{bayes},\pi}) \lesssim \text{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) \lesssim t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})} t_n.$$

(b) If F is hinge loss, then

$$\mathrm{Risk}_\pi(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi}) \lesssim \mathrm{Risk}_{F'}(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_{F'}(f_{\mathrm{bayes},\pi}) \lesssim t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi,\mathcal{N})} t_n,$$

where $\mathrm{Risk}_{F'}(\phi) := \mathbb{E}\left[|\bar{Y}_\pi| F'(\phi(\boldsymbol{X})\mathrm{sgn}\bar{Y}_\pi)\right]$ *denotes the risk evaluated under T-truncated hinge loss* $F' = \min(T,\ (1-z)_+)$*, and* $T = \max(2, J) \geq \max(2, \|\phi_\pi^{(n)}\|_F)$ *is a constant based on Assumption 1(a).*

Proof of Theorem B.1. Write $\rho = \rho(\pi, \mathcal{N})$. Combining Theorem 3.2 and Lemma 3 gives

$$\begin{aligned}
\|\mathrm{sgn}\hat{\phi}_{\pi,F} - f_{\mathrm{bayes},\pi}\|_1 &\lesssim \left[\mathrm{Risk}_\pi(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi})\right]^{\alpha/(\alpha+1)} + \frac{1}{\rho}\left[\mathrm{Risk}_\pi(\hat{\phi}_{F,\pi}) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi})\right] \\
&\lesssim t_n^{\alpha/(\alpha+2)} + \frac{1}{\rho^{\alpha/\alpha+1}} t_n^{\alpha/(\alpha+1)} + \frac{1}{\rho} t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho^2} t_n \\
&\leq 4 t_n^{\alpha/(\alpha+2)} + \frac{4}{\rho^2} t_n,
\end{aligned}$$

where the last line follows from the fact that $a(b^2 + b^{(\alpha+2)/(\alpha+1)} + b + 1) \leq 4a(b^2+1)$ with $a = \rho^{-2} t_n$ and $b = \rho t_n^{-1/(\alpha+2)}$. The proof is complete by specializing t_n in each context. □

We now provide the proofs for the two key Lemmas 2 and 3.

Proof of Lemma 2. Case 1: $F(z) = \mathbb{1}(z < 0)$ is 0-1 loss.

Properties (a) and (b) directly follow from Theorem 3.1. To prove (c), we expand the variance by

$$\begin{aligned}
\mathrm{Var}\left[\ell_\pi(\phi; (\boldsymbol{X}, Y)) - \ell_\pi(f_{\mathrm{bayes},\pi}, (\boldsymbol{X}, Y)\right] &\lesssim \mathbb{E}|\ell_\pi(\phi; (\boldsymbol{X}, Y)) - \ell_\pi(f_{\mathrm{bayes},\pi}, (\boldsymbol{X}, Y)|^2 \\
&\lesssim \mathbb{E}|\ell_\pi(\phi; (\boldsymbol{X}, Y)) - \ell_\pi(f_{\mathrm{bayes},\pi}, (\boldsymbol{X}, Y)| \\
&\lesssim \mathbb{E}\left|\,|\mathrm{sgn}\bar{Y}_\pi - \mathrm{sgn}\phi(\boldsymbol{X})| - |\mathrm{sgn}\bar{Y}_\pi - f_{\mathrm{bayes},\pi}(\boldsymbol{X})|\,\right| \\
&\leq \mathbb{E}|\mathrm{sgn}\phi - f_{\mathrm{bayes},\pi}|, \qquad (42)
\end{aligned}$$

where the second line comes from the boundedness of 0-1 loss, and the third line comes from the boundedness of weight $|\bar{Y}_\pi|$, and fourth line comes from the inequality $||a-b| - |c-b|| \leq |a-b|$ for $a, b, c \in \{-1, 1\}$. Here we have absorbed the constant multipliers in $\lesssim$. Therefore, the conclusion (41) then directly follows by applying Remark 3 to (42).

Case 2: $F(z) = (1-z)_+$ is hinge loss.

Property (a) was firstly introduced in Wang et al. [2008, Lemma 1], and here we provide an alternative proof.

A direct calculation (see Lemma 5) shows that

$$\mathrm{Risk}_{\pi,F}(\phi) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \geq \mathbb{E}|\phi - f_{\mathrm{bayes},\pi}||f - \pi| \geq 0,$$

Therefore, $\inf_{\text{all }\phi} \text{Risk}_{\pi,F}(\phi) = \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})$. Property (40) is from Scott [2011, Corollary 1] (see also Theorem C.1 in Section C).

Case 3: When $F(z) = 2\min(1, (1-z)_+)$ is psi-loss.

Again, the property (a) follows from Wang et al. [2008, Lemma 1]. For the property (40), we use Theorem C.1 to find the transformation function ψ that relates 0-1 risk to F-risk:

$$\psi(\text{Risk}_\pi(\phi) - \text{Risk}_\pi(f_{\text{bayes},\pi})) \le \text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}).$$

To put our problem in the context of Theorem C.1, we need additional notation. For any function measurable $g\colon x \mapsto g(x)$, we write $g = g^+ - g^{-1}$, where g^+ and g^- are two non-negative functions given by

$$g^+(x) = \max\{g(x), 0\} = \begin{cases} g(x), & \text{if } g(x) > 0, \\ 0, & \text{otherwise,} \end{cases} \quad g^-(x) = \max\{-g(x), 0\} = \begin{cases} -g(x), & \text{if } g(x) < 0, \\ 0, & \text{otherwise.} \end{cases}$$

Under this notation, we have $|g| = g^+ + g^{-1}$.

Define the conditional F-risk

$$C_{\pi,F}(\boldsymbol{X}, t) := F(t)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^+ + F(-t)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^-.$$

A direct calculation shows that

$$C_{\pi,F}(\boldsymbol{X}, t) = \begin{cases} 2\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^-, & \text{if } t \ge 1, \\ 2\mathbb{E}_{Y|\boldsymbol{X}}|Y-\pi| - 2t\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^+, & \text{if } t \in [0,1), \\ 2\mathbb{E}_{Y|\boldsymbol{X}}|Y-\pi| + 2t\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^-, & \text{if } t \in [-1,0), \\ 2\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^+, & \text{if } t < -1. \end{cases}$$

Therefore, following the notation of Theorem C.1, we have

$$H_{\pi,F}(\boldsymbol{X}) := \inf_{t\in\mathbb{R}\colon t(f(\boldsymbol{X})-\pi)\le 0} C_{\pi,F}(\boldsymbol{X}, t) - \inf_{t\in\mathbb{R}} C_{\pi,F}(\boldsymbol{X}, t) = 2|f(\boldsymbol{X}) - \pi|.$$

Applying Theorem C.1 to the above setup gives the excess risk transformation rule: $\psi : z \to 2|z|$. Therefore, the property (40) is proved.

To prove (41), notice that

$$\begin{aligned} &\text{Var}\left\{|\bar{Y}_\pi|\left[F(\phi(\boldsymbol{X})\text{sgn}\bar{Y}_\pi) - F(f_{\text{bayes},\pi}(\boldsymbol{X})\text{sgn}\bar{Y}_\pi)\right]\right\} \\ \lesssim\ &\mathbb{E}|\bar{Y}_\pi||F(\phi(\boldsymbol{X})\text{sgn}\bar{Y}_\pi) - F(f_{\text{bayes},\pi}(\boldsymbol{X})\text{sgn}\bar{Y}_\pi)| \\ \lesssim\ &\underbrace{\mathbb{E}\left|1 - \text{sgn}(\phi(\boldsymbol{X})\bar{Y}_\pi) - F(f_{\text{bayes},\pi}(\boldsymbol{X})\text{sgn}\bar{Y}_\pi)\right|}_{=:\text{(i)}} + \underbrace{\mathbb{E}|\bar{Y}_\pi|\left|F(\phi(\boldsymbol{X})\text{sgn}\bar{Y}_\pi) - \left(1 - \text{sgn}(\phi(\boldsymbol{X})\bar{Y}_\pi)\right)\right|}_{=:\text{(ii)}} \end{aligned} \tag{43}$$

