Contents

A Additional theoretical results

A.1 Sign rank and matrix rank

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

Example 1 (Max graphon)**.** Suppose the matrix $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$ takes the form

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

Then

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

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

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

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

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

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

In fact, Example 1 is a special case of the following proposition.

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

$$\mathbf{\Theta}(i,j) = g(\max(x_i, y_j)), \quad (i,j) \in [d]^2. \tag{1}$$

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

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

The same conclusion holds if we use min in place of max in (1).

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

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

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

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

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

  Therefore, $\operatorname{srank}(\boldsymbol{\Theta}) \leq \operatorname{rank}(\operatorname{sgn}(\boldsymbol{\Theta})) = 2$.
- When $r \geq 2$. By continuity, the function $g(z)$ is non-zero and remains an unchanged sign in each of the intervals (z_s, z_{s+1}) for $1 \leq s \leq r-1$. Define the index set

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

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

$$\begin{aligned}\{(i,j) \colon \boldsymbol{\Theta}(i,j) < 0\} &= \{(i,j) \colon g(\max(x_i, y_j)) < 0\} \\ &= \cup_{s \in \mathcal{I}} \{(i,j) \colon \max(x_i, y_j) \in (z_s, z_{s+1})\} \\ &= \cup_{s \in \mathcal{I}} \Big(\{(i,j) \colon x_i < z_{s+1}, y_j < z_{s+1}\} \cap \{(i,j) \colon x_i \leq z_s, y_j \leq z_{s+1}\}^c \Big). \quad (2)\end{aligned}$$

  The equation (2) is equivalent to

$$\mathbb{1}(\boldsymbol{\Theta}(i,j) < 0) = \sum_{s \in \mathcal{I}} \left(\mathbb{1}(x_i < z_{s+1}) \mathbb{1}(y_j < z_{s+1}) - \mathbb{1}(x_i \leq z_s) \mathbb{1}(y_j \leq z_s) \right), \quad (3)$$

  for all $(i,j) \in [d]^2$, where $\mathbb{1}(\cdot) \in \{0,1\}$ denotes the indicator function. The equation (3) implies the low-rank representation of $\operatorname{sgn}(\boldsymbol{\Theta})$,

$$\operatorname{sgn}(\boldsymbol{\Theta}) = 1 - 2 \sum_{s \in \mathcal{I}} \left(\boldsymbol{a}_{s+1} \otimes \boldsymbol{b}_{s+1} - \bar{\boldsymbol{a}}_s \otimes \bar{\boldsymbol{b}}_s \right), \quad (4)$$

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

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

  and $\boldsymbol{b}_{s+1}, \bar{\boldsymbol{b}}_s$ are binary vectors defined similarly by using y_j in place of x_i. Therefore, by (4) and

the assumption $|\mathcal{I}| \leq r-1$, we conclude that

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

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

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

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

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

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

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

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

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

$$\boldsymbol{A} = \boldsymbol{b} \otimes \operatorname{rev}(\boldsymbol{b}) + \operatorname{rev}(\boldsymbol{b}) \otimes \boldsymbol{b}. \tag{5}$$

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

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

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

$$\boldsymbol{A}(i,j) \geq \boldsymbol{A}(i',j'), \quad \text{for all } |i-j| \geq |i'-j'|. \tag{6}$$

Notice that for a given $\pi \in \mathbb{R}$, there exists $\pi' \in \mathbb{R}$ such that $\operatorname{sgn}(\boldsymbol{A} - \pi') = \operatorname{sgn}(\boldsymbol{M} - \pi)$. This is because both $\boldsymbol{A}$ and $\boldsymbol{M}$ are banded matrices satisfying monotonicity (6). By definition (5), $\boldsymbol{A}$ is a rank-2 matrix. Henceforce, $\operatorname{srank}(\boldsymbol{M} - \pi) = \operatorname{srank}(\boldsymbol{A} - \pi') \leq 3$. □

Example 3 (Identity matrices)**.** Let $\boldsymbol{I}$ be a d-by-d identity matrix. Then

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

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

1. $\mathrm{sgn}(\boldsymbol{I} - \pi)$ is a matrix of all 1, or of all -1;
2. $\mathrm{sgn}(\boldsymbol{I} - \pi) = 2\boldsymbol{I} - \mathbf{1}_d \otimes \mathbf{1}_d$.

The former case is trivial, so it suffices to show $\mathrm{srank}(\boldsymbol{I} - \pi) \leq 3$ in the second case. Based on Example 2, the rank-2 matrix $\boldsymbol{A}$ in (5) satisfies

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

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

A.2 Adjusting for additional covariates

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

A.3 Extension to sub-Gaussian noise

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

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

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

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

Assumption 1 (Sub-Gaussian noise)**.**

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

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

Theorem A.3.1 (Extension of Theorem 6 to sub-Gaussian noise)**.** *Consider the same conditions of Theorem 6. Under Assumption 1, with high probability over training data $\boldsymbol{Y}_\Omega$, we have*

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

$$\mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi, \mathrm{sgn}(\boldsymbol{\Theta} - \pi)) \lesssim \left(\frac{r\sigma^2 d \log d}{|\Omega|}\right)^{\frac{\alpha+1}{\alpha+2}} + \frac{1}{\rho(\pi, \mathcal{N})}\left(\frac{r\sigma^2 d \log d}{|\Omega|}\right). \tag{7}$$

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

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

The proof is provided in Section B.6.

Remark 1 (Extending to general non-parametric matrix regression)**.** For matrix nonparametrix regression (Theorem 4 of the main paper), the extension of bounded noise to sub-Gaussian noise incurs an additional $\log n$ factor, where n is the sample size. The techniques of handling sub-Gaussian noise is identical to the above extension, and is thus omitted in the paper.

A.4 Extension to unbounded number of mass points

Theorem 6 of our main paper assumes the bounded $|\mathcal{N}|_{\text{cover}} < c < \infty$ for some constant $c > 0$, where $|\mathcal{N}|_{\text{cover}}$ is defined as the covering number of $\mathcal{N}$ with $2\Delta s$-bin's. Recall that $\mathcal{N}$ corresponds to regions of jumps greater than $\Delta s = 1/d^2$ in the CDF $G(\pi) = \mathbb{P}_{\omega\sim\Pi}(\Theta(\omega) \leq \pi)$. This setup gives a cleaner exposition of our results but may be restricted in some cases. For example, the high-rank matrices in Example 5 and Figure 1(b) are excluded, because $\alpha = \infty$ and $|\mathcal{N}|_{\text{cover}} = d$ in this setup. Fortunately, our framework still applies to this family of matrices with a little amendment.

