Econometrics with Pre-Trained Embeddings for Unstructured Data

Yuya Shimizu*

This version: July 19, 2026
[Click here for the latest version]

Abstract

Unstructured data, such as images and text, are increasingly used in empirical economics. Since training machine-learning models on unstructured data is costly, economists often use off-the-shelf pre-trained deep learning models developed by computer scientists to extract embeddings, which are then used as covariates in target economic analyses. Despite the popularity of this practice, its theoretical foundations remain limited. There are two main difficulties. First, the pre-trained model is usually trained on a different dataset and for a different task. Consequently, it is unclear when such a model can be used reliably for the target task. Second, the embedding function is subject to an identification problem, which makes it difficult to analyze the estimation error of the embedding function and its effect on the target task. In this paper, we provide sufficient conditions to overcome these difficulties and derive the convergence rate of machine learning models with pre-trained embeddings. We illustrate the theory through double machine learning applications for estimating parameters of interest, such as partially linear regression with unstructured controls, price elasticity in demand estimation considering the product quality measured by images and text, missing data imputation with unstructured data, and the average treatment effect with unstructured confounders.

Keywords: transfer learning, double machine learning, deep learning, average treatment effect, demand estimation

JEL codes: C45, C14, C55, C21

*yuya.shimizu@wisc.edu, Department of Economics, University of Wisconsin-Madison. I am grateful to Bruce Hansen, Jack Porter, and Xiaoxia Shi for their invaluable advice and encouragement. I thank Naoki Aizawa, Yong Cai, Harold Chiang, Junho Choi, Hugo Freeman, Woosik Gong, Lorenzo Magnolfi, Ashesh Rambachan, Kensuke Sakamoto, Jing Tao, Keyon Vafa, Kenneth West, Kohei Yata, and seminar participants at Wisconsin and ESIF-AIML for helpful comments and suggestions. This work is supported by Richard E. Stockwell Dissertation Fellowship, Richard A. Meese Dissertation Fellowship, and the summer research fellowship from the University of Wisconsin-Madison.

1 Introduction

Unstructured data, such as images and text, are increasingly used in empirical economics. In these applications, economists often use off-the-shelf pre-trained deep neural network models trained by computer scientists to extract embeddings and then use those embeddings as generated covariates in a downstream target task.[1] Embeddings transform unstructured data into structured numerical representations, typically vectors in a lower-dimensional latent space, that preserve relevant information from the original data. From an econometric perspective, this is a two-step estimation problem (Murphy and Topel, 1985; Pagan, 1984). This paper provides sufficient conditions under which this workflow yields valid inference in econometric analyses.

The structured dimensionality reduction can be substantial. A color image is typically represented by the intensity values of its pixels. For example, an image with resolution 224×224 and three RGB color channels has

$$\underbrace{224 \times 224}_{\text{pixels}} \times \underbrace{3}_{\text{RGB channels}} = 150{,}528$$

raw input values. Similarly, a text document is typically represented as a sequence of tokens drawn from a large vocabulary, corresponding to a vector with more than 30,000 dimensions. These raw vectors are ultra-high-dimensional and lack direct economic interpretation in the sense that individual pixel values or naive aggregations of token vectors generally do not correspond to economically meaningful variables. This makes them difficult to use directly in standard empirical specifications, motivating the use of embeddings as lower-dimensional, structured representations of the same underlying information.

Pre-trained embeddings are increasingly used in empirical economics. In some applications, including embeddings can materially change estimates of parameters of interest, including their sign and statistical significance, relative to specifications that omit them.[2] Since training

[1] In the machine learning literature, embeddings are also commonly referred to as extracted features or learned representations.

[2] For example, Avivi (2025) uses patent-text embeddings to control application quality when studying gender bias in patent examination. Dube, Jacobs, Naidu, and Suri (2020) estimates the elasticity of task vacancy-filling speed with respect to rewards, using task-description embeddings as controls. Bajari et al. (2025), Bach, Chernozhukov, Klaassen, Spindler, Teichert-Kluge, and Vijaykumar (2024), and Compiani, Morozov, and Seiler (2025) use product text and image embeddings to control for product quality in hedonic regression and demand estimation. Klaassen, Teichert-Kluge, Bach, Chernozhukov, Spindler, and Vijaykumar (2024) study causal effects with multimodal data. Across these applications, the common assumption is that embeddings summarize economically relevant information in otherwise unstructured inputs and can be used as covariates to control or to impute missing variables. Other applications use product embeddings to measure similarity in product space (Han and Lee, 2025; Magnolfi, McClure, and Sorensen, 2025) and satellite image embeddings to predict local poverty status (Jean, Burke, Xie, Alampay Davis, Lobell, and Ermon, 2016).

such representation models from scratch typically requires an extremely large sample and substantial computational resources, economists often rely on pre-trained embeddings as structured summaries of raw inputs.

Despite the empirical popularity of this workflow, its theoretical foundations remain limited. There are two main theoretical difficulties. First, the pre-trained model is usually trained on a different dataset and for a different task than the economic application; thus, it is unclear whether the pre-trained representation is valid for the target task. Second, the embedding function is usually not identified: for example, we can insert a matrix and its inverse between the last hidden layer of a deep neural network and the output layer without changing the output.

To address these difficulties, we first formalize the setup containing a source task for the computer scientist and a target task for the economist. The embedding function is estimated from the source task and then used in the target task. In practice, the economist can access the pre-trained embedding function from the computer scientist via standard deep learning libraries such as PyTorch, but it is difficult or impossible for the economist to observe the source sample used to train the embedding function. Our theory allows for this practical setting by treating the source sample as unobserved for the economist.[3]

We then show that two conditions are key to justifying the two-step workflow of using pre-trained embeddings in the target task. First, we assume that the source task and the target task share the same approximately true embedding function by requiring that approximation errors of both tasks are small when a shared embedding function is used for both. This condition formalizes the common (implicit) assumption in the empirical literature that the pre-trained embedding function summarizes the relevant information in the unstructured data for the target task. Second, we impose a transferability condition requiring that a neighborhood of the source task identified set is contained in the corresponding target task neighborhood. If the embedding function is the last hidden layer of a deep neural network and the economist uses a linear model on top of it, the transferability condition requires that the coefficient of the target task lies in the linear span of the source-task coefficients associated with the embedding function in population.

To illustrate the importance of the transferability condition, we compare a partially linear model with unstructured controls under two designs in Figure 1, where one design has the target nuisance lying in the span of the source tasks and the other design does not. The in-span design produces an estimator centered near the true parameter. By contrast, the

[3]In the machine learning literature, this is known as a transfer-learning setup in which knowledge learned in the source task is reused in the target task (Pan and Yang, 2010).

out-of-span design exhibits a clear shift, motivating a transferability condition that rules out such non-transferable directions. See Appendix I for details of the simulation.

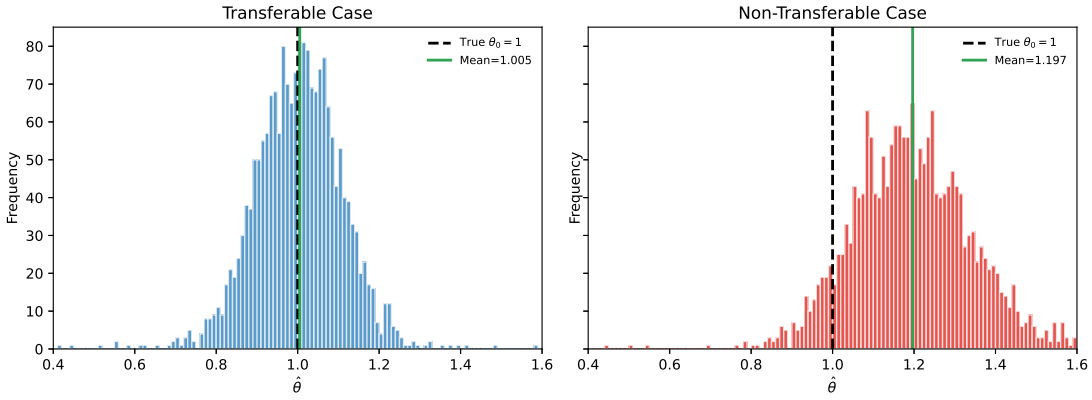

Figure 1: Histogram of $\widehat{\theta}$ for the in-span and out-of-span designs.

Under these conditions, we show that the convergence rate of transfer learning can be faster than $n^{-1/4}$, which is sufficient for valid inference in the double machine learning (DML) framework of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for a size-n target sample. Our convergence rate accounts for randomness in both the source and target samples and depends jointly on the complexities and sample sizes of the source task and the target task. The convergence rate in the target task is determined by the slower of the source-side and target-side rates.

Our theory also provides practical guidance for applied researchers. First, the transferability condition is easier to justify for later hidden layers of a deep neural network than for earlier hidden layers, because earlier hidden layers introduce more complex partial identification issues. Thus, the last hidden layer is preferable as the embedding function if the economist is otherwise indifferent between earlier and last hidden layers. Second, even under the sufficient conditions we impose and when the last hidden layer is selected as the embedding function, the embedding function is still not point identified, and it is only identified up to an invertible linear transformation. Partial identification of the embedding function makes variable selection, such as LASSO, fragile. Instead, we recommend using ridge regression or shallow neural networks for the target model since they are robust to invertible linear transformations of the embeddings.

Related Literature The literature most closely related to ours can be grouped into three strands. First, our paper contributes to the growing literature on inference with variables generated from unstructured data. Fong and Tyler (2021), Angelopoulos, Bates, Fannjiang,

Jordan, and Zrnic (2023), Egami, Hinck, Stewart, and Wei (2023), Zhang, Xue, Yu, and Tan (2026), Battaglia, Christensen, Hansen, and Sacher (2025), Carlson and Dell (2025), and Ludwig, Mullainathan, and Rambachan (2026) provide a framework to correct bias arising from variables generated from unstructured data in the context of missing data. They typically assume a missing completely at random (MCAR) condition or a known propensity score.[4] As discussed by Carlson and Dell (2025), the known propensity score assumption can be relaxed if we can estimate it with sufficiently fast convergence rates. However, the convergence rate of propensity-score estimators based on unstructured data is not studied in these papers. Beyond missing variable imputation, Rambachan, Singh, and Viviano (2024), Modarressi, Spiess, and Venugopal (2025), and Carlson (2025) study settings in which unstructured data are used to construct outcome measures. Our paper departs from this literature by studying a general double machine learning setting in which generated embeddings enter the nuisance learning stage as inputs for both outcome imputation and propensity score estimation.

Second, several recent papers use pre-trained embeddings with theoretical justifications for downstream inference. Vafa, Athey, and Blei (2025) study wage-gap decomposition with foundation models. Bach et al. (2024) show that multimodal embeddings from text and images can improve demand estimation and elasticity analysis. These two papers establish square-root-n consistency for the parameter of interest under high-level conditions. Christensen and Compiani (2026) develop a bias correction in demand counterfactuals when product differentiation is proxied by finite-dimensional nuisance parameters reparameterized from the embedding function and economic parameters.[5] While the scope of Christensen and Compiani (2026) is different from ours, their theory takes convergence of the nuisance parameters as given, and our theory can complement their theory by providing sufficient conditions for the convergence in probability. The closest paper to ours is Schulte, Rügamer, and Nagler (2025). They study the estimation of the average treatment effect (ATE) with unstructured confounders and point out the identification issues of the embedding function. However, they ignore source-task estimation error and assume transferability across tasks implicitly. Our paper differs by deriving a general convergence rate theory under explicit transferability conditions and by incorporating source task estimation error. This source task error is essential for valid target task inference.

Third, our transferability condition is related to the transfer learning literature in machine learning. Tripuraneni, Jordan, and Jin (2020) introduce task diversity as a key condition for

[4]Except for Battaglia et al. (2025), these papers use a missing-variable-imputation framework to handle missing data in the context of Chen, Hong, and Tarozzi (2008).

[5]Such reparameterization is not available for the general DML setup we consider in this paper, which includes Examples 1 to 4.

learning a shared representation that transfers well to new tasks, and Watkins, Ullah, Nguyen-Tang, and Arora (2023) derive optimistic excess-risk bounds for multi-task representation learning under related diversity conditions. The rate in Watkins et al. (2023) is comparable to our rate, but their theory requires a realizability condition that the risk of the true model is zero, which is hard to satisfy in economic applications. Our paper uses a similar task diversity condition to ensure that the source task representation is valid for the target task, but we provide an econometric characterization of the condition through an identification lens. Moreover, we provide a $n^{-1/4}$ convergence rate for the estimated embedding function, which is directly applicable to econometric inference in the target task. In contrast, the machine learning literature typically studies excess-risk bounds for prediction and often obtains slower convergence rates for the estimated embedding function.

Notation For deterministic sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for large n, and we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We define the $L^r(P)$ norm as $\|a\|_{P,r} = \{\sum_{t=1}^T \int |a_t(v)|^r dP(v)\}^{1/r}$ for a function $a : \mathcal{V} \mapsto \mathbb{R}^T$, where a_t is the t-th element of a, and $\|a\|_\infty = \sup_{v \in \mathcal{V}} |a(v)|$ for a function $a : \mathcal{V} \to \mathbb{R}$. For a vector $v \in \mathbb{R}^T$, we define the ℓ^r norm as $\|v\|_r = \{\sum_{t=1}^T |v_t|^r\}^{1/r}$, where v_t is the t-th element of v. For a matrix $A \in \mathbb{R}^{T \times K}$, we denote the t-th largest singular value of A by $\sigma_t(A)$, the largest singular value of A by $\sigma_{\max}(A) = \sigma_1(A)$, and the smallest singular value of A by $\sigma_{\min}(A) = \sigma_{\min\{T,K\}}(A)$. We also define the smallest nonzero singular value of A by $\sigma^+_{\min}(A)$. We define the spectral norm (or operator norm) as $\|A\|_{\text{sp}} = \sigma_{\max}(A)$, the Frobenius norm as $\|A\|_F = \sqrt{\sum_{t=1}^T \sum_{k=1}^K A_{t,k}^2}$, and the nuclear norm (or trace norm) as $\|A\|_* = \sum_{t=1}^{\min\{T,K\}} \sigma_t(A)$. We denote the Moore-Penrose pseudo-inverse of a matrix $A \in \mathbb{R}^{T \times K}$ by $A^+ \in \mathbb{R}^{K \times T}$. For a square matrix $A \in \mathbb{R}^{T \times T}$, we denote the t-th largest eigenvalue of A by $\mu_t(A)$, the largest eigenvalue of A by $\mu_{\max}(A) = \mu_1(A)$, and the smallest eigenvalue of A by $\mu_{\min}(A) = \mu_T(A)$ when all eigenvalues are real.

2 Setup

We first introduce the general DML setup, then four empirical applications, and finally formalize the transfer learning setup for the machine learning components. In the following examples, we consider the setting where the economist has a sample of size n for the target task and the computer scientist provides the pre-trained embeddings $\check{h}(\cdot)$ estimated independently from a different source task. See Section 2.3 for the detailed discussion of the embedding functions.

2.1 Double Machine Learning (DML)

Let $\psi(W_i;\theta,\gamma,\alpha)$ be a moment function, where W_i is an observation from an i.i.d. sample, θ is the parameter of interest, and (γ,α) is an infinite-dimensional nuisance parameter vector. We consider the moment condition:

$$\mathbb{E}[\psi(W_i;\theta^*,\gamma^*,\alpha^*)]=0, \tag{1}$$

where θ^* is the true value of θ, and (γ^*,α^*) is the true value of (γ,α). We assume that γ^* and α^* are unknown functions of Z_i, which is a subvector of W_i containing unstructured or ultra-high-dimensional variables such as images and text. The remaining components of W_i are low-dimensional and structured, such as an outcome variable and a treatment variable.

We assume that the computer scientist provides a pre-trained embedding function $\breve{h}(z)$ estimated from a different source task. In this setting, it is common for the economist to estimate the nuisance parameters (γ^*,α^*) using the pre-trained embeddings $\breve{h}(Z_i)$ as regressors in the target stage models. Our nuisance parameter estimators for γ^* and α^* are $\widehat{f}_\gamma\circ\breve{h}(z)$ and $\widehat{f}_\alpha\circ\breve{h}(z)$, respectively, where $\widehat{f}_\gamma$ and $\widehat{f}_\alpha$ are the estimated target stage models using the pre-trained embeddings $\breve{h}(Z_i)$ as regressors. Here, $\widehat{f}_\gamma\circ\breve{h}(z)$ is the composite function of $\widehat{f}_\gamma$ and $\breve{h}(z)$, defined as $\widehat{f}_\gamma\circ\breve{h}(z)=\widehat{f}_\gamma(\breve{h}(z))$. The DML estimator $\widehat{\theta}$ is the solution of the sample moment equation

$$\frac{1}{n}\sum_{i=1}^{n}\psi(W_i;\widehat{\theta},\widehat{f}_\gamma\circ\breve{h},\widehat{f}_\alpha\circ\breve{h})=0. \tag{2}$$

To use the DML theory, we typically verify three key conditions: (i) Neyman orthogonality with respect to (γ,α),[6] (ii) sample splitting, and (iii) the sufficiently fast convergence rate of the nuisance parameters:[7]

$$\|\gamma^*(Z)-\widehat{f}_\gamma\circ\breve{h}(Z)\|_{P,2}=o_P(n^{-1/4})\quad\text{and}\quad\|\alpha^*(Z)-\widehat{f}_\alpha\circ\breve{h}(Z)\|_{P,2}=o_P(n^{-1/4}), \tag{3}$$

where $\|\cdot\|_{P,2}$ is the $L^2(P)$ norm for the distribution of W_i, and the o_P notation is with respect to the source and target samples used to estimate $(\widehat{f}_\gamma,\widehat{f}_\alpha)$ and $\breve{h}$.

The first condition holds by construction of the score, and the second holds by the sample-splitting procedure. Because the source sample is independent of the target sample and is used

[6] Escanciano and Pérez-Izquierdo (2026) study locally robust estimation with generated regressors. While their theory suggests that the generated regressors affect inference on the parameter of interest in general, this effect of generated regressors does not arise in our setup. We discuss the relationship to their results in Appendix A.2.

[7] We implicitly assume that the Neyman orthogonal score function is smooth enough to use a second-order Taylor expansion. See also Assumption 3.2 in Chernozhukov et al. (2018) and the discussion after that.

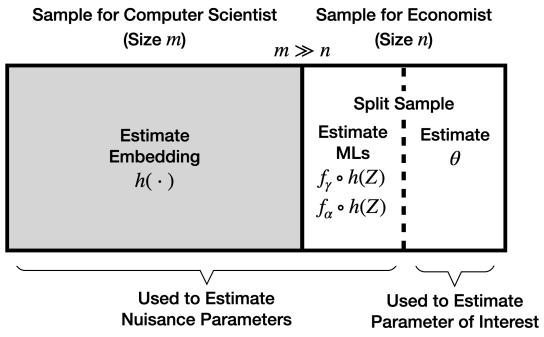

Figure 2: Sample Splitting with Transfer Learning

only to estimate nuisance parameters, the sample-splitting condition holds once the target sample is split between nuisance estimation and parameter estimation (Figure 2). Alternatively, for some machine learning models, we can use the leave-one-out stability condition in Chen, Syrgkanis, and Austern (2022) to avoid sample splitting. Since the remaining DML arguments are the same as in Chernozhukov et al. (2018), we omit the details. Under additional regularity conditions on the moment function in Chernozhukov et al. (2018), the asymptotic normality and consistency of the asymptotic variance estimator are guaranteed. Thus, this paper focuses on the third condition: the convergence rate of the nuisance parameters.

Remark 1. We implicitly assume that the two target stage machine learning models f_γ and f_α use the same embedding function h from the source task for notational simplicity. However, we can easily extend our theory to the case where two target stage machine learning models use different embeddings h_γ and h_α from different source tasks. ■

2.2 Applications

We include four empirical applications to illustrate the DML setup with pre-trained embeddings. Our theory can cover any DML model having the same structure as these examples.

Example 1 (Partially Linear Model with Unstructured Controls.)**.** *We consider the partially linear model* $Y_i = X_i\theta^* + \kappa^*(Z_i) + U_i$ *with* $\mathbb{E}[U_i \mid X_i, Z_i] = 0$*, where* Y_i *is the outcome,* X_i *is a scalar variable,* Z_i *is a vector of unstructured or ultra-high-dimensional control variables such as images and text,* $\kappa^*(\cdot)$ *is an unknown nonparametric function, and* U_i *is the error term. The parameter of interest is the coefficient* θ^* *of* X_i*. This is the classical partially linear setup (Robinson, 1988), now with controls represented by unstructured-data embeddings. We assume that an economist has a sample of size* n*,* $\{(Y_i, X_i, Z_i)\}_{i=1}^n$*, for the target task. The parameter* θ^* *is estimated by solving the Neyman orthogonal score* $\mathbb{E}[\psi(W_i; \theta, \gamma^*, \alpha^*)] = 0$*, where*

$W_i = (Y_i, X_i, Z_i)$, $\psi(W_i; \theta, \gamma, \alpha) = \{(Y_i - \gamma(Z_i)) - (X_i - \alpha(Z_i))\theta\}(X_i - \alpha(Z_i))$, $\gamma^*(Z_i) = \mathbb{E}[Y_i|Z_i]$, *and* $\alpha^*(Z_i) = \mathbb{E}[X_i|Z_i]$.[8] *The DML estimator is given by*

$$\widehat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{f}_\alpha \circ \check{h}(Z_i)\right)^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{f}_\alpha \circ \check{h}(Z_i)\right)\left(Y_i - \widehat{f}_\gamma \circ \check{h}(Z_i)\right)\right). \quad (4)$$

Example 2 (Demand Estimation with Text and Image Data of Products)**.** *Consider a variant of the demand estimation model in Berry (1994), motivated by the recent findings by Bajari et al. (2025) and Compiani et al. (2025) that text and image embeddings of products are informative about consumer choice. Let* θ^* *be the parameter of interest, which is the price elasticity. Let* Y_i *be the logarithm of the relative market share of product* i*,* X_i *be the price of product* i*, and* Z_i *be the unstructured or ultra-high-dimensional control variables of product* i*, such as product images and descriptions. A simple demand estimation model is given by* $Y_i = X_i\theta^* + \kappa^*(Z_i) + U_i$ *and* $\mathbb{E}[U_i \mid X_i, Z_i] = 0$*, where* $\kappa^*(Z)$ *is an unknown nonparametric function and* U_i *is an error term. Under the exogenous price assumption, we can use the partially linear model as in Example 1 to estimate the price elasticity* θ^**. The exogenous price assumption can also be relaxed. Suppose that the price* X_i *is endogenous due to the brand effect, and the economist observes an instrumental variable* V_i *for the price* X_i*, such as a cost shifter. Consider the demand model given by* $Y_i = X_i\theta^* + \gamma^*(Z_i) + U_i$*,* $\mathbb{E}[U_i \mid Z_i, V_i] = 0$*,* $V_i = \alpha^*(Z_i) + e_i$*, and* $\mathbb{E}[e_i \mid Z_i] = 0$*. The parameter of interest* θ^* *is estimated as the solution of the Neyman orthogonal moment equation* $\mathbb{E}[\psi(W_i; \theta, \gamma^*, \alpha^*)] = 0$*, where* $W_i = (Y_i, X_i, V_i, Z_i)$*,* $\psi(W_i; \theta, \gamma, \alpha) = (Y_i - X_i\theta - \gamma(Z_i))(V_i - \alpha(Z_i))$*, and* $\alpha^*(Z_i) = \mathbb{E}[V_i|Z_i]$*.*

Example 3 (Imputation with Unstructured Data)**.** *Suppose that an economist observes a sample of size* n*,* $\{(Y_i, X_i, D_i, Z_i)\}_{i=1}^n$*, and is interested in the regression coefficient* θ^**, which is the solution of the moment* $\mathbb{E}[(Y_i - X_i\theta)X_i] = 0$*, where* Y_i *is the observed outcome and* X_i *is a scalar covariate with missing values.* X_i *is observed if* $D_i = 1$ *and missing if* $D_i = 0$*. The economist can use the unstructured or ultra-high-dimensional variables* Z_i *to predict* X_i*. We assume that* X_i *is missing at random, i.e.,* $X_i \perp\!\!\!\perp D_i \mid Z_i$ *(Rubin, 1976). The parameter* θ^* *is estimated by solving the Neyman orthogonal moment equation* $\mathbb{E}[\psi(W_i; \theta, \gamma^*, \alpha^*)] = 0$*, where* $W_i = (Y_i, X_i, D_i, Z_i)$*,* $\psi(W_i; \theta, \gamma, \alpha) = (D_i/\alpha(Z_i))(Y_iX_i - X_i^2\theta - (\gamma_1(Z_i) - \gamma_2(Z_i)\theta)) + \gamma_1(Z_i) - \gamma_2(Z_i)\theta$*,* $\gamma_1^*(Z_i) = \mathbb{E}[X_iY_i|Z_i]$*,* $\gamma_2^*(Z_i) = \mathbb{E}[X_i^2|Z_i]$*,* $\gamma^* = (\gamma_1^*, \gamma_2^*)$ *and* $\alpha^*(Z_i) = P[D_i = 1|Z_i]$ *is the propensity score of observing* X_i*.*

Example 4 (Average Treatment Effect with Unstructured confounders)**.** *Under the unconfoundedness assumption* $Y_i \perp\!\!\!\perp D_i \mid Z_i$*, the average treatment effect (ATE) is identified by*

[8] Alternatively, we can use the score $\psi(W_i; \theta, \kappa, \alpha) = (Y_i - X_i\theta - \kappa(Z_i))(X_i - \alpha(Z_i))$ with the same nuisance parameters. Chernozhukov et al. (2018) show that both scores are Neyman orthogonal.

$\theta^* = \mathbb{E}[\gamma^*(1, Z_i) - \gamma^*(0, Z_i)]$, *where* Y_i *is the outcome,* D_i *is the binary treatment,* Z_i *denotes the unstructured or ultra-high-dimensional confounders, and* $\gamma^*(d, Z_i) = \mathbb{E}[Y_i \mid D_i = d, Z_i]$ *(Rosenbaum and Rubin, 1983). The ATE is estimated using DML with the Neyman orthogonal moment equation* $\psi(W_i; \theta, \gamma^*, \alpha^*) = 0$*, where* $W_i = (Y_i, D_i, Z_i)$*,* $\psi(W_i; \theta, \gamma, \alpha) = (D_i(Y_i - \gamma(1, Z_i))/(\alpha(Z_i)) - (1 - D_i)(Y_i - \gamma(0, Z_i))/(1 - \alpha(Z_i))) + (\gamma(1, Z_i) - \gamma(0, Z_i) - \theta)$*, and* $\alpha^*(Z_i) = P[D_i = 1|Z_i]$ *is the propensity score of receiving treatment.*

2.3 Transfer Learning

In this subsection, we explain the transfer-learning setup for the DML estimator. We explicitly consider the source task for the computer scientist, together with the target task for the economist.

Hereafter, let P_{so} and P_{ta} denote the source and target task distributions, and let $\mathbb{E}_{\text{so}}$ and $\mathbb{E}_{\text{ta}}$ denote the corresponding expectations. Let P and $\mathbb{E}$ denote the joint distribution of the source and target tasks and its expectation, respectively. Thus, the expectations and probabilities used in the above subsections are under the target task $\mathbb{E}_{\text{ta}}$ and P_{ta} except for the o_P notation, which is with respect to both the source and target tasks.

We consider the setting where a source sample of size m and a target sample of size n share the same approximately true embedding function h_m (Definition 1). Our asymptotics let both the source sample size m and the target sample size n go to infinity, where the rate at which m diverges may depend on n. Formally, we treat the source sample size $m = m_n$ as a sequence indexed by the target sample size n and let $m_n \to \infty$ as $n \to \infty$. Suppose that these samples are generated by $(S_j, Z_j) \sim_{i.i.d.} P_{\text{so}}$ for $j = 1, \ldots, m$ and $(Y_i, Z_i) \sim_{i.i.d.} P_{\text{ta}}$ for $i = 1, \ldots, n$, where (S_j, Z_j) and (Y_i, Z_i) are independent and Z_j and Z_i have different marginal probability distributions. Here, $S_j \in \mathcal{S} \subset \mathbb{R}$ is the outcome for the source task, $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is the outcome for the target task, and $Z_j, Z_i \in \mathcal{Z} \subset \mathbb{R}^d$ denote unstructured or ultra-high-dimensional covariates.

Source Task for Computer Scientist We focus on the most empirically relevant setting where the source task outcome is multi-valued, i.e., $\mathcal{S} = \{1, 2, \ldots, T + 1\}$ for some integer $T \geq 1$. Then, the source task is multi-class classification with $T + 1$ classes and the computer scientist trains the pre-trained model $g \circ h(z) : \mathcal{Z} \mapsto \mathbb{R}^T$ to provide the predicted log-odds for each class $t = 1, \ldots, T$ against the base class $T + 1$. Let $h : \mathcal{Z} \mapsto \mathcal{V} \subset \mathbb{R}^K$ be an embedding function returning $h(z) \in \mathcal{V}$ for some integer $K \geq 1$, and let $g = (g_1, \ldots, g_T) : \mathcal{V} \mapsto \mathbb{R}^T$ be a model for the source task, where $g_t(v)$ is the t-th element of $g(v)$ for $t = 1, \ldots, T$. We normalize the score of the base class as $g_{T+1}(v) = 0$ for every $v \in \mathcal{V}$. Let $\mathcal{G}_m$ denote the

function class of g, $\mathcal{G}_{t,m}$ denote the function class of g_t for each $t = 1, \ldots, T$, and $\mathcal{H}_m$ denote the function class of h for the source task. For example, in a deep neural network model, the model can be decomposed as $g \circ h$, where h is some hidden layer of a deep neural network and g is the remaining transformation from the hidden layer to the output layer. For a given loss function $\ell_{\text{so}} : \mathbb{R}^T \times \mathcal{S} \mapsto \mathbb{R}$, let the set of population minimizers be the solution of the population risk minimization:[9]

$$\mathcal{R}_m = \underset{g \in \mathcal{G}_m, h \in \mathcal{H}_m}{\arg\min} \; \mathbb{E}_{\text{so}}[\ell_{\text{so}}(g \circ h(Z), S)]. \tag{5}$$

For a score vector $a_{\text{so}} = (a_{\text{so},1}, \ldots, a_{\text{so},T})' \in \mathbb{R}^T$, we similarly normalize the base-class score as $a_{\text{so},T+1} = 0$. The most common loss function for multi-class classification is the multinomial logit loss:[10]

$$\ell_{\text{so}}(a_{\text{so}}, S) = -\sum_{t=1}^{T} \mathbb{1}\{S = t\} a_{\text{so},t} + \log\left(1 + \sum_{t=1}^{T} \exp(a_{\text{so},t})\right).$$

The key observation here is that the source task is essentially trained on T different tasks $g_1, \ldots, g_T$ with shared embeddings $h(Z) \in \mathbb{R}^K$ to classify labels with $T + 1$ classes.

The computer scientist observes a sample of size m, $\{(S_j, Z_j)\}_{j=1}^m \sim P_{\text{so}}^m$, for the source task and estimates the pre-trained model $(\breve{g}, \breve{h})$ by some estimation method, such as empirical risk minimization:

$$(\breve{g}, \breve{h}) \in \underset{g \in \mathcal{G}_m, h \in \mathcal{H}_m}{\arg\min} \frac{1}{m} \sum_{j=1}^{m} \ell_{\text{so}}(g \circ h(Z_j), S_j). \tag{6}$$

Target Task for Economist Next, we consider the target task for the economist. The economist will train the target model $f \circ h(z) : \mathcal{Z} \mapsto \mathbb{R}$ to predict the target outcome $Y \in \mathcal{Y}$ using learned embeddings from the source task. Let $f : \mathcal{V} \mapsto \mathbb{R}$ be the model for the target task, where $\mathcal{F}_n$ is the function class of f for the target task. Let $\ell_{\text{ta}} : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}$ be the loss function for the target task such as the least squares loss for regression and the logistic loss for binary classification. Given an embedding function h_m from the source task, let $\mathcal{F}_n^{\text{sub}}(h_m)$ denote the set of solutions to the population risk minimization problem:

$$\mathcal{F}_n^{\text{sub}}(h_m) = \underset{f \in \mathcal{F}_n^{\text{sub}}}{\arg\min} \, \mathbb{E}_{\text{ta}}[\ell_{\text{ta}}(f \circ h_m(Z), Y)], \tag{7}$$

[9] As in Farrell, Liang, and Misra (2021), we assume that minimizer(s) exist for both population and sample risk for simplicity. For the statistical convergence rate theorem, exact existence is not crucial as we can always replace the infimum by a function within a functional class with arbitrarily close risk.

[10] This loss function is also common in the machine learning literature, where it is often called the cross-entropy loss with a softmax output layer.

where $\mathcal{F}_n^{\text{sub}} \subseteq \mathcal{F}_n$ is a subset of a functional class $\mathcal{F}_n$ for the target task, possibly subject to restrictions such as a sparsity restriction for LASSO in a linear model $\mathcal{F}_n$. We often have $\mathcal{F}_n^{\text{sub}} = \mathcal{F}_n$ without additional restrictions, for example for linear regression or shallow neural networks.

The economist observes a sample of size n, $\{Y_i, Z_i\}_{i=1}^n \sim P_{\text{ta}}^n$, and knows the pre-trained model $\breve{h}$ given by the computer scientist, but does not observe the source sample. Since the embedding $\breve{h}(z)$ is often high-dimensional, the economist can use various machine learning models for f in the target task. For example, the economist uses a pre-trained convolutional neural network (CNN) to extract high-dimensional embeddings $\breve{h}(z)$ (e.g., embeddings of dimension 4,096 from a final pooling layer) from ultra-high-dimensional image data (e.g., pixel information with 150,528 dimensions). The extracted embeddings are then used as regressors, for example, in a ridge regression of the outcome Y_i on $\breve{h}(Z_i)$.

Finally, the economist estimates the target model $\widehat{f}$ by empirical risk minimization:[11]

$$\widehat{f} \in \underset{f \in \mathcal{F}_n}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{ta}}(f \circ \breve{h}(Z_i), Y_i). \tag{8}$$

In our theory, we allow small optimization errors for both estimators as shown in (6) and (8) in Theorems 6 and 7.

In summary, a computer scientist provides the pre-trained model $(\breve{g}, \breve{h})$ estimated from the source sample $\{(S_j, Z_j)\}_{j=1}^m$, and an economist uses the embeddings $\breve{h}(Z_i)$ as regressors to estimate the target function $\widehat{f}$ in the target sample $\{(Y_i, Z_i)\}_{i=1}^n$.

Below, we present four examples of the transfer learning setup. Further details of pre-trained deep learning models are provided in Appendix B.

Example 5 (DNN Embeddings for Ultra-High-Dimensional Data)**.** *Consider the setting where the computer scientist trains a deep neural network (DNN) on the source task and the economist uses the learned embeddings as inputs to machine learning models in the target DML task. In this case, the embedding function h is given by the last hidden layer of the DNN, and the pre-trained model g is given by the output layer of the DNN (linear regression). Economists can use various machine learning models, such as shallow neural networks and ridge regression, for the function f in the target task. The data Z can be ultra-high-dimensional, and the DNN can extract high-dimensional embeddings of the data using a source task nonparametric regression problem, where the outcome S is a continuous variable and the loss function ℓ_{so} is the mean squared error loss.*

[11] Throughout the paper, we impose suitable conditions ensuring that minimizers indexed by random variables exist and are measurable.

An autoencoder is a special case of the DNN in Example 5, where the source task is the nonparametric regression task with the same input and output, i.e., $S = Z$, and the loss function ℓ_{so} is a distance function between the input and output, such as the mean squared error loss (Hinton and Salakhutdinov, 2006).

Example 6 (CNN Embeddings for Image Data)**.** *The computer scientist trains a CNN, such as VGG19 (Simonyan and Zisserman, 2015) or ResNet50 (He, Zhang, Ren, and Sun, 2016), on the source task*[12]*, and the economist uses the learned embeddings as inputs to machine learning models in the target DML task as in Example 5. The embedding function h is the last pooling layer of the CNN. The source head g is the classifier attached to that embedding: a three-hidden-layer network with a softmax output in VGG19, or a one-hidden-layer network with a softmax output in ResNet50. The covariate Z is an image, and the CNN can extract high-dimensional embeddings of dimension 4,096 for VGG19 and 1,000 for ResNet50 from the image using the nonparametric classification task in the source task, where the outcome S is an image label (1,000 classes in ImageNet) and the loss function ℓ_{so} is the cross-entropy loss (negative log likelihood).*

Example 7 (Transformer Embeddings for Text Data)**.** *The computer scientist trains a transformer model, such as BERT (Devlin, Chang, Lee, and Toutanova, 2019), on the source task*[13]*, and the economist uses the learned embeddings as inputs to machine learning models in the target DML task as in Example 5. BERT is pre-trained on two tasks: masked language modeling and next sentence prediction. The embedding function $h(\texttt{S}, \texttt{T})$ is given by the last output vectors of the transformer for the input sentence* `S` *and token* `T`*. The token is a masked token* `[MASK]` *in the masked language modeling task or the special token* `[CLS]` *in the next sentence prediction task. To separate the sentences, the special token* `[SEP]` *is inserted between the first sentence and the second sentence. The model g_1 is trained to predict the masked token using the masked language modeling task, and the model g_2 is trained to predict whether the second sentence follows the first sentence using the next sentence prediction task. They are simply linear classifiers on the embeddings $h(\texttt{S}, \texttt{T})$. Both tasks use cross-entropy loss. The masked language modeling label is the original masked token (30,000-label classification), while the next-sentence prediction label is binary. For the target task, economists can use various machine learning models. The data Z are text, and the transformer can extract high-dimensional embeddings of the text using source classification tasks. To extract a text-level embedding of dimension 1,024 for the BERT LARGE model, we can use either the output*

[12] Pre-trained models for both CNNs are available in the torchvision package in PyTorch.
[13] Pre-trained models for BERT are available in the Hugging Face Transformers package in PyTorch.

vector for the special token `[CLS]`*, i.e.,* $h(\texttt{S}, \texttt{[CLS]})$*, or the average of the output vectors for all tokens in the text* `S`*, i.e.,* $|\texttt{S}|^{-1} \sum_{\texttt{T} \in \texttt{S}} h(\texttt{S}, \texttt{T})$.[14]

See Appendix B for details of the functional form of the pre-trained models in the examples above.

