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 **??** from two classes of $\boldsymbol{X}$ to a series of $\boldsymbol{X}$ on a continuous spectrum of π.

Example 1 (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 $N(0,1)$; σ 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., $\sum_{s\in[r]} \operatorname{supp}(\boldsymbol{B}_s) \leq (s_1, s_2)$; and the coefficients $g_s(\cdot)\colon [0,1] \to \mathbb{R}$ are monotonic functions with respect to π 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_π is conditionally independent of $\boldsymbol{X}_\pi$ given π, and $\pi \in [0,1]$ is drawn from an unknown distribution over $[0,1]$. Equivalently, the generative model is expressed in the following scheme:

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

where $\boldsymbol{X}_{pq} \overset{\text{indep.}}{\sim} \mathcal{N}(g_s(\pi), \sigma^2)$ 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 $\operatorname{sign}(\mu(\boldsymbol{X}) - \pi) = \operatorname{sign} f_\pi(\boldsymbol{X})$ 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 π is unobserved. Our level-set approach essentially learns the right "sorting" of $\boldsymbol{X}_\pi$ against the index $\pi \in [0,1]$ (see Figure **??**), 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 1 (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 $\operatorname{sign}(\mu(\boldsymbol{X}) - \pi) = \operatorname{sign}(f_\pi(\boldsymbol{X}))$ 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 π.

Proof of Proposition 1. 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 $S^c(\pi) = \{\boldsymbol{X} : \mu(\boldsymbol{X}) < \pi\}$.

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

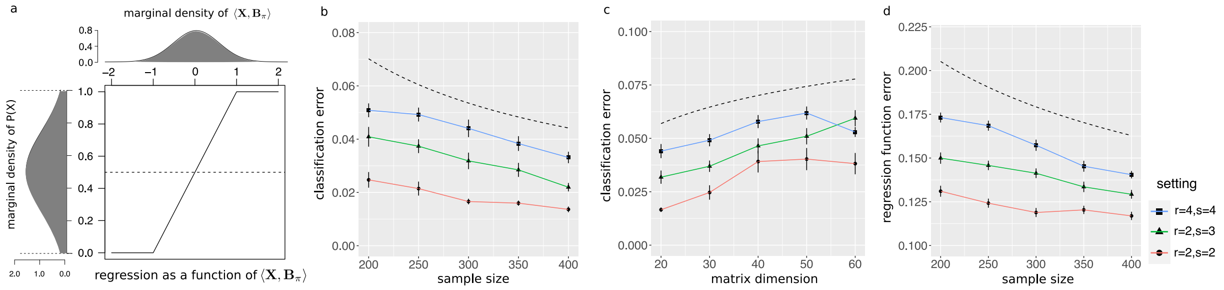
(C1): $S_1(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$ and $S_2(\pi) = \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) < \pi\}$.

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

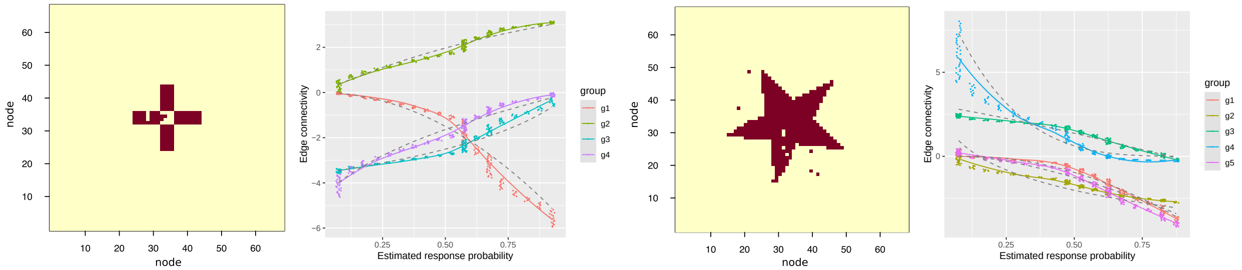
Suppose $S_1(\pi) \subset \{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$. Since $\{\partial S(\pi), S_1(\pi), S_2(\pi)\}$ is a partition of $\mathbb{R}^{d_1 \times d_2}$ and both $\{\boldsymbol{X} \colon \mu(\boldsymbol{X}) > \pi\}$ and $\{\boldsymbol{X} \colon \mu(\boldsymbol{X}) < \pi\}$ are 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 ??