The first term (i) is bounded as follows

$$\begin{aligned}\text{(i)} = \mathbb{E}\left|\operatorname{sgn}(\phi(\boldsymbol{X})\bar{Y}_\pi) - \operatorname{sgn}(f_{\text{bayes},\pi}(\boldsymbol{X})\bar{Y}_\pi)\right| &\lesssim d_\Delta(S_\phi, S_{\text{bayes}}(\pi)) \\ &\lesssim d_\pi^\alpha(S_\phi, S_{\text{bayes}}(\pi)) + \frac{1}{\rho(\pi,\mathcal{N})} d_\pi(S_\phi, S_{\text{bayes}}(\pi)),\end{aligned}$$

where the first line uses the fact that $F(1) = 0$ and $F(-1) = 2$, and last inequality is from Theorem 3.2. Here we define indicator set corresponding ϕ as $S_\phi = \{\boldsymbol{X} \in \mathcal{X} \colon \phi(\boldsymbol{X}) \geq 0\}$. The second term (ii) is bounded as follows

$$\begin{aligned}\text{(ii)} &= \mathbb{E}\left[|\bar{Y}_\pi| F(\phi(\boldsymbol{X})\operatorname{sgn}\bar{Y}_\pi) - |\bar{Y}_\pi|\left(1 - \operatorname{sgn}(\phi(\boldsymbol{X})\bar{Y}_\pi)\right)\right] \\ &= \mathbb{E}\left[|\bar{Y}_\pi| F(\phi(\boldsymbol{X})\operatorname{sgn}\bar{Y}_\pi) - |\bar{Y}_\pi| F(f_{\text{bayes},\pi}(\boldsymbol{X})\operatorname{sgn}\bar{Y}_\pi)\right] \\ &\quad + \mathbb{E}\left[|\bar{Y}_\pi|(1 - \operatorname{sgn}(f_{\text{bayes},\pi}\bar{Y}_\pi)) - |\bar{Y}_\pi|(1 - \operatorname{sgn}(\phi(\boldsymbol{X})\bar{Y}_\pi))\right] \\ &\leq \left[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right] + d_\pi(S_\phi, S_{\text{bayes}}(\pi)),\end{aligned}$$

where the first equality is based on $F(z) = 1 - \operatorname{sgn}(z)$ if $z = 1$ or -1, and the last inequality is from definition of $d_\pi(\cdot,\cdot)$. Notice we have $d_\pi(S_\phi, S_{\text{bayes}}(\pi)) = \text{Risk}_\pi(\phi) - \text{Risk}_\pi(f_{\text{bayes},\pi})$ by definition. Therefore, the proof is complete by combining (43), (40) and bounds (i)-(ii).

Case 4: $F(z) = \min((1-z)_+, T)$ for T-truncated hinge loss, for given $T \geq 2$. A direct calculation (c.f. Remark 7 after Lemma 5) shows that

$$\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) \geq \mathbb{E}|\phi^T - f_{\text{bayes},\pi}||f - \pi| \geq 0,$$

where $\phi^T \colon \mathcal{X} \to [-(T-1),\ (T-1)]$ denotes the $(T-1)$-truncation of ϕ,

$$\phi^T = \begin{cases} T-1 & \text{if } \phi > T-1, \\ \phi, & \text{if } |\phi| \leq T-1, \\ -(T-1), & \text{if } \phi < -(T-1). \end{cases} \tag{44}$$

Therefore, $\inf_{\text{all }\phi} \text{Risk}_{\pi,F}(\phi) = \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})$. To show property (40), we again use Theorem C.1 to find the transformation function ψ that relates 0-1 risk to F-risk:

$$\psi(\text{Risk}_\pi(\phi) - \text{Risk}_\pi(f_{\text{bayes},\pi})) \leq \text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}).$$

Using similar arguments as in Case 3, we obtain the conditional F-risk

$$C_{\pi,F}(\boldsymbol{X}, t) = \begin{cases} \min\left\{T,\ (1+t)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^-\right\}, & \text{if } t \geq 1, \\ \mathbb{E}_{Y|\boldsymbol{X}}|Y-\pi| - t(f(\boldsymbol{X}) - \pi), & \text{if } t \in [0,1), \\ \mathbb{E}_{Y|\boldsymbol{X}}|Y-\pi| + t(f(\boldsymbol{X}) - \pi), & \text{if } t \in [-1,0), \\ \min\left\{T,\ (1-t)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)^+\right\}, & \text{if } t < -1. \end{cases}$$

Therefore, following the notation of Theorem C.1, we have

$$H_{\pi,F}(\boldsymbol{X}) := \inf_{t\in\mathbb{R}:\, t(f(\boldsymbol{X})-\pi)\le 0} C_{\pi,F}(\boldsymbol{X},t) - \inf_{t\in\mathbb{R}} C_{\pi,F}(\boldsymbol{X},t) = |f(\boldsymbol{X})-\pi|.$$

Applying Theorem C.1 to the above setup gives the excess risk transformation rule: $\psi : z \to |z|$. Therefore, the property (40) is proved.

To prove (41), we use Lemma 5 and the boundedness condition of $\|F\|_\infty \le T$. Specifically, we bound the variance using the L-1 distance between ϕ and $f_{\text{bayes},\pi}$,

$$\begin{aligned}
&\text{Var}\left\{|\bar{Y}_\pi|\left[F(\phi(\boldsymbol{X})\text{sgn}\bar{Y}_\pi) - F(f_{\text{bayes},\pi}(\boldsymbol{X})\text{sgn}\bar{Y}_\pi\right]\right\}\\
\le\ & 4\mathbb{E}|F(\phi(\boldsymbol{X})\text{sgn}\bar{Y}_\pi) - F(f_{\text{bayes},\pi}(\boldsymbol{X})\text{sgn}\bar{Y}_\pi)|^2\\
\lesssim\ & T\mathbb{E}|F(\phi(\boldsymbol{X})\text{sgn}\bar{Y}_\pi) - F(f_{\text{bayes},\pi}(\boldsymbol{X})\text{sgn}\bar{Y}_\pi)|\\
\lesssim\ & T\mathbb{E}|\phi^T - f_{\text{bayes},\pi}|,
\end{aligned}$$

where $T > 0$ is the upper bound of truncated hinge loss, the first inequality comes from the boundedness of $|\bar{Y}_\pi|$, the second inequality comes from the boundedness of the T-truncated hinge loss, and the last line comes from the definition of F. Applying Remark 7 in Lemma 5 to the last inequality complete the proof.

□

Proof of Lemma 3. Fix $\pi \notin \mathcal{N}$, and write $\rho = \rho(\pi,\mathcal{N})$, $L_n = t_n^{(\alpha+1)/(\alpha+2)}$. We first consider the (bounded) psi-loss, and then consider the (unbounded) hinge loss. The 0-1 loss incurs only slight difference in the proof, and we address this case at last.

Case 1: psi-loss, $\lambda \asymp L_n + t_n/\rho$, and $a_n \lesssim L_n$.

For any function $\phi \in \Phi(r, s_1, s_2)$ of consideration, define the empirical weighted F-risk

$$\widehat{\text{Risk}}_{\pi,F}(\phi) = \frac{1}{n}\sum_{i=1}^n \ell_{\pi,F}(\phi;(\boldsymbol{X}_i, Y_i)).$$

Under the notation, our estimate $\hat{\phi}_{\pi,F}$ is the minimizer of the regularized empirical F-risk,

$$\hat{\phi}_{\pi,F} = \underset{\phi\in\Phi(r,s_1,s_2)}{\arg\min}\left\{\widehat{\text{Risk}}_{\pi,F}(\phi) + \lambda\|\phi\|_F^2\right\}. \tag{45}$$