We now extend the setup to allow for more general structured matrices including those in Example 5. Redefine $\Delta s = 1/d$. Correspondingly, redefine the smoothness index α and the set $\mathcal{N}$ for the psudo density of $\boldsymbol{\Theta}(\omega)$ with new bin width $2\Delta s$. Let $|\mathcal{N}|_{\text{cover}}$ be the covering number of $\mathcal{N}$ with new

$2\Delta s$-bin's. Under this new setup, the signal matrix in Example 5 has $|\mathcal{N}|_{\text{cover}} = 0$ and $\alpha < \infty$. Following the same line as in Theorem 6 and use the fact that $\Delta s \lesssim t_n$, we obtain that

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

Therefore, setting $H \asymp \sqrt{\frac{1+|\mathcal{N}|_{\text{cover}}}{t_d}}$ yields the error bound

$$\text{MAE}(\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \leq \tilde{\mathcal{O}} \left\{ \left(\frac{dr}{|\Omega|} \right)^{\min(\frac{\alpha}{2+\alpha}, \frac{1}{2})} \right\}. \tag{8}$$

The result (8) applies to cases when the signal matrices belong to $\mathcal{M}_{\text{sgn}}(r)$ and have at most d distinct entries with repetition patterns.

A.5 Connection to structured matrix model with functional coefficients

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

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

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

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

Specifically, we consider a structured matrix model with functional coefficients

$$\boldsymbol{X}_\pi \overset{\text{def}}{=} \boldsymbol{B}_0 + \sum_{s=1}^r g_s(\pi)\boldsymbol{B}_s + \sigma \boldsymbol{E}, \quad Y_\pi \sim \text{Ber}(\pi), \quad \boldsymbol{X}_\pi \perp Y_\pi|\pi, \tag{9}$$

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

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

Proposition 2 (Connection to structured matrix model with functional coefficients)**.** Let $\mathbb{P}_{\boldsymbol{X},Y}$ denote the joint distribution induced by $(\boldsymbol{X}_\pi, Y_\pi)_{\pi \in [0,1]}$ drawn from from (9). In the noiseless case $\sigma = 0$, let $f(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$ denote the regression function based on $\mathbb{P}_{\boldsymbol{X},Y}$. Then $f \in \mathcal{F}_{\text{sgn}}(r, s_1, s_2)$.

Proof. We restrict ourselves to the noiseless case with $\sigma = 0$ in (9). Let

$$\mathcal{X} = \{\boldsymbol{X}_\pi \colon \boldsymbol{X}_\pi \text{ has structure specified in (9) for } \pi \in [0,1]\}$$

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

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

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

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

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

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

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

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

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

Remark 2. The above result shows the connection of our method to joint matrix model (9) $(\boldsymbol{X}_\pi, Y_\pi)_{\pi \in [0,1]}$. We should point out, despite of the seeming similarity, a fundamental challenge arises in our setting when the latent index π is unobserved. Our level-set approach essentially learns the right ordering of $\boldsymbol{X}_\pi$ against the index $\pi \in [0,1]$ (see Figure 2 of the main paper), thereby facilitating the estimation of regression function f.

B Proofs

B.1 Main notation

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

B.2 Proof of Theorem 1

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

$$\begin{aligned}
&\mathrm{Risk}_\pi(\mathrm{sgn}\phi) - \mathrm{Risk}_\pi(\mathrm{sgn}(f-\pi)) \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{X}} \underbrace{\mathbb{E}_{Y|\boldsymbol{X}} \left\{ |Y-\pi| \left[|\mathrm{sgn}(Y-\pi) - \mathrm{sgn}\phi| - |\mathrm{sgn}(Y-\pi) - \mathrm{sgn}(f-\pi)| \right] \right\}}_{\stackrel{\mathrm{def}}{=} I}. \qquad (10)
\end{aligned}$$

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

$$\begin{aligned}
I &= \mathbb{E}_{Y|\boldsymbol{X}} \left[(Y-\pi)(\mathrm{sgn}(f-\pi) - \mathrm{sgn}\phi)\mathbb{1}(Y \geq \pi) + (\pi - Y)(\mathrm{sgn}\phi - \mathrm{sgn}(f-\pi))\mathbb{1}(Y < \pi) \right] \\
&= \mathbb{E}_{Y|\boldsymbol{X}} \left[(\mathrm{sgn}(f-\pi) - \mathrm{sgn}\phi)(Y-\pi) \right] \\
&= \left[\mathrm{sgn}(f-\pi) - \mathrm{sgn}\phi \right] \left[f - \pi \right] \\
&= |\mathrm{sgn}(f-\pi) - \mathrm{sgn}\phi||f-\pi|, \qquad (11)
\end{aligned}$$

where the third line uses the fact $\mathbb{E}_{Y|\boldsymbol{X}} Y = f(\boldsymbol{X})$. Combining (11) with (10), we conclude that, for all $\phi \in \Phi(r)$,

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

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

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

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

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

□

B.3 Proof of Theorem 2

Proof. Fix $\pi \in [-1, 1]$. For ease of notation, we drop the dependence of π in $S_{\mathrm{bayes}}(\pi)$ and simply write S_{bayes}. Based on (11) in the proof of Theorem 1, we have

$$\begin{aligned}
d_\pi(S, S_{\mathrm{bayes}}) &\stackrel{\mathrm{def}}{=} \mathrm{Risk}_\pi(\mathrm{sgn}(S)) - \mathrm{Risk}_\pi(\mathrm{sgn}(S_{\mathrm{bayes}})) \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{X}} \left(|\mathrm{sgn}(S) - \mathrm{sgn}(S_{\mathrm{bayes}})| \, |\pi - f| \right) \\
&= \int_{\boldsymbol{X} \in S \Delta S_{\mathrm{bayes}}} |f(\boldsymbol{X}) - \pi| d\mathbb{P}_{\boldsymbol{X}}. \qquad (12)
\end{aligned}$$

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

Case 1: $\alpha > 0$.

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

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

where the last inequality is from α-globally smoothness condition. Combining the above inequality with the identity (12) yields

$$d_{\pi}(S, S_{\text{bayes}}) \geq t \left(\mathbb{P}_{\boldsymbol{X}}\left(S \Delta S_{\text{bayes}}\right) - C t^{\alpha}\right), \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{13}$$

We maximize the lower bound of (13) with respect to t, and obtain the optimal t_{opt},

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

The corresponding lower bound of the inequality (13) becomes

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

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

$$d_{\Delta}(S, S_{\text{bayes}}) \stackrel{\text{def}}{=} \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}) \lesssim \left[d_{\pi}(S, S_{\text{bayes}})\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})} d_{\pi}(S, S_{\text{bayes}}), \tag{14}$$

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

Case 2: $\alpha = \infty$.

