# Nonparametric ~~trace regression~~ learning with matrix-valued predictors in high dimensions

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

**Abstract**

We consider the problem of learning the relationship between a binary label response and a high-dimensional matrix-valued predictor. Prediction based on matrices or networks has recently surged in brain connectivity studies, sensor network localization, and integrative genomics. Traditional regression methods take a parametric procedure by imposing a priori functional form between variables. These parametric models, however, are inadequate for structure learning and often fail in accurate prediction. Here, we develop a learning reduction framework to address a range of learning tasks from classification to binary regression for matrix-valued predictors. Our proposal achieves interpretable prediction via a low-rank two-way sparse halfspace learning for the target function level sets. Unlike earlier approaches, our method efficiently exploits the important features in the high-dimensional matrices. Statistical accuracy, excess risk bounds, and efficient algorithms are established. We demonstrate the advantage of our method over previous approaches through simulations and applications to human brain connectome data.

*Keywords:* Nonparametric learning, high-dimensional matrices, sparse and low-rank models, classification, regression, feature selection

## 1 Introduction

~~We consider the problem of estimating~~

$$Y = \langle \boldsymbol{X}, \boldsymbol{B} \rangle + \varepsilon, \text{ with } \operatorname{rank}(\boldsymbol{B}) \le r,$$

~~for $i = 1, \ldots, n$. We call the function $\boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{X}\boldsymbol{B}^T)$ the trace function. Extension of trace regression to exponential family has been proposed,~~

$$\mathbb{E}(Y_i|\boldsymbol{X}_i) = g(\langle \boldsymbol{X}_i, \boldsymbol{B} \rangle), \text{ with a known } g\colon \mathbb{R} \to \mathbb{R} \text{ and low-rank } \boldsymbol{B}.$$

[1]Department of Statistics, University of Wisconsin-Madison
[2]Department of Biostatistics and Epidemiology, University of California at Berkeley
[3]Department of Mathematics, University of Arizona

~~Now we propose a nonparametric trace regression.~~

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad \text{for } i = 1, \ldots, n$$

where $f(\cdot)\colon \mathbb{R}^{d\times d} \to \mathbb{R}$ is an unknown low-rank sign-representable function.

~~The function $f(\cdot)\colon \mathbb{R}^{d\times d} \to \mathbb{R}$ is called low-rank sign-representable function at $\pi \in [-1, 1]$, if the function has the same sign as a rank-$r$ trace function; i.e.,~~

$$\operatorname{sign}(f(\boldsymbol{X}) - \pi) = \operatorname{sign}(\langle \boldsymbol{X}, \boldsymbol{B} \rangle), \quad \text{for all } \boldsymbol{X} \in \mathcal{X},$$

~~where $\boldsymbol{B} = \boldsymbol{B}(\pi) \in \mathbb{R}^{d\times d}$ is a rank-$r$ matrix. If the function $f$ is rank-$r$ sign-representable at all $\pi$ except for a finite number of points, then we call the function $f$ the $r$-globally sign-representable.~~

Matrix-valued predictors ubiquitously arise in modern applications. In brain connectivity studies, for example, individuals are represented by their brain networks, and the networks quantify the connectivity patterns over a set of nodes (brain regions of interest). Human connectom project (**?**) has constructed brain networks for over 1,200 individuals using Desikan atlas with 68 brain nodes. Structural connectivity is measured for every pair of nodes, resulting in an adjacency matrix of size $68 \times 68$ for each individual. This connectivity matrix provides important information for disease prediction. Other examples include electroencephalography studies of alcoholism (**?**). Researchers measure the voltage values from 64 channels of electrodes on 256 subjects for 256 time points. The study yields a $256 \times 64$ matrix-valued feature, along with a binary indicator of subject being alcoholic or not. Identifying the relationship between EEG signals and alcoholism is helpful for disease diagnostics.

We consider the problem of binary regression between a matrix-valued predictor $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$ and a binary label response $y \in \{-1, 1\}$. The regression problem concerns the estimation of ~~conditional probability of $Y$ given $\boldsymbol{X}$, $\mathbb{P}(Y = 1|\boldsymbol{X}) = p(\boldsymbol{X})$~~conditional expectation $\mathbb{E}(y|\boldsymbol{X})$, or equivalently in our setting, the conditional probability $\mathbb{P}(y = 1|\boldsymbol{X}) = \mu(\boldsymbol{X})$. There are generally two types of approaches underlying many existing regression methods. The first approach is the parametric method such as linear regression or logistic regression (**??**) that imposes a priori function form on $\mu(\cdot)$. Parametric methods estimate a fixed number of parameters in $\mu(\cdot)$ and often suffers from poor prediction. The second approach is the nonparametric regression (**??**) that adaptively learns the form of $\mu(\cdot)$ from growing data. Nonparametric methods allow the number of parameters to increase with sample size, thereby providing flexibility in the prediction. Current nonparametric methods aim for accurate prediction at the cost of hard interpretability. In the aforementioned and many other scientific studies, however, researchers are interested in *interpretable prediction* (**?**), where the goal is to not only make accurate prediction but also identify important features for descriptive simplicity. Efficient methods that achieve both have yet to be developed.

### 1.1 Our contributions

This paper develops a nonparametric method that efficiently exploits the matrix-valued feature space for interpretable prediction. Our contributions are summarized below.

First, we develop a new nonparametric regression paradigm – learning reduction (**?**) – that solves regression through a ~~sequence~~ series of classifications. Figure 1a illustrates our learning reduction approaches. We convert a hard ~~*What* problem~~ problem *"what is the value of ~~$p(\boldsymbol{X})$"~~ $\mu(\boldsymbol{X})$?"* to an easier ~~*Is*~~*where* problem *"~~Is $p(\boldsymbol{X})$ smaller than $\pi$~~where does $\mu(\boldsymbol{X})$ fall below a given value?"* ~~for a given $\pi \in (0,1)$~~. The latter is related to classification which is statistically easier to address. Most existing nonparametric regression such as nearest neighbors, local polynomials, regularing kernels, etc, extract information from data by exploiting the neighborhood structure in the domain space (**?**). In contrast, our reduction method focuses on the neighborhood structure in the range space, and uses a ~~sequence~~ series of classifications to bridge the range space and the domain space. It is worth noting that the former is a 1-dimensional bounded random variable whereas the latter is a $d_1 d_2$-dimensional random variable. The shifted focus to the simpler range space brings both statistical and computational benefits; see Sections 3 and 4 for details.

Second, we develop a low-rank two-way sparse halfspace learning method to efficiently mitigate the curse of dimensionality in nonparametric prediction. Despite the popularity of nonparametric learning, the problem has unique challenges when predictors are matrices. In the brain network analysis, for example, the matrix-valued predictors represent connectivity networks with global structure (e.g. clusters, community hubs) and local structure (e.g. node degrees, edge connections)(**?**). Simultaneous prediction and structure learning are of practical importance for matrix-valued predictors. We show that our low-rank two-way sparse structure learning enables variable selection in high-dimensional matrices, thereby achieving high interpretability in prediction.

Third, we develop a large-margin based risk minimization for empirical estimation. Prediction accuracy guarantees are established for three matrix problems: classification, level-set estimation, and regression. Our error bound reveals the increased complexity in the three problems and demonstrates the success of learning reduction approach. Specifically, we show that the regression error is bounded by classification error,

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

where $\alpha \in [0,1]$ is a smoothness index of the regression function $\mu(\boldsymbol{X})$, $H \in \mathbb{N}_+$ is a resolution parameter, and $t_n = \mathcal{O}(n^{-1}\log d)$, $a_n \to 0$ under certain matrix models (see details in Sections 3-4). The results show the high-dimensional prediction consistency of our method that allows matrix dimension $d$ to grow sub-exponentially in sample size $n$. Furthermore, we find that the matrix

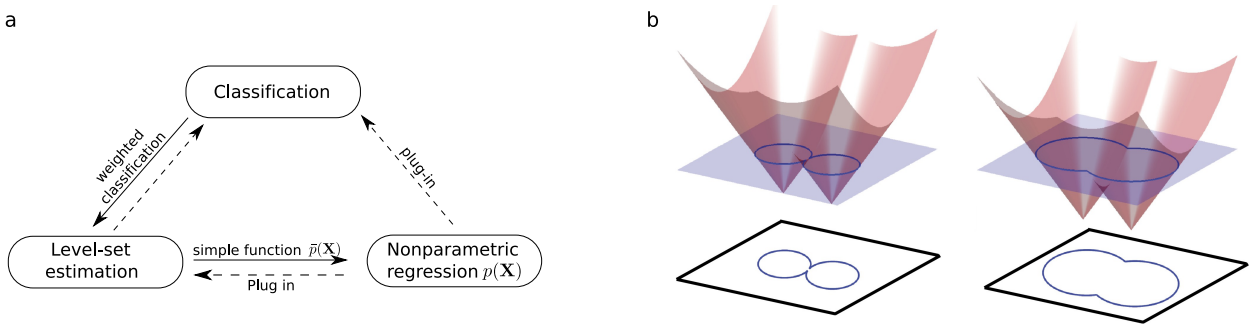



**Figure 1:** (a) Our learning reduction approach (solid line) to the three problems of interest. The classical plug-in approaches are depicted in dashed line. (b) Schematic diagram for nonparametric function estimation via level set aggregations (i.e, a ~~sequence~~ series of weighted classifications). Figure (b) is modified from **?**).

classification has a fast $\mathcal{O}(n^{-1})$ error bound whereas the regression has a slow $\mathcal{O}(n^{-1/2})$ error bound. We expect this general result may benefit other settings beyond matrix learning tasks.

Lastly, we supplement the general statistical properties by developing an alternating direction method of multipliers (ADMM) algorithm. The algorithm leverages recent advances in large-margin solvers and nonconvex optimization for low-rank sparse matrix learning. We illustrate the efficacy of our method through both simulations and data applications.

## 1.2 Related work

Our work is related to but also clearly distinctive from several lines of existing research. We review some of the main ideas that have emerged.

### 1.2.1 Parametric matrix regression

There has been increased interest in studying matrix-valued predictors for regression; for example, trace regression (**??**), network logistic regression (**??**), and regularized matrix regression (**??**). These methods fall into the broad category of parametric approaches that impose a priori functional form between variables. The most common model in this vein is generalized linear model (GLM) which relates conditional mean to predictors though a known link function. Given a matrix predictor $\boldsymbol{X}$ and a response variable $y$, the regression model takes the form

$$p\mathbb{E}(y|\boldsymbol{X}) = g(\langle \boldsymbol{B}, \boldsymbol{X} \rangle), \text{ with a known function } g(\cdot)\colon \mathbb{R} \to \mathbb{R},$$

where $\boldsymbol{B}$ is an unknown matrix with certain structural constraints depending on the specific problems. The GLM is a popular parametric regression, but nevertheless, this approach lacks the flexibility and often leads to inaccurate prediction in high dimensions.

### 1.2.2 Common nonparametric strategies

A variety of nonparametric regressions have been developed to overcome the limitation of parametric regression. For example, single index models (**?**) extend GLMs by allowing an unknown $g$ in order to provide flexibility in prediction. Nearest neighbor methods (**?**) estimate the function by taking the average of responses in the neighbored of predictors. Polynomial regression (**?**) and wavelets (**?**) perform curve fitting through basis expansions. Computational efficiency and statistical accuracy are both real concerns as the predictor dimension grows large, in nonparametric regression. Common strategies to mitigate the curse of dimensionality include imposing structural constraints, such as additivity, monotonicity, and/or sparsity, to the regression function. The nonparametric learning with structural constraints is still very much in development (**??**).

Three key challenges have yet to address when we consider matrix-valued predictors. First, matrix-valued predictors encode far richer information than vectors, as mentioned in the earlier brain network example. A naive approach is to transform the predictors into vectors and apply classical nonparametric regression. The practice of vectorization, however, destroys the two-way structural information in the original predictors. Second, most nonparametric methods rely on some notion of smoothness in the local neighborhood of predictors. In the context of matrix-valued predictors, however, the predictor space is huge and barely explored by small sample size data. Third, matrix-valued predictors often lead to a high dimensional problem with massive features. The relatively weak assumptions made by nonparametric methods bring new challenges to interpretation. Achieving accurate prediction while maintaining descriptive simplicity will be our goal.

### 1.2.3 Nonparametric methods with manifolds

Several attempts have been developed to address nonparametric regression with manifold-valued predictors. The manifolds here refer to general nonconventional predictors, such as matrices, tensors, networks, images, and so on. **?**) extends deep neural network to regression on low-dimensional manifold. **?**) develops a sparse additive model with tensor covariates by extending usual spline basis. **?**) also considers tensor covariates but instead proposes a broadcasting operation to impose nonlinearity to individual tensor entries. In computer vision, convolution neural network (CNN) has been widely adopted as a nonparametric tool for image data (**?**). These methods in some sense address the first and ~~(~~part of the second ~~) challenge~~ challenges mentioned in the earlier paragraph; they use the conventional wisdom of exploiting low-dimensional structure in the input data in order to achieve accurate prediction.

Our method share the same ground as theirs, but we provide a completely different paradigm for extracting information from massive high-dimensional data. We develop a low-rank, sparse halfspace learning in the matrix space, and allow the informative features in the matrix-valued predictor to vary from one level to another. Unlike the aforementioned methods, our method relies

little on the local neighborhood of predictors but rather on the local neighborhood of responses. More importantly, we do not only tailor a specific nonparametric tool for matrix-valued predictors; we provide a learning paradigm that potentially enables state-of-art classification tools, such as neural network, decision trees, and boosting, to address the more challenge regressing problem. The ability to import and adapt existing classification methods is one advantage of the proposed learning reduction framework. We also numerically compare our approach with CNN in Section 6.

### 1.2.4 Regression level set estimation

Level set method has a long history in statistics (**?**) and computational mathematics (**?**). The approach has drawn increased attention because of recent successes in neural network (**?**), density estimation (**?**), classification (**?**), and bandit optimization (**?**). Level sets provide an efficient approach to representing functions. Instead of constructing a function point-wise in the predictor space, the method focuses on the range space (see Figure 1b). The benefit bears the analogy of Riemann vs. Lebesgue integrals in the functional analysis, in the sense that the focus is shifted from domain space to range space. This approach is especially appealing for matrix binary regression, where the feature space is complicated whereas the response is a simple bounded univariate.

Earlier work on level set estimation considers only a single level (**??**). Statistical properties for plug-in level set estimates are studied in **??**). In comparison to our learning reduction, their analysis can be viewed as a *learning induction* approach where the regression function estimates are provided as inputs (Figure 1a). **?**) develops a conditional probability estimation method based on support vector machine (SVM), but their results are restricted to fixed number of features (e.g. $d = 2$ in their example) and vector predictors only. **?**) proposes a partition-based method for multiple level sets extraction, but again their goal is to estimate the level sets per se but not function estimation. None of these methods address the ~~high dimensionality or~~ regression problem or high-dimensional matrix-valued predictors as in our work.

## 1.3 Notation and organization

Let $\mathcal{X} \subset \mathbb{R}^{d_1 \times d_2}$ be the feature space. Given a function $f \colon \mathcal{X} \to \mathbb{R}$, we use $\operatorname{sign} f$ to denote its sign function, such that $\operatorname{sign} f(\boldsymbol{X}) = 1$ if $f(\boldsymbol{X}) > 0$ and $\operatorname{sign} f(\boldsymbol{X}) = -1$ otherwise. The notion of sign function also applies to sets in $\mathcal{X}$. We use $\operatorname{sign}(\boldsymbol{X} \in A)$ to denote the sign function induced by the set $A \subset \mathcal{X}$, i.e., a function taking value 1 on the event $\{\boldsymbol{X} \in A\}$ and -1 otherwise. We use shorthand $[n] := \{1, \ldots, n\}$ to denote the $n$-set for $n \in \mathbb{N}_+$ and use $|\cdot|$ to denote the cardinality of sets. Let $\|\cdot\|_p$ denote the vector $p$-norm for $p \geq 0$, and $\|\cdot\|_F$ be the matrix Frobenious norm. Given a $d_1$-by-$d_2$ matrix $\boldsymbol{B}$, we use $\boldsymbol{B}_i$ to denote the $i$-th row of $\boldsymbol{B}$. The $(p, q)$-norm of a matrix $\boldsymbol{B}$ is defined as $\|\boldsymbol{B}\|_{p,q} = \|\boldsymbol{b}\|_q$, 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 of the rows in $\boldsymbol{B}$. In particular, $\|\boldsymbol{B}\|_{1,0} = |\{i \in [d_1] \colon \boldsymbol{B}_i \neq 0\}|$ denotes the number of non-zero rows in $\boldsymbol{B}$. An event $E$ is said to occur "with high probability" if $\mathbb{P}(E)$ tends to 1 as the matrix

dimension $d_{\min} = \min(d_1, d_2) \to \infty$. We denote $a_n \asymp b_n$ if ~~$\lim_n b_n/a_n = c$~~ $\lim_n b_n/a_n \to c$ for some constant $c >$ and denote $a_n \lesssim b_n$ if ~~$\lim_n b_n/a_n = 0.$~~ $\lim_n b_n/a_n \to 0.$ We use $\mathbb{1}(\cdot)$ to denote the indicator function.

The rest of the paper is organized as follows. Section 2 presents a range of learning tasks, from classification to regression, for matrix-valued predictors. In Section 3, we develop a learning reduction approach by relating the regression to classification, and establish the oracle procedures of nonparamatric matrix regression. In Section 4, we present the empirical estimation and the associated optimization algorithm. Statistical accuracy, high-dimensional consistency, and excess risk bounds are presented in Section 5. We present the simulations in Section 6 and human connectome Project data analyses in Section 6.3. All technical proofs are deferred to Appendix.

## 2 Three learning problems

We present the main learning goals of our interest. Let $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ denote the matrix-valued predictor, $y \in \{-1, 1\}$ denote the binary label response, and $\mathbb{P}_{\boldsymbol{X},y}$ denote the unknown joint probability distribution over the pair $(\boldsymbol{X}, y)$. In the context of binary response, $y$ is a Bernoulli random variable with conditional probability ~~$p(\boldsymbol{X}) \overset{\text{def}}{=} \mathbb{P}(y = 1|\boldsymbol{X})$~~ $\mu(\boldsymbol{X}) \overset{\text{def}}{=} \mathbb{P}(y = 1|\boldsymbol{X})$; we generally make no parametric assumptions on the marginal distribution $\mathbb{P}_{\boldsymbol{X}}$ or form of $\mu(\boldsymbol{X})$.

Suppose that we observe a sample of $n$ training data points, $(\boldsymbol{X}_1, y_1), \ldots, (\boldsymbol{X}_n, y_n)$, identically and independently distributed (i.i.d.) according to $\mathbb{P}_{\boldsymbol{X},y}$. Let $(\boldsymbol{X}_{\text{new}}, y_{\text{new}})$ be a future unseen test point drawn independently from the same distribution. Our goal is to predict $y_{\text{new}}$ based on $\boldsymbol{X}_{\text{new}}$. We often omit the subscript "new" and simply write $(\boldsymbol{X}, y)$ for the prototypical test point. The relevant probabilistic statements should be interpreted as taken jointly with respect to $(\boldsymbol{X}, y)$.

We consider three learning problems: classification, level set estimation, and regression.