We are interested in the convergence rate of $\hat{\phi}_{\pi,F}$'s excess risk, $\text{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})$. Let $L_n \asymp t_n^{(\alpha+1)/(\alpha+2)}$ denote the desired convergence rate to seek. By the definition of $\hat{\phi}_{\pi,F}$, we have

$$\widehat{\text{Risk}}_{\pi,F}(\hat{\phi}_{\pi,F}) + \lambda\|\hat{\phi}_{\pi,F}\|_F^2 \le \widehat{\text{Risk}}_{\pi,F}(\phi_\pi^{(n)}) + \lambda J^2,$$

where $\phi_\pi^{(n)}$ is a sequence of functions in Assumption 1(a). Therefore, we have the following inclusion

of probability events,

$$
\begin{aligned}
&\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \mathrm{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \geq 2L_n \right\} \\
\subset &\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \exists \phi \in \Phi(r, s_1, s_2),\ \text{s.t. } \mathrm{Risk}_{\pi,F}(\phi; (\boldsymbol{X}, Y)) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \geq 2L_n, \right. \\
&\qquad\qquad \left. \text{and } \widehat{\mathrm{Risk}}_{\pi,F}(\phi) + \lambda\|\phi\|_F^2 \leq \widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) + \lambda J^2 \right\} \\
\subset &\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \sup_{\substack{\phi\in\Phi(r,s_1,s_2) \\ \mathrm{Risk}_{\pi,F}(\phi;(\boldsymbol{X},Y))-\mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi})\geq 2L_n}} \left[\widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) + \lambda J^2 - \widehat{\mathrm{Risk}}_{\pi,F}(\phi) - \lambda\|\phi\|_F^2\right] \geq 0 \right\} \\
\subset &\bigcup_{\phi\in A_{s,k}} \left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \sup_{\phi\in A_{s,k}} \left[\widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) + \lambda J^2 - \widehat{\mathrm{Risk}}_{\pi,F}(\phi) - \lambda\|\phi\|_F^2\right] \geq 0 \right\}. \qquad (46)
\end{aligned}
$$

In the last line of (46), we have partitioned the set $\{\phi \in \Phi(r, s_1, s_2) \colon \mathrm{Risk}_{\pi,F}(\phi; (\boldsymbol{X}, Y)) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \geq 2L_n\}$ into a union of $A_{s,k}$, with

$$
A_{s,k} = \{\phi \in \Phi(r, s_1, s_2) \colon (s+1)L_n \leq \mathrm{Risk}_{\pi,F}(\phi) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) < (s+2)L_n, (k-1)J^2 \leq \|\phi\|_F^2 < kJ^2\},
$$

for $s, k = 1, 2, \ldots$.

Let Γ denote the target probability for the first line in (46). To bound Γ, it suffices to bound the sum of probabilities over sets $A_{s,k}$. For each $A_{s,k}$, we consider the centered empirical process,

$$
\begin{aligned}
v_n(\phi) &:= \left[\widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) - \widehat{\mathrm{Risk}}_{\pi,F}(\phi)\right] - \left[\mathrm{Risk}_{\pi,F}(\phi_\pi^{(n)}) - \mathrm{Risk}_{\pi,F}(\phi)\right] \\
&= \frac{1}{n}\sum_{i\in[n]} \left\{\ell_{\pi,F}(\phi_\pi^{(n)}; (\boldsymbol{X}_i, Y_i)) - \ell_{\pi,F}(\phi; (\boldsymbol{X}_i, Y_i)) - \mathbb{E}\left[\ell_{\pi,F}(\phi_\pi^{(n)}; (\boldsymbol{X}_i, Y_i)) - \ell_{\pi,F}(\phi; (\boldsymbol{X}_i, Y_i))\right]\right\}. \qquad (47)
\end{aligned}
$$

Notice that

$$
\begin{aligned}
\mathrm{Risk}_{\pi,F}(\phi) - \mathrm{Risk}_{\pi,F}(\phi_\pi^{(n)}) &= \mathrm{Risk}_{\pi,F}(\phi) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) + \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) - \mathrm{Risk}_{\pi,F}(\phi_\pi^{(n)}) \\
&\geq (s+1)L_n - a_n \\
&\geq sL_n, \qquad (48)
\end{aligned}
$$

where the first inequality is from the fact that $\phi \in A_{s,k}$ and Assumption 1(a), and the last inequality uses the condition that $a_n \lesssim L_n$.

Combining the definition of v_n in (47) and inequality (48) gives (46) as

$$
\Gamma \leq \sum_{s,k=1}^{\infty} \mathbb{P}\left\{\sup_{\phi\in A_{s,k}} \left[v_n(\phi) - \lambda\|\phi\|_F^2\right] \geq sL_n - \lambda J^2\right\}
$$

$$\leq \sum_{s,k=1}^{\infty} \mathbb{P}\left\{ \sup_{\phi \in A_{s,k}} v_n(\phi) \geq sL_n + \lambda(k-2)J^2 =: M(s,k) \right\}, \tag{49}$$

where $M(s,k) > 0$ for all $s, k \in \mathbb{N}$ from the condition $\lambda J^2 \leq L_n/2$ by the choice of (L_n, λ). Verification of this condition is deferred to when we specify (λ, L_n) in (53).

The variance of the empirical process is bounded by

$$\begin{aligned}
&\sup_{\phi \in A_{s,k}} \mathrm{Var}\left[\ell_{\pi,F}(\phi_\pi^{(n)}; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)) \right] \\
&\leq \sup_{\phi \in A_{s,k}} 2\Big\{ \mathrm{Var}\left[\ell_{\pi,F}(\phi_\pi^{(n)}; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(f_{\mathrm{bayes},\pi}; (\boldsymbol{X}, Y) \right] \\
&\qquad\qquad + \mathrm{Var}\left[\ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(f_{\mathrm{bayes},\pi}; (\boldsymbol{X}, Y) \right] \Big\} \\
&\lesssim [M(s,k)]^{\alpha/(1+\alpha)} + \frac{M(s,k)}{\rho} =: V(s,k),
\end{aligned} \tag{50}$$

where the last inequality is from Lemma 2.

We next bound the right-hand-side of (49) by choosing (L_n, λ) that satisfies the conditions in Theorem C.2. (The specification of (L_n, λ) is deferred to the next paragraph). Once such (L_n, λ) is chosen, then it follows from Theorem C.2 that

$$\begin{aligned}
\Gamma &\lesssim \sum_{s,k} \exp\left(-\frac{nM^2(s,k)}{V(s,k) + 2M(s,k)} \right) \\
&\lesssim \sum_{s,k} \exp(-\rho n M(s,k)) = \sum_{s,k} \exp\left(-n\rho s L_n - n\rho\lambda(k-2)J^2 \right) \\
&\leq \left(\frac{e^{-n\rho L_n}}{1 - e^{-n\rho L_n}} \right) \left(\frac{e^{n\rho\lambda J^2}}{1 - e^{-n\rho\lambda J^2}} \right) \\
&\leq \frac{e^{-n\rho L_n/2}}{(1 - e^{-n\rho L_n})(1 - e^{-n\rho\lambda J^2})},
\end{aligned} \tag{51}$$

where the first line uses the boundedness of psi-loss, and the last inequality is from the condition $\lambda J^2 \leq L_n/2$ by the choice of (λ, L_n).

Now, we specify (L_n, λ) that satisfies the condition of Theorem C.2. The pair (L_n, λ) is determined by the solution to the following inequality,

$$\sup_{k \geq 1, s \geq 1} \frac{1}{x} \int_x^{\sqrt{x^{\alpha/(\alpha+1)} + x/\rho}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \Phi^k, \|\cdot\|_2)} d\varepsilon \lesssim n^{1/2}, \quad \text{where } x = sL_n + \lambda(k-2)J^2. \tag{52}$$

In particular, the smallest L_n satisfying (52) yields the best upper bound of the error rate. Here $\mathcal{H}_{[\,]}(\varepsilon, \Phi^k, \|\cdot\|_2)$ denotes the L_2-norm, ε-bracketing number (c.f. Definition 4) for function family Φ^k, and, we have denoted $\Phi^k = \{\phi \in \Phi(r, s_1, s_2) \colon \|\phi\|_F^2 \leq k\}$, i.e., the subset of functions in $\Phi(r, s_1, s_2)$ with magnitudes bounded by k, for $k \geq 1$.

It remains to solve for the smallest possible L_n in (52). Based on Lemma 7, the inequality (52) is satisfied with the choice

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{t_n}{\rho}, \quad \text{and} \quad \lambda = \frac{L_n}{2J^2}, \tag{53}$$

where

$$t_n = \begin{cases} \frac{r d_{\max}}{n}, & \text{low-rank model } \phi \in \Phi(r), \\ \frac{r(s_1+s_2)\log d_{\max}}{n}, & \text{low-rank and two-way sparse model } \phi \in \Phi(r, s_1, s_2). \end{cases} \tag{54}$$

Notice that this choice of (L_n, λ) guarantees the conditions for earlier calculation in (49) and (51). Specifically, we have the assumption $\lambda \asymp t_n^{(\alpha+1)/(\alpha+2)} + t_n/\rho$ from the setup of Theorem B.1. Given this λ, we choose an L_n with a suitable constant factor such that $\lambda J^2 \le L_n/2$. So conditions for earlier calculation in (49) and (51) are verified.

