Optimal Fine-tuning in Deep ReLU Neural Networks for High Dimensional Regression

Jianqing Fan* Cheng Gao* Shange Tang*

December 3, 2024

Abstract

1 Introduction

sec:intro

Transfer learning has gained widespread use in neural networks, and with the advent of large language models (LLMs), it has become essential in AI development, primarily through the banner of *fine-tuning*. In addition to the goal of transferring knowledge to new tasks, another key motivation behind LLMs is that, despite their exceptional ability to process human language, they require substantial computational resources and vast amounts of data for training. In practice, we often perform inference on data with distributions different from those used in training. This challenge is especially significant in the LLM era since training an LLM from scratch can take months. To address this, practitioners typically fine-tune LLMs by optimizing a smaller set of residual parameters while keeping most pre-trained parameters fixed.

Message to Shange: In related literature, please review the minimax optimal rate mentioned in Schmidt-Hieber (2020); Kohler & Langer (2021) involving γ. Point out their limit as they cannot handle high-dimensional covariate input, and this issue is addressed by Fan & Gu (2023). For regression models and methodology, we use the same setting as in Fan & Gu (2023), and we aim to, on their theoretical basis, consider the improvement of convergence rate given additional source data with a pretrained model.

Besides statistical theory papers, also cite Caio Almeida & Tang (2023), highlighting the applications of related studies in financial data. They are fitting residuals from parametric modeling with nonparametric (NN) modeling.

1.1 Related Works

1.2 Contributions

2 The Problem Formulation

We study the problem of neural factor regression. Similar to the framework of general transfer learning or domain adaptation, our central goal is to answer the question of *When is the extra*

*Equal contribution.

source data accelerating the learning of target data given the distribution of P and Q to be unequal? In particular, this work considers the transfer learning problem under the high dimensional semi-parametric regression setup.

Given covariate X^Q and X^P across the target and source data, we consider two linear factor models as follows

$$X^Q = B^Q f^Q + u^Q, \quad X^P = B^P f^P + u^P,$$

where $B^P, B^Q \in R^{p\times r}$ are unknown loading matrices with $p \gg r$, while (f^Q, u^Q) and (f^P, u^P) are unobserved latent factors and idiosyncratic components that are identically distributed across the source and target data, respectively. Factor models have long been applied in the literature of financial econometrics and high dimensional statistics (). In particular, a large body of empirical study over decades has verified that factor modeling provides a good approximation to high dimensional data ().

We assume the covariate data samples $X_1^Q, \ldots, X_{n_Q}^Q, X_1^P, \ldots, X_{n_P}^P$ are observable and their associated latent vectors $\{f_i^Q, u_i^Q\}$ and $\{f_i^P, u_i^P\}$ are i.i.d. copies of (f^Q, u^Q) and (f^P, u^P), respectively. This assumption is reasonable because, although the explicit covariates may differ, the common implicit latent factors driving their realizations are typically the same.

Given the above preparation, we define the regression functions for the source distributions to be

$$\mathbb{E}[y^P|f^P, u^P] = g^P(X^P_{\mathcal{J}^P}), \qquad y_i^P = g^P(X^P_{i,\mathcal{J}^P}) + \epsilon_i^P, \quad \forall i \in [n_P] \tag{1}$$

eq:Pdata

where $\mathcal{J}^P \subset \{1, 2, \ldots, p\}$ is an unknown index subset. We assume that ϵ^P to be the noise that is i.i.d. across data instances. The formulation of the source data is often referred to as the factor augmented regression setup. For the generative model of the target distribution, we assume that

$$\mathbb{E}[y^Q|f^Q, u^Q] = g^Q(f^Q, u^Q_{\mathcal{J}\cup\mathcal{J}^P}) = h\left(f, u_{\mathcal{J}}, g^P(X^Q_{\mathcal{J}^P})\right), \qquad y_i^Q = g^Q(f_i^Q, u^Q_{i,\mathcal{J}\cup\mathcal{J}^P}) + \epsilon_i^Q, \quad \forall i \in [n_Q] \tag{2}$$

eq:Qdata

where $\mathcal{J} \subset \{1, 2, \ldots, p\}$ is also an unknown index subset, and h belongs to a (typically lower) complexity function class. Similarly, we assume that ϵ^Q is also i.i.d. across data instances. The reason for incorporating the idiosyncratic components u^P and u^Q in the input of g^P and g^Q is to consider a more realistic formulation where a sparse subset of u's entries correlate strongly with the response. This formulation has also been empirically justified as a better regression model for various real-world high dimensional datasets arising from problems in econometrics and data sciences (). For a detailed technical illstratuin of how they encompass various statistical models, see Fan & Gu (2023).

Message to Shange: Give some basic special cases of (8), with brief literature review if possible. For example, one can consider the residual modeling case where $h = h_0 + g^P$, which is g^P plus some "epsilon" terms for fine-tuning. What may also be interesting is some regime switching models like $h = h_0 \times ReLU(g^P)$.

2.1 Notations

We introduce some notation to be used throughout the paper. For any positive integer m, define $[m] = \{1, \ldots, m\}$. We use $x = (x_1, \ldots, x_d)^\top$ to denote a d-dimension vector, and $A = [a_{ij}]_{i\in[n],j\in[m]}$ to denote a $n \times m$ matrix. The $d \times d$ identity matrix is denoted by I_d. For any vector $y \in \mathbb{R}^{d_2}$, $[x, y]$ concatenates x and y together to form a $(d+d_2)$-dimensional vector. Let $\|x\|_q = (\sum_i |x_i|^q)^{1/q}$ be the vector l_q norm. Let $\|A\|_2 = \sup_{x\in\mathbb{R}^m, \|x\|_2=1} \|Ax\|_2$, $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, and $\|A\|_{\max} = \max_{i,j}|a_{ij}|$.

Additionally, we use $\lambda_{\min}(A)$ and $\nu_{\min}(A)$ to denote its minimum eigenvalue and singular value, respectively. Let $\mathbf{1}\{\cdot\}$ denote the indicator function. Let $\mathrm{supp}(X)$ denote the support of any random variable X. Let $X \sim Y$ represent that random variables X and Y have the same distribution. Let $a \wedge b$ and $a \vee b$ denote the minimum and maximum of a and b, respectively. Let $a_n \lesssim b_n$ denote $|a_n| \leq c|b_n|$ for some constant $c > 0$ when n is large enough. Let $a_n \gtrsim b_n$ denote $|a_n| \geq c|b_n|$ for some constant $c > 0$ when n is large enough. Let $a_n \asymp b_n$ denote $1/c \leq |a_n|/|b_n| \leq c$ for some constant $c > 1$ when n is large enough. Let $a_n \overset{n\to\infty}{\longrightarrow} \infty$ denote that a_n tends to infinity with n growing to infinity. Let $a_n \ll b_n$ denote $|a_n|/|b_n| \to 0$ when n is large enough. Let $a_n \gg b_n$ denote $|b_n|/|a_n| \to 0$ when n is large enough. In this paper, the symbols $\lesssim$, $\gtrsim$, $\ll$ and $\gg$ suppress constant dependencies that are independent of the parameters on both sides.

ReLU Neural Networks We use neural networks as a scalable, nonparametric technique for model fitting. Specifically, we consider fully connected deep neural networks with ReLU activation $\sigma(\cdot) = \max\{\cdot, 0\}$. Given any positive integer L (depth) and N (width), a deep ReLU neural network maps from $\mathbb{R}^d$ to $\mathbb{R}$ and takes the form:

$$g(x) = \mathcal{L}_{L+1} \circ \bar{\sigma}_L \circ \mathcal{L}_L \circ \bar{\sigma}_{L-1} \circ \cdots \circ \mathcal{L}_2 \circ \bar{\sigma}_1 \circ \mathcal{L}_1(x), \tag{3}$$ eq:nn-def

where $\mathcal{L}_l(z) = W_l z + b_l$ is an affine transformation with weight matrix $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and bias vector $b_l \in \mathbb{R}^{d_l}$, with $(d_0, d_1, \ldots, d_L, d_{L+1}) = (d, N, \ldots, N, 1)$, and $\bar{\sigma}_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ applies the ReLU activation $\sigma(\cdot)$ elementwise to a d_l-dimensional vector. For simplicity, we assume equal width across layers. We refer to both W_l and b_l as the *weights of a deep ReLU network*.

Definition 1 (Deep ReLU network class)**.** *Define the family of deep ReLU networks taking a d-dimensional vector as input, with depth L, width N, output truncated by M, and weights truncated by T as:*

$$\mathcal{G}(d, L, N, M, T) = \{\tilde{g}(x) = \mathrm{trun}_M(g(x)) : g(x) \text{ as in (3) with } \|W_l\|_{\max} \leq T, \|b_l\|_{\max} \leq T\},$$

where $\mathrm{trun}_M : \mathbb{R} \to \mathbb{R}$ *is the truncation operator defined as* $\mathrm{trun}_M(z) = \max\{|z|, M\} \cdot \mathrm{sign}(z)$.

Hölder's (β, C)-smoothness Let $\beta > 0$ and $C > 0$. A d-variate function g is (β, C)-smooth if for every non-negative sequence $\boldsymbol{\alpha} \in \mathbb{N}^d$ such that $\|\boldsymbol{\alpha}\|_1 = \lfloor \beta \rfloor$, the partial derivative $\partial g / \partial x^{\boldsymbol{\alpha}}$ exists and satisfies

$$\left| \frac{\partial g}{\partial x^{\boldsymbol{\alpha}}}(x) - \frac{\partial g}{\partial x^{\boldsymbol{\alpha}}}(z) \right| \leq C||x - z||_2^{\{\beta\}}.$$

We denote the set of all d-variate (β, C)-smooth functions by $\mathcal{F}_{d,\beta,C}$. Here, $\lfloor \beta \rfloor$ and $\{\beta\}$ denote the integer and fraction part of β, respectively. Throughout this paper, given any $A \in \mathbb{R}^{d \times d'}, b \in \mathbb{R}^{d'}$, we treat a linear transformation $Ax + b$ as $(\infty, \|A\|_2)$-smooth.

3 Model

sec:model

3.1 High-dimensional nonparametric regression with factor modeling

Suppose we observe two datasets $\{(X_i^P, y_i^P)\}_{i \in [n_P]} \overset{i.i.d.}{\sim} P$ and $\{(X_i^Q, y_i^Q)\}_{i \in [n_Q]} \overset{i.i.d.}{\sim} Q$ corresponding to the samples coming from the *source* and *target* domains. Here we let $n_P, n_Q \in \mathbb{N}$ to be the sample size of the source and target data and let $(X_i^Q, y_i^Q), (X_j^P, y_j^P) \in \mathbb{R}^p \times \mathbb{R}$ for all

$i \in [n_P], j \in [n_Q]$. Suppose that $n_P \overset{n_Q \to \infty}{\longrightarrow} \infty$. We further consider the following factor decomposition models for X^Q and X^P:

$$X^Q = B^Q f^Q + u^Q, \quad X^P = B^P f^P + u^P, \tag{4}$$ eq:factorModel2

where $B^P, B^Q \in \mathbb{R}^{p \times r}$ are unknown loading matrices with $p \gg r$ in the high-dimensional regime, f^Q and f^P are latent factor vectors, and u^Q and u^P are idiosyncratic components in the target and source domains, respectively. Throughout the paper, we assume that the covariate data samples $X_1^Q, \ldots, X_{n_Q}^Q, X_1^P, \ldots, X_{n_P}^P$ are observable, while their associated latent vectors $\{f_i^Q, u_i^Q\}$ and $\{f_i^P, u_i^P\}$ are i.i.d. copies of (f^Q, u^Q) and (f^P, u^P).

In this work, we focus on the regression setting, where $y_i^P, y_i^Q \in \mathbb{R}$. Given the latent factor structure, we consider the factor-augmented regression setting introduced in Fan & Gu (2023), where we treat the factors with idiosyncratic noise, i.e., (f^Q, u^Q) and (f^P, u^P), as the regressors for the target and source domains. This is a more general regression framework but has the advantage of making the variables much more weakly dependent. Specifically, we assume:

$$\begin{aligned}
&\mathbb{E}[y^Q | f^Q, u^Q] = g^Q(f^Q, u^Q_{\mathcal{J}^Q}), \\
&\mathbb{E}[y^P | f^P, u^P] = g^P([B_{\mathcal{J}^P,:}]f^P + u^P_{\mathcal{J}^P}) = g^P(X^P_{\mathcal{J}^P}),
\end{aligned}$$

where $\mathcal{J}^Q, \mathcal{J}^P \subset \{1, 2, \ldots, p\}$ are two unknown subsets of indices, and $\mathbb{E}$ represents the expectation with respect to the data generating process of (X^Q, y^Q) and (X^P, y^P). Note that here the form of g^P follows a slightly stronger assumption than g^Q, namely that the conditional expectation can be fully determined by $X^P_{\mathcal{J}^P}$. We will show in Section 3.2 that this assumption ensures $g^P(X^Q_{\mathcal{J}^P})$ functions as a key component for fine-tuning, allowing us to directly use X^Q as input with any estimator of g^P.

Denote the noise terms by $\epsilon^Q = y^Q - \mathbb{E}[y^Q | f^Q, u^Q]$ and $\epsilon^P = y^P - \mathbb{E}[y^P | f^P, u^P]$ for the target and source data, respectively. Then, our data-generating process can be summarized as follows:

$$\begin{aligned}
&X_i^Q = B^Q f_i^Q + u_i^Q, \quad y_i^Q = g^Q(f_i^Q, u^Q_{i,\mathcal{J}^Q}) + \epsilon_i^Q, \quad \forall i \in [n_Q], \\
&X_j^P = B^P f_j^P + u_j^P, \quad y_i^P = g^P(X^P_{i,\mathcal{J}^P}) + \epsilon_i^P, \quad \forall i \in [n_P].
\end{aligned} \tag{5}$$ eq:setting

For a more detailed discussion of the factor-augmented regression model, see Fan & Gu (2023).

The goal of this paper is to improve the estimation of the target data by fine-tuning the estimation learned from the source data, transferring useful information from the source domain. The effectiveness of a function $m : \mathbb{R}^p \to \mathbb{R}$ is evaluated by its population L_2 risk on the target distribution:

$$R(m) = \mathbb{E}[|y^Q - m(X^Q)|^2].$$

Since the regression function g^Q minimizes the population risk on the target domain, the performance of any empirical estimator $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ can then be measured by the excess risk (on the target distribution):

$$\mathcal{E}_Q(\widehat{m}) := \mathbb{E}[|\widehat{m}(X^Q) - g^Q(f^Q, u^Q_{\mathcal{J}^Q})|^2] = R(\widehat{m}) - \mathbb{E}[|y^Q - g^Q(f^Q, u^Q_{\mathcal{J}^Q})|^2]. \tag{6}$$ eq:excessRisk

Given the excess risk defined in (6), our objective is to use both target and source data to construct an empirical estimate $\widehat{m}$ that improves the target excess risk $\mathcal{E}_Q(\widehat{m})$, achieved by fine-tuning the estimate $\widehat{g}^P$ obtained from the source data.

3.2 Hierarchical decomposition of regression functions

sec:mainAss

The rate at which the excess risk converges depends on the assumptions made about the family of functions that g^Q and g^P belong to. In non-parametric regression theory, a common assumption is the *hierarchical decomposition model* (HDM) Fan & Gu (2023), which models the hierarchical composition structure of the true regression functions. This structure can be adaptively learned by deep ReLU networks, making them well-suited for capturing low-dimensional structures without requiring explicit guidance about the functional forms. **Message to Shange:** Cite papers about HDMs, see Fan & Gu (2023) page 10.

def:hcm

Definition 2 (Hierarchical composition model)**.** *The function class of the hierarchical composition model $\mathcal{H}(d, l, \mathcal{P}, C)$, with $l, d \in \mathbb{N}^+$, $C > 0$, and $\mathcal{P}$, a subset of $[1, \infty] \times \mathbb{N}^+$ is defined as follows. For $l = 1$,*

$$\mathcal{H}(d, 1, \mathcal{P}, C) = \{h : \mathbb{R}^d \to \mathbb{R} : h(x) = g(x_{\pi(1)}, \ldots, x_{\pi(t)}), \text{ where } g : \mathbb{R}^\top \to \mathbb{R} \text{ is } (\beta, C)\text{-smooth for some } (\beta, t) \in \mathcal{P} \text{ and } \pi : [t] \to [d]\}.$$

For $l > 1$,

$$\mathcal{H}(d, l, \mathcal{P}, C) = \{h : \mathbb{R}^d \to \mathbb{R} : h(x) = g(f_1(x), \ldots, f_t(x)), \text{ where } g : \mathbb{R}^\top \to \mathbb{R} \text{ is } (\beta, C)\text{-smooth for some } (\beta, t) \in \mathcal{P} \text{ and } f_i \in \mathcal{H}(d, l-1, \mathcal{P}, C)\}.$$

Following Kohler & Langer (2021), we restrict our focus to the case where all compositions have a smoothness parameter $\beta \geq 1$ to simplify the presentation. The hierarchical composition model classes that g^Q, g^P, and h belong to are crucial for deducing the learning rate. To evaluate the complexity of the regression function family in terms of the model class, for any $\mathcal{P} \subset [1, \infty] \times \mathbb{N}^+$, we define the function $\gamma(\cdot)$ of the hardest component in the composition as

$$\gamma(\mathcal{P}) = \min_{(\beta,t)\in\mathcal{P}} \frac{\beta}{t}. \tag{7}$$

eq:hardestComponent

$\gamma(\mathcal{P})$ represents the smallest dimensionality-adjusted degree of smoothness and, as such, can be interpreted as a measure of the complexity of the regression function.

With the setup above, we are now in a position to introduce our main working assumption regarding the regression functions g^Q and g^P, along with their relationship, which provides the rationale behind fine-tuning techniques.

ass:transferability

Assumption 1. *The target regression function can be written with the form of*

$$g^Q(f^Q, u^Q_{\mathcal{J}^Q}) = h\left(f^Q, u^Q_{\mathcal{J}}, g^P(X^Q_{\mathcal{J}^P})\right), \tag{8}$$

eq:transfer

where

1. $\mathcal{J}^Q = \mathcal{J} \cup \mathcal{J}^P$;
2. *The function $g^P(x) : \mathbb{R}^p \to \mathbb{R}$ satisfies $g^P \in \mathcal{H}(|\mathcal{J}^P|, l^P, \mathcal{P}^P, c_0)$ with $\gamma^P := \gamma(\mathcal{P}^P) < \infty$;*
3. *The function $h(f, u_{\mathcal{J}}, s) : \mathbb{R}^{r+|\mathcal{J}|+1} \to \mathbb{R}$ satisfies $h \in \mathcal{H}(r + |\mathcal{J}| + 1, l, \mathcal{P}, c_0)$ with $\gamma := \gamma(\mathcal{P}) < \infty$.*

for some $l, \mathcal{P}, l^P, \mathcal{P}^P$, $c_0 > 0$, and an unknown subset of indices $\mathcal{J} \subset \{1, 2, \ldots, p\}$. We call the remaining regression function part h given full information of g^P as the residual regression function.