2.1 *Matrix classification*: Classification is the problem of predicting the label $y \in \{-1, 1\}$ to which the new matrix $\boldsymbol{X}$ belongs. A prediction rule (also called a classifier) decides that $y = 1$ if $\boldsymbol{X} \in S$ and ~~$y = -1$~~ $y = 0$ if $\boldsymbol{X} \notin S$, where $S$ is a Borel subset of $\mathcal{X}$. We formulate the classification problem as choosing a classifier $S \in \mathcal{S}$, from a given set of candidate classifiers $\mathcal{S}$, that minimizes the expected classification error

$$R(S) = \mathbb{P}_{\boldsymbol{X},y}\left[y \neq \text{sign}(\boldsymbol{X} \in S)\right]. \tag{1}$$

The $R(S)$ is also called the classification risk. When the candidate set $\mathcal{S}$ consists of all Borel subsets of $\mathcal{X}$, the minimizer of (2.1) is called the Bayes classifier. It is known that the Bayes classifier can be written as

$$S_{\text{bayes}}(1/2) = \{\boldsymbol{X} \in \mathcal{X} : p\mu(\boldsymbol{X}) \geq 1/2\}. \tag{2}$$

Note that the minimizer of (2.1) is non-unique, because arbitrary prediction rules are allowed on the

boundaries ~~$\partial S_{\text{bayes}}(1/2) = \{\boldsymbol{X} \in \mathcal{X} : p(\boldsymbol{X}) = 1/2\}$~~ $\partial S_{\text{bayes}}(1/2) = \{\boldsymbol{X} \in \mathcal{X} : \mu(\boldsymbol{X}) = 1/2\}$. Without loss of generality, we will use (2.1) as the canonical form of the Bayes classifier.

In practice the population distribution is unknown, so the objective function (2.1) and the minimizer needs to be estimated through the data $(\boldsymbol{X}_i, y_i)_{i=1}^n$. Our first goal is to estimate the Bayes classifier for matrix classification.

**Question 1.** How to perform classification when matrix dimension far exceeds the sample size $n$?

2.2 *Level set estimation*: The problem of level set estimation generalizes the classification task. For a given $\pi \in (0, 1)$, the target $\pi$-level set of the conditional probability function ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ is defined as

$$S_{\text{bayes}}(\pi) = \{\boldsymbol{X} \in \mathcal{X} : p\mu(\boldsymbol{X}) \geq \pi\}.$$

An important fact is that the set $S_{\text{bayes}}(\pi)$ optimizes the weighted classification risk (**???**). Specifically, among all Borel subsets of $\mathcal{X}$, the set $S_{\text{bayes}}(\pi)$ is the global minimizer of the expected $\pi$-weighted classification error,

$$R_\pi(S) = \mathbb{E}\left[w_\pi(y)\mathbb{1}(y \neq \text{sign}(\boldsymbol{X} \in S))\right], \tag{3}$$

where $y \in \{-1, 1\}$ is distributed based on ~~$\mathbb{P}(y = 1|\boldsymbol{X}) = p(\boldsymbol{X})$~~ $\mathbb{P}(y = 1|\boldsymbol{X}) = \mu(\boldsymbol{X})$, and we define $w_\pi(y) = 1 - \pi$ or $\pi$ depending on $y = 1$ or $-1$; that is, $w_\pi(\cdot)$ assigns unequal weights to the two labels. In light of (2.1) and (2.2), the level set problem extends the usual classification from equal weight $\pi = 1/2$ to general weight $\pi \in (0, 1)$. We consider the following question:

**Question 2.** How to simultaneously estimate the level set and identify important features in the matrix-valued predictors, for the goal of interpretable prediction?

2.3 *Nonparametric binary regression*: ~~For a binary response, the~~The problem of nonparametric regression is to estimate the conditional ~~probability function $p(\boldsymbol{X}) = \mathbb{P}(Y = 1|\boldsymbol{X})$.~~ mean $\mathbb{E}(y|\boldsymbol{X})$ as a multivariable function in the predictor space. In our contexts, the nonparametric regression is equivalent to estimating the conditional probability $\mu(\boldsymbol{X}) = \mathbb{P}(y = 1|\boldsymbol{X}) = \frac{1}{2}(\mathbb{E}(y|\boldsymbol{X}) + 1)$. Throughout the paper we will focus on ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ and refer to it as the regression function. The function ~~$p(\boldsymbol{X})$~~$\mu(\boldsymbol{X})$, is the global minimizer, among all measurable functions $f : \mathcal{X} \to [0, 1]$, of the expected squared error,

$$R_{\text{reg}}(f) = \mathbb{E}\left[\frac{1}{2}(y + 1) - f1 - 2f(\boldsymbol{X})\right]^2, \tag{4}$$

where $R_{\text{reg}}(f)$ is also known as regression risk. Our final goal is the function estimation:

**Question 3.** How to learn the binary regression function ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ in the high-dimensional matrix space?

The three problems of our interest represent a range of learning tasks with increasing difficulties. Classification is a special case of level set estimation with $\pi = 1/2$, whereas the level set is a discrete

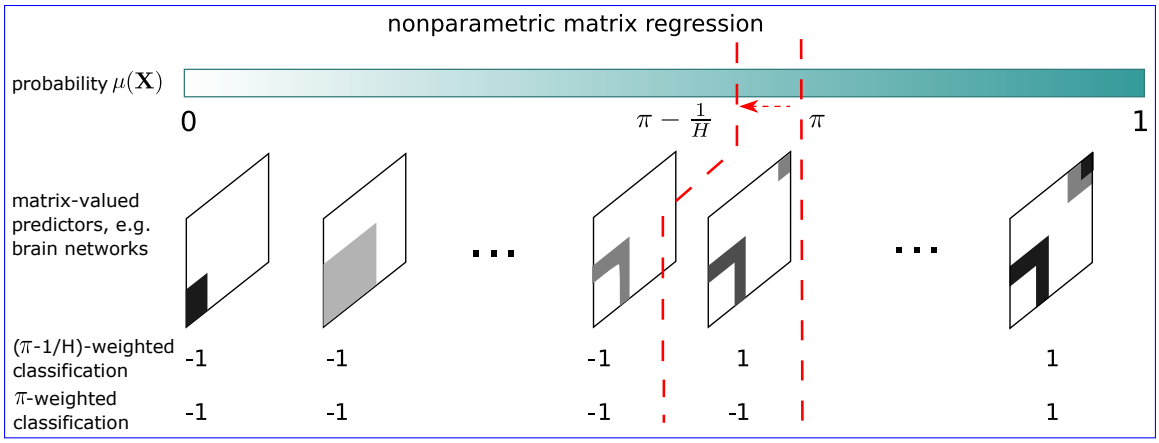



**Figure 2:** Level-set approaches to matrix binary nonparametric regression. We use weighted classification to find the level-set in the matrix space, and then estimate the target regression function via aggregation. ~~(Should the two slices in the middle be swapped?)~~

approximation of the regression function. A common approach is to address regression first, and then solve the earlier two using plug-in estimates (Figure 1a). This procedure, however, undermines the fact that regression is generally harder than the other two. Indeed, as we show in Section 5, regression has a slower convergence rate $\mathcal{O}(n^{-1/2})$ compared to the rate $\mathcal{O}(n^{-1})$ of classification. Ignorance of the increased complexity violates Vapnik's maxim: *When solving a given problem, one should try to avoid solving a more general problem as an intermediate step.*

## 3 From classification to regression: a new deal

We develop a "learning reduction" approach (Figure 1a) by relating the regression to classification. The latter problem is more fundamental and statistically easier to address. In general, regression requires more assumptions than classification. Our learning reduction approach bridges these two tasks using level set estimation, a problem lies somewhere in between. The connection allows us to disentangle complexity and leverage existing algorithms.

In this section we describe the oracle procedure of estimating regression function ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ when the true distribution $\mathbb{P}_{\boldsymbol{X},y}$ is known. This simplified situation leads to a cleaner characterization with deterministic risk functions in (2.1), (2.2), and (2.3). The finite sample estimation will be presented in Section 4, in which we address the general case with unknown distribution $\mathbb{P}_{\boldsymbol{X},y}$, and ~~the only information is through~~ all the information is based on empirical (stochastic) risk estimated from the training set $(\boldsymbol{X}_i, y_i)_{i=1}^n$.

### 3.1 Level set approaches to nonparametric binary regression

Figure 2 illustrates the main idea of our approaches. We use a ~~sequence~~ series of weighted classifications to find the level sets in the matrix space, and then estimate the regression function ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X}) = \frac{1}{2}(\mathbb{E}(y|\boldsymbol{X}) + 1)$ via level set aggregation. Our building block is to use level sets to estimate regression function $p(\cdot)$ through classifications. The level set approach bridges the two sides of a same coin – characteristic (set indicator) functions in functional analysis and weighted classifications in statistical learning.

Specifically, let $\mu(\cdot)\colon \mathcal{X} \to [0,1]$ be the target regression function of interest, and ~~$S_{\text{bayes}}(\pi) = \{\boldsymbol{X} \in \mathcal{X}\colon p(\boldsymbol{X}) \geq \pi\}$~~ $S_{\text{bayes}}(\pi) = \{\boldsymbol{X} \in \mathcal{X}\colon \mu(\boldsymbol{X}) \geq \pi\}$ be the associated $\pi$-level set. Let $\Pi = \{\frac{1}{H}, \frac{2}{H}, \ldots, \frac{H-1}{H}\}$ be a sequence of evenly spaced points in $[0,1]$, where $H \in \mathbb{N}_+$ is the resolution parameter. We introduce an $H$-step function $\bar{\mu}(\cdot)\colon \mathcal{X} \to [0,1]$ by

$$\bar{\mu}(\boldsymbol{X}) = \frac{1}{2H} \sum_{\pi \in \Pi} \text{sign}(\boldsymbol{X} \in \bar{S}(\pi)) + \frac{1}{2}, \quad \text{for all } \boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^{d_1 \times d_2}, \tag{5}$$

where, for every $\pi \in \Pi$, the set $\bar{S}(\pi) \subset \mathcal{X}$ is the classifier that minimizes the $\pi$-weighted classification risk,

$$\bar{S}(\pi) \stackrel{\text{def}}{=} \arg\min_{S \in \mathcal{S}} R_\pi(S), \quad \text{where} \quad R_\pi(S) = \mathbb{E}\left[w_\pi(y)\mathbb{1}(y \neq \text{sign}(\boldsymbol{X} \in S))\right], \tag{6}$$

subject to the constraint $S \in \mathcal{S}$, with $\mathcal{S}$ being a given candidate set of classifiers in $\mathcal{X}$. When the set $\mathcal{S}$ is rich enough, e.g., $\mathcal{S}$ consists of all Borel subsets of $\mathcal{X}$, then $\bar{S}(\pi)$ has the same risk as $S_{\text{bayes}}(\pi)$. We leave the $\mathcal{S}$ in general here; the specific choice of $\mathcal{S}$ will be described in Section 3.2.

In order to address the accuracy ~~between~~ of $\bar{\mu}(\boldsymbol{X})$ ~~and~~ from $\mu(\boldsymbol{X})$, we establish the recovery guarantee of level sets $S(\pi)$ from optimization (3.1). The Bayes classifier $S_{\text{bayes}}(\pi)$ minimizes the weighted classification risk $R_\pi(S)$; the inverse, however, may not be true because of possible multiple global minimizers of $R_\pi(S)$. The uniqueness and stability around $S_{\text{bayes}}(\pi)$ turns out to play a key role in the accurate estimation of ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$.

We introduce the following notion to characterize the behavior of the regression function near the level set boundaries ~~$\partial S_{\text{bayes}}(\pi) = \{p(\boldsymbol{X}) = \pi\}$~~ $\partial S_{\text{bayes}}(\pi) = \{\mu(\boldsymbol{X}) = \pi\}$. The condition essentially quantifies the uniqueness of level sets recovery from weighted classification.

Some additional notation is needed. We call a level $\pi \in [0,1]$ a mass point if the level set boundary $\partial S_{\text{bayes}}(\pi)$ has non-zero measures under $\mathbb{P}_{\boldsymbol{X}}$. Let ~~$\mathcal{N} = \{\pi \in [0,1]\colon \mathbb{P}_{\boldsymbol{X}}[p(\boldsymbol{X}) = \pi] \neq 0\}$~~ $\mathcal{N} = \{\pi \in [0,1]\colon \mathbb{P}_{\boldsymbol{X}}[\mu(\boldsymbol{X}) = \pi] \neq 0\}$ denote the collection of mass points in ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$. Assume $|\mathcal{N}| \leq c < \infty$ for some constant $c > 0$.

**Definition 1** ($\alpha$-regularity)**.** Fix $\pi \notin \mathcal{N}$. A function ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ is called $(\alpha, \pi)$-locally regular, if there exist constants $C = C(\pi) > 0, \alpha = \alpha(\pi) \in [0,1]$, such that,

$$\sup_{0 \leq t < \rho(\pi, \mathcal{N})} \frac{\mathbb{P}_{\boldsymbol{X}}(|p(\boldsymbol{X}) - \pi| \leq t)}{t^{\alpha/(1-\alpha)}} \sup_{0 \leq t < \rho(\pi, \mathcal{N})} \frac{\mathbb{P}_{\boldsymbol{X}}(|\mu(\boldsymbol{X}) - \pi| \leq t)}{t^{\alpha/(1-\alpha)}} \leq C, \tag{7}$$

where $\rho(\pi, \mathcal{N}) \stackrel{\text{def}}{=} \min_{\pi' \in \mathcal{N}} |\pi - \pi'|$ denotes the distance from $\pi$ to the nearest point in $\mathcal{N}$. When $\mathcal{N} = \phi$, we define $\rho(\pi, \mathcal{N}) = 1$. The largest possible $\alpha = \alpha(\pi)$ is called the smoothness index at level $\pi$. In particular $\alpha = \infty$ means the numerator in (2) is zero.

If $\mu(\boldsymbol{X})$ is $(\alpha, \pi)$-locally regular for all $\pi \in [0, 1]$ except for a finite number of points, and $C = \max_\pi C(\pi) < \infty$, then we call $\mu(\boldsymbol{X})$ is $\alpha$-globally regular.

The global regularity controls the uniform behavior of $\mu(\boldsymbol{X})$ over all possible $\pi$'s (except for a finite number of points). The exponent $\alpha$ quantifies the concentration of probability mass $\mu(\boldsymbol{X})$ around level set boundaries. We show that the global regularity implies the identifiability of $S_{\text{bayes}}(\pi)$ from optimization (3.1) for almost all $\pi$'s. For two sets $S_1, S_2 \in \mathbb{R}^{d_1 \times d_2}$, we define the probabilistic set difference

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

and the excess classification risk

$$d_\pi(S_1, S_2) \stackrel{\text{def}}{=} R_\pi(S_1) - R_\pi(S_2).$$

**Proposition 1** (Identifiability)**.** Suppose the regression function $\mu(\boldsymbol{X})$ is $\alpha$-globally regular with $\alpha \in [0, 1]$. Then, ~~there exists a constant $c > 0$ such that~~

$$d_\Delta(S, S_{\text{bayes}}(\pi)) \lesssim d_\pi^\alpha(S, S_{\text{bayes}}(\pi)) + \frac{1}{\rho(\pi, \mathcal{N})} d_\pi(S, S_{\text{bayes}}(\pi)), \tag{8}$$

for all sets $S \in \mathbb{R}^{d_1 \times d_2}$ and all levels $\pi \in [0, 1]$ except for a finite number of points.

The bound (1) controls the worst-case perturbation of classifiers in the probability space $\mathbb{P}_{\boldsymbol{X}}$ with respect to weighted classification risks. When $\alpha \neq 0$, the inequality (1) immediately implies the uniqueness of $S_{\text{bayes}}(\pi)$ up to a measure-zero set in $\mathbb{P}_{\boldsymbol{X}}$, whereas $\alpha = 0$ corresponds to no identifiability. Notice that $\rho(\pi, \mathcal{N})$ becomes a constant for fixed $\pi$, implying that local regularity ensures the identifiability of Bayes set $S_{\text{bayes}}(\pi)$.

Our identifiability characterization generalizes the earlier results for single level set estimation (**??**). Inspection of the proof shows that the set estimation is more difficult at levels where the point mass concentrates (small $\alpha$), as intuition would suggest. Consider a simple case when the entries of matrix $\boldsymbol{X}$ are i.i.d. drawn from Uniform$[-1, 1]$; that is, $\mathbb{P}_{\boldsymbol{X}}$ is the Lebesque measure on $[0, 1]^{d_1 d_2}$. Then, the best rate $\alpha \to 1$ corresponds to a clear separation with no point mass at the boundary, whereas the worst rate $\alpha \to 0$ corresponds to a heavy mass of ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ near the boundary. A typical intermediate case is $\alpha = 1/2$ when ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ is a locally bilipschitz function (see Proposition 3 in Appendix).

Now we reach the main result of this section by putting together the proposal (3.1), (3.1) and Proposition 1. Define the regression excess risk by ~~$R_{\text{reg}}(\bar{p}) - R_{\text{reg}}(p)$~~ $R_{\text{reg}}(\bar{\mu}) - R_{\text{reg}}(\mu)$. The following result bounds the regression excess risk using classification excess risk.

**Theorem 1** (Nonparametric regression via weighted classifications)**.** *Suppose that the regression function ~~$p(\boldsymbol{X})$~~ $\mu(\boldsymbol{X})$ is $\alpha$-globally regular with $\alpha \in [0,1]$. Let ~~$\bar{p}(\boldsymbol{X})$~~ $\bar{\mu}(\boldsymbol{X})$ the step function (3.1) constructed from weighted classifiers. Then, the regression excess risk is ~~bounded~~ bounded by the classification excess risk; i.e.,*

$$R_{\text{reg}}(\bar{\mu}) - R_{\text{reg}}(\mu) \leq 4\mathbb{E}_{\boldsymbol{X}} \left| p\mu(\boldsymbol{X}) - \bar{p}\mu(\boldsymbol{X}) \right| \leq \lesssim \frac{21}{H} + \frac{4}{H}\frac{1}{H} \sum_{\pi \in \Pi} \sum_{\pi \in \Pi \setminus \mathcal{N}} \left\{ d_\pi^\alpha(S_{\text{bayes}}(\pi), \bar{S}(\pi)) + \frac{1}{\rho(\pi, \mathcal{N})} d_\pi(S_{\text{bayes}}(\pi), \bar{S}(\pi)) \right\}, \tag{9}$$

*for all resolution parameter $H = |\Pi| \in \mathbb{N}_+$.*

Theorem 1 shows the key role of $\bar{\mu}(\boldsymbol{X})$ in bridging regression and classification. The results suggest that the estimation of $\mu(\boldsymbol{X})$ can be reduced to estimation of $\bar{\mu}(\boldsymbol{X})$, or equivalently, to a ~~sequence~~ series of weighted classifications $(\bar{S}(\pi))_{\pi \in \Pi}$. The regression excess risk bound (1) has two terms. The first term is the bias due to the step function approximation to the regression function. The second term is the classification excess risk of recovering $S_{\text{bayes}}(\pi)$ from optimization (3.1). In the case of unknown population distribution $\mathbb{P}_{\boldsymbol{X},y}$, the second term should be plugged in using the empirical risk, which results in a variance-bias error due to finite sample size (see Section 4).

Estimating $\bar{\mu}(\boldsymbol{X})$ as a surrogate of $\mu(\boldsymbol{X})$ provides several benefits. From a computational perspective, $\bar{\mu}(\boldsymbol{X})$ is a finite combination of weighted classifiers, which are easier to solve than a direct regression. From the statistical perspective, the function $\bar{\mu}(\boldsymbol{X})$ provides a valid approximation to $\mu(\boldsymbol{X})$ even when $\mu(\boldsymbol{X})$ is non-regular and irregular. In particular, the estimation accuracy relies little on the local neighborhood of $\boldsymbol{X}$ but rather on the local neighborhood of $\mu(\boldsymbol{X})$. Note that the former is a $d_1 d_2$-dimensional random variable whereas the later is a $[0,1]$-valued univariate random variable. The shifted focus of local structure to the range space is especially appealing for matrix-valued predictors, since the predictor space is high dimensional and often barely explored by small sample size data.