Example 8 (Multimodal Embeddings for Image and Text Data)**.** *The economist can also use the multimodal embeddings learned by the computer scientist. For example, the economist can use images and text as unstructured inputs and extract embeddings separately using a pre-trained CNN and a pre-trained transformer.*

3 Transfer Learning Theory

This and the next sections present our main theoretical results on transfer learning for DML. For this purpose, we first introduce the true models for the source and target tasks. Suppose that the source and target tasks admit true models satisfying

$$a_{\text{so}}^* \in \underset{a_{\text{so}}\text{: square-integrable}}{\arg\min} \mathbb{E}_{\text{so}}[\ell_{\text{so}}(a_{\text{so}}(Z), S)]$$

and

$$a_{\text{ta}}^* \in \underset{a_{\text{ta}}\text{: square-integrable}}{\arg\min} \mathbb{E}_{\text{ta}}[\ell_{\text{ta}}(a_{\text{ta}}(Z), Y)],$$

respectively. For example, suppose that the source loss function ℓ_{so} is the multinomial logit loss, $P_{\text{so}}(S = t \mid Z) > 0$ for every $t = 1, \ldots, T+1$, P_{so}-a.s., and the corresponding log-odds functions are square-integrable under P_{so}. Under the normalization $a_{\text{so},T+1}^*(Z) = 0$ and suitable moment conditions, the true source model is almost surely unique and given by the log-odds of the class probabilities, $a_{\text{so},t}^*(Z) = \log(P_{\text{so}}(S = t \mid Z)/P_{\text{so}}(S = T+1 \mid Z))$ for $t = 1, \ldots, T$. For the target task, if the loss is the squared loss $\ell_{\text{ta}}(a_{\text{ta}}(Z), Y) = (1/2) \times (Y - a_{\text{ta}}(Z))^2$ and suitable moment conditions are satisfied, the true model is almost surely unique and given by the conditional expectation of the outcome as $a_{\text{ta}}^*(Z) = \mathbb{E}_{\text{ta}}[Y \mid Z]$, and if the loss is the binary logit loss $\ell_{\text{ta}}(a_{\text{ta}}(Z), Y) = -Y \log(\exp(a_{\text{ta}}(Z))/(1 + \exp(a_{\text{ta}}(Z)))) - (1 - Y) \log(1/(1 + \exp(a_{\text{ta}}(Z))))$ and suitable moment conditions are satisfied, the true model is almost surely unique and given by the log-odds of the conditional probability as $a_{\text{ta}}^*(Z) = \log(P_{\text{ta}}(Y = 1 \mid Z)/P_{\text{ta}}(Y = 0 \mid Z))$. In DML applications, these two loss functions are the most common for the target task.

[14] Whereas the original BERT paper (Devlin et al., 2019) updates all parameters of the pre-trained model for the target task, we only consider the embeddings learned in the source task without updating the parameters of the pre-trained model. Both approaches are called fine-tuning in the machine learning literature. We use the term fine-tuning to mean updating all parameters of the pre-trained model.

In this section, we develop a general convergence rate theory for the target model $\widehat{f} \circ \breve{h}$ by relating it to the source task performance of the pre-trained model $\breve{g} \circ \breve{h}$ relative to the true model a^*_{so}. In Section 3.1, we introduce two key concepts: an approximately true embedding function and task diversity. In Section 3.2, we present the theorem on the convergence rate of the target model in general form. Section 3.3 discusses sufficient and necessary conditions for task diversity.

3.1 Key Concepts

3.1.1 Approximately True Embedding Function

We first introduce the concept of an approximately true embedding function, which is a key definition for the transfer learning theory. In general, a^*_{so} and a^*_{ta} need not be representable as $g_m \circ h_m$ and $f_n \circ h_m$. This mismatch arises from restrictions on the source classes $\mathcal{G}_m, \mathcal{H}_m$ and the target class $\mathcal{F}^{\text{sub}}_n$ even if there exists an embedding function that captures the common features of the source and target tasks well. The following definition quantifies the approximation error for both source and target tasks.

Definition 1 (Approximation Error)**.** For sequences of functions $(g_m, h_m) \in \mathcal{R}_m$ and $f_n \in \mathcal{F}^{\text{sub}}_n(h_m)$, we define the approximation errors for the source and target tasks as

$$\|g_m \circ h_m - a^*_{\text{so}}\|_{P_{\text{so}},2} =: \epsilon_{\text{so},m}, \tag{9}$$

$$\|f_n \circ h_m - a^*_{\text{ta}}\|_{P_{\text{ta}},2} =: \epsilon_{\text{ta},n}. \tag{10}$$

Remark 2. This definition allows $\epsilon_{\text{so},m}$ and $\epsilon_{\text{ta},n}$ to depend on (g_m, h_m) and (f_n, h_m), respectively, but we suppress this dependence in the notation for simplicity. If there exists a true embedding function h^* such that $h^* \in \mathcal{H}_m$ eventually and there exist some measurable functions g^* and f^* such that $a^*_{\text{so}} = g^* \circ h^*$, $a^*_{\text{ta}} = f^* \circ h^*$, then we can interpret h^* as the true embedding function and g^* and f^* as the true source and target models, respectively. Although it is common to assume the existence of the true embedding function in the transfer learning literature and set $\epsilon_{\text{so},m}$ and $\epsilon_{\text{ta},n}$ exactly to zero, we allow the approximate embedding function to cover more general cases where there is no true embedding function, but the source and target models can be well approximated by some embedding function. ■

While we state it as a definition, our result requires that the approximation errors $\epsilon_{\text{so},m}$ and $\epsilon_{\text{ta},n}$ converge to zero as $n \to \infty$ for consistency of the DML estimator in Theorem 6. Decays of $\epsilon_{\text{so},m}$ and $\epsilon_{\text{ta},n}$ are determined by the richness of the function classes $\mathcal{G}_m$ and $\mathcal{F}^{\text{sub}}_n$ and how well the embedding function h_m captures the common features of the source and

target tasks. When the function classes $\mathcal{G}_m$ and $\mathcal{F}_n^{\text{sub}}$ are linear, the approximation errors $\epsilon_{\text{so},m}$ and $\epsilon_{\text{ta},n}$ are determined by the linear projection of a_{so}^* and a_{ta}^* onto the linear span of the embedding function h_m, which is similar to the approximation error in the series regression model (Newey, 1997) using some basis functions instead of the embedding function h_m.

3.1.2 Task Diversity

Next, we introduce the concept of task diversity, which is another key definition. We define task diversity in a way that makes its connection to identification explicit. The reader can refer to Appendix A.1 for a detailed comparison with existing definitions of task diversity in the machine learning literature. For functions $(g_m, h_m) \in \mathcal{R}_m$ and $f_n \in \mathcal{F}_n^{\text{sub}}(h_m)$, define the following approximate identified sets of h_m for $\rho \geq 0$:

$$\mathcal{H}_{\text{so},m}(\rho) := \left\{ h \in \mathcal{H}_m : \min_{g \in \mathcal{G}_m} \|g \circ h - g_m \circ h_m\|_{P_{\text{so}},2} \leq \rho \right\}$$

and

$$\mathcal{H}_{\text{ta},m}(\rho) := \left\{ h \in \mathcal{H}_m : \min_{f \in \mathcal{F}_n^{\text{sub}}} \|f \circ h - f_n \circ h_m\|_{P_{\text{ta}},2} \leq \rho \right\},$$

where we assume that the minimum exists for both the source and target tasks for simplicity. Note that $\mathcal{H}_{\text{so},m}(0)$ and $\mathcal{H}_{\text{ta},m}(0)$ are the identified sets of h_m for the source task and the target task, respectively. If h_m is identified, these sets shrink to $\{h_m\}$ as $\rho \downarrow 0$. In general, however, h_m need not be identified, and these sets can contain multiple elements. It is known that the hidden layers of the deep neural network are identified only up to some transformations, such as the permutation of the order of the hidden units, the scaling of the hidden units, and the rotation of the hidden units (Grigsby, Lindsey, and Rolnick, 2023). See also Example 10 for another example of partial identification of the embedding function for middle hidden layers of a deep neural network.

Definition 2 (Task Diversity)**.** Fix functions $(g_m, h_m) \in \mathcal{R}_m$ and $f_n \in \mathcal{F}_n^{\text{sub}}(h_m)$. For function classes $\mathcal{G}_m$, $\mathcal{H}_m$, and $\mathcal{F}_n^{\text{sub}}$ and some $\rho_{\text{so},m} \geq 0$ and $\rho_{\text{ta},n} \geq 0$, we say that g_m is $(\rho_{\text{so},m}, \rho_{\text{ta},n})$-*diverse over* f_n *with respect to* h_m if

$$\mathcal{H}_{\text{so},m}(\rho_{\text{so},m}) \subseteq \mathcal{H}_{\text{ta},m}(\rho_{\text{ta},n}). \tag{11}$$

Thus, the task diversity condition requires that if h is in a near-identified set of the source task, then h is also in a near-identified set of the target task, where the distances are measured by the $L^2(P_{\text{so}})$- and $L^2(P_{\text{ta}})$-norms, respectively.

The name "task diversity" comes from the intuition that if the T source tasks are sufficiently diverse, then any embedding function h close to the source task identified set should also be close to the target task identified set. This intuition and sufficient and necessary conditions for task diversity are formalized in Section 3.3.

Next, we introduce the concept of ν_n-transferability, which is defined in terms of the task diversity condition.

Definition 3 (ν_n-Transferability)**.** Fix a sequence $\{\nu_n\}_{n=1}^{\infty}$ satisfying $\nu_n > 0$ for every n, sequences of functions $(g_m, h_m) \in \mathcal{R}_m$, and $f_n \in \mathcal{F}_n^{\text{sub}}(h_m)$. For sequences of function classes $\{\mathcal{G}_m\}_{m=1}^{\infty}$, $\{\mathcal{H}_m\}_{m=1}^{\infty}$, and $\{\mathcal{F}_n^{\text{sub}}\}_{n=1}^{\infty}$, we say that g_m is *ν_n-transferable to f_n with respect to h_m at rate δ_m* if for every sequence $\{\rho_{\text{so},m}\}_{m=1}^{\infty}$ such that $\rho_{\text{so},m} = O(\delta_m)$, the $(\rho_{\text{so},m}, \rho_{\text{ta},n})$-diversity condition holds with $\rho_{\text{ta},n} = \rho_{\text{so},m}/\nu_n$ for all large enough n.

ν_n-transferability at rate δ_m requires that the $(\rho_{\text{so},m}, \rho_{\text{ta},n})$-diversity condition holds for all source-side sequences $\rho_{\text{so},m} = O(\delta_m)$ with $\rho_{\text{ta},n} = \rho_{\text{so},m}/\nu_n$. The quantity ν_n captures the identification strength of the target task informed by the source task. Larger ν_n implies that the task g_m is more transferable to f_n since it implies a smaller target task near-identified set given the near-identified set of the source task.

We will relate the sequences $\rho_{\text{so},m}$ to the convergence rates of the source task introduced below. Let $f_n^{\bullet} \in \arg\min_{f \in \mathcal{F}_n^{\text{sub}}} \|f \circ \breve{h} - f_n \circ h_m\|_{P_{\text{ta}},2}$ be the best approximation of the population target model $f_n \circ h_m$ by the estimated embedding function $\breve{h}$. Formally, $f_n^{\bullet}$ depends on $\breve{h}$, but we suppress the dependence for notational simplicity.

Assumption 1 (Convergence Rate of Models)**.**

(i) (Source Model) There exists some sequence $\delta_{\text{so},m} \geq 0$ such that

$$\left\| \breve{g} \circ \breve{h} - g_m \circ h_m \right\|_{P_{\text{so}},2} = O_{P_{\text{so}}}(\delta_{\text{so},m}). \tag{12}$$

(ii) (Target Model given $\breve{h}$) For every sequence of estimated embedding functions $\breve{h} \in \mathcal{H}_m$ satisfying (12), there exists some sequence $r_{\text{ta},n} \geq 0$ such that

$$\left\| \widehat{f} \circ \breve{h} - f_n^{\bullet} \circ \breve{h} \right\|_{P_{\text{ta}},2} = O_P(r_{\text{ta},n}). \tag{13}$$

The first part of this assumption states that the pre-trained model $\breve{g} \circ \breve{h}$ converges to the population source model $g_m \circ h_m$ at some rate $\delta_{\text{so},m}$. The second part states that the target model $\widehat{f} \circ \breve{h}$ converges to the population target model $f_n^{\bullet} \circ \breve{h}$ at some rate $r_{\text{ta},n}$ given the

estimated embedding function $\check{h}$. Note that the O_P notation in the second part is with respect to both the source and target samples since $\check{h}$ depends on the source sample. This assumption is high-level, and we will provide more primitive sufficient conditions in Section 4 to avoid assuming this directly.

Now, we assume the ν_n-transferability condition with the convergence rate of the source task $\delta_{\mathrm{so},m}$.

Assumption 2 (Transferability with Convergence Rates)**.** *For sequences of function classes $\{\mathcal{G}_m\}_{m=1}^{\infty}$, $\{\mathcal{H}_m\}_{m=1}^{\infty}$, and $\{\mathcal{F}_n^{\mathrm{sub}}\}_{n=1}^{\infty}$, there exist $(g_m, h_m) \in \mathcal{R}_m$ and $f_n \in \mathcal{F}_n^{\mathrm{sub}}(h_m)$ such that the task g_m is ν_n-transferable to f_n with respect to h_m at rate $\delta_{\mathrm{so},m}$, where $\delta_{\mathrm{so},m}$ is the convergence rate of the source task defined in Assumption 1.*

Assumption 2 and the quantity ν_n play a key role in determining the convergence rate of the target model $\widehat{f} \circ \check{h}$ as we will see in the next subsection.

3.2 Convergence Rate of Target Model

Our main theorem on the convergence rate of the target model is as follows.

Theorem 1 (Convergence Rate of Target Model)**.** *Suppose that Assumptions 1 and 2 hold. Then, for every sequence of estimated embedding functions $\check{h} \in \mathcal{H}_m$ satisfying (12), we have*

$$\left\| \widehat{f} \circ \check{h} - a_{\mathrm{ta}}^* \right\|_{P_{\mathrm{ta}},2} = O_P\left(r_{\mathrm{ta},n} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n} \right). \tag{14}$$

This theorem states that the convergence rate of the target model $\widehat{f} \circ \check{h}$ to the true model a_{ta}^* is determined by three components: (i) the convergence rate $r_{\mathrm{ta},n}$ of the target model given the estimated embedding function $\check{h}$, (ii) the convergence rate of the source task adjusted by the transferability strength ν_n, and (iii) the approximation error $\epsilon_{\mathrm{ta},n}$ due to an approximately true embedding function. Thus, the rate of the target model is determined by the slowest of these three components. The target-stage rate $r_{\mathrm{ta},n}$ is achieved when transferability is strong, the source task converges sufficiently fast, and the approximation error is sufficiently small.

Remark 3. Applied work using source task embeddings often ignores the estimation error of the pre-trained model and directly applies the DML theory for the target task by treating the estimated embedding function $\check{h}$ as the approximately true embedding function h_m. This theorem provides a theoretical justification for this common practice by showing that the estimation error of the pre-trained model can be negligible if the transferability is strong, the source task converges sufficiently fast, and the approximation error is sufficiently small.

However, there are two important differences between the current practice and the theoretical result in this theorem. First, current practice often ignores the randomness of the estimated embedding function $\check{h}$, while our convergence rate is with respect to joint randomness under P, rather than target sample randomness under P_{ta} alone. Second, applied work often ignores nonidentification of h_m and the associated task diversity condition. Without considering the nonidentification issue, the convergence rate can depend on a specific optimization algorithm used by the computer scientist to estimate the pre-trained model, which is not desirable for the economist. The theorem addresses both issues by imposing transferability and by allowing any estimated embedding function $\check{h}$ in a source task near-identified set.[15] ■

The next theorem gives a useful necessary condition based on the source task identified set $\mathcal{H}_{\text{so},m}(0)$. We will revisit this necessary condition in Section 4.2.2 for the specific case of the LASSO.

Theorem 2 (Necessary Condition for Convergence of Target Model)**.** *Fix sequences of functions* $(g_m, h_m) \in \mathcal{R}_m$ *and* $f_n \in \mathcal{F}_n^{\text{sub}}(h_m)$*. Let* $f^{\bullet}_{n,h_m^\dagger} \in \arg\min_{f \in \mathcal{F}_n^{\text{sub}}} \|f \circ h_m^\dagger - f_n \circ h_m\|_{P_{\text{ta}},2}$ *be the best approximation of the population target model* $f_n \circ h_m$ *by the embedding function* $h_m^\dagger$*. Let* $\widehat{f}_{n,h_m^\dagger} \in \arg\min_{f \in \mathcal{F}_n^{\text{sub}}} \frac{1}{n}\sum_{i=1}^n \ell_{\text{ta}}(f \circ h_m^\dagger(Z_i), Y_i)$ *be the corresponding estimator of the target model. Suppose that, for every deterministic sequence* $h_m^\dagger \in \mathcal{H}_{\text{so},m}(0)$*,*

$$\|\widehat{f}_{n,h_m^\dagger} \circ h_m^\dagger - f^{\bullet}_{n,h_m^\dagger} \circ h_m^\dagger\|_{P_{\text{ta}},2} = o_{P_{\text{ta}}}(1),$$
$$\|\widehat{f}_{n,h_m^\dagger} \circ h_m^\dagger - a^*_{\text{ta}}\|_{P_{\text{ta}},2} = o_{P_{\text{ta}}}(1).$$

Then, $\mathcal{H}_{\text{so},m}(0) \subseteq \mathcal{H}_{\text{ta},m}(\rho_{\text{ta},n})$ *for some sequence* $\rho_{\text{ta},n} = o(1)$*.*

Remark 4. The ν_n-transferability condition in Assumption 2 is stronger than necessary for convergence of the target model for simplicity of presentation. A weaker sufficient condition is that $\min_{f \in \mathcal{F}_n^{\text{sub}}} \left\| f \circ \check{h} - f_n \circ h_m \right\|_{P_{\text{ta}},2} = O_P(\delta_{\text{so},m}/\nu_n)$ for every sequence of estimated embedding functions $\check{h} \in \mathcal{H}_m$ satisfying (12). Indeed, this weaker condition is necessary in the following sense: if $\epsilon_{\text{ta},n} = O(\delta_{\text{so},m}/\nu_n)$, $r_{\text{ta},n} = O(\delta_{\text{so},m}/\nu_n)$, and $\left\| \widehat{f} \circ \check{h} - a^*_{\text{ta}} \right\|_{P_{\text{ta}},2} = O_P\left(\delta_{\text{so},m}/\nu_n\right)$ for every sequence of estimated embedding functions $\check{h} \in \mathcal{H}_m$ satisfying (12), then $\min_{f \in \mathcal{F}_n^{\text{sub}}} \left\| f \circ \check{h} - f_n \circ h_m \right\|_{P_{\text{ta}},2} = O_P(\delta_{\text{so},m}/\nu_n)$. The proof is similar to the one for Theorem 2. ■

[15] Our rate is pointwise with respect to the sequence of estimated embedding functions $\check{h} \in \mathcal{H}_m$ satisfying (12). Extending the result to a uniform convergence rate over the estimated set of embedding functions h_m is an interesting direction for future research, but doing so would require additional assumptions and is beyond the scope of this paper.

3.3 Closer Look at Task Diversity

In this subsection, we discuss sufficient and necessary conditions for task diversity in Definition 2.

3.3.1 Sufficient Condition

We first assume that the density ratio of Z in the source and target tasks is bounded away from zero.

Assumption 3 (Prediction Bound for Source Model)**.** *There exists $\underline{w}_n > 0$ such that $p_{\mathrm{so}}(z)/p_{\mathrm{ta}}(z) \geq \underline{w}_n^2$, P_{ta}-a.s., where $p_{\mathrm{so}}(z)$ and $p_{\mathrm{ta}}(z)$ are the densities of Z in the source and target samples with respect to some measures, respectively.*

Assumption 3 holds only if the support of Z in the target task is included in the support of Z in the source task. We allow $\underline{w}_n$ to decrease slowly to zero as n increases, but not too quickly. Although it is commonly assumed in the machine learning literature (e.g., Johansson, Sontag, and Ranganath, 2019), the density ratio condition is a strong assumption especially when Z is unstructured data. Relaxing this assumption is beyond the scope of this paper and is an important direction for future research.

Theorem 3 (Sufficient Condition for Transferability on the Mean Squared Loss)**.** *Suppose that there exists a Lipschitz function $\Gamma : \mathbb{R}^T \to \mathbb{R}$ with Lipschitz constant L_Γ such that $\|f_n \circ h_m - \Gamma \circ g_m \circ h_m\|_{P_{\mathrm{ta}},2} \leq \varrho_n$ for some tolerance $\varrho_n \geq 0$. For this Γ, we assume that for every $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$, there exists $f \in \mathcal{F}_n^{\mathrm{sub}}$ such that $\|f \circ h - \Gamma \circ g_m^\bullet \circ h\|_{P_{\mathrm{ta}},2} \leq \varrho_n$, where $g_m^\bullet \in \arg\min_{g \in \mathcal{G}_m} \|g \circ h - g_m \circ h_m\|_{P_{\mathrm{so}},2}$. Then, under Assumption 3, g_m is $(\rho_{\mathrm{so},m}, \rho_{\mathrm{ta},n})$-diverse over f_n with respect to h_m for every $\rho_{\mathrm{so},m} \geq 0$ and $\rho_{\mathrm{ta},n} = L_\Gamma \rho_{\mathrm{so},m}/\underline{w}_n + 2\varrho_n$.*

This theorem provides a sufficient condition for $(\rho_{\mathrm{so},m}, \rho_{\mathrm{ta},n})$-diversity based on the existence of a Lipschitz function Γ that relates the target model $f_n \circ h_m$ to the source model $g_m \circ h_m$. The first condition requires that the target model can be approximated by applying the Lipschitz function Γ to the source model with some small error ϱ_n. The second condition requires that for every embedding function h close to the source task identified set, there exists a target model f that can approximate $\Gamma \circ g_m^\bullet \circ h$ with some small error ϱ_n. Under these conditions and the density ratio condition in Assumption 3, the task g_m is $(\rho_{\mathrm{so},m}, L_\Gamma \rho_{\mathrm{so},m}/\underline{w}_n + 2\varrho_n)$-diverse over f_n with respect to h_m for every $\rho_{\mathrm{so},m} \geq 0$. Note that transferability is stronger if the Lipschitz constant L_Γ is small, the approximation error ϱ_n is small, and the constant $\underline{w}_n$ related to the infimum of the density ratio is large.

In the extreme case $L_\Gamma = 0$ (i.e., Γ is constant), we have $(\rho_{\mathrm{so},m}, 2\varrho_n)$-diversity. Also, if L_Γ diverges to infinity, the $(\rho_{\mathrm{so},m}, \rho_{\mathrm{ta},n})$-diversity condition does not hold for every finite $\rho_{\mathrm{ta},n}$. In summary, a small L_Γ leads to better transferability.

In the next example, we verify the sufficient condition in Theorem 3 when both the computer scientist and the economist use linear models.

Example 9 (Linear Model)**.** *Consider the simple linear models for both f_n and g_m, i.e., $f_n \circ h_m(z) = \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} h_m(z)$ and $g_m \circ h_m(z) = \alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m(z)$, where $\alpha_{\mathrm{ta},n} \in \mathbb{R}$, $\beta_{\mathrm{ta},n} \in \mathbb{R}^K$, $\alpha_{\mathrm{so},m} = (\alpha_{\mathrm{so},m,1}, \ldots, \alpha_{\mathrm{so},m,T})' \in \mathbb{R}^T$, and $B_{\mathrm{so},m} = (\beta_{\mathrm{so},m,1}, \ldots, \beta_{\mathrm{so},m,T})' \in \mathbb{R}^{T \times K}$ with $\beta_{\mathrm{so},m,t} \in \mathbb{R}^K$ for $t = 1, \ldots, T$. Let $\mathcal{G}_m$ be the class of linear source heads $g_{\alpha,B}(v) = \alpha + Bv$ with $\alpha \in \mathbb{R}^T$ and $B \in \mathbb{R}^{T \times K}$ and $\mathcal{F}_n^{\mathrm{sub}}$ be the class of linear target heads $f_{\alpha,\beta}(v) = \alpha + \beta' v$ with $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^K$. If $B_{\mathrm{so},m}$ has full column rank, i.e., $\mathrm{rank}(B_{\mathrm{so},m}) = K$, then*

$$\Gamma(x) = \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} B^+_{\mathrm{so},m} (x - \alpha_{\mathrm{so},m})$$

satisfies $f_n \circ h_m = \Gamma \circ g_m \circ h_m$ for any linear f_n, where $B^+_{\mathrm{so},m} := (B'_{\mathrm{so},m} B_{\mathrm{so},m})^{-1} B'_{\mathrm{so},m}$ is the Moore-Penrose inverse of $B_{\mathrm{so},m}$.

On the other hand, if $\mathrm{rank}(B_{\mathrm{so},m}) < K$, g_m is not diverse for all directions. An affine function Γ with an intercept satisfying $f_n \circ h_m = \Gamma \circ g_m \circ h_m$ exists for every value of h_m if and only if $\beta_{\mathrm{ta},n}$ belongs to the row span of $B_{\mathrm{so},m}$. That is, there exists $b \in \mathbb{R}^T$ such that $\beta'_{\mathrm{ta},n} = b' B_{\mathrm{so},m}$.[16] Then, we can choose $\Gamma(x) = \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} B^+_{\mathrm{so},m} (x - \alpha_{\mathrm{so},m})$ by the definition of the Moore-Penrose inverse $B_{\mathrm{so},m} B^+_{\mathrm{so},m} B_{\mathrm{so},m} = B_{\mathrm{so},m}$ and $\beta'_{\mathrm{ta},n} = b' B_{\mathrm{so},m}$. Since the intercept does not affect the Lipschitz constant, $L_\Gamma = \|\beta'_{\mathrm{ta},n} B^+_{\mathrm{so},m}\|_2 \leq \|\beta_{\mathrm{ta},n}\|_2 / \sigma^+_{\min}(B_{\mathrm{so},m})$ in both cases.[17] Thus, by Theorem 3, g_m is $(\rho_{\mathrm{so},m}, L_\Gamma \rho_{\mathrm{so},m} / \underline{w}_n)$-diverse over f_n with respect to h_m if, for every $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$ and every corresponding best linear source head $(\alpha^\bullet_{\mathrm{so},m}, B^\bullet_{\mathrm{so},m}) \in \arg\min_{\alpha \in \mathbb{R}^T, B \in \mathbb{R}^{T \times K}} \|\alpha + Bh - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m)\|_{P_{\mathrm{so}},2}$, the linear model

$$v \mapsto \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} B^+_{\mathrm{so},m} (\alpha^\bullet_{\mathrm{so},m} - \alpha_{\mathrm{so},m} + B^\bullet_{\mathrm{so},m} v)$$

belongs to $\mathcal{F}_n^{\mathrm{sub}}$. In particular, the displayed linear model belongs to $\mathcal{F}_n^{\mathrm{sub}}$ if $\mathcal{F}_n^{\mathrm{sub}}$ contains all linear models.

Example 9 shows that, when the target class contains all linear heads with intercepts,

[16] More generally, by allowing some approximation error $\varrho_n \geq 0$, we can instead assume that there exists $b \in \mathbb{R}^T$ such that $\|(\beta'_{\mathrm{ta},n} - b' B_{\mathrm{so},m}) h_m(Z)\|_{P_{\mathrm{ta}},2} \leq \varrho_n$. Choosing $\Gamma(x) = \alpha_{\mathrm{ta},n} + b'(x - \alpha_{\mathrm{so},m})$ gives Lipschitz constant $L_\Gamma = \|b\|_2$ and approximation error at most ϱ_n. In this case, g_m is $(\rho_{\mathrm{so},m}, L_\Gamma \rho_{\mathrm{so},m} / \underline{w}_n + 2\varrho_n)$-diverse over f_n with respect to h_m, provided that the functional class condition holds. Note that the approximation error ϱ_n induced by Γ is always zero in the full-column-rank case.

[17] The inequality follows from Lemmas 17 and 18.

the row-span condition is sufficient for transferability with $\varrho_n = 0$. Under the regularity conditions in Theorem 4, the same condition is also necessary for $(0, 0)$-diversity. Intuitively, if the coefficients of the source tasks are informative enough to represent the coefficients of the target task, the transferability condition holds. Thus, if we want to guarantee transferability uniformly over all target tasks for every possible β_{ta}, this requires at least K source tasks ($T \geq K$). The next corollary states sufficient conditions for general Lipschitz target models, which include the linear target case as a special case.

Corollary 1 (Sufficient Condition for Transferability for Linear Source Models)**.** *Let $\mathcal{G}_m$ be the class of linear source heads $g_{\alpha,B}(v) = \alpha + Bv$ with $\alpha \in \mathbb{R}^T$ and $B \in \mathbb{R}^{T\times K}$, and suppose that $g_m \circ h_m = \alpha_{\text{so},m} + B_{\text{so},m}h_m$, where $\alpha_{\text{so},m} \in \mathbb{R}^T$ and $B_{\text{so},m} \in \mathbb{R}^{T\times K}$. Consider any Lipschitz target model class $\mathcal{F}_n^{\text{sub}}$ with Lipschitz constant $L_{\mathcal{F}}$. Suppose that, for every $\rho_{\text{so},m} \geq 0$, every $h \in \mathcal{H}_{\text{so},m}(\rho_{\text{so},m})$, and every corresponding best linear source head*

$$(\alpha^{\bullet}_{\text{so},m}, B^{\bullet}_{\text{so},m}) \in \underset{\alpha\in\mathbb{R}^T,\, B\in\mathbb{R}^{T\times K}}{\arg\min} \|\alpha + Bh - (\alpha_{\text{so},m} + B_{\text{so},m}h_m)\|_{P_{\text{so}},2}\,,$$

there exists $f \in \mathcal{F}_n^{\text{sub}}$ such that

$$f \circ h = f_n\left(B^{+}_{\text{so},m}(\alpha^{\bullet}_{\text{so},m} - \alpha_{\text{so},m} + B^{\bullet}_{\text{so},m}h)\right), \qquad P_{\text{ta}}\text{-a.s.}$$

Assume Assumption 3 holds. If $B_{\text{so},m}$ has full column rank, then g_m is $(\rho_{\text{so},m}, \rho_{\text{ta},n})$-diverse over f_n with respect to h_m for every $\rho_{\text{so},m} \geq 0$, where

$$\rho_{\text{ta},n} = \frac{L_{\mathcal{F}}}{\underline{w}_n \sigma_{\min}(B_{\text{so},m})}\rho_{\text{so},m}.$$

The corollary implies that, when the source task is linear and the target task is Lipschitz, the ν_n-transferability condition holds if the source task coefficients are informative enough to represent the target task coefficients. Moreover, the transferability measure $\nu_n = \underline{w}_n\sigma_{\min}(B_{\text{so},m})/L_{\mathcal{F}}$ is larger if the source task has a larger minimum singular value $\sigma_{\min}(B_{\text{so},m})$, the target task has a smaller Lipschitz constant $L_{\mathcal{F}}$, and the density ratio of Z in the source and target tasks is bounded away from zero with a larger value of $\underline{w}_n$.

3.3.2 Necessary Condition

The following theorems provide necessary conditions for task diversity. We start with the linear case as in Example 9. When $B_{\text{so},m}$ is rank deficient, there is a direction $v(Z)$ in the embedding space that is not captured by any of the source tasks, but it can be relevant for

the target task. Intuitively, if the source tasks are not diverse enough to cover all directions of the target task, transferability generally fails for the uncovered directions of the source task coefficients. See the next theorem for a formal statement.

Theorem 4 (Necessary Condition for Linear Models)**.** *Suppose that the population source and target heads are linear as in Example 9, so that* $g_m \circ h_m = \alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m$ *and* $f_n \circ h_m = \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} h_m$*. Assume that* $\mathcal{G}_m$ *contains every linear source head* $v \mapsto \alpha + Bv$*, where* $\alpha \in \mathbb{R}^T$ *and* $B \in \mathbb{R}^{T \times K}$*, and that every* $f \in \mathcal{F}_n^{\mathrm{sub}}$ *is a linear target head* $v \mapsto \alpha + \beta' v$*, where* $\alpha \in \mathbb{R}$ *and* $\beta \in \mathbb{R}^K$*. Assume also that* $\mathbb{E}_{\mathrm{ta}}[(1, h_m(Z)')'(1, h_m(Z)')]$ *is positive definite and that, for every unit vector* $v \in \mathbb{R}^K$*, there exists some* $c \in \mathbb{R}^K$ *such that* $(I_K - vv')h_m + c \in \mathcal{H}_m$*. If* g_m *is* $(0,0)$*-diverse over* f_n *with respect to* h_m*, then there exists a vector* $b \in \mathbb{R}^T$ *such that* $\beta'_{\mathrm{ta},n} = b' B_{\mathrm{so},m}$*, and there exists an affine function* $\Gamma(x) = \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} B^{+}_{\mathrm{so},m}(x - \alpha_{\mathrm{so},m})$ *such that* $f_n \circ h_m(Z) = \Gamma \circ g_m \circ h_m(Z)$*,* P_{ta}*-a.s.*

The assumption $(I_K - vv')h_m + c \in \mathcal{H}_m$ is a technical condition that allows us to construct a function h that attains the same source prediction as h_m but a different target prediction from h_m. For some functional classes $\mathcal{H}_m$ such as the class of ReLU activated hidden layer functions, $\mathcal{H}_m$ only contains nonnegative functions, but it satisfies this assumption by choosing c to be sufficiently large if $\mathcal{H}_m$ is sufficiently rich.

The following theorem provides necessary conditions for transferability in a general model with a nonlinear source task under sufficiently rich function classes $\mathcal{H}_m$ and $\mathcal{G}_m$ so that the nonlinear analog of the construction h in the proof of Theorem 4 is possible. Note that we do not put any restriction on the function class $\mathcal{F}_n^{\mathrm{sub}}$.

Theorem 5 (Necessary Condition for Transferability in General Models)**.** *Assume that* $\mathcal{F}_n$ *and* $\mathcal{G}_{m,t}$ *for* $t = 1, \ldots, T$ *are classes of square-integrable functions with respect to* P_{ta} *and* P_{so}*. We also assume that there exists* $h_0 \in \mathcal{H}_{\mathrm{so},m}(0)$ *such that* $h_0 \neq h_m$*. Then, if* g_m *is* $(0,0)$*-diverse over* f_n *with respect to* h_m*, there exists a measurable function* $\Gamma_0 : \mathbb{R}^K \to \mathbb{R}$ *such that* $f_n \circ h_m(Z) = \Gamma_0 \circ h_0(Z)$*,* P_{ta}*-a.s.*

In particular, suppose that

$$\sigma(h_0(Z)) \subseteq \overline{\sigma(g_m \circ h_m(Z))}^{P_{\mathrm{ta}}},$$

where the right-hand side is the P_{ta}*-completion defined in Lemma 3. Then there exists a measurable function* $\Gamma : \mathbb{R}^T \to \mathbb{R}$ *such that* $f_n \circ h_m(Z) = \Gamma \circ g_m \circ h_m(Z)$*,* P_{ta}*-a.s.*

The existence of a distinct h_0 in Theorem 5 means that the population source model does not uniquely identify the embedding. Such source nonidentification commonly arises in deep

neural networks, where different hidden representations can be paired with compensating source heads. As $(0,0)$-diversity requires $\mathcal{H}_{\mathrm{so},m}(0) \subseteq \mathcal{H}_{\mathrm{ta},m}(0)$, the existence of Γ_0 is needed to adapt the nonidentified embedding function. The sigma-field inclusion condition is an additional information condition for the target distribution. For the same h_0, it states that, up to P_{ta}-null sets, the population source output $g_m \circ h_m$ contains all the information on $h_0(Z)$. Sufficiently rich classes $\mathcal{G}_m$ and $\mathcal{H}_m$ are needed for the construction of a source-equivalent h_0 satisfying this condition. Under this additional condition, the existence of Γ is necessary for $(0,0)$-diversity.

Under the head-class and regularity conditions in Theorems 4 and 5, the necessary conditions show that the corresponding sufficient condition is also necessary in these settings. These theorems suggest that the sufficient condition in Theorem 3 with $\varrho_n = 0$, namely $f_n \circ h_m(Z) = \Gamma \circ g_m \circ h_m(Z)$, P_{ta}-a.s., is also necessary for the transferability condition in these cases. However, the next example shows that the transferability condition may not hold even if both f_n and g_m are linear but the pre-training model class $\mathcal{G}_m$ is nonlinear.