Assumption 1 considers a scenario where the information from g^P is beneficial for estimating g^Q, as g^P serves as the input for the residual regression function h, capturing partial information necessary for estimating the full target regression function g^Q. Model (8) encompasses many useful models in the literature of transfer learning, domain adaptation, and model fine-tuning. **Message to Shange:** Expand with technical examples if needed.

The purpose of setting a special case $g^P(X^Q_{\mathcal{J}^P})$ instead of the full factor structure $g^P(f^Q, u^Q_{\mathcal{J}^P})$ serves for several purposes: **Message to Shange:** Rewrite and polish it up

- after getting an estimate of $g^P(X^P)$, we can directly use the X^Q as the input the generate plug-in pretrained signals.
- This also avoids measurement error from $B^Q \neq B^P$
- Intuitively, the source regression function $g^P(X^Q_{\mathcal{J}^P})$ is uncorrelated with f^Q, so no need to write g^P depending on f^Q.

The following assumption demonstrates that both g^Q and g^P belong to a certain hierarchical decomposition model family when considering the latent factor and idiosyncratic noise as the input, with the complexity of the family for g^Q determined by the more complex part between g^P and h.

prop:transferabilityBenefit

Proposition 1. *Under Assumption 1, we have*

$$g^P([B^P_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P}) \in \mathcal{H}(r + |\mathcal{J}^P|, l^P + 1, \mathcal{P}^P \cup \{\infty, r + |\mathcal{J}^P|\}, c_0 \vee \|B^P_{\mathcal{J}^P,:}\|_2,$$
$$g^Q(f, u_{\mathcal{J}\cup\mathcal{J}^P}) \in \mathcal{H}(r + |\mathcal{J} \cup \mathcal{J}^P|, l + l^P + 1, \mathcal{P} \cup \mathcal{P}^P \cup \{\infty, r + |\mathcal{J}^P|\}, c_0 \vee \|B^P_{\mathcal{J}^P,:}\|_2).$$

Moreover, we have

$$\gamma(\mathcal{P} \cup \mathcal{P}^P \cup \{\infty, r + |\mathcal{J}^P|\}) = \min\{\gamma, \gamma^P\}.$$

Since h is only a portion of the full regression function g^Q, intuitively, the estimation process should be easier with faster convergence if h belongs to a less complex function family. Proposition 1 demonstrates that g^Q follows a hierarchical decomposition model, stacking g^P and h. The complexity measure of g^Q is thus determined by the more complex function between g^P and h. Therefore, given a good estimate for g^P, if h belongs to a less complex class than g^P, we wish a fine-tuning procedure to improve the convergence rate of the excess risk by focusing on estimating h rather than the full function g^Q.

4 Methodology

sec:method

In this section, we describe our method for handling both covariate shift, i.e. how to let the source data aid in a better estimate of the latent factor f^Q, and posterior shift, i.e. how to let the pretrained model for the source regression function g^P help improve the estimation of the target regression function g^Q.

4.1 Diversified projection matrix

Compared with the standard neural network that directly uses the covariate X as input, our method introduces an additional module ahead of the input layer of a deep ReLU network: a pre-defined *diversified projection matrix* W to estimate the factor f.

Let $\overline{r}$ be an integer satisfying $\overline{r} \geq r$. We first introduce the idea of *diversified projection matrix* proposed by Fan & Liao (2022) for both the target and source domains.

def:dpm

Definition 3 (Diversified projection matrix). *Let $\overline{r} \geq r$, and C_1 be a universal positive constant. For any $* \in \{Q, P\}$, a matrix $W^* \in \mathbb{R}^{p\times\overline{r}}$ is called a* target (or source) diversified projection matrix *if it satisfies*

1. *(Boundedness) $\|W^*\|_{\max} \leq C_1$;*
2. *(Exogeneity) W is independent of $X_1^*, \ldots, X_{n_*}^*$ in (5);*
3. *(Significance) The matrix $H^* = p^{-1}(W^*)^\top B^* \in \mathbb{R}^{\overline{r}\times r}$ satisfies $\nu_{\min}(H^*) \gg p^{-1/2}$.*
   Each column of W^* is called as diversified weight, *and $\overline{r}$ is the number of diversified weights.*

Given the target and source diversified projection matrices W^Q and W^P, we can define the following surrogates for f^Q and f^P in the downstream regression task:

$$\begin{aligned} \widetilde{f}^Q &= p^{-1}(W^Q)^\top X^Q, \\ \widetilde{f}^P &= p^{-1}(W^P)^\top X^P, \end{aligned} \tag{9}$$

eq:surrogateF

where $\widetilde{f}^Q$ and $\widetilde{f}^P$ serve as surrogates for the latent factors f^Q and f^P. **Message to Shange:** Elaborate more on this idea if needed. See Fan & Gu (2023) page 7, but no need to be that detailed.

4.2 Transfer factor estimate for covariate shift

To understand how we leverage source data to improve the surrogate estimate $\widetilde{f}^Q$ for f^Q, we first introduce the method for determining the diversified projection matrices W^Q and W^P in practice.

Define the sample covariance matrices for the target data, source data, and pooled data as follows:

$$\begin{cases} \widehat{\Sigma}^Q &= \frac{1}{n_Q}\sum_{i=1}^{n_Q} X_i^Q (X_i^Q)^\top, \\ \widehat{\Sigma}^P &= \frac{1}{n_P}\sum_{i=1}^{n_P} X_i^P (X_i^P)^\top, \\ \widehat{\Sigma}^A &= \frac{1}{n_Q+n_P}\Big(\sum_{i=1}^{n_Q} X_i^Q (X_i^Q)^\top + \sum_{i=1}^{n_P} X_i^P (X_i^P)^\top\Big) = \frac{n_Q}{n_Q+n_P}\widehat{\Sigma}^Q + \frac{n_P}{n_Q+n_P}\widehat{\Sigma}^P. \end{cases}$$

Next, for any $* \in \{Q, P, A\}$, let $\widehat{v}_1^*, \ldots, \widehat{v}_{\overline{r}}^* \in \mathbb{R}^p$ be the top-$\overline{r}$ eigenvectors of $\widehat{\Sigma}^*$, and $\widehat{W}^*$ be $p^{1/2}[\widehat{v}_1^*, \ldots, \widehat{v}_{\overline{r}}^*]$. Without the availability of the source data, a common method is to choose $\widehat{W}^Q$ to be the diversified projection matrix for the target data. **Message to Shange:** Cite papers, see Fan & Gu (2023) end of Section 3.1. Here, we omit the standard sample-splitting procedure for simplicity, which ensures exogeneity in Definition 3.

However, this method does not leverage the information from the source data. Given the availability of source data, we instead propose a model-selection process for the sample covariance estimation. In this process, we continue using the conventional covariance estimate $\widehat{\Sigma}^Q$ when $\|\widehat{\Sigma}^Q - \widehat{\Sigma}^A\|_F$ is large, indicating significant differences between the target and pooled data. However, when $\|\widehat{\Sigma}^Q - \widehat{\Sigma}^A\|_F$ is small, we assume the source data provides valuable information for estimating the covariance matrix of the target data and use $\widehat{\Sigma}^A$ instead. Specifically, we extract the factors by

$$\widehat{\Sigma}^{TL} = \begin{cases} \widehat{\Sigma}^A, & p^{-1}\|\widehat{\Sigma}^Q - \widehat{\Sigma}^A\|_F \leq \delta, \\ \widehat{\Sigma}^Q, & \text{otherwise.} \end{cases}$$

For some pre-determined parameter δ. This model selection process follows a similar philosophy first introduced as the "Transfer Around Boundary (TAB)" classifier in Fan et al. (2023), where the

source data or estimate is utilized only when a focal value falls below a certain threshold, indicating low confidence in the estimation obtained from the target data.

Next, we let $\widehat{v}_1^{TL}, \ldots, \widehat{v}_{\overline{r}}^{TL} \in \mathbb{R}^p$ be the top-$\overline{r}$ eigenvectors of $\widehat{\Sigma}^{TL}$. The surrogate factor $\widetilde{f}^Q$ is then defined as

$$\widetilde{f}^Q = p^{-1}(\widehat{W}^{TL})^\top X^Q, \tag{10}$$

where $\widehat{W}^{TL} = p^{1/2}[\widehat{v}_1^{TL}, \ldots, \widehat{v}_{\overline{r}}^{TL}]$. Equivalently, the diversified projection matrix estimate for the target data follows a model-selected construction, i.e.

$$\widetilde{f}^Q = \begin{cases} p^{-1}(\widehat{W}^A)^\top X^Q, & p^{-1}\|\widehat{\Sigma}^Q - \widehat{\Sigma}^A\|_F \le \delta \\ p^{-1}(\widehat{W}^Q)^\top X^Q, & \text{otherwise.} \end{cases}$$

The part of transferring factors of the source source data from the target data is symmetric. Specifically, for some pre-determined parameter $\delta > 0$, we let

$$\widetilde{f}^P = \begin{cases} p^{-1}(\widehat{W}^A)^\top X^P, & p^{-1}\|\widehat{\Sigma}^P - \widehat{\Sigma}^A\|_F \le \delta \\ p^{-1}(\widehat{W}^P)^\top X^P, & \text{otherwise.} \end{cases}$$

For simplicity, we assume W_Q and W_P are given as prior knowledge and pre-defined for the non-parametric regression throughout the rest of this paper.

4.3 Transfer regression function for posterior shift

Define the clipped-L_1 function with threshold $\tau > 0$ as

$$\psi_\tau(x) = \frac{|x|}{\tau} \wedge 1$$

as an approximator of $\mathbf{1}\{x \neq 0\}$. **Message to Shange:** Elaborate briefly on its motivation. See Fan & Gu (2023).

We explore the fine-tuning approach in transfer learning estimation. The idea behind this approach is straightforward: the complex part of the regression function structure, g^P, is well-learned using the source data, which typically has a large sample size. The resulting estimate $\widehat{g}^P$ is then used as an additional input when estimating g^Q, reducing the estimation complexity. This is because only the simpler component, h, needs to be learned from the target data.

Specifically, our estimator for the source regression function is defined as follows:

$$\widehat{g}^P(\cdot), \widehat{\Theta}^P = \underset{g \in \mathcal{G}(L^P, r+N^P, N^P, M, T^P), \Theta^P \in \mathbb{R}^{p \times N^P}}{\arg\min} \frac{1}{n^P} \sum_{i=1}^{n^P} \left(Y_i^P - g\left(\left[\widetilde{f}_i^P, \mathrm{trun}_M\left((\Theta^P)^\top X_i^P \right) \right] \right) \right)^2 + \lambda^P \sum_{i,j} \psi_\tau(\Theta_{i,j}^P) \tag{11}$$

Here, we follow the approach in Fan & Gu (2023) to estimate the throughput u_i^P. This involves creating sparse linear combinations to select a subset of idiosyncratic throughputs for estimation. Upon acquiring the source neural network regression function, the pretrained model generated by applying this fitted network is determined as follows:

$$\widehat{s}(x) = \widehat{g}^P\left(\left[p^{-1}(W^P)^\top x, \mathrm{trun}_M\left((\widehat{\Theta}^P)^\top x \right) \right] \right) \tag{12}$$

with

$$\widehat{s}_i = \widehat{s}(X_i^Q), \quad \forall i \in [n_Q]. \tag{13}$$

Next, we incorporate the derived transferable pretrained signals, denoted as $\{\widehat{s}_i\}_{i=1,\ldots,n_Q}$, as part of the input in fitting the target neural network regression function.

$$\widehat{h}(\cdot), \widehat{\Theta} = \underset{h\in\mathcal{G}(L,r+N+1,N,M,T),\Theta\in\mathbb{R}^{p\times N}}{\arg\min} \frac{1}{n_Q}\sum_{i=1}^{n^Q}\Big(Y_i - h(\Big[\widehat{f}_i^Q, \mathrm{trun}_M(\Theta^\top X_i^Q), \widehat{s}_i\Big])\Big)^2 + \lambda\sum_{i,j}\psi_\tau(\Theta_{i,j}) \tag{14}$$

eq:FTFAST4

All of $M, \tau, L, L^P, N, N^P, T, T^P, \lambda, \lambda^P$ in (11)-(14) are tuning hyper-parameters. The final fine-tuning factor augmented estimator is thus defined by

$$\widehat{m}_{FT}(x) = \widehat{h}(\Big[p^{-1}(W^Q)^\top x, \mathrm{trun}_M(\widehat{\Theta}^\top x), \widehat{s}(x)\Big]). \tag{15}$$

eq:FTFAST5

5 Theory for factor transferring

sec:factorTransfer

In this section, we present the theoretical results for the methods in Section 4. The following assumption is essential for transferring useful source information within the factor decomposition structure. It controls the maximum discrepancy between the factor loading matrices B^Q and B^P, thereby ensuring the similarity between the generating distributions of the covariates X^Q and X^P.

ass:factorLoadingDiff

Assumption 2 (Factor loading similarity)**.** *We assume the factor decomposition* (4) *satisfies that*

$$p^{-1}\|B^Q(B^Q)^\top - B^P(B^P)^\top\|_F \le \varepsilon$$

for some constant $\varepsilon > 0$.

Next, we impose several regularity conditions on the covariate and factor model settings.

ass:boundedness

Assumption 3 (Boundedness)**.** *There exist universal constants* c_1 *and* b *such that*

1. $\|B^Q\|_{max} \vee \|B^P\|_{max} \le c_1$;
2. $\mathbb{E}[f^*] = 0,\ \mathbb{E}[f^*(f^*)\top] = I_r,\ \mathrm{supp}(f^*) \subset [-b,b]^r$ *for any* $* \in \{Q,P\}$;
3. $\mathbb{E}[u^*] = 0,\ \mathrm{supp}(u^*) \subset [-b,b]^p$ *for any* $* \in \{Q,P\}$.

Assumption 4 (Weak dependence)**.** *There exists some universal constant* c_1 *such that for any* $* \in \{Q,P\}$,

1. *(Between entries of* u^**)* $\sum_{1\le j<k\le p}|\mathbb{E}[u_j^* u_k^*]| \le c_1 \cdot p$;
2. *(Between* f^* *and* u^**)* $\|B^*\mathbb{E}[f^*(u^*)^\top]\|_F \le c_1\sqrt{p}$.

ass:weakDependence

Assumption 5 (Pervasiveness)**.** *There exists some universal constant* $c_1 \ge 1$ *such that*

$$p/c_1 \le \lambda_{\min}((B^Q)^\top B^Q) \wedge \lambda_{\min}((B^P)^\top B^P)$$

and

$$\lambda_{\max}((B^Q)^\top B^Q) \vee \lambda_{\max}((B^P)^\top B^P) \le c_1 \cdot p.$$

ass:pervasiveness

Assumptions 3–5 are standard in factor models. We replace the sub-Gaussian condition for f^Q, f^P, u^Q, and u^P with uniform boundedness in Assumption 3. Assumptions 4 and 5 on the distribution of (f^*, u^*) for any $* \in \{Q, P\}$ lay the foundation of the validity of the chosen diversified projection matrices, ensuring that the condition $\nu_{\min}(H^Q) \asymp 1$ holds.

The following theorem demonstrates that $\nu_{\min}(H^*) \asymp 1$ holds with high probability for both n^Q and n^P.

thm:factorTransfer

Theorem 1 (Factor transfering)**.** *Under Assumptions 2-5, the matrix $\widehat{W}^{TL}$ proposed in* (10) *satisfies*

$$\nu_{max}(p^{-1}\widehat{W}^{TL}B^Q) \leq c_3. \tag{16}$$

Moreover, for any $t > 0$, we define the following quantities:

$$\varepsilon^Q(t) = r\sqrt{\frac{\log p + t}{n_Q}} + r^2\sqrt{\frac{\log r + t}{n_Q}} + \frac{1}{\sqrt{p}},$$

$$\varepsilon^A(t) = r\sqrt{\frac{\log p + t}{n_Q + n_P}} + r^2\sqrt{\frac{\log r + t}{n_Q + n_P}} + \frac{1}{\sqrt{p}} + \varepsilon.$$

Then given t such that $\delta \geq c_3\varepsilon^Q(t)$, we have

eq:thmFactorResult

$$\nu_{min}(p^{-1}\widehat{W}^{TL}B^Q) \geq c_1 - c_2\delta \wedge \varepsilon^A(t). \tag{17}$$

with probability at least $1 - 6e^{-t}$ with respect to the target and source data for some universal constant $c_1 - c_3$ that are independent of $n_Q, n_P, p, t, r, \overline{r}$.

Theorem 1 indicates that the source data can enhance transfer learning of the target latent factor loadings by enabling the construction of a more statistically robust diversified projection matrix, $\widehat{W}^{TL}$, compared to $\widehat{W}^Q$ that is derived without access to the source data. The choice of t can be of the order $\log n_Q$. The process of learning the source diversified projection matrix W^P is symmetric, incorporating the target data. However, since n^P is typically large, it is often sufficient to use only $\widehat{W}^P$. For the choice of δ, since we only require the right-hand side of (17) to be lower bounded by a positive constant, a precise or tight choice of δ is unnecessary.

In the high-dimensional regime with $p \gg 1$, and without source data, we can construct surrogates of the latent factors $\widetilde{f}^Q$ when $n_Q \gg (\log p + t)$. However, even when n^Q is small, a robust estimate can still be achieved if $n^P \gg (\log p + t)$ and the true factor loading difference ϵ is sufficiently small. Thus, a small subset of the target and source data can be used to learn the diversified weights with $\nu_{\min}(H^Q) \asymp 1$, depending on the specific conditions of n_Q, n_P, and any prior information about ϵ. When ϵ is believed to be small or zero, splitting samples from the source data alone is sufficient. Otherwise, it is recommended to select samples from both the target and source domains, and we can sample more source data since the target data is typically rare.