### 3.2 Sparse and low-rank function boundaries

We describe the choice of candidate sets $\mathcal{S}$ in (3.1). A desirable $\mathcal{S}$ should balance the prediction and interpretability; i.e., $\mathcal{S}$ should be flexible enough for accurate prediction while being simple enough for high interpretability. We propose to optimize (3.1) over the family of linear halfspaces; i.e.,

$$\bar{S}(\pi) = \{\boldsymbol{X} \colon \bar{f}(\boldsymbol{X}) \geq 0\}, \quad \text{with} \quad \bar{f}(\boldsymbol{X}) = \underset{f(r,s_1,s_2) \in \mathcal{F}}{\arg\min} \, \mathbb{E}\left[ w_\pi(y) \mathbb{1}(y \neq \operatorname{sign} f(\boldsymbol{X})) \right], \tag{10}$$

where the function family $\mathcal{F}(r, s_1, s_2)$ consists of linear classifiers with low-rank two-way sparse matrix coefficients,

$$\mathcal{F}(r, s_1, s_2) = \{f\colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle + b \;\big|\; \mathrm{rank}(\boldsymbol{B}) \le r,\ \mathrm{supp}(\boldsymbol{B}) \le (s_1, s_2),\ \boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2},\ b \in \mathbb{R}\} \quad (11)$$

Here $\mathrm{rank}(\boldsymbol{B})$ denotes the rank of the coefficient matrix, and $\mathrm{supp}(\boldsymbol{B})$ denotes the two-way sparsity parameter with $s_1 = \|\boldsymbol{B}\|_{1,0}$ and $s_2 = \|\boldsymbol{B}^T\|_{1,0}$ being the numbers of non-zero rows and columns of $\boldsymbol{B}$, respectively. For the theory, we assume that $(r, s_1, s_2)$ are known; the adaptation to unknown $(r, s_1, s_2)$ in practice is described in Section 6. Combining formulations (3.1), (3.2) and (3.2) yields our (oracle procedure of ) "learning ~~deduction~~reduction" approach to nonparametric matrix regression.

The low-rank two-way sparse classifier (3.2) enables efficient variable selection in high-dimensional matrices, thereby achieving high interpretability in prediction. In the brain network analysis, scientists are interested in identifying important nodes attached to at least one active edges with non-zero effects. Classical entrywise sparsity essentially treats $\boldsymbol{X}$ as "a bag of non-ordered edges", and loses the two-way paring information among entries. In contrast, our two-way sparsity efficiently identifies the underlying active nodes by making use of matrix structure in the predictors.

It is worthy noting that the linearity in the classifiers $\mathcal{F}$ does not preclude the global nonlinearity in the regression function $\mu(\boldsymbol{X})$ or its variant $\bar{\mu}(\boldsymbol{X})$. As shown in the following examples, many nonlinear regression functions in existing literature are special cases of our representation (3.2) with (3.2), in the sense that their level sets are precisely halfspaces.

**Example 1** (Monotonic single index models (**??**))**.** Suppose the true regression function can be expressed as $\mu(\boldsymbol{X}) = g(\langle \boldsymbol{X}, \boldsymbol{B}\rangle)$, where $g(\cdot)\colon \mathbb{R} \to [0, 1]$ is an arbitrary monotonic link function, and $\boldsymbol{B}$ is a low-rank two-way sparse matrix. Then, for every $\pi \in (0, 1)$, there exists $f_\pi \in \mathcal{F}(r, s_1, s_2)$, such that $\mathrm{sign}(\mu(\boldsymbol{X}) - \pi) = \mathrm{sign} f_\pi(\boldsymbol{X})$. Our method generalizes nonlinear single index model to high dimensional matrices by joint learning matrix coefficient $\boldsymbol{B}$ and nonlinear function $g$.

Single index models with (unknown) monotonic links, such as logistic function $g(z) = (1+\exp(-z))^{-1}$, arctangent function $g(z) = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$, truncated rectified linear unit (ReLU) function $g(z) = z\mathbb{1}(z \in [0, 1]) + \mathbb{1}(z > 1)$, and any arbitrary inverse cumulative distribution functions (CDF) are included in our classifier functions. In particular, our model extends parametric matrix regression with known link functions (**???**) to the nonparametric regression with unknown link functions.

There are usually two approaches for analyzing association between matrix data and binary labels. One approach is to analyze the conditional distribution $y|\boldsymbol{X}$ from the prospective model. The other approach is to analyze the conditional distribution $\boldsymbol{X}|y$ from the retrospective model. Although our proposal is built on the former, our classifier function also incorporates the retrospective models from matrix linear discriminant analysis (LDA) (**?**).

**Example 2** (Multivariate normal mixtures)**.** Suppose the matrix-valued predictor $\boldsymbol{X}$ follows a

Gaussian mixture distribution, $\boldsymbol{X}|\{y=-1\} = \boldsymbol{B}_1 + \boldsymbol{E}_1$ and $\boldsymbol{X}|\{y=1\} = \boldsymbol{B}_2 + \boldsymbol{E}_2$, where $(\boldsymbol{B}_1 - \boldsymbol{B}_2)$ is a low-rank two-way sparse matrix, and $\boldsymbol{E}_1, \boldsymbol{E}_2$ are two mutually independent noise matrices with i.i.d. $N(0,1)$ entries. Then, for every $\pi \in (0,1)$, $\text{sign}(\mu(\boldsymbol{X}) - \pi) = \text{sign} f_\pi(\boldsymbol{X})$ for some $f_\pi \in \mathcal{F}(r, s_1, s_2)$.

More generally, we have established the characterization by extending two classes of $\boldsymbol{X}$ to a series of $\boldsymbol{X} = \boldsymbol{X}(\pi)$ over a continuous spectrum of $\pi \in (0,1)$ (see Appendix A). Our level-set approach essentially learns the right "sorting" of $\boldsymbol{X}(\pi)$ against the index $\pi \in (0,1)$ (see Figure 2), thereby facilitating the estimation of relationship $\pi = \pi(\boldsymbol{X})$.

In principle, more complicated classifiers, such as neural network, decision trees, and boosting, can also be brought to bear on the level set construction (3.2). The ability to import and adapt existing classification methods is one advantage of the proposed learning reduction framework. We find that, in our motivating brain network analysis, the low-rank two-way sparse classifiers (3.2) are able to provide the benefit of interpretable predictions with high accuracy (see Section 6.3).

# 4 Estimation

In previous sections we have established the oracle procedure from classification to regression (Figure 2). In this section we address the empirical learning problems when the true distribution $\mathbb{P}_{\boldsymbol{X},y}$ is unknown. The objective function in the earlier optimization now becomes empirical (stochastic) risks calculated from high dimensional training data $(\boldsymbol{X}_i, y_i)_{i=1}^n$.

## 4.1 Large-margin learning with high dimensional matrices

When the distribution $\mathbb{P}_{\boldsymbol{X},y}$ is unknown, we propose the regression function estimate $\hat{\mu}(\cdot)\colon \mathcal{X} \to [0,1]$ by

$$\hat{\mu}(\boldsymbol{X}) = \frac{1}{2H} \sum_{\pi \in \Pi} \text{sign}(\boldsymbol{X} \in \hat{S}(\pi)) + \frac{1}{2}, \quad \text{for all } \boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^{d_1 \times d_2}. \tag{12}$$

Here, for every $\pi \in \Pi$, the set $\hat{S}(\pi) \subset \mathcal{X}$ is the estimated classifier from empirical surrogate risk minimization,

$$\hat{S}(\pi) = \{\boldsymbol{X}\colon \hat{f}_\pi(\boldsymbol{X}) \geq 0\}, \quad \text{with} \quad \hat{f}_\pi = \min_{f \in \mathcal{F}(r,s_1,s_2)} \left\{ \frac{1}{n} \sum_{i=1}^n w_\pi(y_i) \ell\left(y_i f(\boldsymbol{X}_i)\right) + \lambda \|f\|_F^2 \right\}, \tag{13}$$

where $w_\pi(y) = 1 - \pi$ if $y = 1$ and $w_\pi(y) = \pi$ if $y = -1$ is the label-dependent weight; $\ell(z)\colon \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is the surrogate classification loss defined as a function of margin $z = yf(\boldsymbol{X})$; $\lambda > 0$ is the penalty parameter; and we define the penalization term $\|f\|_F = \|\boldsymbol{B}\|_F$, with $\boldsymbol{B}$ being the coefficient matrix associated with $f \in \mathcal{F}(r, s_1, s_2)$. Examples of large-margin loss functions are hinge loss $\ell(z) = (1-z)_+$ for support vector machines, logistic loss $\ell(z) = \log(1 + e^{-z})$ for important vector

machines, and $\psi$-loss $\ell(z) = 2\min(1, (1-z)_+)$ with $z_+ = \max(z, 0)$. Our algorithm implements the hinge loss for illustration, although our framework and theory are applicable to general large-margin losses (**?**).

The estimation (4.1) generalizes the oracle formulation (3.2) from three aspects. First, the population expectation in (3.2) is replaced by the empirical sample average, which is common in statistical learning problems with i.i.d. assumption. Second, we add the ridge penalization $\lambda\|f\|_F^2$ to control the magnitude of the classifiers. The tuning parameter $\lambda$ depends on the sample size and the problem dimension as we will describe in Section 5. The resulting sieve estimate enjoys numerical stability and statistical accuracy. In practice, we choose $\lambda$ in a data-adaptive fashion via cross validation. Third, we replace the binary loss in (3.2) by a more manageable large-margin loss. This relaxation allows us to leverage efficient large-margin algorithms while maintaining desirable statistical performance under mild assumptions.

### 4.2 Alternating optimization for structural risk minimization

We describe the optimization algorithm for solving matrix classification and regression. We focus on the general $\pi$-weighted classification (4.1) because both classification and regression naturally follow by setting $\pi = \frac{1}{2}$, and $\pi \in \{\frac{1}{H}, \ldots, \frac{1-H}{H}\}$, respectively. For brevity, we present the algorithm assuming zero intercept in the classifier functions (3.2), and use $\mathcal{F}(r, s_1, s_2)$ to denote the set of matrices satisfying $\text{rank}(\boldsymbol{B}) \le r$ and $\text{supp}(\boldsymbol{B}) \le (s_1, s_2)$. The estimation problem (4.1) is formulated as an optimization over matrices,

$$\min_{\boldsymbol{B} \in \mathcal{F}(r,s_1,s_2)} L(\boldsymbol{B}), \quad \text{where } L(\boldsymbol{B}) = \frac{1}{n}\sum_{i=1}^{n} w_\pi(y_i)\ell(y_i\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle) + \lambda\|\boldsymbol{B}\|_F^2, \tag{14}$$

where the objective function can be either convex (such as hinge loss, logistic loss) or non-convex ($\psi$-loss). The optimization (4.2) has a non-convex feasible region because of the low-rank and sparse constraint.

We propose an alternating direction method of multipliers (ADMM) approach to solve problem of this type. ADMM introduces a dual variable and an additional feasibility constraint to perform coordinate descent in the corresponding augmented Lagrangian function. The augmented ADMM objective in our context is given by

$$L(\boldsymbol{B}, \boldsymbol{S}, \boldsymbol{\Lambda}, \rho) = \frac{1}{n}\sum_{i=1}^{n} w_\pi(y_i)\ell(y_i\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle) + \lambda\|\boldsymbol{B}\|_F^2 + \rho\|\boldsymbol{B} - \boldsymbol{S}\|_F^2 + \langle \boldsymbol{\Lambda}, \boldsymbol{B} - \boldsymbol{S}\rangle, \tag{15}$$

and $\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}$ is the unconstrained primal variable, $\boldsymbol{S} \in \mathcal{F}(r, s_1, s_2)$ is the constrained dual variable, $\boldsymbol{\Lambda} \in \mathbb{R}^{d_1 \times d_2}$ is the Lagrangian multiplier, and $\rho > 0$ is the step-size parameter. Note that formulation (4.2) has moved the non-convexity from the first two terms in $\boldsymbol{B}$ to the last two simpler terms in $\boldsymbol{S}$. This separability of ADMM makes the optimization efficient for a wide range of loss

functions and constraints.

We optimize the ADMM objective (4.2) via coordinate descent, by iteratively update one variable at a time while holding others fixed. Each update in the ADMM reduces to a simpler problem and can be efficiently solved by standard algorithms. Specifically, given variables $(\boldsymbol{S}, \boldsymbol{\Lambda}, \rho)$ and $\bar{\boldsymbol{S}} \stackrel{\text{def}}{=} \frac{1}{2(\rho+\lambda)}(2\rho\boldsymbol{S} - \boldsymbol{\Lambda})$, the objective with respect to $\boldsymbol{B}$ is

$$L(\boldsymbol{B}|\boldsymbol{S}, \boldsymbol{\Lambda}, \rho) = \frac{1}{n}\sum_{i=1}^{n} w_\pi(y_i)\ell(y_i\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle) + (\lambda+\rho)\|\boldsymbol{B} - \bar{\boldsymbol{S}}\|_F^2. \tag{16}$$

This unconstrained optimization is a usual vector-based classification with ridge penalty and an offset $\bar{\boldsymbol{S}}$. Therefore, various loss functions and fast software can be adopted into (4.2) such as weighted SVM, logistic, and $\psi$-learning. Similarly, given $(\boldsymbol{B}, \boldsymbol{\Lambda}, \rho)$ and $\bar{\boldsymbol{B}} \stackrel{\text{def}}{=} \frac{1}{2\rho}(2\rho\boldsymbol{B} + \boldsymbol{\Lambda})$, the objective with respect to $\boldsymbol{S}$ is

$$L(\boldsymbol{S}|\boldsymbol{B}, \boldsymbol{\Lambda}, \rho) = \|\boldsymbol{S} - \bar{\boldsymbol{B}}\|_F^2, \quad \text{subject to } \boldsymbol{S} \in \mathcal{F}(r, s_1, s_2). \tag{17}$$

This formulation is equivalent to the best sparse and low rank approximation, in the least-square sense, to the matrix $\boldsymbol{B}$. Compared to the original objective (4.2), the F-norm based objective makes the optimization easier to handle. A number of algorithms have been designated to approximately solve this problem, including sparse PCA, sparse SVD, and projection pursuit. We use the recently-developed double projection algorithm for (4.2) which has provably better performance than convex alternatives in high dimensional regimes (**?**). Finally, the Lagrangian multiplier $\boldsymbol{\Lambda}$ is updated by standard scheme $\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + 2\rho(\boldsymbol{B} - \boldsymbol{S})$. These steps are performed until the algorithm convergence within tolerance. The value $\rho$ controls the closeness between dual and primal variables. We initialize $\rho$ from 0.1 and increases its value geometrically throughout iterations. In practice, we observed this scheme gives a good balance between the variable feasibility and convergence speed, although other self-tuning methods are also possible (**?**). Algorithm 1 gives the full description for the $\pi$-classification.

**Algorithm 1: Matrix classification and level-set estimation via ADMM**

**Input:** Data $\{(\boldsymbol{X}_i, y_i) \in \mathbb{R}^{d_1\times d_2} \times \{-1, 1\} : i \in [n]\}$, rank $r$, support $(s_1, s_2)$, ridge parameter $\lambda$, a target level $\pi \in (0, 1)$.

**Output:** Estimated $\pi$-level set $\hat{S}(\pi) = \{\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2} : \hat{f}(\boldsymbol{X}) \geq 0\}$.

**Initialize:** primal variable $\boldsymbol{B}$, dual variable $\boldsymbol{S}$, Lagrangian multiplier $\boldsymbol{\Lambda} = \boldsymbol{0}$, step size $\rho = 0.1$.

**Do until converges**

**Update** $\boldsymbol{B}$ fixing $(\boldsymbol{S}, \boldsymbol{\Lambda}, \rho)$: $\boldsymbol{B} \leftarrow \arg\min_{\boldsymbol{B}} L(\boldsymbol{B}|\boldsymbol{S}, \boldsymbol{\Lambda}, \rho)$, where $L(\boldsymbol{B}|\boldsymbol{S}, \boldsymbol{\Lambda}, \rho) = \frac{1}{n}\sum_{i=1}^{n} w_\pi(y_i)\ell(y_i\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle) + (\lambda+\rho)\|\boldsymbol{B} - \bar{\boldsymbol{S}}\|_F^2$ and $\bar{\boldsymbol{S}} \stackrel{\text{def}}{=} \frac{1}{2(\rho+\lambda)}(2\rho\boldsymbol{S} - \boldsymbol{\Lambda})$.

**Update** $\boldsymbol{S}$ fixing $(\boldsymbol{B}, \boldsymbol{\Lambda}, \rho)$: $\boldsymbol{S} \leftarrow \arg\min \|\boldsymbol{S} - \bar{\boldsymbol{B}}\|_F^2$ subject to $\boldsymbol{S} \in \mathcal{F}(r, s_1, s_2)$, where $\bar{\boldsymbol{B}} \stackrel{\text{def}}{=} \frac{1}{2\rho}(2\rho\boldsymbol{B} + \boldsymbol{\Lambda})$.

**Update** $\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + 2\rho(\boldsymbol{B} - \boldsymbol{S})$.

**Update** $\rho \leftarrow 1.1\rho$.

We develop the nonparametric matrix regression in a similar way with little modification. The nonparametric regression (4.1) estimates $\pi$-classifiers at a ~~sequence~~ series of weights $\pi \in \{\frac{1}{H}, \ldots, \frac{H-1}{H}\}$. In principal, one can optimize all classifiers jointly subject to ~~nestedness~~ nested constraints among sequential level sets. However, this strategy would lead to increased computational burden, and the gain in accuracy is often little with moderate sample size. We choose to use parallel processing to obtain $\pi$-classifiers separately to speed up the computation. The procedure is summarized in Algorithm 2. The software for both matrix classification and regression will be available at CRAN.

**Algorithm 2: Nonparamatrix matrix regression**

**Input:** Data $\{(\boldsymbol{X}_i, y_i) \in \mathbb{R}^{d_1 \times d_2} \times \{-1, 1\} \colon i \in [n]\}$, rank $r$, support $(s_1, s_2)$, ridge parameter $\lambda$, resolution parameter $H$.

**Output:** Level sets $\hat{S}(\pi)$ for $\pi \in \{\frac{1}{H}, \ldots, \frac{H-1}{H}\}$ and regression function $\hat{\mu}(\boldsymbol{X})$.

**for** $\pi = \frac{1}{H}$ **to** $\frac{H-1}{H}$ **do**

Obtain estimated $\pi$-level set $\hat{S}(\pi)$ by performing weighted classification using Algorithm 1.

$\hat{S}(\pi) \leftarrow \{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2} \colon \hat{f}_\pi(\boldsymbol{X}) \geq 0\}$.

**end for**

Estimated regression function $\hat{\mu}(\boldsymbol{X}) \leftarrow \frac{1}{2H} \sum_{\pi \in \Pi} \operatorname{sign}(\boldsymbol{X} \in \hat{S}(\pi)) + \frac{1}{2}$.

## 5 Statistical learning theory

In this section we establish excess risk bounds for the global optimal estimates in the weighted classification (4.1) and matrix regression (4.1). Our learning reduction approach successfully bridges these two tasks based on Vapnik's maxim and achieves theoretical guarantee for both problems.

We introduce some additional notation to establish the accuracy for weighted classification (4.1). Let $f_{\text{bayes},\pi}(\boldsymbol{X}) = \operatorname{sign}(\mu(\boldsymbol{X}) - \pi)$ be the Bayes classifier corresponding to the $\pi$-level set. Let $R_\pi(f) = \mathbb{E}\left[w_\pi(y)\mathbb{1}\{y \neq \operatorname{sign} f(\boldsymbol{X})\}\right]$ denote the weighted classification risk, and $R_{\ell,\pi}(f) = \mathbb{E}\left[w_\pi(y)\ell(yf(\boldsymbol{X}))\right]$ denote the surrogated weighted classification risk as the counterpart of the empirical surrogate risk in (4.1). We use the following important facts about hinge loss and $\psi$-loss. For all $\pi \in [0, 1]$, the excess surrogate risk satisfies Fisher consistency (**?**),

$$d_\pi(S_f, S_{\text{bayes}}(\pi)) \stackrel{\text{def}}{=} R_\pi(f) - R_\pi(f_{\text{bayes},\pi}) \leq R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes},\pi}), \tag{18}$$

for all measurable functions $f$ where $S_f = \{\boldsymbol{X} \in \mathcal{X} \colon f(\boldsymbol{X}) \geq 0\}$. In other words, the convergence in surrogate risk $R_{\ell,\pi}$ implies convergence in classification risk $R_\pi$. (For aforementioned other surrogate losses, the excess risk bound (5) holds up to a constant term and an additional exponent in $(0, 1)$ on the right hand side (**?**)). The bound (5) also implies that, among all measurable functions $f \colon \mathcal{X} \to \mathbb{R}$, the level set function $f_{\text{bayes},\pi}$ minimizes the surrogate loss $R_{\ell,\pi}(f)$. Conversely, this minimizer is essentially unique as shown in Proposition 1.