The inequality (13) now becomes

$$d_{\pi}(S, S_{\text{bayes}}) \geq t \mathbb{P}_{\boldsymbol{X}}(S \Delta S_{\text{bayes}}) = t d_{\Delta}(S, S_{\text{bayes}}), \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{15}$$

The conclusion (14) follows by taking $t = \frac{\rho(\pi, \mathcal{N})}{2}$ in the inequality (15).

□

Remark 3 (Bounding L_1 distance by classification risk)**.** The bound controls the L_1 distance to $f_{\text{bayes},\pi} = \text{sgn}(f - \pi)$ using the classification excess risk to $\text{Risk}_{\pi}(f_{\text{bayes},\pi})$. The result applies

uniformly to $\pi \in [-1,1]$ if f is globally-α smooth; i.e., the bound

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

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

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

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

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

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

  where ϕ^T is a truncation of function ϕ (see formal definition in (22)).

See Lemma 4.

B.4 Proofs of Theorem 3, Part (a) in Theorems 5 and 7

We provide a unified framework that incorporates Theorem 3, Part (a) in Theorems 5 and 7 in the main paper. For any $\pi \in [-1,1]$, write $\bar{Y}_\pi = Y - \pi$, and let $\ell_{\pi,F}(\phi;(\boldsymbol{X},Y))$ denote the weighted F-loss

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

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

$$\hat{\phi}_{\pi,F} = \mathop{\arg\min}_{\phi \in \Phi(r,s_1,s_2)} \left\{ \frac{1}{n}\sum_{i=1}^n \ell_{\pi,F}(\phi;(\boldsymbol{X}_i,Y_i)) + \lambda\|\phi\|_F^2 \right\}. \tag{16}$$

The following theorem states the accuracy for sign function estimate $\mathrm{sgn}(\hat{\phi}_{\pi,F})\colon \mathcal{X} \to \{-1,1\}$. We say that an event E occurs "with high probability" if $\mathbb{P}(E)$ tends to 1 as the dimension $d \to \infty$.

Theorem B.4.1 (Large-margin sign estimation)**.** *Fix $\pi \notin \mathcal{N}$. Suppose the regression function $f \in \mathcal{F}_{\mathrm{sgn}}(r,s_1,s_2)$ is (π,α)-smooth over $\mathcal{X}$. Then, with high probability at least $1-\exp(-nt_n)$ over training data $(\boldsymbol{X}_i,Y_i)_{i\in[n]}$, the estimate (16) satisfies*

$$\|\mathrm{sgn}\hat{\phi}_{\pi,F} - f_{\mathrm{bayes},\pi}\|_1 \lesssim t_n^{\alpha/(2+\alpha)} + \frac{1}{\rho^2(\pi,\mathcal{N})}t_n, \tag{17}$$

under the following three specifications:

(a) (Theorem 3) 0-1 loss $F(z) = \mathbb{1}(z > 0)$*, no penalization* $\lambda = 0$*,* $(s_1, s_2) = (d_1, d_2)$*, and* $t_n = \frac{1}{n} r d_{\max}$.

(b) (Theorem 5) 0-1 loss $F(z) = \mathbb{1}(z > 0)$*, no penalization* $\lambda = 0$*, constant* (s_1, s_2)*, and* $t_n = \frac{1}{n} r (s_1 + s_2) \log d_{\max}$.

(c) (Theorem 7) Surrogate loss satisfying Assumption 1, constant (s_1, s_2)*,* $t_n = \frac{1}{n} r (s_1 + s_2) \log d_{\max}$*, penalization* $\lambda \asymp (t_n^{(\alpha+1)/(\alpha+2)} + \rho t_n)$*, approximation error* $a_n^{(\alpha+1)/(\alpha+2)} \leq t_n$.

Here, the constants absorbed in the $\lesssim$ *of* (17) *are independent of* π.

Remark 4 (Ridge penalization)**.** The estimation under 0-1 loss requires no penalization, because only the sign, but not the magnitude, of ϕ affects the 0-1 risk. One can constrain $\|\phi\|_F = 1$ in the empirical 0-1 risk minimization without altering the solution. In contrast, the surrogate loss such as as hinge loss is scale-sensitive, rending the possible unboundedness of ϕ. We impose penalization to control the magnitude of the $\|\phi\|_F$ and thus the local complexity. The resulting estimation enjoys the fast convergence as in sieve estimate [Shen and Wong, 1994] under well tuned λ.

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

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

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

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

The following result summarizes the key properties of four common losses: 0-1 loss, hinge loss, T-truncated hinge loss, and psi-loss. Here, the T-truncated hinge loss is defined as $F(z) = \min((1 - z)_+, T)$ for a given $T \geq 2$. We will use T-truncated hinge loss to faciliate the proofs of Lemma 2 and Theorem B.4.1.

Lemma 1 (Conversion inequalities)**.** *Suppose the regression function* f *is* (π, α)*-smooth, and denote*

$f_{\text{bayes},\pi} = \text{sgn}(f-\pi)$ *for* $\pi \in [-1,1]$. *Let* F *be 0-1 loss, hinge loss,* T*-truncated hinge loss, or psi-loss. Then, the following three properties hold for all* $\pi \in [-1,1]$.

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

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

$$\text{Risk}_{\pi}(\phi) - \text{Risk}_{\pi}(f_{\text{bayes},\pi}) \le C\left[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right], \tag{18}$$

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

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

$$\begin{aligned}
&\text{Var}\left[\ell_{\pi,F}(\phi;(\boldsymbol{X},Y)) - \ell_{\pi,F}(f_{\text{bayes},\pi};(\boldsymbol{X},Y))\right] \\
\lesssim\ &\left[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right]^{\alpha/(1+\alpha)} + \frac{1}{\rho(\pi,\mathcal{N})}\left[\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})\right].
\end{aligned} \tag{19}$$

Remark 5. The property (c) holds only for bounded loss functions, i.e, excluding hinge loss.