In addition, we provide the pertubation results regarding the covariance matrix estimate $\widehat{\Sigma}^{TL}$ and the eigenspaces.

prop:pertubation

Proposition 2 (Pertubation bounds)**.** *Under the assumptions and notations of Theorem 1, for some universal constant c_4 we have*

1. $\|\widehat{\Sigma}^{TL} - B^Q(B^Q)^\top\|_F \lesssim p\Big(\delta \wedge \varepsilon^A(t)\Big)$;
2. $\min_{R\in\mathcal{O}(r)}\|\widehat{V}_r^{TL}R - V^Q\|_F \lesssim \delta \wedge \varepsilon^A(t)$.

with probability at least $1 - 6e^{-t}$ given $t > 0$ such that $\delta \geq c_4\varepsilon^Q(t)$. Here, $\mathcal{O}(r)$ denotes the space of orthogonal matrices of size $r \times r$, and $\widehat{V}_r^{TL}$ and V^Q represent the top-r eigenvectors of $\widehat{\Sigma}^{TL}$ and $B^Q(B^Q)^\top$, respectively.

6 Theory for regression task fine-tuning

sec:fineTuning

In this section, we aim to establish high probability bounds for both the out-of-sample excess risk $\mathcal{E}_Q(\widehat{m}_{FT})$ and the in-sample mean squared error, denoted as $\widehat{\mathcal{E}}_Q(\widehat{m}_{FT})$ where

$$\widehat{\mathcal{E}}_Q(m) := \frac{1}{n_Q}\sum_{i=1}^{n_Q}|m(X_i^Q) - g^Q(f_i^Q, u^Q_{i,\mathcal{J}\cup\mathcal{J}^P})|^2.$$

For justifications on the importance of deriving high-probability error bounds for both in-sample and out-of-sample l_2 errors, see Farrell et al. (2021).

6.1 Oracle-type upper bound

Before presenting the results about how the fine-tuning factor augmented estimator in (15) improves the risk convergence rate, we impose two additional assumptions as follows

ass:subGaussianNoiseAndBoundedFunction

Assumption 6. *The model inputs satisfy that*

1. *(Sub-Gaussian noise) There exists a universal constant c_1 such that $\mathbb{P}(|\epsilon^*|\geq t|X^*) \leq 2\exp(-c_1t^2)$ for all $t>0$ and $*\in\{Q,P\}$ almost surely.*
2. *(Bounded Regression function) We have $\max\{\|g^P\|_\infty, \|h\|_\infty\} \leq M^*$ for some universal constant M^*. The choice of M satisfies that $1\leq M^* \leq M \leq c_2M^*$ for some universal constant $c_2>1$. Additionally, g^P and h are all c_1-Lipschitz for some universal constant c_1.*

ass:absCon

Assumption 7 (Absolutely continuity)**.** *The Radon-Nikodym derivative satisfies $\frac{d\mu_{X^Q}}{d\mu_{X^P}}(x) \leq c_1$ almost surely for some universal constant c_1.*

Assumption 6 is standard in the literature of non-parametric regression. Assumption 7 ensures that any good source population estimate $\widehat{s}$ for g^P, with a small $\mathbb{E}\Big[|\widehat{s}(X^P) - g^P(X^P_{\mathcal{J}^P})|^2\Big]$, also enables proportionally accurate learning for the target population estimate, i.e.,

$$\mathbb{E}\Big[|\widehat{s}(X^Q) - g^P(X^Q_{\mathcal{J}^P})|^2\Big] \lesssim \mathbb{E}\Big[|\widehat{s}(X^P) - g^P(X^P_{\mathcal{J}^P})|^2\Big].$$

The intuition behind is that, the source marginal distribution for X^P should provide full coverage over the target marginal distribution for X^Q, allowing the accuracy of non-parametric regression to transfer effectively. An assumption similar to the second statement is made in Fan et al. (2023).

The following theorem presents an orcale-type inequality for the error bound.

thm:fineTuning

Theorem 2 (Oracle-type bound for fine-tuning factor augmented estimator)**.** *Suppose that Assumptions 3, 4, 6 and 7 hold. Consider the fine-tuning factor augmented model in* (11)-(15) *with*

- $N^P \geq 2(r+|\mathcal{J}^P|),\ N \geq 2(r+|\mathcal{J}|)$
- $T^P \geq c_1[\nu_{min}(H^P)]^{-1}|\mathcal{J}^P|r,\ T \geq c_1[\nu_{min}(H^Q)]^{-1}|\mathcal{J}|(r+1)$
- $\lambda^P \geq c_2 n_p^{-1}\Big(\log(pn_P(N^P+\overline{r}))+L^P\log(T^PN^P)\Big),\ \lambda \geq c_2 n_Q^{-1}\Big(\log(pn_Q(N+\overline{r}))+L\log(TN)\Big)$
- $\tau^{-1} \geq c_3(r+1)p\Big[\Big((T^PN^P)^{L^P+1}(N^P+\overline{r})n_P\Big)\vee\Big((TN)^{L+1}(N+\overline{r})n_Q\Big)\Big]$

for some universal constants c_1-c_3 and $\overline{r}\geq r$. Define

- $\delta_a^h = \inf_{g\in\mathcal{G}(L-1,r+|\mathcal{J}|+1,N,M,T)} \sup_{\kappa\in[-M,M]} \|g(f^Q, u^Q_{\mathcal{J}}, \kappa) - h(f^Q, u^Q_{\mathcal{J}}, \kappa)\|^2_\infty$
- $\delta_a^Q = \inf_{g\in\mathcal{G}(L-1,r+|\mathcal{J}\cup\mathcal{J}^P|,N,M,T)} \|g(f^Q, u^Q_{\mathcal{J}\cup\mathcal{J}^P}) - g^Q(f^Q, u^Q_{\mathcal{J}\cup\mathcal{J}^P})\|^2_\infty$
- $\delta_a^P = \inf_{g\in\mathcal{G}(L^P-1,r^P+|\mathcal{J}^P|,N^P,M^P,T^P)} \|g(f^P, u^P_{\mathcal{J}^P}) - g^P(X^P_{\mathcal{J}^P})\|^2_\infty$
- $\delta_s^h = n_Q^{-1}(N^2L + N\bar{r})L\log(TNn_Q) + \lambda|\mathcal{J}|$, $\delta_s^Q = n_Q^{-1}(N^2L + N\bar{r})L\log(TNn_Q) + \lambda|\mathcal{J}\cup\mathcal{J}^P|$, $\delta_s^P = n_p^{-1}\Big((N^P)^2L^P + N^P\bar{r}\Big)L^P\log(TN^Pn_P) + \lambda^P|\mathcal{J}^P|$
- $\delta_f^h = \frac{|\mathcal{J}|r\cdot\bar{r}}{p\cdot\nu^2_{min}(H^Q)}$, $\delta_f^Q = \frac{|\mathcal{J}\cup\mathcal{J}^P|r\cdot\bar{r}}{p\cdot\nu^2_{min}(H^Q)}$, $\delta_f^P = \frac{|\mathcal{J}^P|r\cdot\bar{r}}{p\cdot\nu^2_{min}(H^P)}$

Then, with probability at least $1-6e^{-t}$ *with respect to the target and source data, for* n_Q *(and thus* n_P*) large enough, we have*

$$\begin{aligned}\mathcal{E}_Q(\widehat{m}_{FT}) + \widehat{\mathcal{E}}_Q(\widehat{m}_{FT}) \leq& c_4\Big\{\Delta^Q \wedge \Big(\Delta^P + \frac{t}{n_P}\Big) + \Delta^h + \frac{t}{n_Q}\Big\} \\ \leq& 2c_4\Big\{\Delta^Q \wedge \Delta^P + \Delta^h + \frac{t}{n_Q \wedge n_P}\Big\}\end{aligned} \tag{18}$$

where

$$\Delta^* = \delta_a^* + \delta_f^* + \delta_s^*$$

for any $* \in \{Q, P, h\}$*, and* c_4 *is a universal constant that depends only on the constants in Assumptions 3, 4 and 6.*

Message to Shange: δ_a^Q and $\delta_a^{TL} + \delta_s^P + \delta_f^P$ correspond to two instances that seek to minimize the approximation error. The first instance uses only the h network structure, while the second considers both h and g^P, but considers the error of $\widehat{s}$. The second term dominates the first when n_P is large compared with n_Q.

Theorem 6 establishes a high-probability bound on both the out-of-sample squared error and the in-sample mean squared error. The error bound is composed of Δ^Q, Δ^P, and Δ^h, which represent the errors from estimating g^Q, g^P, and h, respectively. Notably, the convergence rate takes the form $\Delta^Q \wedge \Delta^P + \Delta^h$, indicating that if Δ^P is sufficiently small due to a good estimate of g^P (possibly from a large n_P), the fine-tuning procedure can improve the convergence rate from Δ^Q to Δ^h, leveraging the useful information from g^P to enhance the estimation of g^Q. However, a term of Δ^h is unavoidable, as h still needs to be estimated from the target data.

Each aggregate error bound Δ^* consists of three components: the neural network approximation error δ_a^* to the underlying regression function, the stochastic error δ_s^* due to empirical risk minimization, and the error δ_f^* from inferring the latent factors from observed covariates.

6.2 Optimal upper bound

An optimal rate can be further achieved by carefully selecting the hyperparameters to balance the trade-off between the approximation error δ_a^* and the stochastic error δ_s^*. This choice requires understanding the neural network's approximation capability as follows:

$$\sup_{\substack{m^*\in\mathcal{H}(d,l,\mathcal{P},C)\\ \gamma(\mathcal{P})=\gamma^*, \|m^*\|_\infty\leq M^*}} \inf_{g\in\mathcal{G}(d,c1,c_2N,\infty,c_3N^{c_4})} \|g - m^*\|^2_\infty \lesssim N^{-4\gamma^*}$$

for some universal constants c_1-c_4, when d is upper bouneded by a universal constant. See Theorem 4 of Fan & Gu (2023) for a rigorous statement. The following condition lists our choice of hyperparamters.

cond:hyper

Condition 1. *The deep ReLU neural network hyperparameters satisfy that*

1. $r \le \bar{r} \lesssim r+1$;
2. $c_1\{\log n_Q \vee \log[\nu_{\min}(H^Q)]^{-1}\} \le \log T \lesssim \log n_Q,\ c_1\{\log n_P \vee \log[\nu_{\min}(H^P)]^{-1}\} \le \log T^P \lesssim \log n_P$;
3. $c_1 \le L \lesssim 1,\ c_1(n_Q/\log n_Q)^{1/(4\Gamma+2)} \le N \lesssim (n_Q/\log n_Q)^{1/(4\Gamma+2)}$;
4. $c_1 \le L^P \lesssim 1,\ c_1(n_P/\log n_P)^{1/(4\gamma^P+2)} \le N^P \lesssim (n_P/\log n_P)^{1/(4\gamma^P+2)}$.

where $\Gamma = \gamma \mathbf{1}\{n_P \ge n_Q\} + (\gamma \wedge \gamma^P)\mathbf{1}\{n_P < n_Q\}$. *Here,* c_1 *is some universal constant which only depends on* $l, \mathcal{P}, l^P, \mathcal{P}^P$.

The choice of hyperparameters in Condition 1 aligns with the parameters specified in the approximation result in Theorem 4 of Fan & Gu (2023). Note that the asymptotic order of N depends on the sample size, and requires a possibly larger order when $n_Q > n_P$. The intuition is that, when $n_P < n_Q$, the source data sample is too small to significantly accelerate the excess risk convergence rate, and we need a more complex architecture to directly estimate the full target regression function g^Q instead of just h.

Given the choice of hyperparameters, the following theorem shows the optimal rate for our fine-tuning factor-augmented estimator.

thm:fineTuning2

Theorem 3 (Optimal upper bound for fine-tuning factor augmented estimator)**.** *Suppose that Assumptions 1, 3, 4, 6 and 7 hold. Consider the fine-tuning factor-augmented model in* (11)-(15) *with* $r + |\mathcal{J} \cup \mathcal{J}^P| \le c_1$ *for some universal constant* c_1 *that satisfies Condition 1, and*

$$c_2\frac{\log(pn_Q)}{n_Q} \le \lambda \le c_3\frac{\log(pn_Q)}{n_Q}, \quad c_2\frac{\log(pn_P)}{n_P} \le \lambda^P \le c_3\frac{\log(pn_Q)}{n_Q}, \quad \tau^{-1} \ge (n_Q \vee n_P)^{c_4}p$$

for some universal constants c_2-c_4 *independent of* n_Q, n_P *and* p. *Furthermore, suppose that* $n_P \gtrsim (\log p)^{2\gamma^P+1}$. *Then, for any universal constant* $K \ge 1$, *with probability at least* $1-p^{-6K}$ *with respect to the target and source data, for* n_Q *(and thus* n_P*) large enough, we have*

$$\mathcal{E}_Q(\widehat{m}_{FT}) + \widehat{\mathcal{E}}_Q(\widehat{m}_{FT}) \le c_5 K\left\{\Big[\frac{\log(n_Q+n_P)}{n_Q+n_P}\Big]^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2\gamma}{2\gamma+1}} + \frac{\log p}{n_Q} + \frac{1\wedge r}{[\nu^2_{\min}(H^Q)\wedge\nu^2_{\min}(H^P)]p}\right\},$$

where c_5 *is a universal constant that depends only on the constants in Assumptions 3, 4, 6 and Condition 1.*

Theorem 3 indicates that the asymptotic convergence rate for our proposed fine-tuning factor-augmented estimator is determined by

$$\Big[\frac{\log(n_Q+n_P)}{n_Q+n_P}\Big]^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2\gamma}{2\gamma+1}} + \frac{\log p}{n_Q} + \frac{1\wedge r}{[\nu^2_{\min}(H^Q)\wedge\nu^2_{\min}(H^Q)]p}. \tag{19}$$

eq:fineTuningUpperBound

We now explain every term in the upper bound (19).

- The term $\Big[\frac{\log(n_Q+n_P)}{n_Q+n_P}\Big]^{\frac{2\gamma^P}{2\gamma^P+1}}$ represents the risk from estimating the source regression function g^P, with complexity component γ^P. It benefits from both the target and source data samples, which is a key advantage of our fine-tuning procedure. When $n_P \gg n_Q$, we benefit from fine-tuning, while we still conserve the conventional rate while n_P is small.

- The term $\left(\frac{\log n_Q}{n_Q}\right)^{\frac{2\gamma}{2\gamma+1}}$ captures the risk from estimating the remaining component of g^Q besides g^P, specifically h, which has complexity component γ. The estimation h only relies on the target data.
- The term $\frac{\log p}{n_Q}$ reflects the variable selection uncertainty for $\mathcal{J}$ in the target regression function. The variable selection uncertainty for the source regression function is dominated by the main terms and is omitted due to the condition $n_P \gtrsim (\log p)^{2\gamma^P+1}$.
- The term $\frac{1\wedge r}{[\nu^2_{\min}(H^Q)\wedge\nu^2_{\min}(H^Q)]p}$ represents the risk associated with inferring the latent factors from observed covariates in the target and source domains. When $r \neq 0$ and $\nu^2_{\min}(H^*) \asymp 1$ for $* \in \{Q, P\}$, this p^{-1} term is typically dominated by the other main terms under the high-dimensional regime.

Additionally, Theorem 3 clearly shows how our fine-tuning procedure improves the statistical convergence rate. Without source data, the optimal asymptotic risk rate we can obtain is given in Fan & Gu (2023) as:

$$\Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2(\gamma\wedge\gamma^P)}{2(\gamma\wedge\gamma^P)+1}} + \frac{\log p}{n_Q} + \frac{1\wedge r}{[\nu^2_{\min}(H^Q)\wedge\nu^2_{\min}(H^Q)]p}. \tag{20}$$ eq:fastUpperBound

Comparing (19) with this formula, we observe that when:

$$n_P \gg n_Q, \quad \gamma^P < \gamma, \tag{21}$$ eq:fasterCond

i.e., when the source sample size is large enough, and the source regression function g^P is indeed more complex than the remaining h component, we can accelerate the risk convergence via the fine-tuning estimator. On the other hand, even if the condition (21) is not satisfied, we maintain the optimal convergence rate as in (20), without involving source data. In this sense, our estimator is both "efficient" in improving the rate as in (19) and "robust" in avoiding negative transfer.

Remark 1. *Given $\mathcal{J} = \emptyset$, the term $\frac{\log p}{n_Q}$ in (19) is eliminated since $\lambda|\mathcal{J}| = 0$ in Theorem 2. This case corresponds to a scenario where the idiosyncratic noise does not contribute to learning h, which is possibly because the information of the idiosyncratic noise is fully captured through g^P. Additionally, by letting $\lambda \to \infty$ in (14), (14) simplifies to*

$$\widehat{h}(\cdot), \widehat{\Theta} = \underset{h\in\mathcal{G}(L,r+1,N,M,T)}{\arg\min} \frac{1}{n_Q}\sum_{i=1}^{n^Q}\Big(Y_i - h(\Big[\widetilde{f}_i^Q, \widehat{s}_i\Big])\Big)^2.$$

This optimization corresponds to the FAR-NN procedure mentioned in Fan & Gu (2023), with an added entry $\widehat{s}_i$. Following the proof structure of Theorem 3 and treating $\lambda \to \infty$ with $\mathcal{J} = \emptyset$, we reach the upper bound:

$$\Big[\frac{\log(n_Q+n_P)}{n_Q+n_P}\Big]^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2\gamma}{2\gamma+1}} + \frac{1\wedge r}{[\nu^2_{\min}(H^Q)\wedge\nu^2_{\min}(H^Q)]p}.$$

The formula above mitigates the curse of high dimensionality and still converges as n_Q grows to infinity, even if n_Q is not large compared with p.