The following assumption quantifies the representation capability of candidate classifiers $\mathcal{F}(r, s_1, s_2)$. For simplicity of notation, we assume $d_1 = d_2 = d$ and $\|\boldsymbol{X}\|_F \leq 1$ with probability 1.

**Assumption 1** (Approximation error)**.** Assume there exists a sequence of functions $f_n^* \in \mathcal{F}(r, s_1, s_2)$ for which the surrogate excess risk vanishes; i.e., $R_{\ell,\pi}(f_n^*) - R_{\ell,\pi}(f_{\text{bayes},\pi}) \leq a_n$ for some sequence $a_n \to 0$ as $n, d \to \infty$. Let $J_n = \|f_n^*\|_F^2$, and we allow $J_n$ to grow with $n$.

**Theorem 2** (Accuracy for weighted matrix classification)**.** *Fix a level $\pi \in (0,1)$. Consider the problem of $\pi$-level set estimation for a regression function $\mu(\boldsymbol{X})$. Suppose that the function $\mu(\boldsymbol{X})$ is $(\pi, \alpha)$-locally regular with $\alpha \in [0,1]$, and that Assumption 1 holds. Let $\hat{S}(\pi)$ be the level set estimate in (4.1) with penalty parameter $\lambda \asymp \left(\frac{r(s_1+s_2)\log d}{nJ_n}\right)^{1/(2-\alpha)}$ Then, with high probability over training set, the classification excess risk is bounded by*

$$d_\pi(\hat{S}(\pi), S_{\text{bayes}}(\pi)) \lesssim \underbrace{C(\pi)\left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)}}_{\text{statistical error}} + \underbrace{a_n}_{\text{approximation error}} \tag{19}$$

*where $C(\pi) > 0$ is a constant that depends on $\pi$ and $\mu(\boldsymbol{X})$. Notice that the usual classification corresponds to $\pi = 1/2$.*

Theorem 2 reveals the weak dependence on matrix dimension of our estimates. Consider the case when the statistical error (first term) dominates the approximation error (second term). Then, the bound (2) immediately implies the classification consistency in the high dimensional regime $d, n \to \infty$, as long as the matrix dimension $d$ grows sub-exponentially in sample size $n$; i.e., $d = o(e^n)$. This sample complexity shows the advantage of proposed low-rank two-way sparse structural models. Furthermore, we find that classification (2) reaches a fast rate $1/n$ when $\alpha = 1$, and in general the risk has rate no slower than $1/\sqrt{n}$. This observation extends the asymptotic for usual vector-based classification (**???**).

Based on the identifiability Proposition 1, our estimate $\hat{S}(\pi)$ in (4.1) accurately recovers the level set in the high-dimensional matrix space. The following corollary follows the same proof as in Theorem 2, except that we provide the explicit form of $C(\pi)$ in preparation for Theorem 3.

**Corollary 1** (Accuracy for level set estimation)**.** Assume the same setup in Theorem 2. Denote $t_n = \frac{r(s_1+s_2)\log d}{n}$ for $n \in \mathbb{N}_+$. Then, with high probability over training set, the probabilistic set different between $\hat{S}(\pi)$ and $S_{\text{bayes}}(\pi)$ is bounded by

$$d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) \lesssim t_n^{\alpha/(2-\alpha)} + \frac{1}{\underline{\rho}^2(\pi, \mathcal{N})} t_n + a_n^\alpha, \quad \text{for all levels } \pi \notin \mathcal{N}, \tag{20}$$

where the probability measure involved in $d_\Delta(\cdot,\cdot)$ is taken respect to a prototypical test point $\boldsymbol{X}$ i.i.d. from the training set $(\boldsymbol{X}_i)_{i=1}^n$. For a fixed $\pi$, the first two terms in the right hand side of (1) can be combined as $C(\pi)t_n^{\alpha/(2-\alpha)}$.

We reach the main theorem in this section for our nonparametric matrix regression.

**Theorem 3** (Accuracy for nonparametric matrix regression)**.** *Suppose the regression function $\mu(\boldsymbol{X})$*

*is $\alpha$-globally regular with $\alpha \in [0,1]$. Consider the same setup as in Theorem 2. Furthermore, assume Assumption 1 holds for all $\pi \in \Pi \setminus \mathcal{N}$. Then with high probability, the estimate (4.1) is bounded by*

$$\mathbb{E}|\hat{\mu}(\boldsymbol{X}) - \mu(\boldsymbol{X})| \lesssim \underbrace{\left(\frac{r(s_1+s_2)\log d}{n}\right)^{\frac{\alpha}{2-\alpha}} + a_n^{\alpha}}_{\text{estimation error inherited from classification}} + \underbrace{\frac{1}{H}}_{\text{reduction bias}} + \underbrace{H\left(\frac{r(s_1+s_2)\log d}{n}\right)}_{\text{reduction variance}},$$

Theorem 3 demonstrates the high dimensional convergence of our nonparametric matrix regression. Our results reveal three sources of errors: the estimation errors (including statistical error and approximation error) inherited from classification, the bias, and the variance due to reduction from classification to regression. The resolution parameter $H$ controls the bias-variance tradeoff.

**Corollary 2** (High-dimensional consistency)**.** Consider the same set-up as in Theorem 3. Assume $a_n \lesssim \left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)}$ and set $H \asymp \left(\frac{n}{r(s_1+s_2)\log d}\right)^{1/2}$. Then, with high probability,

$$\mathbb{E}|\hat{\mu}(\boldsymbol{X}) - \mu(\boldsymbol{X})| \lesssim \left(\frac{r(s_1+s_2)\log d}{n}\right)^{\min\{\alpha/(2-\alpha),\ 1/2\}} \quad \text{as } d, n \to \infty \text{ while } d = o(e^n). \tag{21}$$

We apply the convergence rate in Theorem 3 to two specific examples.

**Example 3** (Piece-wise constant model)**.** Consider piece-wise constant probability function $\mu(\boldsymbol{X}) = \sum_{t=1}^{T} c_t \mathbb{1}(\langle \boldsymbol{X}, \boldsymbol{B}_t \rangle = 0)$ with nonequal $c_1 < c_2 < \cdots < c_T$. In particular, ~~$T = 1, \boldsymbol{B}_1 = \boldsymbol{0}$ reduces~~ the constant model $\mu(\boldsymbol{X}) \equiv c$ is a special case when $T = 1, \boldsymbol{B}_1 = 0$. We have $\alpha = 1$ in both cases. Theorem 3 gives an convergence rate $\mathcal{O}(n^{-1/2})$ by setting $H \asymp n^{-1/2}$. This rate achieves minimax optimality as in parametric models.

**Example 4** (Monotonic single index model)**.** Consider the parametric model $\mu(\boldsymbol{X}) = g(\langle \boldsymbol{X}, \boldsymbol{B} \rangle)$ as in Example 1. For common link functions $g$ (such as $g(z) = z, (1+\exp(-z))^{-1}, \frac{1}{\pi}\arctan(z) + \frac{1}{2}$, etc) that have non-degenerate first-order Taylor expansion, we have $\alpha = 1/2$ (see Proposition 3 in Appendix). Choosing ~~$H \asymp n^{1/2}$~~ $H \asymp n^{1/3}$ yields the convergence rate $\mathcal{O}(n^{-1/3})$, which is consistent to the most recent results (**?**).

We conclude this section by comparing regression and classification. The regression bound (2) reaches the fastest rate $1/\sqrt{n}$ when $\alpha = 1$. This error rate is generally slower than the corresponding classification rate in (2). The fact confirms our earlier premise that classification is an easier problem than regression. Our level set approach successfully bridges theses two tasks and achieves theoretical guarantee for both problems. The connection allows us to disentangle complexity and leverage existing algorithms. We expect this general principle may also benefit other settings beyond matrix learning tasks.

# 6 Numerical experiments

In this section, we evaluate the empirical performance of our method, and compare the accuracy with other common approaches, on both synthetic and real datasets. The comparison covers a range of nonlinear, nonregular models which do not necessarily follow the assumptions in our proposal. This allows us to fairly assess the performance of various approaches under practical applications.

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

We examine the prediction accuracy of our method using four experiments. The dataset is generated from multiple-index model which extends the single index model (see Example 1) by allowing a multivariate latent response $(z_1, z_2) = (\langle \boldsymbol{B}_1, \boldsymbol{X} \rangle, \langle \boldsymbol{B}_2, \boldsymbol{X} \rangle)$. We simulate random matrices $\boldsymbol{X} \in \mathbb{R}^{d \times d}$ with i.i.d. Uniform[0,1] entries, and draw $\boldsymbol{B}_1, \boldsymbol{B}_2$ from $\mathcal{F}(r, s, s)$. The response label is simulated from $y \sim \text{Ber}(\mu(\boldsymbol{X}))$, where the regression function $\mu(\boldsymbol{X})$ is generated from the following chain scheme,

$$\mathbf{X} \longrightarrow (z_1, z_2) \underset{\text{where } \bar{z}_i = \Phi^{-1} \circ \Phi_i(z_i) \text{ for } i = 1, 2.}{\longrightarrow} (\bar{z}_1, \bar{z}_2) \longrightarrow p(\mathbf{X}) = \begin{cases} g(\bar{z}_1), & \text{if } \bar{z}_1 > 0, \\ g(\bar{z}_2), & \text{otherwise.} \end{cases}$$

We set $\Phi_i$ = empirical CDF of $z_i$ for $i = 1, 2$; $\Phi$ = CDF of standard normal; $g(z) = (1 + \exp(z))^{-1}$; matrix dimension $d = 20, 30, \ldots, 60$; and training sample size $n = 100, 150, \ldots, 400$. The construction of ~~$p$~~ $\mu$ amounts to a high nonlinearity from $\boldsymbol{X}$ to $\mu(\boldsymbol{X})$. Unlike parametric methods, the functional form of ~~$p$~~ $\mu$ is set unknown to the algorithm.

The first experiment assesses the impact of sample size to classification. We fix the matrix dimension $d = 30$ and let the sample size $n$ increase. Figure 3a plots the resulting density of $\mu(\boldsymbol{X})$ induced from the nonlinear function $\mu(\cdot)$ and distribution of $\boldsymbol{X}$. The classification error is measured by the excess risk $R(\hat{S}_{\text{bayes}}) - R(S_{\text{bayes}})$ evaluated on test data. As seen from Figure 3b, the classification error decays polynomially in sample size, which is consistent with our theoretical results. We find that a higher rank or a higher number of active nodes leads to a higher classification error, as reflected by the upward shift of the curve as $(r, s)$ increases. Indeed, a higher $(r, s)$ implies a higher complexity in the model space, thus increasing the generalization error of classification.

The second experiment evaluates the impact of matrix dimension to classification accuracy. We fix the sample size $n = 200$ and let the matrix dimension $d$ increase. Figure 3c plots the classification error versus the dimension $d$ for each of three model settings $(r, s) = (2, 2), (2, 3)$ and $(4, 4)$. We find that the error increases slowly with matrix dimension, and the growth appears well controlled by the log rate. Note that, as the dimension increases, the number of active nodes remain unchanged, but the combinatoric complexity increases in the model space. The error seems unavoidable because of the price one needs to pay for not knowing the positions of the $s$ active

nodes. The ability to effectively control massive noisy features highlights the benefit of our method in high dimensions.

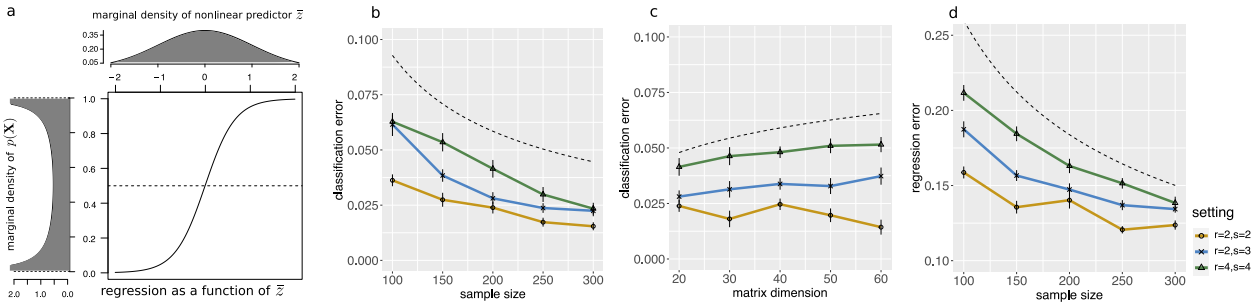



**Figure 3:** Finite sample accuracy of matrix classification and regression. (a) simulation setup. (b) classification error with sample size when $d = 30$. (c) classification error with matrix dimension when $n = 200$. (d) regression error with sample size. The dashed line in panels (b)-(d) represent theoretical rates $\mathcal{O}(n^{-2/3})$, $\mathcal{O}(\log d)$, and $\mathcal{O}(n^{-1/2})$, respectively. The reported statistics are averaged across 30 simulation replicates, with standard error given in the error bar.

The third experiment investigates similar aspects as before but to the task of matrix regression. We set the ~~regularing~~ resolution parameter $H = 20$ and aggregated the multiple level-sets as in our proposal (4.1). Figure 3d shows the regression error measured by $L$-1 risk, $\mathbb{E}|\hat{\mu}(\boldsymbol{X}) - \mu(\boldsymbol{X})|$, evaluated on test data. Again, we see that the regression error decays polynomially in sample size. Note that our matrix-valued feature has ambient dimension $30 \times 30 = 900$ whereas the sample size is on the order of hundreds. This scenario of high feature dimension low sample size is prevalent in brain network analysis. Nevertheless, our nonparametric method consistently learns the function ~~p~~ $\mu$ from limited data without the need to specify a priori functional form.

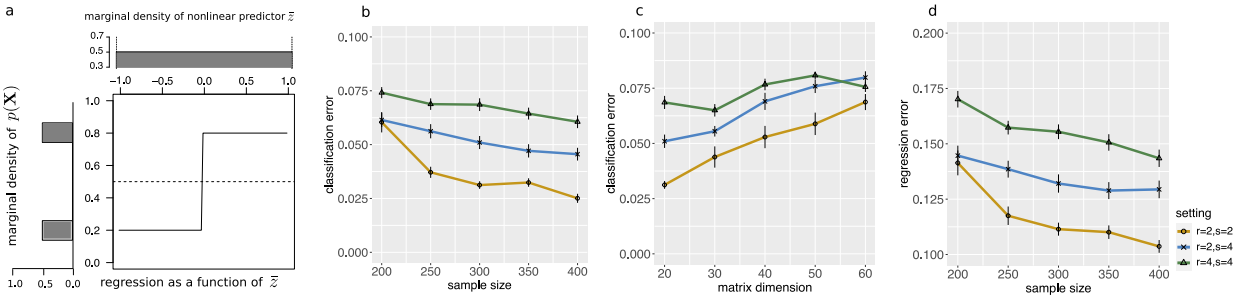



**Figure 4:** Finite sample accuracy under a different setting. (a) simulation setup. (b) classification error with sample size when $d = 30$. (c) classification error with matrix dimension when $n = 200$. (d) regression error with sample size.

The fourth experiment investigates the impact of target function to the prediction accuracy. We have shown that the probabilistic behavior of the random variable $\mu(\boldsymbol{X})$ plays a key role in our learning reduction (see Section 3). Here we assess the empirical performance by repeating all the above experiments using a variety of $\mu(\boldsymbol{X})$. For space consideration, only one representative example is presented in the main texts, and the rest in the appendix B. Figure 4a shows a model setting that falls on the other end of the spectrum. The random variable $\mu(\boldsymbol{X})$ concentrates

at two mass points $\pi = 0.2$ and 0.8. This makes the $\pi$-level set estimation challenging around $\pi = 0.2$ and 0.8, because of the nonidentifiability in the weighted classification. Interestingly, we find that our method maintains good performance on classification at $\pi = 0.5$ (Figures 4b-c) and the overall regression (Figure 4d). One possible reason of this robustness is that we aggregate in total $(H-1)$ classifiers from $\Pi = \{\frac{1}{H}, \frac{2}{H}, \ldots, \frac{H-1}{H}\}$, each of which incurs at most $\frac{1}{H}$ error to the function estimation. Therefore, the estimation is robust against off-target level sets, as long as the majority are accurate.

### 6.2 Comparison with other methods

Next, we compare our method with several popular alternative methods:

• Unstructured regular lasso (**Lasso**). We vectorize the network predictor into a high-dimensional feature and then use elastic net (**?**) with logistic loss to fit the vector-valued predictors.

• Parametric regression for network predictors with group lasso (**LogisticM**, **?**)). The original proposal is designated for network classification, and we adopt the fitted value from logistic loss as the probability estimation.

• Convolutional Neural Network (**CNN**) with two hidden layers implemented in Keras (**?**). We apply 64 filters with $3 \times 3$ convolutional kernels to the matrix-valued predictor, followed by a pooling layer with size $5 \times 5$. The resulting features (neurons) are fed to a fully connected layer of neural network with ReLU activation.

• **Non**parametric **MA**trix **R**egression (**NonMAR**). This is our method that uses level set approaches to estimate the regression function for matrix-valued predictors.

We choose a range of representative methods and investigate the benefit of each approach. The **Lasso** serves as a baseline to assess the gain of matrix-valued predictors over vector-valued predictors. The methods **CNN** and **NonMAR** are nonparametric approaches and **LogisticM** is a parametric solution for matrix based prediction.

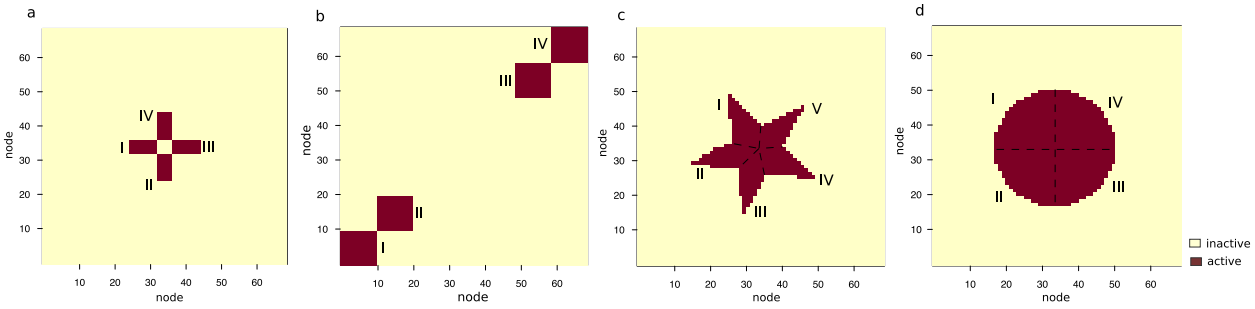



**Figure 5:** Four active pattern in simulations. The active region is divided into four or five subregions (denoted I, II, ...), each of which has its own edge connectivity signal $g_{pq}(\pi)$.

For fair comparison, we adopt similar simulation setup as in **?**), except that we add more challenging network patterns in order to assess model misspecification. We simulate the data $(\boldsymbol{X}_i, y_i)_{i=1}^n$ from latent variable model $(\boldsymbol{X}, y)|\pi$ based on the following scheme (see details in Appendix A),

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

The edge connectivity signal, $g_{pq}(\pi)$, varies depending on the response probability $\pi$ and location of $(p,q) \in [d]^2$. Figure 5 illustrates the active pattern which specifies the locations of active edges. The active region is further divided into several subregions, each of which has its own signal function $g_{pq}(\cdot)\colon [0,1] \to \mathbb{R}$. The function form of $g_{pq}(\cdot)$ is randomly drawn from a pre-specified library consisting of common functions such as $g(z) = \log(5z+1), 3\tan(z), 6z^2, \ldots$. We set $d = 68$, a training size $n = 160$, and a test size 80.