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

Properties (a) and (b) directly follow from Theorem 1. To prove (c), we expand the variance by

$$\begin{aligned}
\text{Var}\left[\ell_{\pi}(\phi;(\boldsymbol{X},Y)) - \ell_{\pi}(f_{\text{bayes},\pi},(\boldsymbol{X},Y)\right] &\lesssim \mathbb{E}|\ell_{\pi}(\phi;(\boldsymbol{X},Y)) - \ell_{\pi}(f_{\text{bayes},\pi},(\boldsymbol{X},Y)|^2 \\
&\lesssim \mathbb{E}|\ell_{\pi}(\phi;(\boldsymbol{X},Y)) - \ell_{\pi}(f_{\text{bayes},\pi},(\boldsymbol{X},Y)| \\
&\lesssim \mathbb{E}\left||\text{sgn}\bar{Y}_{\pi} - \text{sgn}\phi(\boldsymbol{X})| - |\text{sgn}\bar{Y}_{\pi} - f_{\text{bayes},\pi}(\boldsymbol{X})|\right| \\
&\le \mathbb{E}|\text{sgn}\phi - f_{\text{bayes},\pi}|,
\end{aligned} \tag{20}$$

where the second line comes from the boundedness of 0-1 loss, and the third line comes from the boundedness of weight $|\bar{Y}_{\pi}|$, and fourth line comes from the inequality $||a-b| - |c-b|| \le |a-b|$ for $a,b,c \in \{-1,1\}$. Here we have absorbed the constant multipliers in $\lesssim$. Therefore, the conclusion (19) then directly follows by applying Remark 3 to (20).

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

Property (a) was firstly introduced in Wang et al. [2008, Lemma 1], and here we provide an alternative proof. A direct calculation (see Lemma 4) shows that

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

Therefore, $\inf_{\text{all}\ \phi} \text{Risk}_{\pi,F}(\phi) = \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})$. Property (18) is from Scott [2011, Corollary 1].

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

Again, the property (a) follows from Wang et al. [2008, Lemma 1]. For the property (18), we

use Scott [2011, Theorem 1] to find the transformation function ψ that relates 0-1 risk to F-risk:

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

To put our problem in the context of Scott [2011], we need additional notation. For any function measurable $g\colon x \mapsto g(x)$, we write $g = g^+ - g^{-1}$, where g^+ and g^- are two non-negative functions given by

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

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

Define the conditional F-risk

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

A direct calculation shows that

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

Therefore, following the notation of Scott [2011], we have

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

Applying Scott [2011, Theorem 1] to the above setup gives the excess risk transformation rule: $\psi : z \to 2|z|$. Therefore, the property (18) is proved.

To prove (19), notice that

$$\begin{aligned} &\mathrm{Var}\left\{|\bar{Y}_\pi|\left[F(\phi(\boldsymbol{X})\mathrm{sgn}\bar{Y}_\pi) - F(f_{\mathrm{bayes},\pi}(\boldsymbol{X})\mathrm{sgn}\bar{Y}_\pi)\right]\right\} \\ &\lesssim \mathbb{E}|\bar{Y}_\pi||F(\phi(\boldsymbol{X})\mathrm{sgn}\bar{Y}_\pi) - F(f_{\mathrm{bayes},\pi}(\boldsymbol{X})\mathrm{sgn}\bar{Y}_\pi)| \\ &\lesssim \underbrace{\mathbb{E}\left|1 - \mathrm{sgn}(\phi(\boldsymbol{X})\bar{Y}_\pi) - F(f_{\mathrm{bayes},\pi}(\boldsymbol{X})\mathrm{sgn}\bar{Y}_\pi)\right|}_{=:\text{(i)}} + \underbrace{\mathbb{E}|\bar{Y}_\pi|\left|F(\phi(\boldsymbol{X})\mathrm{sgn}\bar{Y}_\pi) - \left(1 - \mathrm{sgn}(\phi(\boldsymbol{X})\bar{Y}_\pi)\right)\right|}_{=:\text{(ii)}} \end{aligned} \tag{21}$$

The first term (i) is bounded as follows

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

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

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

where the first equality is based on $F(z) = 1 - \operatorname{sgn}(z)$ if $z = 1$ or -1, and the last inequality is from definition of $d_\pi(\cdot,\cdot)$. Notice we have $d_\pi(S_\phi, S_{\text{bayes}}(\pi)) = \text{Risk}_\pi(\phi) - \text{Risk}_\pi(f_{\text{bayes},\pi})$ by definition. Therefore, the proof is complete by combining (21), (18) and bounds (i)-(ii).

Case 4: $F(z) = \min((1-z)_+, T)$ for T-truncated hinge loss, for given $T \geq 2$. A direct calculation (c.f. Remark 6 after Lemma 4) shows that

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

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

$$\phi^T = \begin{cases} T-1 & \text{if } \phi > T-1, \\ \phi, & \text{if } |\phi| \leq T-1, \\ -(T-1), & \text{if } \phi < -(T-1). \end{cases} \tag{22}$$

Therefore, $\inf_{\text{all }\phi} \text{Risk}_{\pi,F}(\phi) = \text{Risk}_{\pi,F}(f_{\text{bayes},\pi})$. To show property (18), we again use Scott [2011, Theorem 1] to find the transformation function ψ that relates 0-1 risk to F-risk:

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

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

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

Therefore, following the notation of Scott [2011], we have

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

Applying Scott [2011, Theorem 1] to the above setup gives the excess risk transformation rule: $\psi : z \to |z|$. Therefore, the property (18) is proved.

To prove (19), we use Lemma 4 and the boundedness condition of $\|F\|_\infty \le T$. Specifically, we bound the variance using the L-1 distance between ϕ and $f_{\text{bayes},\pi}$,

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

where $T > 0$ is the upper bound of truncated hinge loss , the second inequality comes from the boundedness of the T-truncated hinge loss, and the last line comes from 1-Lipschitz continuity and the definition of F. Applying Remark 6 in Lemma 4 on the last inequality complete the proof.

□

Below we establish the estimation convergence rate for $\hat{\phi}_{\pi,F}$'s excess F-risk. The variance-to-mean relationship in Lemma 1 plays a key role in determining the convergence rate based on empirical process theory [Shen and Wong, 1994]. The proof adopts the local iterative techniques from Wang et al. [2008, Theorem 3]. Similar techniques have been used in Bartlett et al. [2006, Theorem 4] for similar estimate but without ridge penalization.

Lemma 2 (Excess F-risk error)**.** *Consider the set-up as in Theorem B.4.1. Then, with high probability and t_n specified in Theorem B.4.1, the following holds for all $\pi \notin \mathcal{N}$.*

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

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

(b) If F is hinge loss, then

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

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

Proof of Lemma 2. In Theorem 3 and 5, we have $\lambda = a_n = 0$. We consider the most general case $\lambda, a_n \ge 0$ as in Theorem 7. We first consider the bounded loss (0-1 loss or psi-loss).The modification for unbounded loss (hinge loss) incurs only slight difference in the proof.