6.3 Minimax optimal lower bound

Define a family of distributions of (f, u, X) as

$$\mathcal{P}(p,r,\rho) = \Big\{\mu(f,u,X) : \text{supp}(f) \subset [-1,1]^r, \text{supp}(u) \subset [-1,1]^p, \mathbb{E}[f] = \mathbb{E}[u] = 0, \\ f \text{ and } u \text{ independent, both have independent components,} \\ X = Bf + u \text{ with } \|B\|_{\max} \le 1 \text{ and } \lambda_{\min}(B^\top B) \ge \rho\Big\}. \tag{22}$$

eq:lb-factor-dist

Besides the optimal upper bound obtained in Theorem 3, we also provide the following minimax lower bound for $\mathcal{E}_Q(\widehat{m})$, where $\widehat{m}$ is any general estimator depending on the target and source data.

thm:minimaxLower

Theorem 4 (Minimax optimal lower bound)**.** *Consider the i.i.d. samples* $\{(X_i^Q, y_i^Q)\}_{i\in[n_Q]} \cup \{(X_i^P, y_i^P)\}_{i\in[n_P]}$ *from the model* (5) *with* $\{\epsilon_i^Q\}_{i\in[n_Q]} \cup \{\epsilon_i^P\}_{j\in[n_P]} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$*, the standard normal distribution. Suppose further that*

1. *There exists some* $(\beta^h, d^h) \in \mathcal{P}$ *such that* $\gamma = \gamma(\mathcal{P}) = \frac{\beta^h}{d^h}$ *and* $d^h \le r+1 \lesssim 1$.
2. *There exists some* $(\beta^P, d^P) \in \mathcal{P}$ *such that* $\gamma^P = \gamma(\mathcal{P}^P) = \frac{\beta^P}{d^P}$ *and* $d^P \le r+1 \lesssim 1$.

for some universal constant c_1*. Then, for* n_Q *(and thus* n_P*) large enough, we have*

$$\inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}=\mu_{(f^P,u^P,X^P)}\in\mathcal{P}(p,r,\rho) \\ \text{Assumption 1 holds with } |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) \ge c_2 \left\{\left[(n_Q+n_P)^{-\frac{2\gamma^P}{2\gamma^P+1}} + n_Q^{-\frac{2\gamma}{2\gamma+1}} + \frac{1}{n_Q} + \frac{1}{\rho}\right] \wedge \frac{\log n_Q}{n_Q}\right\} \tag{23}$$

eq:fineLowerBound

for some universal constant c_2 *independent of* n_Q, n_P *and* ρ*, and the infimum is taken over all estimators based on the observations* $\{(X_i^Q, y_i^Q)\}_{i\in[n_Q]} \cup \{(X_i^P, y_i^P)\}_{i\in[n_P]}$.

The fourth term $\frac{1}{\rho}$ in the lower bound (23) reduces to $\frac{1}{p}$ under the pervasiveness assumption (Assumption 5) where

$$\lambda_{\min}((B^Q)^\top B^Q) \asymp \lambda_{\min}((B^P)^\top B^P) \asymp p.$$

It thus matches the term $\frac{1\wedge r}{[\nu^2_{\min}(H^Q)\wedge\nu^2_{\min}(H^P)]p}$ in (19) when $r \neq 0$ and $\nu^2_{\min}(H^*) \asymp 1$ for $* \in \{Q, P\}$. Even when $r = 0$, the term $\frac{1}{p}$ is typically dominated by other main terms in (23) under the high-dimensional regime. Therefore, the upper bound we obtained in Theorem 3 matches the optimal lower bound up to logarithmic factors in n_Q and n_P.

References

Caio Almeida, Jianqing Fan, G. F. & Tang, F. (2023). Can a machine correct option pricing models? *Journal of Business & Economic Statistics*, 41(3), 995–1009.

Cape, J., Tang, M., & Priebe, C. E. (2017). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*.

Fan, J., Gao, C., & Klusowski, J. M. (2023). Robust transfer learning with unreliable source data.

Fan, J. & Gu, Y. (2023). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *Journal of the American Statistical Association*, 0(0), 1–15.

Fan, J. & Liao, Y. (2022). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association*, 117(538), 909–924.

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181–213.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York.

Kohler, M. & Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4), 1875–1897.

Stewart, G. W. & guang Sun, J. (1990). *Matrix Perturbation Theory*. Academic Press.

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York.

Weyl, H. V. (1909). Über beschränkte quadratische formen, deren differenz vollstetig ist. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27, 373–392.

A Overview of Appendix

The appendix is organized as follows

Section B contains all the proofs for the transferred factor estimate in Section 5.

Section C contains all the proofs for the fine-tuning estimator in Section 6.

Section D includes the proofs for all auxiliary results, along with technical lemmas presented in Sections B and C.

B Proofs for the transferred factor estimate in Section 5

app:covariate

B.1 Proof of Theorem 1

For any $* \in \{Q, P, A, TL\}$, define

$$\widehat{V}^* := [\widehat{v}_1^*, \ldots, \widehat{v}_{\overline{r}}^*], \quad \widehat{W}^* := p^{1/2}[\widehat{v}_1^*, \ldots, \widehat{v}_{\overline{r}}^*].$$

Let $S = B^Q(B^Q)^\top$, which has the eigen-decomposition $S = V^Q\Lambda(V^Q)^\top$, where Λ is an $r \times r$ diagonal matrix and $(V^Q)^\top V^Q = I_r$. It follows from the identification condition that $B^Q = V^Q\Lambda^{1/2}$.

A key component of the proof is the Frobenius norm of the matrix $\widehat{\Sigma}^{TL} - S$, i.e., the pertubation bound of the covariance matrix. We need the following technical lemma to bound it.

lemma:factorTransferLemma

Lemma 1. *Under Assumptions 2-5, for some universal constant c_4 we have*

$$\|\widehat{\Sigma}^{TL} - S\|_F \lesssim p\Big(\delta \wedge \varepsilon^A(t)\Big) \tag{24}$$

eq:factorTransferLemma

with probability at least $1 - 6e^{-t}$ for any $t > 0$ such that $\delta \geq c_4\varepsilon^Q(t)$.

Proof of Theorem 1. We first prove the upper bound. Note that

$$\begin{aligned}
\nu_{max}(p^{-1}\widehat{W}^{TL}B^Q) &= \sup_{u \in \mathbb{S}^{\overline{r}-1}} p^{-1}\|(B^Q)^\top \widehat{W}^{TL}u\|_2 \\
&\leq p^{-1/2}\|\Lambda^{1/2}\|_2 \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|(V^{TL})^\top \widehat{V}^{TL}u\|_2 \\
&\leq p^{-1/2}\|\Lambda^{1/2}\|_2 \max\Big\{ \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|(V^Q)^\top \widehat{V}^Q u\|_2, \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|(V^A)^\top \widehat{V}^A u\|_2 \Big\}.
\end{aligned}$$

The pervasiveness assumption (Assumption 5) implies that

$$\Lambda_{ii} \asymp p, \quad \forall i = 1, \ldots, r. \tag{25}$$

eq:LambdaEigenvalues

Hence, it follows from (25) and orthogonality of $V^Q, \widehat{V}^Q, V^A, \widehat{V}^A$ that

$$\begin{aligned}
\nu_{max}(p^{-1}\widehat{W}^{TL}B^Q) &\leq p^{-1/2}\|\Lambda^{1/2}\|_2 \max\Big\{ \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|(V^Q)^\top \widehat{V}^Q u\|_2, \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|(V^A)^\top \widehat{V}^A u\|_2 \Big\} \\
&\lesssim \max\Big\{ \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|(V^Q)^\top \widehat{V}^Q u\|_2, \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|(V^A)^\top \widehat{V}^A u\|_2 \Big\} \\
&\lesssim \sup_{u \in \mathbb{S}^{\overline{r}-1}} \|u\|_2 \\
&\lesssim 1.
\end{aligned}$$

Next, we prove the lower bound. Let $\widehat{V}_r^{TL} \in \mathbb{R}^{p\times r}$ be the leftmost $r \le \overline{r}$ columns, i.e. top-r eigenvectors, of $\widehat{V}^{TL}$, then

$$\nu_{min}(p^{-1}\widehat{W}^{TL}B^Q) \ge p^{-1/2}\nu_{min}(\widehat{V}_r^{TL}B^Q) \ge p^{-1/2}\nu_{min}(\widehat{V}_r^{TL}V^Q)\nu_{min}(\Lambda^{1/2}) \gtrsim \nu_{min}(\widehat{V}_r^{TL}V^Q). \quad (26)$$

where the last inequality follows from (25). Therefore, it suffices to bound $\nu_{min}(\widehat{V}_r^{TL}V^Q)$ from below.

Let $\lambda_1, \dots, \lambda_p$ denote the eigenvalues of S in a non-increasing order, and $\widehat{\lambda}_1^{TL}, \dots, \widehat{\lambda}_p^{TL}$ denote the eigenvalues of $\widehat{\Sigma}^{TL}$ in a non-increasing order. It is obvious that

$$\lambda_r \ge C_1 p, \quad \lambda_{r+1} = 0. \quad (27)$$

For simplicity, we rewrite $p^{-1}\|\widehat{\Sigma}^{TL} - S\|_F$ by $\widehat{\delta}$. Furthermore, the Weyl's Theorem (Weyl, 1909) indicates that

$$|\widehat{\lambda}_i^{TL} - \lambda_i| \le \|\widehat{\Sigma}^{TL} - S\|_2 \le \|\widehat{\Sigma}^{TL} - S\|_F = p\widehat{\delta}, \quad \forall i = 1, \dots, p. \quad (28)$$

We claim that

$$\nu_{min}(\widehat{V}_r^{TL}V^Q) \ge 1 - \frac{2\widehat{\delta}}{C_1}. \quad (29)$$

which is trivial when $\widehat{\delta} > C_1/2$ since in this case $\nu_{min}(\widehat{V}_r^{TL}V^Q)$ is always non-negative. It thus suffices to prove (29) when $\widehat{\delta} \le C_1/2$. Combining (27) and (28), we bound the eigen-gap between $\widehat{\Sigma}^{TL}$ and S by

$$\begin{aligned}
\widetilde{\delta} &= \inf\left\{|\widehat{\lambda}^{TL} - \lambda| \colon \widehat{\lambda}^{TL} \in (-\infty, \widehat{\lambda}_{r+1}^{TL}],\ \lambda \in [\lambda_1, \lambda_r]\right\} \\
&= \lambda_r - \widehat{\lambda}_{r+1}^{TL} \\
&\ge |\lambda_r - \lambda_{r+1}| - |\lambda_{r+1} - \widehat{\lambda}_{r+1}^{TL}| \\
&\ge C_1 p - p\widehat{\delta} \ge C_1 p/2.
\end{aligned}$$

It follows from Theorem V.3.6 in Stewart & guang Sun (1990) that

$$\|\sin\Theta(\widehat{V}_r^{TL}, V^Q)\|_F \le \frac{p\widehat{\delta}}{\widetilde{\delta}} \le \frac{2\widehat{\delta}}{C_1}, \quad (30)$$

Here $\sin\Theta(\widehat{V}_r^{TL}, V^Q))$ is a $r \times r$ diagnoal matrix satisfying

$$[\sin\Theta(\widehat{V}_r^{TL}, V^Q))]^2 + [\cos\Theta(\widehat{V}_r^{TL}, V^Q))]^2 = I_r$$

where $\cos\Theta(\widehat{V}_r^{TL}, V^Q))$ is a $r \times r$ diagnoal matrix of singular values of $(\widehat{V}_r^{TL})^\top V^Q)$. Hence,

$$\begin{aligned}
\nu_{min}^2(\widehat{V}_r^{TL}V^Q) = \|\cos\Theta(\widehat{V}_r^{TL}, V^Q))\|_2^2 &\ge 1 - \|\sin\Theta(\widehat{V}_r^{TL}, V^Q))\|_2^2 \\
&\ge 1 - \|\sin\Theta(\widehat{V}_r^{TL}, V^Q))\|_F^2 \ge 1 - \left(\frac{2\widehat{\delta}}{C_1}\right)^2 \\
&\ge 1 - \frac{2\widehat{\delta}}{C_1}.
\end{aligned}$$

by the fact that $\sqrt{1-x^2} \ge x^2$ for any $x \in [0, 1]$. Thus we have proven (29), and the theorem directly follows by plugging in Lemma 1 into (29). □

B.2 Proof of Proposition 2

Proof. The first statement was proved in Lemma 1. The second statement is obvious by combining (30) and the following known inequality

$$\min_{R\in\mathcal{O}(r)}\|\widehat{V}_r^{TL}R - V^Q\|_F \le \sqrt{2}\|\sin\Theta(\widehat{V}_r^{TL}, V^Q)\|_F.$$

The minimizer, i.e. the best rotation of basis can be given by the singular value decomposition (SVD) of $(\widehat{V}_r^{TL})^\top \widehat{V}_r^{TL}$. See Cape et al. (2017) for details. □

C Proofs for the fine-tuning estimator in Section 6

app:posterior

C.1 Additional technical notations

For any two functions m_1, m_2, define

$$\|\cdot\|_2^2 = \mathbb{E}_{(f^Q,u^Q)}\Big[|m_1 - m_2|^2\Big].$$

Similar to $\mathcal{E}_Q(m)$ and $\widehat{\mathcal{E}}_Q(m)$, for any function $m:\mathbb{R}^p\to\mathbb{R}$, we define

$$\mathcal{E}_P(m) := \mathbb{E}_{(f^P,u^P)}\Big[|m(X^P) - g^P(X^P_{\mathcal{J}^P})|^2\Big],$$
$$\widehat{\mathcal{E}}_P(m) := \frac{1}{n_P}\sum_{i=1}^{n_P}|m(X_i^P) - g^P(X^P_{i,\mathcal{J}^P})|^2.$$

For any function $s:\mathbb{R}^p\to\mathbb{R}$, Define the function class $\mathcal{F}_s$,

$$\mathcal{F}_s = \{m(x; W^Q, g, \Theta, s) = g(\Big[p^{-1}(W^Q)^\top x, \mathrm{trun}_M(\Theta^\top x), s(x)\Big]) :$$
$$g\in\mathcal{G}(L, \overline{r}+N+1, N, M, T), \Theta\in\mathbb{R}^{p\times N}, \|\Theta\|_{\max}\le T\}$$

and

$$\mathcal{F}_{s,\kappa} = \Big\{m\in\mathcal{F}_s : \sum_{i,j}\psi_\tau(\Theta_{i,j})\le\kappa\Big\}$$

for any $\kappa>0$. Similar to the representation of deep ReLU networks, for any $m\in\mathcal{F}_s$, we can write

$$m(x) = \mathcal{L}_{L+1}\circ\bar{\sigma}\circ\mathcal{L}_L\circ\bar{\sigma}\circ\cdots\circ\mathcal{L}_2\circ\bar{\sigma}\circ\mathcal{L}_1\circ\phi\circ\mathcal{L}_0(x, s(x)).$$

Here, $\phi:\mathbb{R}^{\overline{r}+N+1}\to\mathbb{R}^{\overline{r}+N+1}$ satisfies

$$[\phi(v)]_i = \begin{cases} T_M(v_i) & \overline{r} < i \le \overline{r}+N, \\ v_i & \text{otherwise,}\end{cases}$$

and $\mathcal{L}_0$ satisfies that $\mathcal{L}_0(x, s(x)) = (p^{-1}x^\top W, x^\top\Theta, s(x))^\top$. Additionally, define

$$\mathcal{F}_s^0 = \{m\in\mathcal{F}_s : \text{the last column of } W_1 \text{ is all zeros}\},\quad \mathcal{F}_{s,\kappa}^0 = \Big\{m\in\mathcal{F}_s^0 : \sum_{i,j}\psi_\tau(\Theta_{i,j})\le\kappa\Big\}.$$

It is obvious that $\mathcal{F}_s^0\subset\mathcal{F}_s$ and $\mathcal{F}_{s,\kappa}^0\subset\mathcal{F}_{s,\kappa}$. The definition of $\mathcal{F}_s^0$ specifies a sub-family of $\mathcal{F}_s$ where the choice of s does not affect its elements.

Similar to $\mathbb{E}$, let $\mathbb{P}$ denote the probability measure with respect to the target and source data. Finally, we define the following two quantities of interest:

$$v_{n_Q} = \left(N^2 L + N\overline{r}\right) \frac{L \log(TNn_Q)}{n_Q}, \tag{31}$$

eq:fast-rate-v_n

$$\varrho_{n_Q} = \frac{\log(np(N+\overline{r})) + L\log(TN)}{n_Q}. \tag{32}$$

eq:fast-rate-varrho_n

C.2 Technical lemmas

Before starting the proofs fo the main results, we first list a few necessary techncial lemmas as follows

lemma:m-lip-weight

Lemma 2. *Let* $\theta(m) = \left\{\theta, (W_l, b_l)_{l=1}^{L+1}\right\}$ *be the set of parameters for function* $m \in \mathcal{F}_{s,\kappa}$, $\theta(\breve{m}) = \left\{\breve{\theta}, (\breve{W}_l, \breve{b}_l)_{l=1}^{L+1}\right\}$ *be the set of parameters for function* $\breve{m} \in \mathcal{F}_{s,\kappa}$. *Define the parameter distance*

$$d(\theta(m), \theta(\breve{m})) = \max_{1 \le l \le L+1} \left(\|b_l - \breve{b}_l\|_\infty \vee \|W_l - \breve{W}_l\|_{\max} \right).$$

If $M \ge 1$, *then the following holds*

$$\begin{aligned} \|m(x) - \tilde{m}(x)\|_{\infty,[-K,K]^p} \le\ & (M \vee K\|W\|_{\max})(L+1)T^L(N+1)^{L+1} d(\theta(m), \theta(\breve{m})) \\ & + KT^{L+1}N^L(N+\overline{r}+1)p\|\theta - \breve{\theta}\|_{\max}. \end{aligned} \tag{33}$$

eq:lemma:m-lip-weight:result

The following three technical lemmas are restated from Fan & Gu (2023), covering results on the ϵ-net covering number and the empirical process. Their proofs are the same as in Fan & Gu (2023), with only minor changes to the exact values of universal constants, so we omit them here. These changes come from the small difference between (33) and Lemma 8 of Fan & Gu (2023), as we add $s(\cdot)$ as an extra feature entry. Refer to Fan & Gu (2023) for the full proof and further illustrations.

lemma:g-infty-cover-number

Lemma 3 (Restatement of Lemma 7 of Fan & Gu (2023))**.** *There exists a universal constant* c_1 *such that for any* $K > 0$, $\delta > 2\tau KT^{L+1}N^{L+1}(N+\overline{r}+1)p$ *and* $N, L \ge 2$,

$$\begin{aligned} \log \mathcal{N}(\delta, \mathcal{F}_{s,\kappa}, \|\cdot\|_{\infty,[-K,K]^p}) \le c_1 \Bigg\{ & \left(N^2L + N\overline{r}\right)\left[L\log BN + \log\left(\frac{M \vee K\|W\|_{\max}}{\delta} \vee 1\right)\right] \\ & + s\left[L\log(BN) + \log p + \log\left(\frac{K(N+\overline{r})}{\delta} \vee 1\right)\right] \Bigg\}. \end{aligned}$$

lemma:equivalence-population-empirical-l2-regularized

Lemma 4 (Restatement of Lemma 9 of Fan & Gu (2023))**.** *Suppose* $\tilde{m}$ *is a fixed function in* $\mathcal{F}_s$. *Under the conditions of Lemma 6, there exists some universal constants* c_1*-*c_3 *such the event*

$$\mathcal{C}_t = \left\{ \forall m(x; W^Q, g, \Theta, s) \in \mathcal{F}_s,\ \ \frac{1}{2}\|m - \tilde{m}\|_2^2 \le \|m - \tilde{m}\|_{n_Q}^2 + 2\lambda \sum_{i,j} \psi_\tau(\Theta_{ij}) + c_1\left(v_{n_Q} + \rho_{n_Q} + \frac{t}{n_Q}\right) \right\}$$

satisfies $\mathbb{P}[\mathcal{C}_t] \ge 1 - e^{-t}$ *for any* $t > 0$ *as long as* $\lambda \ge c_2\varrho_{n_Q}$ *and* $\tau^{-1} \ge c_3(r+1)b(TN)^{L+1}(N+\overline{r})pn_Q$.

lemma:weighted-empirical-process-regularized

Lemma 5 (Restatement of Lemma 10 of Fan & Gu (2023)). *Suppose $\tilde{m}$ is a fixed function in $\mathcal{F}_s$. Under the conditions of Lemma 6, there exists some universal constants c_1–c_2 such that for any fixed $\epsilon \in (0,1)$, the event*

$$\mathcal{B}_{t,\epsilon} = \Big\{ \forall m(x; W^Q, g, \Theta, s) \in \mathcal{F}_s,\ \frac{4}{n_Q} \sum_{i=1}^{n_Q} \varepsilon_i (m(X_i^Q) - \tilde{m}(X_i^Q)) - \lambda \sum_{i,j} \psi_\tau(\Theta_{ij}) \\ \leq \epsilon \|m - \tilde{m}\|_{n_Q}^2 + \frac{c_1}{\epsilon} \left(v_{n_Q} + \varrho_{n_Q} + \frac{t}{n_Q} \right) \Big\}$$

occurs with probability at least $1 - e^{-t}$ for any $t > 0$ as long as $\lambda \geq \frac{c_2 \varrho_{n_Q}}{\epsilon}$ and $\tau^{-1} \geq 4(r+1) b (TN)^{L+1} (N + \overline{r}) p n_Q$.