Plugging (53) into (51) gives that

$$\begin{aligned} \Gamma &= \mathbb{P}\left[\mathrm{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \ge L_n\right] \\ &\le \frac{e^{-n\rho L_n/2}}{(1-e^{-n\rho L_n})(1-e^{-n\rho\lambda J^2})} \\ &\lesssim e^{-n\rho L_n} \le e^{-nt_n}, \end{aligned}$$

where the last line uses the fact that $\rho\lambda J^2 \asymp \rho L_n \gtrsim t_n \gtrsim n^{-1}$ by (53) and (54). The proof is then complete by bounding the 0-1 risk by F-risk.

Case 2: hinge loss, $\lambda \asymp L_n + t_n/\rho$, and $a_n \lesssim L_n$.

For unbounded hinge loss, we seek to bound the F'-risk of $\hat{\phi}_{\pi,F}$, where F' is T-truncated version of F. The general strategy is to evaluate $\hat{\phi}_{\pi,F}$'s error using F'-risk. Note that the estimate $\hat{\phi}_{\pi,F}$ (45) is defined under unbounded loss F. Therefore, the inclusion (46) changes to

$$\begin{aligned} &\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \mathrm{Risk}_{\pi,F'}(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_{\pi,F'}(f_{\mathrm{bayes},\pi}) \ge 2L_n\right\} \\ \subset &\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \exists \phi \in \Phi(r, s_1, s_2), \text{ s.t. } \mathrm{Risk}_{\pi,F'}(\phi; (\boldsymbol{X}, Y)) - \mathrm{Risk}_{\pi,F'}(f_{\mathrm{bayes},\pi}) \ge 2L_n, \right. \\ &\qquad\qquad \left. \text{and } \widehat{\mathrm{Risk}}_{\pi,F}(\phi) + \lambda\|\phi\|_F^2 \le \widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) + \lambda J^2\right\} \\ \subset &\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \exists \phi \in \Phi(r, s_1, s_2), \text{ s.t. } \mathrm{Risk}_{\pi,F'}(\phi; (\boldsymbol{X}, Y)) - \mathrm{Risk}_{\pi,F'}(f_{\mathrm{bayes},\pi}) \ge 2L_n, \right. \\ &\qquad\qquad \left. \text{and } \widehat{\mathrm{Risk}}_{\pi,F'}(\phi) + \lambda\|\phi\|_F^2 \le \widehat{\mathrm{Risk}}_{\pi,F'}(\phi_\pi^{(n)}) + \lambda J^2\right\}, \end{aligned}$$

where the last line comes from

$$\widehat{\mathrm{Risk}}_{\pi,F'}(\phi) \le \widehat{\mathrm{Risk}}_{\pi,F}(\phi) \text{ for all } \phi \in \Phi(r, s_1, s_2) \quad \text{and} \quad \widehat{\mathrm{Risk}}_{\pi,F'}(\phi_\pi^{(n)}) = \widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}),$$

because the truncation constant is $T = \max(2, J) > \max(2, \sup_n \|\phi_\pi^{(n)}\|_F)$. Notice that the last

line is exactly the same with (46) except F being replaced by F'. The remaining proof follows the same line of argument as in Case 1. In particular, we invoke Lemma 2 to control the variance-to-mean relationship for bounded F'-loss in (50). The final conclusion follows from the excess bound inequality for T-truncated risk (c.f. Lemma 2).

Case 3: 0-1 loss, $\lambda = 0$ and $a_n = 0$.

Under 0-1 loss, only the sign, but not the magnitude, of ϕ affects the 0-1 risk. Without loss of generality, we assume $\|\phi\|_F \le 1$. Then, we have the following inclusion of probability events,

$$\begin{aligned}
\Gamma &:= \left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \mathrm{Risk}_\pi(\hat{\phi}_\pi) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi}) \ge L_n\right\} \\
&\subset \Bigg\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \exists \phi \in \Phi(r, s_1, s_2),\ \text{s.t. } \mathrm{Risk}_\pi(\phi; (\boldsymbol{X}, Y)) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi}) \ge L_n \\
&\qquad\qquad \text{and } \widehat{\mathrm{Risk}}_\pi(\phi) \le \widehat{\mathrm{Risk}}_\pi(f_{\mathrm{bayes},\pi})\Bigg\} \\
&\subset \left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \sup_{\substack{\phi\in\Phi(r,s_1,s_2) \\ \mathrm{Risk}_\pi(\phi;(\boldsymbol{X},Y))-\mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi})\ge L_n}} \left[\widehat{\mathrm{Risk}}_\pi(f_{\mathrm{bayes},\pi}) - \widehat{\mathrm{Risk}}_\pi(\phi)\right] \ge 0\right\} \\
&\subset \bigcup_{\phi\in A_s} \left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \sup_{\phi\in A_s} \left[\widehat{\mathrm{Risk}}_\pi(f_{\mathrm{bayes},\pi}) - \widehat{\mathrm{Risk}}_\pi(\phi)\right] \ge 0\right\},
\end{aligned}$$

where we have partitioned $\{\phi \in \Phi(r, s_1, s_2) \colon \mathrm{Risk}_\pi(\phi; (\boldsymbol{X}, Y)) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi}) \ge L_n\}$ into a union of A_s with

$$A_s = \{\phi \in \Phi(r, s_1, s_2) \colon sL_n \le \mathrm{Risk}_\pi(\phi) - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi}) < (s+1)L_n\},$$

for $s = 1, 2, \ldots$. Similar to Case 1, we consider empirical process,

$$v_n(\phi) := [\widehat{\mathrm{Risk}}_\pi(f_{\mathrm{bayes},\pi}) - \widehat{\mathrm{Risk}}_\pi(\phi)] - \mathrm{Risk}_\pi(f_{\mathrm{bayes},\pi}) - \mathrm{Risk}_\pi(\phi)].$$

Then, our goal is to bound

$$\Gamma \le \sum_{s=1}^{\infty} \mathbb{P}\left\{\sup_{\phi\in A_s} v_n(\phi) \ge sL_n := M(s)\right\}. \tag{55}$$

Notice the variance of empirical process is bounded by

$$\sup_{\phi\in A_s} \mathrm{Var}\left[\ell_{\pi,F}(f_{\mathrm{bayes},\pi}; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)\right] \lesssim [M(s)]^{\alpha/(1+\alpha)} + \frac{M(s)}{\rho} =: V(s)$$

where F is 0-1 loss and the inequality is from Lemma 2. Applying Lemma 7 with finite $k = 1$ and

$\lambda = 0$ shows that $L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + t_n/\rho$ satisfies the conditions Theorem C.2, where

$$t_n = \begin{cases} \frac{r d_{\max}}{n}, & \text{low-rank model } \phi \in \Phi(r), \\ \frac{r(s_1+s_2)\log d_{\max}}{n}, & \text{low-rank and two-way sparse model } \phi \in \Phi(r, s_1, s_2). \end{cases}$$

Therefore, it follows from Theorem C.2 and (55) that

$$\begin{aligned} \Gamma &\lesssim \sum_s \exp\left(\frac{-nM^2(s)}{V(s)+M(s)}\right) \\ &\lesssim \sum_s \exp(-\rho s n L_n) \\ &\leq \left(\frac{e^{-n\rho L_n}}{1-e^{-n\rho L_n}}\right) \\ &\lesssim e^{-nt_n}, \end{aligned}$$

where the last line uses the fact that $\rho L_n \gtrsim t_n \gtrsim \frac{1}{n}$ by our choice of L_n and t_n.

□

B.5 Proofs of Theorem 3.4, Theorem 4.1, and Part (b) in Theorem 5.1

Proof of Theorem 3.4. For any $t \geq t_n$ with t_n specified in Theorem B.1, define the event

$$A = \left\{ \|\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)\|_1 \leq t^{\alpha/(2+\alpha)} + \frac{t}{\rho^2(\pi, \mathcal{N})} \text{ for all } \pi \in \mathcal{H} \right\}.$$

We first show that the event A implies

$$\|\hat{f} - f\|_1 \lesssim t^{\alpha/(\alpha+2)} + \frac{1}{H} + tH. \tag{56}$$

It follows from the definition of $\hat{f}$ that

$$\begin{aligned} \|\hat{f} - f\|_1 &= \mathbb{E}\left| \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}\hat{\phi}_\pi - f \right| \\ &\leq \mathbb{E}\left| \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} (\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)) \right| + \mathbb{E}\left| \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(f-\pi) - f \right| \\ &\leq \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \|\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)\|_1 + \frac{1}{H}, \end{aligned} \tag{57}$$

where the last line comes from the triangle inequality and the inequality

$$\left| \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(f(\boldsymbol{X}) - \pi) - f(\boldsymbol{X}) \right| \leq \frac{1}{H}, \quad \text{for all } \boldsymbol{X} \in \mathcal{X}.$$

It suffices to bound the first term in (57).