Our simulation reflects the challenging heterogeneity commonly arisen in brain network analysis. The sample consists of a mixture of individual groups with varying response probability, and the network patterns vary from one group to another. Active brain regions are supported on a submatrix with typically unknown rank. In the noiseless case, the cross and block patterns are low-rank ($r = 3$ and 5, respectively), whereas the star and circle patterns are nearly full-rank (numerical rank $r \approx 30$ on the supported submatrix). In our simulation with noisy observation, we select the rank and sparsity parameters $(r, s)$ by 5-fold cross validation. The hyperparameters for the other three methods are selected by either default setting (**LogisticM**) or cross validation (**Lasso**, **CNN**). We provide each algorithm the labeled networks as inputs after randomly permuting the node indices in the network. Because the software for **LogisticM** supports symmetric matrices only, we provide the algorithm $\frac{1}{2}(\boldsymbol{X} + \boldsymbol{X}^T)$.

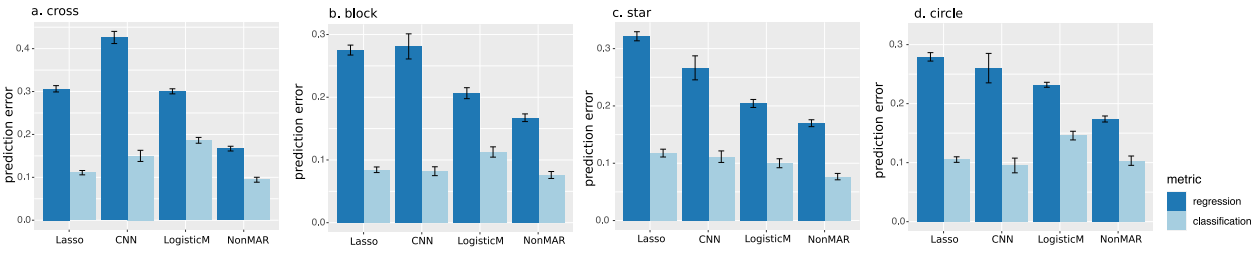



**Figure 6:** Performance comparison between various methods under four different active patterns.

Figure 6 compares the out-of-sample prediction between different methods. We focus on the regression problem and assess the performance using test data. We find that **NonMAR** consistently outperforms others, and the reduction in error is substantial. For example, the relative reduction using **NonMAR** over the next best approach, **LogisticM**, is over 20% for patterns a and d, and over 15% for patterns b and c. The results show the benefit of our nonparametric approaches by allowing more flexible functional space. Furthermore, we find that neither **Lasso** nor **CNN** has satisfactory regression performance. One possible reason is that these two methods fail to appro-

priately incorporate the network structure of the predictors. The **Lasso** takes vectorized matrices as inputs and therefore losses the two-way pairing information. On the other hand, **CNN** assumes spacial ordering within row/column indices. Although local similarity is an appropriate model for common imaging analysis, the row/column indices are meaningless for networks. Indeed, adjacency matrices differ by row/column permutation represent the same network, and methods that are index-invariant (**LogisticM** and **NonMAR**) show better performance. Our simulations cover a reasonably broad range of network models with rich structure. The results demonstrate the advantages of our nonparametric approach on the task of network regression.

We also report the accuracy on classification which is an intermediate step of regression. Figure 6 shows the favorable performance of our method especially when compared to **LogisticM**. Among the four models of active regions, our method performs the best in all three. The only exception is the circle pattern where the **CNN** has a lower classification error by a slight margin. This is perhaps due to the fact that circle pattern is nearly full rank which favors complicated models such as **CNN**. Nevertheless, our method **NonMAR** achieves stable performance in spite of its simplicity. Interestingly, we find that the benefit of our method is more substantial in regression than in classification. One possible reason is that classification is an easier problem and less sensitive to various approaches. The result also suggests that the main reason of our method's superior regression performance may be attributable to level set aggregations.

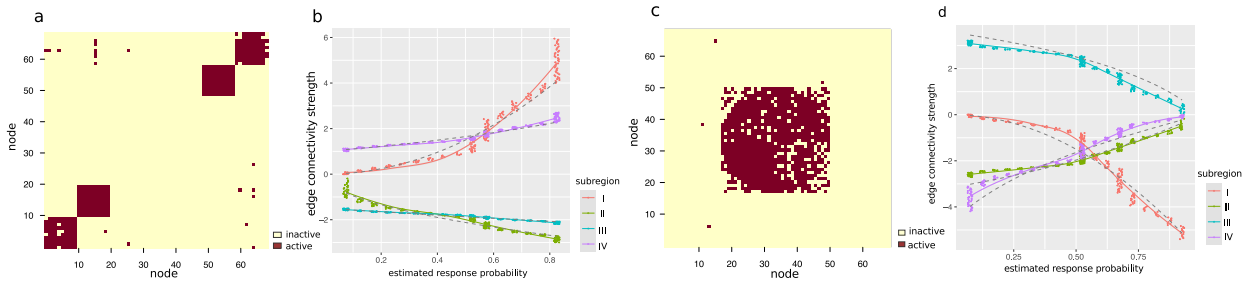



**Figure 7:** Example outputs returned by **NonMAR**. Panels (a) and (c) plot the top edges selected by our method. Panels (b) and (d) are scatter plots of the edge connectivity strength (averaged by subregion) versus the estimated response probability. The ground truth function is depicted in dashed curve.

We provide illustrate examples to show the outputs returned by **NonMAR**. Figures 7a and c plot the top edges selected by **NonMAR** based on the moving averages of feature weights ~~$(\hat{\boldsymbol{B}}_\pi)_{\pi\in[0,1]}$ with a window size $\Delta\pi = 0.2$~~$(\hat{\boldsymbol{B}}_\pi)_{\pi\in\Pi}$ with $\Pi = \{0.2, \ldots, 0.8\}$. The selected region agrees well with the ground truth (Figures 6a and c). We also investigate the relationship between edge connectivity for individual $i$ and the estimated response probability $\hat{\pi}_i$ (Figures 7b and d). The trajectory of the edge connectivity accurately resembles the ground truth function in each subregion. The results demonstrate that our method is able to recover the right "sorting" of individuals with respect to the response probability on a continuous spectrum. The successful recovery of complicated unknown functions makes our method **NonMAR** appealing in applications.

### 6.3 human brain connectome data

We apply our method to brain network data from Human Connectome Project (HCP) (**?**). The HCP is a recent consortial effort that aims to understand the relationships between brain connectivity and human traits. The dataset consists of brain connectivity networks from a sample of healthy individuals. The connectivity networks were preprocessed following a standard pipeline (**?**), and the brain was parcellated to 68 regions-of-interest following the Desikan atlas (**?**). Each brain network is represented as a 68-by-68 binary matrix, where the entries encode the presence or absence of fiber connections between the 68 brain regions. In addition, a range of cognitive, motor, sensory, and emotional scores are measured for each individual.

We analyze the Variable Short Penn Line Orientation Test (VSPLOT) score which measures the individual's visuospatial processing ability. In VSPLOT, two line segments are presented on the screen, and individuals are asked to rotate a movable line to make it parallel to the fixed line (**?**). We use the dataset processed by **?**), and analyze $n = 212$ individuals whose VSPLOT scores are either high ($y = 1$) or low ($y = -1$). We adjust age and gender as additional covariates in the prediction, and use a random 60-20-20 split of the data for training, validation, and testing.

a

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

Note: CNN dost not report summary statistics for node selection. Numbers in parentheses are standard errors over 5-fold cross validations.

b

| Rank | Node | Node | $p$-value* |
|---|---|---|---|
| 1 | r-inferiortemp** | r-middletemp | 0.01 |
| 2 | r-pars | r-supra | 3e-5 |
| 3 | r-posterior | l-precentral | 0.01 |
| 4 | l-caudal | l-isthmus | 2e-5 |

* calculated from two sample test based on two label groups.
** Node names are shown in abbreviations, with "r" and "l" indicating the right and left hemisphere, respectively.

**Figure 8:** HCP analysis results based on our method **NonMAR**. (a) Comparison of prediction accuracy. (b) Top edges selected by our method **NonMAR-p**.

We compare our performance to other methods using the same procedure as in the previous section. Figure 8a shows that our method achieves high regression accuracy, measured by area under receiver operating characteristic (AUC). As common in high-dimensional settings, we observe that models with optimal cross-validation accuracy tend to include many noise variables. A useful heuristic is the so-called "one-standard-error rule" (**?**), in which one 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 **NonMAR-p**. It is remarkable to see that **NonMAR-p** results in 12% reduction of active nodes but still achieves excellent accuracy (AUC = 0.73).

Figure 8b lists the top brain edges identified by our method. Edges are ranked by their maximal values in the feature weights $(\hat{\boldsymbol{B}}_\pi)_{\pi\in[0,1]}$ via moving averaging. We find that the top edges involve connections between frontal and occipital regions in the right hemisphere (Figure 9a). This seems consistent with recent evidence of dysfunction in right posterior regions for deficits in visuospatial processing (**?**). We also find nonlinear relationship between edge connection strength and response probability. In Figure 9b, the connection (r-parstriangularis, r-supramarginal) grows slowly when

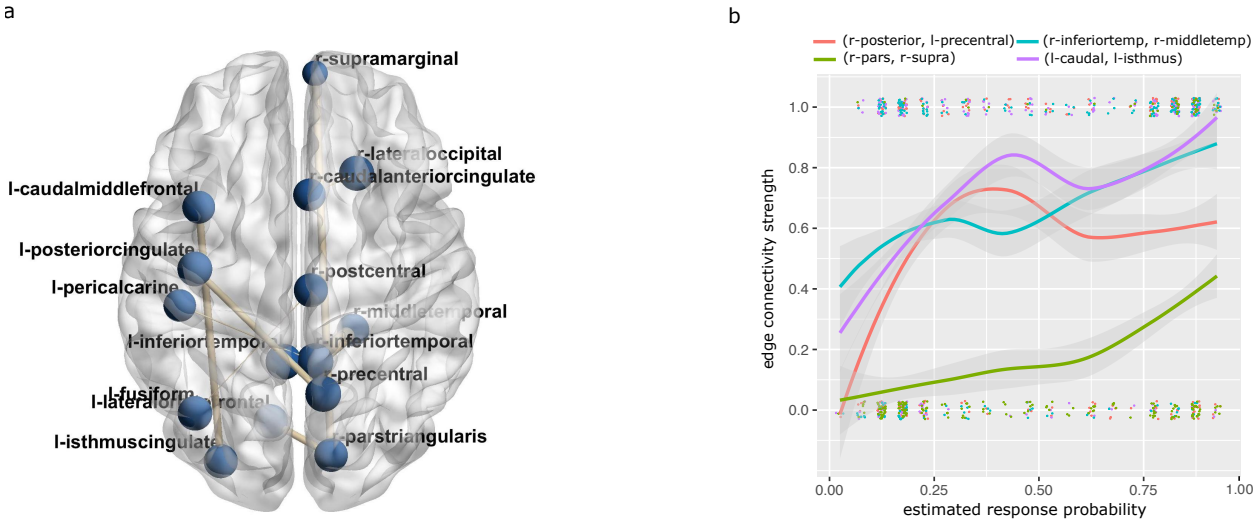



**Figure 9:** HCP analysis results based on our method **NonMAR**. (a) Top edges overlaid on brain template. (b) Edge connectivity strength versus estimated response probability. Colored curves represent the moving averages of connectivity strengths, gray bands represent one standard error, and jitter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

$\pi$ is low but fast when $\pi$ is high. In contrary, the connection (r-posteriorcingulate, r-precentral) grows fast initially and then reaches a plateau as $\pi$ increaases. The detected pattern reveals the heterogeneous changes in brain connectivities with respect to visuospatial processing ability.

# 7 Discussion

We have developed the learning framework for the relationship between a binary label response and a high-dimensional matrix-valued predictor. Our method respects the matrix structure of the predictors and provide interpretable prediction via a nonparametric approach. The theoretical and numerical results demonstrate the competitive performance of our method.

Our work unlocks several directions of future research. Although we have presented the work in the context of matrix-valued predictors, we may consider extensions to other nonconventional predictors, such as images, graphs, tensors, functional time-series, and so on. Recent research has shown fruitful results in classification tasks with nonconventional predictors (**??**). Our proposed reduction approach provides a potential building block for more challenging regression tasks. It is also worthy noting that each of the aforementioned predictors may processes its own special structure depending on the applications. Exploiting the benefits and properties of nonparametric regression in the specialized tasks warrants future research.

We may also ask whether the results here, provided in the setting of binary regression with $y \in \{-1, 1\}$, may be extended to a continuous response $y \in \mathbb{R}$. The answer is affirmative if we assume the response $y$ is bounded, e.g. $y \in [-L, L]$, for $L > 0$. One possible solution is to use response-

dependent weight in place of response-dependent weight in the classification. Specifically, for a fixed target level $\pi \in [-L, L]$, define a new binary response $\tilde{y} = \text{sign}(y - \pi)$, a response-dependent weight $\tilde{w}_\pi(y) = |y - \pi|$, and a general weighted classification risk

$$\tilde{R}_\pi(S) \stackrel{\text{def}}{=} \mathbb{E}\left[\tilde{w}_\pi(y)\mathbb{1}(\tilde{y} \neq \text{sign}(\boldsymbol{X} \in S))\right]. \tag{22}$$

The risk (7) extends the $\pi$-weighted classification risk (2.2) for a continuous response, where the weight $\tilde{w}_\pi(y) = |y - \pi|$ is the distance from the response $y$ to the target level $\pi$. Importantly, the level set $S(\pi) = \{\boldsymbol{X} \in \mathcal{X} \colon \mu(\boldsymbol{X}) \geq \pi\}$ is the global minimum of (7) under conditional model of the type $y|\boldsymbol{X} = \mu(\boldsymbol{X}) + \varepsilon$, where the noise $\varepsilon$ is a mean-zero random variable whose distribution is allowed to depend on $\boldsymbol{X}$ (**??**). With this statistical characterization, our main result on excess regression risk bound in Section 3 still holds. Therefore, our learning reduction approach equally applies to a continuous response by using $\tilde{w}_\pi(y)$ and $\tilde{y}$ in place of $w_\pi(y)$ and $y$, respectively. For an unbounded response, it is unclear whether the level set approach still achieves accuracy guarantees. We leave these questions for future work.

# Appendix

## A Connection to joint matrix decomposition with functional coefficients

In the main text, we have shown that our classifier functions $\mathcal{F}(r, s_1, s_2)$ incorporates certain retrospective models $\boldsymbol{X}|y$. Here we extend the retrospective model to a more general family which we coin as "functionally decomposable matrices". The model essentially extends Example 2 from two classes of $\boldsymbol{X}$ to a series of $\boldsymbol{X}$ on a continuous spectrum of $\pi$.

**Example 5** (Functionally decomposable matrices)**.** We consider a matrix model $\boldsymbol{X}_\pi \overset{\text{def}}{=} \boldsymbol{B}_0 + \sum_{s\in[r]} g_s(\pi)\boldsymbol{B}_s + \sigma\boldsymbol{E}$, where $\pi \in [0,1]$ is a real-valued index; $\boldsymbol{E}$ is a noise matrix consisting of i.i.d. entries in $N(0,1)$; $\sigma$ is the noise level; $\boldsymbol{B}_0$ is an arbitrary baseline matrix; $(\boldsymbol{B}_s)_{s\in[r]}$ is a set of rank-1 matrices in $\{0,1\}^{d_1\times d_2}$ that satisfy (i) non-overlapping supports, i.e., $\langle \boldsymbol{B}_s, \boldsymbol{B}_{s'}\rangle = 0$ for all $s \neq s'$, and (ii) bounded total support, i.e., $\sum_{s\in[r]} \text{supp}(\boldsymbol{B}_s) \leq (s_1, s_2)$; and the coefficients $g_s(\cdot)\colon [0,1] \to \mathbb{R}$ are monotonic functions with respect to $\pi$ for all $s \in [r]$.

Now, suppose the observation takes the form of pair $(\boldsymbol{X}_\pi, y_\pi)$ with $y_\pi \sim \text{Ber}(\pi)$, where $y_\pi$ is conditionally independent of $\boldsymbol{X}_\pi$ given $\pi$, and $\pi \in [0,1]$ is drawn from an unknown distribution over $[0,1]$. Equivalently, the generative model is expressed in the following scheme:

$$\pi \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1] \overset{\text{conditional on } \pi}{\longrightarrow} \begin{cases} y \sim \text{Ber}(\pi),\ y \perp \boldsymbol{X}|\pi, \\ \boldsymbol{X} = \boldsymbol{B}_0 + [\![\boldsymbol{X}_{pq}]\!], \end{cases}$$

where ~~$s = s(p,q)$ is the index such that the~~ $\boldsymbol{X}_{pq} \overset{\text{indep.}}{\sim} \mathcal{N}(g_s(\pi), \sigma^2)$ if the position $(p,q)$ falls in the support of $\boldsymbol{B}_s$ and $\boldsymbol{X}_{pq} \overset{\text{indep.}}{\sim} \mathcal{N}(0,\sigma^2)$ otherwise. Recall that $\mu(\boldsymbol{X}) = \mathbb{E}(y|\boldsymbol{X})$. In the noiseless case $\sigma = 0$, there exists a sequence of functions $(f_\pi)_{\pi\in(0,1)}$ with $f_\pi \in \mathcal{F}(r, s_1, s_2)$, such that $\text{sign}(\mu(\boldsymbol{X}) - \pi) = \text{sign} f_\pi(\boldsymbol{X})$ for all $\pi \in (0,1)$.

The above example shows the connection of our method to joint decomposition of matrices $(\boldsymbol{X}_\pi)_{\pi\in[0,1]}$. We should point out, despite of the seeming similarity, a fundamental challenge arises in our setting when the latent index $\pi$ is unobserved. Our level-set approach essentially learns the right "sorting" of $\boldsymbol{X}_\pi$ against the index $\pi \in [0,1]$ (see Figure 2), thereby facilitating the joint estimation of matrix basis $\boldsymbol{B}_s$ and relationship $\pi = \pi(\boldsymbol{X})$.

Furthermore, we provide additional results on the low-rank two-way sparse classifiers. The following proposition gives a sufficient condition for exact recovery of level sets through halfspaces.

**Proposition 2** (Low-rank and sparse boundaries)**.** Let $\mu(\boldsymbol{X})$ be a regression function continuous in $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$. Suppose that there exists a low-rank two-way sparse matrix $\boldsymbol{B}_\pi$ and a real value $b_\pi \in \mathbb{R}$, such that the boundary set $\partial S(\pi) = \{\boldsymbol{X}\colon \mu(\boldsymbol{X}) = \pi\}$ is included in the set $\{\boldsymbol{X}\colon \langle \boldsymbol{X}, \boldsymbol{B}_\pi\rangle = b_\pi\}$. Then $\text{sign}(\mu(\boldsymbol{X}) - \pi) = \text{sign}(f_\pi(\boldsymbol{X}))$ for some $f_\pi(\boldsymbol{X}) \in \mathcal{F}(r, s_1, s_2)$. Note that $(\boldsymbol{B}_\pi, b_\pi)$ is allowed to vary depending on $\pi$.

*Proof of Proposition 2.* For given $\pi \in \Pi$, suppose $\partial S(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) = \pi\} = \{\boldsymbol{X} : \langle \boldsymbol{B}_\pi, \boldsymbol{X} \rangle = b_\pi\}$. Define two open sets $S_1$ and $S_2$ as

$$S_1(\pi) = \bar{S}(\pi) \setminus \partial S(\pi) = \{\boldsymbol{X} \colon \langle \boldsymbol{B}_\pi, \boldsymbol{X} \rangle > b_\pi\} \quad \text{and} \quad S_2(\pi) = \bar{S}^c(\pi) = \{\boldsymbol{X} \colon \langle \boldsymbol{B}_\pi, \boldsymbol{X} \rangle < b_\pi\}.$$

First, we prove that there is no $\boldsymbol{X}_1, \boldsymbol{X}_2 \in S_i(\pi)$ such that $p(\boldsymbol{X}_1) > \pi$ while $p(\boldsymbol{X}_2) < \pi$ for $i = 1, 2$. To be specific, $S_i(\pi)$ is included either $S(\pi) \setminus \partial S(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$ or $S^c(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) < \pi\}$.