Fix $\pi \notin \mathcal{N}$, and write $\rho = \rho(\pi, \mathcal{N})$. For any function $\phi \in \Phi(r, s_1, s_2)$ of consideration, define

empirical weighted F-risk

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

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

$$\hat{\phi}_{\pi,F} = \operatorname*{arg\,min}_{\phi\in\Phi(r,s_1,s_2)} \left\{\widehat{\mathrm{Risk}}_{\pi,F}(\phi) + \lambda\|\phi\|_K^2\right\}. \tag{23}$$

We are interested in the convergence of $\hat{\phi}_{\pi,F}$'s excess risk, $\mathrm{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi})$.

Let L_n denote the convergence rate to seek. By the definition of $\hat{\phi}_{\pi,F}$, we have

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

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

$$\begin{aligned}
&\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \mathrm{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \ge 2L_n\right\} \\
\subset &\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \exists\phi\in\Phi(r,s_1,s_2), \text{s.t. } \mathrm{Risk}_{\pi,F}(\phi;(\boldsymbol{X},Y)) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \ge 2L_n,\right. \\
&\qquad\qquad \left.\text{and } \widehat{\mathrm{Risk}}_{\pi,F}(\hat{\phi}_{\pi,F}) + \lambda\|\hat{\phi}_{\pi,F}\|_K^2 \le \widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) + \lambda J\right\} \\
\subset &\left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \sup_{\substack{\phi\in\Phi(r,s_1,s_2)\\ \mathrm{Risk}_{\pi,F}(\phi;(\boldsymbol{X},Y))-\mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi})\ge 2L_n}} \left[\widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) - \widehat{\mathrm{Risk}}_{\pi,F}(\phi)\right] \ge \lambda\|\phi\|_K^2 - \lambda J\right\} \\
\subset &\bigcup_{\phi\in A_{s,k}} \left\{(\boldsymbol{X}_i, Y_i)_{i\in[n]} \colon \sup_{\phi\in A_{s,k}} \left[\widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) - \widehat{\mathrm{Risk}}_{\pi,F}(\phi)\right] \ge \lambda\|\phi\|_K^2 - \lambda J\right\}.
\end{aligned} \tag{24}$$

In the last line of (24), we have partitioned the set $\{\phi\in\Phi(r,s_1,s_2)\colon \mathbb{E}\Delta(\phi;(\boldsymbol{X},Y)) \ge L_n\}$ into a union of $A_{s,k}$, with

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

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

Let Γ denote the target probability for the first line in (24). To bound Γ, it suffices to bound the sum of probabilities over sets $A_{s,k}$. For each $A_{s,k}$, we consider the scaled empirical process,

$$\begin{aligned}
v_n(\phi) &:= \left[\widehat{\mathrm{Risk}}_{\pi,F}(\phi_\pi^{(n)}) - \widehat{\mathrm{Risk}}_{\pi,F}(\phi)\right] - \left[\mathrm{Risk}_{\pi,F}(\phi_\pi^{(n)}) - \mathrm{Risk}_{\pi,F}(\phi)\right] \qquad (25)\\
&= \frac{1}{n}\sum_{i}^{n}\left(\ell_{\pi,F}(\phi_\pi^{(n)};(\boldsymbol{X}_i,Y_i)) - \ell_{\pi,F}(\phi;(\boldsymbol{X}_i,Y_i)) - \mathbb{E}\left[\ell_{\pi,F}(\phi_\pi^{(n)};(\boldsymbol{X}_i,Y_i)) - \ell_{\pi,F}(\phi;(\boldsymbol{X}_i,Y_i))\right]\right).
\end{aligned}$$

Notice that

$$\begin{aligned}\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(\phi_\pi^{(n)}) &= \text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) + \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) - \text{Risk}_{\pi,F}(\phi_\pi^{(n)}) \\ &\geq (s+1)L_n - a_n \\ &\geq sL_n, \end{aligned} \tag{26}$$

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

Combining the definition of v_n in (25) and (26) gives (24) as

$$\begin{aligned}\Gamma &\leq \sum_{s,k=1}^{\infty} \mathbb{P}\left\{ \sup_{\phi \in A_{s,k}} v_n(\phi) \geq sL_n + \lambda \|\phi\|_K^2 - \lambda J \right\} \\ &\leq \sum_{s,k=1}^{\infty} \mathbb{P}\left\{ \sup_{\phi \in A_{s,k}} v_n(\phi) \geq sL_n + \lambda(k-2)J =: M(s,k) \right\}, \end{aligned} \tag{27}$$

where $M(s,k) > 0$ for all $s, k \in \mathbb{N}$ from the condition $\lambda J \leq L_n/2$.

The variance of the empirical process is bounded by,

$$\begin{aligned}&\sup_{A_{s,k}} \text{Var}\left[\ell_{\pi,F}(\phi_\pi^{(n)}; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)\right] \qquad (28) \\ &\leq \sup_{A_{s,k}} 2\left\{ \text{Var}\left[\ell_{\pi,F}(\phi_\pi^{(n)}; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(f_{\text{bayes},\pi}; (\boldsymbol{X}, Y)\right] + \text{Var}\left[\ell_{\pi,F}(\phi; (\boldsymbol{X}, Y)) - \ell_{\pi,F}(f_{\text{bayes},\pi}; (\boldsymbol{X}, Y)\right] \right\} \\ &\lesssim [M(s,k)]^{\alpha/(1+\alpha)} + \frac{M(s,k)}{\rho} =: V(s,k), \end{aligned}$$

where the last inequality is from Lemma 1.

We next bound the right-hand-side of (27) choosing an L_n that satisfies the conditions in Shen and Wong [1994, Theorem 3]. (The specification of L_n is deferred to the next paragraph). Once such L_n is chosen, then it follows from Shen and Wong [1994, Theorem 3] that

$$\begin{aligned}\Gamma &\lesssim \sum_{s,k} \exp\left(-\frac{nM^2(s,k)}{V(s,k) + M(s,k)} \right) \\ &\lesssim \sum_{s,k} \exp(-\rho n M(s,k)) \\ &= \sum_{s,k} \exp\left(-n\rho s L_n - n\rho J\lambda(k-2)\right) \\ &\leq \left(\frac{e^{-n\rho L_n}}{1 - e^{-n\rho L_n}} \right) \left(\frac{e^{n\rho\lambda J}}{1 - e^{n\rho\lambda J}} \right) \leq e^{-n\rho L_n/2}. \end{aligned} \tag{29}$$

where the last inequality uses the condition $\lambda J \leq L_n/2$. Notice that this bound for Γ has a sub-exponential decay in L_n.