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

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

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

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

Combining the above inequality with the identity (1) yields

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

We maximize the lower bound of (2) with respect to t and obtain the optimal $0 \leq t_{\text{opt}} < \rho(\pi,\mathcal{N})$,

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

The corresponding lower bound of the inequality (2) becomes

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

Combining both cases gives

$$d_\Delta(S, S_{\text{bayes}}) \overset{\text{def}}{=} \mathbb{P}(S\Delta S_{\text{bayes}}) \leq c\left(d^\alpha(S, S_{\text{bayes}}) + \frac{1}{\rho(\pi,\mathcal{N})}d(S, S_{\text{bayes}})\right), \tag{3}$$

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

Case 2: $\alpha = 1$.

The inequality (2) now becomes

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

The conclusion (3) follows by taking $t = 1/\rho(\pi, \mathcal{N})$ into the above inequality.

Therefore, Proposition **??** holds. □

C.2 Proof of Theorem ??

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

$$\begin{aligned}|\mu(\boldsymbol{X}) - \bar{\mu}(\boldsymbol{X})| &\leq \frac{1}{2H}\sum_{\pi\in\Pi}\left|I(\boldsymbol{X}\in\bar{S}(\pi)) - I(\boldsymbol{X}\in S_{\text{bayes}}(\pi))\right| + \left|\mu(\boldsymbol{X}) - \frac{1}{2} - \frac{1}{2H}\sum_{\pi\in\Pi}I(\boldsymbol{X}\in S_{\text{bayes}}(\pi))\right| \\ &= \frac{1}{H}\sum_{\pi\in\Pi}\mathbb{1}(\boldsymbol{X}\in\bar{S}(\pi)\Delta S_{\text{bayes}}(\pi)) + \left|\mu(\boldsymbol{X}) - \frac{1}{2H}\left(1 + 2\sum_{\pi\in\Pi}\mathbb{1}(\boldsymbol{X}\in S_{\text{bayes}}(\pi))\right)\right| \\ &\leq \frac{1}{H}\sum_{\pi\in\Pi}\mathbb{1}(\boldsymbol{X}\in\bar{S}(\pi)\Delta S_{\text{bayes}}(\pi)) + \frac{1}{2H} \\ &\leq \frac{1}{H}\sum_{\pi\in\Pi\setminus\mathcal{N}}\mathbb{1}(\boldsymbol{X}\in\bar{S}(\pi)\Delta S_{\text{bayes}}(\pi)) + \frac{1+2c}{2H},\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}|\mu(\boldsymbol{X}) - \bar{\mu}(\boldsymbol{X})| &\leq \frac{1}{H}\sum_{\pi\in\Pi\setminus\mathcal{N}}d_{\Delta}(\bar{S}(\pi), S_{\text{bayes}}(\pi)) + \frac{1+2c}{2H} \\ &\lesssim \frac{1}{H}\sum_{\pi\in\Pi\setminus\mathcal{N}}\left[d_{\pi}^{\alpha}(\bar{S}(\pi), S_{\text{bayes}}(\pi)) + \frac{1}{\rho(\pi,\mathcal{N})}d_{\pi}(\bar{S}(\pi), S_{\text{bayes}}(\pi))\right] + \frac{1}{H},\end{aligned}$$

where the last line comes from Proposition **??**. Furthermore, the definition of regression risk (**??**) 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 2

Proposition 2 (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 $t \in [0, c_2)$. 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

$$\int_{\{\boldsymbol{x}: l(t) \leq \langle \boldsymbol{b}, \boldsymbol{x} \rangle \leq u(t)\}} d\boldsymbol{x} \leq c_d t^{\frac{\alpha}{1-\alpha}} \tag{4}$$

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.