C.3 The non-parametric regression error

In this section, we aim to present a general error bound for non-parametric regression. The generality lies in its independence from the fine-tuning procedure. Since fine-tuning is performed in two stages in Section 6, with each stage using independent training data, we can first fix the pretrained signal $\widehat{s}$ to be used as input and focus on deriving the error bound for the remaining estimation steps.

Fixing the pretrained model output plays a crucial role in the proof. The following lemma provides an oracle-type inequality for non-parametric regression, given any pretrained model $s(\cdot)$. This result is inspired by Theorem 2 of Fan & Gu (2023), which establishes an oracle-type inequality for the so-called FAST-NN estimator. However, our result extends it by incorporating an additional pretrained term $s(\cdot)$.

lemma:fast

Lemma 6 (Oracle-type bound for non-parametric regression). *Suppose that Assumptions 3, 4 and 6 holds. Given any $s(\cdot) : \mathbb{R}^p \to [-M, M]$, we consider the factor augmented model as*

$$\widehat{h}(\cdot), \widehat{\Theta} = \underset{h \in \mathcal{G}(L, r+N+1, N, M, T), \Theta \in \mathbb{R}^{p \times N}}{\arg\min} \frac{1}{n_Q} \sum_{i=1}^{n^Q} \Big(Y_i - h(\Big[\widetilde{f}_i^Q, \mathrm{trun}_M(\Theta^\top x_i^Q), s(x_i^Q) \Big]) \Big)^2 + \lambda \sum_{i,j} \psi_\tau(\Theta_{i,j}) \tag{34}$$

eq:FTFAST6

with

$$\widehat{m}(x) = \widehat{h}(\Big[p^{-1} (W^Q)^\top x, \mathrm{trun}_M(\widehat{\Theta}^\top x), s(x) \Big]). \tag{35}$$

eq:FTFAST7

Furthermore, we suppose that

- $N \geq 2(r + |\mathcal{J}|)$,
- $T \geq c_1 [\nu_{min}(H^Q)]^{-1} |\mathcal{J}| (r+1)$
- $\lambda \geq c_2 n_Q^{-1} \Big(\log(p n_Q (N + \overline{r})) + L \log(TN) \Big)$
- $\tau^{-1} \geq c_3 (r+1) p \Big((TN)^{L+1} (N + \overline{r}) n_Q \Big)$

for some universal constants c_1-c_3 and $\overline{r} \geq r$. Define

- $\delta_a = \inf_{g \in \mathcal{G}(L-1, r+|\mathcal{J}|+1, N, M, T)} \sup_{\kappa \in [-M,M]} \|g(f^Q, u_{\mathcal{J}}^Q, \kappa) - h(f^Q, u_{\mathcal{J}}^Q, \kappa)\|_\infty^2 + \|s(X^Q) - g^P(X_{\mathcal{J}^P}^Q)\|_2^2$
- $\delta_a^0 = \inf_{g \in \mathcal{G}(L-1, r+|\mathcal{J} \cup \mathcal{J}^P|, N, M, T)} \|g(f^Q, u_{\mathcal{J} \cup \mathcal{J}^P}^Q) - g^Q(f^Q, u_{\mathcal{J} \cup \mathcal{J}^P}^Q)\|_\infty^2$

- $\delta_f = \frac{|\mathcal{J}|r\cdot\overline{r}}{p\cdot\nu^2_{min}(H^Q)}, \delta_f^0 = \frac{|\mathcal{J}\cup\mathcal{J}^P|r\cdot\overline{r}}{p\cdot\nu^2_{min}(H^Q)}$
- $\delta_s = \frac{(N^2L+N\overline{r})L\log(TNn_Q)}{n_Q} + \lambda|\mathcal{J}|,\ \delta_s^0 = \frac{(N^2L+N\overline{r})L\log(TNn_Q)}{n_Q} + \lambda|\mathcal{J}\cup\mathcal{J}^P|$

Then, with probability at least $1-3e^{-t}$ with respect to the target data, for n_Q large enough, we have

$$\mathcal{E}_Q(\widehat{m}) + \widehat{\mathcal{E}}_Q(\widehat{m}) \le c_4\Big\{\Big(\delta_a+\delta_f+\delta_s\Big)\wedge\Big(\delta_a^0+\delta_f^0+\delta_s^0\Big) + \frac{t}{n_Q}\Big\}$$

where c_4 is a universal constant that depends only on the constants in Assumptions 3, 4 and 6.

A straightforward corollary of Lemma 6 is the case where the original $s = g^P = 0$, $\mathcal{J}^P = \emptyset$, and h in the lemma's context is replaced by g^P, representing the scenario without the extra feature entry $s(\cdot)$. To facilitate its application to the proof of Theorem 2, we convert the non-parametric regression on the target data to that on the source data. This requires no additional effort given the generality of the result for any data sample.

col:fast

Corollary 1. *Suppose that Assumptions 3, 4 and 6 holds. Suppose that*

- $N^P \ge 2(r+|\mathcal{J}^P|)$,
- $T^P \ge c_1[\nu_{min}(H^P)]^{-1}|\mathcal{J}^P|(r+1)$
- $\lambda \ge c_2 n_P^{-1}\Big(\log(pn_P(N^P+\overline{r})) + L^P\log(T^PN^P)\Big)$
- $\tau^{-1} \ge c_3(r+1)p\Big((T^PN^P)^{L^P+1}(N^P+\overline{r})n_P\Big)$

for some universal constants c_1-c_3 and $\overline{r} \ge r$. Then, with probability at least $1-3e^{-t}$ with respect to the source data, for n_P large enough, the pretrained model $\widehat{s}(x)$ satisfies

$$\mathcal{E}_P(\widehat{s}) + \widehat{\mathcal{E}}_P(\widehat{s}) \le c_4\Big\{\Delta^P + \frac{t}{n_P}\Big\},$$

where Δ^P is defined in Theorem 2 and c_4 is a universal constant that depends only on the constants in Assumptions 3, 4 and 6.

C.4 Proof of Theorem 2

Proof. Define the following two quantities

$$\widehat{\delta}_a = \inf_{g\in\mathcal{G}(L-1,r+|\mathcal{J}|+1,N,M,T)}\sup_{\kappa\in[-M,M]}\|g(f^Q,u^Q_{\mathcal{J}},\kappa)-h(f^Q,u^Q_{\mathcal{J}},\kappa)\|^2_\infty + \|\widehat{s}(X^Q)-g^P(X^Q_{\mathcal{J}^P})\|_2^2,$$
$$\widehat{\Delta} = \widehat{\delta}_a + \delta_f^h + \delta_s^h,$$

and the following two events

$$\mathcal{A} = \left\{\mathcal{E}_Q(\widehat{m}_{FT}) + \widehat{\mathcal{E}}_Q(\widehat{m}_{FT}) \le c_4\Big\{\Delta^Q\wedge\widehat{\Delta} + \frac{t}{n_Q}\Big\}\right\},$$
$$\mathcal{B} = \left\{\mathcal{E}_P(\widehat{s}) + \widehat{\mathcal{E}}_P(\widehat{s}) \le c_4\Big\{\Delta^P + \frac{t}{n_P}\Big\}\right\}.$$

By Lemma 6 and Corollary 1, we see that each of $\mathcal{A}$ and $\mathcal{B}$ holds with probability at least $1-3e^{-t}$. Hence, the event $\mathcal{A}\cap\mathcal{B}$ holds with probability at least $1-6e^{-t}$.

Note that by Assumption 7, when $\mathcal{A} \cap \mathcal{B}$ holds,

$$\begin{aligned}\|\widehat{s}(X^Q) - g^P(X^Q_{\mathcal{J}^P})\|_2^2 =& \int |\widehat{s}(X^Q) - \bar{g}^P(X^Q_{\mathcal{J}^P})|^2 d\mu_{X_Q} \\ \lesssim& \int |\widehat{s}(X^P) - \bar{g}^P(X^Q_{\mathcal{J}^P})|^2 d\mu_{X_P} \\ \lesssim& \Delta^P + \frac{t}{n_P}\end{aligned}$$

which further implies $\widehat{\Delta} \lesssim \Delta^h + \Delta^P + \frac{t}{n_P}$. It is easy to check that $\Delta^h \lesssim \Delta^Q$ by the definition. Therefore, when $\mathcal{A} \cap \mathcal{B}$ holds, we have

$$\begin{aligned}\mathcal{E}_Q(\widehat{m}_{FT}) + \widehat{\mathcal{E}}_Q(\widehat{m}_{FT}) \leq& c_4\Big\{\Delta^Q \wedge \widehat{\Delta} + \frac{t}{n_Q}\Big\} \\ \lesssim& \Delta^Q \wedge \Big(\Delta^h + \Delta^P + \frac{t}{n_P}\Big) + \frac{t}{n_Q} \\ \lesssim& \Delta^Q \wedge \Big(\Delta^P + \frac{t}{n_P}\Big) + \Delta^h + \frac{t}{n_Q} \\ \leq& 2\Big\{\Delta^Q \wedge \Delta^P + \Delta^h + \frac{t}{n_Q \wedge n_P}\Big\},\end{aligned}$$

which completes the proof. □

C.5 Proof of Theorem 3

Proof. We treat K as a parameter throughout the proof. The problem setting indicates $r \vee |\mathcal{J}^P| \leq c_1$, which further implies

$$\|B^P_{\mathcal{J}^P}\|_2 \leq \|B^P_{\mathcal{J}^P}\|_F \leq \sqrt{|\mathcal{J}^P| r} \leq c_1.$$

Hence, by Proposition 1, we have

$$\begin{aligned}&h \in \mathcal{H}(r + |\mathcal{J}| + 1, l, \mathcal{P}, c_0), \\ &g^P([B^P_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P}) \in \mathcal{H}(r + |\mathcal{J}^P|, l^P + 1, \mathcal{P}^P \cup \{\infty, r + |\mathcal{J}^P|\}, c_0 \vee c_1, \\ &g^Q(f, u_{\mathcal{J} \cup \mathcal{J}^P}) \in \mathcal{H}(r + |\mathcal{J} \cup \mathcal{J}^P|, l + l^P + 1, \mathcal{P} \cup \mathcal{P}^P \cup \{\infty, r + |\mathcal{J}^P|\}, c_0 \vee c_1), \\ &\gamma(\mathcal{P}) = \gamma, \quad \gamma(\mathcal{P}^P \cup \{\infty, r + |\mathcal{J}^P|) = \gamma^P, \quad \gamma(\mathcal{P} \cup \mathcal{P}^P \cup \{\infty, r + |\mathcal{J}^P|\}) = \gamma \wedge \gamma^P.\end{aligned} \tag{36}$$

Applying Theorem 4 of Fan & Gu (2023) to (36), we have

$$\delta_a^h \lesssim N^{-4\gamma}, \quad \delta_a^Q \lesssim N^{-4\gamma \wedge \gamma^P}, \quad \delta_a^P \lesssim (N^P)^{-4\gamma^P}.$$

By plugging in our choice of hyperparameters in Condition 1, we obtain

$$\begin{aligned}\Delta^h \lesssim& \Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2\gamma}{2\Gamma+1}} + \Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2\Gamma}{2\Gamma+1}} + \frac{1 \wedge r}{\nu^2_{\min}(H^Q)p}, \\ \Delta^P \lesssim& \Big(\frac{\log n_P}{n_P}\Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \frac{\log n_P}{n_P} + \frac{\log p}{n_P} + \frac{1 \wedge r}{\nu^2_{\min}(H^P)p} \\ \lesssim& \Big(\frac{\log n_P}{n_P}\Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \frac{1 \wedge r}{\nu^2_{\min}(H^P)p}, \\ \Delta^Q \lesssim& \Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2(\gamma \wedge \gamma^P)}{2\Gamma+1}} + \Big(\frac{\log n_Q}{n_Q}\Big)^{\frac{2\Gamma}{2\Gamma+1}} + \frac{\log p}{n_Q} + \frac{1 \wedge r}{\nu^2_{\min}(H^Q)p}.\end{aligned} \tag{37}$$

By letting $t = K \log P$, (18) in Theorem 2 becomes

$$\begin{aligned}\mathcal{E}_Q(\widehat{m}_{FT}) + \widehat{\mathcal{E}}_Q(\widehat{m}_{FT}) \leq & c_4 \Big\{ \Delta^Q \wedge \Big(\Delta^P + \frac{K \log p}{n_P} \Big) + \Delta^h + \frac{K \log p}{n_Q} \Big\} \\ \lesssim & K \Big\{ \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2(\gamma \wedge \gamma^P)}{2\Gamma+1}} \wedge \Big(\frac{\log n_P}{n_P} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\Gamma+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\Gamma}{2\Gamma+1}} \\ & + \frac{\log p}{n_Q} + \frac{1 \wedge r}{[\nu_{\min}^2(H^Q) \wedge \nu_{\min}^2(H^Q)]p} \Big\}\end{aligned}$$

with probability $1 - p^{-6K}$ with respect to the target and source data.

It remains to show that $A \lesssim B$, where

$$\begin{aligned} A :=& \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2(\gamma \wedge \gamma^P)}{2\Gamma+1}} \wedge \Big(\frac{\log n_P}{n_P} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\Gamma+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\Gamma}{2\Gamma+1}}, \\ B :=& \Big[\frac{\log(n_Q + n_P)}{n_Q + n_P} \Big]^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma+1}}.\end{aligned}$$

We divide its proof into three cases as follows

Case 1: $\gamma \leq \gamma^P$**.** In this case, we have $\gamma \wedge \gamma^P = \Gamma = \gamma$, which further implies

$$\begin{aligned} A \lesssim & \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma+1}} \wedge \Big(\frac{\log n_P}{n_P} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma+1}} \\ \lesssim & \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma+1}} \lesssim B.\end{aligned}$$

Case 2: $\gamma > \gamma^P, n_Q > n_P$**.** In this case, we have $\gamma \wedge \gamma^P = \Gamma = \gamma^P$. Hence, we have

$$\begin{aligned} A \lesssim & \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} \wedge \Big(\frac{\log n_P}{n_P} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} \\ \lesssim & \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma^P+1}} \\ \lesssim & \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} \\ \lesssim & \Big[\frac{\log(2n_Q)}{2n_Q} \Big]^{\frac{2\gamma^P}{2\gamma^P+1}} \\ \lesssim & \Big[\frac{\log(n_Q + n_P)}{n_Q + n_P} \Big]^{\frac{2\gamma^P}{2\gamma^P+1}} \lesssim B.\end{aligned}$$

Case 2: $\gamma > \gamma^P, n_Q \leq n_P$**.** In this case, we have $\gamma \wedge \gamma^P = \gamma^P, \Gamma = \gamma$, which implies

$$\begin{aligned} A \lesssim & \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma^P}{2\gamma+1}} \wedge \Big(\frac{\log n_P}{n_P} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma+1}} \\ \lesssim & \Big(\frac{\log n_P}{n_P} \Big)^{\frac{2\gamma^P}{2\gamma^P+1}} + \Big(\frac{\log n_Q}{n_Q} \Big)^{\frac{2\gamma}{2\gamma+1}} \lesssim B.\end{aligned}$$

Combining Cases 1-3 above completes the proof. □

C.6 Proof of Theorem 4

Proof. We always assume that Assumption 1 holds throughout the proof. By setting $g^P \equiv 0$, $g^Q(f, u_{\mathcal{J}}) = h$ belongs to the family $\mathcal{H}(r + |\mathcal{J}|, l, \mathcal{P}, c_0)$, where c_0 is the universal constant defined in Assumption 1. Thus, Lemma 1 of Fan & Gu (2023) shows that

$$\inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}=\mu_{(f^P,u^P,X^P)}\in\mathcal{P}(p,r,\rho) \\ \text{Assumption 1 holds with } |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) \geq \inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}\in\mathcal{P}(p,r,\rho) \\ g^Q(f,u_{\mathcal{J}})\in\mathcal{H}(r+1,l,\mathcal{P},c_0),\ |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) \gtrsim \frac{1}{\rho}.$$

Therefore, it suffices to show that

$$\inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}=\mu_{(f^P,u^P,X^P)}\in\mathcal{P}(p,r,\rho) \\ \text{Assumption 1 holds with } |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) \gtrsim (n_Q + n_P)^{-\frac{2\gamma^P}{2\gamma^P+1}} + n_Q^{-\frac{2\gamma}{2\gamma+1}} + \frac{\log p}{n_Q}. \tag{38}$$

We divide the proof of (38) into the following two cases:

Case 1: $g^P \equiv 0$**.** In this case, the source data has no informative power for g^Q, and we have

$$\begin{aligned}\inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}=\mu_{(f^P,u^P,X^P)}\in\mathcal{P}(p,r,\rho) \\ \text{Assumption 1 holds with } |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) &\geq \inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}\in\mathcal{P}(p,r,\rho) \\ g^Q(f,u_{\mathcal{J}})\in\mathcal{H}(r+1,l,\mathcal{P},c_0),\ |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}(X^Q)) \\ &\geq \inf_{\breve{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}\in\mathcal{P}(p,r,\rho) \\ g^Q(f,u_{\mathcal{J}})\in\mathcal{H}(r+1,l,\mathcal{P},c_0),\ |\mathcal{J}|=1}} \mathcal{E}_Q(\breve{m}(f^Q, u^Q)),\end{aligned}$$

where the last inequality holds because, when constructing $\breve{m}$, we have additional access to the latent factor structures (f^Q, u^Q) and the factor loading matrix B^Q, which provides strictly more information than only knowing X^Q used to construct $\widehat{m}$.