Theorem B.1 shows that the sign function accuracy depends on the closeness of $\pi \in \mathcal{H}$ to the mass points in $\mathcal{H}$. Therefore, we partition the level set $\pi \in \mathcal{H}$ based on their closeness to $\mathcal{H}$. Specifically, let $\mathcal{N}_H \stackrel{\text{def}}{=} \bigcup_{\pi' \in \mathcal{N}} \left(\pi' - \frac{1}{H}, \pi' + \frac{1}{H}\right)$ denote the set of levels at least $\frac{1}{H}$-close to the mass points. We expand left hand side of (57) by

$$\begin{aligned}
&\frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \|\operatorname{sgn}\hat{\phi}_\pi - \operatorname{sgn}(f-\pi)\|_1 \\
&= \frac{1}{2H+1} \sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H} \|\operatorname{sgn}\hat{\phi}_\pi - \operatorname{sgn}(f-\pi)\|_1 + \frac{1}{2H+1} \sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H^c} \|\operatorname{sgn}\hat{\phi}_\pi - \operatorname{sgn}(f-\pi)\|_1. \qquad (58)
\end{aligned}$$

By assumption, the first term of (58) involves only finite number of summands and thus can be bounded by $4C/(2H+1)$ where $C > 0$ is a constant such that $|\mathcal{N}| \le C$. We bound the second term using the explicit forms of $\rho(\pi, \mathcal{N})$ in the sequence $\pi \in \Pi \cap \mathcal{N}_H^c$.

$$\begin{aligned}
\frac{1}{2H+1} \sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H^c} \|\operatorname{sgn}\hat{\phi}_\pi - \operatorname{sgn}(f-\pi)\|_1 &\lesssim \frac{1}{2H+1} \sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H^c} t^{\alpha/(2+\alpha)} + \frac{t}{2H+1} \sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H^c} \frac{1}{\rho^2(\pi, \mathcal{N})} \qquad (59) \\
&\le t^{\alpha/(2+\alpha)} + \frac{t}{2H+1} \sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H^c} \sum_{\pi' \in \mathcal{N}} \frac{1}{|\pi - \pi'|^2} \\
&\le t^{\alpha/(2+\alpha)} + \frac{t}{2H+1} \sum_{\pi' \in \mathcal{N}} \sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H^c} \frac{1}{|\pi - \pi'|^2} \\
&\le t^{\alpha/(2+\alpha)} + 2CHt,
\end{aligned}$$

where the first inequality uses the property of event A, and the last inequality follows from Lemma 4. Combining (57), (58) and (59) comletes the proof of (56); that is

$$\mathbb{P}\left(\|\hat{f} - f\|_1 \lesssim t^{\alpha/(\alpha+2)} + \frac{1}{H} + tH \right) \ge \mathbb{P}(A), \quad \text{for all } t \ge t_n. \qquad (60)$$

Based on Remark 4 and union bound over $\pi \in \mathcal{H}$, we have,

$$\begin{aligned}
\mathbb{P}(A) &\ge 1 - \sum_{\pi \in \mathcal{H}} \mathbb{P}\left(\|\operatorname{sgn}\hat{\phi}_\pi - \operatorname{sgn}(f-\pi)\|_1 \le t^{\alpha/(2+\alpha)} + \frac{t}{\rho^2(\pi, \mathcal{N})} \text{ for a given } \pi \right) \\
&\gtrsim 1 - (2H+1)\exp(-nt) \gtrsim 1 - \exp(-nt + \log H). \qquad (61)
\end{aligned}$$

We choose $t \asymp t_n \log H$ in (61) so that $\log H$ is negligible compared to nt. It then follows from (60) and (61) that

$$\|\hat{f} - f\|_1 \lesssim (t_n \log H)^{\alpha/(\alpha+2)} + \frac{1}{H} + t_n H \log H,$$

with probability at least $1 - \exp(-nt_n \log H) \ge 1 - \exp(-nt_n)$. Setting $H \asymp t_n^{-1/2}$ yields the desired conclusion.

Proofs of Theorem 4.1 and Part (b) in Theorem 5.1 follow the same argument with t_n specified in Theorem B.1. □

Lemma 4. *Fix $\pi' \in \mathcal{N}$ and a sequence $\mathcal{H} = \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$ with $H \geq 2$. Then,*

$$\sum_{\pi \in \mathcal{H} \cap \mathcal{N}_H^c} \frac{1}{|\pi - \pi'|^2} \leq 4H^2.$$

Proof of Lemma 4. Notice that all points $\pi \in \mathcal{H} \cap \mathcal{N}_H^c$ satisfy $|\pi - \pi'| > \frac{1}{H}$ for all $\pi' \in \mathcal{N}$. We use this fact to compute the sum

$$\begin{aligned}
\sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \frac{1}{|\pi - \pi'|^2} &= \sum_{\frac{h}{H} \in \mathcal{H} \cap \mathcal{N}_H^c} \frac{1}{|\frac{h}{H} - \pi'|^2} \\
&\leq 2H^2 \sum_{h=1}^{H} \frac{1}{h^2} \\
&\leq 2H^2 \left\{ 1 + \int_1^2 \frac{1}{x^2} dx + \int_2^3 \frac{1}{x^2} dx + \cdots + \int_{H-1}^{H} \frac{1}{x^2} dx \right\} \\
&= 2H^2 \left(1 + \int_1^H \frac{1}{x^2} dx \right) \leq 4H^2,
\end{aligned}$$

where the third line uses the monotonicity of $\frac{1}{x^2}$ for $x \geq 1$. □

B.6 Proofs of Theorem 4.2 and Theorem A.1

Proof of Theorem 4.2. Theorem 4.2 follows from the same line of proof as in Theorem 3.4, with slight modification to account for discrete measure space. For any matrix $\boldsymbol{Z} \in \mathbb{R}^{d \times d}$, we use $f_{\boldsymbol{Z}} \colon [d]^2 \to \mathbb{R}$ to denote the function induced by matrix $\boldsymbol{Z}$ such that $f_{\boldsymbol{Z}}(\omega) = \boldsymbol{Z}(\omega)$ for $\omega \in [d]^2$. Set $\mathcal{X} = \{\boldsymbol{e}_i^T \boldsymbol{e}_j \colon (i,j) \in [d]^2\}$ be the discrete feature space, and $n = |\Omega|$ the sample size. Under this set up, $\|\hat{f} - f\|_1 = \mathbb{E}_{\boldsymbol{X}} |\hat{f}(\boldsymbol{X}) - f(\boldsymbol{X})| = \mathbb{E}_\omega |\hat{\boldsymbol{\Theta}}(\omega) - \boldsymbol{\Theta}(\omega)| = \mathrm{MAE}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})$. Notice that the small tolerance $\Delta s = 1/d^2$ in the pseudo density is dominated by the derived convergence rate. Applying Theorem 3.4 to this setting finishes the proof. □

Proof of Theorem A.1. By setting $s = \log(d_{\max})$ in Lemma 8, we have

$$\mathbb{P}(\|\boldsymbol{E}\|_\infty \geq \sqrt{4\sigma^2 \log d}) \leq 2d^{-2}.$$

We divide the sample space into two exclusive events:

- Event I: $\|\boldsymbol{E}\|_\infty \geq \sqrt{4\sigma^2 \log d}$;
- Event II: $\|\boldsymbol{E}\|_\infty < \sqrt{4\sigma^2 \log d}$.