Example 10 (Counterexample for Nonlinear Pre-training Model)**.** *Suppose that $h_m(Z) \in \mathbb{R}_+$ and both tasks use the mean squared loss. Suppose that $\mathcal{F}_n^{\mathrm{sub}}$ is a set of linear models as in Example 9, but $\mathcal{G}_m$ includes any shallow neural networks with one hidden layer with ReLU activation of width 2. Suppose that $Z \sim$ Unif $[-1,1]$, $h_m(Z) = |Z|$, $f_n(\cdot) = g_m(\cdot) = (\cdot)$ are identity maps and $T = 1$. If we choose $h(Z) = Z + 1$, then we have*

$$\min_{g \in \mathcal{G}_m} \|g \circ h(Z) - g_m \circ h_m(Z)\|_{P_{\mathrm{so}},2}^2 = 0,$$

since $g \in \mathcal{G}_m$ can approximate the function $g_m \circ h_m(Z) = |Z|$, for example, by $g(x) = \mathrm{ReLU}(x-1) + \mathrm{ReLU}(-(x-1))$. On the other hand,

$$\min_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - f_n \circ h_m\|_{P,2}^2 = \mathbb{E}[(1/2 - |Z|)^2] = 1/12 > 0,$$

since $f \in \mathcal{F}_n^{\mathrm{sub}}$ is linear and cannot approximate $f_n \circ h_m(Z) = |Z|$ perfectly. Thus, g_m is not $(0,0)$-diverse over f_n with respect to h_m.

From this section, we have seen that the transferability condition is substantially stronger for nonlinear source heads $\mathcal{G}_m$ than for linear source heads $\mathcal{G}_m$. This is because the nonlinear source head $\mathcal{G}_m$ can approximate a much richer class of functions than the linear source head $\mathcal{G}_m$, which makes it more difficult to satisfy the diversity condition. Therefore, if the economist is indifferent between linear and nonlinear source heads, the linear source head is preferable for guaranteeing the transferability condition.

4 Convergence Rate of Transfer Learning under Lower-Level Conditions

4.1 Low-dimensional Case

In this section, we provide lower-level primitive conditions for the convergence rates for the transfer-learning problem without the high-level conditions of Assumption 1. We first introduce some notation. For a function class $\mathcal{F}$, let $\mathsf{V}(\mathcal{F})$ be the VC dimension of the set of subgraphs of functions in $\mathcal{F}$.[18] To avoid technical complications concerning the measurability of the supremum, we assume that the functional classes $\mathcal{F}_n$, $\mathcal{F}_n^{\text{sub}}$, and $\mathcal{G}_m \circ \mathcal{H}_m$ are pointwise measurable.[19]

We start with the convergence rate of the target model given the estimated embedding function $\check{h}$ from the source task. Before we state the theorem on the convergence rate of the target model $\widehat{f} \circ \check{h}$, we assume one more condition on the target loss function.

Assumption 4 (Conditions on Target Task)**.**

1. *For every sequence* $\{\rho_{\text{so},m}\}_{m=1}^{\infty}$ *such that* $\rho_{\text{so},m} = O(\delta_{\text{so},m})$*, every* $f \in \mathcal{F}_n$*, and every* $h \in \mathcal{H}_{\text{so},m}(\rho_{\text{so},m})$*, we have*

$$\|f \circ h - a_{\text{ta}}^*\|_{P_{\text{ta}},2}^2 \lesssim \mathbb{E}_{\text{ta}}[\ell_{\text{ta}}(f \circ h(Z), Y)] - \mathbb{E}_{\text{ta}}[\ell_{\text{ta}}(a_{\text{ta}}^*(Z), Y)] \lesssim \|f \circ h - a_{\text{ta}}^*\|_{P_{\text{ta}},2}^2$$

   for large enough m *and* n*.*

2. *The loss function* $\ell_{\text{ta}}(\cdot, \cdot)$ *is Lipschitz continuous in the first argument with some constant* $C_{\ell,\text{ta}} > 0$*.*

This assumption requires that the excess risk of the target loss function is bounded above and below by the squared L^2-norm. This assumption is standard in statistical learning theory and is imposed in many existing works (e.g., Farrell et al., 2021). Lemma 4 verifies the excess-risk comparison and the required Lipschitz comparison over uniformly bounded candidate-score ranges for least squares and binary logistic loss.

[18] The *subgraph* of a function $f : \mathcal{X} \to \mathbb{R}$ is the subset of $\mathcal{X} \times \mathbb{R}$ given by $\{(x, y) : y < f(x)\}$. The VC dimension of a set of subgraphs is defined as the largest integer such that there exists a set of points with cardinality equal to the integer that can be shattered by the subgraphs. See van der Vaart and Wellner (2023), Section 2.6 for more details.

[19] A measurable functional class $\mathcal{F}$ is *pointwise measurable* if it contains a countable subset $\mathcal{F}^0$ such that for every $f \in \mathcal{F}$ there exists a sequence $f_j \in \mathcal{F}^0$ with $f_j(v) \to f(v)$ for every v, where the convergence is with respect to the Euclidean norm. See van der Vaart and Wellner (2023), Section 2.3 for more details.

Theorem 6 (Convergence Rate of Target Model)**.** *Let*

$$r_{\mathrm{ta},n} = \sqrt{\frac{\mathsf{V}(\mathcal{F}_n)\log(n)}{n}}. \tag{15}$$

For every estimated embedding function $\breve{h}$ *satisfying Assumption 1 (i), let* $\widehat{f} \in \mathcal{F}_n$ *be an estimated target model satisfying*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{ta}}(\widehat{f} \circ \breve{h}(Z_i), Y_i) \leq \min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{ta}}(f \circ \breve{h}(Z_i), Y_i) \\ + O_P(r_{\mathrm{ta},n}^2 + (\delta_{\mathrm{so},m}/\nu_n)^2 + \epsilon_{\mathrm{ta},n}^2).\end{aligned} \tag{16}$$

Suppose that $\|f\|_\infty \leq M_F < \infty$ *for every* $f \in \mathcal{F}_n$*. Then, under Assumption 1 (i), Assumptions 2 to 4, we have*

$$\left\| \widehat{f} \circ \breve{h} - a_{\mathrm{ta}}^* \right\|_{P_{\mathrm{ta}},2} = O_P\left(r_{\mathrm{ta},n} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}\right). \tag{17}$$

The theorem shows that, under standard conditions on the function class $\mathcal{F}_n$ and the target loss function, the convergence rate of the target model $\widehat{f} \circ \breve{h}$ to the true model a_{ta}^* is governed by the VC dimension of $\mathcal{F}_n$ and the target sample size n.

Remark 5 (Fine-tuning)**.** Our theory accommodates fine-tuning by allowing $\breve{h}$ to differ from the source task estimator. This remains valid provided $\breve{h}$ is estimated on a sample independent of the final parameter estimation sample and $\widehat{f} \circ \breve{h}$ converges to $f_n \circ h_m$ faster than $n^{-1/4}$. Thus, the economist can fine-tune the embedding function using the same sample to train $\widehat{f}$, provided fine-tuning improves the target model convergence rate relative to the unfine-tuned embedding function, whose convergence rate is given by Theorem 6. Since fine-tuning does not necessarily improve the accuracy of the target model (Kumar, Raghunathan, Jones, Ma, and Liang, 2022, Table 1), it is important to check the convergence rate of the fine-tuned embedding function to ensure that the fine-tuning is not harmful to the convergence rate of the target model. For example, the economist can check by cross-validation whether the loss of $\widehat{f} \circ \breve{h}$ with fine-tuning is smaller than the loss without fine-tuning. The formal hypothesis testing procedure for this comparison is proposed by Fava (2025). ■

Next, we provide a theorem on the convergence rate of the source model $\breve{g} \circ \breve{h}$ to the population source model $g_m \circ h_m$. The rate of convergence for deep learning models is an active area of research and can be improved in future work. Importantly, our transfer learning rate of the target task uses the convergence rate of the source model as an input; thus, any

improvement in the convergence rate of the source model can be directly translated into an improved convergence rate of the target model by our theory.

Assumption 5 (Conditions on Source Task)**.**

1. *For every $g \in \mathcal{G}_m$ and $h \in \mathcal{H}_m$,*

$$\tau_m^2 \left\| g \circ h - a_{\mathrm{so}}^* \right\|_{P_{\mathrm{so}},2}^2 \lesssim \mathbb{E}_{\mathrm{so}}[\ell_{\mathrm{so}}(g \circ h(Z), S)] - \mathbb{E}_{\mathrm{so}}[\ell_{\mathrm{so}}(a_{\mathrm{so}}^*(Z), S)] \lesssim \left\| g \circ h - a_{\mathrm{so}}^* \right\|_{P_{\mathrm{so}},2}^2 \tag{18}$$

   with some sequence $\tau_m \in (0, 1]$ for large enough m and n.

2. *The loss function $\ell_{\mathrm{so}}(\cdot, \cdot)$ is Lipschitz continuous in the first argument with some constant $C_{\ell,\mathrm{so}} > 0$.*

3. *There exists a sequence $L_{\mathcal{G},m} \geq 1$ such that every $g \in \mathcal{G}_m$ is $L_{\mathcal{G},m}$-Lipschitz with respect to the Euclidean norm, uniformly over $\mathcal{G}_m$. That is, for every $v, \widetilde{v} \in \mathcal{V}$,*

$$\|g(v) - g(\widetilde{v})\|_2 \leq L_{\mathcal{G},m} \|v - \widetilde{v}\|_2.$$

The loss conditions in parts (i)-(ii) are satisfied with $\tau_m = T^{-1}$ if the loss function is the multinomial logistic loss and the Euclidean-norm bound M in Lemma 5 holds uniformly in m and T over all $g \in \mathcal{G}_m$ and $h \in \mathcal{H}_m$ and for a_{so}^*. Part (iii) is a regularity condition on the source-head class rather than the loss function. It rules out discontinuous heads and allows the entropy of the composite class $\mathcal{G}_m \circ \mathcal{H}_m$ to be controlled by the coordinatewise entropies of $\mathcal{G}_m$ and $\mathcal{H}_m$.

The next theorem relates the convergence rate of the source model in Assumption 1 (i) to the VC dimensions of the function classes $\mathcal{H}_m$ and $\mathcal{G}_m$.

Theorem 7 (Convergence Rate of Source Model)**.** *Let*

$$r_{\mathrm{so},m} = \sqrt{\frac{\sum_{k=1}^{K} \mathsf{V}(\mathcal{H}_{m,k}) \log(m L_{\mathcal{G},m} K) + \sum_{t=1}^{T} \mathsf{V}(\mathcal{G}_{m,t}) \log(mT)}{\tau_m^2 m}}. \tag{19}$$

Let $\breve{h} \in \mathcal{H}_m$ and $\breve{g} \in \mathcal{G}_m$ be any estimated embedding function and source model estimator satisfying

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \ell_{\mathrm{so}}(\breve{g} \circ \breve{h}(Z_i), S_i) \leq & \min_{g \in \mathcal{G}_m, h \in \mathcal{H}_m} \frac{1}{m} \sum_{i=1}^{m} \ell_{\mathrm{so}}(g \circ h(Z_i), S_i) \\ & + O_{P_{\mathrm{so}}}(r_{\mathrm{so},m}^2 + \epsilon_{\mathrm{so},m}^2), \end{aligned} \tag{20}$$

Suppose that $\sup_{z\in\mathcal{Z}} \|h(z)\|_2 \leq M_H < \infty$ *for every* $h \in \mathcal{H}_m$*, and* $\sup_{v\in\mathcal{V}} \|g(v)\|_2 \leq M_G < \infty$ *for every* $g \in \mathcal{G}_m$*. Then, under Assumption 5, we have*

$$\left\|\breve{g}\circ\breve{h} - g_m \circ h_m\right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}\left(r_{\mathrm{so},m}/\tau_m + \epsilon_{\mathrm{so},m}/\tau_m\right), \tag{21}$$

$$\left\|\breve{g}\circ\breve{h} - a^*_{\mathrm{so}}\right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}\left(r_{\mathrm{so},m}/\tau_m + \epsilon_{\mathrm{so},m}/\tau_m\right). \tag{22}$$

This theorem states that under some standard conditions on the function classes $\mathcal{H}_m$ and $\mathcal{G}_m$ and the source loss function, the convergence rate of the source model $\breve{g}\circ\breve{h}$ to the population model $g_m \circ h_m$ (or the true model a^*_{so}) is determined by the coordinatewise VC dimensions of $\mathcal{H}_m$ and $\mathcal{G}_m$, the logarithmic factors induced by the embedding dimension K, the output dimension T, and the head Lipschitz constant $L_{\mathcal{G},m}$, the source sample size m, and the curvature term τ_m of the source loss function.

Theorems 6 and 7 together imply that, under the transferability assumption, the convergence rate of the target model $\widehat{f}\circ\breve{h}$ to the true model a^*_{ta} is governed mainly by five components: the ratio of the VC dimension of the target model class $\mathcal{F}_n$ to the target sample size n; the ratio of the source complexities for $\mathcal{H}_m$ and $\mathcal{G}_m$ to the source sample size m; the transferability rate ν_n; the approximation error terms $\epsilon_{\mathrm{ta},n}$ and $\epsilon_{\mathrm{so},m}$; and the curvature term τ_m of the source loss function. For example, $\mathsf{V}(\mathcal{F}_n) = O(K)$ if $\mathcal{F}_n$ is a class of linear models with K regressors, and $\mathsf{V}(\mathcal{F}_n) = O(W_n L_n \log(W_n))$ if $\mathcal{F}_n$ is a class of DNNs with width W_n and depth L_n (Bartlett, Harvey, Liaw, and Mehrabian, 2019, Theorem 7). In particular, the rate improves when both function class complexities are small relative to sample size and the transferability and approximation error terms are negligible. If the complexity of the function classes is smaller than the square root of the sample sizes, and the transferability and approximation error terms are sufficiently small, the target model can achieve a convergence rate faster than $n^{-1/4}$.

The derived rate is faster than the one in Tripuraneni et al. (2020) since we effectively use the curvature condition and the boundedness of the loss function. Also, our rate is comparable to the result in Watkins et al. (2023) although their fast rate assumes $\mathbb{E}_{\mathrm{so}}[\ell_{\mathrm{so}}(a^*_{\mathrm{so}}(Z), S)] = 0$, which is a strong and unrealistic assumption in practice. Without this assumption, their rate becomes slower than ours and comparable to the one in Tripuraneni et al. (2020). Importantly, the slow rate in Tripuraneni et al. (2020) is not enough to apply the DML framework since it is always slower than $n^{-1/4}$.

We can directly apply the above theorems to derive the convergence rates of transfer learning when the target sample size n is large enough compared to its input dimension K. This strategy is taken in, for example, Bajari et al. (2025) and Bach et al. (2024). Since the VC dimension is well studied for many specific models such as linear models and DNNs

(Bartlett et al., 2019), the above theorems can be applied to derive the convergence rates of transfer learning for these specific models under some conditions on the transferability and approximation errors. For DNNs, the approximation error can be small if the true model is sufficiently smooth and the DNN has enough width and depth (Yarotsky, 2017).

4.2 High-dimensional Case

4.2.1 Convergence of h for linear source heads g

For the low-dimensional case, the VC dimension of the function class $\mathcal{F}_n$ is small, and Theorem 6 is directly applicable to derive the convergence rate of the target model $\widehat{f} \circ \check{h}$ to the true model a^*_{ta}. An important feature of the VC dimension is that it is a pure functional complexity measure and does not depend on the distribution of the input data. On the other hand, Theorem 6 is not directly applicable to the high-dimensional case since the VC dimension of $\mathcal{F}_n$ is large, and we need to assume some additional structure on the input data, i.e., embeddings, to derive the convergence rate of the target model.

In this section, we focus on the case in which the source model has a linear head as in Example 9 and $h_m(Z) \in \mathbb{R}^K$ is the last hidden layer of the source model. In this case, we have a convergence rate of the estimated embedding function $\check{h}$ to an approximately true embedding function h_m up to a linear transformation. Recall that the population source model is $g_m \circ h_m(Z) = \alpha_{\text{so},m} + B_{\text{so},m} h_m(Z)$ for some vector $\alpha_{\text{so},m} \in \mathbb{R}^T$, some matrix $B_{\text{so},m} \in \mathbb{R}^{T \times K}$, and an embedding function $h_m(Z) \in \mathbb{R}^K$. The pre-trained source model is $\check{g} \circ \check{h}(Z) = \check{\alpha}_{\text{so}} + \check{B}_{\text{so}} \check{h}(Z)$ for some vector $\check{\alpha}_{\text{so}} \in \mathbb{R}^T$, some matrix $\check{B}_{\text{so}} \in \mathbb{R}^{T \times K}$, and an embedding function $\check{h}(Z) \in \mathbb{R}^K$ estimated from the source task. Define a transformed embedding function $\check{h}^{\text{proj}}$ as the projection of $\check{h}$ onto the row space of $\check{B}_{\text{so}}$ as follows:

$$\check{h}^{\text{proj}} := \check{B}^+_{\text{so}} \check{B}_{\text{so}} \check{h} \tag{23}$$

The following lemma gives an L_2 convergence rate for the projected embedding $\check{h}^{\text{proj}}$ toward an affine transformation $\check{q} + \check{Q} h_m$ of h_m, where $\check{q} := \check{B}^+_{\text{so}}(\alpha_{\text{so},m} - \check{\alpha}_{\text{so}})$ is the shift term and $\check{Q} := \check{B}^+_{\text{so}} B_{\text{so},m}$ is the linear transformation of h_m.

Lemma 1 (Convergence of the projected embedding)**.** *Suppose that the population and pre-trained source heads are linear, as defined in Example 9, and that Assumption 1 (i) holds. If, for some deterministic sequence $\underline{\sigma}_m > 0$,*

$$\|\check{B}^+_{\text{so}}\|_{\text{sp}} = O_{P_{\text{so}}}(\underline{\sigma}_m^{-1}), \tag{24}$$

then

$$\left\| \breve{h}^{\text{proj}} - (\breve{q} + \breve{Q} h_m) \right\|_{P_{\text{so}},2} = O_{P_{\text{so}}}(\delta_{\text{so},m}/\underline{\sigma}_m). \tag{25}$$

In particular, the condition (24) holds if [20]

$$P_{\text{so}}\left(\text{rank}(\breve{B}_{\text{so}}) \geq 1,\ \sigma^{+}_{\min}(\breve{B}_{\text{so}}) \geq \underline{\sigma}_m\right) \to 1.$$

If $\breve{B}_{\text{so}}$ has full column rank, $\breve{B}^{+}_{\text{so}} = (\breve{B}'_{\text{so}}\breve{B}_{\text{so}})^{-1}\breve{B}'_{\text{so}}$ is the left inverse of $\breve{B}_{\text{so}}$ and $\breve{B}^{+}_{\text{so}}\breve{B}_{\text{so}} = I_K$ is the identity matrix; thus, $\breve{h}^{\text{proj}} = \breve{h}$. On the other hand, we have $\breve{B}^{+}_{\text{so}}\breve{B}_{\text{so}} \neq I_K$ in general if $\breve{B}_{\text{so}}$ does not have full column rank.

We can interpret $\breve{h}^{\text{proj}}$ as the projection of $\breve{h}$ onto the row space of $\breve{B}_{\text{so}}$ since $\breve{B}^{+}_{\text{so}}\breve{B}_{\text{so}} = P_{\breve{B}'_{\text{so}}} := \breve{B}'_{\text{so}}(\breve{B}_{\text{so}}\breve{B}'_{\text{so}})^{+}\breve{B}_{\text{so}}$ by Harville (2008), Corollary 20.3.8. Note that $\breve{h}^{\text{proj}}$ and $\breve{h}$ return the same output for the source task since $\breve{B}_{\text{so}}\breve{h}^{\text{proj}}(z) = \breve{B}_{\text{so}}\breve{h}(z)$. The projection preserves the source output while removing components of the estimated embedding in the null space of $\breve{B}_{\text{so}}$.[21]

Lemma 1 implies that we can estimate h_m only up to a shifted linear transformation represented by $\breve{h}^{\text{proj}}$. The matrix $\breve{Q}$ describes the linear map from an approximately true embedding function h_m to the component of $\breve{h}$ that is identified from the source task. In the special case in which $\breve{\alpha}_{\text{so}} = \alpha_{\text{so},m}$, $\breve{B}_{\text{so}} = B_{\text{so},m}A^{-1}$ and $\breve{h} = Ah_m$ for an invertible matrix $A \in \mathbb{R}^{K\times K}$, and $B_{\text{so},m}$ has full column rank, we have $\breve{B}_{\text{so}}\breve{h} = B_{\text{so},m}A^{-1}Ah_m = B_{\text{so},m}h_m$ and the exact equality $\breve{h}^{\text{proj}} = \breve{Q}h_m$, where $\breve{Q} = \breve{B}^{+}_{\text{so}}B_{\text{so},m} = A$. Since this holds for every invertible matrix, the embedding function h_m is identified from the source task only up to an invertible linear transformation A such that $Ah_m \in \mathcal{H}_m$. This special case is consistent with the discussion in Schulte et al. (2025) that the embedding function is identified only up to an invertible linear transformation under the restrictive assumption that there is no estimation error for the embedding function. More generally, the matrix $\breve{Q}$ captures the identified set of the embedding function h_m from the source task output. This extends the invertible reparameterization partial identification to settings in which $B_{\text{so},m}$ and $\breve{B}_{\text{so}}$ may be rank deficient.

[20] The event $\{\text{rank}(\breve{B}_{\text{so}}) \geq 1\}$ is only used to ensure that the smallest singular value $\sigma_{\min}(\breve{B}_{\text{so}})$ is well defined.

[21] An indeterminacy can remain. For every constant vector $c \in \mathbb{R}^K$, the pairs $(\breve{\alpha}_{\text{so}}, \breve{h}^{\text{proj}})$ and $(\breve{\alpha}_{\text{so}} + \breve{B}_{\text{so}}c, \breve{h}^{\text{proj}} - c)$ produce the same source output algebraically. However, this indeterminacy does not affect the convergence of the target model as long as $\|c\|_2$ is bounded and the target model is nonparametric or includes a slope coefficient.

4.2.2 LASSO Target Model

We do not recommend using LASSO for the target model for the following reason. As we have seen in Section 4.2.1, $\mathcal{H}_{\text{so},m}(0)$ includes invertible linear transformations of an approximately true embedding function h_m. Theorem 2 suggests that, to ensure $\|\widehat{f}_{n,h_m^\dagger} \circ h_m^\dagger - a_{\text{ta}}^*\|_{P_{\text{ta}},2} = o_{P_{\text{ta}}}(1)$ for any sequence $h_m^\dagger \in \mathcal{H}_{\text{so},m}(0)$, it is necessary that $\mathcal{H}_{\text{so},m}(0) \subseteq \mathcal{H}_{\text{ta},m}(\rho_{\text{ta},n})$ for some sequence $\rho_{\text{ta},n} = o(1)$ when the target model has small estimation and approximation errors. That is, for every invertible matrix $Q \in \mathbb{R}^{K\times K}$ and every vector $q \in \mathbb{R}^K$ such that $q + Qh_m \in \mathcal{H}_m$, there must exist some $f \in \mathcal{F}_n^{\text{sub}}$ such that $\|f_n \circ h_m - f \circ (q + Qh_m)\|_{P_{\text{ta}},2}$ converges to zero as $n \to \infty$. Note that linear regression, ridge regression, and DNNs satisfy this condition since they transform the input vector by a linear transformation.

Therefore, our theory recommends avoiding methods that rely on assumptions, such as sparsity, that are not preserved under the unidentified invertible linear transformations of the embedding function h_m (e.g., Wainwright, 2019, Chapter 7). Suppose the economist uses LASSO for the target model given the estimated embedding function $\breve{h}$ from the source task. To use LASSO, the existing theory typically requires an approximate-sparsity assumption. On the other hand, since $\mathcal{H}_{\text{so},m}(0)$ includes invertible linear transformations of an approximately true embedding function h_m, Theorem 1 in Kolesár, Müller, and Roelsgaard (2026) implies that the approximate sparsity assumption is violated in general. Thus, even if the economist can justify the approximate sparsity assumption for an embedding function h_m, the approximate sparsity assumption is hard to justify for other $h \in \mathcal{H}_{\text{so},m}(0)$. Since we cannot identify the embedding function h_m from the source task, any method relying on variable selection can be fragile.

4.2.3 Ridge Target Model

Section 4.2.2 shows that approximate sparsity is not preserved under the linear reparameterizations left unidentified by the source task. We therefore study an ℓ_2-penalized linear target model as a tractable alternative.

The scale of a linear source head is not identified separately from the scale of its embedding. Thus, the condition in (24) is not invariant to observationally equivalent rescalings. Multiplying the embedding by a large constant and dividing the source-head matrix by the same constant can make its smallest positive singular value arbitrarily small. This scale indeterminacy is not a problem for the low-dimensional case since the VC dimension is invariant to scaling. However, the scale indeterminacy is typically a problem for the high-dimensional case since we use the structure on the embeddings and the source head to derive the convergence rate of the

target model. The ridge theory below therefore imposes fixed singular value bounds on the source-head matrices used in the analysis.

Fix constants $C_{\alpha,\mathrm{so}} < \infty$ and $0 < \underline{\sigma} \leq \overline{\sigma} < \infty$, and let $\Theta_{\mathrm{so},m}$ be a deterministic, nonempty, compact subset of

$$\left\{(\alpha, B) \in \mathbb{R}^T \times \mathbb{R}^{T\times K} : \|\alpha\|_2 \leq C_{\alpha,\mathrm{so}}, \mathrm{rank}(B) \geq 1, \|B\|_{\mathrm{sp}} \leq \overline{\sigma}, \sigma^+_{\min}(B) \geq \underline{\sigma}\right\}.$$

We assume that the population source-head parameters $(\alpha_{\mathrm{so},m}, B_{\mathrm{so},m})$ and the pre-trained source-head parameters $(\breve{\alpha}_{\mathrm{so}}, \breve{B}_{\mathrm{so}})$ belong to $\Theta_{\mathrm{so},m}$ with probability approaching one. The economist can always obtain them through a transformation in Appendix D if they want to ensure this condition explicitly.

For each $(\alpha, B) \in \Theta_{\mathrm{so},m}$, define

$$g_{\alpha,B} : v \mapsto \alpha + Bv. \tag{26}$$

For this subsection, let $\mathcal{G}_m := \{g_{\alpha,B} : (\alpha, B) \in \mathbb{R}^T \times \mathbb{R}^{T\times K}\}$ be the class of linear source heads, so that $\{g_{\alpha,B} : (\alpha, B) \in \Theta_{\mathrm{so},m}\} \subseteq \mathcal{G}_m$. Thus, $\mathcal{G}_m$ may also contain linear heads with parameters outside $\Theta_{\mathrm{so},m}$. For every $(\alpha, B) \in \Theta_{\mathrm{so},m}$, Lemma 18 gives $\|B^+\|_{\mathrm{sp}} \leq \underline{\sigma}^{-1}$. Moreover, on the event that the chosen population and pre-trained source-head parameters belong to $\Theta_{\mathrm{so},m}$, we have $\|\breve{Q}\|_{\mathrm{sp}} = \|\breve{B}^+_{\mathrm{so}} B_{\mathrm{so},m}\|_{\mathrm{sp}} \leq \overline{\sigma}/\underline{\sigma}$.

Define the projected-embedding class used in the ridge analysis by

$$\mathcal{H}^{\mathrm{proj}}_m := \left\{B^+Bh : h \in \mathcal{H}_m, (\alpha, B) \in \Theta_{\mathrm{so},m} \text{ for some } \alpha \in \mathbb{R}^T\right\}. \tag{27}$$

For every $\rho \geq 0$, let $\mathcal{H}^{\mathrm{proj}}_{\mathrm{so},m}(\rho) := \{h \in \mathcal{H}^{\mathrm{proj}}_m : \min_{g\in\mathcal{G}_m} \|g \circ h - g_m \circ h_m\|_{P_{\mathrm{so}},2} \leq \rho\}$. Fix constants $C_\alpha, C_\beta < \infty$, and let

$$\Theta_n := \left\{(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^K : |\alpha| \leq C_\alpha, \|\beta\|_2 \leq C_\beta\right\}.$$

Let $\mathcal{F}_n = \mathcal{F}^{\mathrm{sub}}_n = \{f(v) = \alpha + \beta' v : (\alpha, \beta) \in \Theta_n\}$ and let $f_n(v) = \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} v \in \mathcal{F}^{\mathrm{sub}}_n(h_m)$ be the population target model. For a deterministic sequence $\lambda_n > 0$, define the ridge estimator on the projected embedding by

$$\widehat{f}^{\mathrm{ridge}}_n(v) := \widehat{\alpha}^{\mathrm{ridge}}_n + \widehat{\beta}^{\mathrm{ridge}'}_n v \in \underset{\alpha+\beta' v\in\mathcal{F}_n}{\arg\min} \left\{\frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{ta}}\left(\alpha + \beta'\breve{h}^{\mathrm{proj}}(Z_i), Y_i\right) + \lambda_n^2\|\beta\|_2^2\right\}. \tag{28}$$

We assume the following conditions on the target model and the source parameters.

Assumption 6 (Conditions for Ridge Regression)**.**

(i) (Source Parameter and Embedding Restrictions) The population source parameters satisfy $(\alpha_{\mathrm{so},m}, B_{\mathrm{so},m}) \in \Theta_{\mathrm{so},m}$ *and* $h_m \in \mathcal{H}_m$*. The pre-trained source parameters satisfy* $P_{\mathrm{so}}((\breve{\alpha}_{\mathrm{so}}, \breve{B}_{\mathrm{so}}) \in \Theta_{\mathrm{so},m}, \breve{h} \in \mathcal{H}_m) \to 1$*. Consequently,* $\breve{h}^{\mathrm{proj}} \in \mathcal{H}_m^{\mathrm{proj}}$ *with probability approaching one. For every* m*, there exists a deterministic constant* $C_m < \infty$*, possibly diverging with* n*, such that* $\max\{\mathbb{E}_{\mathrm{so}}[\|h_m(Z)\|_2^2], \sup_{h \in \mathcal{H}_m^{\mathrm{proj}}} \mathbb{E}_{\mathrm{so}}[\|h(Z)\|_2^2]\} \le C_m$*.*

(ii) (Target-Slope Representation for Transferability) There exist a sequence $b_n \in \mathbb{R}^T$ *and a constant* $C_b < \infty$ *such that* $\sup_n \|b_n\|_2 \le C_b$ *and* $\beta'_{\mathrm{ta},n} = b'_n B_{\mathrm{so},m}$*, where* $B_{\mathrm{so},m}$ *denotes the population source head.*

(iii) (Target Parameter Restrictions) The target parameter bounds satisfy $\sup_n |\alpha_{\mathrm{ta},n}| + 2C_b C_{\alpha,\mathrm{so}} \le C_\alpha$ *and* $C_b \overline{\sigma} \le C_\beta$*.*

For any embedding h, let

$$\Sigma_{h,n} := \mathbb{E}_{\mathrm{ta}}\left[(h(Z) - \mathbb{E}_{\mathrm{ta}}[h(Z)])(h(Z) - \mathbb{E}_{\mathrm{ta}}[h(Z)])'\right].$$

In particular, $\Sigma_{h_m,n}$ denotes the covariance matrix of the population embedding. Define the effective degrees of freedom of the ridge regression based on the embedding h by

$$\mathcal{N}_{h,n}(\lambda) := \mathrm{tr}\left(\Sigma_{h,n}(\Sigma_{h,n} + \lambda^2 I_K)^{-1}\right) = \sum_{j=1}^{K} \frac{\mu_j(\Sigma_{h,n})}{\mu_j(\Sigma_{h,n}) + \lambda^2}, \tag{29}$$

where $\mu_1(\Sigma_{h,n}) \ge \cdots \ge \mu_K(\Sigma_{h,n}) \ge 0$ are the eigenvalues of $\Sigma_{h,n}$.

Theorem 8 (Convergence Rate of Ridge Target Model)**.** *Suppose that Assumption 1 (i) and Assumptions 3 and 6 hold. We also assume Assumption 4 holds with* $\mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$ *replaced by* $\mathcal{H}_{\mathrm{so},m}^{\mathrm{proj}}(\rho_{\mathrm{so},m})$*. Let* $\lambda_n > 0$ *be deterministic and let* $\widehat{f}_n^{\mathrm{ridge}} \circ \breve{h}^{\mathrm{proj}}$ *be defined in (28). Then,*

$$\begin{aligned}
&\left\| \widehat{f}_n^{\mathrm{ridge}} \circ \breve{h}^{\mathrm{proj}} - a^*_{\mathrm{ta}} \right\|_{P_{\mathrm{ta}},2} \\
=&O_P\left(\sqrt{\frac{1 + \mathcal{N}_{h_m,n}(\lambda_n)}{n}} + \lambda_n + \frac{\delta_{\mathrm{so},m}/\underline{w}_n + (\delta_{\mathrm{so},m}/\underline{w}_n)^{1/2}\,\mathrm{tr}(\Sigma_{h_m,n})^{1/4}}{\sqrt{n}\lambda_n} + \frac{\delta_{\mathrm{so},m}}{\underline{w}_n} + \epsilon_{\mathrm{ta},n} \right).
\end{aligned} \tag{30}$$

The first two terms in the above rate are the estimation errors of ridge regression, and the last two terms are the transferability and approximation errors. The middle term arises from the difference between $\mathcal{N}_{\breve{h}^{\mathrm{proj}},n}(\lambda_n)$ and $\mathcal{N}_{h_m,n}(\lambda_n)$.

For a concrete rate, suppose that, uniformly over n and $j = 1, \ldots, K$,

$$\mu_j(\Sigma_{h_m,n}) \leq C_\mu j^{-2a} \qquad \text{for constants } C_\mu < \infty \text{ and } a > 1/2,$$

and $\delta_{\text{so},m}/\underline{w}_n + \epsilon_{\text{ta},n} = O(n^{-a/(2a+1)})$. Then $\mathcal{N}_{h_m,n}(\lambda) \lesssim \lambda^{-1/a}$. Choosing $\lambda_n \asymp n^{-a/(2a+1)}$ gives

$$\left\| \widehat{f}_n^{\text{ridge}} \circ \breve{h}^{\text{proj}} - a_{\text{ta}}^* \right\|_{P_{\text{ta}},2} = O_P\left(\max\left\{ n^{-a/(2a+1)}, n^{-(a+1)/(2(2a+1))} \right\} \right).$$

If $a > 1/2$, the above rate is $o_P(n^{-1/4})$, which is sufficient for the DML framework. Thus, ridge is a recommended regularized alternative when the target model is linear and the embedding is nearly identified up to a well-conditioned linear distortion.

In practice, the tuning parameter λ_n for ridge is typically selected by K-fold cross-validation that directly targets prediction risk; see, for example, Hastie, Montanari, Rosset, and Tibshirani (2022) for a recent detailed discussion of cross-validation for ridge regression, including in over-parameterized settings.

A natural question is whether a similar convergence rate can be achieved by neural network models with regularization such as random dropout or ℓ_2 regularization.[22] We conjecture that the answer is positive. For linear regression, dropout can be interpreted as a form of ridge regularization (Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov, 2014, Section 9.1). Wei, Kakade, and Ma (2020) demonstrate both theoretically and empirically that dropout can work as regularization for deep neural networks. However, even when h_m is directly observed, the convergence rate of the regularized neural network estimator is not well understood in the literature at the level of generality of Farrell et al. (2021), to the best of our knowledge. Establishing the convergence rate of regularized neural network estimators for the target model in our setting is an important direction for future research.

References

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-powered inference," Science, 382, 669–674. [Cited on page 4]

[22] Another line of research in deep learning theory on convergence rates makes use of functional shape assumptions on the function of interest (Bhattacharya, Fan, and Mukherjee, 2024; Kohler and Langer, 2021) or data structure assumptions such as approximate manifold structure (Jiao, Shen, Lin, and Huang, 2023; Schulte et al., 2025). However, their theory relies on the VC-dimension bound derived by Bartlett et al. (2019), which is always larger than the dimension of the raw input data. Thus, the convergence rate is not informative unless the dimension of the raw input data is treated as fixed relative to the sample size.