From the theorem setting, we can find $(\beta^h, d^h) \in \mathcal{P}$ such that $\gamma = \frac{\beta^h}{d^h}$ and $d^h \leq r+1$. Therefore, it follows from the well-known minimax optimal lower bound result in Theorem 3.2 of Györfi et al. (2002) for the family $\mathcal{F}_{d^h,\beta^h,c_0} \subset \mathcal{H}(r+1, l, \mathcal{P}, c_0)$ that

$$\inf_{\breve{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)}\in\mathcal{P}(p,r,\rho) \\ g^Q(f,u_{\mathcal{J}})\in\mathcal{H}(r+1,l,\mathcal{P},c_0),\ |\mathcal{J}|=1}} \mathcal{E}_Q(\breve{m}(f^Q, u^Q)) \geq \inf_{\breve{m}} \sup_{\substack{\mu_x\sim\mathrm{Uni}[-1,1]^{d^h} \\ g^Q(x)\in\mathcal{F}(d^h,\beta^h,c_0)}} \mathcal{E}_Q(\breve{m}(x)) \gtrsim n_Q^{-\frac{2\gamma}{2\gamma+1}}, \tag{39}$$

where $\mathrm{Uni}[-1,1]^d$ is the uniform distribution over $[-1,1]^d$ for any $d \geq 1$.

Define the quantity $\delta := \sqrt{\frac{\log p}{2n_Q}}$. For any $j \in [p]$, let $\mathbb{P}_j$ be the probability measure under which $f^Q \sim \mathrm{Uni}[-1,1]^r$ and $u^Q \sim \mathrm{Uni}[-1,1]^p$ are independent random variables, and

$$y^Q = m^{(j)}(u^Q) + \epsilon^Q, \quad m^{(j)}(u) := \delta u_{(j)} \mathbf{1}\{j \geq 1\} \in \mathcal{H}(l, r+1, \mathcal{P}),$$

where $u_{(j)}$ is the j-entry of $u \in \mathbb{R}^p$. It follows from the KL-divergence of two Gaussian random variables that

$$KL(\mathbb{P}_0 \| \mathbb{P}_j) = \frac{1}{2} n_Q \mathbb{E}|m^{(0)} - m^{(j)}(u)|^2 \leq \frac{1}{6} n_Q \delta^2 \Longrightarrow \frac{1}{p}\sum_{j=1}^{p} KL(\mathbb{P}_0 \| \mathbb{P}_j) \leq \frac{1}{6} n_Q \cdot \frac{\log p}{2n_Q} < \frac{1}{8}\log p.$$

Therefore, for any $j \neq j'$, we have

$$\|m^{(j)} - m^{(j')}\| = \sqrt{\mathbb{E}|m^{(j)} - m^{(j')}|^2} = \sqrt{\frac{2}{3}}\delta \geq \frac{\log p}{3n_Q}.$$

By Theorem 2.7 of Tsybakov (2008) where M is replaced by p that

$$\inf_{\breve{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)} \in \mathcal{P}(p,r,\rho) \\ g^Q(f,u_{\mathcal{J}}) \in \mathcal{H}(r+1,l,\mathcal{P},c_0),\ |\mathcal{J}|=1}} \mathcal{E}_Q(\breve{m}(f^Q,u^Q)) \gtrsim \delta^2 \gtrsim \frac{\log p}{n_Q}. \tag{40}$$

Combining (39) and (40), we have

$$\inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)} = \mu_{(f^P,u^P,X^P)} \in \mathcal{P}(p,r,\rho) \\ \text{Assumption 1 holds with } |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) \gtrsim n_Q^{-\frac{2\gamma}{2\gamma+1}} + \frac{\log p}{n_Q}. \tag{41}$$

Case 2: $h(f, u_{\mathcal{J}}, s) \equiv s$**.** In this case, it holds that $g^Q \equiv g^P$, and $(X^Q, y^Q) \sim (X^P, y^P)$ when $\mu_{(f^Q,u^Q,X^Q)} = \mu_{(f^P,u^P,X^P)} \in \mathcal{P}(p,r,\rho)$. Note that

$$\begin{aligned}
\inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)} = \mu_{(f^P,u^P,X^P)} \in \mathcal{P}(p,r,\rho) \\ \text{Assumption 1 holds with } |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) &\geq \inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)} = \mu_{(f^P,u^P,X^P)} \in \mathcal{P}(p,r,\rho) \\ g^Q = g^P,\ g^P(X_{\mathcal{J}^P}) \in \mathcal{H}(|\mathcal{J}^P|,l^P,\mathcal{P}^P,c_0)}} \mathcal{E}_Q(\widehat{m}) \\
&\geq \inf_{\breve{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)} = \mu_{(f^P,u^P,X^P)} \in \mathcal{P}(p,r,\rho) \\ g^Q = g^P,\ g^P(X_{\mathcal{J}^P}) \in \mathcal{H}(|\mathcal{J}^P|,l^P,\mathcal{P}^P,c_0)}} \mathcal{E}_Q(\breve{m}(f^Q,u^Q)) \\
&\geq \inf_{\breve{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)} = \mu_{(f^P,u^P,X^P)} \in \mathcal{P}(p,r,\rho) \\ g^Q = g^P,\ g^P(X_{\mathcal{J}^P}) \in \mathcal{H}(|\mathcal{J}^P|,l^P,\mathcal{P}^P,c_0) \\ B^Q=0, f^Q \equiv 0}} \mathcal{E}_Q(\breve{m}(f^Q,u^Q)) \\
&\geq \inf_{\breve{m}} \sup_{\substack{x^Q,x^P \sim \mathrm{Uni}[-1,1]^p \\ g^Q = g^P,\ g^P(x_{\mathcal{J}^P}) \in \mathcal{H}(|\mathcal{J}^P|,l^P,\mathcal{P}^P,c_0)}} \mathcal{E}_Q(\breve{m}(x^Q))
\end{aligned}$$

where the second inequality holds due to the same reasoning applied in Case 1.

From the theorem setting, we can find $(\beta^P, d^P) \in \mathcal{P}^P$ such that $\gamma^P = \frac{\beta^P}{d^P}$ and $d^P \leq r+1$. Applying Theorem 3.2 of Györfi et al. (2002) again for the family $\mathcal{F}_{d^P,\beta^P,c_0}$, we see that

$$\begin{aligned}
\inf_{\widehat{m}} \sup_{\substack{\mu_{(f^Q,u^Q,X^Q)} = \mu_{(f^P,u^P,X^P)} \in \mathcal{P}(p,r,\rho) \\ \text{Assumption 1 holds with } |\mathcal{J}|=1}} \mathcal{E}_Q(\widehat{m}) &\geq \inf_{\breve{m}} \sup_{\substack{x^Q,x^P \sim \mathrm{Uni}[-1,1]^p \\ g^Q = g^P,\ g^P(x_{\mathcal{J}^P}) \in \mathcal{H}(|\mathcal{J}^P|,l^P,\mathcal{P}^P,c_0)}} \mathcal{E}_Q(\breve{m}(x^Q)) \\
&\geq \inf_{\breve{m}} \sup_{\substack{x^Q,x^P \sim \mathrm{Uni}[-1,1]^p \\ g^Q = g^P,\ g^P(x_{\mathcal{J}^P}) \in \mathcal{H}(|\mathcal{J}^P|,l^P,\mathcal{P}^P,c_0) \\ \mathcal{J}^P = [d^P]}} \mathcal{E}_Q(\breve{m}(x^Q)) \\
&\geq \inf_{\breve{m}} \sup_{\substack{x^Q,x^P \sim \mathrm{Uni}[-1,1]^p \\ g^Q(x) = g^P(x) \in \mathcal{F}_{d^P,\beta^P,c_0}}} \mathcal{E}_Q(\breve{m}(x^Q)) \gtrsim (n_Q + n_P)^{-\frac{2\gamma^P}{2\gamma^P+1}}.
\end{aligned} \tag{42}$$

Therefore, we obtain (38) by combining (41) and (42), thereby completing the proof. □

D Proofs of auxiliary results

app:aux

D.1 Proof of Proposition 1

Proof of Proposition 1. Given the definition of h, there exists $\left\{h_0^{(1)}, \ldots, h_0^{(t)}\right\} \in \mathcal{H}(r+|\mathcal{J}|+1, 1, \mathcal{P}, c_0)$ and $h_1 \in \mathcal{H}(t, l-1, \mathcal{P}, c_0)$ for some positive integer t such that

$$h(f, u_{\mathcal{J}}, z) = h_1\Big(h_0^{(1)}(f, u_{\mathcal{J}}, z), \ldots, h_0^{(t)}(f, u_{\mathcal{J}}, z)\Big)$$

for any $z \in \mathbb{R}$.

Recall that we treat $(f, u_{\mathcal{J}\cup\mathcal{J}^P})$ as the function input. Thus, each coordinate of $(f, u_{\mathcal{J}^P})$ (as a subset of $(f, u_{\mathcal{J}\cup\mathcal{J}^P})$) can be viewed as an element in $\mathcal{H}(r+|\mathcal{J}\cup\mathcal{J}^P|, l^P, \mathcal{P}\cup\mathcal{P}^P, c_0)$. Note that

$$[B^P_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P} \in \mathcal{F}_{r+|\mathcal{J}^P|,\infty,\|B^P_{\mathcal{J}^P,:}\|_2}$$

by the definition of the Hölder's smoothness. Therefore, by treating $g^P(X^Q) = g^P([B^P_{\mathcal{J}^P,:}]f^Q + u^Q_{\mathcal{J}^P})$ as a function of $(f^Q, u^Q_{\mathcal{J}^P})$, we have

$$\begin{aligned} g^P([B^P_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P}) \in& \mathcal{H}(r+|\mathcal{J}^P|, l^P+1, \mathcal{P}^P\cup\{\infty, r+|\mathcal{J}^P|\}, c_0 \vee \|B^P_{\mathcal{J}^P,:}\|_2) \\ \subset& \mathcal{H}(r+|\mathcal{J}\cup\mathcal{J}^P|, l^P+1, \mathcal{P}\cup\mathcal{P}^P\cup\{\infty, r+|\mathcal{J}^P|\}, c_0 \vee \|B^P_{\mathcal{J}^P,:}\|_2). \end{aligned}$$

Thus, each coordinate of $f, u_{\mathcal{J}}, g^P([B^Q_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P})$ is an element in $\mathcal{H}(r+|\mathcal{J}\cup\mathcal{J}^P|, l^P, \mathcal{P}\cup\mathcal{P}^P\cup\{\infty, r+|\mathcal{J}^P|\}, c_0 \vee \|B^P_{\mathcal{J}^P,:}\|_2)$. We further define

$$\begin{aligned} m^{(j)}\Big(f, u_{\mathcal{J}\cup\mathcal{J}^P}\Big) :=& h^{(j)}\Big(f, u_{\mathcal{J}}, g^P([B^Q_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P})\Big) \\ \in& \mathcal{H}(r+|\mathcal{J}\cup\mathcal{J}^P|, l^P+2, \mathcal{P}\cup\mathcal{P}^P\cup\{\infty, r+|\mathcal{J}^P|\}, c_0 \vee \|B^P_{\mathcal{J}^P,:}\|_2). \end{aligned}$$

By Definition 2, we can rewrite g^Q as

$$g^Q\Big(f, u_{\mathcal{J}}, g^P([B^Q_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P})\Big) = h_1\Big(m^{(1)}(f, u_{\mathcal{J}\cup\mathcal{J}^P}), \ldots, m^{(t)}(f, u_{\mathcal{J}\cup\mathcal{J}^P})\Big),$$

and we see that $g^Q\Big(f, u_{\mathcal{J}}, g^P([B^Q_{\mathcal{J}^P,:}]f + u_{\mathcal{J}^P})\Big) \in \mathcal{H}(r+|\mathcal{J}\cup\mathcal{J}^P|, l+l^P+1, \mathcal{P}\cup\mathcal{P}^P\cup\{\infty, r+|\mathcal{J}^P|\}, c_0 \vee \|B^P_{\mathcal{J}^P,:}\|_2)$. The equality $\gamma(\mathcal{P}\cup\mathcal{P}^P\cup\{\infty, r+|\mathcal{J}^P|\}) = \min\{\gamma, \gamma^P\}$ is trvial by the definition of $\gamma(\cdot)$. □

D.2 Proof of Lemma 1

Proof. Define $\varepsilon^P(t) = r\sqrt{\frac{\log p + t}{n_P}} + r^2\sqrt{\frac{\log r + t}{n_P}} + \frac{1}{\sqrt{p}}$. Firstly, from Lemma 5 of Fan & Gu (2023), we already obtain that there exists a universal constant C_1 such that, by the simple union bound, with probability at least $1 - 6e^{-t}$ both of the following inequalities hold

$$\|\widehat{\Sigma}^Q - S\|_F \le C_1 p \cdot \varepsilon^Q(t), \tag{43}$$

eq:fanLemma5eq1

$$\|\widehat{\Sigma}^P - B^P(B^P)^\top\|_F \le C_1 p \cdot \varepsilon^P(t). \tag{44}$$

eq:fanLemma5eq2

Denote the event that both (43) and (44) hold by E_0. We have that the probability that E_0 holds is at least $1 - 6e^{-t}$. Let S^A be $\frac{n_Q}{n_Q+n_P}S + \frac{n_P}{n_Q+n_P}B^P(B^P)^\top$. By Assumption 2, it is easy to see that

$$\|S^A - S\|_F \le p \cdot \varepsilon. \tag{45}$$

eq:factorTransferLemmaProof1

Then, it follows from (43) and (44) that under the event E_0, we have

$$\begin{aligned}\|\widehat{\Sigma}^P - S^A\|_F &= \|\frac{n_Q}{n_Q+n_P}(\widehat{\Sigma}^Q - S) + \frac{n_P}{n_Q+n_P}(\widehat{\Sigma}^P - B^P(B^P)^\top)\|_F \\ &\le \frac{n_Q}{n_Q+n_P}\|(\widehat{\Sigma}^Q - S)\|_F + \frac{n_P}{n_Q+n_P}\|(\widehat{\Sigma}^P - B^P(B^P)^\top)\|_F \\ &\le C_1 p \cdot (\frac{n_Q}{n_Q+n_P}\varepsilon^Q(t) + \frac{n_P}{n_Q+n_P}\varepsilon^P(t)) \\ &\le C_1 p \cdot \Big(r\sqrt{\frac{\log p + t}{n_Q+n_P}} + r^2\sqrt{\frac{\log r + t}{n_Q+n_P}} + \frac{1}{\sqrt{p}}\Big),\end{aligned} \tag{46}$$

eq:factorTransferLemmaProof2

From (45) and (46), we obtain that under the event E_0 we have

$$\|\widehat{\Sigma}^A - S\|_F \ \le \|\widehat{\Sigma}^P - S^A\|_F \ + \ \|S^A - S\|_F \ \le (C_1 \vee 1)p \cdot \varepsilon^A(t) \tag{47}$$

eq:factorTransferLemmaProof3

Fix $c_4 = 2C_1$, then $\delta \ge 2C_1\varepsilon^Q(t)$ gives $C_1\varepsilon^Q(t) \le \delta/2$. We finish the proof by considering the following two cases separately.

Case I: Suppose $\varepsilon^A(t) > \frac{1}{2(C_1\vee 1)}\delta$. By definition of $\widehat{\Sigma}^{TL}$ we have

$$\|\widehat{\Sigma}^{TL} - \widehat{\Sigma}^Q\|_F \ \le \|\widehat{\Sigma}^Q - \widehat{\Sigma}^P\|_F \mathbf{1}\{p^{-1}\|\widehat{\Sigma}^Q - \widehat{\Sigma}^P\|_F \le \delta\} \le p\delta,$$

so from (43) we have that with probablity at least $1 - 6e^{-t}$

$$\|\widehat{\Sigma}^{TL} - S\|_F \le \|\widehat{\Sigma}^{TL} - \widehat{\Sigma}^Q\|_F + \|\widehat{\Sigma}^Q - S\|_F \ \le p\delta + C_1 p \cdot \varepsilon^Q(t) \le 3p\delta/2 \lesssim p\Big(\delta \wedge \varepsilon^A(t)\Big). \tag{48}$$

eq:factorTransferLemmaProof4

Case II: Suppose $\varepsilon^A(t) \le \frac{1}{2(C_1\vee 1)}\delta$. From (47), we obtain that under the event E_0 we have

$$\begin{aligned}\|\widehat{\Sigma}^A - \widehat{\Sigma}^Q\| &\le \|\widehat{\Sigma}^A - S\|_F + \|\widehat{\Sigma}^Q - S\|_F \\ &\le (C_1 \vee 1)p \cdot \varepsilon^A(t) + C_1 p \cdot \varepsilon^Q(t) \\ &\le \delta/2 + \delta/2 = \delta,\end{aligned} \tag{49}$$

eq:factorTransferLemmaProof5

which implies $\widehat{\Sigma}^{TL} = \widehat{\Sigma}^A$. Therefore, it follows from (47) that, under the event E_0 we have

$$\|\widehat{\Sigma}^{TL} - S\|_F = \|\widehat{\Sigma}^A - S\|_F \le (C_1 \vee 1)p \cdot \varepsilon^A(t) \lesssim p\Big(\delta \wedge \varepsilon^A(t)\Big). \tag{50}$$

eq:factorTransferLemmaProof6

Combining two inequalities (48) and (50) in the different two cases leads to the result. □

D.3 Proof of Lemma 2

The proof follows the same approach as Lemma 8 of Fan & Gu (2023), with modifications only to the first layer $\mathcal{L}_0$, where this paper additionally accounts for the $s(x)$ entry. For the convenience of the reader, we provide the complete proof below.