Now, we specify an L_n that satisfies the condition of Shen and Wong [1994, Theorem 3]. The convergence rate $L_n > 0$ is determined by the solution to the following inequality,

$$\sup_{k\geq 2} \frac{1}{x} \int_x^{\sqrt{x^{\alpha/(\alpha+1)}+\rho^{-1}x}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \Phi^k, \|\cdot\|_2)} d\varepsilon \lesssim n^{1/2}, \quad \text{where } x = L_n + \lambda J(k-2). \tag{30}$$

In particular, the smallest L_n satisfying (30) yields the best upper bound of the error rate. Here $\mathcal{H}_{[\,]}(\varepsilon, \Phi^k, \|\cdot\|_2)$ denotes the L_2-norm, ε-bracketing number (c.f. Definition 1) for function family Φ^k, and, we have denoted $\Phi^k = \{\phi \in \Phi(r, s_1, s_2) \colon \|\phi\|_K^2 \leq k\}$, i.e., the subset of functions in $\Phi(r, s_1, s_2)$ with magnitudes bounded by k, for $k \in \mathbb{N}_+$.

It remains to solve for the smallest possible L_n in (30). Based on Lemma 6, the inequality (30) is satisfied with

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho} t_n, \quad \text{and} \quad \lambda \asymp \frac{L_n}{J}, \tag{31}$$

where

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

Plugging (31) into (29) gives that,

$$\begin{aligned} \Gamma &= \mathbb{P}\left[\mathrm{Risk}_{\pi,F}(\hat{\phi}_{\pi,F}) - \mathrm{Risk}_{\pi,F}(f_{\mathrm{bayes},\pi}) \geq L_n\right] \\ &\lesssim \exp(-n\rho L_n) \lesssim \exp(-n t_n). \end{aligned}$$

where the last inequality uses the fact that $\rho L_n \geq t_n$. The proof is then complete by bounding the 0-1 risk by F-risk.

For unbounded loss (hinge loss), we seek to bound the F'-risk of $\hat{\phi}_{\pi,F}$, where F' is T-truncated version of F. The general strategy is to evaluate $\hat{\phi}_{\pi,F}$'s error using F'-risk. Note that the estimate $\hat{\phi}_{\pi,F}$ (23) is defined under unbounded loss F. Therefore, the inclusion (24) changes to

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

where the second line is newly added, and the last line comes from the fact that

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

setting $T = \max(2, J) > \max(2, \sup_n \|\phi_\pi^{(n)}\|_K)$, which is a constant. The remaining proof follows a similar line of argument as the bounded case. In particular, we invoke the mean-to-variance relationship for bounded F'-loss in (28). The final conclusion follows from the excess bound inequality for T-truncated risk (c.f. Lemma 1). □

Proof of Theorem B.4.1. Write $\rho = \rho(\pi, \mathcal{N})$. Combining Theorem 2 and Lemma 2 gives

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

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

B.5 Proofs of Theorem 4, Part (b) in Theorems 5 and 7

Proof of Theorem 4. It follows from the definition of $\hat{f}$ that

$$\begin{aligned}
\|\hat{f} - f\|_1 &= \mathbb{E}\left|\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} \mathrm{sgn}\hat{\phi}_\pi - f\right| \\
&\le \mathbb{E}\left|\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} (\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi))\right| + \mathbb{E}\left|\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} \mathrm{sgn}(f-\pi) - f\right| \\
&\le \frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} \|\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)\|_1 + \frac{1}{H},
\end{aligned} \tag{32}$$

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

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

It suffices to bound the first term in (32). We shall prove that, with probability at least $1 - \exp(-nt\log H)$,

$$\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} \|\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)\|_1 \le t^{\alpha/(\alpha+2)} + \frac{1}{H} + Ht, \tag{33}$$

for any $t \ge t_n$, where t_n is specified in Theorem B.4.1.

For any $t > t_n$, define the union of events

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

We first bound the probability of the event E, and then show E implies (33). Based on Theorem B.4.1 and union bound over $\pi \in \mathcal{H}$, we have that, for any $t \ge t_n$,

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

One the other hand, we now show that the E implies (33). Theorem B.4.1 shows that the sign function accuracy depends on the closeness of $\pi \in \mathcal{H}$ to the mass points in $\mathcal{H}$. Therefore, we partition the level set $\pi \in \mathcal{H}$ based on their closeness to $\mathcal{H}$. Specifically, let $\mathcal{N}_H \stackrel{\text{def}}{=} \bigcup_{\pi'\in\mathcal{N}} \left(\pi' - \frac{1}{H}, \pi' + \frac{1}{H}\right)$ denote the set of levels at least $\frac{1}{H}$-close to the mass points. We expand left hand side of (33) by

$$\begin{aligned}&\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}} \|\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)\|_1 \\ &= \frac{1}{2H+1}\sum_{\pi\in\mathcal{H}\cap\mathcal{N}_H} \|\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)\|_1 + \frac{1}{2H+1}\sum_{\pi\in\mathcal{H}\cap\mathcal{N}_H^c} \|\mathrm{sgn}\hat{\phi}_\pi - \mathrm{sgn}(f-\pi)\|_1. \qquad (34)\end{aligned}$$

By assumption, the first term of (34) involves only finite number of summands and thus can be bounded by $4C/(2H+1)$ where $C > 0$ is a constant such that $|\mathcal{N}| \le C$. We bound the second term using the explicit forms of $\rho(\pi,\mathcal{N})$ in the sequence $\pi \in \Pi \cap \mathcal{N}_H^c$.