Avivi, H. (2025): "Are Patent Examiners Gender Neutral," Unpublished manuscript. [Cited on page 2]

Bach, P., V. Chernozhukov, S. Klaassen, M. Spindler, J. Teichert-Kluge, and S. Vijaykumar (2024): "Adventures in demand analysis using AI," arXiv preprint arXiv:2501.00382. [Cited on pages 2, 5, and 28]

Bajari, P., Z. Cen, V. Chernozhukov, M. Manukonda, S. Vijaykumar, J. Wang, R. Huerta, J. Li, L. Leng, G. Monokroussos, and S. Wang (2025): "Hedonic prices and quality adjusted price indices powered by AI," Journal of Econometrics, 251, 106052. [Cited on pages 2, 9, and 28]

Bartlett, P. L., N. Harvey, C. Liaw, and A. Mehrabian (2019): "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," Journal of Machine Learning Research, 20, 1–17. [Cited on pages 28, 29, and 34]

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2025): "Inference for Regression with Variables Generated by AI or Machine Learning," arXiv preprint arXiv:2402.15585. [Cited on page 5]

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," The RAND Journal of Economics, 242–262. [Cited on page 9]

Bhattacharya, S., J. Fan, and D. Mukherjee (2024): "Deep neural networks for nonparametric interaction models with diverging dimension," Annals of Statistics, 52, 2738–2766. [Cited on page 34]

Carlson, J. (2025): "Making Interpretable Discoveries from Unstructured Data: A High-Dimensional Multiple Hypothesis Testing Approach," arXiv preprint arXiv:2511.01680. [Cited on page 5]

Carlson, J. and M. Dell (2025): "A Unifying Framework for Robust and Efficient Inference with Unstructured Data," arXiv preprint arXiv:2505.00282. [Cited on page 5]

Chen, Q., V. Syrgkanis, and M. Austern (2022): "Debiased machine learning without sample-splitting for stable estimators," Advances in Neural Information Processing Systems, 35, 3096–3109. [Cited on page 8]

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models with Auxiliary Data," Annals of Statistics, 808–843. [Cited on page 5]

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," The Econometrics Journal, C1–C68. [Cited on pages 4, 7, 8, and 9]

Christensen, T. and G. Compiani (2026): "From Unstructured Data to Demand Counterfactuals: Theory and Practice," arXiv preprint arXiv:2601.05374. [Cited on page 5]

Compiani, G., I. Morozov, and S. Seiler (2025): "Demand estimation with text and image data," arXiv preprint arXiv:2503.20711. [Cited on pages 2 and 9]

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019): "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 4171–4186. [Cited on pages 13, 14, and 46]

Dube, A., J. Jacobs, S. Naidu, and S. Suri (2020): "Monopsony in online labor markets," American Economic Review: Insights, 2, 33–46. [Cited on page 2]

Dudley, R. M. (2002): Real analysis and probability, Cambridge University Press. [Cited on page 77]

Egami, N., M. Hinck, B. Stewart, and H. Wei (2023): "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models," Advances in Neural Information Processing Systems, 36, 68589–68601. [Cited on page 5]

Escanciano, J. C. and T. Pérez-Izquierdo (2026): "Automatic Locally Robust GMM with Machine-Learning-Generated Regressors," arXiv preprint arXiv:2301.10643v4. [Cited on pages 7, 41, 43, 44, and 45]

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep neural networks for estimation and inference," Econometrica, 89, 181–213. [Cited on pages 11, 25, 34, 55, and 78]

Fava, B. (2025): "Training and testing with multiple splits: A central limit theorem for split-sample estimators," arXiv preprint arXiv:2511.04957. [Cited on page 26]

Folland, G. B. (1999): Real Analysis: Modern Techniques and Their Applications, John Wiley & Sons, 2 ed. [Cited on page 77]

Fong, C. and M. Tyler (2021): "Machine learning predictions as regression covariates," Political Analysis, 29, 467–484. [Cited on page 4]

Grigsby, E., K. Lindsey, and D. Rolnick (2023): "Hidden symmetries of ReLU networks," in International Conference on Machine Learning, PMLR, 11734–11760. [Cited on page 16]

Hahn, J. and G. Ridder (2013): "Asymptotic variance of semiparametric estimators with generated regressors," Econometrica, 81, 315–340. [Cited on page 45]

Han, S. and K. Lee (2025): "Copyright and Competition: Estimating Supply and Demand with Unstructured Data," arXiv preprint arXiv:2501.16120. [Cited on page 2]

Harville, D. A. (2008): Matrix Algebra From a Statistician's Perspective, Springer Science & Business Media. [Cited on pages 30 and 91]

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022): "Surprises in high-dimensional ridgeless least squares interpolation," Annals of Statistics, 50, 949–986. [Cited on page 34]

He, K., X. Zhang, S. Ren, and J. Sun (2016): "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778. [Cited on pages 13 and 46]

Hinton, G. E. and R. R. Salakhutdinov (2006): "Reducing the dimensionality of data with neural networks," Science, 313, 504–507. [Cited on page 13]

Horn, R. and C. R. Johnson (1994): Topics in matrix analysis, Cambridge University Press Cambridge, UK. [Cited on page 90]

Jean, N., M. Burke, M. Xie, W. M. Alampay Davis, D. B. Lobell, and S. Ermon (2016): "Combining satellite imagery and machine learning to predict poverty," Science, 353, 790–794. [Cited on page 2]

Jiao, Y., G. Shen, Y. Lin, and J. Huang (2023): "Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors," Annals of Statistics, 51, 691–716. [Cited on page 34]

Johansson, F. D., D. Sontag, and R. Ranganath (2019): "Support and invertibility in domain-invariant representations," in The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 527–536. [Cited on page 20]

Klaassen, S., J. Teichert-Kluge, P. Bach, V. Chernozhukov, M. Spindler, and S. Vijaykumar (2024): "Doublemldeep: Estimation of causal effects with multimodal data," arXiv preprint arXiv:2402.01785. [Cited on page 2]

Kohler, M. and S. Langer (2021): "On the rate of convergence of fully connected deep neural network regression estimates," Annals of Statistics, 49, 2231–2249. [Cited on page 34]

Kolesár, M., U. K. Müller, and S. T. Roelsgaard (2026): "The fragility of sparsity," arXiv preprint arXiv:2311.02299. [Cited on page 31]

Kumar, A., A. Raghunathan, R. Jones, T. Ma, and P. Liang (2022): "Fine-tuning can distort pretrained features and underperform out-of-distribution," arXiv preprint arXiv:2202.10054. [Cited on page 26]

Ledoux, M. and M. Talagrand (1991): Probability in Banach Spaces: isoperimetry and processes, vol. 23, Springer. [Cited on page 90]

Ludwig, J., S. Mullainathan, and A. Rambachan (2026): "Large language models: An applied econometric framework," Annual Review of Economics, 18. [Cited on page 5]

Magnolfi, L., J. McClure, and A. Sorensen (2025): "Triplet embeddings for demand estimation," American Economic Journal: Microeconomics, 17, 282–307. [Cited on page 2]

Mammen, E., C. Rothe, and M. Schienle (2016): "Semiparametric estimation with generated covariates," Econometric Theory, 32, 1140–1177. [Cited on page 45]

Modarressi, I., J. Spiess, and A. Venugopal (2025): "Causal inference on outcomes learned from text," arXiv preprint arXiv:2503.00725. [Cited on page 5]

Murphy, K. M. and R. H. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," Journal of Business & Economic Statistics, 3, 370–379. [Cited on page 2]

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," Journal of econometrics, 79, 147–168. [Cited on page 16]

Pagan, A. (1984): "Econometric issues in the analysis of regressions with generated regressors," International Economic Review, 221–247. [Cited on page 2]

Pan, S. J. and Q. Yang (2010): "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, 22, 1345–1359. [Cited on page 3]

Rambachan, A., R. Singh, and D. Viviano (2024): "Program Evaluation with Remotely Sensed Outcomes," arXiv preprint arXiv:2411.10959. [Cited on page 5]

Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," Econometrica, 931–954. [Cited on page 8]

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," Biometrika, 70, 41–55. [Cited on page 10]

Rubin, D. B. (1976): "Inference and missing data," Biometrika, 63, 581–592. [Cited on page 9]

Schulte, R., D. Rügamer, and T. Nagler (2025): "Adjustment for confounding using pre-trained representations," arXiv preprint arXiv:2506.14329. [Cited on pages 5, 30, and 34]

Simonyan, K. and A. Zisserman (2015): "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015. [Cited on pages 13 and 46]

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, 15, 1929–1958. [Cited on page 34]

Stewart, G. W. and J.-g. Sun (1990): Matrix Perturbation Theory, Computer Science and Scientific Computing, Boston: Academic Press. [Cited on pages 61 and 90]

Tripuraneni, N., M. Jordan, and C. Jin (2020): "On the theory of transfer learning: The importance of task diversity," Advances in Neural Information Processing Systems, 33, 7852–7862. [Cited on pages 5, 28, 42, 43, and 57]

Vafa, K., S. Athey, and D. M. Blei (2025): "Estimating wage disparities using foundation models," Proceedings of the National Academy of Sciences, 122, e2427298122. [Cited on page 5]

van der Vaart, A. and J. A. Wellner (2023): Weak Convergence and Empirical Processes: With Applications to Statistics, Springer Nature. [Cited on pages 25, 58, 59, 69, 81, 82, and 85]

Wainwright, M. J. (2019): High-dimensional statistics: A non-asymptotic viewpoint, vol. 48, Cambridge University Press. [Cited on page 31]

Watkins, A., E. Ullah, T. Nguyen-Tang, and R. Arora (2023): "Optimistic rates for multi-task representation learning," Advances in Neural Information Processing Systems, 36, 2207–2251. [Cited on pages 6, 28, and 43]

Wei, C., S. Kakade, and T. Ma (2020): "The implicit and explicit regularization effects of dropout," in International Conference on Machine Learning, PMLR, 10181–10192. [Cited on page 34]

Xu, Z. and A. Tewari (2021): "Representation learning beyond linear prediction functions," Advances in Neural Information Processing Systems, 34, 4792–4804. [Cited on page 42]

Yarotsky, D. (2017): "Error bounds for approximations with deep ReLU networks," Neural Networks, 94, 103–114. [Cited on page 29]

Zhang, J., W. Xue, Y. Yu, and Y. Tan (2026): "Debiasing ML-or AI-generated regressors in partially linear models," Information Systems Research. [Cited on page 5]

Supplement to "Econometrics with Pre-trained Embeddings for Unstructured Data"

Yuya Shimizu
University of Wisconsin-Madison

Contents

This supplementary appendix contains proofs of the results in the main text as well as auxiliary results.

Additional Notation In the supplementary appendix, we use the following notation in addition to the notation introduced in the main text. For deterministic sequences a_n and b_n and for a scalar ε, we write $a_n \lesssim_\varepsilon b_n$ if there exists a constant $C_\varepsilon > 0$, depending only on ε, such that $a_n \leq C_\varepsilon b_n$ for large n. In the proofs, $C_{\varepsilon,1}, C_{\varepsilon,2}, \ldots$ denote positive constants that depend only on ε, while $C_1, C_2, \ldots$ denote positive absolute constants. These constants may denote different values in different proof sections. P^* and $\mathbb{E}^*$ denote the outer probability and outer expectation, respectively. For two symmetric matrices $A \in \mathbb{R}^{K\times K}$ and $\widetilde{A} \in \mathbb{R}^{K\times K}$, we write $A \preceq \widetilde{A}$ if $\widetilde{A} - A$ is positive semi-definite.

A Related Results in the Literature

A.1 Transfer Learning in the Machine Learning Literature

While a version of the diversity condition in Definition 2 is considered in the machine learning literature on the excess-risk scale (e.g., Tripuraneni et al., 2020; Xu and Tewari, 2021), we characterize the diversity condition by a novel relationship between the diversity condition and the set inclusion condition on the near-identified sets of the embedding function. Although the machine learning literature provides some sufficient conditions for task diversity and rate

bounds for transfer learning, this paper provides results that are more general and tailored to economic applications in the following three aspects.

First, our sufficient condition in Theorem 3 is more flexible than the existing results. The existing results focus on the full-rank coefficient matrix $B_{\text{so},m}$ for the linear source head, but our sufficient condition does not require the full-rank condition on $B_{\text{so},m}$ and allows for rank deficiency of $B_{\text{so},m}$ as long as the row span of $B_{\text{so},m}$ includes the target coefficient $\beta_{\text{ta},n}$. In addition, our sufficient condition in Theorem 3 is more flexible than the existing results as it allows different source and target covariate distributions and more general relationships between source and target models, including Lipschitz transformations.

Second, the necessity results in Theorems 4 and 5 are new to the best of our knowledge. The existing literature does not provide necessity results for the diversity condition, and our theorems show, under their respective head-class and regularity conditions, that our sufficient condition is also necessary for the linear head case and the rich source head case, respectively.

Finally, our rate of convergence of the target model is faster than the existing rates in the literature. Importantly, existing rates in the literature are not fast enough to guarantee the $o_P(n^{-1/4})$ rate required for the DML framework without the realizability assumption, which requires the risks for both the source and target tasks to converge to zero. The realizability assumption is too strong for many economic applications since it requires a very small error variance for nonparametric regressions. For example, Theorem 3 of Tripuraneni et al. (2020) shows that the excess risk of the target model is bounded by a rate always slower than $O_P(1/\sqrt{n})$, which implies a rate slower than $O_P(n^{-1/4})$ for the target model itself under the convex loss assumption in Assumption 4 (i). Theorems 1 and 2 of Watkins et al. (2023) derive excess risk bounds for the target model under the realizability and non-realizability assumptions, respectively, but the rates are not fast enough to guarantee the $o_P(n^{-1/4})$ rate for the target model without the realizability assumption. On the other hand, our target model convergence rate matches the fast rate of Watkins et al. (2023) under realizability, but does not require realizability.

A.2 Relations to Escanciano and Pérez-Izquierdo (2026)

Escanciano and Pérez-Izquierdo (2026) study locally robust estimation with generated regressors. While their theory suggests that the generated regressors affect inference on the parameter of interest in general, this effect of generated regressors does not arise in our setup. Return to the setup in Section 2.1, where we have a target parameter θ^* and nuisance functions $\eta^* = (\gamma^*, \alpha^*)$. The goal is to estimate θ^* using the estimated nuisance functions $\widehat{\eta} = (\widehat{f}_\gamma, \widehat{f}_\alpha) \circ \check{h}$ while accounting for the estimation error in the embeddings.

To clarify the relationship, we first consider the case with no approximation error in the target task $\epsilon_{\text{ta},n} = 0$, and we next consider the case with approximation error. For exposition, we also focus on a partially linear model (Example 1), but the same logic works for the other applications with closed-form Neyman orthogonal scores and nuisance functions taking conditional expectation forms, including Examples 2 to 4.

With no approximation error $\epsilon_{\text{ta},n} = 0$, the population nuisance function in the target task is $\eta^* = f_{\eta,n} \circ h_m$. The partially linear model in Example 1 has $\eta^*(Z_i) = (\gamma^*(Z_i), \alpha^*(Z_i)) = (\mathbb{E}_{\text{ta}}[Y_i|Z_i], \mathbb{E}_{\text{ta}}[X_i|Z_i])$ and $f_{\eta,n} \circ h_m(Z_i) = (f_{\gamma,n} \circ h_m(Z_i), f_{\alpha,n} \circ h_m(Z_i)) = (\mathbb{E}_{\text{ta}}[Y_i|h_m(Z_i)], \mathbb{E}_{\text{ta}}[X_i|h_m(Z_i)])$. Thus, our setup considers the case where $\mathbb{E}_{\text{ta}}[Y_i|Z_i] = \mathbb{E}_{\text{ta}}[Y_i|h_m(Z_i)]$ and $\mathbb{E}_{\text{ta}}[X_i|Z_i] = \mathbb{E}_{\text{ta}}[X_i|h_m(Z_i)]$. On the other hand, Escanciano and Pérez-Izquierdo (2026) consider a more general setup allowing $\mathbb{E}_{\text{ta}}[Y_i|Z_i] \neq \mathbb{E}_{\text{ta}}[Y_i|h_m(Z_i)] = f_{\gamma,n} \circ h_m(Z_i)$ and $\mathbb{E}_{\text{ta}}[X_i|Z_i] \neq \mathbb{E}_{\text{ta}}[X_i|h_m(Z_i)] = f_{\alpha,n} \circ h_m(Z_i)$. In our setting, $h_m(Z_i)$ captures sufficient statistics of Z_i for the target task, making the Neyman-orthogonal score insensitive to small perturbations in both f_η and h. In the more general setting considered by Escanciano and Pérez-Izquierdo (2026), the same Neyman-orthogonal score can be sensitive to perturbations in generated regressors $h(Z_i)$.

More concretely, recall that the Neyman orthogonality condition is defined as follows:

Definition 4 (Neyman Orthogonality)**.** A moment function $\psi(W; \theta, \eta)$ is *Neyman orthogonal* at (θ^*, η^*) if

$$\frac{\partial}{\partial r}\mathbb{E}_{\text{ta}}[\psi(W; \theta^*, \eta^* + r(\eta - \eta^*))]\Big|_{r=0} = 0$$

for all η in some neighborhood of η^*.

Consider a Neyman-orthogonal score for the partially linear model: $\psi(W_i; \theta, \gamma, \alpha) = \{(Y_i - \gamma(Z_i)) - (X_i - \alpha(Z_i))\theta\}(X_i - \alpha(Z_i))$. Under the regularity conditions for differentiability,

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}_{\text{ta}}[\psi(W; \theta^*, f_{\eta,n} \circ h_m + r(f_\eta \circ h - f_{\eta,n} \circ h_m))]\Big|_{r=0} \\
&\quad = \mathbb{E}_{\text{ta}}[\{(f_{\gamma,n} \circ h_m(Z) - f_\gamma \circ h(Z)) - (f_{\alpha,n} \circ h_m(Z) - f_\alpha \circ h(Z))\theta^*\}(X - f_{\alpha,n} \circ h_m(Z))] \\
&\qquad + \mathbb{E}_{\text{ta}}[\{(Y - f_{\gamma,n} \circ h_m(Z)) - (X - f_{\alpha,n} \circ h_m(Z))\theta^*\}(f_{\alpha,n} \circ h_m(Z) - f_\alpha \circ h(Z))].
\end{aligned}$$

By the law of iterated expectations, the derivative equals zero in our setup by taking conditional expectation given Z. Without these conditional-mean sufficiency restrictions, orthogonality with respect to the second-step regressions does not generally eliminate the direct evaluation effect induced by perturbing the generated regressor. This is the effect addressed by the

first-step correction in Escanciano and Pérez-Izquierdo (2026).[23] Our setup is the DML analog of the index restriction in the literature on semiparametric models (Hahn and Ridder, 2013; Mammen, Rothe, and Schienle, 2016).

When the embedding sufficiency is approximate ($\epsilon_{\text{ta},n} > 0$), Neyman orthogonality remains defined at the true nuisance vector η^*, rather than at its approximation $f_{\eta,n} \circ h_m$. Approximation and source-task estimation errors enter through the total errors of the composite nuisance estimators. Standard DML inference therefore remains valid when these total errors satisfy the required product-rate and sample-splitting conditions. In this sense, the two approaches are complementary. Our analysis derives source-to-target convergence rates for independently pre-trained embeddings, whereas Escanciano and Pérez-Izquierdo (2026) construct influence-function corrections for direct first-step effects that are not absorbed by orthogonality of the composite nuisance.

B Pre-trained Deep Learning

B.1 Fully Connected Feedforward Neural Networks (FNNs)

Let us consider a fully connected feedforward neural network with L_{NN} layers. The input layer is denoted as layer 0, and the output layer is layer L_{NN}. Each layer ℓ has H_ℓ hidden units (or neurons), and the output dimension is T. The weights connecting layer $\ell - 1$ to layer ℓ are represented by a weight matrix $W^{(\ell)} \in \mathbb{R}^{H_\ell \times H_{\ell-1}}$, where $H_0 = d$ and $H_{L_{\text{NN}}} = T$. The constant terms ("biases") for layer ℓ are represented by a vector $b^{(\ell)} \in \mathbb{R}^{H_\ell}$. Let $\sigma(\cdot)$ denote an element-wise activation function, such as the ReLU $\sigma(a) = \max\{a, 0\}$. To avoid confusion with the target functions a^*_{so} and a^*_{ta} used elsewhere in the paper, write the layer states as

$$u^{(0)}(Z) = Z$$

and, for $\ell = 1, \ldots, L_{\text{NN}} - 1$,

$$u^{(\ell)}(Z) = \sigma\left(W^{(\ell)} u^{(\ell-1)}(Z) + b^{(\ell)}\right).$$

If the economist uses the last hidden layer as the embedding, the embedding and source head in the notation of the main text are

$$h_m(Z) := u^{(L_{\text{NN}}-1)}(Z) \in \mathbb{R}^{H_{L_{\text{NN}}-1}}$$

[23] While Escanciano and Pérez-Izquierdo (2026) propose debiasing estimators for lower-dimensional generated regressors, debiasing for high-dimensional generated regressors is not studied in the literature.

and

$$g_m(v) := W^{(L_{\text{NN}})} v + b^{(L_{\text{NN}})} \in \mathbb{R}^T.$$

Our theory is also applicable to other choices of the embeddings. For example, the economist may use the second-to-last layer as the embedding, in which case

$$h_m(Z) := u^{(L_{\text{NN}}-2)}(Z) \in \mathbb{R}^{H_{L_{\text{NN}}-2}}$$

and

$$g_m(v) := W^{(L_{\text{NN}})} \sigma\left(W^{(L_{\text{NN}}-1)} v + b^{(L_{\text{NN}}-1)}\right) + b^{(L_{\text{NN}})} \in \mathbb{R}^T.$$

B.2 Convolutional Neural Networks (CNNs)

CNNs are designed for grid-structured inputs such as images. Let $Z \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}} \times C_{\text{img}}}$ denote an image with height H_{img}, width W_{img}, and C_{img} channels (e.g., RGB color scale). A CNN replaces dense linear maps by local convolutional filters and pooling operations, so that early layers detect local patterns and deeper layers aggregate them into more abstract features. Denote by

$$h_m(Z) \in \mathbb{R}^K$$

the output of the last pooling layer or penultimate layer. The remaining fully connected classification layer defines the source-task head g_m. Therefore, a pre-trained CNN, such as VGG19 (Simonyan and Zisserman, 2015) or ResNet50 (He et al., 2016), again has the form $g_m \circ h_m$. In applications, economists typically extract $\breve{h}(Z_i)$ from the final pooling layer and use these learned embeddings as regressors in the target task. Alternatively, the economist may use the last hidden layer as the embedding, in which case the source head g_m includes the final activation function and the final classification layer. Since the latter choice of embedding requires weaker assumptions for transferability, we recommend using the last hidden layer as the embedding in practice if the economist is indifferent between the two choices.

B.3 Transformer (BERT)

BERT is a bidirectional transformer encoder pre-trained on large text corpora (Devlin et al., 2019). Its original pre-training combines masked language modeling (MLM) and next sentence prediction (NSP). This structure is useful here because it makes the shared-embedding structure explicit. A packed BERT input has the form

$$Z = (\texttt{[CLS]},\ \text{span A},\ \texttt{[SEP]},\ \text{span B},\ \texttt{[SEP]}),$$

where the two spans may represent either a genuine neighboring pair or an artificially constructed pair. BERT tokenizes text into WordPiece units with a vocabulary size of $T \approx 30{,}000$. If p denotes a token position in the packed sequence, the input to the encoder is

$$u_m^{(0)}(Z, p) = e_{\text{tok}}(z_p) + e_{\text{seg}}(s_p) + e_{\text{pos}}(p),$$

where e_{tok}, e_{seg}, and e_{pos} denote token, segment, and position embeddings, respectively. Let L_{tr} denote the number of transformer blocks. Let $Z = (z_1, \ldots, z_M)$ denote a tokenized sequence of length M. For each block ℓ and token position p, let $u_m^{(\ell)}(Z, p) \in \mathbb{R}^K$ denote the hidden state at position p after block ℓ in the fitted source encoder. The transformer encoder maps $\{u_m^{(0)}(Z, p)\}_{p=1}^M$ into contextualized states $\{u_m^{(L_{\text{tr}})}(Z, p)\}_{p=1}^M$.

For MLM, let $\mathcal{M}(Z) \subset \{1, \ldots, M\}$ denote the masked positions and let Z^{corr} be the corrupted sequence used as input. The MLM head is a shallow neural network classifier

$$g_{\text{MLM},m} : \mathbb{R}^K \to \mathbb{R}^T,$$

applied to each contextualized token state $u_m^{(L_{\text{tr}})}(Z^{\text{corr}}, p)$ for $p \in \mathcal{M}(Z)$. Let

$$S^{\text{BERT}} = \big(\{S_p\}_{p \in \mathcal{M}(Z)}, S_{\text{NSP}}\big)$$

collect the MLM and NSP labels for one sequence. If S_p denotes the original token at a masked position p, the MLM loss for one sequence is

$$\ell_{\text{MLM}}(Z, S^{\text{BERT}}) = -\sum_{p \in \mathcal{M}(Z)} \log \Pr(S_p \mid Z^{\text{corr}}, p).$$

For NSP, the head is a binary classifier

$$g_{\text{NSP},m} : \mathbb{R}^K \to \mathbb{R}^2,$$

applied to the `[CLS]` state $u_m^{(L_{\text{tr}})}(Z, \texttt{[CLS]})$. If $S_{\text{NSP}} \in \{0, 1\}$ indicates whether span B truly follows span A, and

$$\pi_{\text{NSP}}(Z) = \texttt{softmax}(g_{\text{NSP},m}(u_m^{(L_{\text{tr}})}(Z, \texttt{[CLS]}))),$$

then the NSP loss is

$$\ell_{\text{NSP}}(Z, S^{\text{BERT}}) = -S_{\text{NSP}} \log \pi_{\text{NSP},1}(Z) - (1 - S_{\text{NSP}}) \log \pi_{\text{NSP},0}(Z).$$

The overall BERT pre-training loss is therefore

$$\ell_{\text{BERT}}(Z, S^{\text{BERT}}) = \ell_{\text{MLM}}(Z, S^{\text{BERT}}) + \ell_{\text{NSP}}(Z, S^{\text{BERT}}).$$

From the perspective of our theory, the important feature is that both pre-training tasks share the same encoder. The contextualized states $\{u_m^{(L_{\text{tr}})}(Z,p)\}_{p=1}^M$ provide the common embedding, while the MLM and NSP prediction layers are task-specific heads. Hence BERT can be viewed as a pre-trained model with a shared embedding map h_m and a vector-valued source task g_m. In downstream econometric applications, a common choice is either the `[CLS]` embedding

$$\breve{h}(Z) = u_m^{(L_{\text{tr}})}(Z, \texttt{[CLS]})$$

or the pooled token embedding

$$\breve{h}(Z) = \frac{1}{M}\sum_{p=1}^{M} u_m^{(L_{\text{tr}})}(Z,p).$$

The economist then treats $\breve{h}(Z)$ as the generated regressor in the target sample and estimates the downstream model $f \circ \breve{h}$.[24] Alternatively, as with CNNs, the economist may use the average of the last hidden layer outputs for the MLM task as the embedding, in which case the source head g_m includes the final activation function and the final classification layer. Our recommendation is to use the last hidden layer as the embedding:

$$\breve{h}(Z) = \frac{1}{M}\sum_{p=1}^{M} u_{\text{MLM}} \circ u_m^{(L_{\text{tr}})}(Z,p),$$

where u_{MLM} denotes the transformation from $u_m^{(L_{\text{tr}})}(Z,p)$ to the MLM head input.

Across fully connected networks, CNNs, and transformers, the same structure recurs: a shared embedding map h_m is learned in the source task, a task-specific head g_m is attached to this embedding to predict source labels, and the estimated embedding $\breve{h}$ is carried to the target sample. This common decomposition is what makes the transfer-learning framework in the main text directly applicable to standard pre-trained deep learning models.

[24] The original BERT paper fine-tunes the full network for the target task. In this paper, unless stated otherwise, we treat the pre-trained encoder as fixed and analyze the target estimator conditional on using the learned embedding $\breve{h}$.

C Transferability Condition for Deep Learning Models

In this section, we examine the transferability condition for CNNs and transformers. This appendix explains how the sufficient conditions in Theorem 3 and Corollary 1 translate into common pre-trained deep learning models. The main message is that transferability depends on how informative the source-task head is about the learned embedding. When the last source head is linear, its slope matrix has enough rank, and the target class satisfies the required functional class condition, the sufficient conditions follow from Corollary 1. When the slope matrix is low-rank or the head is nonlinear, transferability can hold only for target tasks that depend on the embedding through the part preserved by the source outputs.

C.1 Transferability Condition for Pre-trained CNN

For a standard CNN classifier, the last source head is linear:

$$g_m(v) = W_{\text{cnn},m} v + b_{\text{cnn},m},$$

where $v \in \mathbb{R}^K$ is the last pooling or penultimate embedding, $W_{\text{cnn},m} \in \mathbb{R}^{T \times K}$, and $b_{\text{cnn},m} \in \mathbb{R}^T$. This includes the empirically common case in which the computer scientist trains the CNN on a multi-class image-classification task and the economist uses the frozen final pooling layer as the downstream embedding.

Suppose that the target model class $\mathcal{F}_n = \mathcal{F}_n^{\text{sub}}$ is Lipschitz with constant $L_{\mathcal{F}}$ and that Assumption 3 holds. If $W_{\text{cnn},m}$ has full column rank, then

$$W_{\text{cnn},m}^{+} = (W_{\text{cnn},m}' W_{\text{cnn},m})^{-1} W_{\text{cnn},m}'$$

is well defined, and for every target model f_n we can define

$$\Gamma(x) = f_n \left(W_{\text{cnn},m}^{+} (x - b_{\text{cnn},m}) \right).$$

Then

$$\Gamma \circ g_m \circ h_m(Z) = f_n \left(W_{\text{cnn},m}^{+} (W_{\text{cnn},m} h_m(Z) + b_{\text{cnn},m} - b_{\text{cnn},m}) \right) = f_n \circ h_m(Z),$$

and the Lipschitz constant of Γ is bounded by

$$L_\Gamma \leq L_{\mathcal{F}} \| W_{\text{cnn},m}^{+} \|_{\text{sp}}.$$

Therefore, by Corollary 1, the CNN source task is $(\rho_{\mathrm{so},m}, \rho_{\mathrm{ta},n})$-diverse over f_n with

$$\rho_{\mathrm{ta},n} = \frac{L_{\mathcal{F}}}{\sigma^{+}_{\min}(W_{\mathrm{cnn},m})\underline{w}_n}\rho_{\mathrm{so},m},$$

provided that the functional class requirement in Corollary 1 is satisfied. Specifically, for every candidate embedding h and corresponding best linear source head $b^{\bullet}_{\mathrm{cnn},m} + W^{\bullet}_{\mathrm{cnn},m}(\cdot)$, the target class must contain the function

$$u \mapsto f_n\left(W^{+}_{\mathrm{cnn},m}(b^{\bullet}_{\mathrm{cnn},m} - b_{\mathrm{cnn},m} + W^{\bullet}_{\mathrm{cnn},m}u)\right).$$

The bias difference accounts for translations of the candidate embedding that can be absorbed by the source intercept. This functional class condition holds, for example, when the target class is closed under the displayed linear transformations with intercept shifts. Transferability is stronger when the smallest singular value of $W_{\mathrm{cnn},m}$ is bounded away from zero, because then $\|W^{+}_{\mathrm{cnn},m}\|_{\mathrm{sp}}$ is small.

If $W_{\mathrm{cnn},m}$ is rank deficient, then universal transferability fails in general. When the target subclass consists of linear heads and the other conditions of Theorem 4 hold, $(0,0)$-diversity can hold only if the target coefficient belongs to the row span of $W_{\mathrm{cnn},m}$. More generally, Theorem 5 implies that transferability requires the existence of a measurable map Γ such that

$$f_n \circ h_m(Z) = \Gamma \circ g_m \circ h_m(Z),$$

P_{ta}-a.s. Hence, a CNN with a wide hidden embedding and relatively few source labels cannot be expected to transfer to arbitrary downstream targets; it transfers only to targets that depend on the embedding through the information preserved by the source classifier.

C.2 Transferability Condition for Pre-trained Transformer

For BERT-style transformers, the relevant source tasks depend on how the downstream embedding is constructed and on the fact that pre-training is carried out on a corrupted sequence $\widetilde{Z} = C(Z, R)$ rather than on the original sequence Z. Two common choices are the `[CLS]` embedding and the mean-pooled token embedding described in Appendix B. The main distinction is that the next sentence prediction task is a single sequence-level task, whereas masked language modeling generates many token-level source tasks.

`[CLS]` embedding. Suppose that the economist uses

$$h_m(Z) = u_m^{(L_{\text{tr}})}(Z, \texttt{[CLS]}).$$

If the only relevant source task is next sentence prediction, then the source head is linear of the form

$$g_{\text{NSP},m}(v) = W_{\text{NSP},m} v + b_{\text{NSP},m},$$

where $W_{\text{NSP},m} \in \mathbb{R}^{2\times K}$. The full-rank sufficient condition in Corollary 1 can then hold only if $\text{rank}(W_{\text{NSP},m}) = K$, which requires $K \leq 2$. For empirically relevant transformer embeddings such as BERT, K is much larger than 2. Therefore, next sentence prediction alone cannot generically identify a high-dimensional `[CLS]` embedding. Transferability based only on this binary source task can hold only for target functions that depend on $h_m(Z)$ through the low-dimensional information preserved by $g_{\text{NSP},m} \circ h_m(Z)$.

Mean-pooled token embedding. The more favorable case is when the economist uses a pooled embedding built from the token-level states relevant for masked language modeling. The BERT architecture may insert a small transform layer before the final vocabulary projection. Let

$$u_m(Z, p) = u_{\text{MLM},m} \circ u_m^{(L_{\text{tr}})}(Z, p) \in \mathbb{R}^K$$

denote the token state after this optional MLM-specific transform. If the economist uses the mean-pooled token embedding, the relevant target-side representation is

$$h_m(Z) = \frac{1}{M} \sum_{p=1}^{M} u_m(Z, p).$$

The MLM head then takes the linear form

$$g_{\text{MLM},m}(v) = W_{\text{MLM},m} v + b_{\text{MLM},m},$$

where $W_{\text{MLM},m} \in \mathbb{R}^{T\times K}$. In BERT, this matrix is often tied to the input token-embedding matrix, but this does not change the linear form of the final projection. Because the same vocabulary projection is applied to each token position, averaging the token-level logits on the corrupted source input yields

$$\frac{1}{M} \sum_{p=1}^{M} g_{\text{MLM},m}(u_m(\widetilde{Z}, p)) = W_{\text{MLM},m} h_m(\widetilde{Z}) + b_{\text{MLM},m},$$

where

$$h_m(\widetilde{Z}) = \frac{1}{M}\sum_{p=1}^{M} u_m(\widetilde{Z}, p).$$

Thus, for the pooled representation induced by the corrupted sequence, the relevant source head is

$$\overline{g}_{\mathrm{MLM},m}(v) = W_{\mathrm{MLM},m} v + b_{\mathrm{MLM},m}.$$

If $W_{\mathrm{MLM},m}$ has full column rank, then the full-rank sufficient condition in Corollary 1 applies to $\overline{g}_{\mathrm{MLM},m}$ on the source side. In particular, for every Lipschitz target model f_n in the function class, there exists

$$\Gamma(x) = f_n\left(W^{+}_{\mathrm{MLM},m}(x - b_{\mathrm{MLM},m})\right)$$

such that

$$f_n \circ h_m(\widetilde{Z}) = \Gamma \circ \overline{g}_{\mathrm{MLM},m} \circ h_m(\widetilde{Z}),$$

provided that the target class satisfies the functional class condition in Corollary 1. In particular, for every candidate embedding h and corresponding best linear source head $b^{\bullet}_{\mathrm{MLM},m} + W^{\bullet}_{\mathrm{MLM},m}(\cdot)$, the target class must contain

$$u \mapsto f_n\left(W^{+}_{\mathrm{MLM},m}(b^{\bullet}_{\mathrm{MLM},m} - b_{\mathrm{MLM},m} + W^{\bullet}_{\mathrm{MLM},m} u)\right).$$

Under this functional class condition,

$$\rho_{\mathrm{ta},n} = \frac{L_{\mathcal{F}}}{\sigma^{+}_{\min}(W_{\mathrm{MLM},m})\underline{w}_n}\rho_{\mathrm{so},m}.$$

This condition is substantially more plausible than in the `[CLS]`-NSP case because the MLM task has a much larger output dimension, and thus the final vocabulary projection can carry richer information about the pooled embedding. The optional dense-GELU-layer-normalization block in the BERT prediction head is not an obstacle here, because it can be absorbed into the definition of the embedding map u_m. If $W_{\mathrm{MLM},m}$ is rank deficient, universal transferability again fails in general. In that case, only target functions that depend on the pooled embedding through the information preserved by the row span of $W_{\mathrm{MLM},m}$ can be expected to transfer.

One difficulty with transformers is that BERT pre-training is carried out on a random corruption of the clean sequence rather than on the clean sequence itself. Let $\mathcal{I}_{\mathrm{pred}} \subseteq \{1, \ldots, M\}$ denote the set of token positions eligible for masked language modeling, excluding positions occupied by the special tokens `[CLS]`, `[SEP]`, and `[PAD]`. Draw the entries of a mask vector $R = (R_1, \ldots, R_M)$ independently across positions $p \in \mathcal{I}_{\mathrm{pred}}$. For these positions, let

$R_p = 1$ with probability π and $R_p = 0$ otherwise, and set $R_p = 0$ for $p \notin \mathcal{I}_{\text{pred}}$. In BERT, $\pi = 0.15$. Let $\mathcal{U}(R) = \{p \in \mathcal{I}_{\text{pred}} : R_p = 1\}$ denote the set of masked positions. The corrupted sequence $\widetilde{Z} = C(Z, R)$ is then defined coordinatewise by

$$\widetilde{Z}_p = \begin{cases} \texttt{[MASK]}, & p \in \mathcal{U}(R) \text{ with probability } 0.8, \\ Z_p^{\text{rnd}}, & p \in \mathcal{U}(R) \text{ with probability } 0.1, \\ Z_p, & p \in \mathcal{U}(R) \text{ with probability } 0.1, \\ Z_p, & p \notin \mathcal{U}(R), \end{cases}$$

where Z_p^{rnd} is drawn uniformly from the BERT vocabulary. For a fixed clean sequence and a masked position, averaging the source loss over this corruption distribution yields the masked conditional objective used in pre-training. Therefore, the observed MLM loss is an unbiased Monte Carlo approximation to that population criterion. The 80/10/10 rule also prevents the source task from depending only on the literal `[MASK]` token and thereby mitigates the train-target mismatch between masked source inputs and unmasked target inputs.