Proof. For $m \in \mathcal{F}_{s,\kappa}$, we first recursively define

$$m_+^{(l)}(x) = \begin{cases} \phi \circ \mathcal{L}_0(x, s(x)) & l = 1 \\ \mathcal{L}_1 \circ m^{(l-1)}(x) & l = 2 \\ \mathcal{L}_{l-1} \circ \bar{\sigma} \circ m^{(l-1)}(x) & l \in \{3, \cdots, L+1\} \end{cases},$$

and

$$m_-^{(l)}(z) = \begin{cases} \mathcal{L}_{L+1} \circ \bar{\sigma}(z) & l = L+1 \\ m^{(l+1)} \circ \mathcal{L}_l \circ \bar{\sigma}(z) & l \in \{2, \cdots, L\} \\ m^{(l+1)} \circ \mathcal{L}_l(z) & l = 1 \end{cases}.$$

Similarly, we can define the corresponding functions $\breve{m}_+^{(l)}(x) : \mathbb{R}^p \to \mathbb{R}^{d_{l-1}}$ and $\breve{m}_-^{(l)}(z) : \mathbb{R}^{d_{l-1}} \to \mathbb{R}$ for $\breve{m}$ that

$$\breve{m}(x) = \breve{\mathcal{L}}_{L+1} \circ \bar{\sigma} \circ \breve{\mathcal{L}}_L \circ \bar{\sigma} \circ \cdots \circ \breve{\mathcal{L}}_2 \circ \bar{\sigma} \circ \breve{\mathcal{L}}_1 \circ \phi \circ \breve{\mathcal{L}}_0(x, s(x)).$$

We make the following two claims:

$$\|m_-^{(l)}(u) - m_-^{(l)}(v)\|_\infty \leq \begin{cases} (TN)^{L+2-l}\|u-v\|_\infty & l \in \{2, \cdots, L+1\} \\ T^{L+1}N^L(N+\overline{r}+1)\|u-v\|_\infty & l = 1 \end{cases}, \tag{51}$$

eq:m-lip-weight-claim1

and

$$\|m_+^{(l)}(x)\|_\infty \leq (M \vee K\|W\|_{\max})(T(N+1))^{l-1} \qquad \forall l \in \{1, \cdots, L+1\}. \tag{52}$$

eq:m-lip-weight-claim2

Proof of Claim (51). For the first claim (51), note that for any $l \in \{1, \cdots, L\}$,

$$\|\mathcal{L}_l(u) - \mathcal{L}_l(v)\|_\infty \leq \|W_l^\top(u-v)\|_\infty \leq d_{l-1}T\|u-v\|_\infty, \tag{53}$$

eq:m-lip-weight-claim1:proof-induction-ineq

provided that $\|W_l\|_{\max} \leq T$. We prove the claim (51) by induction. For the base case $l = L+1$, the result follows directly from (53) as

$$\|m_-^{(l)}(u) - m_-^{(l)}(v)\|_\infty = \|\mathcal{L}_{L+1}(\bar{\sigma}(u) - \bar{\sigma}(v))\|_\infty \leq d_L T\|u-v\| = TN\|u-v\|.$$

If the claim (51) holds for $l+1$ with $l \in \{2, 3, \cdots, L\}$, then we further have

$$\begin{aligned} \|m_-^{(l)}(u) - m_-^{(l)}(v)\|_\infty &= \|m_-^{(l+1)}(u) \circ \mathcal{L}_l \circ \bar{\sigma}(u) - m_-^{(l+1)}(u) \circ \mathcal{L}_l \circ \bar{\sigma}(v)\|_\infty \\ &\leq (TN)^{L+1-l}\|\mathcal{L}_l \circ \bar{\sigma}(u) - \mathcal{L}_l \circ \bar{\sigma}(v)\|_\infty \\ &\leq (TN)^{L+1-l}(TN)\|\bar{\sigma}(u) - \bar{\sigma}(v)\|_\infty \\ &\leq (TN)^{L+2-l}\|u-v\|_\infty. \end{aligned}$$

Here, the first inequality follows from the induction hypothesis, the second inequality follows from (53), and the third inequality follows from the fact that $|\sigma(x)| \leq |x|$ for the ReLU activation function. The case for $l = 1$ proceeds by

$$\begin{aligned} \|m_-^{(1)}(u) - m_-^{(1)}(v)\|_\infty &= \|m_-^{(2)}(u) \circ \mathcal{L}_1 \circ \bar{\sigma}(u) - m_-^{(2)}(u) \circ \mathcal{L}_1 \circ \bar{\sigma}(v)\|_\infty \\ &\leq (TN)^L\|\mathcal{L}_l \circ \bar{\sigma}(u) - \mathcal{L}_1 \circ \bar{\sigma}(v)\|_\infty \\ &\leq (TN)^L T(N+\overline{r}+1)\|\bar{\sigma}(u) - \bar{\sigma}(v)\|_\infty \\ &\leq (TN)^L T(N+\overline{r}+1)\|u-v\|_\infty \end{aligned}$$

thereby concluding the proof. □

Proof of Claim (52). Our proof will use the fact that

$$\|\mathcal{L}(x)\|_\infty = \|W_l x + b_l\|_\infty \le B + d_{l-1}B\|x\|_\infty \tag{54}$$

provided $l \in \{1, \cdots, L\}$ repeatedly. We also prove (52) by induction. For the base case $l = 1$, it follows from the definition of $\mathcal{L}_0$ and ϕ that

$$\|m_+^{(l)}(x)\|_\infty \le \|\mathcal{L}_0(x, s(x))\|_\infty \le K\|W\|_{\max} \vee M.$$

If (52) holds for $l-1$ with $l \in \{2, \ldots, L+1\}$, then combining (54) with the fact $|\sigma(x)| \le |x|$ gives

$$\begin{aligned}\|m_+^{(l)}(x)\|_\infty \le T + d_{l-1}T\|m_+^{(l-1)}(x)\|_\infty &\le T + (TN)(T(N+1))^{l-2}(K\|W\|_{\max} \vee M) \\ &\le (K\|W\|_{\max} \vee M)(T(N+1))^{l-1}.\end{aligned}$$

which completes the proof of claim (52). □

Now, we return to the proof of Lemma 2. The above construction of $m_+^{(l)}$ and $m_-^{(l)}$ implies that

$$m(x) = m_-^{(l)} \circ m_+^{(l)}(x) = m_-^{(l)} \circ \mathcal{L}_{l-1} \circ \bar{\sigma} \circ m_+^{(l-1)}(x)$$

Consequently, it follows from a series of triangle inequalities that

$$\begin{aligned}|m(x) - \breve{m}(x)| = &\left|\mathcal{L}_{L+1} \circ \bar{\sigma} \circ m_+^{(L+1)}(x) - \breve{\mathcal{L}}_{L+1} \circ \bar{\sigma} \circ m_+^{(L+1)}(x)\right| \\ &+ \sum_{l=1}^{L} \left|\breve{m}_-^{(l+1)} \circ \mathcal{L}_l \circ \bar{\sigma} \circ m_+^{(l)}(x) - \breve{m}_-^{(l+1)} \circ \breve{\mathcal{L}}_l \circ \bar{\sigma} \circ m_+^{(l)}(x)\right| \\ &+ \left|\breve{m}_-^{(1)} \circ \phi \circ \mathcal{L}_0(x, s(x)) - \breve{m}_-^{(1)} \circ \phi \circ \breve{\mathcal{L}}_0(x, s(x))\right|.\end{aligned} \tag{55}$$

Using Claim (51) with $l = 1$ gives

$$\begin{aligned}\|\breve{m}_-^{(1)} \circ \phi \circ \mathcal{L}_0(x, s(x)) - \breve{m}_-^{(1)} \circ \phi \circ \breve{\mathcal{L}}_0(x, s(x))\| &\le T^{L+1}N^L(N + \overline{r} + 1)\|\phi \circ \mathcal{L}_0(x, s(x)) - \phi \circ \breve{\mathcal{L}}_0(x, s(x))\|_\infty \\ &\le T^{L+1}N^L(N + \overline{r} + 1)\|\mathcal{L}_0(x, s(x)) - \breve{\mathcal{L}}_0(x, s(x))\|_\infty.\end{aligned}$$

For any $i \in \{1, \cdots, N\}$ and $x \in [-K, K]^p$, we have

$$\|[\mathcal{L}_0(x, s(x)]_{i+\overline{r}} - [\breve{\mathcal{L}}_0(x, s(x))]_{i+\overline{r}}\| = \|[\Theta]_{:,i}^\top x - [\breve{\Theta}]_{:,i}^\top x\| \le \|x\|_\infty \|[\Theta]_{:,i} - [\breve{\Theta}]_{:,i}\|_1 \le Kp\|\Theta - \breve{\Theta}\|_{\max}, \tag{56}$$

which further indicates

$$\|\breve{m}_-^{(1)} \circ \phi \circ \mathcal{L}_0(x, s(x)) - \breve{m}_-^{(1)} \circ \phi \circ \breve{\mathcal{L}}_0(x, s(x))\| \le KT^{L+1}N^L(N + \overline{r} + 1)p\|\Theta - \breve{\Theta}\|_{\max}. \tag{57}$$

For any $l \in \{1, \cdots, L+1\}$, it follows from a similar argument that

$$\begin{aligned}\|\mathcal{L}_l \circ \bar{\sigma}(z) - \breve{\mathcal{L}}_l \circ \bar{\sigma}(z)\|_\infty &= \|W_l\bar{\sigma}(z) - \breve{W}_l\bar{\sigma}(z) + b_l - \breve{b}_l\|_\infty \\ &\le \|W_l\bar{\sigma}(z) - \breve{W}_l\bar{\sigma}(z)\|_\infty + \|b_l - \breve{b}_l\|_\infty \\ &= \max_{i \in \{1, \cdots, d_l\}} \left|[W_l]_{i,:}\bar{\sigma}(z) - [\breve{W}_l]_{i,:}\bar{\sigma}(z)\right| + \|b_l - \breve{b}_l\|_\infty \\ &\le \max_{i \in \{1, \cdots, d_l\}} \|[W_l]_{i,:} - [\breve{W}_l]_{i,:}\|_1 \|\bar{\sigma}(z)\|_\infty + \|b_l - \breve{b}_l\|_\infty \\ &\le \|W_l - \breve{W}_l\|_{\max} d_{l-1}\|z\|_\infty + \|b_l - \breve{b}_l\|_\infty,\end{aligned}$$

which implies

$$\|\mathcal{L}_l \circ \bar{\sigma}(z) - \check{\mathcal{L}}_l \circ \bar{\sigma}(z)\|_\infty \leq d(\theta(m), \theta(\check{m}))(1 + d_{l-1}\|z\|_\infty). \tag{58}$$

eq:m-lip-weight-lell-sigma-bound

It follows from (52) and (58) with $l = L + 1$ that

$$\begin{aligned}\left|\mathcal{L}_{L+1} \circ \bar{\sigma} \circ m_+^{(L+1)}(x) - \check{\mathcal{L}}_{L+1} \circ \bar{\sigma} \circ m_+^{(L+1)}(x)\right| &\leq d(\theta(m), \theta(\check{m}))(1 + N\|m_+^{(L+1)}(x)\|_\infty) \\ &\leq (M \vee K\|W\|_{\max})T^L(N+1)^{L+1} d(\theta(m), \theta(\check{m})),\end{aligned} \tag{59}$$

eq:proof:lemma:m-lip-weight:decomp2

given that $(T(N+1))^L(M \vee K\|W\|_{\max}) \geq 1$.

In addition, for any $l \in \{1, \cdots, L\}$, combing (51), (52) and (58) gives

$$\begin{aligned}&\left|\check{m}_-^{(l+1)} \circ \mathcal{L}_l \circ \bar{\sigma} \circ m_+^{(l)}(x) - \check{m}_-^{(l+1)} \circ \check{\mathcal{L}}_l \circ \bar{\sigma} \circ m_+^{(l)}(x)\right| \\ \leq &(TN)^{L+1-l}\|\mathcal{L}_l \circ \bar{\sigma} \circ m_+^{(l)}(x) - \check{\mathcal{L}}_l \circ \bar{\sigma} \circ m_+^{(l)}(x)\|_\infty \\ \leq &(TN)^{L+1-l} d(\theta(m), \theta(\check{m}))(1 + d_{l-1}\|m_+^{(l)}(x)\|_\infty) \\ \leq &(TN)^{L+1-l} d(\theta(m), \theta(\check{m}))\left(1 + N(M \vee K\|W\|_{\max})(T(N+1))^{l-1}\right) \\ \leq &(M \vee K\|W\|_{\max})T^L(N+1)^L d(\theta(m), \theta(\check{m})).\end{aligned} \tag{60}$$

eq:proof:lemma:m-lip-weight:decomp3

Plugging (57), (59) and (60) into (55) completes the proof.

□

D.4 Proofs of Lemma 6 and Corollary 1

Proof of Lemma 6. Since the conditions and results do not involve any source data, we omit the superscript Q on $W^Q, B^Q, H^Q, y^Q, f^Q, u^Q, n_Q, \widetilde{f}^Q, \epsilon^Q$ and the subscript Q on $\mathcal{E}_Q, \widehat{\mathcal{E}}_Q$, and n_Q for notational simplicity, emphasizing that this general result is independent of any fine-tuning setting. Furthermore, we abbreviate x^Q as x for simplicity.

Step I: Bound the approximation error for g^Q. The goal of this step is to show that there exists some $m(x; W, \widetilde{\Theta}, \widetilde{g}, s) \in \mathcal{F}_s$ with

$$\|\widetilde{\Theta}\|_0 \leq |\mathcal{J}|\mathbf{1}\{\delta_f + \delta_a \leq \delta_f^0 + \delta_a^0\} + |\mathcal{J} \cup \mathcal{J}^P|\mathbf{1}\{\delta_f + \delta_a > \delta_f^0 + \delta_a^0\}$$

such that

$$\mathcal{E}(m) \lesssim (\delta_f + \delta_a) \wedge (\delta_f^0 + \delta_a^0).$$

From Theorem 2 of Fan & Gu (2023) and the first step of its proof, we observe that there always exists some $m^0(x; W, \widetilde{\Theta}, \widetilde{g}, s) \in \mathcal{F}_s^0 \subset \mathcal{F}_s$ with $\|\widetilde{\Theta}\|_0 \leq |\mathcal{J} \cup \mathcal{J}^P|$ such that

$$\mathcal{E}(m) \lesssim \delta_f^0 + \delta_a^0.$$

This result corresponds to the approximation error bound without involving s when we approximate $g^Q(f^Q, u^Q_{\mathcal{J}\cup\mathcal{J}^P})$. Therefore, it suffices to show that there exists some $m(x; W, \widetilde{\Theta}, \widetilde{g}, s) \in \mathcal{F}_s$ with $\|\widetilde{\Theta}\|_0 \leq |\mathcal{J}|$ such that

$$\mathcal{E}(m) \lesssim \delta_f + \delta_a$$

to obtain the goal of this step. We divide the derivation into two cases regarding whether $r = 0$.

Case 1: $r \geq 1$. Let

$$\widetilde{m}^*(x) = h(H^\dagger \widetilde{f}, x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f}, g^P(x_{\mathcal{J}^P})),$$
$$m^*(x) = h(H^\dagger \widetilde{f}, x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f}, s(x)),$$

where $H^\dagger$ is the Moore–Penrose inverse of H. For any f, u, from the definition of W and H, we have

$$\widetilde{f} = Hf + \xi, \quad \xi = p^{-1}W^\top u,$$

which indicates that

$$\begin{aligned} |\widetilde{m}^*(x) - g^Q(f, u_{\mathcal{J}\cup\mathcal{J}^P})| &\lesssim \|H^\dagger\xi\|_2 + \|[B_{\mathcal{J},:}]H^\dagger\xi\|_2 \\ &\leq (\|[B_{\mathcal{J},:}]\|_2 + 1)\|H^\dagger\|_2\|\xi\|_2 \\ &\leq (\|[B_{\mathcal{J},:}]\|_2 + 1)\|\xi\|_2/\nu_{\min}(H) \\ &\leq (\|[B_{\mathcal{J},:}]\|_F + 1)\|\xi\|_2/\nu_{\min}(H) \\ &\lesssim \frac{\sqrt{|\mathcal{J}|r}}{\nu_{\min}(H)}\|\xi\|_2 \end{aligned} \tag{61}$$

where the first inequality is due to the Lipschitz condition of g^Q in Assumption 6, and the last inequality comes from the boundedness of $\|B\|_{\max}$ in Assumpiton 3. Following the linearity of expectation, we can bound the expectation of $\|\xi\|_2^2$ by

$$\begin{aligned} \mathbb{E}\|\xi\|_2^2 &= p^{-2}\mathbb{E}\Big[\sum_{k=1}^{\overline{r}}\Big(\sum_{j=1}^{p} W_{jk}u_j\Big)^2\Big] \\ &= p^{-2}\sum_{k=1}^{\overline{r}}\sum_{j=1}^{p} W_{jk}^2\mathbb{E}[u_j^2] + \sum_{j\neq j'} W_{jk}W_{j'k}\mathbb{E}[u_j u_{j'}] \\ &\leq \frac{\overline{r}}{p}\max_{j,k}|W_{jk}|\max_j \mathbb{E}[u_j^2] + \frac{\overline{r}}{p^2}\max_{j,k}|W_{jk}|^2\sum_{j\neq j'}|\mathbb{E}[u_j u_{j'}]| \\ &\lesssim \frac{\overline{r}}{p}, \end{aligned} \tag{62}$$

where the last inequality applies Assumptions 3 and 4. From (61), (62), and the Lipschitz continuity of h, we have

$$\begin{aligned} \mathcal{E}(m^*) &= \mathbb{E}|m^*(x) - g^Q(f, u_{\mathcal{J}\cup\mathcal{J}^P})|^2 \\ &\lesssim \mathbb{E}|m^*(x) - \widetilde{m}^*(x)|^2 + \mathbb{E}|\widetilde{m}^*(x) - g^Q(f, u_{\mathcal{J}\cup\mathcal{J}^P})|^2 \\ &\lesssim \|s(x) - g^P(x_{\mathcal{J}^P})\|_2^2 + \frac{|\mathcal{J}|r\overline{r}}{(\nu_{\min}(H))p} \\ &\lesssim \delta_a + \delta_f. \end{aligned} \tag{63}$$