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

where the first inequality uses the property of E, and the last inequality follows from Lemma 3. Combining the bounds for the last two terms in (34) completes the proof for inequality (33). Finally, plugging (33) into (32) yields

$$\mathbb{P}\left(\|\hat{f} - f\|_1 \lesssim t^{\alpha/(\alpha+2)} + \frac{1}{H} + tH \right) \ge 1 - \exp(-nt\log H),$$

for any $t \ge t_n$. In particular, taking $t = t_n \log H$ gives

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

with high probability at least $1-\exp(-nt)$. Setting $H \asymp t_n^{-1/2}$ yields the desired upper bound. □

Lemma 3. *Fix $\pi' \in \mathcal{N}$ and a sequence $\mathcal{H} = \{-1, \dots, -1/H, 0, 1/H, \dots, 1\}$ with $H \geq 2$. Then,*

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

Proof of Lemma 3. Notice that all points $\pi \in \mathcal{H} \cap \mathcal{N}_H^c$ satisfy $|\pi - \pi'| > \frac{1}{H}$ for all $\pi' \in \mathcal{N}$. We use this fact to compute the sum

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

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

B.6 Proofs of Theorem 6 and Theorem A.3.1

Proof of Theorem 6. Theorem 6 follows from the same line of proof as in Theorem 4, with slight modification to account for discrete measure space. For any matrix $\boldsymbol{Z} \in \mathbb{R}^{d \times d}$, we use $f_{\boldsymbol{Z}} \colon [d]^2 \to \mathbb{R}$ to denote the function induced by matrix $\boldsymbol{Z}$ such that $f_{\boldsymbol{Z}}(\omega) = \boldsymbol{Z}(\omega)$ for $\omega \in [d]^2$. Set $\mathcal{X} = \{\boldsymbol{e}_i \otimes \boldsymbol{e}_j \colon (i,j) \in [d]^2\}$ be the discrete feature space, and $n = |\Omega|$ the sample size. Under this set up, $\|\hat{f} - f\|_1 = \mathbb{E}_{\boldsymbol{X}} |\hat{f}(\boldsymbol{X}) - f(\boldsymbol{X})| = \mathbb{E}_\omega |\hat{\boldsymbol{\Theta}}(\omega) - \boldsymbol{\Theta}(\omega)| = \mathrm{MAE}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})$. Notice that the small tolerance $\Delta s = 1/d^2$ in the pseudo density is dominated by the derived convergence rate. Applying Theorem 4 to this setting finishes the proof. □

Proof of Theorem A.3.1. By setting $s = \log(d_{\max})$ in Lemma 7, we have

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

We divide the sample space into two exclusive events:

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

Because the Event I occurs with probability tending to zero, we restrict ourselves to the Event II only by following the proof of Theorem 3. We summarize the key difference compared to Section 4.

Let $\ell_\omega(\cdot)$ denote the 0-1 loss evaluated at the ω-th value of matrix. We expand the variance by

$$\begin{aligned}\mathrm{Var}\left[\ell_\omega\left(\boldsymbol{Z}, \bar{\boldsymbol{Y}}_\Omega\right) - \ell_\omega\left(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{Y}}_\Omega\right)\right] &\le \mathbb{E}|\ell_\omega(\boldsymbol{Z}(\omega), \bar{\boldsymbol{Y}}(\omega)) - \ell_\omega(\bar{\boldsymbol{\Theta}}(\omega), \bar{\boldsymbol{Y}}(\omega))|^2 \\ &= \mathbb{E}|\bar{\boldsymbol{Y}}(\omega) - \bar{\boldsymbol{\Theta}}(\omega) + \bar{\boldsymbol{\Theta}}(\omega)|^2 |\mathrm{sgn}\boldsymbol{Z}(\omega) - \mathrm{sgn}\bar{\boldsymbol{\Theta}}(\omega)| \\ &\le 2\left(4\sigma^2 \log d_{\max} + 2\right) \mathbb{E}|\mathrm{sgn}\boldsymbol{Z} - \mathrm{sgn}\bar{\boldsymbol{\Theta}}| \\ &\lesssim (\sigma^2 \log d_{\max}) \mathrm{MAE}(\mathrm{sgn}\boldsymbol{Z}, \mathrm{sgn}\bar{\boldsymbol{\Theta}}), \end{aligned} \tag{35}$$

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

Based on (35), the (α, π)-smoothness of $\boldsymbol{\Theta}$ implies that for all measurable functions $f_{\boldsymbol{Z}}$, we have

$$\mathrm{Var}\Delta_i(f_{\boldsymbol{Z}}, \bar{\boldsymbol{Y}}) \lesssim \left(\sigma^2 \log d_{\max}\right) \left\{ \left[\mathbb{E}\Delta_i(f_{\boldsymbol{Z}}, \bar{\boldsymbol{Y}})\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho}\mathbb{E}\Delta_i(f_{\boldsymbol{Z}}, \bar{\boldsymbol{Y}}) \right\}. \tag{36}$$

The empirical process with variance-to-mean relationship (36) gives that

$$\mathbb{P}\left(\mathrm{Risk}(\hat{\boldsymbol{Z}}) - \mathrm{Risk}(\bar{\boldsymbol{\Theta}}) \ge L_n\right) \lesssim \exp(-nt_n), \tag{37}$$

where the convergence rate L_n is obtained by the same way in the proof of Lemma 6,

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho} t_n, \quad \text{with } t_n = \frac{r\sigma^2 d_{\max} \log d_{\max}}{n}. \tag{38}$$

Combining (37) and (38), we obtain that, with high probability,

$$\mathrm{Risk}(\hat{\boldsymbol{Z}}) - \mathrm{Risk}(\bar{\boldsymbol{\Theta}}) \lesssim \left(\frac{r\sigma^2 d_{\max} \log d_{\max}}{|\Omega|}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})} \left(\frac{r\sigma^2 d_{\max} \log d_{\max}}{|\Omega|}\right), \tag{39}$$

where constants have been absorbed into the $\lesssim$ relationship. Therefore, combining (39) and (12) completes the proof for sign matrix estimation error in (7). The signal estimation error follows the same proof of Theorem 4. □

C Auxiliary lemmas

Lemma 4 (Hinge loss and L-1 distance)**.** *Consider the same set-up as in Theorem 7. Let $F(z) = (1-z)_+$ be the hinge loss. Then, the L-1 distance between ϕ and $f_{\text{bayes},\pi}$ is bounded by their excess risk; i.e,*

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

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

Remark 6. With little modification in the proof, similar inequality also holds for T-truncated hinge loss $F(z) = \min(T, (1-z)_+)$ with $T \geq 2$. Specifically,

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

where $\phi^T\colon \mathcal{X} \to [-(T-1),\ (T-1)]$ is the truncated ϕ defined in (22).

Proof of Lemma 4. For ease of notation, we drop the random variable $\boldsymbol{X}$ in the function expression, and simply use ϕ, $f_{\text{bayes},\pi}$, f, to represent the trace function, Bayes rule, and the regression function, respectively. The meaning should be clear given the contexts.