In our framework, the source task therefore identifies the representation through $h_m(\widetilde{Z})$ under the joint law induced by the clean source sequence and the corruption draw, whereas the economist uses $h_m(Z)$ in the target sample. A concrete route to a positive $\underline{w}_n$ comes from the fact that the clean sequence is observed with positive probability under the BERT corruption scheme. Indeed, each eligible token is left unchanged with probability $1 - 0.9\pi$, so

$$p_M = \Pr(\widetilde{Z} = Z \mid Z) \geq (1 - 0.9\pi)^M > 0.$$

On the event $\{\widetilde{Z} = Z\}$, the pooled embedding is unchanged. Therefore, if the uncorrupted source distribution of Z already satisfies the density ratio condition in Assumption 3, then the corrupted-source comparison above also holds with a constant $\underline{w}_n$ equal to the clean-sample constant multiplied by $p_M^{1/2}$. This bound is likely to be too conservative because it uses only the event that no effective corruption occurs. The mean-pooling structure suggests that a larger constant may be available under additional stability conditions on the transformer states, since many tokens remain unchanged and the average embedding can still preserve the target-relevant variation even when some positions are corrupted. Thus, the corruption operator need not remove the components of the pooled embedding that matter for the downstream econometric task, and when this comparison holds, the sufficient transferability arguments above continue to apply with the corrupted source covariate $\widetilde{Z}$ in place of the original unmasked sequence.

D Source Head Transformation for Ridge Regression

For any $B \in \mathbb{R}^{T\times K}$, define

$$\begin{aligned} R_B &:= (B'B)^{1/2} + I_K - B^+B, \\ B^{\mathrm{N}} &:= BR_B^{-1}, \\ h_B^{\mathrm{N}} &:= R_B h. \end{aligned} \tag{31}$$

Under these linear transformations of the source head and embedding function, we obtain the same source output $g_{\alpha,B} \circ h = g_{\alpha,B^{\mathrm{N}}} \circ h_B^{\mathrm{N}}$ and the desirable properties $B^{\mathrm{N}}(B^{\mathrm{N}})'B^{\mathrm{N}} = B^{\mathrm{N}}$ and $(B^{\mathrm{N}})^+ = (B^{\mathrm{N}})'$.[25] Note that these transformations are simple and only require the source head matrix as well as the estimated embedding function $\breve{h}$ from the source task. The source head matrix is typically available to the economist in the pre-trained source model. After this transformation, if $\mathrm{rank}(B) \geq 1$, then the transformed source head B^{N} satisfies $\|B^{\mathrm{N}}\|_{\mathrm{sp}} \leq \overline{\sigma}$ and $\sigma^+_{\min}(B^{\mathrm{N}}) \geq \underline{\sigma}$ with $\overline{\sigma} = \underline{\sigma} = 1$.

After the transformation, we can take the projected estimated embedding as $\breve{h}^{\mathrm{proj}} := (\breve{B}^{\mathrm{N}}_{\mathrm{so}})^+ \breve{B}^{\mathrm{N}}_{\mathrm{so}} \breve{h}^{\mathrm{N}}$. Thus, every occurrence of $\breve{h}^{\mathrm{proj}}$ in the ridge estimator refers to this projection. We also form $\breve{q}$ and $\breve{Q}$ after transformation. If the transformation is used, the population and pre-trained composite source predictions, and hence the convergence rate in Assumption 1 (i), remain unchanged.

The transformation also preserves target-slope compatibility. Because R_B is symmetric, if $\beta' = b'B$ before transformation, then $\beta^{\mathrm{N}} := R_B^{-1}\beta$ satisfies

$$\begin{aligned} \beta^{\mathrm{N}\prime} h_B^{\mathrm{N}} &= \beta' h, \\ \beta^{\mathrm{N}\prime} &= b' B^{\mathrm{N}}. \end{aligned}$$

[25] If $\mathrm{rank}(B) = r \geq 1$ and $B = U_B D_B V_B'$ is the compact singular-value decomposition, where $D_B \in \mathbb{R}^{r\times r}$ contains the positive singular values, then

$$\begin{aligned} R_B &= V_B D_B V_B' + I_K - V_B V_B', \\ R_B^{-1} &= V_B D_B^{-1} V_B' + I_K - V_B V_B', \\ B^{\mathrm{N}} &= U_B V_B'. \end{aligned}$$

Thus, R_B is invertible even when B is rank deficient. The transformation preserves the composite source prediction and satisfies

$$\begin{aligned} g_{\alpha,B^{\mathrm{N}}} \circ h_B^{\mathrm{N}} &= g_{\alpha,B} \circ h, \\ B^{\mathrm{N}}(B^{\mathrm{N}})'B^{\mathrm{N}} &= B^{\mathrm{N}}, \\ (B^{\mathrm{N}})^+ &= (B^{\mathrm{N}})'. \end{aligned}$$

Every positive singular value of B^{N} therefore equals one.

Thus, the transformation changes neither the target score nor the norm of the linking coefficient b. When the transformation is applied to the population source matrix, we reuse $\beta_{\mathrm{ta},n}$ for this transformed target slope.

E Technical Lemmas

Lemma 2 (Transferability Decomposition)**.** *Suppose that $\mathcal{F}_n^{\mathrm{sub}}$ is a class of square-integrable functions and that $\mathcal{H}_m$ is a class of measurable functions. For every $f_n \in L^2(P)$ and every $h \in \mathcal{H}_m$,*

$$\begin{aligned}&\inf_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - f_n \circ h_m\|_{P,2}^2 \\ &\quad = \mathbb{E}\left[\mathrm{Var}\left(f_n \circ h_m(Z) \mid h(Z)\right)\right] + \inf_{f \in \mathcal{F}_n^{\mathrm{sub}}} \mathbb{E}\left[\left(\mathbb{E}\left[f_n \circ h_m(Z) \mid h(Z)\right] - f \circ h(Z)\right)^2\right].\end{aligned}$$

Remark 6. The first term on the right-hand side represents the variation of $f_n \circ h_m(Z)$ that cannot be explained by $h(Z)$; the second term represents the best approximation of the conditional expectation of $f_n \circ h_m(Z)$ given $h(Z)$ by functions in $\mathcal{F}_n^{\mathrm{sub}}$. Our sufficient condition for transferability in Theorem 3 ensures that both terms are small when $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$. Our necessary conditions in Theorems 4 and 5 make the first term equal to zero for some $h \in \mathcal{H}_{\mathrm{so},m}(0)$. Example 10 illustrates the importance of the second term as it can be positive even when the first term is zero, which leads to failure of transferability. ■

Lemma 3 (Doob-Dynkin Factorization under Completion)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space, let $U : \Omega \to \mathcal{U}$ be a random element taking values in a measurable space, and let $V : \Omega \to \mathbb{R}^q$ be a q-dimensional random vector. Let $\overline{\sigma(U)}^P$ be the P-completion of $\sigma(U)$, which is defined by $\overline{\sigma(U)}^P = \{A \cup N : A \in \sigma(U), N \subset B \text{ for some } B \in \sigma(U) \text{ with } P(B) = 0\}$. If*

$$\sigma(V) \subseteq \overline{\sigma(U)}^P,$$

then there exists a measurable function $\Gamma : \mathcal{U} \to \mathbb{R}^q$ such that $V = \Gamma \circ U$, P-a.s.

Lemma 4 (Properties of Loss Functions, Farrell et al., 2021, Lemma 8)**.** *Let P denote the distribution of Z. In each setup below, suppose that $\|f\|_\infty, \|a\|_\infty, \|f^*\|_\infty \leq M$ for some constant $M > 0$ and all candidate functions f and a. For each setup, there exist constants $c_L > 0$, $c_U > 0$, and $C_\ell > 0$ such that*

$$c_L \|f - f^*\|_{P,2}^2 \leq \mathbb{E}[\ell(f(Z), Y)] - \mathbb{E}[\ell(f^*(Z), Y)] \leq c_U \|f - f^*\|_{P,2}^2$$

and

$$|\ell(f(z), y) - \ell(a(z), y)| \leq C_\ell |f(z) - a(z)|$$

for all candidate functions f and a, every $z \in \mathcal{Z}$, and every y in the range of Y.

(a) ***Least squares.*** *Suppose that $|Y| \leq M$ almost surely, $f^*(Z) = \mathbb{E}[Y \mid Z]$, and $\ell(u, y) = (1/2)(y - u)^2$. Then, we can take $c_L = c_U = 1/2$ and $C_\ell = 2M$.*

(b) ***Binary logistic regression.*** *Suppose that $Y \in \{0, 1\}$ almost surely, $P(Y = 1 \mid Z) > 0$ almost surely, and $P(Y = 0 \mid Z) > 0$ almost surely. Let*

$$f^*(Z) = \log\left(\frac{P(Y = 1 \mid Z)}{P(Y = 0 \mid Z)}\right)$$

and $\ell(u, y) = -yu + \log(1 + \exp(u))$. Then, we can take $c_L = 1/(2(\exp(M) + \exp(-M) + 2))$, $c_U = 1/8$, and $C_\ell = 1$.

Lemma 5 (Properties of Loss Functions for Multinomial Logit)**.** *Consider the $T + 1$-class classification problem with the multinomial-logit negative log-likelihood loss. Let P denote the distribution of Z. Suppose that $Y \in \{1, \ldots, T + 1\}$ almost surely and that $P(Y = t \mid Z) > 0$ almost surely for every $t = 1, \ldots, T + 1$. Take class $T + 1$ as the baseline category and let $f^* : \mathcal{Z} \to \mathbb{R}^T$ be a measurable version of the population log-odds vector $f^* = (f_1^*, \ldots, f_T^*)'$ defined by*

$$f_t^*(Z) = \log\left(\frac{P(Y = t \mid Z)}{P(Y = T + 1 \mid Z)}\right), \qquad t = 1, \ldots, T, \quad P\text{-a.s.},$$

and define the multinomial logistic loss by

$$\ell(u, y) = -\sum_{t=1}^{T} \mathbb{1}\{y = t\} u_t + \log\left(1 + \sum_{t=1}^{T} \exp(u_t)\right), \qquad u \in \mathbb{R}^T.$$

Suppose that, for some constant $M > 0$ and all candidate functions f and a,

$$\sup_{z \in \mathcal{Z}} \|f(z)\|_2, \sup_{z \in \mathcal{Z}} \|a(z)\|_2, \sup_{z \in \mathcal{Z}} \|f^*(z)\|_2 \leq M.$$

Then,

$$c_L = \frac{1}{2(1 + T\exp(M))^2}, \qquad c_U = \frac{\exp(M)}{2(1 + (T - 1)\exp(-M) + \exp(M))}, \qquad C_\ell = \sqrt{2}$$

satisfy

$$c_L \|f - f^*\|_{P,2}^2 \leq \mathbb{E}[\ell(f(Z), Y)] - \mathbb{E}[\ell(f^*(Z), Y)] \leq c_U \|f - f^*\|_{P,2}^2$$

and

$$|\ell(f(z), y) - \ell(a(z), y)| \leq C_\ell \|f(z) - a(z)\|_2$$

for all candidate functions f and a, every $z \in \mathcal{Z}$, and every $y \in \{1, \ldots, T+1\}$.

Remark 7. For fixed $M > 0$, $c_L \asymp T^{-2}$ as $T \to \infty$, and this order cannot generally be improved. To see this, take Z to be deterministic and $f^*(z) = \mathbf{0}_T \in \mathbb{R}^T$, so that $P(Y = t \mid Z) = 1/(T+1)$ for $t = 1, \ldots, T+1$. Let $\delta_T = MT^{-1/2}$ and choose $f(z) = \delta_T \mathbf{1}_T$, which satisfies $\|f - f^*\|_{P,2}^2 = M^2$ and $\sup_{z \in \mathcal{Z}} \|f(z)\|_2 = M$. Then,

$$\begin{aligned}
\mathbb{E}[\ell(f(Z), Y)] - \mathbb{E}[\ell(f^*(Z), Y)] &= -\frac{T\delta_T}{T+1} + \log(1 + T\exp(\delta_T)) - \log(1+T) \\
&= \frac{\delta_T}{T+1} + \log\left(1 + \frac{\exp(-\delta_T) - 1}{T+1}\right) \\
&= \frac{\delta_T + \exp(-\delta_T) - 1}{T+1} + o(T^{-2}) \\
&= \frac{M^2}{2T(T+1)} + o(T^{-2}) \asymp T^{-2},
\end{aligned}$$

where the third equality uses $\log(1+x) = x + O(x^2)$ and the fourth equality uses $\exp(-x) = 1 - x + x^2/2 + o(x^2)$ as $x \to 0$. ■

For a norm $\|\cdot\|$ and $\varepsilon > 0$, let $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ denote the covering number of $\mathcal{F}$, i.e., the minimum number of balls of radius ε with respect to the norm $\|\cdot\|$ needed to cover the set $\mathcal{F}$. The following two lemmas are also proved as part of the proof of Theorem 7 in Tripuraneni et al. (2020), but we state them here for completeness and clarity.

Lemma 6 (Properties of Products of Covering Numbers)**.** *Let $\mathcal{H}$ be a measurable class of functions with K-dimensional outputs. Suppose that $\mathcal{H}_k$ is the k-th coordinate projection of $\mathcal{H}$, i.e., $\mathcal{H}_k = \{h_k : h = (h_1, \ldots, h_K) \in \mathcal{H}\}$. Then, we have*

$$N\left(\varepsilon, \mathcal{H}, L^2(Q)\right) \leq \prod_{k=1}^{K} N\left(\varepsilon/K, \mathcal{H}_k, L^2(Q)\right).$$

Lemma 7 (Properties of Composition of Covering Numbers)**.** *Let $\mathcal{F}$ and $\mathcal{H}$ be measurable classes of functions. Suppose that $\mathcal{F}$ is Lipschitz with constant $L_{\mathcal{F}}$. Then, we have*

$$N\left(\varepsilon, \mathcal{F} \circ \mathcal{H}, L^2(Q)\right) \leq N\left(\varepsilon/(2L_{\mathcal{F}}), \mathcal{H}, L^2(Q)\right) \cdot \sup_{h' \in \mathcal{H}} N\left(\varepsilon/2, \mathcal{F}, L^2(Q_{h'})\right),$$

where $Q_{h'}$ is the measure such that $Q_{h'}(A) = Q(h'^{-1}(A))$ for any measurable set A.

Definition 5 (Uniform Entropy Integral)**.** For a class $\mathcal{F}$ of measurable functions with measurable envelope function F, define the *uniform entropy integral* by

$$J\left(\delta, \mathcal{F}|F, L_2\right) = \sup_Q \int_0^\delta \sqrt{1 + \log N\left(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right)} d\varepsilon,$$

where the supremum is taken over all discrete probability measures Q with $\|F\|_{Q,2} > 0$.

Importantly, the uniform entropy integral depends only on the function class and its domain, not on a particular probability measure.

Lemma 8 (Localized Maximal Inequality for Lipschitz Loss)**.** *Let $\mathcal{F}_n$ be a pointwise measurable class of measurable functions mapping into $\mathbb{R}^T$ with $\sup_v \|f(v)\|_2 \leq M < \infty$ for every $f \in \mathcal{F}_n$, let $\mathcal{H}_m$ be a class of measurable functions, and let $h \mapsto f^*_{n,h} \in \mathcal{F}_n$ be a mapping. Let $m_{f,h}(z, y) = -\ell(f \circ h(z), y)$, $\mathbb{M}_n(f, h) = \mathbb{P}_n m_{f,h}$, $M_n(f, h) = P m_{f,h}$, $\mathbb{G}_n g = \sqrt{n}(\mathbb{P}_n - P)g$, and, for every $h \in \mathcal{H}_m$ and $\delta > 0$, define $d_{n,h}(f, f') = \|f \circ h - f' \circ h\|_{P,2}$, $B_{n,h}(\delta) = \left\{f \in \mathcal{F}_n : d_{n,h}\left(f, f^*_{n,h}\right) < \delta\right\}$, and $\mathcal{M}_{n,h}(\delta) = \left\{m_{f,h} - m_{f^*_{n,h},h} : f \in B_{n,h}(\delta)\right\}$.*

Suppose that $\ell : \mathbb{R}^T \times \mathcal{Y} \mapsto \mathbb{R}$ is Lipschitz continuous in the first argument with respect to $\|\cdot\|_2$ with constant C_ℓ. Then,

$$\sup_{h \in \mathcal{H}_m} \mathbb{E}_P \left[\sup_{f \in B_{n,h}(\delta)} \sqrt{n} \left| \left(\mathbb{M}_n - M_n\right)(f, h) - \left(\mathbb{M}_n - M_n\right)\left(f^*_{n,h}, h\right) \right| \right]$$
$$\lesssim C_\ell M \left(J\left(\delta/M, \mathcal{F}_n|M, L_2\right) + \frac{M^2 J^2\left(\delta/M, \mathcal{F}_n|M, L_2\right)}{\delta^2 \sqrt{n}} \right).$$

Lemma 9 (Localized Pointwise Measurable Class)**.** *Let $T \geq 1$ and let $\mathcal{F}$ be a pointwise measurable class of measurable functions mapping into $\mathbb{R}^T$ with measurable envelope function F such that $\|f(x)\|_2 \leq F(x)$ for every $f \in \mathcal{F}$ and $PF^2 < \infty$. Then, the localized class $B(\delta) = \{f \in \mathcal{F} : \|f - f^*\|_{P,2} < \delta\}$ is also pointwise measurable for every $\delta > 0$ and every measurable function f^* mapping into $\mathbb{R}^T$ such that $\|f^*(x)\|_2 \leq F(x)$.*

Lemma 10 (Covering Number for VC Function Classes, van der Vaart and Wellner, 2023, Theorem 2.6.7)**.** *For a VC-class of functions with measurable envelope function F and $r \geq 1$, every probability measure Q with $\|F\|_{Q,r} > 0$ satisfies*

$$N\left(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right) \lesssim \mathsf{V}(\mathcal{F})(16e)^{\mathsf{V}(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r\mathsf{V}(\mathcal{F})},$$

for $0 < \varepsilon < 1$.

Lemma 11 (Rate of Convergence)**.** *For each n, let $\mathcal{F}_n$ be a set of measurable functions, let $\mathcal{F}_n^*$ be a set of measurable functions, and let $\mathcal{H}_m$ be a set of measurable functions h. Assume that for every fixed $h \in \mathcal{H}_m$, there exist mappings*

$$h \mapsto f_{n,h} \in \mathcal{F}_n, \quad h \mapsto f_{n,h}^* \in \mathcal{F}_n^*.$$

For each n, let $\mathbb{M}_n$ be a stochastic process and let M_n be a deterministic function indexed by a set $(\mathcal{F}_n \cup \mathcal{F}_n^*) \times \mathcal{H}_m$. For every fixed $h \in \mathcal{H}_m$, let $f \mapsto d_{n,h}(f, f_{n,h}^*)$ be a deterministic function from $\mathcal{F}_n$ to $[0, \infty)$, and let $\underline{\delta}_n \geq 0$ and $\tau_n \in (0, 1]$ be some sequences. We also assume that for every fixed $h \in \mathcal{H}_m$ and $\delta > 0$, the set $\{f \in \mathcal{F}_n : d_{n,h}(f, f_{n,h}^*) < \delta\}$ is pointwise measurable. Let $\check{h} \in \mathcal{H}_m$ be some possibly random function estimated using a sample of size m independent of $\mathbb{M}_n$, M_n, and $\widehat{f}_{n,h}$ defined below. For every $\varepsilon > 0$, let $\mathcal{H}_{m,\varepsilon} \subseteq \mathcal{H}_m$ be a subset of $\mathcal{H}_m$ such that $\limsup_{m\to\infty} P(\check{h} \notin \mathcal{H}_{m,\varepsilon}) \leq \varepsilon$. Suppose that, for every n, every $\varepsilon > 0$, and every positive $\delta \geq \tau_n^{-1} C_{\varepsilon,0} \underline{\delta}_n$ for some constant $C_{\varepsilon,0} > 0$ depending on ε,

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \sup_{f \in \mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f_{n,h}^*\right) \leq \delta} M_n(f, h) - M_n\left(f_{n,h}^*, h\right) \lesssim_\varepsilon -\tau_n^2 \delta^2, \tag{32}$$

and

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{f \in \mathcal{F}_n : d_{n,h}\left(f, f_{n,h}^*\right) < \delta} \sqrt{n}\left|\left(\mathbb{M}_n - M_n\right)(f, h) - \left(\mathbb{M}_n - M_n\right)\left(f_{n,h}^*, h\right)\right|\right] \lesssim_\varepsilon \phi_n(\delta) \tag{33}$$

for some increasing functions $\phi_n : [\underline{\delta}_n, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\xi$ is decreasing for some $\xi < 2$. Let δ_n satisfy

$$\phi_n\left(\delta_n/\tau_n\right) \lesssim_\varepsilon \sqrt{n}\delta_n^2, \quad \delta_n^2 \gtrsim_\varepsilon \sup_{h \in \mathcal{H}_{m,\varepsilon}} M_n\left(f_{n,h}^*, h\right) - M_n\left(f_{n,h}, h\right), \quad \delta_n \gtrsim_\varepsilon \underline{\delta}_n. \tag{34}$$

Let the sequence $\widehat{f}_{n,h}$ take values in $\mathcal{F}_n$ for every fixed $h \in \mathcal{H}_m$. If $\mathbb{M}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) \geq \mathbb{M}_n\left(f_{n,\check{h}}, \check{h}\right) - O_P\left(\delta_n^2\right)$ is satisfied, then $d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right) = O_P\left(\delta_n/\tau_n\right)$.

Remark 8. The pointwise-measurability assumptions are required to ensure the measurability of the supremum within the expectations. We can eliminate this requirement by using outer expectations and probabilities instead (see van der Vaart and Wellner, 2023, Sections 1.2 and 2.3). ■

Lemma 12 (Rate of Convergence for Regularized Estimator)**.** *For each n, let $\mathcal{F}_n$ be a set of measurable functions, let $\mathcal{F}_n^*$ be a set of measurable functions, and let $\mathcal{H}_m$ be a set of measurable*

functions h. Assume that for every fixed $h \in \mathcal{H}_m$, there exist mappings

$$h \mapsto f_{n,h} \in \mathcal{F}_n, \quad h \mapsto f^*_{n,h} \in \mathcal{F}^*_n.$$

For each n, let $\mathbb{M}_n$ be a stochastic process and let M_n be a deterministic function indexed by a set $(\mathcal{F}_n \cup \mathcal{F}^*_n) \times \mathcal{H}_m$. For every fixed $h \in \mathcal{H}_m$, let $f \mapsto d_{n,h}(f, f^*_{n,h})$ be a deterministic function from $\mathcal{F}_n$ to $[0,\infty)$, and let $\underline{\delta}_n \geq 0$, $\tau_n \in (0,1]$, and $\lambda_n > 0$ be some sequences. Let $\mathcal{J}_n : \mathcal{F}_n \to [0,\infty)$ be some complexity measure of f. We also assume that for every fixed $h \in \mathcal{H}_m$ and $\delta > 0$, the set $\{f \in \mathcal{F}_n : d_{n,h}(f, f^*_{n,h}) < \delta, \mathcal{J}_n(f) < \tau_n \delta/\lambda_n\}$ is pointwise measurable. Let $\check{h} \in \mathcal{H}_m$ be some possibly random function estimated using a sample of size m that is independent of $\mathbb{M}_n$, M_n, $\widehat{f}_{n,h}$, and $\widehat{\lambda}_{n,h}$ defined below. For every $\varepsilon > 0$, let $\mathcal{H}_{m,\varepsilon} \subseteq \mathcal{H}_m$ be a subset of $\mathcal{H}_m$ such that $\limsup_m P(\check{h} \notin \mathcal{H}_{m,\varepsilon}) \leq \varepsilon$. Suppose that, for every n, every $\varepsilon > 0$, and every positive $\delta \geq \tau_n^{-1} C_{\varepsilon,0} \underline{\delta}_n$ for some constant $C_{\varepsilon,0} > 0$ depending on ε,

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \sup_{f \in \mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f^*_{n,h}\right) \leq \delta} M_n(f,h) - M_n\left(f^*_{n,h}, h\right) \lesssim_\varepsilon -\tau_n^2 \delta^2, \tag{35}$$

and

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{\substack{f \in \mathcal{F}_n : d_{n,h}\left(f, f^*_{n,h}\right) < \delta \\ \mathcal{J}_n(f) < \tau_n \delta/\lambda_n}} \sqrt{n} \left| \left(\mathbb{M}_n - M_n\right)(f,h) - \left(\mathbb{M}_n - M_n\right)\left(f^*_{n,h}, h\right) \right| \right] \lesssim_\varepsilon \phi_n(\delta) \tag{36}$$

for some increasing functions $\phi_n : [\underline{\delta}_n, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\xi$ is decreasing for some $\xi < 2$. Let δ_n satisfy

$$\begin{gathered} \phi_n\left(\delta_n/\tau_n\right) \lesssim_\varepsilon \sqrt{n}\delta_n^2, \quad \delta_n^2 \gtrsim_\varepsilon \sup_{h \in \mathcal{H}_{m,\varepsilon}} M_n\left(f^*_{n,h}, h\right) - M_n\left(f_{n,h}, h\right), \quad \delta_n \gtrsim_\varepsilon \underline{\delta}_n, \\ \sup_{h \in \mathcal{H}_{m,\varepsilon}} \lambda_n \mathcal{J}_n\left(f_{n,h}\right) < \delta_n, \quad \sup_{h \in \mathcal{H}_{m,\varepsilon}} d_{n,h}\left(f_{n,h}, f^*_{n,h}\right) < \delta_n/\tau_n. \end{gathered} \tag{37}$$

Let the sequence $\widehat{f}_{n,h}$ take values in $\mathcal{F}_n$ for every fixed $h \in \mathcal{H}_m$ and $\widehat{\lambda}_{n,h}$ be a sequence of nonnegative random variables for every fixed $h \in \mathcal{H}_m$. If

$$\mathbb{M}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \widehat{\lambda}^2_{n,\check{h}} \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) \geq \mathbb{M}_n\left(f_{n,\check{h}}, \check{h}\right) - \widehat{\lambda}^2_{n,\check{h}} \mathcal{J}_n^2\left(f_{n,\check{h}}\right) - O_P\left(\delta_n^2\right), \tag{38}$$

then, on the event $\{\widehat{\lambda}_{n,\check{h}} \geq \lambda_n\}$, we have $d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f^*_{n,\check{h}}\right) = O_P(1) \times \left(\delta_n/\tau_n + \widehat{\lambda}_{n,\check{h}} \mathcal{J}_n\left(f_{n,\check{h}}\right)/\tau_n\right)$.

The next lemma is useful when we want to derive the convergence rate of $\widehat{f}_{n,h}$ around $f_{n,h}$

instead of $f^*_{n,h}$.

Lemma 13 (Curvature Condition for Approximate Maximizer)**.** *Let $d_{n,h}(\cdot,\cdot)$ be a norm on a set containing $\mathcal{F}_n \cup \mathcal{F}^*_n$ for every fixed $h \in \mathcal{H}_m$. Let $\tau_n \in (0,1]$ be some sequence. For every $\varepsilon > 0$, let $\mathcal{H}_{m,\varepsilon} \subseteq \mathcal{H}_m$ be a subset of $\mathcal{H}_m$. We assume that there exists some sequence $\underline{\delta}_n \geq 0$ such that $\sup_{h\in\mathcal{H}_{m,\varepsilon}} d_{n,h}\left(f_{n,h}, f^*_{n,h}\right) \lesssim_\varepsilon \underline{\delta}_n$. Suppose that for every n, every $\varepsilon > 0$, and every $\delta > 0$,*

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \sup_{f\in\mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f^*_{n,h}\right) \leq \delta} M_n(f,h) - M_n\left(f^*_{n,h}, h\right) \lesssim_\varepsilon -\tau_n^2\delta^2 \tag{39}$$

and

$$d_{n,h}\left(f_{n,h}, f^*_{n,h}\right)^2 \gtrsim_\varepsilon M_n\left(f^*_{n,h}, h\right) - M_n\left(f_{n,h}, h\right) \tag{40}$$

for every $h \in \mathcal{H}_{m,\varepsilon}$. Then, for every n, every $\varepsilon > 0$, and every positive $\delta \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n$ for some constant $C_{\varepsilon,0} > 0$ depending on ε,

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \sup_{f\in\mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f_{n,h}\right) \leq \delta} M_n(f,h) - M_n\left(f_{n,h}, h\right) \lesssim_\varepsilon -\tau_n^2\delta^2. \tag{41}$$

Lemma 14 (Inequality for Rademacher Complexity)**.** *Let $\xi_1, \ldots, \xi_n$ be i.i.d. Rademacher random variables, i.e., $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$ for every $i = 1, \ldots, n$, and suppose that they are independent of the data. Let $\mathcal{U}$ be a bounded subset of $\mathbb{R}^n$ and for every $i = 1, \ldots, n$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfy $\phi_i(0) = 0$ and be Lipschitz continuous with constant L_ϕ, i.e., $|\phi_i(u) - \phi_i(v)| \leq L_\phi|u - v|$ for every $u, v \in \mathbb{R}$. Then, for $u = (u_1, \ldots, u_n)' \in \mathcal{U}$, we have*

$$\mathbb{E}\left[\sup_{u\in\mathcal{U}}\left|\sum_{i=1}^n \xi_i\phi_i(u_i)\right|\right] \leq 2L_\phi \cdot \mathbb{E}\left[\sup_{u\in\mathcal{U}}\left|\sum_{i=1}^n \xi_i u_i\right|\right].$$

Lemma 15 (Matrix Norm Inequalities, Stewart and Sun, 1990, Theorem II.3.9)**.** *Let $\|\cdot\|$ be a unitarily invariant norm such as the Frobenius norm $\|\cdot\|_F$, the spectral norm $\|\cdot\|_{\text{sp}}$, and the nuclear norm $\|\cdot\|_*$. For any matrices $A \in \mathbb{R}^{T\times K}$, $\widetilde{A} \in \mathbb{R}^{K\times T'}$, we have*

$$\|A\widetilde{A}\| \leq \|A\|_{\text{sp}} \cdot \|\widetilde{A}\| \quad \text{and} \quad \|A\widetilde{A}\| \leq \|A\| \cdot \|\widetilde{A}\|_{\text{sp}}.$$

Lemma 16. *For every matrix $A \in \mathbb{R}^{T\times K}$ with rank at most one, we have $\|A\|_* = \|A\|_F = \|A\|_{\text{sp}} = \sqrt{\text{tr}(A'A)}$. If $A = uv'$ for some vectors u and v, then $\|A\|_* = \|A\|_F = \|A\|_{\text{sp}} = \|u\|_2\|v\|_2$.*

Lemma 17. *For every function $a : \mathcal{V} \to \mathbb{R}^K$, every matrix $A \in \mathbb{R}^{T\times K}$, and every probability*

measure P on $\mathcal{V}$, we have

$$\|Aa\|_{P,2} \leq \|A\|_{\mathrm{sp}} \cdot \|a\|_{P,2}.$$

Lemma 18. *For every matrix $A \in \mathbb{R}^{T\times K}$ with $\mathrm{rank}(A) \geq 1$, the spectral norm of its Moore-Penrose inverse satisfies*

$$\|A^+\|_{\mathrm{sp}} = 1/\sigma^+_{\min}(A).$$

Lemma 19 (Bound on Effective Degrees of Freedom)**.** *Let $\mathcal{N}_A(\lambda)$ be the effective degrees of freedom defined as $\mathcal{N}_A(\lambda) = \mathrm{tr}\left(A(A+\lambda^2 I)^{-1}\right)$ for a positive semi-definite matrix $A \in \mathbb{R}^{K\times K}$ and $\lambda > 0$. Let $\widetilde{A} \in \mathbb{R}^{K\times K}$ be another positive semi-definite matrix. Then,*

$$\left|\mathcal{N}_A(\lambda) - \mathcal{N}_{\widetilde{A}}(\lambda)\right| \leq \|A - \widetilde{A}\|_*/\lambda^2.$$

Lemma 20 (Transformation on Effective Degrees of Freedom)**.** *Let $\mathcal{N}_A(\lambda)$ be the effective degrees of freedom defined as in Lemma 19. Let $Q \in \mathbb{R}^{K\times K}$ be another matrix. Then,*

$$\mathcal{N}_{QAQ'}(\lambda) \leq \max\{1, \|Q\|^2_{\mathrm{sp}}\}\mathcal{N}_A(\lambda).$$

F Proofs for Main Results

F.1 Proof for Theorem 1

Proof. By the triangle inequality, we have

$$\begin{aligned}\left\|\widehat{f}\circ\check{h} - a^*_{\mathrm{ta}}\right\|_{P_{\mathrm{ta}},2} &\leq \left\|\widehat{f}\circ\check{h} - f^\bullet_n\circ\check{h}\right\|_{P_{\mathrm{ta}},2} + \left\|f^\bullet_n\circ\check{h} - f_n\circ h_m\right\|_{P_{\mathrm{ta}},2} + \|f_n\circ h_m - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \\ &=: \mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)}. \end{aligned} \tag{42}$$

We bound each term separately.

By (13) in Assumption 1 (ii),

$$\mathrm{(I)} = O_P(r_{\mathrm{ta},n}).$$

Next, we bound the second term (II). By (12) in Assumption 1 (i),

$$\left\|\check{g}\circ\check{h} - g_m\circ h_m\right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}).$$

Since $\check{g} \in \mathcal{G}_m$, we have

$$\min_{g\in\mathcal{G}_m}\left\|g\circ\check{h} - g_m\circ h_m\right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}).$$

Hence, for every $\varepsilon > 0$, there exists a constant $M_\varepsilon < \infty$ such that

$$\liminf_{m\to\infty} P_{\mathrm{so}} \left(\min_{g\in\mathcal{G}_m} \left\| g \circ \check{h} - g_m \circ h_m \right\|_{P_{\mathrm{so},2}} \le M_\varepsilon \delta_{\mathrm{so},m} \right) \ge 1 - \varepsilon.$$

Thus, we have

$$\liminf_{m\to\infty} P_{\mathrm{so}} \left(\check{h} \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m}) \right) \ge 1 - \varepsilon,$$

with $\rho_{\mathrm{so},m} = M_\varepsilon \delta_{\mathrm{so},m}$. Since the event $\{\check{h} \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})\}$ depends only on the source sample, the above lower bound also holds under the joint law P. On this event, Assumption 2 and (11) imply that $\check{h} \in \mathcal{H}_{\mathrm{ta},m}(\rho_{\mathrm{so},m}/\nu_n)$, i.e.,

$$\min_{f\in\mathcal{F}_n^{\mathrm{sub}}} \left\| f \circ \check{h} - f_n \circ h_m \right\|_{P_{\mathrm{ta},2}} \le \rho_{\mathrm{so},m}/\nu_n = M_\varepsilon \delta_{\mathrm{so},m}/\nu_n.$$

Therefore,

$$\min_{f\in\mathcal{F}_n^{\mathrm{sub}}} \left\| f \circ \check{h} - f_n \circ h_m \right\|_{P_{\mathrm{ta},2}} = O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}/\nu_n).$$

Hence, the second term satisfies

$$\text{(II)} = O_P(\delta_{\mathrm{so},m}/\nu_n).$$

By Definition 1,

$$\text{(III)} = O(\epsilon_{\mathrm{ta},n}).$$

Combining the three bounds yields

$$\left\| \widehat{f} \circ \check{h} - a^*_{\mathrm{ta}} \right\|_{P_{\mathrm{ta},2}} = O_P\left(r_{\mathrm{ta},n} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n} \right),$$

which proves the claim. □

F.2 Proof for Theorem 2

Proof. For every $h \in \mathcal{H}_{\mathrm{so},m}(0)$, define

$$A_n(h) := \min_{f\in\mathcal{F}_n^{\mathrm{sub}}} \| f \circ h - f_n \circ h_m \|_{P_{\mathrm{ta},2}}.$$

First, consider the deterministic sequence $h_m^\dagger = h_m$. By assumption, $f_n \circ h_m = f^\bullet_{n,h_m} \circ h_m$ P_{ta}-almost surely. The assumptions and the triangle inequality then imply

$$\begin{aligned}\epsilon_{\mathrm{ta},n} &= \|f_n \circ h_m - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \\ &\leq \|f_n \circ h_m - \widehat{f}_{n,h_m} \circ h_m\|_{P_{\mathrm{ta}},2} + \|\widehat{f}_{n,h_m} \circ h_m - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \\ &= o_{P_{\mathrm{ta}}}(1).\end{aligned}$$

Since $\epsilon_{\mathrm{ta},n}$ is deterministic, this implies $\epsilon_{\mathrm{ta},n} = o(1)$.

Now fix any deterministic sequence $h_m^\dagger \in \mathcal{H}_{\mathrm{so},m}(0)$. By the triangle inequality,

$$\begin{aligned}A_n(h_m^\dagger) &= \|f^\bullet_{n,h_m^\dagger} \circ h_m^\dagger - f_n \circ h_m\|_{P_{\mathrm{ta}},2} \\ &\leq \|f^\bullet_{n,h_m^\dagger} \circ h_m^\dagger - \widehat{f}_{n,h_m^\dagger} \circ h_m^\dagger\|_{P_{\mathrm{ta}},2} \\ &\quad + \|\widehat{f}_{n,h_m^\dagger} \circ h_m^\dagger - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} + \epsilon_{\mathrm{ta},n} \\ &= o_{P_{\mathrm{ta}}}(1).\end{aligned}$$

Because $A_n(h_m^\dagger)$ is deterministic, $A_n(h_m^\dagger) = o(1)$ for every such deterministic sequence.