Because the Event I occurs with probability tending to zero, we restrict ourselves to the Event II only by following the proof of Theorem 3.3. We summarize the key difference compared to

Section 3.4. For ease of notation, define $\bar{\boldsymbol{Y}} = \boldsymbol{Y} - \pi$ and $\bar{\boldsymbol{\Theta}} = \boldsymbol{\Theta} - \pi$. Let $\ell_\omega(\cdot,\cdot)$ denote the 0-1 loss evaluated at the ω-th value of two matrices. We expand the variance by

$$\begin{aligned}
\mathrm{Var}\left[\ell_\omega\left(\boldsymbol{Z}, \bar{\boldsymbol{Y}}_\Omega\right) - \ell_\omega\left(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{Y}}_\Omega\right)\right] &\le \mathbb{E}|\ell_\omega(\boldsymbol{Z}(\omega), \bar{\boldsymbol{Y}}(\omega)) - \ell_\omega(\bar{\boldsymbol{\Theta}}(\omega), \bar{\boldsymbol{Y}}(\omega))|^2 \\
&= \mathbb{E}|\bar{\boldsymbol{Y}}(\omega) - \bar{\boldsymbol{\Theta}}(\omega) + \bar{\boldsymbol{\Theta}}(\omega)|^2 |\mathrm{sgn}\boldsymbol{Z}(\omega) - \mathrm{sgn}\bar{\boldsymbol{\Theta}}(\omega)| \\
&\le 2\left(4\sigma^2 \log d + 2\right) \mathbb{E}|\mathrm{sgn}\boldsymbol{Z} - \mathrm{sgn}\bar{\boldsymbol{\Theta}}| \\
&\lesssim (\sigma^2 \log d)\mathrm{MAE}(\mathrm{sgn}\boldsymbol{Z}, \mathrm{sgn}\bar{\boldsymbol{\Theta}}),
\end{aligned} \tag{62}$$

where the third line uses the facts $\|\bar{\boldsymbol{\Theta}}\|_\infty \le 2$ and $\|\bar{\boldsymbol{Y}} - \bar{\boldsymbol{\Theta}}\|_\infty^2 = \|\boldsymbol{E}\|_\infty^2 < 4\sigma^2 \log d$ within the Event II; the last line comes from the definition of MAE and the asymptotic $\sigma^2 \log d \gg 1$ provided that $\sigma > 0$ with d sufficiently large.

Based on (62), the (α, π)-smoothness of $\boldsymbol{\Theta}$ implies that for all measurable functions $f_{\boldsymbol{Z}}$, we have

$$\begin{aligned}
&\mathrm{Var}\left[\ell_\omega\left(\boldsymbol{Z}, \bar{\boldsymbol{Y}}_\Omega\right) - \ell_\omega\left(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{Y}}_\Omega\right)\right] \\
&\lesssim (\sigma^2 \log d)\left\{\left[\mathbb{E}\left[\ell_\omega\left(\boldsymbol{Z}, \bar{\boldsymbol{Y}}_\Omega\right) - \ell_\omega\left(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{Y}}_\Omega\right)\right]\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho}\mathbb{E}\left[\ell_\omega\left(\boldsymbol{Z}, \bar{\boldsymbol{Y}}_\Omega\right) - \ell_\omega\left(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{Y}}_\Omega\right)\right]\right\}.
\end{aligned} \tag{63}$$

The empirical process with variance-to-mean relationship (63) gives that

$$\mathbb{P}\left(\mathrm{Risk}(\hat{\boldsymbol{Z}}) - \mathrm{Risk}(\bar{\boldsymbol{\Theta}}) \ge L_d\right) \lesssim \exp(-|\Omega| t_d), \tag{64}$$

where the convergence rate L_d is obtained by the same way in the proof of Lemma 7 to make sure the conditions hold in Theorem C.2,

$$L_d \asymp t_d^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho} t_d, \quad \text{with } t_d = \frac{r\sigma^2 d \log d}{|\Omega|}. \tag{65}$$

Combining (64) and (65), we obtain that, with high probability,

$$\mathrm{Risk}(\hat{\boldsymbol{Z}}) - \mathrm{Risk}(\bar{\boldsymbol{\Theta}}) \lesssim \left(\frac{r\sigma^2 d \log d}{|\Omega|}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})}\left(\frac{r\sigma^2 d \log d}{|\Omega|}\right), \tag{66}$$

where constants have been absorbed into the $\lesssim$ relationship. Therefore, combining (66) and the proof of Theorem B.1 completes the proof for sign matrix estimation error in (27). The signal estimation error follows the same proof of Theorem 3.4. □

C Auxiliary lemmas

Lemma 5 (Hinge loss and L-1 distance)**.** *Consider the same set-up as in Theorem 5.1. Let $F(z) = (1-z)_+$ be the hinge loss. Then, the L-1 distance between ϕ and $f_{\text{bayes},\pi}$ is bounded by their excess risk; i.e,*

$$\|\phi - f_{\text{bayes},\pi}\|_1 \leq [\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi,\mathcal{N})}[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})],$$

for all functions $\phi\colon \mathcal{X} \to \mathbb{R}$.

Remark 7 (Truncated hinge loss and L-1 distance)**.** With little modification in the proof, similar inequality also holds for T-truncated hinge loss $F(z) = \min(T, (1-z)_+)$ with $T \geq 2$. Specifically,

$$\|\phi^T - f_{\text{bayes},\pi}\|_1 \leq [\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi,\mathcal{N})}[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})],$$

where $\phi^T\colon \mathcal{X} \to [-(T-1),\ (T-1)]$ is the truncated ϕ defined in (44).

Proof of Lemma 5. For ease of notation, we drop the random variable $\boldsymbol{X}$ in the function expression, and simply use ϕ, $f_{\text{bayes},\pi}$, f, to represent the trace function, Bayes rule, and the regression function, respectively. The meaning should be clear given the contexts.

We expand the excess risk using the definition of hinge loss,

$$\begin{aligned}
&\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) \\
&= \mathbb{E}[|\bar{Y}_\pi|(1 - \phi \text{sgn}\bar{Y}_\pi)_+] - \mathbb{E}[|\bar{Y}_\pi|(1 - f_{\text{bayes},\pi}\text{sgn}\bar{Y}_\pi)_+] \\
&= \int_{\boldsymbol{X}} (1-\phi)_+ \int_{y>\pi} (y-\pi) dy d\mathbb{P}_{\boldsymbol{X}} + \int_{\boldsymbol{X}} (1+\phi)_+ \int_{y\leq\pi} (\pi - y) dy d\mathbb{P}_{\boldsymbol{X}} \\
&\quad - \int_{\boldsymbol{X}} (1 - f_{\text{bayes},\pi})_+ \int_{y>\pi} (y-\pi) dy d\mathbb{P}_{\boldsymbol{X}} - \int_{\boldsymbol{X}} (1 + f_{\text{bayes},\pi})_+ \int_{y\leq\pi} (\pi - y) dy d\mathbb{P}_{\boldsymbol{X}}. \qquad (67)
\end{aligned}$$

In order to evaluate the integral, we divide the domain $\boldsymbol{X}$ into four exclusive regions:

- Region I $= \{\boldsymbol{X}\colon f < \pi \text{ and } \phi \geq -1\}$. In this region, $f_{\text{bayes},\pi} = -1$, and the integrant in (67) reduces to

$$\begin{aligned}
\Phi_{\text{I}} &:= [(1-\phi)_+ - 2]\,\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)\mathbb{1}(Y>\pi) + (\phi+1)_+\mathbb{E}_{Y|\boldsymbol{X}}(\pi - Y)\mathbb{1}(Y\leq\pi) \\
&\geq -(\phi+1)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)\mathbb{1}(Y>\pi) - (\phi+1)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)\mathbb{1}(Y\leq\pi) \\
&= (\phi+1)(\pi - f) = |\phi - f_{\text{bayes},\pi}||f-\pi|.
\end{aligned}$$

- Region II $= \{\boldsymbol{X}\colon f < \pi \text{ and } \phi < -1\}$. In this region, $f_{\text{bayes},\pi} = -1$, and the integrant in (67) reduces to

$$\Phi_{\text{II}} := -(\phi+1)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)\mathbb{1}(Y>\pi) \geq -|\phi+1|(f-\pi) = |\phi - f_{\text{bayes},\pi}||f-\pi|.$$