It suffices to show the case $S_1(\pi)$. Without loss of generality, suppose that there is $\boldsymbol{X}_1, \boldsymbol{X}_2 \in S_1(\pi)$ such that $p(\boldsymbol{X}_1) > \pi$ and $p(\boldsymbol{X}_2) < \pi$. We can always find $\boldsymbol{X}_3 = \lambda \boldsymbol{X}_1 + (1 - \lambda)\boldsymbol{X}_2$ for some $\lambda \in (0, 1)$ such that $p(\boldsymbol{X}_3) = \pi$ by continuity of $\mu(\boldsymbol{X})$. Because $p(\boldsymbol{X}_3) = \pi$, we have $\boldsymbol{X}_3 \in \partial S(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) = \pi\} = \{\boldsymbol{X} \colon \langle \boldsymbol{B}_\pi, \boldsymbol{X} \rangle = b_\pi\}$. However, this contradict the fact

$$\langle \boldsymbol{B}_\pi, \boldsymbol{X}_3 \rangle = \lambda \langle \boldsymbol{B}_\pi, \boldsymbol{X}_1 \rangle + (1 - \lambda) \langle \boldsymbol{B}_\pi, \boldsymbol{X}_2 \rangle > \lambda b_\pi + (1 - \lambda) b_\pi = b_\pi.$$

Therefore, there is no $\boldsymbol{X}_2 \in \bar{S}_1(\pi)$ such that $p(\boldsymbol{X}_1) < \pi$.

Second, we prove that there are only two cases that

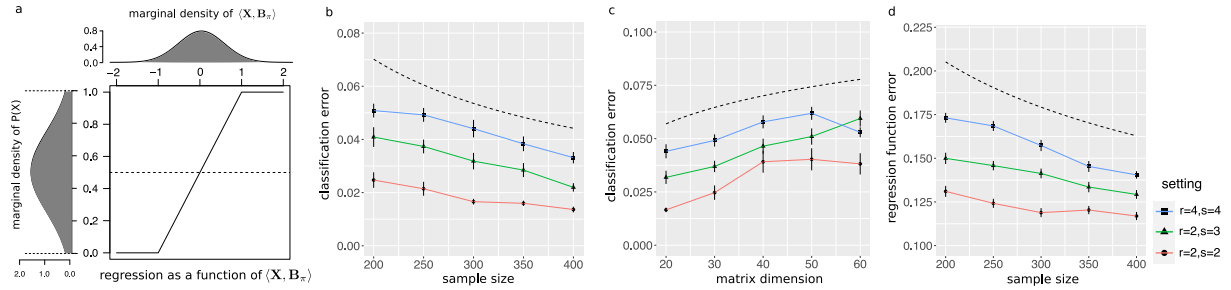
(C1): $S_1(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$ and $S_2(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) < \pi\}$.

(C2): $S_1(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) < \pi\}$ and $S_2(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$.

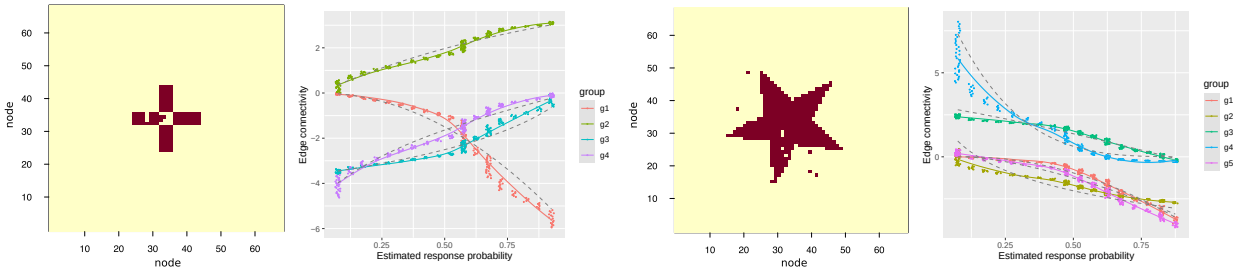
Suppose $S_1(\pi) \subset \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$. Since $\{\partial S(\pi), S_1(\pi), S_2(\pi)\}$ is a partition of $\mathbb{R}^{d_1 \times d_2}$ and both $\{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$ and $\{\boldsymbol{X} \colon \mu(\boldsymbol{X}) < \pi\}$ are non-empty set, only possible case is (C1). Similarly, if $S_2(\pi) \subset \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) < \pi\}$, only possible case is (C2).

Notice that (C1) is equivalent to $S(\pi) = \bar{S}(\pi)$ while (C2) is equivalent to $S(\pi) = \bar{S}^c(\pi)$. □

## B Supplementary figures




**Supplemental Figure S1:** Finite sample accuracy of matrix classification and regression. (a) simulation setup. (b) classification error with sample size when $d = 30$. (c) classification error with matrix dimension when $n = 200$. (d) regression error with sample size. The dash line in panels (b)-(d) represent theoretical rate $\mathcal{O}(n^{-2/3})$, $\mathcal{O}(\log d)$, and $\mathcal{O}(n^{-1/3})$, respectively.




**Supplemental Figure S2:** Example outputs returned by **NonMAR**. Panels (a) and (c) plot the top edges selected by our method. Panels (b) and (d) are scatter plots of the edge connectivity strength (averaged by subregion) versus the estimated response probability. The ground truth function is depicted in dashed curve.

## C Proofs

### C.1 Proof of Proposition 1

*Proof of Proposition 1.* For ease of notation, we drop the argument $\pi$ from $S_{\text{bayes}}(\pi)$, $R_\pi(\cdot)$, and $d_\pi(\cdot,\cdot)$, and simply write $S_{\text{bayes}}$, $R(\cdot)$, and $d(\cdot,\cdot)$ respectively. The following identity is useful to relate the excess risk and set difference in classifiers,

$$\begin{aligned} d(S, S_{\text{bayes}}) &\stackrel{\text{def}}{=} R(S) - R(S_{\text{bayes}}) \\ &= \mathbb{E}_{\boldsymbol{X},y}\left[w(y)\mathbb{1}(y \neq I(S))\right] - \mathbb{E}_{\boldsymbol{X},y}\left[w(y)\mathbb{1}(y \neq I(S_{\text{bayes}}))\right] \\ &= \mathbb{E}_{\boldsymbol{X}}\left[(\pi - \mu(\boldsymbol{X}))\left(I(S) - I(S_{\text{bayes}})\right)\right] \\ &= 2\int_{\boldsymbol{X} \in S \Delta S_{\text{bayes}}} |\mu(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}}. \end{aligned} \tag{23}$$

~~Without loss of generality, suppose $\rho(\pi,\mathcal{N}) \leq C_2$. If $\rho(\pi,\mathcal{N}) > C_2$, we can replace $\rho(\pi,\mathcal{N})$ by $C_2$ in . With a modification of the constant $c$, still holds.~~ We divide the proof into two cases: $\alpha \in (0,1)$ and $\alpha = 1$.

Case 1: $\alpha \in (0,1)$.

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 $A = \{\boldsymbol{X} \colon |\mu(\boldsymbol{X}) - \pi| > t\}$.

$$\begin{aligned} \int_{\boldsymbol{X} \in S \Delta S_{\text{bayes}}} |\mu(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}} &\geq t\left[\mathbb{P}_{\boldsymbol{X}}\left((S \Delta S_{\text{bayes}}) \cap A\right)\right] \\ &\geq t\left(\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) - \mathbb{P}_{\boldsymbol{X}}(A^c)\right) \\ &\geq t\left(\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) - C_1 t C t^{\frac{\alpha}{1-\alpha}}\right), \quad \text{for all } 0 \leq t \leq \in (0, \rho(\pi,\mathcal{N})). \end{aligned}$$

Combining the above inequality with the identity (C.1) yields

$$d(S, S_{\text{bayes}}) \geq 2t\left(\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) - C_1 t C t^{\frac{\alpha}{1-\alpha}}\right), \quad \text{for all } t \in (0, \rho(\pi,\mathcal{N})]). \tag{24}$$

We maximize the lower bound of (C.1) with respect to $t$ and obtain the optimal ~~$0 \leq t_{\text{opt}} < \rho(\pi,\mathcal{N})$,~~ $t_{\text{opt}} \in (0, \rho(\pi,\mathcal{N})])$,

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

The corresponding lower bound of the inequality (C.1) becomes

$$d(S, S_{\text{bayes}}) \geq \begin{cases} 2\alpha\rho(\pi,\mathcal{N})\mathbb{P}(S \Delta S_{\text{bayes}}), & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) \geq \frac{C}{1-\alpha}\rho(\pi,\mathcal{N})^{\frac{\alpha}{1-\alpha}}, \\ 2\alpha\left(\frac{1-\alpha}{C}\right)^{\frac{1-\alpha}{\alpha}}\mathbb{P}_{\boldsymbol{X}}^{\frac{1}{\alpha}}\left(S \Delta S_{\text{bayes}}\right), & \text{if} \quad \mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) < \frac{C}{1-\alpha}\rho(\pi,\mathcal{N})^{\frac{\alpha}{1-\alpha}}. \end{cases}$$

Combining both cases gives

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

where we take ~~$c = \max \left( \frac{1}{2\alpha}, \left( \frac{C_1}{1-\alpha} \right)^{1-\alpha} \left( \frac{1}{2\alpha} \right)^{\alpha} \right)$~~ $c = \max \left( \frac{1}{2\alpha}, \left( \frac{C}{1-\alpha} \right)^{1-\alpha} \left( \frac{1}{2\alpha} \right)^{\alpha} \right)$.

Case 2: $\alpha = 1$.

The inequality (C.1) now becomes

$$d(S, S_{\text{bayes}}) \geq 2t \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}) = 2t d_\Delta(S, S_{\text{bayes}}), \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}),$$

$$d(S, S_{\text{bayes}}) \geq 2t \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}) = 2t d_\Delta(S, S_{\text{bayes}}), \quad \text{for all } t \in [0, \rho(\pi, \mathcal{N})], \pi \notin \mathcal{N}.$$

~~Therefore the conclusion follows by taking $t = \frac{\rho(\pi, \mathcal{N})}{2}$ in the inequality~~ The inequality (??) holds with $c = \frac{1}{2}$ plugging $t = 1/\rho(\pi, \mathcal{N})$ into the above inequality.

Therefore, Proposition 1 holds. □

### C.2 Proof of Theorem 1

*Proof of Theorem 1.* It follows from the definition of $\bar{\mu}(\boldsymbol{X})$ that

$$|\mu(\boldsymbol{X}) - \bar{\mu}(\boldsymbol{X})| \leq \frac{1}{2H} \sum_{\pi \in \Pi} \left| I(\boldsymbol{X} \in \bar{S}(\pi)) - I(\boldsymbol{X} \in S_{\text{bayes}}(\pi)) \right| + \left| \mu(\boldsymbol{X}) - \frac{1}{2} - \frac{1}{2H} \sum_{\pi \in \Pi} I(\boldsymbol{X} \in S_{\text{bayes}}(\pi)) \right|$$

$$= \frac{1}{H} \sum_{\pi \in \Pi} \mathbb{1}(\boldsymbol{X} \in \bar{S}(\pi) \Delta S_{\text{bayes}}(\pi)) + \left| \mu(\boldsymbol{X}) - \frac{1}{2H} \left( 1 + 2 \sum_{\pi \in \Pi} \mathbb{1}(\boldsymbol{X} \in S_{\text{bayes}}(\pi)) \right) \right|$$

$$\leq \frac{1}{H} \sum_{\pi \in \Pi} \mathbb{1}(\boldsymbol{X} \in \bar{S}(\pi) \Delta S_{\text{bayes}}(\pi)) + \frac{1}{2H}$$

$$\leq \frac{1}{H} \sum_{\pi \in \Pi \setminus \mathcal{N}} \mathbb{1}(\boldsymbol{X} \in \bar{S}(\pi) \Delta S_{\text{bayes}}(\pi)) + \frac{1 + 2||2c}{2H} \leq \max_{\pi \in \Pi \setminus \mathcal{N}} \mathbb{1}(\in S(\pi) \Delta(\pi)) + 1 + 2C' 2H.,$$

where the last inequality follows from the assumption that $|\mathcal{N}| \leq c$. Taking expectation with respect to $\boldsymbol{X}$ on both sides gives

$$\mathbb{E}|\mu(\boldsymbol{X}) - \bar{\mu}(\boldsymbol{X})| \leq \max_{\pi \in \Pi \setminus \mathcal{N}} \frac{1}{H} \sum_{\pi \in \Pi \setminus \mathcal{N}} d_\Delta(\bar{S}(\pi), S_{\text{bayes}}(\pi)) + \frac{1 + 2C' 2c}{2H}$$

$$\leq c \max_{\pi \in \Pi \setminus \mathcal{N}} \lesssim \frac{1}{H} \sum_{\pi \in \Pi \setminus \mathcal{N}} \left[ d_\pi^\alpha(\bar{S}(\pi), S_{\text{bayes}}(\pi)) + Hd \frac{1}{\rho(\pi, \mathcal{N})} d_\pi(\bar{S}(\pi), S_{\text{bayes}}(\pi)) \right] + \frac{1+2C'}{2HH},$$

where the last line comes from ~~condition~~ Proposition 1. Furthermore, the definition of regression risk (2.3) implies

$$\begin{aligned} R_{\text{reg}}(\bar{\mu}) - R_{\text{reg}}(p) &= \mathbb{E}_{(\boldsymbol{X},y)} \left( |2\bar{\mu}(\boldsymbol{X}) - y - 1|^2 - |2\mu(\boldsymbol{X}) - y - 1|^2 \right) \\ &= 4\mathbb{E}_{\boldsymbol{X}} \left[ p^2(\boldsymbol{X}) + \bar{\mu}^2(\boldsymbol{X}) - 2\mu(\boldsymbol{X})\bar{\mu}(\boldsymbol{X}) \right] \\ &= 4\mathbb{E}_{\boldsymbol{X}} [\mu(\boldsymbol{X}) - \bar{\mu}(\boldsymbol{X})]^2 \\ &\leq 4\mathbb{E}_{\boldsymbol{X}} |\mu(\boldsymbol{X}) - \bar{\mu}(\boldsymbol{X})|. \end{aligned}$$

□

### C.3 Proof of Proposition 3

**Proposition 3** (Polynomial continuity of inverse links)**.** Consider the parametric model $p = g \circ f$ as in Example 1, where we assume random predictors $\boldsymbol{X}$ with i.i.d. Uniform$[0,1]$ entries. Suppose the inverse link function has a constant $c_1, c_2 > 0$ and $\alpha \in [0,1]$ such that $g^{-1}(\pi+t) - g^{-1}(\pi-t) \leq c_1 t^{\frac{\alpha}{1-\alpha}}$ for all ~~$t \in (0, c_2)$~~ $t \in [0, c_2)$. Then,

$$\mathbb{P}(|\mu(\boldsymbol{X}) - \pi| \leq t) \leq c_3 t^{\frac{\alpha}{1-\alpha}}, \quad \text{for all } t \in ([0, c_2),$$

where $c_3 > 0$ is a constant depending on $\pi \in (0,1)$. In particular, when the function $g$ is the identity or logistic link, we have $\alpha = 1/2$.

*Proof.* Let $f(\boldsymbol{x}) = \langle \boldsymbol{b}, \boldsymbol{x} \rangle$ where $\boldsymbol{b} = (b_1, \ldots, b_d) \in \mathbb{R}^d$ and $\boldsymbol{x} = (x_1, \ldots, x_d) \in [0,1]^d$. First we prove that

$$\int_{\{\boldsymbol{x}: l(t) \leq \langle \boldsymbol{b}, \boldsymbol{x} \rangle \leq u(t)\}} d\boldsymbol{x} \leq c_d t^{\frac{\alpha}{1-\alpha}} \tag{25}$$

for any $l(t)$ and $u(t)$ such that $u(t) - l(t) \leq c t^{\frac{\alpha}{1-\alpha}}$ for some constant $c > 0$.

1. Consider the case when $d = 1$ where $\boldsymbol{x} = x_1 \in [0,1]$ and $\boldsymbol{b} = b_1 \in \mathbb{R}$.

$$\int_{\{\boldsymbol{x}: l(t) \leq \langle \boldsymbol{b}, \boldsymbol{x} \rangle \leq u(t)\}} d\boldsymbol{x} = \int_{\frac{l(t)}{b_1} \leq x_1 \leq \frac{u(t)}{b_1}} dx_1 \leq c_1 t^{\frac{\alpha}{1-\alpha}}.$$

2. Suppose that when $d = k$ where $\boldsymbol{x} = (x_1, \ldots, x_k) \in [0,1]^k$ and $\boldsymbol{b} = (b_1, \ldots, b_k) \in \mathbb{R}^k$, the

following is satisfied.

$$\int_{\{\boldsymbol{x}: l(t) \le \langle \boldsymbol{b}, \boldsymbol{x} \rangle \le u(t)\}} d\boldsymbol{x} \le c_k t^{\frac{\alpha}{1-\alpha}}.$$

3. Consider the case when $d = k+1$ where $\boldsymbol{x} = (x_1, \ldots, x_{k+1}) \in [0,1]^{k+1}$ and $\boldsymbol{b} = (b_1, \ldots, b_{k+1}) \in \mathbb{R}^{k+1}$.

$$\begin{aligned}
\int_{\{\boldsymbol{x}: l(t) \le \langle \boldsymbol{b}, \boldsymbol{x} \rangle \le u(t)\}} d\boldsymbol{x} &= \int_{\{\boldsymbol{x}: l(t) - b_{k+1}x_{k+1} \le b_1 x_1 + \ldots b_k x_k \le u(t) - b_{k+1}x_{k+1}\}} d\boldsymbol{x}_{1:k} dx_{k+1} \\
&\overset{=}{\underset{\sim}{\le}} \int_0^1 \int_{l(t) - b_{k+1}x_{k+1} \le b_1 x_1 + \ldots b_k x_k \le u(t) - b_{k+1}x_{k+1}} d\boldsymbol{x}_{1:k} dx_{k+1} \\
&\le \int_0^1 c_k t^{\frac{\alpha}{1-\alpha}} \, dx_{k+1} \\
&\le c_{k+1} t^{\frac{\alpha}{1-\alpha}}.
\end{aligned}$$

4. By mathematical induction, (C.3) holds for arbitrary $d$.

Notice that

$$\begin{aligned}
\mathbb{P}(|\mu(\boldsymbol{X}) - \pi| \le t) &= \mathbb{P}(\pi - t \le \mu(\boldsymbol{X}) \le \pi + t) \\
&= \mathbb{P}(\pi - t \le g(\langle \boldsymbol{B}, \boldsymbol{X} \rangle) \le \pi + t) \\
&= \mathbb{P}(g^{-1}(\pi - t) \le \langle \boldsymbol{B}, \boldsymbol{X} \rangle \le g^{-1}(\pi + t)).
\end{aligned}$$

Combining the assumption that $g^{-1}(\pi + t) - g^{-1}(\pi - t) \le c_1 t^{\frac{\alpha}{1-\alpha}}$ for all ~~$t \in (0, c_2)$~~ $t \in [0, c_2)$ and (C.3) gives

$$\mathbb{P}(|\mu(\boldsymbol{X}) - \pi| \le t/H) \le c_3 t^{\frac{\alpha}{1-\alpha}}, \text{ for all } t \in [0, c_2).$$

In particular, when the link function is the identity or logistic link we have $\alpha = 1/2$.