It remains to find some $m \in \mathcal{F}_s$ that approximates m^* well. We choose $g \in \mathcal{G}(L-1, r+|\mathcal{J}|+1, N, M, T)$ that minimzes $\sup_{\kappa\in[-M,M]}\|g(f, u_{\mathcal{J}}, \kappa) - h(f, u_{\mathcal{J}}, \kappa)\|_\infty^2$. Then, we have

$$|g(f, u_{\mathcal{J}}, s(x)) - h(f, u_{\mathcal{J}}, s(x)| \lesssim \sqrt{\delta_a}, \quad \forall (f, u_{\mathcal{J}}) \in [-2b, 2b]^r \times [-2b, 2b]^{|\mathcal{J}|} \tag{64}$$

following the definition of δ_a. Since g is a ReLU network, we can write g as

$$g(f, u_{\mathcal{J}}, s(x)) = \mathcal{L}_{L+1}^g \circ \bar{\sigma} \circ \mathcal{L}_L^g \circ \bar{\sigma} \circ \cdots \circ \bar{\sigma} \circ \mathcal{L}_3^g \circ \bar{\sigma} \circ \mathcal{L}_2^g(f, u_{\mathcal{J}}, s(x)).$$

Denote $\mathcal{J} = \{l_1, \ldots, l_{|\mathcal{J}|}\} \subset \{1, \ldots, p\}$. Then, we construct the approximation m as

$$m(x) = \mathcal{L}_{L+1}^g \circ \bar{\sigma} \circ \mathcal{L}_L^g \circ \bar{\sigma} \circ \cdots \circ \bar{\sigma} \circ \mathcal{L}_3^g \circ \bar{\sigma} \circ \mathcal{L}_2 \circ \bar{\sigma} \circ \mathcal{L}_1 \circ \phi \circ \mathcal{L}_0(x, s(x)),$$

where the $\mathcal{L}_0, \mathcal{L}_1$ and $\mathcal{L}_2$ are constructed as follows

1. For $\mathcal{L}_0 : \mathbb{R}^{p+1} \to \mathbb{R}^{\bar{r}+|\mathcal{J}|+1}$, let

$$\mathcal{L}_0(x, s(x)) = (p^{-1}x^\top W, x^\top \widetilde{\Theta}, s(x))^\top = (\widetilde{f}^\top, x_{\mathcal{J}}^\top, s(x))^\top, \quad \widetilde{\Theta}_{ij} = \mathbf{1}\{i \le |\mathcal{J}|,\ j = l_i\}.$$

   Given the definition of ϕ, we have $\phi \circ \mathcal{L}_0(x, s(x))$ given $M \ge r(b+1) \ge \|x_{\mathcal{J}}\|_\infty$. Moreover, it is trivial that $\|\widetilde{\Theta}\|_0 = |\mathcal{J}|$.

2. For $\mathcal{L}_1 : \mathbb{R}^{\bar{r}+|\mathcal{J}|+1} \to \mathbb{R}^{2(\bar{r}+|\mathcal{J}|+1)}$, let

$$\mathcal{L}_1 \begin{bmatrix} \widetilde{f} \\ x_{\mathcal{J}} \\ s(x) \end{bmatrix} = \begin{bmatrix} H^\dagger & 0 & 0 \\ -[B_{\mathcal{J},:}]H^\dagger & I & 0 \\ -H^\dagger & 0 & 0 \\ [B_{\mathcal{J},:}]H^\dagger & -I & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \widetilde{f} \\ x_{\mathcal{J}} \\ s(x) \\ -s(x) \end{bmatrix} + 0 = \begin{bmatrix} H^\dagger \widetilde{f} \\ x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f} \\ -H^\dagger \widetilde{f} \\ -(x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f}) \\ s(x) \end{bmatrix}$$

3. Suppose that the weights $\mathcal{L}_2^g$ are W_2^g and b_2^g. For $\mathcal{L}_2 : \mathbb{R}^{2(\bar{r}+\mathcal{J}+1)} \to \mathbb{R}^N$, given $u \in \mathbb{R}^r,\ v \in \mathbb{R}^{|\mathcal{J}|}$ let

$$\mathcal{L}_2 \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} W_2^g & -W_2^g \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + b_2^g.$$

It follows from the above construction that

$$\begin{aligned} m(x) &= g\Big(\sigma(H^\dagger \widetilde{f}) - \sigma(-H^\dagger \widetilde{f}), \sigma(x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f}) - \sigma(-(x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f})), \sigma(s(x)) - \sigma(-s(x))\Big) \\ &= g(H^\dagger \widetilde{f}, x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f}, s(x)). \end{aligned}$$

Moreover, all weights of $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_{L+1}$ is bounded by $T \vee (C_1 \frac{|\mathcal{J}|r}{\nu_{\min}(H)})$ for some constant C_1 as $\|\cdot\|_{\max} \le \|\cdot\|_2$.

We are now in the position to upper bound $\mathbb{E}|m(x) - m^*(x)|^2$. Define the event

$$E = \Big\{H^\dagger \widetilde{f} \in [-2b, 2b]^r, x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f} \in [-2b, 2b]^{|\mathcal{J}|}\Big\}.$$

We have

$$\begin{aligned} \mathbb{E}|m(x) - m^*(x)|^2 =& \mathbb{E}|m(x) - m^*(x)|^2 \mathbf{1}_E + \mathbb{E}|m(x) - m^*(x)|^2 \mathbf{1}_{E^c} \\ \le& \mathbb{E}|g(H^\dagger \widetilde{f}, x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f}, s(x)) - h(H^\dagger \widetilde{f}, x_{\mathcal{J}} - [B_{\mathcal{J},:}]H^\dagger \widetilde{f}, s(x))|^2 \mathbf{1}_E \\ &+ \sup\Big(|m(x)| + |m^*(x)|\Big)^2 \mathbb{P}(E^c) \\ \lesssim& \delta_a + (M + M^*)^2 \mathbb{P}(E^c) \qquad \texttt{eq:appEq5} \quad (65) \\ \lesssim& \delta_a + \mathbb{P}(\sqrt{\|H^\dagger \xi\|_2^2 + \|[B_{\mathcal{J},:}]H^\dagger \xi\|_2^2} \ge b) \\ \lesssim& \delta_a + \frac{1}{b^2}\mathbb{E}[\|H^\dagger \xi\|_2^2 + \|[B_{\mathcal{J},:}]H^\dagger \xi\|_2^2] \\ \lesssim& \delta_a + \delta_f. \end{aligned}$$

Combining (63) and (65) yields

$$\mathbb{E}|m(x) - g^Q(f, u_{\mathcal{J}\cup\mathcal{J}^P})|^2 \lesssim \delta_a + \delta_f.$$

To complete Case 1, it remains to verify that $C_1\frac{|\mathcal{J}|r}{\nu_{\min}(H)} \le T$ when $2(|\mathcal{J}|+r) \le N$ to ensure that m belongs to $\mathcal{F}_s$. This condition is established through the padding argument in Section I.1 of Fan & Gu (2023).

Case 1: $r = 0$**.** We have $\delta_f = 0$ and $x = u$ in this case. We choose $g \in \mathcal{G}(L-1, r+|\mathcal{J}|+1, N, M, T)$ that minimzes $\sup_{\kappa\in[-M,M]}\|g(x_{\mathcal{J}}, \kappa) - h(x_{\mathcal{J}}, \kappa)\|_\infty^2$. Then, we have

$$|g(x_{\mathcal{J}}, s(x)) - h(x_{\mathcal{J}}, s(x))| \lesssim \sqrt{\delta_a}, \quad \forall x_{\mathcal{J}} \in [-2b, 2b]^{|\mathcal{J}|}$$

following the definition of δ_a. The construction of m is similar to Case 1. If $|\mathcal{J}| \le N$, it follows from the padding argument in Section I.1 of Fan & Gu (2023) that there exists some $m \in \mathcal{F}_s$ and $\widetilde{g} \in \mathcal{G}(L, \overline{r} + N + 1, N, M, T)$ such that

$$m(x) = \widetilde{g}(\widetilde{f}, \widetilde{\Theta}^\top x, s(x)) = g(x_{\mathcal{J}}, s(x)), \quad \widetilde{\Theta}_{ij} = \mathbf{1}\{i \le |\mathcal{J}|, j = l_i\}.$$

Therefore, we have

$$\mathbb{E}|m(x) - g^Q(x_{\mathcal{J}\cup\mathcal{J}^P})|^2 = \mathbb{E}|m(x) - h(x_{\mathcal{J}}, s(x))|^2 \lesssim \delta_a,$$

which completes the proof of Case 2. We denote this m as $\widetilde{m}$ for the rest of the proof.

Combining Cases 1 and 2 above, we know that there must exist some $\widetilde{m}(x; W, \widetilde{\Theta}, \widetilde{g}, s) \in \mathcal{F}_s$ with

$$\|\widetilde{\Theta}\|_0 \le |\mathcal{J}|\mathbf{1}\{\delta_f + \delta_a \le \delta_f^0 + \delta_a^0\} + |\mathcal{J}\cup\mathcal{J}^P|\mathbf{1}\{\delta_f + \delta_a > \delta_f^0 + \delta_a^0\}$$

such that

$$\mathcal{E}(\widetilde{m}) \lesssim (\delta_f + \delta_a) \wedge (\delta_f^0 + \delta_a^0).$$

Step II: Derive the basic inequality. Define the quantity

$$\Pi := |\mathcal{J}|\mathbf{1}\{\delta_f + \delta_a \le \delta_f^0 + \delta_a^0\} + |\mathcal{J}\cup\mathcal{J}^P|\mathbf{1}\{\delta_f + \delta_a > \delta_f^0 + \delta_a^0\}. \tag{66}$$ eq:fast-rate-Pi

It follows from (34) and (35) that

$$\frac{1}{n}\sum_{i=1}^n (y_i - \widehat{m}(x_i))^2 + \lambda\sum_{i,j}\psi_\tau(\widehat{\Theta}_{ij}) \le \frac{1}{n}\sum_{i=1}^n (y_i - \widetilde{m}(x_i))^2 + \lambda\Pi.$$

Plugging in the formula $y_i = g^Q(f, u_{\mathcal{J}\cup\mathcal{J}^P}) + \epsilon_i$, we further have

$$\|\widehat{m} - g^Q\|_n^2 + \lambda\sum_{i,j}\psi_\tau(\widehat{\Theta}_{ij}) \le \|\widetilde{m} - g^Q\|_n^2 + \frac{2}{n}\sum_{i=1}^n \epsilon_i(\widehat{m}(x_i) - \widetilde{m}(x_i)) + \lambda\Pi. \tag{67}$$ eq:estEq1

From the triangle inequality we have

$$\|\widehat{m} - \widetilde{m}\|_n^2 \le 2\|\widehat{m} - g^Q\|_n^2 + 2\|\widetilde{m} - g^Q\|_n^2. \tag{68}$$ eq:estEq2

Combining (67) and (68), we derive the basic inequality

$$\|\widehat{m} - \widetilde{m}\|_n^2 + 2\lambda\sum_{i,j}\psi_\tau(\widehat{\Theta}_{ij}) \le 4\|\widetilde{m} - g^Q\|_n^2 + \frac{4}{n}\sum_{i=1}^n \epsilon_i(\widehat{m}(x_i) - \widetilde{m}(x_i)) + 2\lambda\Pi. \tag{69}$$ eq:estBasic

Step III: Bound the empirical approximation error. Let $z_i = |\widetilde{m}(x_i) - g^Q(f_i, u_{\mathcal{J}\cup\mathcal{J}^P})|$. It is obvious that $\{z_i\}_{i=1,\ldots,n}$ are i.i.d. samples with

$$|z_i| \leq (M + M^*)^2, \quad \mathrm{var}(z_i) \leq \mathbb{E}|z_i|^2 = \mathcal{E}(\widetilde{m}).$$

Since W is independent of $\{x_i\}_{i=1}^n$, $\{\widetilde{m}(x)_i\}_{i=1}^n$ are also mutually indepedent. Therefore, the Bernstein inequality shows that for any $t > 0$, with probability at least $1 - e^{-t}$ with respect to the target data, we have

$$\begin{aligned}\|\widetilde{m} - g^Q\|_n^2 =& \frac{1}{n}\sum_{i=1}^n |z_i|^2 \lesssim (\frac{1}{n}\sum_{i=1}^n \mathbb{E}[z_i])^2 + C_2\Big(\sqrt{\mathcal{E}(\widetilde{m})\frac{t}{n}} + \frac{t}{n}\Big)\\ \lesssim& \mathcal{E}(\widetilde{m}) + \frac{t}{n} \leq C_3\Big((\delta_f + \delta_a) \wedge (\delta_f^0 + \delta_a^0) + \frac{t}{n}\Big)\end{aligned}$$

for some universal constants C_2 and C_3. Define the event

$$E_1(t) = \Big\{\|\widetilde{m} - g^Q\|_n^2 \leq C_3\Big((\delta_f + \delta_a) \wedge (\delta_f^0 + \delta_a^0) + \frac{t}{n}\Big)\Big\}, \tag{70}$$

then $\mathbb{P}(E_1(t)) \geq 1 - e^{-t}$.

Step IV: Bound the empirical excess risk. Define the event

$$\begin{aligned}E_2(t) = \Big\{\forall m(x; W, g, \Theta, s) \in \mathcal{F}_s,\ \frac{4}{n}\sum_{i=1}^n \varepsilon_i(m(x_i) - \tilde{m}(x_i)) - \lambda\sum_{i,j}\psi_\tau(\Theta_{ij})\\ \leq \frac{1}{2}\|m - \tilde{m}\|_n^2 + 2c_1\left(v_n + \varrho_n + \frac{t}{n}\right)\Big\},\end{aligned}$$

where c_1 is the universal constant defined in Lemma 5. By Lemma 5, we have $\mathbb{P}(E_2(t)) \geq 1 - e^{-t}$ provided $\lambda \geq C_4\varrho_n$ for some universal constant C_4. This is achievable by setting the universal constant c_2 in Lemma 6 sufficiently large.

Hence, given the definition of v_n, ϱ_n, Π (Equation (32), (31) and (66)), It is notable that

$$\varrho_n \lesssim \lambda, \quad (\delta_f + \delta_a) \wedge (\delta_f^0 + \delta_a^0) + v_n + \lambda\Pi \lesssim (\delta_f + \delta_a + \delta_s) \wedge (\delta_f^0 + \delta_a^0 + \delta_s^0).$$

Under $E_1(t) \cap E_2(t)$, (69) thus reduces to

$$\begin{aligned}&\|\widehat{m} - \widetilde{m}\|_n^2 + \lambda\sum_{i,j}\psi_\tau(\widehat{\Theta}_{ij}) \leq 4C_3\Big((\delta_f + \delta_a) \wedge (\delta_f^0 + \delta_a^0)\Big) + \frac{1}{2}\|m - \tilde{m}\|_n^2 + 2c_1\Big(v_n + \varrho_n + \frac{t}{n}\Big) + 2\lambda\Pi\\ \Longrightarrow&\|\widehat{m} - \widetilde{m}\|_n^2 + 2\lambda\sum_{i,j}\psi_\tau(\widehat{\Theta}_{ij}) \leq 8C_3\Big((\delta_f + \delta_a) \wedge (\delta_f^0 + \delta_a^0)\Big) + 4c_1\Big(v_n + \varrho_n + \frac{t}{n}\Big) + 4\lambda\Pi\\ &\lesssim (\delta_f + \delta_a + \delta_s) \wedge (\delta_f^0 + \delta_a^0 + \delta_s^0) + \frac{t}{n},\end{aligned} \tag{71}$$

Combining (70) and (71), we have

$$\|\widehat{m} - g^Q\|_n^2 \lesssim \|\widetilde{m} - g^Q\|_n^2 + \|\widehat{m} - \widetilde{m}\|_n^2 \lesssim (\delta_f^0 + \delta_a^0 + \delta_s^0) + \frac{t}{n} \tag{72}$$

under $E_1(t) \cap E_2(t)$, which completes the proof of upper bounding the empirical excess risk since $\mathbb{P}(E_1(t) \cap E_2(t)) \geq 1 - 2e^{-t}$.

Step IV: Bound the excess risk. It remains the upper bound the excess risk (population error) $\mathcal{E}(\widehat{m}) = \|\widehat{m} - g^Q\|_2^2$ to complete the proof of Lemma 6. Define the event

$$E_3(t) = \left\{ \forall m(x; W^Q, g, \Theta, s) \in \mathcal{F}_s, \quad \frac{1}{2}\|m - \tilde{m}\|_2^2 \leq \|m - \tilde{m}\|_n^2 + 2\lambda \sum_{i,j} \psi_\tau(\Theta_{ij}) + c_1 \left(v_n + \rho_n + \frac{t}{n} \right) \right\},$$

where c_1 is the universal constant defined in Lemma 4. By Lemma 4, as long as $\lambda \geq C_5 \varrho_n$ for some universal constant C_5, we have $\mathbb{P}(E_3(t)) \geq 1 - e^{-t}$. This is achievable by setting the universal constant c_2 in Lemma 6 sufficiently large.

Combining (71) and the definition of $E_3(t)$, we have that under $E_1(t) \cap E_2(t) \cap E_3(t)$,

$$\begin{aligned} &\|\widehat{m} - \widetilde{m}\|_2^2 \lesssim (\delta_f + \delta_a + \delta_s) \wedge (\delta_f^0 + \delta_a^0 + \delta_s^0) + \frac{t}{n} \\ \Longrightarrow &\|\widehat{m} - g^Q\|_2^2 \lesssim \|\widetilde{m} - g^Q\|_2^2 + \|\widehat{m} - \widetilde{m}\|_2^2 \lesssim (\delta_f + \delta_a + \delta_s) \wedge (\delta_f^0 + \delta_a^0 + \delta_s^0) + \frac{t}{n}, \end{aligned}$$

which completes the proof as $\mathbb{P}(E_1(t) \cap E_2(t) \cap E_3(t)) \geq 1 - 3e^{-t}$. □

Proof of Corollary 1. The result follows directly by setting $s = g^P = 0$ and $\mathcal{J}^P = \emptyset$ in Lemma 6. Under this setting, $\delta_a + \delta_f + \delta_s$ is equal to $\delta_a^0 + \delta_f^0 + \delta_s^0$, and the result holds by treating the sample for non-parametric regression as the source data and letting h in Lemma 6 be g^P. □