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

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

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

$$\leq c_{k+1} t^{\frac{\alpha}{1-\alpha}}.$$

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

Notice that

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

Combining the assumption that $g^{-1}(\pi + t) - g^{-1}(\pi - t) \leq c_1 t^{\frac{\alpha}{1-\alpha}}$ for all $t \in [0, c_2)$ and (4) gives

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

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

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

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

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 ??

Proof of Theorem **??**. 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 3, (π, α)-local regularity implies

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

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

$$\mathbb{P}\left[R_{\ell,\pi}(\hat{f}) - R_{\ell,\pi}(f_{\text{bayes}}) \geq \max\{a_n, t_n\}\right] \leq \frac{7}{2}\exp\left(-Cn(\lambda J_n)^{2-\alpha}\right), \tag{5}$$

where $C > 0$. Here $\{a_n\}$ is the vanishing sequence specified in Assumption **??**. Combining (**??**) and (5) 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,

$$\sup_{k\geq 2}\frac{1}{L}\int_{L}^{L^{\alpha/2}}\sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}^k, \|\cdot\|_2)}d\varepsilon \leq n^{1/2}, \quad \text{where } L = t_n + \lambda J_n(k/2-1). \tag{6}$$

In particular, the smallest t_n satisfying (6) yields the best upper bound of the error rate. Here $\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}^k, \|\cdot\|_2)$ denotes the L_2-norm, ε-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 (6). Based on Lemma 6, the inequality (6) is satisfied with

$$t_n \asymp \left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)} \quad \text{and} \quad \lambda \asymp \frac{t_n}{J_n}, \tag{7}$$

Plugging (7) into (5) gives

$$\begin{aligned}\mathbb{P}\left[d_\pi\left(\hat{S}_{\text{bayes}}(\pi), S_{\text{bayes}}(\pi)\right) \geq \max\left\{a_n, C_1\left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)}\right\}\right] &\leq \frac{7}{2}\exp\left(-C_2 r(s_1+s_2)\log d\right)\\ &\leq d^{-C_3 r(s_1+s_2)},\end{aligned}$$

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

C.5 Proofs of Theorem ??

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

Lemma 1 (Multiple level-sets estimation). *Suppose $\mu(\boldsymbol{X})$ is α-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* **??**, *we have*

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

Furthermore,

$$\frac{1}{H}\sum_{\pi\in\Pi} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) \lesssim \frac{1}{H} + t_n^{\alpha/(2-\alpha)} + Ht_n. \tag{8}$$

Proof of Lemma 1. The α-global regularity of $\mu(\boldsymbol{X})$ implies that, for all $\pi \notin \mathcal{N}$, the conversion inequality holds by Proposition 3,

$$d_\Delta(S_{\text{bayes}}(\pi), S) \lesssim \begin{cases} d_\pi^\alpha(S_{\text{bayes}}(\pi), S), & \text{if } S \in \text{I}, \\ \frac{1}{\rho(\pi,\mathcal{N})} d_\pi(S_{\text{bayes}}, S), & \text{otherwise}, \end{cases} \tag{9}$$

where region $\text{I} = \{S \colon d_\Delta(S_{\text{bayes}}(\pi), S) \leq \frac{C}{1-\alpha}\rho(\pi,\mathcal{N})^{\alpha/(1-\alpha)}\}$. In addition, the mean-to-variance inequality also holds by Proposition 3,

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

Applying Theorem **??** 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 (9), we obtain