When $g(z) = z$, we have

$$g^{-1}(\pi + t) - g^{-1}(\pi - t) = 2t,.$$

~~where the constant is independent on $\pi \in (0,1)$. Therefore, we have $\alpha = 1/2$ and $\beta = 1$.~~

When $g(z) = \frac{e^z}{1+e^z}$, we have

$$\begin{aligned}
g^{-1}(\pi + t) - g^{-1}(\pi - t) &= \log\left(\frac{\pi - t}{1 - (\pi - t)}\right) - \log\left(\frac{\pi + t}{1 - (\pi + t)}\right) \\
&= \frac{2}{\pi^*(1 - \pi^*)} t,
\end{aligned}$$

for some $\pi^* \in [\pi - t, \pi + t]$ by mean value theorem.

Therefore, we have $\alpha = 1/2$ when the link function is the identity or logistic link. □

### C.4 Proof of Theorem 2

*Proof of Theorem 2.* Our proof adopts the techniques of **?**, Theorem 3) to the contexts of classifier functions $\mathcal{F}(r, s_1, s_2)$. We summarize only the key difference here but refer to **?**) for complete proof.

Let $\hat{f} = I(\hat{S}_{\text{bayes}}(\pi)), f_{\text{bayes}} = I(S_{\text{bayes}}(\pi))$ be the indicator functions corresponding to the set $\hat{S}_{\text{bayes}}(\pi), S_{\text{bayes}}(\pi) \subset \mathbb{R}^{d_1 \times d_2}$, respectively . From Proposition 4, ~~Assumption 1(ii) implies~~ $(\pi, \alpha)$-local regularity implies

$$\text{Var}[w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))] \leq c' \lesssim [R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes},\pi})]^{\alpha}.$$

Notice $\rho(\pi, \mathcal{N})$ is a constant in the proposition because we only consider a given $\pi$. Applying the second order condition to Theorem 3 of **?**) gives that, with the choice $\lambda \leq \frac{t_n}{4J_n}$,

$$\mathbb{P}\left[R_\ell(\hat{f}) - R_\ell(f_{\text{bayes}}) \geq \max\{a_n, t_n\}\right] \leq \frac{7}{2}\exp\left(-Cn(\lambda J_n)^{2-\alpha}\right), \tag{26}$$

where $C > 0$. Here $\{a_n\}$ is the vanishing sequence specified in Assumption 1. Combining (5) and (C.4) gives

$$\mathbb{P}\left[d_\pi\left(\hat{S}_{\text{bayes}}(\pi), S_{\text{bayes}}(\pi)\right) \geq \max\{a_n, t_n\}\right] \leq \frac{7}{2}\exp\left(-Cn(\lambda J_n)^{2-\alpha}\right).$$

The rate of convergence $t_n > 0$ is determined by the solution to the following inequality,

$$\sup_{k \geq 2} \frac{1}{L}\int_L^{L^{\alpha/2}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}^k, \|\cdot\|_2)}d\varepsilon \leq n^{1/2}, \quad \text{where } L = t_n + \lambda J_n(k/2 - 1). \tag{27}$$

In particular, the smallest $t_n$ satisfying (C.4) yields the best upper bound of the error rate. Here $\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}^k, \|\cdot\|_2)$ denotes the $L_2$-norm, $\varepsilon$-bracketing number for function family $\mathcal{F}^k$, and for all $k \in \mathbb{N}_+$, we define $\mathcal{F}^k = \{f \in \mathcal{F}(r, s_1, s_2) \colon \|f\|_F^2 \leq k\}$; i.e., the subset of functions in $\mathcal{F}(r, s_1, s_2)$ with magnitudes bounded by $k$.

It remains to solve for the smallest possible $t_n$ in (C.4). Based on Lemma 6, the inequality (C.4) is satisfied with

$$t_n \asymp \left(\frac{r(s_1 + s_2)\log d}{n}\right)^{1/(2-\alpha)} \quad \text{and} \quad \lambda \asymp \frac{t_n}{J_n}, \tag{28}$$

Plugging (C.4) into (C.4) gives

$$\mathbb{P}\left[d_\pi\left(\hat{S}_{\text{bayes}}(\pi), S_{\text{bayes}}(\pi)\right) \geq \max\left\{a_n, C_1\left(\frac{r(s_1 + s_2)\log d}{n}\right)^{1/(2-\alpha)}\right\}\right] \leq \frac{7}{2}\exp\left(-C_2 r(s_1 + s_2)\log d\right)$$

$$\leq d^{-C_3 r(s_1+s_2)},$$

where $C_1, C_2, C_3 > 0$ are constants. □

### C.5 Proofs of Theorem 3

We first suggest lemmas and proposition that will be used to prove Theorem 3.

**Lemma 1** (Multiple level-sets estimation). *Suppose $\mu(\boldsymbol{X})$ is $\alpha$-globally regular with $\alpha \in [0,1]$. Denote $t_n = \frac{r(s_1+s_2)\log d}{n}$ for $n \in \mathbb{N}_+$. Then, under the same condition and high probability specified in Theorem 2, we have*

$$d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) \lesssim t_n^{\alpha/(2-\alpha)} + \frac{1}{\rho^2(\pi, \mathcal{N})} t_n, \quad \textit{for all levels } \pi \notin \mathcal{N},$$

*where $\lesssim$ denotes the inequality up to multiplicative constants. Furthermore,*

$$\frac{1}{H}\sum_{\pi \in \Pi} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) \lesssim \frac{1}{H} + t_n^{\alpha/(2-\alpha)} + H t_n. \tag{29}$$

*Proof of Lemma 1.* The ~~global~~ $\alpha$-global regularity of $\mu(\boldsymbol{X})$ implies that, for all $\pi \notin \mathcal{N}$, the conversion inequality holds ~~,~~ by Proposition 4,

$$d_\Delta(S_{\text{bayes}}(\pi), S) \leq \lesssim \begin{cases} d_\pi^\alpha(S_{\text{bayes}}(\pi), S), & \text{if } S \in \mathrm{I}, \\ \frac{1}{\rho(\pi,\mathcal{N})} d_\pi(S_{\text{bayes}}, S), & \text{otherwise}, \end{cases} \tag{30}$$

where region ~~$\mathrm{I} = \{S : d_\Delta(S_{\text{bayes}}(\pi), S) \leq H^{-\alpha/(1-\alpha)}\}$. By proposition 4~~ $\mathrm{I} = \{S : d_\Delta(S_{\text{bayes}}(\pi), S) \leq \frac{C}{1-\alpha}\rho(\pi,\mathcal{N})^{\alpha/(1-}$

In addition, the mean-to-variance inequality also holds ~~,~~ by Proposition 4,

$$\mathrm{Var}\left[w(y)\ell(f(\boldsymbol{X})) - w(y)\ell(f_{\text{bayes}}(\boldsymbol{X}))\right] \leq \lesssim \left[R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})\right]^\alpha + \frac{1}{\rho(\pi,\mathcal{N})}\left[R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})\right].$$

Applying Theorem 2 and Lemma 6 to the above mean-to-variance relationship, we obtain the classification risk bound,

$$d_\pi(S_{\text{bayes}}(\pi), \hat{S}(\pi)) \lesssim t_n^{1/(2-\alpha)} + \frac{1}{\rho(\pi,\mathcal{N})} t_n, \quad \text{where } t_n = \frac{r(s_1+s_2)\log d}{n}.$$

Plugging the above bound into (C.5), we obtain

$$\begin{aligned} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) \leq \lesssim & \; d_\pi^\alpha(S_{\text{bayes}}(\pi), S) + \frac{1}{\rho(\pi,\mathcal{N})} d_\pi(S_{\text{bayes}}(\pi), S) \\ \lesssim & \; t_n^{\alpha/(2-\alpha)} + \frac{1}{\rho^\alpha(\pi,\mathcal{N})} t_n^\alpha + \frac{1}{\rho(\pi,\mathcal{N})} t_n^{1/(2-\alpha)} + \frac{1}{\rho^2(\pi,\mathcal{N})} t_n \\ \leq & \; 4 t_n^{\alpha/(2-\alpha)} + \frac{4}{\rho^2(\pi,\mathcal{N})} t_n. \end{aligned}$$

where the last line follows from the fact that $a(b^2 + b^{2-\alpha} + b + 1) \le 4a(b^2 + 1)$ with $a := \frac{t_n}{\rho^2(\pi,\mathcal{N})}$ and $b := \rho(\pi,\mathcal{N})t_n^{(\alpha-1)/(2-\alpha)}$. Therefore, we obtain the first conclusion.

To prove the second conclusion (1), we write

$$\frac{1}{H}\sum_{\pi\in\Pi} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) = \frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) + \frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)),$$

$$\frac{1}{H}\sum_{\pi\in\Pi} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) = \frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) + \frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)),$$

where $\mathcal{N}_H \stackrel{\text{def}}{=} \bigcup_{\pi'\in\mathcal{N}} (\pi' - \frac{1}{H}, \pi' + \frac{1}{H})$. The first term involves only finite number of summands and thus can be ~~absorbed in to $C'/H$~~bounded by $2c/H$ 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 $\pi \in \Pi \cap \mathcal{N}_H^c$,

$$\begin{aligned}
\frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} d_{\Delta,}(S_{\text{bayes}}(\pi), \hat{S}(\pi)) &\lesssim \frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} t_n^{\alpha/(2-\alpha)} + \frac{t_n}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} \frac{1}{\rho^2(\pi,\mathcal{N})} \\
&\le t_n^{\alpha/(2-\alpha)} + \frac{t_n}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c}\sum_{\pi'\in\mathcal{N}} \frac{1}{|\pi-\pi'|^2} \\
&\le t_n^{\alpha/(2-\alpha)} + \frac{t_n}{H}\sum_{\pi'\in\mathcal{N}}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} \frac{1}{|\pi-\pi'|^2} \\
&\le t_n^{\alpha/(2-\alpha)} + \frac{t_n}{H}\sum_{\pi'\in\mathcal{N}} 2H^2 \\
&\le t_n^{\alpha/(2-\alpha)} + 2C'Ht2cHt_n,
\end{aligned}$$

where the last inequality follows from the Lemma 2. Combining the bounds we obtained for the last two terms in (??) completes the second conclusion. As seen from the calculation, the distance $\rho^2(\pi,\mathcal{N})$ grows quadratically as the level $\pi$ moves away from the mass points in $\mathcal{N}$. This leads to a fast linear rate in terms of $H$, provided that there are finitly many mass points in $\mathcal{N}$. □

**Lemma 2.** *Fix a $\pi' \in \mathcal{N}$ and a sequence $\Pi = \{1/H, \ldots, (H-1)/H\}$ with $H \ge 2$. Then,*

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

*Proof.* Notice that all points $\pi \in \Pi \cap \mathcal{N}_H^c$ satisfy $|\pi - \pi'| > \frac{1}{H}$ for all $\pi' \in \mathcal{N}$.

$$\begin{aligned}
\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} \frac{1}{|\pi-\pi'|^2} &= \sum_{\frac{h}{H}\in\Pi\cap\mathcal{N}_H^c} \frac{1}{|\frac{h}{H}-\pi'|^2} \\
&\le H^2 \sum_{h=1}^{H} \frac{1}{h^2}
\end{aligned}$$

$$\leq H^2 \left\{ 1 + \int_1^2 \frac{1}{x^2} dx + \int_2^3 \frac{1}{x^2} dx + \cdots + \int_{H-1}^H \frac{1}{x^2} dx \right\}$$
$$= H^2 \left( 1 + \int_1^H \frac{1}{x^2} dx \right) \leq 2H^2,$$

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

**Proposition 4.** Let $\pi \in (0,1)$ denote a given level value, and $R_{\ell,\pi}(f) = \mathbb{E}[w(y)\ell(yf(\boldsymbol{X}))]$ denote the weighted hinge risk or $\psi$ risk for the decision function $f$. Suppose Definition 2 holds with $\alpha \in (0,1]$ and $C > 0$. Then, the following two properties hold for bounded functions $f \in \mathcal{F}(r, s_1, s_2)$.

1. Conversion inequality~~under hinge loss~~:

$$d_\Delta(S_{\text{bayes}}(\pi), S) \lesssim \begin{cases} d_\pi^\alpha(S_{\text{bayes}}(\pi), S), & \text{if } S \in \mathrm{I}, \\ \frac{1}{\rho(\pi, \mathcal{N})} d_\pi(S_{\text{bayes}}(\pi), S), & \text{otherwise}, \end{cases}$$

   where the region ~~$\mathrm{I} = \{S : d_\Delta(S_{\text{bayes}}(\pi), S) \leq H^{-\alpha/(1-\alpha)}\}$~~ $\mathrm{I} = \{S : d_\Delta(S_{\text{bayes}}(\pi), S) \leq \frac{C}{1-\alpha} \rho(\pi, \mathcal{N})^{\alpha/(1-\alpha)}\}$.

2. Mean-variance relationship:

$$\mathrm{Var}\left[w(y)\ell(f(\boldsymbol{X})) - w(y)\ell(f_{\text{bayes}}(\boldsymbol{X}))\right] \lesssim \left[R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})\right]^\alpha + \frac{1}{\rho(\pi, \mathcal{N})} \left[R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})\right].$$

*Proof.* Property 1 follows directly from the ~~excess risk bound .~~ proof in Proposition 1. For Property 2, we devide the proof into two cases: when $\ell(\cdot)$ is hinge loss and $\psi$ loss.

Case 1: When $\ell(z) = (1-z)_+$.

Property 2 follows from Lemma 3 and the boundedness of $f$. Specifically, we bound the variance using the $L$-1 distance between $f$ and $f_{\text{bayes},\pi}$,

$$\begin{aligned} \mathrm{Var}[w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))] &\leq L\mathbb{E}|\ell(yf(\boldsymbol{X})) - \ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))| \\ &\leq L\mathbb{E}|f(\boldsymbol{X}) - f_{\text{bayes},\pi}(\boldsymbol{X})| \leq LCH^\alpha R_{\ell,\pi}(f) - R_{\ell,\pi}()^\alpha \end{aligned}$$

where $L > 0$ is a constant that bounds the magnitude of $f$ in the local neighborhood of $R_{\ell,\pi}(f_{\text{bayes},\pi})$ (c.f. Assumption 1), the second line comes form the Lipschitz continuity of the hinge loss~~, and the last line comes from~~ . Applying Lemma 3 ~~.~~ on the last inequality complete the proof.

Case 2: When $\ell(z) = 2\min(1, (1-z)_+)$. Notice that

$$\begin{aligned} \mathrm{Var}\left[w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] &\leq \mathbb{E}|w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))|^2 \\ &\leq L\mathbb{E}|w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))| \end{aligned}$$

$$\lesssim L \underbrace{\mathbb{E}\left|w(y)\left(1-\operatorname{sign}(yf(\boldsymbol{X}))\right)-w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))\right|}_{=:\text{(i)}}$$
$$+L\underbrace{\mathbb{E}\left|w(y)\ell(yf(\boldsymbol{X}))-w(y)\left(1-\operatorname{sign}(yf(\boldsymbol{X}))\right)\right|}_{=:\text{(ii)}},$$

(i) is bounded as follows

$$\text{(i)}=\mathbb{E}\left[w(y)\left|\operatorname{sign}(yf(\boldsymbol{X}))-\operatorname{sign}(yf_{\text{bayes}}(\boldsymbol{X}))\right|\right]\leq 2d_{\Delta}(S_{\text{bayes}}(\pi),S)$$
$$\lesssim d_{\pi}^{\alpha}(S_{\text{bayes}}(\pi),S)+\frac{1}{\rho(\pi,\mathcal{N})}d_{\pi}(S_{\text{bayes}}(\pi),S),$$

where the last inequality is from Property 1. (ii) bounded as follows

$$\begin{aligned}\text{(ii)}&=\mathbb{E}\left[w(y)\ell(yf(\boldsymbol{X}))-w(y)\left(1-\operatorname{sign}(yf(\boldsymbol{X}))\right)\right]\\&=\mathbb{E}\left[w(y)\ell(yf(\boldsymbol{X}))-w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))\right]+\mathbb{E}\left[w(y)(\operatorname{sign}(f(\boldsymbol{X}))-\operatorname{sign}(f_{\text{bayes},\pi}(\boldsymbol{X}))\right]\\&\lesssim[R_{\ell,\pi}(f)-R_{\ell,\pi}(f_{\text{bayes}})]+2d_{\Delta}(S_{\text{bayes}}(\pi),S)\\&\lesssim[R_{\ell,\pi}(f)-R_{\ell,\pi}(f_{\text{bayes}})]+d_{\pi}^{\alpha}(S_{\text{bayes}}(\pi),S)+\frac{1}{\rho(\pi,\mathcal{N})}d_{\pi}(S_{\text{bayes}}(\pi),S),\end{aligned}$$

where the last inequality is form Property 1. Combining (5) and two bounds (i),(ii) completes the proof.

□

*Proof of Theorem 3.* Note that the proof of Theorem 1 still holds by replacing $\bar{S}(\pi)$ with $\hat{S}(\pi)$. Based on Lemma 1 and Theorem 2, with $1-C_1d^{-C_2r(s_1+s_2)}$,

$$\mathbb{E}|\mu(\boldsymbol{X})-\hat{\mu}(\boldsymbol{X})|\lesssim\frac{1}{H}+\left(\frac{r(s_1+s_2)\log d}{n}\right)^{\alpha/(2-\alpha)}+H\left(\frac{r(s_1+s_2)\log d}{n}\right)+a_n,$$

where the constants have been suppressed in the asymptotical order relationship $\lesssim$. □

## D Auxiliary lemmas

**Lemma 3** (Hinge excess loss and $L$-1 distance)**.** *Consider the same set-up as in Theorem 3. Then, the L-1 distance between $f$ and $f_{\text{bayes}}$ is bounded by their hinge excess risk; i.e,*

$$\mathbb{E}|f(\boldsymbol{X}) - f_{\text{bayes},\pi}(\boldsymbol{X})| \leq \lesssim \left[R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})\right]^{\alpha} + \frac{1}{\rho(\pi,\mathcal{N})}\left[R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})\right].$$

~~*where the region* $\mathrm{I} = \{S : d(S, S_{\text{bayes}}) \leq H^{-\alpha/(1-\alpha)}\}$.~~

*Proof.* For ease of notation, we drop the subscript $\pi$ and simply write $f_{\text{bayes}}$ and $R_\ell(f_{\text{bayes}})$ in place of $f_{\text{bayes},\pi}$ and $R_{\ell,\pi}(f_{\text{bayes},\pi})$. We also drop the random variable $\boldsymbol{X}$ in the function expression, and simply use $f, f_{\text{bayes}}, p\mu$, to represent the decision function, Bayes rule, and the probability function, respectively. The meaning should be clear given the contexts.

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

$$\begin{aligned} R_\ell(f) - R_\ell(f_{\text{bayes}}) &= \mathbb{E}[w(y)(1 - yf)_+] - \mathbb{E}[w(y)(1 - yf_{\text{bayes}})_+] \\ &= \int_{\boldsymbol{X}} w(1)p\mu(1 - f)_+ d\mathbb{P}_{\boldsymbol{X}} + \int_{\boldsymbol{X}} w(-1)(1 - p1 - \mu)(1 + f)_+ d\mathbb{P}_{\boldsymbol{X}} \\ &\quad - \int_{\boldsymbol{X}} w(1)p\mu(1 - f_{\text{bayes}})_+ d\mathbb{P}_{\boldsymbol{X}} - \int_{\boldsymbol{X}} w(-1)(1 - p1 - \mu)(1 + f_{\text{bayes}})_+ d\mathbb{P}_{\boldsymbol{X}}, \end{aligned} \tag{31}$$

where $w(1) = 1 - \pi$ and $w(-1) = \pi$.