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

$$\begin{aligned}
&\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) \\
&= \mathbb{E}[|\bar{Y}_\pi|(1 - \phi \text{sgn}\bar{Y}_\pi)_+] - \mathbb{E}[|\bar{Y}_\pi|(1 - f_{\text{bayes},\pi}\text{sgn}\bar{Y}_\pi)_+] \\
&= \int_{\boldsymbol{X}} (1-\phi)_+ \int_{y>\pi} (y-\pi) dy d\mathbb{P}_{\boldsymbol{X}} + \int_{\boldsymbol{X}} (1+\phi)_+ \int_{y\leq\pi} (\pi - y) dy d\mathbb{P}_{\boldsymbol{X}} \\
&\quad - \int_{\boldsymbol{X}} (1 - f_{\text{bayes},\pi})_+ \int_{y>\pi} (y-\pi) dy d\mathbb{P}_{\boldsymbol{X}} - \int_{\boldsymbol{X}} (1 + f_{\text{bayes},\pi})_+ \int_{y\leq\pi} (\pi - y) dy d\mathbb{P}_{\boldsymbol{X}}. \qquad (40)
\end{aligned}$$

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

- Region I $= \{\boldsymbol{X}\colon f < \pi \text{ and } \phi \geq -1\}$. In this region, $f_{\text{bayes},\pi} = -1$, and the integrant in (40) reduces to

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

- Region II $= \{\boldsymbol{X}\colon f < \pi \text{ and } \phi < -1\}$. In this region, $f_{\text{bayes},\pi} = -1$, and the integrant in (40) reduces to

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

- Region III $= \{\boldsymbol{X}\colon f \geq \pi \text{ and } \phi \leq 1\}$. In this region, $f_{\text{bayes},\pi} = 1$, and the integrant in (40) reduces to

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

- Region IV $= \{\boldsymbol{X}\colon f \geq \pi \text{ and } \phi > 1\}$. In this region, $f_{\text{bayes},\pi} = 1$, and the integrant in (40) reduces to

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

Therefore, the integral is evaluated as

$$\begin{aligned}\text{Risk}_{\pi,F}(\phi) - \text{Risk}_{\pi,F}(f_{\text{bayes},\pi}) &= \int_{\text{I}} \Phi_{\text{I}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\text{II}} \Phi_{\text{II}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\text{III}} \Phi_{\text{III}} d\mathbb{P}_{\boldsymbol{X}} + \int_{\text{IV}} \Phi_{\text{IV}} d\mathbb{P}_{\boldsymbol{X}} \\ &\geq \mathbb{E}|\phi - f_{\text{bayes},\pi}||f-\pi|. \end{aligned} \tag{41}$$

Note that the function $|f-\pi|$ is α-smooth. Using the same techniques as in Theorem 2 to the last line of (41), we conclude

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

□

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

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

and

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

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

Lemma 5 (Bracketing number for bounded functions in $\Phi(r,s_1,s_2)$)**.** *Let $\Phi(r,s_1,s_2)$ denote the collection of trace functions,*

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

Assume, for simplicity, that the intercept b is known and that the function domain satisfies $\mathbb{P}\left(\|\boldsymbol{X}\|_F \leq 1\right) = 1$. For any given $k \in \mathbb{N}_+$, consider the subset of functions in $\Phi(r,s_1,s_2)$ with magnitudes bounded

by k, denoted by $\Phi^k = \{f \in \Phi(r, s_1, s_2) \colon \|f\|_F^2 \le k\}$. Then, there exists a constant $C > 0$ such that

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

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

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

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

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

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

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

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

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

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

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

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

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

where $C, C' > 0$ are constants. □

Lemma 6 (Local complexity of $\Phi(r, s_1, s_2)$)**.** *Define $\Phi^k = \{f \in \Phi(r, s_1, s_2) \colon \|f\|_F^2 \le k\}$ for all*

$k \in \mathbb{N}_+$; *i.e.,* Φ^k *is the subset of functions in* $\Phi(r, s_1, s_2)$ *with magnitudes bounded by* k. *Set*

$$L_n \asymp \left(\frac{r(s_1+s_2)\log d}{n}\right)^{\frac{\alpha+1}{\alpha+2}} + \frac{1}{\rho(\pi,\mathcal{N})}\left(\frac{r(s_1+s_2)\log d}{n}\right), \quad \text{and } \lambda \asymp \frac{L_n}{J}.$$

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

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

Proof. To simplify the notation, we write $\rho = \rho(\pi, \mathcal{N})$, and define

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

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

$$\begin{aligned} g(x,k) &\le \frac{\sqrt{r(s_1+s_2)}}{L}\int_x^{\sqrt{x^{\alpha/(\alpha+1)} + \rho^{-1}x}} \sqrt{\log\left(\frac{kd}{x}\right)}d\varepsilon \\ &\le \sqrt{r(s_1+s_2)(\log k + \log d - \log x)}\left(\frac{\sqrt{x^{\alpha/(2\alpha+2)}} + \sqrt{\rho^{-1}x}}{x} - 1\right) \\ &\le \sqrt{r(s_1+s_2)(\log k + \log d)}\left(\frac{1}{x^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho x}}\right), \end{aligned} \tag{42}$$

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

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

for some universal constant $C > 0$. Plugging the above expression into the last line of (42) gives

$$g(L_n + \lambda J(k/2-1), k) \le n^{1/2}\sqrt{\frac{\log k + \log d}{(k/2)^{(\alpha+2)/(\alpha+1)}\log d}} + n^{1/2}\sqrt{\frac{\log k + \log d}{(k/2)\log d}} \le C' n^{1/2},$$

for all $k \in \{2, 3, \ldots\}$, where $C' > 0$ is a constant independent of k and d. The proof is therefore complete. □

Theorem C.0.1 (Theorem 3 in Shen and Wong [1994]). *Let* $\mathcal{F}$ *be a class of functions defined on* $\mathcal{X}$ *with* $\sup_{f\in\mathcal{F}}\|f\|_\infty \le T$. *Let* $(\boldsymbol{X}_i)_{i=1}^n$ *be i.i.d. random variables with distribution* $\mathbb{P}_{\boldsymbol{X}}$ *over* $\mathcal{X}$. *Set* $\sup_{f\in\mathcal{F}}\mathrm{Var} f(\boldsymbol{X}) = V < \infty$. *Define the empirical process* $\hat{\mathbb{E}}f = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i)$. *Define* x_n^* *be the solution of the equation to the following equation*

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

Suppose

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

Then we have

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

Remark 7. Both upper bounds T and V are allowed to depend on n. By applying Theorem C.0.1 to $\mathcal{G} = \{g\colon g = \mathbb{E}f - f, f \in \mathcal{F}\}$, we obtain the other side inequality,

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

Lemma 7 (sub-Gaussian maximum)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian zero-mean random variables with variance proxy σ^2. Then, for any $s > 0$,*

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

Proof. The conclusion follows from

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

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

References

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

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

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

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

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

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