- Region III $= \{\boldsymbol{X}\colon f \geq \pi \text{ and } \phi \leq 1\}$. In this region, $f_{\text{bayes},\pi} = 1$, and the integrant in (67) reduces to

$$\begin{aligned}\Phi_{\text{III}} &:= (1-\phi)_+ \mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)\mathbb{1}(Y>\pi) + [(1+\phi)_+ - 2]\,\mathbb{E}_{Y|\boldsymbol{X}}(\pi - Y)\mathbb{1}(Y \leq \pi) \\ &\geq (1-\phi)\mathbb{E}_{Y|\boldsymbol{X}}(Y-\pi)\mathbb{1}(Y>\pi) + (\phi-1)\mathbb{E}_{Y|\boldsymbol{X}}(\pi - Y)\mathbb{1}(Y \leq \pi) \\ &= (1-\phi)(f-\pi) = |\phi - f_{\text{bayes},\pi}||f-\pi|.\end{aligned}$$

- Region IV $= \{\boldsymbol{X}\colon f \geq \pi \text{ and } \phi > 1\}$. In this region, $f_{\text{bayes},\pi} = 1$, and the integrant in (67) reduces to

$$\Phi_{\text{IV}} := (\phi-1)\mathbb{E}_{Y|\boldsymbol{X}}(\pi - Y)\mathbb{1}(Y \leq \pi) \geq (\phi-1)(f-\pi) = |\phi - f_{\text{bayes},\pi}||f-\pi|.$$

Therefore, the integral is evaluated as

$$\begin{aligned}\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) &= \int_{\text{I}} \Phi_{\text{I}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\text{II}} \Phi_{\text{II}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\text{III}} \Phi_{\text{III}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\text{IV}} \Phi_{\text{IV}} d\mathbb{P}_{\boldsymbol{X}} \\ &\geq \mathbb{E}|\phi - f_{\text{bayes},\pi}||f-\pi|. \end{aligned} \tag{68}$$

Note that the function $|f-\pi|$ is α-smooth. Using the same techniques as in Theorem 3.2 to the last line of (68), we conclude

$$\mathbb{E}|\phi - f_{\text{bayes},\pi}| \lesssim [\text{Risk}_{\pi,F}(f) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi,\mathcal{N})}\left[\text{Risk}_{\pi,F}(f) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right].$$

□

Definition 4 (Bracketing number)**.** Consider a function set Φ, and let $\varepsilon > 0$. We call $\{(f_m^l, f_m^u)\}_{m=1}^M$ an L_2-metric, ε-bracketing function set of Φ, if for every $f \in \Phi$, there exists an $m \in [M]$ such that

$$f_m^l(\boldsymbol{X}) \leq f(\boldsymbol{X}) \leq f_m^u(\boldsymbol{X}), \quad \text{for all } \boldsymbol{X} \in \mathbb{R}^{d\times d},$$

and

$$\|f_m^l - f_m^u\|_2 \stackrel{\text{def}}{=} \sqrt{\mathbb{E}_{\boldsymbol{X}}|f_m^l(\boldsymbol{X}) - f_m^u(\boldsymbol{X})|^2} \leq \varepsilon, \text{ for all } m = 1,\ldots,M.$$

The bracketing number with L_2-metric, $\mathcal{H}_{[\,]}(\varepsilon,\ \Phi,\ \|\cdot\|_2)$, is defined as the logarithm of the smallest cardinality of the ε-bracketing function set of Φ.

Lemma 6 (Bracketing number for bounded functions in $\Phi(r,s_1,s_2)$ and $\Phi(r)$)**.** *Let $\Phi(r,s_1,s_2)$ denote the trace function family*

$$\Phi(r,s_1,s_2) = \{\phi\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle + b \,\big|\text{rank}(\boldsymbol{B}) \leq r, \text{supp}(\boldsymbol{B}) \leq (s_1,s_2), |b| \leq \|\boldsymbol{B}\|_F + 1\},$$

We use $\|\phi\|_F \stackrel{\text{def}}{=} \|\boldsymbol{B}\|_F$ to denote the coefficient magnitude. Assume, for simplicity, $\mathbb{P}(\|\boldsymbol{X}\|_F \leq 1) = 1$. For any given $k \geq 1$, let $\Phi^k = \{f \in \Phi(r,s_1,s_2)\colon \|\phi\|_F^2 \leq k\}$ denote the sub-class of functions

with coefficient magnitudes bounded by k. Then,

$$\mathcal{H}_{[\,]}(\varepsilon,\ \Phi^k,\ \|\cdot\|_2) \lesssim r(s_1+s_2)\log\frac{kd_{\max}}{\varepsilon}.$$

Furthermore, when we consider $\Phi^k = \{\phi \in \Phi(r)\colon \|\phi\|_F^2 \le k\}$ with $(s_1, s_2) = (d_1, d_2)$, then

$$\mathcal{H}_{[\,]}(\varepsilon,\ \Phi^k,\ \|\cdot\|_2) \lesssim rd_{\max}\log\frac{k}{\varepsilon}.$$

Proof of Lemma 6. For any given $k \ge 1$, define a matrix family

$$\mathcal{B} = \left\{ \begin{bmatrix} \boldsymbol{B} & 0 \\ 0 & b \end{bmatrix} \in \mathbb{R}^{(d_1+1)\times(d_2+1)}\colon \operatorname{rank}(\boldsymbol{B}) \le r,\ \operatorname{supp}(\boldsymbol{B}) \le (s_1,s_2),\ |b| \le \sqrt{k}+1,\ \|\boldsymbol{B}\|_F \le \sqrt{k} \right\}$$

By definition of trace functions, there is an onto mapping from matrices in $\mathcal{B}$ to functions in Φ^k; i.e.

$$\Phi^k \subset \left\{ \phi\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle + b \;\middle|\; \begin{bmatrix} \boldsymbol{B} & 0 \\ 0 & b \end{bmatrix} \in \mathcal{B} \right\}.$$

Furthermore, every pair of functions $\phi_1 = \langle \boldsymbol{X}, \boldsymbol{B}_1\rangle + b_1,\ \phi_2 = \langle \boldsymbol{X}, \boldsymbol{B}_2\rangle + b_2 \in \Phi^k$ satisfies the norm relationship

$$\|\phi_1 - \phi_2\|_2 \le \|\phi_1 - \phi_2\|_\infty = \sup_{\|\boldsymbol{X}\|_F \le 1} |\langle \boldsymbol{X}, \boldsymbol{B}_1\rangle + b_1 - \langle \boldsymbol{X}, \boldsymbol{B}_2\rangle - b_2| \le \sqrt{\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2 + |b_1 - b_2|^2}.$$

Based on Kosorok [2007, Theorem 9.23], the L_2-metric, (2ε)-bracketing number in Φ^k is bounded by

$$\mathcal{H}_{[\,]}(2\varepsilon,\ \Phi^k,\ \|\cdot\|_2) \le \mathcal{H}\left(\varepsilon,\ \mathcal{B},\ \|\cdot\|_F\right)$$

where $\mathcal{H}$ denotes the log covering number for the (non-bracketing) set. Therefore, it suffices to bound $\mathcal{H}(\varepsilon,\ \mathcal{B},\ \|\cdot\|_F)$ where $\mathcal{B}$ is included in a $(\sqrt{5k})$-ball by definition of $\mathcal{B}$. Now fix two subsets $S_1, S_2 \subset [d]$ with $|S_1| = s_1$ and $|S_2| = s_2$, where $|\cdot|$ denotes the cardinality of the sets. Let $\mathcal{B}_{S_1,S_2} \subset \mathcal{B}$ denote the subset of matrices satisfying $\boldsymbol{B}(i,j) = 0$ whenever $(i,j) \notin S_1 \times S_2$. Based on Candes and Plan [2011, Lemma 3.1], the log covering number for $\mathcal{B}_{S_1,S_2}$ is

$$\mathcal{H}\left(\varepsilon,\ \mathcal{B}_{S_1,S_2},\ \|\cdot\|_F\right) \lesssim r(s_1+s_2+1)\log\left(\frac{k}{\varepsilon}\right). \tag{69}$$

In view of the construction $\mathcal{B} \subset \bigcup\{\mathcal{B}_{S_1,S_2}\colon S_1 \times S_2 \subset [d_1]\times[d_2], |S_1| = s_1, |S_2| = s_2\}$, an ε-covering set $\mathcal{B}$ is then given by the union of ε-covering set of $\mathcal{B}_{S_1,S_2}$. Using Stirling's bound, we derive that

$$\begin{aligned}
\mathcal{H}(\varepsilon,\ \mathcal{B},\ \|\cdot\|_F) &\le \log\left\{ \binom{d_{\max}}{s_1}\binom{d_{\max}}{s_2} \exp\left[\mathcal{H}(\varepsilon, \mathcal{B}_{S_1,S_2}, \|\cdot\|_F)\right] \right\} \\
&\le s_1\log\frac{d_{\max}}{s_1} + s_2\log\frac{d_{\max}}{s_2} + C'r(s_1+s_2+1)\log\frac{k}{\varepsilon} \\
&\le Cr(s_1+s_2)\log\frac{kd_{\max}}{\varepsilon},
\end{aligned}$$

where $C, C' > 0$ are constants.