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

- Region I ~~$= \{\boldsymbol{X} : p < \pi \text{ and } f \geq -1\}$~~ $= \{\boldsymbol{X} : \mu < \pi \text{ and } f \geq -1\}$. In this region, $f_{\text{bayes}} = -1$, and the integrant in (D) reduces to

$$\begin{aligned} \Phi_{\mathrm{I}} &:= (1 - \pi)p\mu(1 - f)_+ + \pi(1 - p1 - \mu)(1 + f)_+ - 2(1 - \pi)p\mu \\ &\geq (1 - \pi)p\mu(1 - f) + \pi(1 - p1 - \mu)(1 + f) - 2(1 - \pi)p\mu \\ &= (f + 1)(\pi - p - \mu) = |f - f_{\text{bayes}}||\pi - p - \mu|. \end{aligned}$$

- Region II ~~$= \{\boldsymbol{X} : p < \pi \text{ and } f < -1\}$~~ $= \{\boldsymbol{X} : \mu < \pi \text{ and } f < -1\}$. In this region, $f_{\text{bayes}} = -1$, and the integrant in (D) reduces to

$$\Phi_{\mathrm{II}} := (1 - \pi)p\mu(1 - f) - 2(1 - \pi)p\mu = (-1 - f)(p - p\mu - \mu\pi) = |f - f_{\text{bayes}}|(1 - \pi)p\mu.$$

- Region III ~~$= \{\boldsymbol{X} : p \geq \pi \text{ and } f \leq 1\}$~~ $= \{\boldsymbol{X} : \mu \geq \pi \text{ and } f \leq 1\}$. In this region, $f_{\text{bayes}} = 1$, and

the integrant in (D) reduces to

$$\begin{aligned}\Phi_{\mathrm{III}} &:= (1-\pi)p\mu(1-f)_+ + \pi(1-p1-\mu)(1+f)_+ - 2\pi(1-p1-\mu)\\ &\geq (1-\pi)p\mu(1-f) + \pi(1-p1-\mu)(1+f) - 2\pi(1-p1-\mu)\\ &= (1-f)(p-\mu-\pi) = |f-f_{\mathrm{bayes}}||\pi-p-\mu|\end{aligned}$$

- Region IV ~~$= \{\boldsymbol{X} : p \geq \pi \text{ and } f > 1\}$~~ $= \{\boldsymbol{X} : \mu \geq \pi \text{ and } f > 1\}$. In this region, $f_{\mathrm{bayes}} = 1$, and the integrant in (D) reduces to

$$\Phi_{\mathrm{IV}} := \pi(1-p1-\mu)(1+f) - 2\pi(1-p1-\mu) = (f-1)(\pi-p-\mu\pi) = |f-f_{\mathrm{bayes}}|(1-p1-\mu)\pi$$

Therefore, the integral is evaluated as

$$\begin{aligned}R_\ell(f) - R_\ell(f_{\mathrm{bayes}}) &= \int_{\mathrm{I}} \Phi_{\mathrm{I}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\mathrm{II}} \Phi_{\mathrm{II}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\mathrm{III}} \Phi_{\mathrm{III}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\mathrm{IV}} \Phi_{\mathrm{IV}} d\mathbb{P}_{\boldsymbol{X}}\\ &\geq \mathbb{E}|f-f_{\mathrm{bayes}}||\pi-p-\mu|\mathbb{1}(|f|\leq 1) + \mathbb{E}|f-f_{\mathrm{bayes}}|(\pi-p-\mu\pi)\mathbb{1}(f>1)\\ &\quad + \mathbb{E}|f-f_{\mathrm{bayes}}|(p-p\mu-\mu\pi)\mathbb{1}(f<-1).\end{aligned}$$

Note that the functions ~~$|\pi-p|, (\pi-p\pi), (p-p\pi)$~~ $|\pi-\mu|, (\pi-\mu\pi), (\mu-\mu\pi)$ are non-negative, $[0,1]$-valued, and satisfy (4) by ~~Assumption 2. Applying~~ $\alpha$-global regularity (local regularity follows the same argument). Now we use Lemma 4 to bound the three terms in the last equation. To make sure that each term has the same order in Lemma 4~~to the above integral gives~~

$$\begin{aligned}R_\ell(f) - R_\ell(f_{\mathrm{bayes}}) &\geq CH^{-1}[\mathbb{E}|f-f_{\mathrm{bayes}}|\mathbb{1}(|f|\leq 1)]^{1/\alpha} + C[\mathbb{E}|f-f_{\mathrm{bayes}}|\mathbb{1}(f>1)]^{1/\alpha}\\ &\quad + C[\mathbb{E}|f-f_{\mathrm{bayes}}|\mathbb{1}(f<-1)]^{1/\alpha}\\ &\geq CH^{-1}[\mathbb{E}|f-f_{\mathrm{bayes},\pi}|]^{1/\alpha},\end{aligned}$$

, we rescale the functions $|f-f_{\mathrm{bayes}}|\mathbb{1}(|f|\leq 1), |f-f_{\mathrm{bayes}}|\mathbb{1}(f>1), |f-f_{\mathrm{bayes}}|\mathbb{1}(f<-1)$ to have expectation 1 and set upper bound of the functions as

$$L = \max\left(\||f-f_{\mathrm{bayes}}|\mathbb{1}(|f|\leq 1)\|_\infty, \||f-f_{\mathrm{bayes}}|\mathbb{1}(f>1)\|_\infty, \||f-f_{\mathrm{bayes}}|\mathbb{1}(f<-1)\|_\infty\right).$$

Then, the last inequality is bounded either Case 1 or Case 2 from Lemma 4:

Case 1:

$$R_\ell(f) - R_\ell(f_{\mathrm{bayes}}) \gtrsim [\mathbb{E}|f-f_{\mathrm{bayes}}|\mathbb{1}(|f|\leq 1)]^{1/\alpha} + [\mathbb{E}|f-f_{\mathrm{bayes}}|\mathbb{1}(f>1)]^{1/\alpha} + [\mathbb{E}|f-f_{\mathrm{bayes}}|\mathbb{1}(f<-1)]^{1/\alpha}$$

$$\gtrsim [\mathbb{E}|f - f_{\text{bayes},\pi}|]^{1/\alpha},$$

where the last inequality uses the property that $x^{1/\alpha} + y^{1/\alpha} \geq (x+y)^{1/\alpha}$ for $x, y \geq 0$ and $\alpha \in (0, 1]$. ~~Therefore, we conclude that~~

$$\mathbb{E}|f - f_{\text{bayes}}| \leq C' H^{\alpha} [R_{\ell}(f) - R_{\ell}(f_{\text{bayes}})]^{\alpha},$$

Case 2:

$$R_{\ell}(f) - R_{\ell}(f_{\text{bayes}}) \gtrsim \rho(\pi, \mathcal{N}) \left[\mathbb{E}\left(|f - f_{\text{bayes}}| \mathbb{1}(|f| \leq 1) + |f - f_{\text{bayes}}| \mathbb{1}(f > 1) + |f - f_{\text{bayes}}| \mathbb{1}(f < -1)\right)\right]$$

$$= \rho(\pi, \mathcal{N}) \mathbb{E}|f - f_{\text{bayes},\pi}|.$$

~~where $C' > 0$ is a constant.~~

Therefore, combining two cases gives

$$\mathbb{E}|f(\boldsymbol{X}) - f_{\text{bayes},\pi}(\boldsymbol{X})| \lesssim [R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})]^{\alpha} + \frac{1}{\rho(\pi, \mathcal{N})} [R_{\ell,\pi}(f) - R_{\ell,\pi}(f_{\text{bayes}})].$$

□

**Lemma 4** (Expectation of function products)**.** *Let $\boldsymbol{X} \in \mathcal{X}$ be a random variable. Let $g \colon \mathcal{X} \to [0, 1]$ be a function taking values on $[0, 1]$ and satisfying*

$$\mathbb{P}[g(\boldsymbol{X}) \leq t] \leq C(Ht) C t^{\alpha/1-\alpha}, \quad \text{for all } t \in (0, \delta\rho], \tag{32}$$

*for some constants $\alpha \in (0, 1]$ and ~~$t_* \in (0, 1], C > 0$~~ $C \geq 0$. (When $\alpha = 1$, the right hand side of (4) is interpreted as being zero.) Then, for all nonnegative, bounded functions $f \colon \mathcal{X} \to \mathbb{R}_{\geq 0}$ with $\|f\|_{\infty} \leq L$, the ~~expected $f(\boldsymbol{X})$ is upper bounded by the $\alpha$-th power of the expected $f(\boldsymbol{X})g(\boldsymbol{X})$; i.e.,~~ following holds true,*

$$\mathbb{E}f() \leq C H^{\alpha} \mathbb{E}f()g()^{\alpha}, \mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] \gtrsim \begin{cases} \rho \mathbb{E} f(\boldsymbol{X}), & \text{if } \frac{\mathbb{E}f(\boldsymbol{X})}{L} \geq \frac{C}{1-\alpha} \rho^{\alpha/1-\alpha}, \\ [\mathbb{E} f(\boldsymbol{X})]^{1/\alpha}, & \text{if } \frac{\mathbb{E}f(\boldsymbol{X})}{L} < \frac{C}{1-\alpha} \rho^{\alpha/1-\alpha}. \end{cases}$$

~~*where $C' > 0$ is a constant.*~~

**Remark 1.** Roughly speaking, Lemma 4 bounds the function $f$ by the function product $fg$ in expectation, provided that the multiplier $g$ has controlled probability mass around zero. Note that we always have the lower bound $\mathbb{E}f(\boldsymbol{X})g(\boldsymbol{X}) \leq \mathbb{E}f(\boldsymbol{X})$ because of the boundedness of $g$.

*Proof.* Let $t > 0$ be an arbitrary number in the interval ~~$(0, t_*]$~~ $(0, \rho]$. We bound the expectation of

the product $fg$ using the integral over the region ~~$\{\boldsymbol{X}: g(\boldsymbol{X}) > t/H\}$~~, $\{\boldsymbol{X}: g(\boldsymbol{X}) \geq t\}$,

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] &\geq \mathbb{E}[f(\boldsymbol{X})\mathbb{1}(g(\boldsymbol{X}) > t/H)]\mathbb{E}[f(\boldsymbol{X})\mathbb{1}(g(\boldsymbol{X}) > t)] \\
&= tH\mathbb{E}[f(\boldsymbol{X}) - f(\boldsymbol{X})\mathbb{1}(g(\boldsymbol{X}) \leq t/H)]\mathbb{E}[f(\boldsymbol{X}) - f(\boldsymbol{X})\mathbb{1}(g(\boldsymbol{X}) \leq t)] \qquad (33) \\
&\geq tH\left(\mathbb{E}f(\boldsymbol{X}) - L\mathbb{P}[g(\boldsymbol{X}) \leq t/H]\mathbb{P}[g(\boldsymbol{X}) \leq t]\right) \\
&\geq 1Ht\mathbb{E}f(\boldsymbol{X}) - CLt^{1/1-\alpha}, \quad \text{for all } t \in (0, t_*\rho],
\end{aligned}$$

where the third line uses the fact that $f(\boldsymbol{X}) \in [0, L]$, and the last line follows from (4). We maximize the lower bound (D) with respect to ~~$t \in (0, t_*]$~~ $t \in (0, \rho]$ and obtain the optimal ~~$t_{\text{opt}} \in (0, t_*]$~~, $t_{\text{opt}} \in (0, \rho]$,

$$t_{\text{opt}} = \begin{cases} \rho, & \text{if} \quad \mathbb{E}f(\boldsymbol{X}) \geq \frac{CL}{1-\alpha}\rho^{\alpha/1-\alpha}, \\ \left[\frac{1-\alpha}{CL}\mathbb{E}f(\boldsymbol{X})\right]^{(1-\alpha)/\alpha}, & \text{if} \quad \mathbb{E}f(\boldsymbol{X}) < \frac{CL}{1-\alpha}\rho^{\alpha/1-\alpha}. \end{cases}$$

The corresponding lower bound of the inequality (D) becomes

$$\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] \geq \begin{cases} \alpha\rho\mathbb{E}f(\boldsymbol{X}), & \text{if} \quad \mathbb{E}f(\boldsymbol{X}) \geq \frac{CL}{1-\alpha}\rho^{\alpha/1-\alpha}, \\ \alpha\left(\frac{1-\alpha}{CL}\right)^{(1-\alpha)/\alpha}[\mathbb{E}f(\boldsymbol{X})]^{1/\alpha}, & \text{if} \quad \mathbb{E}f(\boldsymbol{X}) < \frac{CL}{1-\alpha}\rho^{\alpha/1-\alpha}. \end{cases}$$

~~Hence, we conclude~~

$$\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] \geq \frac{C_2}{H}\min\left\{\mathbb{E}f(\boldsymbol{X}),\ [\mathbb{E}f(\boldsymbol{X})]^{1/\alpha}\right\},$$

~~where $C_2 > 0$ is a constantindependent of $f$. When $\mathbb{E}f(\boldsymbol{X}) \leq 1$, the right hand side of the above inequality takes value $C_2[\mathbb{E}(f(\boldsymbol{X})]^{1/\alpha}/H$. When $\mathbb{E}f(\boldsymbol{X}) > 1$, we rescale $f$ by $f/L$ and obtain~~

$$\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] \geq \frac{C_2}{H}L^{-(\alpha-1)/\alpha}[\mathbb{E}f(\boldsymbol{X})]^{1/\alpha}.$$

~~Combining both cases gives~~ Since, $\alpha, C$, and $L$ are given constant, we have the desired results.

□

**Definition 2** (Bracketing number)**.** Consider a function set $\mathcal{F}$, and let $\varepsilon > 0$. We call $\{(f_m^l, f_m^u)\}_{m=1}^M$ an $L_2$-metric, $\varepsilon$-bracketing function set of $\mathcal{F}$, if for every $f \in \mathcal{F}$, 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, $\mathcal{H}_{[\,]}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_2)$, is defined as the logarithm of the smallest

cardinality of the $\varepsilon$-bracketing function set of $\mathcal{F}$.

**Lemma 5** (Bracketing number for bounded functions in $\mathcal{F}(r, s_1, s_2)$)**.** *Let $\mathcal{F}(r, s_1, s_2)$ denote the classifier functions in (3.2), where we assume the intercept $b = 0$ is known and that the function domain satisfies $\mathbb{P}(\|\boldsymbol{X}\|_F \leq 1) = 1$. For any given $k \in \mathbb{N}_+$, consider the subset of functions in $\mathcal{F}(r, s_1, s_2)$ with magnitudes bounded by $k$, denoted by $\mathcal{F}^k = \{f \in \mathcal{F}(r, s_1, s_2) \colon \|f\|_F^2 \leq k\}$. Then, there exists a constant $C > 0$ such that*

$$\mathcal{H}_{[\,]}(\varepsilon,\ \mathcal{F}^k,\ \|\cdot\|_2) \leq Cr(s_1 + s_2) \log \frac{kd}{\varepsilon}.$$

*Proof.* For any given $k \in \mathbb{N}_+$, define the matrix class

$$\mathcal{B} = \{\boldsymbol{B} \in \mathbb{R}^{d\times d} \colon \operatorname{rank}(\boldsymbol{B}) \leq r,\ \operatorname{supp}(\boldsymbol{B}) \leq (s_1, s_2),\ \|\boldsymbol{B}\|_F^2 \leq k\}.$$

Based on the assumption of known $b$, there is an one-to-one correspondence between functions in $\mathcal{F}^k$ and matrices in $\mathcal{B}$,

$$\mathcal{F}^k = \{f \colon \boldsymbol{X} \mapsto \langle \boldsymbol{X}, \boldsymbol{B}\rangle + b \;\big|\; \boldsymbol{B} \in \mathcal{B}\}.$$

Furthermore, every pair of two functions $f_1 = \langle \boldsymbol{X}, \boldsymbol{B}_1\rangle,\ f_2 = \langle \boldsymbol{X}, \boldsymbol{B}_2\rangle \in \mathcal{F}(r, s_1, s_1)$ satisfies the norm relationship

$$\|f_1 - f_2\|_2 \leq \|f_1 - f_2\|_\infty = \sup_{\|\boldsymbol{X}\|_F \leq 1} |\langle \boldsymbol{X}, \boldsymbol{B}_1\rangle - \langle \boldsymbol{X}, \boldsymbol{B}_2\rangle| \leq \|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F.$$

Based on **?**, Theorem 9.23), the $L_2$-metric, $(2\varepsilon)$-bracketing number in $\mathcal{F}^k$ is bounded by

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

where $\mathcal{H}$ denotes the log covering number for the (non-bracketing) set. Therefore, it suffices to bound $\mathcal{H}(\varepsilon,\ \mathcal{B},\ \|\cdot\|_F)$. 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 **?**, Lemma 3.1), the log covering number for $\mathcal{B}_{S_1,S_2}$ is

$$\mathcal{H}\left(\varepsilon,\ \mathcal{B}_{S_1,S_2},\ \|\cdot\|_F\right) \leq r(s_1 + s_2 + 1) \log\left(\frac{9\sqrt{k}}{\varepsilon}\right).$$

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 $\varepsilon$-covering set $\mathcal{B}$ is then given by the union of $\varepsilon$-covering set of $\mathcal{B}_{S_1,S_2}$. Using Stirling's bound, we derive that

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

where $C, C' > 0$ are constants. □

**Lemma 6** (Local complexity of $\mathcal{F}(r, s_1, s_2)$)**.** *Define* $\mathcal{F}^k = \{f \in \mathcal{F}(r, s_1, s_2) \colon \|f\|_F^2 \le k\}$ *for all* $k \in \mathbb{N}_+$*; i.e.,* $\mathcal{F}^k$ *is the subset of functions in* $\mathcal{F}(r, s_1, s_2)$ *with magnitudes bounded by* $k$*. Set*

$$t_n \asymp \left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)} + \frac{1}{\rho(\pi,\mathcal{N})}\left(\frac{r(s_1+s_2)\log d}{n}\right) \text{ and } \lambda_n \asymp \frac{1}{J_n}\left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)}.$$

*Then, the following inequality is satisfied for all* $k \in \{2, 3, \ldots\}$*,*

$$\frac{1}{L_n}\int_{L_n}^{\sqrt{L_n^\alpha + \frac{1}{\rho(\pi,\mathcal{N})}L_n}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon,\ \mathcal{F}^k,\ \|\cdot\|_2)}d\varepsilon \le n^{1/2}, \quad \text{where } L_n := t_n + \lambda_n J_n(k/2-1).$$

*Proof.* To simplify the notation, we denote $L = t + \lambda J_n(k/2-1) > 0$, $\rho = \rho(\pi,\mathcal{N})$, and define

$$g(L,k) = \frac{1}{L}\int_L^{\sqrt{L^\alpha + \rho^{-1}L}} \sqrt{r(s_1+s_2)\log\left(\frac{kd}{\varepsilon}\right)}d\varepsilon, \quad \text{for all } k \in \{2,3,\ldots\},$$

where we have inserted the bracketing number based on Lemma 5. Notice that

$$\begin{aligned} g(L,k) &\le \frac{\sqrt{r(s_1+s_2)}}{L}\int_L^{\sqrt{L^\alpha+\rho^{-1}L}} \sqrt{\log\left(\frac{kd}{L}\right)}d\varepsilon \\ &\le \sqrt{r(s_1+s_2)(\log k + \log d - \log L)}\left(\frac{\sqrt{L^\alpha}+\sqrt{\rho^{-1}L}}{L} - 1\right) \\ &\le \sqrt{r(s_1+s_2)(\log k + \log d)}\left(\frac{1}{L^{(2-\alpha)/2}} + \frac{1}{\sqrt{\rho L}}\right), \end{aligned} \tag{34}$$

where the second line follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$. It remains to verify that $g(L_n, k) \le n^{1/2}$ for all $k \in \{2, 3, \ldots, \}$, where

$$L_n = t_n + \lambda_n J_n(k/2-1) = (k/2+C)\left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)} + \frac{1}{\rho}\left(\frac{r(s_1+s_2)\log d}{n}\right),$$

for some universal constant $C > 0$. Plugging $L_n$ into the last line of (D) gives

$$g(L_n,k) \le n^{1/2}\sqrt{\frac{\log k + \log d}{(k/2)^{(2-\alpha)}\log d}} + n^{1/2}\sqrt{\frac{\log k + \log d}{\log d}} \le C' n^{1/2}, \quad \text{for all } k \in \{2,3,\ldots\},$$

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