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

where the last line follows from the fact that $a(b^2 + b^{2-\alpha} + b + 1) \leq 4a(b^2+1)$ 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 (8), we write

$$\frac{1}{H}\sum_{\pi\in\Pi} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) = \frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)) + \frac{1}{H}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c} d_\Delta(S_{\text{bayes}}(\pi), \hat{S}(\pi)), \tag{10}$$

where $\mathcal{N}_H \stackrel{\text{def}}{=} \bigcup_{\pi'\in\mathcal{N}} \left(\pi' - \frac{1}{H}, \pi' + \frac{1}{H}\right)$. The first term involves only finite number of summands and thus can be bounded by $2c/H$ where $c > 0$ is a constant such that $|\mathcal{N}| \leq c$. We bound the second term using the explicit forms of $\rho(\pi,\mathcal{N})$ in the sequence $\pi \in \Pi \cap \mathcal{N}_H^c$,

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

$$
\begin{aligned}
&\le t_n^{\alpha/(2-\alpha)} + \frac{t_n}{H} \sum_{\pi' \in \mathcal{N}} 2H^2 \\
&\le t_n^{\alpha/(2-\alpha)} + 2cHt_n,
\end{aligned}
$$

where the last inequality follows from the Lemma 2. Combining the bounds for the last two terms in (10) completes the second conclusion. As seen from the calculation, the distance $\rho^2(\pi, \mathcal{N})$ grows quadratically as the level π moves away from the mass points in $\mathcal{N}$. This leads to a fast linear rate in terms of H, provided that there are finitely many mass points in $\mathcal{N}$. □

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

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

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

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

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

Proposition 3. 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 ψ risk for the decision function f. Suppose Definition **??** 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:

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

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

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 proof in Proposition **??**. For Property 2, we devide the proof into two cases: when $\ell(\cdot)$ is hinge loss and ψ 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 $f_{\text{bayes},\pi}$,

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

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 **??**), the second line comes form the Lipschitz continuity of the hinge loss. Applying Lemma 3 on the last inequality complete the proof.

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

$$\begin{aligned}\mathrm{Var}\left[w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes}}(\boldsymbol{X}))\right] \le& \mathbb{E}|w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))|^2 \\ \le& L\mathbb{E}|w(y)\ell(yf(\boldsymbol{X})) - w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))| \\ \le& L\,\underbrace{\mathbb{E}\left|w(y)\left(1 - \mathrm{sign}(yf(\boldsymbol{X}))\right) - w(y)\ell(yf_{\text{bayes},\pi}(\boldsymbol{X}))\right|}_{=:(\text{i})} \\ &+ L\,\underbrace{\mathbb{E}\left|w(y)\ell(yf(\boldsymbol{X})) - w(y)\left(1 - \mathrm{sign}(yf(\boldsymbol{X}))\right)\right|}_{=:(\text{ii})},\end{aligned}$$

(i) is bounded as follows

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

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

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

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

□

Proof of Theorem **??**. Note that the proof of Theorem **??** still holds by replacing $\bar{S}(\pi)$ with $\hat{S}(\pi)$.

Based on Lemma 1 and Theorem **??**, with $1 - C_1 d^{-C_2 r(s_1+s_2)}$,

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

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* **??**. *Then, the L-1 distance between f and f_{bayes} is bounded by their hinge excess risk; i.e,*

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

Proof. For ease of notation, we drop the subscript π and simply write f_{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_{bayes}, μ, 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,

$$\begin{aligned} R_{\ell}(f) - R_{\ell}(f_{\text{bayes}}) &= \mathbb{E}[w(y)(1-yf)_{+}] - \mathbb{E}[w(y)(1-yf_{\text{bayes}})_{+}] \\ &= \int_{\boldsymbol{X}} w(1)\mu(1-f)_{+} d\mathbb{P}_{\boldsymbol{X}} + \int_{\boldsymbol{X}} w(-1)(1-\mu)(1+f)_{+} d\mathbb{P}_{\boldsymbol{X}} \\ &\quad - \int_{\boldsymbol{X}} w(1)\mu(1-f_{\text{bayes}})_{+} d\mathbb{P}_{\boldsymbol{X}} - \int_{\boldsymbol{X}} w(-1)(1-\mu)(1+f_{\text{bayes}})_{+} d\mathbb{P}_{\boldsymbol{X}}, \end{aligned} \tag{11}$$