We claim that $\sup_{h \in \mathcal{H}_{\mathrm{so},m}(0)} A_n(h) = o(1)$. Otherwise, there would exist $\varepsilon > 0$, a subsequence $\{n_j\}$, and deterministic functions $h^\dagger_{m_{n_j}} \in \mathcal{H}_{\mathrm{so},m_{n_j}}(0)$ such that $A_{n_j}(h^\dagger_{m_{n_j}}) \geq \varepsilon$ for every j. Setting $h_m^\dagger = h_m$ outside this subsequence would produce a deterministic sequence contradicting $A_n(h_m^\dagger) = o(1)$. Finally, define $\rho_{\mathrm{ta},n} := \sup_{h \in \mathcal{H}_{\mathrm{so},m}(0)} A_n(h)$. Then $\rho_{\mathrm{ta},n} = o(1)$ and, by definition, $\mathcal{H}_{\mathrm{so},m}(0) \subseteq \mathcal{H}_{\mathrm{ta},m}(\rho_{\mathrm{ta},n})$. □

F.3 Proof for Theorem 3

Proof. Pick any $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$. Then, by the triangle inequality and $\|f_n \circ h_m(Z) - \Gamma \circ g_m \circ h_m(Z)\|_{P_{\mathrm{ta}},2} \leq \varrho_n$, we have for every $f \in \mathcal{F}_n^{\mathrm{sub}}$,

$$\begin{aligned}&\|f \circ h - f_n \circ h_m\|_{P_{\mathrm{ta}},2} \\ &\quad \leq \|f \circ h - \Gamma \circ g_m \circ h_m\|_{P_{\mathrm{ta}},2} + \|\Gamma \circ g_m \circ h_m - f_n \circ h_m\|_{P_{\mathrm{ta}},2} \\ &\quad \leq \|f \circ h - \Gamma \circ g_m \circ h_m\|_{P_{\mathrm{ta}},2} + \varrho_n \\ &\quad \leq \|f \circ h - \Gamma \circ g^\bullet_m \circ h\|_{P_{\mathrm{ta}},2} + \|\Gamma \circ g^\bullet_m \circ h - \Gamma \circ g_m \circ h_m\|_{P_{\mathrm{ta}},2} + \varrho_n.\end{aligned}$$

Taking the minimum over $f \in \mathcal{F}_n^{\mathrm{sub}}$ on both sides, we have

$$\begin{aligned}
&\min_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - f_n \circ h_m\|_{P_{\mathrm{ta}},2} \\
&\quad \le \min_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - \Gamma \circ g_m^\bullet \circ h\|_{P_{\mathrm{ta}},2} + \|\Gamma \circ g_m^\bullet \circ h - \Gamma \circ g_m \circ h_m\|_{P_{\mathrm{ta}},2} + \varrho_n \\
&\quad \le \|\Gamma \circ g_m^\bullet \circ h - \Gamma \circ g_m \circ h_m\|_{P_{\mathrm{ta}},2} + 2\varrho_n \\
&\quad \le L_\Gamma \|g_m^\bullet \circ h - g_m \circ h_m\|_{P_{\mathrm{ta}},2} + 2\varrho_n \\
&\quad \le L_\Gamma \rho_{\mathrm{so},m} / \underline{w}_n + 2\varrho_n,
\end{aligned}$$

where the second inequality follows from the existence of $f \in \mathcal{F}_n^{\mathrm{sub}}$ such that $\|f \circ h - \Gamma \circ g_m^\bullet \circ h\|_{P_{\mathrm{ta}},2} \le \varrho_n$. The third inequality follows from the Lipschitz continuity of Γ, and the last inequality follows from Assumption 3 and the definition of $\mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$. □

F.4 Proof for Corollary 1

Proof. Define the function $\Gamma : \mathbb{R}^T \to \mathbb{R}$ by

$$\Gamma(x) = f_n\left(B_{\mathrm{so},m}^+(x - \alpha_{\mathrm{so},m})\right).$$

Since $B_{\mathrm{so},m}$ has full column rank, $B_{\mathrm{so},m}^+ B_{\mathrm{so},m} = I_K$, and hence

$$\Gamma \circ g_m \circ h_m = f_n\left(B_{\mathrm{so},m}^+ B_{\mathrm{so},m} h_m\right) = f_n \circ h_m.$$

Moreover, Γ has Lipschitz constant $L_\Gamma \le L_{\mathcal{F}} \|B_{\mathrm{so},m}^+\|_{\mathrm{sp}}$ since

$$\begin{aligned}
|\Gamma(x_1) - \Gamma(x_2)| &\le L_{\mathcal{F}} \left\|B_{\mathrm{so},m}^+(x_1 - x_2)\right\|_2 \\
&\le L_{\mathcal{F}} \left\|B_{\mathrm{so},m}^+\right\|_{\mathrm{sp}} \|x_1 - x_2\|_2
\end{aligned}$$

for every $x_1, x_2 \in \mathbb{R}^T$, where the last inequality follows from Lemma 17.

Fix $\rho_{\mathrm{so},m} \ge 0$ and $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$, and let $(\alpha_{\mathrm{so},m}^\bullet, B_{\mathrm{so},m}^\bullet)$ be the minimizer in the statement of the corollary. Define $g_m^\bullet(v) = \alpha_{\mathrm{so},m}^\bullet + B_{\mathrm{so},m}^\bullet v$. By the assumption, there exists $f \in \mathcal{F}_n^{\mathrm{sub}}$ such that

$$\begin{aligned}
f \circ h &= f_n\left(B_{\mathrm{so},m}^+(\alpha_{\mathrm{so},m}^\bullet - \alpha_{\mathrm{so},m} + B_{\mathrm{so},m}^\bullet h)\right) \\
&= \Gamma \circ g_m^\bullet \circ h,
\end{aligned}$$

P_{ta}-a.s. Therefore, by Theorem 3, g_m is

$$\left(\rho_{\mathrm{so},m}, \frac{L_{\mathcal{F}}\|B_{\mathrm{so},m}^{+}\|_{\mathrm{sp}}}{\underline{w}_n}\rho_{\mathrm{so},m}\right)$$

diverse over f_n with respect to h_m. Finally, Lemma 18 gives $\|B_{\mathrm{so},m}^{+}\|_{\mathrm{sp}} = 1/\sigma_{\min}(B_{\mathrm{so},m})$ because $B_{\mathrm{so},m}$ has full column rank. □

F.5 Proof for Theorem 4

Proof. Suppose that there is no such function b. This implies that $\beta_{\mathrm{ta},n} \notin \mathrm{span}\{B'_{\mathrm{so},m}\} := S$. Thus, there exists a unit vector $v \in S^{\perp}$, where $S^{\perp}$ is the orthogonal complement of S, such that $\beta'_{\mathrm{ta},n}v \neq 0$. Define $P_v = I_K - vv'$ as the projection matrix onto the linear subspace orthogonal to v. Let $h(Z) = P_v h_m(Z) + c$, where $c \in \mathbb{R}^K$ is the vector such that $h \in \mathcal{H}_m$, and let $g(u) = \alpha_{\mathrm{so},m} - B_{\mathrm{so},m}c + B_{\mathrm{so},m}u$. The head-class assumption in the theorem ensures that $g \in \mathcal{G}_m$. Then, we have

$$\begin{aligned}
\|g \circ h - g_m \circ h_m\|_{P_{\mathrm{so}},2}^2 &= \mathbb{E}_{\mathrm{so}}\left[\|\alpha_{\mathrm{so},m} + B_{\mathrm{so},m}(P_v h_m(Z) + c - c) - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m}h_m(Z))\|_2^2\right] \\
&= \mathbb{E}_{\mathrm{so}}\left[\|B_{\mathrm{so},m}(P_v h_m(Z) - h_m(Z))\|_2^2\right] \\
&= \mathbb{E}_{\mathrm{so}}\left[\|B_{\mathrm{so},m}vv'h_m(Z)\|_2^2\right] = 0,
\end{aligned}$$

where the last equality follows from $v \in S^{\perp}$. Thus, $h \in \mathcal{H}_{\mathrm{so},m}(0)$. On the other hand, every $f \in \mathcal{F}_n^{\mathrm{sub}}$ can be written as $f(u) = \alpha + \beta' u$ for some $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^K$. Hence,

$$\begin{aligned}
&\min_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - f_n \circ h_m\|_{P_{\mathrm{ta}},2}^2 \\
&\geq \min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^K} \mathbb{E}_{\mathrm{ta}}\left[\left(\begin{pmatrix} \alpha + \beta' c - \alpha_{\mathrm{ta},n} \\ P_v\beta - \beta_{\mathrm{ta},n} \end{pmatrix}' \begin{pmatrix} 1 \\ h_m(Z) \end{pmatrix}\right)^2\right] \\
&\geq \mu_{\min}\left(\mathbb{E}_{\mathrm{ta}}\left[\begin{pmatrix} 1 \\ h_m(Z) \end{pmatrix}\begin{pmatrix} 1 \\ h_m(Z) \end{pmatrix}'\right]\right) \min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^K} \left\|\begin{pmatrix} \alpha + \beta' c - \alpha_{\mathrm{ta},n} \\ P_v\beta - \beta_{\mathrm{ta},n} \end{pmatrix}\right\|_2^2 \\
&\geq \mu_{\min}\left(\mathbb{E}_{\mathrm{ta}}\left[\begin{pmatrix} 1 \\ h_m(Z) \end{pmatrix}\begin{pmatrix} 1 \\ h_m(Z) \end{pmatrix}'\right]\right)(v'\beta_{\mathrm{ta},n})^2 > 0,
\end{aligned}$$

where the first inequality follows because the target subclass contains only linear heads, and the second inequality follows from the property of the minimum eigenvalue $\mu_{\min}(\cdot)$, and the

third inequality follows from

$$\begin{aligned}\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^K}\left\|\begin{pmatrix}\alpha+\beta' c-\alpha_{\mathrm{ta},n}\\ P_v\beta-\beta_{\mathrm{ta},n}\end{pmatrix}\right\|_2^2 &\geq \min_{\beta\in\mathbb{R}^K}\|P_v\beta-\beta_{\mathrm{ta},n}\|_2^2\\ &= \min_{\beta\in\mathbb{R}^K}\|P_v(\beta-\beta_{\mathrm{ta},n})-vv'\beta_{\mathrm{ta},n}\|_2^2\\ &= \min_{\beta\in\mathbb{R}^K}\|P_v(\beta-\beta_{\mathrm{ta},n})\|_2^2+(v'\beta_{\mathrm{ta},n})^2\\ &\geq (v'\beta_{\mathrm{ta},n})^2>0,\end{aligned}$$

and the positive definiteness of the second moment matrix of $(1, h_m(Z))$ under P_{ta}. Thus, $h\notin\mathcal{H}_{\mathrm{ta},m}(0)$. □

F.6 Proof for Theorem 5

Proof. Suppose that there is no such function Γ_0. Thus, for every measurable function $\Gamma_0:\mathbb{R}^K\to\mathbb{R}$, we have $P_{\mathrm{ta}}\left(f_n\circ h_m(Z)\neq\Gamma_0\circ h_0(Z)\right)>0$. On the other hand, by Lemma 2,

$$\begin{aligned}&\inf_{f\in\mathcal{F}_n^{\mathrm{sub}}}\|f\circ h_0-f_n\circ h_m\|_{P_{\mathrm{ta}},2}^2\\ &\quad=\mathbb{E}_{\mathrm{ta}}\left[\mathrm{Var}_{\mathrm{ta}}\left(f_n\circ h_m(Z)\mid h_0(Z)\right)\right]+\inf_{f\in\mathcal{F}_n^{\mathrm{sub}}}\mathbb{E}_{\mathrm{ta}}\left[\left(\mathbb{E}_{\mathrm{ta}}\left[f_n\circ h_m(Z)\mid h_0(Z)\right]-f\circ h_0(Z)\right)^2\right]\\ &\quad\geq\mathbb{E}_{\mathrm{ta}}\left[\mathrm{Var}_{\mathrm{ta}}\left(f_n\circ h_m(Z)\mid h_0(Z)\right)\right]\\ &\quad=\mathbb{E}_{\mathrm{ta}}\left[\left(f_n\circ h_m(Z)-\mathbb{E}_{\mathrm{ta}}\left[f_n\circ h_m(Z)\mid h_0(Z)\right]\right)^2\right].\end{aligned}$$

The last term is strictly positive since $P_{\mathrm{ta}}\left(f_n\circ h_m(Z)\neq\Gamma_0\circ h_0(Z)\right)>0$ for every measurable function Γ_0. Thus, $h_0\notin\mathcal{H}_{\mathrm{ta},m}(0)$, while we have $h_0\in\mathcal{H}_{\mathrm{so},m}(0)$ by assumption, which contradicts the $(0,0)$-diversity of g_m over f_n with respect to h_m. This proves the first claim. For the second claim, suppose that the additional sigma-field inclusion condition holds. Since $h_0(Z)$ and $g_m\circ h_m(Z)$ take values in the standard Borel spaces $\mathbb{R}^K$ and $\mathbb{R}^T$, respectively, Lemma 3, applied with $U=g_m\circ h_m(Z)$, $V=h_0(Z)$, and $P=P_{\mathrm{ta}}$, implies that there exists a measurable function $\Gamma_1:\mathbb{R}^T\to\mathbb{R}^K$ such that $h_0(Z)=\Gamma_1\circ g_m\circ h_m(Z)$, P_{ta}-a.s. The second claim then follows by defining $\Gamma:=\Gamma_0\circ\Gamma_1$. □

F.7 Proof for Theorem 6

Proof. By Theorem 1, it suffices to verify (13) with $r_{\mathrm{ta},n}$ there replaced by $r_{\mathrm{ta},n}+\delta_{\mathrm{so},m}/\nu_n+\epsilon_{\mathrm{ta},n}$, since adding the transfer and approximation terms again does not change the resulting stochastic

order. For every $h \in \mathcal{H}_m$, write

$$f^\bullet_{n,h} \in \arg\min_{f \in \mathcal{F}^{\text{sub}}_n} \|f \circ h - f_n \circ h_m\|_{P_{\text{ta}},2},$$

and retain the shorthand $f^\bullet_n = f^\bullet_{n,\check{h}}$. By definition and the inclusion $\mathcal{F}^{\text{sub}}_n \subseteq \mathcal{F}_n$, we have

$$f^\bullet_{n,h} \in \mathcal{F}^{\text{sub}}_n \subseteq \mathcal{F}_n.$$

Moreover, $\widehat{f} \in \mathcal{F}_n$ by the statement of the theorem. By (12) in Assumption 1 (i), for every $\varepsilon > 0$, there exists a constant $M_\varepsilon < \infty$ such that

$$\limsup_{m \to \infty} P_{\text{so}} \left(\check{h} \notin \mathcal{H}_{\text{so},m}(M_\varepsilon \delta_{\text{so},m}) \right) \leq \varepsilon.$$

We will apply Lemma 11 with $\mathcal{H}_{m,\varepsilon} = \mathcal{H}_{\text{so},m}(M_\varepsilon \delta_{\text{so},m})$, $\mathcal{F}_n$ as given in the theorem, $f^*_{n,h} = f_{n,h} = f^\bullet_{n,h}$, $\widehat{f}_{n,h} = \widehat{f}$, $P = P_{\text{ta}}$, $M_n(f,h) = -\mathbb{E}_{\text{ta}}[\ell_{\text{ta}}(f \circ h(Z), Y)]$, $\mathbb{M}_n(f,h) = -\frac{1}{n}\sum_{i=1}^n \ell_{\text{ta}}(f \circ h(Z_i), Y_i)$, $d_{n,h}(f,f') = \|f \circ h - f' \circ h\|_{P_{\text{ta}},2}$, and $\tau_n = 1$, where the notation on the left-hand side is that of Lemma 11.[26]

By Lemma 9, the pointwise measurability condition in Lemma 11 is satisfied. Note that $f^\bullet_{n,h}$ and $\widehat{f}$ depend on h through the minimization problem given h.

Since $f^\bullet_n \circ \check{h}$ need not equal a^*_{ta}, we cannot directly apply the curvature condition in Assumption 4 to $f^\bullet_n \circ \check{h}$ to verify (32). For this purpose, we first apply Lemma 13 with $\mathcal{H}_{m,\varepsilon} = \mathcal{H}_{\text{so},m}(M_\varepsilon \delta_{\text{so},m})$, $\mathcal{F}_n$ as given in the theorem, $f_{n,h} = f^\bullet_{n,h}$ and $f^*_{n,h} = a^*_{\text{ta}}$. We can bound

$$\begin{aligned}
&\sup_{h \in \mathcal{H}_{m,\varepsilon}} \|f^\bullet_{n,h} \circ h - a^*_{\text{ta}}\|_{P_{\text{ta}},2} \\
&\leq \sup_{h \in \mathcal{H}_{\text{ta},m}(M_\varepsilon \delta_{\text{so},m}/\nu_n)} \|f^\bullet_{n,h} \circ h - f_n \circ h_m\|_{P_{\text{ta}},2} + \|f_n \circ h_m - a^*_{\text{ta}}\|_{P_{\text{ta}},2} \\
&\leq M_\varepsilon \delta_{\text{so},m}/\nu_n + \epsilon_{\text{ta},n}, \qquad (43)
\end{aligned}$$

where the first inequality follows from the triangle inequality and the inclusion $\mathcal{H}_{m,\varepsilon} \subseteq \mathcal{H}_{\text{ta},m}(M_\varepsilon \delta_{\text{so},m}/\nu_n)$, which follows from the definition of $\mathcal{H}_{m,\varepsilon}$ and Assumption 2. The second inequality follows from $f^\bullet_{n,h} \in \arg\min_{f \in \mathcal{F}^{\text{sub}}_n} \|f \circ h - f_n \circ h_m\|_{P_{\text{ta}},2}$, the definition of $\mathcal{H}_{\text{ta},m}(\cdot)$, and Definition 1. If we set $\underline{\delta}_n = \delta_{\text{so},m}/\nu_n + \epsilon_{\text{ta},n}$, we obtain the conditions in Lemma 13 with

[26] To apply Lemma 13 below, extend these definitions to the true target score a^*_{ta} by setting, for every $f \in \mathcal{F}_n$ and $h \in \mathcal{H}_m$, $M_n(a^*_{\text{ta}}, h) = -\mathbb{E}_{\text{ta}}[\ell_{\text{ta}}(a^*_{\text{ta}}(Z), Y)]$, $\mathbb{M}_n(a^*_{\text{ta}}, h) = -\frac{1}{n}\sum_{i=1}^n \ell_{\text{ta}}(a^*_{\text{ta}}(Z_i), Y_i)$, $d_{n,h}(f, a^*_{\text{ta}}) = d_{n,h}(a^*_{\text{ta}}, f) = \|f \circ h - a^*_{\text{ta}}\|_{P_{\text{ta}},2}$, and $d_{n,h}(a^*_{\text{ta}}, a^*_{\text{ta}}) = 0$.

$\tau_n = 1$ by the curvature condition in Assumption 4. Thus, Lemma 13 implies that there exists a constant $C_{\varepsilon,0} > 0$ such that for every $\delta > C_{\varepsilon,0}(\delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n})$, (32) in Lemma 11 holds with $f_{n,h} = f^*_{n,h} = f^\bullet_{n,h}$ and $\tau_n = 1$.

Next, we verify (33). By Lemma 8 with $P = P_{\mathrm{ta}}$, $\ell = \ell_{\mathrm{ta}}$, $C_\ell = C_{\ell,\mathrm{ta}}$, and $M = M_F$,

$$\sup_{h\in\mathcal{H}_m} \mathbb{E}\left[\sup_{f\in\mathcal{F}_n: d_{n,h}\left(f, f^*_{n,h}\right)<\delta} \sqrt{n}\left|\left(\mathbb{M}_n - M_n\right)(f,h) - \left(\mathbb{M}_n - M_n\right)\left(f^*_{n,h}, h\right)\right|\right]$$
$$\lesssim C_{\ell,\mathrm{ta}} M_F \left(J\left(\delta/M_F, \mathcal{F}_n | M_F, L_2\right) + \frac{M_F^2 J^2\left(\delta/M_F, \mathcal{F}_n | M_F, L_2\right)}{\delta^2\sqrt{n}}\right).$$

By the VC dimension assumption and Lemma 10, we have

$$J\left(\delta/M_F, \mathcal{F}_n | M_F, L_2\right) \lesssim \int_0^{\delta/M_F} \sqrt{\mathsf{V}(\mathcal{F}_n)\log(1/\epsilon)}d\epsilon$$
$$\lesssim \delta M_F^{-1}\sqrt{\mathsf{V}(\mathcal{F}_n)\log(M_F/\delta)}.$$

Thus,

$$\sup_{h\in\mathcal{H}_m} \mathbb{E}\left[\sup_{f\in\mathcal{F}_n: d_{n,h}\left(f, f^*_{n,h}\right)<\delta} \sqrt{n}\left|\left(\mathbb{M}_n - M_n\right)(f,h) - \left(\mathbb{M}_n - M_n\right)\left(f^*_{n,h}, h\right)\right|\right]$$
$$\lesssim C_{\ell,\mathrm{ta}}\left(\delta\sqrt{\mathsf{V}(\mathcal{F}_n)\log(M_F/\delta)} + \frac{M_F\mathsf{V}(\mathcal{F}_n)\log(M_F/\delta)}{\sqrt{n}}\right).$$

Thus, by setting $\phi_n(\delta) = \delta\sqrt{\mathsf{V}(\mathcal{F}_n)\log(M_F/\delta)} + M_F\mathsf{V}(\mathcal{F}_n)\log(M_F/\delta)/\sqrt{n}$, we can verify (33).[27] Finally, the condition $\phi_n(\delta) \lesssim \sqrt{n}\delta^2$ holds when $\delta \gtrsim \sqrt{(\mathsf{V}(\mathcal{F}_n)\log(n))/n}$. By Lemma 11, we obtain the following convergence rate using the condition (34):

$$\left\|\widehat{f}\circ\breve{h} - f^\bullet_n\circ\breve{h}\right\|_{P_{\mathrm{ta}},2} = O_P\left(\sqrt{\frac{\mathsf{V}(\mathcal{F}_n)\log(n)}{n}} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}\right). \tag{44}$$

Thus, Assumption 1 (ii) holds with rate $r_{\mathrm{ta},n} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}$ rather than $r_{\mathrm{ta},n}$ alone. Applying Theorem 1 yields (17), since adding the transfer and approximation terms once more does not change the resulting stochastic order. This completes the proof. □

[27] The function $\phi_n(\delta)$ is not increasing around $\delta = 0$. However, we can replace $\phi_n(\delta)$ with an increasing function without affecting the convergence rate in Lemma 11. See the footnote on page 433 of van der Vaart and Wellner (2023) for details.

F.8 Proof for Theorem 7

Proof. As in the proof of Theorem 6, we will apply Lemma 11, but here the argument is unconditional on $\mathcal{I}$. Thus, we set $f_m^* = f_m = g_m \circ h_m$, $\widehat{f}_m = \breve{g} \circ \breve{h}$, and $P = P_{\text{so}}$. We define the objective functions for joint optimization over g and h, rather than optimization over f for fixed h. For every $a, \widetilde{a} \in (\mathcal{G}_m \circ \mathcal{H}_m) \cup \{a_{\text{so}}^*\}$, define $M_m(a) = -\mathbb{E}_{\text{so}}[\ell_{\text{so}}(a(Z), S)]$, $\mathbb{M}_m(a) = -\frac{1}{m}\sum_{j=1}^m \ell_{\text{so}}(a(Z_j), S_j)$, and $d_m(a, \widetilde{a}) = \|a - \widetilde{a}\|_{P_{\text{so}},2}$.

The curvature condition (32) is verified as in the proof of Theorem 6, by applying Lemma 13 with $f_m = g_m \circ h_m$, $f_m^* = a_{\text{so}}^*$, and $\underline{\delta}_m = \epsilon_{\text{so},m}$. That is, there exists a constant $C > 0$ such that for every $\delta > \tau_m^{-1} C \epsilon_{\text{so},m}$, (32) holds.

The constant function M_G is an envelope for the composite function class $\mathcal{G}_m \circ \mathcal{H}_m$. For every discrete probability measure Q on $\mathcal{Z}$ and $h \in \mathcal{H}_m$, let Q_h denote the image of Q under h. By the uniform Lipschitz condition in Assumption 5 (iii), the uniform entropy integral satisfies

$$
\begin{aligned}
&J\left(\delta, \mathcal{G}_m \circ \mathcal{H}_m | M_G, L_2\right) \\
&\quad \leq \int_0^\delta \sup_{Q \in \mathcal{P}_{fd}} \sqrt{1 + \log N\left(\varepsilon M_G, \mathcal{G}_m \circ \mathcal{H}_m, \|\cdot\|_{Q,2}\right)} d\varepsilon \\
&\quad \leq \int_0^\delta \sup_{Q \in \mathcal{P}_{fd}} \sqrt{1 + \sum_{k=1}^K \log N\left(\tfrac{\varepsilon M_G}{2 L_{\mathcal{G},m} K}, \mathcal{H}_{m,k}, \|\cdot\|_{Q,2}\right) + \sup_{h \in \mathcal{H}_m} \sum_{t=1}^T \log N\left(\tfrac{\varepsilon M_G}{2T}, \mathcal{G}_{m,t}, \|\cdot\|_{Q_h,2}\right)} d\varepsilon \\
&\quad \lesssim \int_0^\delta \sqrt{\sum_{k=1}^K \mathsf{V}(\mathcal{H}_{m,k}) \log\left(\frac{e L_{\mathcal{G},m} K}{\varepsilon}\right) + \sum_{t=1}^T \mathsf{V}(\mathcal{G}_{m,t}) \log\left(\frac{eT}{\varepsilon}\right)} d\varepsilon \\
&\quad \lesssim \delta \sqrt{\sum_{k=1}^K \mathsf{V}(\mathcal{H}_{m,k}) \log\left(\frac{L_{\mathcal{G},m} K}{\delta}\right) + \sum_{t=1}^T \mathsf{V}(\mathcal{G}_{m,t}) \log\left(\frac{T}{\delta}\right)},
\end{aligned}
$$

where the first inequality follows from Lemmas 6 and 7. The second inequality follows from Lemma 10, and the last inequality follows from calculations similar to those in the proof for Theorem 6. Using the preceding entropy bound, we apply Lemma 8 with $\mathcal{F}_n = \mathcal{G}_m \circ \mathcal{H}_m$, $C_\ell = C_{\ell,\text{so}}$, and $M = M_G$ to verify (33) with[28]

$$
\begin{aligned}
\phi_m(\delta) = \delta \sqrt{\sum_{k=1}^K \mathsf{V}(\mathcal{H}_{m,k}) \log\left(\frac{L_{\mathcal{G},m} K}{\delta}\right) + \sum_{t=1}^T \mathsf{V}(\mathcal{G}_{m,t}) \log\left(\frac{T}{\delta}\right)} \\
+ \frac{\sum_{k=1}^K \mathsf{V}(\mathcal{H}_{m,k}) \log\left(L_{\mathcal{G},m} K/\delta\right) + \sum_{t=1}^T \mathsf{V}(\mathcal{G}_{m,t}) \log\left(T/\delta\right)}{\sqrt{m}}.
\end{aligned}
$$

[28] See Footnote 27.

Hence, the condition $\phi_m(\delta/\tau_m) \lesssim \sqrt{m}\delta^2$ holds when

$$\delta \gtrsim \sqrt{\frac{\sum_{k=1}^{K} \mathsf{V}(\mathcal{H}_{m,k}) \log\left(L_{\mathcal{G},m} K m\right) + \sum_{t=1}^{T} \mathsf{V}(\mathcal{G}_{m,t}) \log\left(Tm\right)}{\tau_m^2 m}} = r_{\mathrm{so},m}.$$

By Lemma 11, we obtain the following convergence rate using the condition (34):

$$\left\| \breve{g} \circ \breve{h} - g_m \circ h_m \right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}\left(r_{\mathrm{so},m}/\tau_m + \epsilon_{\mathrm{so},m}/\tau_m\right). \tag{45}$$

Since $\|g_m \circ h_m - a^*_{\mathrm{so}}\|_{P_{\mathrm{so}},2} \le \epsilon_{\mathrm{so},m}$ and $\tau_m \le 1$, the triangle inequality gives (22). This completes the proof. □

F.9 Proof for Lemma 1

Proof. By the definitions of $\breve{h}^{\mathrm{proj}}$, $\breve{q}$, and $\breve{Q}$,

$$\begin{aligned}
\left\| \breve{h}^{\mathrm{proj}} - (\breve{q} + \breve{Q} h_m) \right\|_{P_{\mathrm{so}},2} &= \left\| \breve{B}^{+}_{\mathrm{so}} \left(\breve{\alpha}_{\mathrm{so}} + \breve{B}_{\mathrm{so}} \breve{h} - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m) \right) \right\|_{P_{\mathrm{so}},2} \\
&\le \|\breve{B}^{+}_{\mathrm{so}}\|_{\mathrm{sp}} \left\| \breve{\alpha}_{\mathrm{so}} + \breve{B}_{\mathrm{so}} \breve{h} - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m) \right\|_{P_{\mathrm{so}},2}.
\end{aligned} \tag{46}$$

The inequality follows from Lemma 17. By Assumption 1 (i), the second factor on the right-hand side of (46) is $O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m})$. Combining this rate with $\|\breve{B}^{+}_{\mathrm{so}}\|_{\mathrm{sp}} = O_{P_{\mathrm{so}}}(\underline{\sigma}_m^{-1})$ proves the stated result. Finally, on the event in the sufficient condition, Lemma 18 gives $\|\breve{B}^{+}_{\mathrm{so}}\|_{\mathrm{sp}} \le \underline{\sigma}_m^{-1}$, which verifies the condition in (24). □

F.10 Proof for Theorem 8

Proof. For each $h \in \mathcal{H}_m^{\mathrm{proj}}$, select a source minimizer

$$(\alpha^{\bullet}_{\mathrm{so},m,h}, B^{\bullet}_{\mathrm{so},m,h}) \in \operatorname*{arg\,min}_{(\alpha,B)\in\Theta_{\mathrm{so},m}} \|\alpha + Bh - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m)\|_{P_{\mathrm{so}},2}.$$

For every $h \in \mathcal{H}_m^{\mathrm{proj}}$, define the target-head parameters through the selected source minimizer by

$$\begin{aligned}
\alpha^{\bullet}_{\mathrm{ta},n,h} &:= \alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} B^{+}_{\mathrm{so},m} (\alpha^{\bullet}_{\mathrm{so},m,h} - \alpha_{\mathrm{so},m}), \\
\beta^{\bullet\prime}_{\mathrm{ta},n,h} &:= \beta'_{\mathrm{ta},n} B^{+}_{\mathrm{so},m} B^{\bullet}_{\mathrm{so},m,h}, \\
f^{\bullet}_{n,h}(v) &:= \alpha^{\bullet}_{\mathrm{ta},n,h} + \beta^{\bullet\prime}_{\mathrm{ta},n,h} v.
\end{aligned}$$

By definition, $(\alpha^{\bullet}_{\mathrm{so},m,h}, B^{\bullet}_{\mathrm{so},m,h}) \in \Theta_{\mathrm{so},m}$. Indeed, Assumption 6 implies, uniformly over $h \in \mathcal{H}^{\mathrm{proj}}_m$,

$$
\begin{aligned}
|\alpha^{\bullet}_{\mathrm{ta},n,h}| &\le |\alpha_{\mathrm{ta},n}| + \|b_n' B_{\mathrm{so},m} B^{+}_{\mathrm{so},m}\|_2 \|\alpha^{\bullet}_{\mathrm{so},m,h} - \alpha_{\mathrm{so},m}\|_2 \le |\alpha_{\mathrm{ta},n}| + 2C_b C_{\alpha,\mathrm{so}} \le C_\alpha, \\
\|\beta^{\bullet}_{\mathrm{ta},n,h}\|_2 &\le \|b_n' B_{\mathrm{so},m} B^{+}_{\mathrm{so},m}\|_2 \|B^{\bullet}_{\mathrm{so},m,h}\|_{\mathrm{sp}} \le C_b \overline{\sigma} \le C_\beta.
\end{aligned} \tag{47}
$$

Hence, $f^{\bullet}_{n,h} \in \mathcal{F}_n$ for every $h \in \mathcal{H}^{\mathrm{proj}}_m$. For each fixed $h \in \mathcal{H}^{\mathrm{proj}}_m$, also define the estimator family

$$
\widehat{f}_{n,h} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{ta}}(f \circ h(Z_i), Y_i) + \lambda_n^2 \|\beta\|_2^2 \right\},
$$

such that $\widehat{f}_{n,\breve{h}^{\mathrm{proj}}} = \widehat{f}^{\mathrm{ridge}}_n$, and let $(\widehat{\alpha}_{\mathrm{ta},n,h}, \widehat{\beta}_{\mathrm{ta},n,h})$ denote the corresponding parameters.

As in (42), we decompose the estimation error of the target model into three terms.

$$
\begin{aligned}
&\|\widehat{f}^{\mathrm{ridge}}_n \circ \breve{h}^{\mathrm{proj}} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \\
&\quad \le \|\widehat{f}^{\mathrm{ridge}}_n \circ \breve{h}^{\mathrm{proj}} - f^{\bullet}_{n,\breve{h}^{\mathrm{proj}}} \circ \breve{h}^{\mathrm{proj}}\|_{P_{\mathrm{ta}},2} \\
&\qquad + \|f^{\bullet}_{n,\breve{h}^{\mathrm{proj}}} \circ \breve{h}^{\mathrm{proj}} - f_n \circ h_m\|_{P_{\mathrm{ta}},2} + \|f_n \circ h_m - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \\
&\quad =: \mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)}.
\end{aligned}
$$

By Definition 1, $\mathrm{(III)} = \epsilon_{\mathrm{ta},n}$. To bound (II), first observe that

$$
\beta'_{\mathrm{ta},n} B^{+}_{\mathrm{so},m} = b_n' B_{\mathrm{so},m} B^{+}_{\mathrm{so},m}, \qquad \|\beta'_{\mathrm{ta},n} B^{+}_{\mathrm{so},m}\|_2 \le C_b.
$$

The same identity implies $\beta'_{\mathrm{ta},n} B^{+}_{\mathrm{so},m} B_{\mathrm{so},m} = \beta'_{\mathrm{ta},n}$. Hence, for every $h \in \mathcal{H}^{\mathrm{proj}}_m$,

$$
\begin{aligned}
&\|\alpha^{\bullet}_{\mathrm{ta},n,h} + \beta^{\bullet\prime}_{\mathrm{ta},n,h} h - (\alpha_{\mathrm{ta},n} + \beta'_{\mathrm{ta},n} h_m)\|_{P_{\mathrm{ta}},2} \\
&\quad = \|\beta'_{\mathrm{ta},n} B^{+}_{\mathrm{so},m} (\alpha^{\bullet}_{\mathrm{so},m,h} + B^{\bullet}_{\mathrm{so},m,h} h - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m))\|_{P_{\mathrm{ta}},2} \\
&\quad \le \frac{C_b}{\underline{w}_n} \|\alpha^{\bullet}_{\mathrm{so},m,h} + B^{\bullet}_{\mathrm{so},m,h} h - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m)\|_{P_{\mathrm{so}},2},
\end{aligned}
$$

where the inequality follows from Assumption 3 and Lemma 17. Moreover,

$$
\breve{B}_{\mathrm{so}} \breve{h}^{\mathrm{proj}} = \breve{B}_{\mathrm{so}} \breve{B}^{+}_{\mathrm{so}} \breve{B}_{\mathrm{so}} \breve{h} = \breve{B}_{\mathrm{so}} \breve{h}.
$$

By Assumption 6 (i), $(\breve{\alpha}_{\mathrm{so}}, \breve{B}_{\mathrm{so}}) \in \Theta_{\mathrm{so},m}$ on the event under consideration and is therefore feasible in the source minimization defining $(\alpha^{\bullet}_{\mathrm{so},m,\breve{h}^{\mathrm{proj}}}, B^{\bullet}_{\mathrm{so},m,\breve{h}^{\mathrm{proj}}})$. Consequently, Assumption 1

(i) gives

$$\begin{aligned}\|g_{\alpha^\bullet_{\mathrm{so},m,\check{h}^{\mathrm{proj}}},B^\bullet_{\mathrm{so},m,\check{h}^{\mathrm{proj}}}} \circ \check{h}^{\mathrm{proj}} - g_m \circ h_m\|_{P_{\mathrm{so}},2} &\le \|\check{g} \circ \check{h}^{\mathrm{proj}} - g_m \circ h_m\|_{P_{\mathrm{so}},2} \\ &= \|\check{g} \circ \check{h} - g_m \circ h_m\|_{P_{\mathrm{so}},2} \\ &= O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}). \end{aligned} \tag{48}$$

It follows that (II) $= O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}/\underline{w}_n)$.