The result for the case of $(s_1, s_2) = (d_1, d_2)$ directly follows from (69). □

Lemma 7 (Local complexity of $\Phi(r, s_1, s_2)$ and $\Phi(r)$)**.** *Define* $\Phi^k = \{f \in \Phi(r, s_1, s_2) \colon \|f\|_F^2 \le k\}$ *for all* $k \ge 1$*; i.e.,* Φ^k *is the subset of functions in* $\Phi(r, s_1, s_2)$ *with coefficient magnitudes bounded by* k*. Set*

$$L_n \gtrsim \left(\frac{r(s_1+s_2)\log d_{\max}}{n}\right)^{\frac{\alpha+1}{\alpha+2}} + \frac{1}{\rho(\pi,\mathcal{N})}\left(\frac{r(s_1+s_2)\log d_{\max}}{n}\right), \quad \text{and } \lambda = \frac{L_n}{2J^2}. \tag{70}$$

Then, the following inequality is satisfied for all $k \ge 1$ *and* $s \ge 1$*.*

$$\frac{1}{x}\int_x^{\sqrt{x^{\alpha/(\alpha+1)} + \frac{x}{\rho(\pi,\mathcal{N})}}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon,\ \Phi^k,\ \|\cdot\|_2)}\, d\varepsilon \lesssim n^{1/2}, \quad \text{where } x := sL_n + \lambda(k-2)J^2.$$

The result for $\Phi(r)$ *is the same except* $\log d_{\max}$ *being removed from* L_n *in* (70)*.*

Proof of Lemma 7. To simplify the notation, we write $\rho = \rho(\pi, \mathcal{N})$, $d = d_{\max}$, and define

$$g(x,k) = \frac{1}{x}\int_x^{\sqrt{x^{\alpha/(\alpha+1)} + x/\rho}} \sqrt{r(s_1+s_2)\log\left(\frac{kd}{\varepsilon}\right)}\, d\varepsilon, \quad \text{for all } k \ge 1,$$

where we have inserted the bracketing number based on Lemma 6. Notice that

$$\begin{aligned} g(x,k) &\le \frac{\sqrt{r(s_1+s_2)}}{L}\int_x^{\sqrt{x^{\alpha/(\alpha+1)} + x/\rho}} \sqrt{\log\left(\frac{kd}{x}\right)}\, d\varepsilon \\ &\le \sqrt{r(s_1+s_2)(\log k + \log d - \log x)}\left(\frac{\sqrt{x^{\alpha/(2\alpha+2)}} + \sqrt{x/\rho}}{x} - 1\right) \\ &\le \sqrt{r(s_1+s_2)(\log k + \log d)}\left(\frac{1}{x^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho x}}\right) =: \bar{g}(x,k), \end{aligned} \tag{71}$$

where the second line follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$. Since the upper bound $\bar{g}(x,k)$ is decreasing function with respect to $x > 0$, it suffices to show that $\bar{g}(x,k) \le n^{1/2}$ for all $k \ge 1$ and $s = 1$; that is, to show $\bar{g}(\bar{x},k) \lesssim n^{1/2}$ for all $k \ge 1$ under the choice

$$\bar{x} := L_n + \lambda(k-2)J^2 \ge \frac{k}{2}\left\{\left(\frac{r(s_1+s_2)\log d}{n}\right)^{\frac{\alpha+1}{\alpha+2}} + \frac{1}{\rho}\left(\frac{r(s_1+s_2)\log d}{n}\right)\right\}.$$

Plugging the above expression into the last line of (71) gives

$$\bar{g}(\bar{x},k) \le n^{1/2}\sqrt{\frac{\log k + \log d}{(k/2)^{(\alpha+2)/(\alpha+1)}\log d}} + n^{1/2}\sqrt{\frac{\log k + \log d}{(k/2)\log d}} \le C' n^{1/2}, \quad \text{for all } k \ge 1,$$

where $C' > 0$ is a constant independent of k and d. The proof is therefore complete. □

Lemma 8 (sub-Gaussian maximum). *Let $X_1, \ldots, X_n$ be independent sub-Gaussian zero-mean random variables with variance proxy σ^2. Then, for any $s > 0$,*

$$\mathbb{P}\left\{\max_{1\le i\le n} |X_i| \ge \sqrt{2\sigma^2(\log n + s)}\right\} \le 2e^{-s}.$$

Proof of Lemma 8. The conclusion follows from

$$\mathbb{P}[\max_{1\le i\le n} |X_i| \ge u] \le \sum_{i=1}^{n} \mathbb{P}[|X_i| \ge u] \le 2ne^{-\frac{u^2}{2\sigma^2}} = 2e^{-s},$$

where we set $u = \sqrt{2\sigma^2(\log n + s)}$. □

We state the results from Scott [2011, Theorem 1] in our contexts.

Theorem C.1 (Theorem 1 in Scott [2011]). *Let $\mathrm{Risk}_{\pi,F}(\cdot)$ be weighted F-risk defined in Section 5.4 of the main paper with $\pi \in [-1, 1]$. Define the conditional risk*

$$C_{\pi,F}(\boldsymbol{X}, t) := F(t)\mathbb{E}_{Y|\boldsymbol{X}}(Y - \pi)^+ + F(-t)\mathbb{E}_{Y|\boldsymbol{X}}(Y - \pi)^-,$$

and associated function $H_{\pi,F}$:

$$H_{\pi,F}(\boldsymbol{X}) = \inf_{t\in\mathbb{R}:\, t(f(\boldsymbol{X})-\pi)\le 0} C_{\pi,F}(\boldsymbol{X}, t) - \inf_{t\in\mathbb{R}} C_{\pi,F}(\boldsymbol{X}, t).$$

Let $f(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$. For any $\varepsilon \ge 0$, define

$$g(\varepsilon) = \begin{cases} \inf_{\boldsymbol{X}\in\mathcal{X}:|f(\boldsymbol{X})-\pi|\ge\varepsilon} H_{\pi,F}(\boldsymbol{X}), & \varepsilon > 0, \\ 0 & \varepsilon = 0. \end{cases}$$

Now set $\psi = g^{**}$ where g^{**} denotes the Fenchel-Legendre biconjugate of g. Then, for any decision function $\phi\colon \mathcal{X} \to \mathbb{R}$ and any distribution of $(\boldsymbol{X}, Y)$, we have

$$\psi\left(\mathrm{Risk}_\pi(\phi) - \inf_{all\ \phi} \mathrm{Risk}_\pi(\phi)\right) \le \mathrm{Risk}_{\pi,F}(\phi) - \inf_{all\ \psi} \mathrm{Risk}_{\pi,F}(\phi).$$

Theorem C.2 (Theorem 3 in Shen and Wong [1994]). *Let $\mathcal{F}$ be a class of functions defined on $\mathcal{X}$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \le T$. Let $(\boldsymbol{X}_i)_{i=1}^n$ be i.i.d. random variables with distribution $\mathbb{P}_{\boldsymbol{X}}$ over $\mathcal{X}$. Set $\sup_{f\in\mathcal{F}} \mathrm{Var} f(\boldsymbol{X}) = V < \infty$. Define the empirical process $\hat{\mathbb{E}}f = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i)$. Define x_n^* to be the solution to the following inequality*

$$\frac{1}{x}\int_x^{\sqrt{V}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_2)}d\varepsilon \lesssim \sqrt{n}.$$

Suppose $\sqrt{V} \le T$ and

$$x_n^* \lesssim \frac{V}{T}, \quad \text{and} \quad \mathcal{H}_{[\,]}(\sqrt{V}, \mathcal{F}, \|\cdot\|_2) \lesssim \frac{n(x_n^*)^2}{V}.$$

Then, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\hat{\mathbb{E}}f-\mathbb{E}f\geq x_n^*\right)\lesssim\exp\left(-\frac{n(x_n^*)^2}{V+Tx_n^*}\right).$$