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 $= \{\boldsymbol{X} \colon \mu < \pi \text{ and } f \geq -1\}$. In this region, $f_{\text{bayes}} = -1$, and the integrant in (11) reduces to

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

- Region II $= \{\boldsymbol{X} \colon \mu < \pi \text{ and } f < -1\}$. In this region, $f_{\text{bayes}} = -1$, and the integrant in (11) reduces to

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

- Region III $= \{\boldsymbol{X} \colon \mu \geq \pi \text{ and } f \leq 1\}$. In this region, $f_{\text{bayes}} = 1$, and the integrant in (11) reduces to

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

- Region IV $= \{\boldsymbol{X} \colon \mu \geq \pi$ and $f > 1\}$. In this region, $f_{\text{bayes}} = 1$, and the integrant in (11) reduces to

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

Therefore, the integral is evaluated as

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

Note that the functions $|\pi - \mu|, (\pi - \mu\pi), (\mu - \mu\pi)$ are non-negative, $[0,1]$-valued, and satisfy (12) by α-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, we rescale the functions $|f - f_{\text{bayes}}|\mathbb{1}(|f| \leq 1), |f - f_{\text{bayes}}|\mathbb{1}(f > 1), |f - f_{\text{bayes}}|\mathbb{1}(f < -1)$ to have expectation 1 and set upper bound of the functions as

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

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

Case 1:

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

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 $\alpha \in (0,1]$.

Case 2:

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

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*

$$\mathbb{P}[g(\boldsymbol{X}) \leq t] \leq C t^{\alpha/1-\alpha}, \quad \textit{for all } t \in (0,\rho], \tag{12}$$

for some constants $\alpha \in (0,1]$ *and* $C > 0$*. (When* $\alpha = 1$*, the right hand side of* (12) *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 following holds true,*

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

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 $(0,\rho]$. We bound the expectation of the product fg using the integral over the region $\{\boldsymbol{X}\colon g(\boldsymbol{X}) > t\}$,

$$\begin{aligned} \mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] &\geq \mathbb{E}[f(\boldsymbol{X})\mathbb{1}(g(\boldsymbol{X}) > t)] \\ &= t\mathbb{E}[f(\boldsymbol{X}) - f(\boldsymbol{X})\mathbb{1}(g(\boldsymbol{X}) \leq t)] \\ &\geq t\left(\mathbb{E}f(\boldsymbol{X}) - L\mathbb{P}[g(\boldsymbol{X}) \leq t]\right) \\ &\geq t\mathbb{E}f(\boldsymbol{X}) - CLt^{1/1-\alpha}, \quad \text{for all } t \in (0,\rho], \end{aligned} \tag{13}$$

where the third line uses the fact that $f(\boldsymbol{X}) \in [0,L]$, and the last line follows from (12). We maximize the lower bound (13) with respect to $t \in (0,\rho]$ and obtain the optimal $t_{\text{opt}} \in (0,\rho]$,

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

The corresponding lower bound of the inequality (13) becomes

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

Since, α, C, and L are given constant, we have the desired results.

□

Definition 1 (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, ε-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

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

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 ε-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 (**??**), 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

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

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

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

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, (2ε)-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 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 $\mathcal{F}(r,s_1,s_2)$)**.** *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 $k \in \{2,3,\ldots\}$,

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

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

$$\begin{aligned}g(L,k) &\le \frac{\sqrt{r(s_1+s_2)}}{L}\int_L^{\sqrt{L^\alpha+\rho^{-1}L}}\sqrt{\log\left(\frac{kd}{L}\right)}d\varepsilon\\ &\le \sqrt{r(s_1+s_2)(\log k+\log d-\log L)}\left(\frac{\sqrt{L^\alpha}+\sqrt{\rho^{-1}L}}{L}-1\right)\\ &\le \sqrt{r(s_1+s_2)(\log k+\log d)}\left(\frac{1}{L^{(2-\alpha)/2}}+\frac{1}{\sqrt{\rho L}}\right),\end{aligned} \tag{14}$$

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 (14) 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. $\square$