It remains to bound (I) uniformly over a high-probability source-side set. For each $h \in \mathcal{H}_m^{\mathrm{proj}}$, select

$$(q_h, Q_h) \in \operatorname*{arg\,min}_{\substack{q\in\mathbb{R}^K,\ Q\in\mathbb{R}^{K\times K}:\\ \|Q\|_{\mathrm{sp}}\le\overline{\sigma}/\underline{\sigma}}} \|h - (q + Qh_m)\|_{P_{\mathrm{so}},2}.$$

The source parameter restriction implies $\|\check{Q}\|_{\mathrm{sp}} \le \overline{\sigma}/\underline{\sigma}$, so $(\check{q}, \check{Q})$ is feasible. It also implies $\|\check{B}^+_{\mathrm{so}}\|_{\mathrm{sp}} \le \underline{\sigma}^{-1}$. The minimizing property of $(q_{\check{h}^{\mathrm{proj}}}, Q_{\check{h}^{\mathrm{proj}}})$ and Lemma 1, applied with $\underline{\sigma}_m = \underline{\sigma}$, give

$$\begin{aligned}\|\check{h}^{\mathrm{proj}} - q_{\check{h}^{\mathrm{proj}}} - Q_{\check{h}^{\mathrm{proj}}} h_m\|_{P_{\mathrm{so}},2} &\le \|\check{h}^{\mathrm{proj}} - (\check{q} + \check{Q}h_m)\|_{P_{\mathrm{so}},2} \\ &= O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}/\underline{\sigma}) \\ &= O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}), \end{aligned} \tag{49}$$

where the last equality uses that $\underline{\sigma}$ is fixed. For every $\varepsilon > 0$, (48) and (49) therefore give a constant $M_\varepsilon < \infty$ such that

$$\limsup_{m\to\infty} P_{\mathrm{so}}(\check{h}^{\mathrm{proj}} \notin \mathcal{H}_{m,\varepsilon}) \le \varepsilon,$$

where

$$\begin{aligned}\mathcal{H}_{m,\varepsilon} := \big\{h \in \mathcal{H}_m^{\mathrm{proj}} : &\|g_{\alpha^\bullet_{\mathrm{so},m,h},B^\bullet_{\mathrm{so},m,h}} \circ h - g_m \circ h_m\|_{P_{\mathrm{so}},2} \le M_\varepsilon \delta_{\mathrm{so},m}, \\ &\|h - (q_h + Q_h h_m)\|_{P_{\mathrm{so}},2} \le M_\varepsilon \delta_{\mathrm{so},m}\big\}. \end{aligned} \tag{50}$$

For every $h \in \mathcal{H}_{m,\varepsilon}$, the selected source head belongs to $\mathcal{G}_m$. Therefore, the first restriction defining $\mathcal{H}_{m,\varepsilon}$ implies

$$\min_{g\in\mathcal{G}_m} \|g \circ h - g_m \circ h_m\|_{P_{\mathrm{so}},2} \le M_\varepsilon \delta_{\mathrm{so},m}.$$

Consequently, $\mathcal{H}_{m,\varepsilon} \subseteq \mathcal{H}^{\mathrm{proj}}_{\mathrm{so},m}(M_\varepsilon \delta_{\mathrm{so},m})$.

For each fixed h, the finite-dimensional parameterization of $\mathcal{F}_n$ is pointwise continuous,

and every subset of Θ_n is separable. Hence, the localized classes required by Lemma 12 are pointwise measurable. We apply Lemma 12 with $\mathcal{H}_m = \mathcal{H}_m^{\mathrm{proj}}$, $\breve{h} = \breve{h}^{\mathrm{proj}}$, $\mathcal{F}_n^* = \mathcal{F}_n$, $f_{n,h} = f_{n,h}^* = f_{n,h}^\bullet$, $\tau_n = 1$, the estimator family $\widehat{f}_{n,h}$ defined above, $\mathcal{J}_n(f_{\alpha,\beta}) = \|\beta\|_2$, and $\widehat{\lambda}_{n,h} = \lambda_n$.[29] Set $M_n(f,h) := -\mathbb{E}_{\mathrm{ta}}[\ell_{\mathrm{ta}}(f \circ h(Z), Y)]$, $\mathbb{M}_n(f,h) := -\frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{ta}}(f \circ h(Z_i), Y_i)$, and $d_{n,h}(f,f') := \|f \circ h - f' \circ h\|_{P_{\mathrm{ta}},2}$. We next verify curvature directly around $f_{n,h}^\bullet$. The transported-head bound above and Definition 1 give

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \|f_{n,h}^\bullet \circ h - a_{\mathrm{ta}}^*\|_{P_{\mathrm{ta}},2} \le C_b M_\varepsilon \frac{\delta_{\mathrm{so},m}}{\underline{w}_n} + \epsilon_{\mathrm{ta},n}.$$

Set

$$\underline{\delta}_n := \frac{\delta_{\mathrm{so},m}}{\underline{w}_n} + \epsilon_{\mathrm{ta},n}.$$

For $\delta \ge C_{\varepsilon,0}\underline{\delta}_n$ with a sufficiently large $C_{\varepsilon,0}$ and any $f \in \mathcal{F}_n$ satisfying

$$\delta/2 < \|f \circ h - f_{n,h}^\bullet \circ h\|_{P_{\mathrm{ta}},2} \le \delta,$$

the reverse triangle inequality gives $\|f \circ h - a_{\mathrm{ta}}^*\|_{P_{\mathrm{ta}},2} \ge \delta/4$. Let $c_L, c_U > 0$ denote uniform constants in the lower and upper excess-risk bounds in the first condition of Assumption 4. Then,

$$\begin{aligned} M_n(f,h) - M_n(f_{n,h}^\bullet, h) &\le -c_L \|f \circ h - a_{\mathrm{ta}}^*\|_{P_{\mathrm{ta}},2}^2 + c_U \|f_{n,h}^\bullet \circ h - a_{\mathrm{ta}}^*\|_{P_{\mathrm{ta}},2}^2 \\ &\lesssim_\varepsilon -\delta^2, \end{aligned}$$

where the last inequality follows by increasing $C_{\varepsilon,0}$ if necessary. This proves (35).

The sample ridge estimator satisfies

$$\mathbb{M}_n(\widehat{f}_{n,h}, h) - \lambda_n^2 \mathcal{J}_n^2(\widehat{f}_{n,h}) \ge \mathbb{M}_n(f_{n,h}^\bullet, h) - \lambda_n^2 \mathcal{J}_n^2(f_{n,h}^\bullet),$$

so (38) holds with zero approximation error.

We next verify the empirical-process modulus. For each h, let

$$A_{h,\lambda,n} := \Sigma_{h,n} + \lambda_n^2 I_K.$$

Let $\xi_1, \dots, \xi_n$ be i.i.d. Rademacher random variables independent of the target and source

[29] The event in Assumption 6 (i) has probability approaching one. For the formal application of Lemma 12, redefine $\breve{h}^{\mathrm{proj}}$ on the complement of this event as $B_{\mathrm{so},m}^+ B_{\mathrm{so},m} h_m \in \mathcal{H}_m^{\mathrm{proj}}$ and define the corresponding ridge estimator by the same criterion. These modified objects agree with the original objects with probability approaching one, so this extension does not affect the claimed stochastic order.

samples, and define

$$\mathfrak{R}_n(a) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i a(Z_i)$$

for every measurable function a. For $f_{\alpha,\beta} \in \mathcal{F}_n$, write $\Delta\alpha := \alpha - \alpha^{\bullet}_{\mathrm{ta},n,h}$ and $\Delta\beta := \beta - \beta^{\bullet}_{\mathrm{ta},n,h}$. For each fixed i and h, the loss-increment map evaluated between $f_{\alpha,\beta} \circ h(Z_i)$ and $f^{\bullet}_{n,h} \circ h(Z_i)$ vanishes at zero and is $C_{\ell,\mathrm{ta}}$-Lipschitz by the Lipschitz condition in Assumption 4. Symmetrization and Lemma 14, applied before enlarging the parameter set, yield

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}_{\mathrm{ta}} \left[\sup_{\substack{f\in\mathcal{F}_n: d_{n,h}(f,f^{\bullet}_{n,h})<\delta \\ \mathcal{J}_n(f)<\delta/\lambda_n}} \sqrt{n} \left| (\mathbb{M}_n - M_n)(f,h) - (\mathbb{M}_n - M_n)(f^{\bullet}_{n,h},h) \right| \right] \\
&\quad \le 4C_{\ell,\mathrm{ta}} \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}_{\xi,\mathrm{ta}} \left[\sup_{\substack{f_{\alpha,\beta}\in\mathcal{F}_n: \|\Delta\alpha+\Delta\beta' h\|_{P_{\mathrm{ta}},2}<\delta \\ \lambda_n\|\beta\|_2<\delta}} |\mathfrak{R}_n(\Delta\alpha + \Delta\beta' h)| \right] \\
&\quad \le 4C_{\ell,\mathrm{ta}} \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}_{\xi,\mathrm{ta}} \left[\sup_{\substack{(\Delta\alpha,\Delta\beta): \|\Delta\alpha+\Delta\beta' h\|_{P_{\mathrm{ta}},2}<\delta \\ \lambda_n\|\Delta\beta+\beta^{\bullet}_{\mathrm{ta},n,h}\|_2<\delta}} |\mathfrak{R}_n(\Delta\alpha + \Delta\beta' h)| \right].
\end{aligned} \tag{51}$$

The last inequality is the only step that enlarges the compact coefficient class. For every increment in the supremum,

$$\mathfrak{R}_n(\Delta\alpha + \Delta\beta' h) = \mathbb{E}_{\mathrm{ta}}[\Delta\alpha + \Delta\beta' h]\mathfrak{R}_n(1) + \Delta\beta'\mathfrak{R}_n(h - \mathbb{E}_{\mathrm{ta}}[h]).$$

The first coefficient is bounded by δ by the Cauchy-Schwarz inequality, and $\mathbb{E}_{\xi,\mathrm{ta}}[|\mathfrak{R}_n(1)|] \le 1$. For the second term,

$$|\Delta\beta'\mathfrak{R}_n(h - \mathbb{E}_{\mathrm{ta}}[h])| \le \|A^{1/2}_{h,\lambda,n}\Delta\beta\|_2 \|A^{-1/2}_{h,\lambda,n}\mathfrak{R}_n(h - \mathbb{E}_{\mathrm{ta}}[h])\|_2.$$

The variance identity gives

$$\Delta\beta'\Sigma_{h,n}\Delta\beta = \mathrm{Var}_{\mathrm{ta}}(\Delta\beta' h(Z)) \le \|\Delta\alpha + \Delta\beta' h\|^2_{P_{\mathrm{ta}},2} < \delta^2.$$

In addition, (47) implies

$$\lambda_n\|\Delta\beta\|_2 \le \lambda_n\|\Delta\beta + \beta^{\bullet}_{\mathrm{ta},n,h}\|_2 + \lambda_n\|\beta^{\bullet}_{\mathrm{ta},n,h}\|_2 < \delta + C_\beta\lambda_n.$$

Consequently,

$$\|A_{h,\lambda,n}^{1/2}\Delta\beta\|_2 \lesssim \delta + \lambda_n.$$

Moreover, Jensen's inequality and the Rademacher second-moment identity imply

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathrm{ta}}\left[\|A_{h,\lambda,n}^{-1/2}\mathfrak{R}_n(h - \mathbb{E}_{\mathrm{ta}}[h])\|_2\right] \\
&\quad \le \operatorname{tr}\left(A_{h,\lambda,n}^{-1/2}\Sigma_{h,n}A_{h,\lambda,n}^{-1/2}\right)^{1/2} \\
&\quad = \mathcal{N}_{h,n}(\lambda_n)^{1/2}.
\end{aligned}$$

It remains to control this effective dimension uniformly over $h \in \mathcal{H}_{m,\varepsilon}$.

Define $r_h := h - q_h - Q_h h_m$ and $u_h := r_h - \mathbb{E}_{\mathrm{ta}}[r_h]$. Conditional on the source sample, q_h and Q_h are constant under P_{ta}, and

$$\begin{aligned}
\Sigma_{h,n} &= Q_h\Sigma_{h_m,n}Q_h' + \Delta_{h,n}, \\
\Delta_{h,n} &= \mathbb{E}_{\mathrm{ta}}[u_h u_h'] + Q_h\mathbb{E}_{\mathrm{ta}}[(h_m - \mathbb{E}_{\mathrm{ta}}[h_m])u_h'] + \mathbb{E}_{\mathrm{ta}}[u_h(h_m - \mathbb{E}_{\mathrm{ta}}[h_m])']Q_h'.
\end{aligned}$$

By Assumption 3 and (50),

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \|r_h\|_{P_{\mathrm{ta}},2} \le M_\varepsilon \frac{\delta_{\mathrm{so},m}}{\underline{w}_n}.$$

Because $\mathbb{E}_{\mathrm{ta}}\|u_h\|_2^2 \le \mathbb{E}_{\mathrm{ta}}\|r_h\|_2^2$, $\|Q_h\|_{\mathrm{sp}} \le \overline{\sigma}/\underline{\sigma}$, and $\|h_m - \mathbb{E}_{\mathrm{ta}}[h_m]\|_{P_{\mathrm{ta}},2} = \operatorname{tr}(\Sigma_{h_m,n})^{1/2}$, the nuclear-norm bounds in Lemmas 15 and 16 and the Cauchy-Schwarz inequality give

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \|\Delta_{h,n}\|_* \lesssim_\varepsilon \left(\frac{\delta_{\mathrm{so},m}}{\underline{w}_n}\right)^2 + \operatorname{tr}(\Sigma_{h_m,n})^{1/2}\frac{\delta_{\mathrm{so},m}}{\underline{w}_n}.$$

Applying Lemmas 19 and 20 and absorbing the constants depending on $\overline{\sigma}, \underline{\sigma}$ now yields

$$\begin{aligned}
\sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathcal{N}_{h,n}(\lambda_n) &\lesssim_\varepsilon \mathcal{N}_{h_m,n}(\lambda_n) + \frac{(\delta_{\mathrm{so},m}/\underline{w}_n)^2 + \operatorname{tr}(\Sigma_{h_m,n})^{1/2}\delta_{\mathrm{so},m}/\underline{w}_n}{\lambda_n^2}, \\
\sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathcal{N}_{h,n}(\lambda_n)^{1/2} &\lesssim_\varepsilon \mathcal{N}_{h_m,n}(\lambda_n)^{1/2} + \frac{\delta_{\mathrm{so},m}/\underline{w}_n + (\delta_{\mathrm{so},m}/\underline{w}_n)^{1/2}\operatorname{tr}(\Sigma_{h_m,n})^{1/4}}{\lambda_n}.
\end{aligned}$$

Combining these bounds with (51) verifies (36) with

$$\phi_n(\delta) := (\delta + \lambda_n)\left(1 + \mathcal{N}_{h_m,n}(\lambda_n)^{1/2} + \frac{\delta_{\mathrm{so},m}/\underline{w}_n + (\delta_{\mathrm{so},m}/\underline{w}_n)^{1/2}\operatorname{tr}(\Sigma_{h_m,n})^{1/4}}{\lambda_n}\right).$$

This function is increasing, and $\phi_n(\delta)/\delta$ is decreasing, so the modulus requirement in Lemma 12

holds with $\xi = 1 < 2$. Choose δ_n to be a sufficiently large constant multiple of

$$\sqrt{\frac{1+\mathcal{N}_{h_m,n}(\lambda_n)}{n}} + \lambda_n + \frac{\delta_{\mathrm{so},m}/\underline{w}_n + (\delta_{\mathrm{so},m}/\underline{w}_n)^{1/2}\operatorname{tr}(\Sigma_{h_m,n})^{1/4}}{\sqrt{n}\lambda_n} + \frac{\delta_{\mathrm{so},m}}{\underline{w}_n} + \epsilon_{\mathrm{ta},n}.$$

Choose the multiplicative constant large enough that $C_\beta\lambda_n < \delta_n$. Then $\phi_n(\delta_n) \lesssim_\varepsilon \sqrt{n}\delta_n^2$, $\delta_n \gtrsim_\varepsilon \underline{\delta}_n$, and

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \lambda_n \mathcal{J}_n(f^\bullet_{n,h}) \le C_\beta\lambda_n < \delta_n.$$

Thus, the remaining requirements in (37) hold, and Lemma 12 gives

$$\begin{aligned}
\text{(I)} &= \|\widehat{f}_{n,\breve{h}^{\mathrm{proj}}} \circ \breve{h}^{\mathrm{proj}} - f^\bullet_{n,\breve{h}^{\mathrm{proj}}} \circ \breve{h}^{\mathrm{proj}}\|_{P_{\mathrm{ta},2}} \\
&= O_P\left(\delta_n + \lambda_n \mathcal{J}_n(f^\bullet_{n,\breve{h}^{\mathrm{proj}}})\right) \\
&= O_P(\delta_n).
\end{aligned}$$

Combining this result with $\text{(II)} = O_P(\delta_{\mathrm{so},m}/\underline{w}_n)$ and $\text{(III)} = \epsilon_{\mathrm{ta},n}$ proves the theorem. □

G Proofs for Appendix

G.1 Proof for Lemma 2

Proof. Since the loss function is the mean squared loss,

$$\begin{aligned}
&\inf_{f\in\mathcal{F}_n^{\mathrm{sub}}} \mathbb{E}\left[(f\circ h(Z) - f_n \circ h_m(Z))^2\right] \\
&\quad= \inf_{f\in\mathcal{F}_n^{\mathrm{sub}}} \mathbb{E}\left[\left(\mathbb{E}\left[f_n\circ h_m(Z) \mid h(Z)\right] - f_n\circ h_m(Z) + f\circ h(Z) - \mathbb{E}\left[f_n\circ h_m(Z)\mid h(Z)\right]\right)^2\right] \\
&\quad= \mathbb{E}\left[\operatorname{Var}\left(f_n\circ h_m(Z)\mid h(Z)\right)\right] + \inf_{f\in\mathcal{F}_n^{\mathrm{sub}}} \mathbb{E}\left[\left(\mathbb{E}\left[f_n\circ h_m(Z)\mid h(Z)\right] - f\circ h(Z)\right)^2\right],
\end{aligned}$$

where the last equality follows from the definition of the conditional variance and the law of iterated expectation. □

G.2 Proof for Lemma 3

Proof. The sigma-field inclusion implies that each coordinate V_j is measurable with respect to $\overline{\sigma(U)}^P$. By Proposition 2.12 of Folland (1999), there exists a $\sigma(U)$-measurable random variable $\widetilde{V}_j$ such that $\widetilde{V}_j = V_j$, P-a.s. For each $j = 1,\dots,q$, Theorem 4.2.8 of Dudley (2002) yields a measurable function $\Gamma_j : \mathcal{U} \to \mathbb{R}$ such that $\widetilde{V}_j = \Gamma_j \circ U$. Hence, $\Gamma = (\Gamma_1,\dots,\Gamma_q)' : \mathcal{U} \to \mathbb{R}^q$ is measurable and satisfies $\widetilde{V} = \Gamma \circ U$. Since $V = \widetilde{V}$, P-a.s., the result follows. □

G.3 Proof for Lemma 4

Proof. See Farrell et al. (2021), Lemma 8. □

G.4 Proof for Lemma 5

Proof. The assumed Euclidean-norm bounds imply that

$$\sup_{z\in\mathcal{Z}}|f_t(z)|, \sup_{z\in\mathcal{Z}}|f_t^*(z)| \leq M, \qquad t = 1,\dots,T.$$

The constants c_L and c_U can therefore be derived as in Lemma 9 of Farrell et al. (2021).

It remains to verify the Lipschitz constant. For $u \in \mathbb{R}^T$, define

$$p_t(u) := \frac{\exp(u_t)}{1+\sum_{j=1}^T \exp(u_j)}, \qquad t = 1,\dots,T.$$

Write $p(u) = (p_1(u),\dots,p_T(u))'$. For $y = 1,\dots,T$, let $e_y \in \mathbb{R}^T$ be the y-th standard basis vector, and let $e_{T+1} = \mathbf{0}_T$. Then,

$$\nabla_1 \ell(u,y) = p(u) - e_y.$$

For $y = 1,\dots,T$,

$$\begin{aligned}
\|\nabla_1\ell(u,y)\|_2^2 &= (1-p_y(u))^2 + \sum_{j\neq y} p_j(u)^2 \\
&\leq (1-p_y(u))^2 + \left(\sum_{j\neq y} p_j(u)\right)^2 \\
&\leq 2(1-p_y(u))^2 \\
&\leq 2,
\end{aligned}$$

where the first inequality follows from $p_j(u) \geq 0$ and the second inequality follows from $\sum_{j\neq y} p_j(u) \leq 1 - p_y(u)$. We also have $\|\nabla_1\ell(u,T+1)\|_2 = \|p(u)\|_2 \leq 1$. The fundamental theorem of calculus and the Cauchy-Schwarz inequality now give

$$\begin{aligned}
|\ell(f(z),y) - \ell(a(z),y)| &\leq \sup_{s\in[0,1]} \|\nabla_1\ell(a(z)+s(f(z)-a(z)),y)\|_2 \|f(z)-a(z)\|_2 \\
&\leq \sqrt{2}\|f(z)-a(z)\|_2.
\end{aligned}$$

□

G.5 Proof for Lemma 6

Proof. For each $k = 1, \ldots, K$, let $\{h_{k,j} : j = 1, \ldots, N_k\}$ be an ε/K-covering set of $\mathcal{H}_k$ with respect to the norm $L^2(Q)$, where $N_k = N\left(\varepsilon/K, \mathcal{H}_k, L^2(Q)\right)$. Then, for every $h \in \mathcal{H}$, there exists $j_k \in \{1, \ldots, N_k\}$ such that $\|h_k - h_{k,j_k}\|_{2,Q} < \varepsilon/K$ for each $k = 1, \ldots, K$. Thus, we have

$$\|h - (h_{1,j_1}, \ldots, h_{K,j_K})\|_{2,Q} \leq \sum_{k=1}^{K} \|h_k - h_{k,j_k}\|_{2,Q} < K \cdot (\varepsilon/K) = \varepsilon.$$

Therefore, the set $\{(h_{1,j_1}, \ldots, h_{K,j_K}) : j_k = 1, \ldots, N_k, k = 1, \ldots, K\}$ forms an ε-covering set of $\mathcal{H}$ with respect to the norm $\|\cdot\|_{2,Q}$, and its cardinality is $\prod_{k=1}^{K} N_k$. This completes the proof. □

G.6 Proof for Lemma 7

Proof. Let $\{h_j : j = 1, \ldots, N_H\}$ be an $\varepsilon/(2L_{\mathcal{F}})$-cover of $\mathcal{H}$. For each j, let $\{f_{j,k} : k = 1, \ldots, N_j\}$ be an $\varepsilon/2$-cover of $\mathcal{F}$ under $L^2(Q_{h_j})$. For every $f \circ h \in \mathcal{F} \circ \mathcal{H}$, choose h_j and $f_{j,k}$ from these covers. The triangle inequality and the Lipschitz property give

$$\|f \circ h - f_{j,k} \circ h_j\|_{2,Q} \leq L_{\mathcal{F}} \|h - h_j\|_{2,Q} + \|f - f_{j,k}\|_{2,Q_{h_j}} \leq \varepsilon.$$

Thus, $\{f_{j,k} \circ h_j : j = 1, \ldots, N_H,\ k = 1, \ldots, N_j\}$ is an ε-cover of $\mathcal{F} \circ \mathcal{H}$. Therefore,

$$N\left(\varepsilon, \mathcal{F} \circ \mathcal{H}, L^2(Q)\right) \leq \sum_{j=1}^{N_H} N_j \leq N\left(\frac{\varepsilon}{2L_{\mathcal{F}}}, \mathcal{H}, L^2(Q)\right) \sup_{h' \in \mathcal{H}} N\left(\frac{\varepsilon}{2}, \mathcal{F}, L^2(Q_{h'})\right).$$

□

G.7 Proof for Lemma 8

Proof. For every discrete probability measure Q and every $h \in \mathcal{H}_m$, define

$$d_{Q,h}(f, g) = \|f \circ h - g \circ h\|_{Q,2}.$$

For every $h \in \mathcal{H}_m$ and $f \in B_{n,h}(\delta)$,

$$\sqrt{n}\left[(\mathbb{M}_n - M_n)(f, h) - (\mathbb{M}_n - M_n)\left(f^*_{n,h}, h\right)\right] = \mathbb{G}_n\left(m_{f,h} - m_{f^*_{n,h},h}\right).$$

Hence,

$$\sup_{f \in B_{n,h}(\delta)} \sqrt{n} \left| (\mathbb{M}_n - M_n)(f,h) - (\mathbb{M}_n - M_n)\left(f^*_{n,h}, h\right) \right| = \|\mathbb{G}_n\|_{\mathcal{M}_{n,h}(\delta)}.$$

Fix $h \in \mathcal{H}_m$ and let Q be any discrete probability measure. Let Q_h denote the image of Q under h. For every $f, \widetilde{f} \in B_{n,h}(\delta)$, the Lipschitz continuity of ℓ implies that, for every (z,y),

$$\left| m_{f,h}(z,y) - m_{\widetilde{f},h}(z,y) \right| = \left| \ell\left(f \circ h(z), y\right) - \ell\left(\widetilde{f} \circ h(z), y\right) \right| \leq C_\ell \left\| f \circ h(z) - \widetilde{f} \circ h(z) \right\|_2.$$

Hence,

$$\begin{aligned} \left\| \left(m_{f,h} - m_{f^*_{n,h},h}\right) - \left(m_{\widetilde{f},h} - m_{f^*_{n,h},h}\right) \right\|_{Q,2} &= \left\| m_{f,h} - m_{\widetilde{f},h} \right\|_{Q,2} \\ &\leq C_\ell \left\| f \circ h - \widetilde{f} \circ h \right\|_{Q,2} \\ &= C_\ell d_{Q,h}(f, \widetilde{f}). \end{aligned} \tag{52}$$

Define the normalized class

$$\widetilde{\mathcal{M}}_{n,h}(\delta) = \left\{ \frac{\ell_{\text{diff}}}{2C_\ell M} : \ell_{\text{diff}} \in \mathcal{M}_{n,h}(\delta) \right\}.$$

We have

$$\begin{aligned} N\left(\varepsilon/(2C_\ell M), \widetilde{\mathcal{M}}_{n,h}(\delta), L_2(Q)\right) &= N\left(\varepsilon, \mathcal{M}_{n,h}(\delta), L_2(Q)\right) \\ &\leq N\left(\varepsilon/C_\ell, B_{n,h}(\delta), d_{Q,h}\right) \\ &\leq N\left(\varepsilon/C_\ell, \mathcal{F}_n, d_{Q,h}\right), \end{aligned} \tag{53}$$

where the first inequality follows from (52) and the second inequality follows from the fact that $B_{n,h}(\delta) \subseteq \mathcal{F}_n$. By the definition of Q_h,

$$d_{Q,h}(f,g) = \|f \circ h - g \circ h\|_{Q,2} = \|f - g\|_{Q_h,2}.$$

Also, for every $\ell_{\text{diff}} = m_{f,h} - m_{f^*_{n,h},h} \in \mathcal{M}_{n,h}(\delta)$ with $f \in B_{n,h}(\delta)$,

$$\begin{aligned} |\ell_{\text{diff}}(z,y)| \leq C_\ell \left\| f \circ h(z) - f^*_{n,h} \circ h(z) \right\|_2 &\leq C_\ell \left\| f \circ h(z) \right\|_2 + C_\ell \left\| f^*_{n,h} \circ h(z) \right\|_2 \\ &\leq 2C_\ell M. \end{aligned}$$

Hence, $\widetilde{\mathcal{M}}_{n,h}(\delta)$ has envelope function equal to 1, and, for every $\widetilde{\ell}_{\text{diff}} = \left(m_{f,h} - m_{f^*_{n,h},h}\right)/(2C_\ell M) \in \widetilde{\mathcal{M}}_{n,h}(\delta)$, the Lipschitz continuity of ℓ implies that

$$\begin{aligned} P\widetilde{\ell}^2_{\text{diff}} &= \frac{1}{4C_\ell^2 M^2} P\left(m_{f,h} - m_{f^*_{n,h},h}\right)^2 \\ &\leq \frac{1}{4M^2} \left\| f \circ h - f^*_{n,h} \circ h \right\|^2_{P,2} \\ &< \frac{\delta^2}{4M^2}. \end{aligned}$$

Because $B_{n,h}(\delta)$ is pointwise measurable by Lemma 9, there exists a countable subset $B^0_{n,h}(\delta) \subseteq B_{n,h}(\delta)$ such that every $f \in B_{n,h}(\delta)$ is the pointwise limit of a sequence in $B^0_{n,h}(\delta)$. Since h is fixed and $\ell(\cdot, y)$ is continuous on $\mathbb{R}^T$ for every y by Lipschitz continuity, the class $\mathcal{M}_{n,h}(\delta)$ is pointwise measurable. Multiplication by $(2C_\ell M)^{-1}$ preserves pointwise measurability; thus, $\widetilde{\mathcal{M}}_{n,h}(\delta)$ is pointwise measurable. Because the map $x \mapsto x^2$ is continuous, the classes $\mathcal{M}^2_{n,h}(\delta)$ and $\widetilde{\mathcal{M}}^2_{n,h}(\delta)$ are also pointwise measurable. Therefore, these classes are P-measurable. Theorem 2.14.2 in van der Vaart and Wellner (2023), applied to $\widetilde{\mathcal{M}}_{n,h}(\delta)$ with local radius $\delta/(2M)$, implies

$$\begin{aligned} \mathbb{E}_P \left\| \mathbb{G}_n \right\|_{\widetilde{\mathcal{M}}_{n,h}(\delta)} \lesssim\ & J\left(\delta/(2M), \widetilde{\mathcal{M}}_{n,h}(\delta) | 1, L_2\right) \\ & + \frac{J^2\left(\delta/(2M), \widetilde{\mathcal{M}}_{n,h}(\delta) | 1, L_2\right)}{(\delta/(2M))^2 \sqrt{n}}. \end{aligned}$$

Next, for any discrete probability measure Q' on the domain of $\mathcal{F}_n$, we have

$$\begin{aligned} & J\left(\delta/(2M), \widetilde{\mathcal{M}}_{n,h}(\delta) | 1, L_2\right) \\ &= \sup_Q \int_0^{\delta/(2M)} \sqrt{1 + \log N\left(\varepsilon, \widetilde{\mathcal{M}}_{n,h}(\delta), L_2(Q)\right)} d\varepsilon \\ &\leq \sup_Q \int_0^{\delta/(2M)} \sqrt{1 + \log N\left(2M\varepsilon, \mathcal{F}_n, d_{Q,h}\right)} d\varepsilon \\ &\leq \sup_{Q'} \int_0^{\delta/(2M)} \sqrt{1 + \log N\left(2\varepsilon M, \mathcal{F}_n, L_2(Q')\right)} d\varepsilon \\ &= \frac{1}{2} J\left(\delta/M, \mathcal{F}_n | F, L_2\right), \end{aligned}$$

where the first equality follows from the definition of the uniform entropy integral, the first inequality follows from (53), the second inequality uses the fact that every image measure Q_h is a discrete probability measure on the domain of $\mathcal{F}_n$, and the last equality uses the change

of variables $u = 2\varepsilon$. Moreover,

$$\mathbb{E}_P \|\mathbb{G}_n\|_{\mathcal{M}_{n,h}(\delta)} = 2C_\ell M \mathbb{E}_P \|\mathbb{G}_n\|_{\widetilde{\mathcal{M}}_{n,h}(\delta)}.$$

Combining these bounds yields

$$\mathbb{E}_P \|\mathbb{G}_n\|_{\mathcal{M}_{n,h}(\delta)} \lesssim MC_\ell \left(J\left(\delta/M, \mathcal{F}_n | M, L_2\right) + \frac{M^2 J^2\left(\delta/M, \mathcal{F}_n | M, L_2\right)}{\delta^2 \sqrt{n}} \right).$$

Taking the supremum over $h \in \mathcal{H}_m$ completes the proof. □

G.8 Proof for Lemma 9

Proof. Let $\mathcal{F}^0 \subset \mathcal{F}$ be a countable subset such that for every $f \in \mathcal{F}$, there exists a sequence $\{f_j\}_{j=1}^\infty$ in $\mathcal{F}^0$ such that $\|f_j(x) - f(x)\|_2 \to 0$ for every x. Pick an arbitrary $f \in B(\delta)$. We have $P\|f - f^*\|_2^2 < \delta^2$. The continuity of the squared Euclidean norm implies that $\|f_j(x) - f^*(x)\|_2^2 \to \|f(x) - f^*(x)\|_2^2$ for every x. Moreover, since $\|f_j(x)\|_2 \le F(x)$ and $\|f^*(x)\|_2 \le F(x)$, we have $\|f_j(x) - f^*(x)\|_2^2 \le 4F(x)^2$. Since $PF^2 < \infty$, applying the dominated convergence theorem to the scalar functions $\|f_j(x) - f^*(x)\|_2^2$ yields

$$P\|f_j - f^*\|_2^2 \to P\|f - f^*\|_2^2 < \delta^2.$$

Therefore, $f_j \in B(\delta)$ eventually. Hence, every $f \in B(\delta)$ is the pointwise limit of a sequence in the countable set $\mathcal{F}^0 \cap B(\delta)$, which proves that $B(\delta)$ is pointwise measurable. □

G.9 Proof for Lemma 10

Proof. See Theorem 2.6.7 in van der Vaart and Wellner (2023). □

G.10 Proof for Lemma 11

Proof. The proof follows the structure of Lemmas 3.2.5 and 3.4.1 in van der Vaart and Wellner (2023), with additional care required for the pre-trained estimator $\check{h}$ and local curvature τ_n. Suppose that $\mathbb{M}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) \ge \mathbb{M}_n\left(f_{n,\check{h}}, \check{h}\right) - R_n$ for some $R_n = O_P\left(\delta_n^2\right)$. For every $h \in \mathcal{H}_m$ and $j \in \mathbb{N}$, define the following sets, called "shells":

$$S_{n,h,j} = \left\{ f \in \mathcal{F}_n : 2^{j-1}\tau_n^{-1}\delta_n < d_{n,h}\left(f, f_{n,h}^*\right) \le 2^j \tau_n^{-1}\delta_n \right\}.$$

Because $d_{n,h}\left(\cdot, f^*_{n,h}\right) \geq 0$, it suffices to show that for every $\varepsilon > 0$, there exists an integer J, possibly depending on ε but not on $\check{h}$, n, and m, such that

$$\begin{aligned}
&\limsup_{n\to\infty} P\left(d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f^*_{n,\check{h}}\right) > 2^J \tau_n^{-1}\delta_n\right) \\
&\quad \leq \limsup_{n\to\infty} P\left(\widehat{f}_{n,\check{h}} \in \bigcup_{j>J} S_{n,\check{h},j}\right) \\
&\quad \leq \limsup_{n\to\infty} \sum_{j>J} P\left(\widehat{f}_{n,\check{h}} \in S_{n,\check{h},j}, R_n \leq 2^J\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\right) \\
&\qquad + \limsup_{n\to\infty} P\left(R_n > 2^J\delta_n^2\right) + \limsup_{m\to\infty} P\left(\check{h} \notin \mathcal{H}_{m,\varepsilon}\right) \\
&\quad \leq 3\varepsilon.
\end{aligned}$$

The second term on the right-hand side can be made arbitrarily small by choosing J large enough since $R_n = O_P\left(\delta_n^2\right)$. The third term on the right-hand side is less than or equal to ε by the assumption on $\mathcal{H}_{m,\varepsilon}$. We will show that the first term on the right-hand side can also be made arbitrarily small by choosing J large enough.

Take $h \in \mathcal{H}_{m,\varepsilon}$ and $f \in S_{n,h,j}$. By (32), if $2^j\tau_n^{-1}\delta_n \geq \tau_n^{-1}C_{\varepsilon,0}\underline{\delta}_n$, or equivalently $2^j\delta_n \geq C_{\varepsilon,0}\underline{\delta}_n$, then

$$M_n(f,h) - M_n\left(f^*_{n,h}, h\right) \lesssim_\varepsilon -2^{2j}\delta_n^2.$$

Now consider $M_n(f,h) - M_n\left(f_{n,h}, h\right)$. Add and subtract $M_n\left(f^*_{n,h}, h\right)$ to get

$$\begin{aligned}
&M_n(f,h) - M_n\left(f_{n,h}, h\right) \\
&\quad = M_n(f,h) - M_n\left(f^*_{n,h}, h\right) + M_n\left(f^*_{n,h}, h\right) - M_n\left(f_{n,h}, h\right) \\
&\quad \leq -2^{2j}C_{\varepsilon,1}\delta_n^2 + M_n\left(f^*_{n,h}, h\right) - M_n\left(f_{n,h}, h\right).
\end{aligned}$$

By (34), the second term is bounded above by a constant multiple of δ_n^2. Thus,

$$M_n(f,h) - M_n\left(f_{n,h}, h\right) \leq -2^{2j}C_{\varepsilon,1}\delta_n^2 + C_{\varepsilon,2}\delta_n^2.$$

Thus, for every $j > J$ with sufficiently large J, we can bound

$$M_n(f,h) - M_n\left(f_{n,h}, h\right) \lesssim_\varepsilon -2^{2j}\delta_n^2 \quad \text{for every } h \in \mathcal{H}_{m,\varepsilon} \text{ and } f \in S_{n,h,j}. \tag{54}$$

Define the empirical process $\mathbb{G}_n(f,h) = \sqrt{n}\left(\mathbb{M}_n - M_n\right)(f,h)$. Then,

$$\begin{aligned}
&\mathbb{G}_n\left(\widehat{f}_{n,\check{h}},\check{h}\right) - \mathbb{G}_n\left(f_{n,\check{h}},\check{h}\right) \\
&\quad = \sqrt{n}\left[\mathbb{M}_n\left(\widehat{f}_{n,\check{h}},\check{h}\right) - \mathbb{M}_n\left(f_{n,\check{h}},\check{h}\right)\right] - \sqrt{n}\left[M_n\left(\widehat{f}_{n,\check{h}},\check{h}\right) - M_n\left(f_{n,\check{h}},\check{h}\right)\right].
\end{aligned}$$

The first term on the right-hand side is bounded below by the approximate maximization property of $\widehat{f}_{n,h}$:

$$\sqrt{n}\left[\mathbb{M}_n\left(\widehat{f}_{n,\check{h}},\check{h}\right) - \mathbb{M}_n\left(f_{n,\check{h}},\check{h}\right)\right] \geq -\sqrt{n}R_n.$$

Suppose that the event $\check{h} \in \mathcal{H}_{m,\varepsilon}$ and $\widehat{f}_{n,\check{h}} \in S_{n,\check{h},j}$ occurs. The second term on the right-hand side is bounded below by (54),

$$-\sqrt{n}\left[M_n\left(\widehat{f}_{n,\check{h}},\check{h}\right) - M_n\left(f_{n,\check{h}},\check{h}\right)\right] \gtrsim_\varepsilon \sqrt{n}2^{2j}\delta_n^2.$$

Combining these two bounds, we have

$$\mathbb{G}_n\left(\widehat{f}_{n,\check{h}},\check{h}\right) - \mathbb{G}_n\left(f_{n,\check{h}},\check{h}\right) \geq \sqrt{n}(C_{\varepsilon,3}2^{2j}\delta_n^2 - R_n) \quad \text{if } \check{h} \in \mathcal{H}_{m,\varepsilon} \text{ and } \widehat{f}_{n,\check{h}} \in S_{n,\check{h},j}.$$

Thus, by the set inclusion, we have for every $j > J$ with sufficiently large J,

$$\begin{aligned}
&P\left(\widehat{f}_{n,\check{h}} \in S_{n,\check{h},j}, R_n \leq 2^J\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\right) \\
&\quad \leq P^*\left(\sup_{f \in S_{n,\check{h},j}} \mathbb{G}_n\left(f,\check{h}\right) - \mathbb{G}_n\left(f_{n,\check{h}},\check{h}\right) \gtrsim_\varepsilon \sqrt{n}2^{2j}\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\right) \\
&\quad \leq \sup_{h \in \mathcal{H}_{m,\varepsilon}} P\left(\sup_{f \in S_{n,h,j}} \mathbb{G}_n\left(f,h\right) - \mathbb{G}_n\left(f_{n,h},h\right) \gtrsim_\varepsilon \sqrt{n}2^{2j}\delta_n^2\right),
\end{aligned}$$

where the last inequality follows because $\check{h}$ is independent of $\mathbb{G}_n\left(f,h\right) - \mathbb{G}_n\left(f_{n,h},h\right)$.

By Markov's inequality, the main term on the right-hand side is bounded above by a constant multiple of

$$\begin{aligned}
&\sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{f \in S_{n,h,j}} \left|\mathbb{G}_n(f,h) - \mathbb{G}_n\left(f_{n,h},h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2) \\
&\lesssim_\varepsilon \sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{f \in \mathcal{F}_n : d_{n,h}\left(f, f_{n,h}^*\right) \leq 2^j\tau_n^{-1}\delta_n} \left|\mathbb{G}_n(f,h) - \mathbb{G}_n\left(f_{n,h}^*,h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2) \qquad (55) \\
&\quad + \sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\left|\mathbb{G}_n\left(f_{n,h}^*,h\right) - \mathbb{G}_n\left(f_{n,h},h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2), \qquad (56)
\end{aligned}$$

where the inequality follows from the triangle inequality and the definition of $S_{n,h,j}$. By the modulus bound in (33), the first term (55) is proportionally bounded above by $\phi_n\left(2^{j+1}\tau_n^{-1}\delta_n\right)/(\sqrt{n}2^{2j}\delta_n^2)$. To bound the second term (56), note that $d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \lesssim_\varepsilon \tau_n^{-1}\delta_n$ since either $d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \leq \tau_n^{-1}C_{\varepsilon,0}\underline{\delta}_n$ or $d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) > \tau_n^{-1}C_{\varepsilon,0}\underline{\delta}_n$ holds. For the former case, we have $d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \lesssim_\varepsilon \tau_n^{-1}\delta_n$ directly from $\delta_n \gtrsim_\varepsilon \underline{\delta}_n$. For the latter case, apply (32) with $\delta = d_{n,h}\left(f_{n,h}, f_{n,h}^*\right)$ to get

$$M_n\left(f_{n,h}, h\right) - M_n\left(f_{n,h}^*, h\right) \lesssim_\varepsilon -\tau_n^2 d_{n,h}\left(f_{n,h}, f_{n,h}^*\right)^2.$$

By (34), the left-hand side is bounded below by a constant multiple of $-\delta_n^2$. Thus, $-\delta_n^2 \lesssim_\varepsilon -\tau_n^2 d_{n,h}\left(f_{n,h}, f_{n,h}^*\right)^2$, which implies $d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \lesssim_\varepsilon \tau_n^{-1}\delta_n$. We can apply (33) to bound the second term (56) proportionally by $\phi_n\left(\tau_n^{-1}\delta_n\right)/(\sqrt{n}2^{2j}\delta_n^2)$ since $\phi_n(\delta)/\delta^\xi$ is decreasing.[30] Combining these bounds, we have

$$P\left(\widehat{f}_{n,\check{h}} \in S_{n,\check{h},j}, R_n \leq 2^J\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\right) \lesssim_\varepsilon \frac{\phi_n\left(2^{j+1}\tau_n^{-1}\delta_n\right)}{\sqrt{n}2^{2j}\delta_n^2}.$$

Using the properties of ϕ_n in (34), we have

$$P\left(\widehat{f}_{n,\check{h}} \in S_{n,\check{h},j}, R_n \leq 2^J\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\right) \lesssim_\varepsilon \frac{\phi_n\left(\tau_n^{-1}\delta_n\right)}{\sqrt{n}\delta_n^2} \cdot \frac{2^\xi}{2^{j(2-\xi)}} \lesssim_\varepsilon \frac{1}{2^{j(2-\xi)}},$$

where the first inequality follows from the decreasing property of $\delta \mapsto \phi_n(\delta)/\delta^\xi$ and the second inequality follows from $\phi_n\left(\delta_n/\tau_n\right) \lesssim_\varepsilon \sqrt{n}\delta_n^2$. Recall that $\lesssim_\varepsilon$ means that the constant in the bound may depend on ε but not on $\check{h}$, n, m, and j. Therefore, by summing over $j > J$, we can take J not depending on $\check{h}$, n, and m such that the right-hand side of

$$\limsup_{n\to\infty} \sum_{j>J} P\left(\widehat{f}_{n,\check{h}} \in S_{n,\check{h},j}, R_n \leq 2^J\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\right) \lesssim_\varepsilon \sum_{j>J} \frac{1}{2^{j(2-\xi)}}$$

is arbitrarily small. This completes the proof. □

G.11 Proof for Lemma 12

Proof. The proof follows the structure of Addendum 3.4.7 in van der Vaart and Wellner (2023), with additional care required for the pre-trained estimator $\check{h}$ and local curvature τ_n. For every

[30] If we have $d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \leq C_{\varepsilon,4}\tau_n^{-1}\delta_n$ for some constant $C_{\varepsilon,4} \leq 1$, the bound follows directly from (33). If $C_{\varepsilon,4} > 1$, apply the decreasing property and get $\phi_n\left(C_{\varepsilon,4}\tau_n^{-1}\delta_n\right) \leq C_{\varepsilon,4}^\xi \phi_n\left(\tau_n^{-1}\delta_n\right)$.

$h \in \mathcal{H}_m$ and $j \in \mathbb{N}$, define

$$S_{n,h,j} = \{(f,\lambda) \in \mathcal{F}_n \times [\lambda_n, \infty) : 2^{2j-2}\delta_n^2 < \tau_n^2 d_{n,h}\left(f, f_{n,h}^*\right)^2 + \lambda^2 \mathcal{J}_n^2(f) \leq 2^{2j}\delta_n^2,$$
$$2^{2J}\lambda^2 \mathcal{J}_n^2(f_{n,h}) < \tau_n^2 d_{n,h}\left(f, f_{n,h}^*\right)^2 + \lambda^2 \mathcal{J}_n^2(f)\Big\}$$

for some large enough integer J to be specified later. We show that for every $\varepsilon > 0$, there exists an integer J such that

$$\begin{aligned}
&\limsup_{n\to\infty} P\Bigg(\tau_n^2 d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right)^2 + \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) > 2^{2J}\delta_n^2\,, \\
&\qquad \tau_n^2 d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right)^2 + \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) > 2^{2J}\widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(f_{n,\check{h}}\right), \\
&\qquad\qquad d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right) \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n, \widehat{\lambda}_{n,\check{h}} \geq \lambda_n\Bigg) \\
&\leq \limsup_{n\to\infty} P\left((\widehat{f}_{n,\check{h}}, \widehat{\lambda}_{n,\check{h}}) \in \bigcup_{j>J} S_{n,\check{h},j}, d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right) \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n, \widehat{\lambda}_{n,\check{h}} \geq \lambda_n\right) \\
&\leq \varepsilon. \qquad (57)
\end{aligned}$$

For simplicity, we assume that $\mathbb{M}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \mathbb{M}_n\left(f_{n,\check{h}}, \check{h}\right) \geq \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) - \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(f_{n,\check{h}}\right) - 2^J\delta_n^2$ for some large enough integer J, which holds with arbitrarily high probability by $\mathbb{M}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \mathbb{M}_n\left(f_{n,\check{h}}, \check{h}\right) \geq \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) - \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(f_{n,\check{h}}\right) - O_P\left(\delta_n^2\right)$, and that $\check{h} \in \mathcal{H}_{m,\varepsilon}$, which holds with probability at least $1-\varepsilon$ by the assumption on $\mathcal{H}_{m,\varepsilon}$.[31]

Take $(f,\lambda) \in S_{n,h,j}$ with $d_{n,h}\left(f, f_{n,h}^*\right) \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n$. Apply (35) with $\delta = d_{n,h}\left(f, f_{n,h}^*\right)$ to get

$$M_n(f,h) - M_n\left(f_{n,h}^*, h\right) \lesssim_\varepsilon -\tau_n^2 d_{n,h}\left(f, f_{n,h}^*\right)^2.$$

Then, for every $j > J$ with sufficiently large J, we can bound

$$\begin{aligned}
&M_n(f,h) - M_n\left(f_{n,h}, h\right) - \lambda^2\mathcal{J}_n^2(f) + \lambda^2\mathcal{J}_n^2\left(f_{n,h}\right) \\
&= \left(M_n(f,h) - M_n\left(f_{n,h}^*, h\right)\right) + \left(M_n\left(f_{n,h}^*, h\right) - M_n\left(f_{n,h}, h\right)\right) - \lambda^2\mathcal{J}_n^2(f) + \lambda^2\mathcal{J}_n^2\left(f_{n,h}\right) \\
&\leq -\tau_n^2 C_{\varepsilon,1} d_{n,h}\left(f, f_{n,h}^*\right)^2 + C_{\varepsilon,2}\delta_n^2 - \lambda^2\mathcal{J}_n^2(f) + \lambda^2\mathcal{J}_n^2\left(f_{n,h}\right) \\
&\leq -(\tau_n^2 d_{n,h}\left(f, f_{n,h}^*\right)^2 + \lambda^2\mathcal{J}_n^2(f))(\min\{C_{\varepsilon,1}, 1\} - 2^{-2J}) + C_{\varepsilon,2}\delta_n^2 \\
&\leq -(2^{2j-2}(\min\{C_{\varepsilon,1}, 1\} - 2^{-2J}) - C_{\varepsilon,2})\delta_n^2 \\
&\lesssim_\varepsilon -2^{2j}\delta_n^2, \qquad (58)
\end{aligned}$$

[31] We can take $R_n = 2^J\delta_n^2$ as in the proof of Lemma 11 and take J large enough that $\limsup_{n\to\infty} P\left(R_n > 2^J\delta_n^2\right)$ can be made arbitrarily small and the proportional bounds hold. The remaining probabilities in the proof can be evaluated after intersecting with the event $\{R_n \leq 2^J\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\}$.

where the first equality follows from adding and subtracting $M_n\left(f^*_{n,h}, h\right)$, the first inequality follows from the above display and $M_n\left(f^*_{n,h}, h\right) - M_n(f_{n,h}, h) \lesssim_\varepsilon \delta_n^2$ in (37), the second inequality follows from $2^{2J}\lambda^2\mathcal{J}_n^2(f_{n,h}) < \tau_n^2 d_{n,h}\left(f, f^*_{n,h}\right)^2 + \lambda^2\mathcal{J}_n^2(f)$ in the definition of $S_{n,h,j}$, and the third inequality follows from $2^{2j-2}\delta_n^2 \leq \tau_n^2 d_{n,h}\left(f, f^*_{n,h}\right)^2 + \lambda^2\mathcal{J}_n^2(f)$ in the definition of $S_{n,h,j}$.

Define the empirical process $\mathbb{G}_n(f,h) = \sqrt{n}\left(\mathbb{M}_n - M_n\right)(f,h)$. Suppose that $(\widehat{f}_{n,\check{h}}, \widehat{\lambda}_{n,\check{h}}) \in S_{n,\check{h},j}$ with $d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f^*_{n,\check{h}}\right) \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n$ and $\widehat{\lambda}_{n,\check{h}} \geq \lambda_n$. Then,

$$
\begin{aligned}
&\mathbb{G}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \mathbb{G}_n\left(f_{n,\check{h}}, \check{h}\right) \\
&\quad = \sqrt{n}\left[\mathbb{M}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \mathbb{M}_n\left(f_{n,\check{h}}, \check{h}\right)\right] - \sqrt{n}\left[M_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - M_n\left(f_{n,\check{h}}, \check{h}\right)\right] \\
&\quad \geq \sqrt{n}\left(\widehat{\lambda}^2_{n,\check{h}}\mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) - \widehat{\lambda}^2_{n,\check{h}}\mathcal{J}_n^2\left(f_{n,\check{h}}\right)\right) - \sqrt{n}2^J\delta_n^2 - \sqrt{n}\left[M_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - M_n\left(f_{n,\check{h}}, \check{h}\right)\right] \\
&\quad \geq \sqrt{n}C_{\varepsilon,3}2^{2j}\delta_n^2 - \sqrt{n}2^J\delta_n^2,
\end{aligned}
$$

where the first inequality follows from the approximate-maximization property of $(\widehat{f}_{n,h}, \widehat{\lambda}_{n,h})$ in (38) and the second inequality follows from (58). Here, for large enough J,

$$
\begin{aligned}
&P\left(\mathbb{G}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \mathbb{G}_n\left(f_{n,\check{h}}, \check{h}\right) \geq \sqrt{n}C_{\varepsilon,3}2^{2j}\delta_n^2 - \sqrt{n}2^J\delta_n^2\right) \\
&\quad = P\left(\mathbb{G}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \mathbb{G}_n\left(f_{n,\check{h}}, \check{h}\right) \gtrsim_\varepsilon \sqrt{n}2^{2j}\delta_n^2\right).
\end{aligned}
$$

Also, $(\widehat{f}_{n,\check{h}}, \widehat{\lambda}_{n,\check{h}}) \in S_{n,\check{h},j}$ with $\widehat{\lambda}_{n,\check{h}} \geq \lambda_n$ implies that $d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f^*_{n,\check{h}}\right) \leq 2^j\tau_n^{-1}\delta_n$ and $\mathcal{J}_n\left(\widehat{f}_{n,\check{h}}\right) < 2^j\delta_n/\lambda_n$. Thus, by the set inclusion, we have

$$
\begin{aligned}
&P\left((\widehat{f}_{n,\check{h}}, \widehat{\lambda}_{n,\check{h}}) \in S_{n,\check{h},j}, d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f^*_{n,\check{h}}\right) \geq \tau_n^{-1}C_{\varepsilon,0}\underline{\delta}_n, \widehat{\lambda}_{n,\check{h}} \geq \lambda_n\right) \\
&\quad \leq P^*\left(\sup_{\substack{f\in\mathcal{F}_n: d_{n,\check{h}}\left(f, f^*_{n,\check{h}}\right) \leq 2^j\tau_n^{-1}\delta_n, \\ \mathcal{J}_n(f) < 2^j\delta_n/\lambda_n}} \left(\mathbb{G}_n\left(f, \check{h}\right) - \mathbb{G}_n\left(f_{n,\check{h}}, \check{h}\right)\right) \gtrsim_\varepsilon \sqrt{n}2^{2j}\delta_n^2\right) \\
&\quad \leq \sup_{h\in\mathcal{H}_{m,\varepsilon}} P\left(\sup_{\substack{f\in\mathcal{F}_n: d_{n,h}\left(f, f^*_{n,h}\right) \leq 2^j\tau_n^{-1}\delta_n, \\ \mathcal{J}_n(f) < 2^j\delta_n/\lambda_n}} \left(\mathbb{G}_n(f,h) - \mathbb{G}_n\left(f_{n,h}, h\right)\right) \gtrsim_\varepsilon \sqrt{n}2^{2j}\delta_n^2\right).
\end{aligned}
$$

By Markov's inequality, the main term on the right-hand side is bounded above by some

constant depending on ε times

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}_n: d_{n,h}\left(f,f^*_{n,h}\right)\leq 2^j\tau_n^{-1}\delta_n,\\ \mathcal{J}_n(f)<2^j\delta_n/\lambda_n}} \left|\mathbb{G}_n\left(f,h\right)-\mathbb{G}_n\left(f_{n,h},h\right)\right|\right]/(\sqrt{n}2^{2j}\delta_n^2)$$

$$\leq \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}_n: d_{n,h}\left(f,f^*_{n,h}\right)\leq 2^j\tau_n^{-1}\delta_n,\\ \mathcal{J}_n(f)<2^j\delta_n/\lambda_n}} \left|\mathbb{G}_n\left(f,h\right)-\mathbb{G}_n\left(f^*_{n,h},h\right)\right|\right]/(\sqrt{n}2^{2j}\delta_n^2) \tag{59}$$

$$+ \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\left|\mathbb{G}_n\left(f^*_{n,h},h\right)-\mathbb{G}_n\left(f_{n,h},h\right)\right|\right]/(\sqrt{n}2^{2j}\delta_n^2). \tag{60}$$

By the modulus bound in (36), the first term (59) is proportionally bounded above by $\phi_n\left(2^{j+1}\tau_n^{-1}\delta_n\right)/(\sqrt{n}2^{2j}\delta_n^2)$. To bound the second term (60), note that $d_{n,h}\left(f_{n,h},f^*_{n,h}\right) < \tau_n^{-1}\delta_n$ and $\mathcal{J}_n(f_{n,h}) < \delta_n/\lambda_n$ by (37). Apply (36) to bound the second term (60) proportionally by $\phi_n\left(\tau_n^{-1}\delta_n\right)/(\sqrt{n}2^{2j}\delta_n^2)$. Combining these bounds, we have

$$P\left((\widehat{f}_{n,\check{h}},\widehat{\lambda}_{n,\check{h}})\in S_{n,\check{h},j}, d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}},f^*_{n,\check{h}}\right)\geq\tau_n^{-1}C_{\varepsilon,0}\underline{\delta}_n,\widehat{\lambda}_{n,\check{h}}\geq\lambda_n\right)\lesssim_\varepsilon \frac{\phi_n\left(2^{j+1}\tau_n^{-1}\delta_n\right)}{\sqrt{n}2^{2j}\delta_n^2}.$$

The remainder of the proof proceeds as in the proof of Lemma 11, using the properties of ϕ_n in (37) and taking J large enough. We obtain (57), which implies that for large enough J,

$$\limsup_{n\to\infty} P\left(\tau_n d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}},f^*_{n,\check{h}}\right)>2^J\max\left\{\delta_n,\widehat{\lambda}_{n,\check{h}}\mathcal{J}_n\left(f_{n,\check{h}}\right)\right\},\right.$$
$$\left. d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}},f^*_{n,\check{h}}\right)\geq\tau_n^{-1}C_{\varepsilon,0}\underline{\delta}_n,\widehat{\lambda}_{n,\check{h}}\geq\lambda_n\right)\leq\varepsilon$$

since $\widehat{\lambda}^2_{n,\check{h}}\mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right)\geq 0$. As $\delta_n\gtrsim_\varepsilon\underline{\delta}_n$, we can take J large enough that $2^J\delta_n\geq C_{\varepsilon,0}\underline{\delta}_n$. Hence, the event in the next display implies $d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}},f^*_{n,\check{h}}\right)>\tau_n^{-1}C_{\varepsilon,0}\underline{\delta}_n$, and we obtain

$$\limsup_{n\to\infty} P\left(\tau_n d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}},f^*_{n,\check{h}}\right)>2^J\max\left\{\delta_n,\widehat{\lambda}_{n,\check{h}}\mathcal{J}_n\left(f_{n,\check{h}}\right)\right\},\widehat{\lambda}_{n,\check{h}}\geq\lambda_n\right)\leq\varepsilon.$$

Thus, on the event $\{\widehat{\lambda}_{n,\check{h}}\geq\lambda_n\}$, we have

$$\tau_n d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}},f^*_{n,\check{h}}\right)=O_P(1)\left(\delta_n+\widehat{\lambda}_{n,\check{h}}\mathcal{J}_n\left(f_{n,\check{h}}\right)\right).$$

This completes the proof. □

G.12 Proof for Lemma 13

Proof. Pick arbitrary $\varepsilon > 0$. By the assumptions, there exist constants $C_{\varepsilon,1} > 0$, $C_{\varepsilon,2} > 0$, and $C_{\varepsilon,3} > 0$ such that

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \leq C_{\varepsilon,1}\underline{\delta}_n, \tag{61}$$

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \sup_{f \in \mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f_{n,h}^*\right) \leq \delta} M_n(f,h) - M_n\left(f_{n,h}^*, h\right) \leq -C_{\varepsilon,2}\tau_n^2\delta^2, \tag{62}$$

and

$$M_n\left(f_{n,h}^*, h\right) - M_n\left(f_{n,h}, h\right) \leq C_{\varepsilon,3} d_{n,h}\left(f_{n,h}, f_{n,h}^*\right)^2 \tag{63}$$

for every $h \in \mathcal{H}_{m,\varepsilon}$. Set

$$C_{\varepsilon,0} := C_{\varepsilon,1} \max\left\{4, \sqrt{32 C_{\varepsilon,3}/C_{\varepsilon,2}}\right\},$$

which depends on ε but not on n. Pick arbitrary $\delta \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n$, $h \in \mathcal{H}_{m,\varepsilon}$, and $f \in \mathcal{F}_n$ such that $\delta/2 < d_{n,h}\left(f, f_{n,h}\right) \leq \delta$. By the choice of $C_{\varepsilon,0}$, (61), and $\tau_n \leq 1$, we have

$$d_{n,h}\left(f, f_{n,h}\right) > \delta/2 \geq 2 d_{n,h}\left(f_{n,h}, f_{n,h}^*\right), \tag{64}$$

$$d_{n,h}\left(f, f_{n,h}\right) > \delta/2 \geq \sqrt{8 C_{\varepsilon,3}/C_{\varepsilon,2}}\, \tau_n^{-1} d_{n,h}\left(f_{n,h}, f_{n,h}^*\right). \tag{65}$$

The reverse triangle inequality and (64) imply

$$d_{n,h}\left(f, f_{n,h}^*\right) \geq d_{n,h}\left(f, f_{n,h}\right) - d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \geq \frac{1}{2} d_{n,h}\left(f, f_{n,h}\right) > 0, \tag{66}$$

Adding and subtracting $M_n\left(f_{n,h}^*, h\right)$, applying (62) with shell radius $d_{n,h}\left(f, f_{n,h}^*\right)$, and using (63), we obtain

$$\begin{aligned}
M_n(f,h) - M_n\left(f_{n,h}, h\right) &\leq -C_{\varepsilon,2}\tau_n^2 d_{n,h}\left(f, f_{n,h}^*\right)^2 + C_{\varepsilon,3} d_{n,h}\left(f_{n,h}, f_{n,h}^*\right)^2 \\
&\leq -(C_{\varepsilon,2}/4)\tau_n^2 d_{n,h}\left(f, f_{n,h}\right)^2 + C_{\varepsilon,3} d_{n,h}\left(f_{n,h}, f_{n,h}^*\right)^2 \\
&\leq -(C_{\varepsilon,2}/8)\tau_n^2 d_{n,h}\left(f, f_{n,h}\right)^2 \\
&\leq -(C_{\varepsilon,2}/32)\tau_n^2\delta^2,
\end{aligned}$$

where the second inequality follows from (66), the third inequality follows from (65), and the last inequality follows from $d_{n,h}\left(f, f_{n,h}\right) \geq \delta/2$. Taking the supremum over the stated shell and over $h \in \mathcal{H}_{m,\varepsilon}$ proves the result. □

G.13 Proof for Lemma 14

Proof. If $L_\phi = 0$, the Lipschitz condition and $\phi_i(0) = 0$ imply that $\phi_i(u) = 0$ for every $u \in \mathbb{R}$ and every i, so the conclusion follows immediately. Suppose henceforth that $L_\phi > 0$. By the contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12), for any contraction $\phi_{0,i} : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}\left[\sup_{u\in\mathcal{U}}\left|\sum_{i=1}^n \xi_i\phi_{0,i}(u_i)\right|\right] \le 2\cdot\mathbb{E}\left[\sup_{u\in\mathcal{U}}\left|\sum_{i=1}^n \xi_i u_i\right|\right].$$

For a Lipschitz function $\phi_i : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant L_ϕ, we can define $\phi_{0,i}(x) = \phi_i(x)/L_\phi$, which is a contraction. Then,

$$\begin{aligned}\mathbb{E}\left[\sup_{u\in\mathcal{U}}\left|\sum_{i=1}^n \xi_i\phi_i(u_i)\right|\right] &= L_\phi\cdot\mathbb{E}\left[\sup_{u\in\mathcal{U}}\left|\sum_{i=1}^n \xi_i\phi_{0,i}(u_i)\right|\right] \\ &\le 2L_\phi\cdot\mathbb{E}\left[\sup_{u\in\mathcal{U}}\left|\sum_{i=1}^n \xi_i u_i\right|\right].\end{aligned}$$

This completes the proof. □

G.14 Proof for Lemma 15

Proof. The inequalities directly follow from Theorem II.3.9 of Stewart and Sun (1990). The Frobenius norm $\|\cdot\|_F$ and the spectral norm $\|\cdot\|_{\text{sp}}$ are unitarily invariant (Stewart and Sun, 1990, p.74). The nuclear norm $\|\cdot\|_*$ is also unitarily invariant (Horn and Johnson, 1994, p.211, Problem 5). □

G.15 Proof for Lemma 16

Proof. By the singular value decomposition, we can write $A = \sigma uv'$ for some $\sigma \ge 0$ and unit vectors u and v. Then, $\|A\|_* = \|A\|_F = \|A\|_{\text{sp}} = \sigma$ since A has either one nonzero singular value σ or no nonzero singular values. Also, $\sqrt{\text{tr}(A'A)} = \sqrt{\text{tr}(\sigma^2 vu'uv')} = \sqrt{\sigma^2\,\text{tr}(vv')} = \sigma$ since u and v are unit vectors. □

G.16 Proof for Lemma 17

Proof. By the definition of the $L_2(P)$ norm,

$$\|Aa\|_{P,2}^2 = \int \|Aa(v)\|_2^2\,dP(v) \le \int \|A\|_{\text{sp}}^2\|a(v)\|_2^2\,dP(v) = \|A\|_{\text{sp}}^2\|a\|_{P,2}^2,$$

where the inequality follows from $\|Aa(v)\|_2 = \|Aa(v)\|_F$ and Lemma 15. This gives the desired inequality after taking square roots of both sides. □

G.17 Proof for Lemma 18

Proof. By Harville (2008), Theorem 20.5.6, the Moore-Penrose inverse from the singular value decomposition $A = U\Sigma V'$ is given by $A^+ = V\Sigma^+ U'$, where Σ is the singular value matrix of A and Σ^+ is the Moore-Penrose inverse of Σ obtained by replacing each nonzero singular value σ with $1/\sigma$ and leaving the zero singular values unchanged. Hence, we have $\|A^+\|_{\text{sp}} = \|\Sigma^+\|_{\text{sp}} = 1/\sigma^+_{\min}(A)$ as the spectral norm is equal to the largest singular value. □

G.18 Proof for Lemma 19

Proof. By the definition of $\mathcal{N}_A(\lambda)$, we have

$$\begin{aligned}
&\mathcal{N}_{\widetilde{A}}(\lambda) - \mathcal{N}_A(\lambda) \\
&\quad = \lambda^2 \operatorname{tr}\left((A + \lambda^2 I)^{-1} - (\widetilde{A} + \lambda^2 I)^{-1}\right) \\
&\quad = \lambda^2 \operatorname{tr}\left((\widetilde{A} + \lambda^2 I)^{-1}(\widetilde{A} - A)(A + \lambda^2 I)^{-1}\right),
\end{aligned}$$

where the second equality follows from the matrix identity $X^{-1} - Y^{-1} = Y^{-1}(Y - X)X^{-1}$ for any invertible matrices X and Y.

We bound the absolute value of the above expression:

$$\begin{aligned}
&\lambda^2 \left|\operatorname{tr}\left((\widetilde{A} + \lambda^2 I)^{-1}(\widetilde{A} - A)(A + \lambda^2 I)^{-1}\right)\right| \\
&\quad \le \lambda^2 \|(\widetilde{A} + \lambda^2 I)^{-1}(\widetilde{A} - A)(A + \lambda^2 I)^{-1}\|_* \\
&\quad \le \lambda^2 \|(\widetilde{A} + \lambda^2 I)^{-1}\|_{\text{sp}} \|\widetilde{A} - A\|_* \|(A + \lambda^2 I)^{-1}\|_{\text{sp}} \\
&\quad \le \|A - \widetilde{A}\|_*/\lambda^2,
\end{aligned}$$

where the first inequality follows from the singular value decomposition of $(\widetilde{A} + \lambda^2 I)^{-1}(\widetilde{A} - A)(A + \lambda^2 I)^{-1}$, the cyclic property of the trace, and the fact that the matrices of left and right singular vectors are unitary, the second inequality follows from applying Lemma 15 twice, and the last inequality follows from $\|(A + \lambda^2 I)^{-1}\|_{\text{sp}} \le 1/\lambda^2$ for any positive semi-definite matrix A. □

G.19 Proof for Lemma 20

Proof. Define $Q^* = Q/\sqrt{\max\{1, \|Q\|_{\text{sp}}^2\}}$. Then $\|Q^*\|_{\text{sp}} \leq 1$ and

$$QAQ' = \max\{1, \|Q\|_{\text{sp}}^2\} Q^* A Q^{*'}.$$

Since $\|Q^*\|_{\text{sp}} \leq 1$, we have $Q^{*\prime} Q^* \preceq I$. Therefore, for every $x \in \mathbb{R}^K$,

$$x' A^{1/2} Q^{*\prime} Q^* A^{1/2} x \leq x' A x.$$

Hence, $A^{1/2} Q^{*\prime} Q^* A^{1/2} \preceq A$ and by the order-reversing property of the matrix inverse for positive definite matrices,

$$(A + \lambda^2 I)^{-1} \preceq (A^{1/2} Q^{*\prime} Q^* A^{1/2} + \lambda^2 I)^{-1}.$$

Since $Q^* A Q^{*\prime}$ and $A^{1/2} Q^{*\prime} Q^* A^{1/2}$ have the same nonzero eigenvalues, we have

$$\mathcal{N}_{Q^* A Q^{*\prime}}(\lambda) = \mathcal{N}_{A^{1/2} Q^{*\prime} Q^* A^{1/2}}(\lambda).$$

Using the identity,

$$\mathcal{N}_A(\lambda) = \operatorname{tr}\left(I - \lambda^2 (A + \lambda^2 I)^{-1}\right) = K - \lambda^2 \operatorname{tr}\left((A + \lambda^2 I)^{-1}\right)$$

we get

$$\mathcal{N}_{Q^* A Q^{*\prime}}(\lambda) = \operatorname{tr}\left(I - \lambda^2 (A^{1/2} Q^{*\prime} Q^* A^{1/2} + \lambda^2 I)^{-1}\right) \leq \operatorname{tr}\left(I - \lambda^2 (A + \lambda^2 I)^{-1}\right) = \mathcal{N}_A(\lambda).$$

Finally, let $\mu_1, \ldots, \mu_K$ denote the eigenvalues of $Q^* A Q^{*\prime}$. Using the definition of Q^*, we obtain

$$\begin{aligned}
\mathcal{N}_{QAQ'}(\lambda) &= \mathcal{N}_{\max\{1, \|Q\|_{\text{sp}}^2\} Q^* A Q^{*\prime}}(\lambda) \\
&= \sum_{j=1}^{K} \frac{\max\{1, \|Q\|_{\text{sp}}^2\} \mu_j}{\max\{1, \|Q\|_{\text{sp}}^2\} \mu_j + \lambda^2} \\
&\leq \max\{1, \|Q\|_{\text{sp}}^2\} \sum_{j=1}^{K} \frac{\mu_j}{\mu_j + \lambda^2} \\
&= \max\{1, \|Q\|_{\text{sp}}^2\} \mathcal{N}_{Q^* A Q^{*\prime}}(\lambda) \\
&\leq \max\{1, \|Q\|_{\text{sp}}^2\} \mathcal{N}_A(\lambda).
\end{aligned}$$

This completes the proof. □

H Multimodal Transfer Learning

In applications, the economist may have access to several pre-trained models, each trained on a different modality. For example, a computer scientist may train a CNN on images and a transformer on text, after which the economist uses both learned embeddings as inputs to a target model. This section extends our transfer-learning theory to this multimodal setting.

Let $r = 1, \dots, R$ index the modalities, where R is fixed as the sample sizes diverge. For each r, write $m_r = m_{r,n}$ and suppose that $m_{r,n} \to \infty$ as $n \to \infty$. The target observations satisfy

$$(Y_i, Z_{0,i}, Z_{1,i}, \dots, Z_{R,i}) \sim_{i.i.d.} P_{\text{ta}}, \qquad i = 1, \dots, n,$$

where $Z_{0,i}$ denotes the structured covariates and $Z_{r,i} \in \mathcal{Z}_r$ denotes the unstructured input for modality r. Write $Z_i = (Z_{0,i}, Z_{1,i}, \dots, Z_{R,i})$. The structured covariates may be low- or high-dimensional. For each modality r, the computer scientist observes a source sample

$$(S_{r,j}, Z_{r,j}) \sim_{i.i.d.} P_{\text{so},r}, \qquad j = 1, \dots, m_r,$$

where the modality-specific source samples and the target sample are mutually independent. Let P denote the joint distribution of these samples, and let $\mathbb{E}_{\text{so},r}$ denote expectation under $P_{\text{so},r}$.

Let $\ell_{\text{so},r}$ be the source loss for modality r, and let $\mathcal{G}_{r,m_r}$ and $\mathcal{H}_{r,m_r}$ be the corresponding classes of source heads and embedding functions. For each modality, define the set of population source-risk minimizers by

$$\mathcal{R}_{r,m_r} = \underset{g_r \in \mathcal{G}_{r,m_r}, h_r \in \mathcal{H}_{r,m_r}}{\arg\min} \mathbb{E}_{\text{so},r}[\ell_{\text{so},r}(g_r \circ h_r(Z_r), S_r)]. \tag{67}$$

For each modality r, let

$$(\breve{g}_r, \breve{h}_r) \in \underset{g_r \in \mathcal{G}_{r,m_r}, h_r \in \mathcal{H}_{r,m_r}}{\arg\min} \frac{1}{m_r} \sum_{j=1}^{m_r} \ell_{\text{so},r}(g_r \circ h_r(Z_{r,j}), S_{r,j}). \tag{68}$$

Thus, $\breve{h}_r$ is the pre-trained embedding estimated from the modality-r source sample.

Define $\mathbf{m} = (m_1, \dots, m_R)$ and $\mathcal{H}_{\mathbf{m}} = \prod_{r=1}^{R} \mathcal{H}_{r,m_r}$. For every multimodal embedding

$h = (h_1, \ldots, h_R) \in \mathcal{H}_{\mathbf{m}}$, define

$$(f \circ h)(Z) := f(Z_0, h_1(Z_1), \ldots, h_R(Z_R)).$$

After introducing the multimodal representation, we define the population target-risk minimizer correspondence. For every $h \in \mathcal{H}_{\mathbf{m}}$, let

$$\mathcal{F}_n^{\mathrm{sub}}(h) = \operatorname*{arg\,min}_{f \in \mathcal{F}_n^{\mathrm{sub}}} \mathbb{E}_{\mathrm{ta}}[\ell_{\mathrm{ta}}((f \circ h)(Z), Y)]. \tag{69}$$

Given the estimated multimodal embedding $\breve{h} = (\breve{h}_1, \ldots, \breve{h}_R)$, the economist estimates

$$\widehat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{ta}}((f \circ \breve{h})(Z_i), Y_i).$$

Definition 6 (Modality-Wise Diversity and Transferability)**.** Fix population minimizers $(g_{r,m_r}, h_{r,m_r}) \in \mathcal{R}_{r,m_r}$ for $r = 1, \ldots, R$ and $f_n \in \mathcal{F}_n^{\mathrm{sub}}(h_{\mathbf{m}})$, where $h_{\mathbf{m}} = (h_{1,m_1}, \ldots, h_{R,m_R})$. For each modality r and $\rho \geq 0$, define the source-side approximate identified set by

$$\mathcal{H}_{\mathrm{so},r,m_r}(\rho) := \left\{ h_r \in \mathcal{H}_{r,m_r} : \min_{g_r \in \mathcal{G}_{r,m_r}} \|g_r \circ h_r - g_{r,m_r} \circ h_{r,m_r}\|_{P_{\mathrm{so},r},2} \leq \rho \right\}.$$

For $h_{-r} = (h_1, \ldots, h_{r-1}, h_{r+1}, \ldots, h_R) \in \prod_{s \neq r} \mathcal{H}_{s,m_s}$, let (h_r, h_{-r}) denote the multimodal embedding whose r-th component is h_r. A jointly selected family of population target models is a collection $\{f_{n,h} : h \in \mathcal{H}_{\mathbf{m}}\}$ satisfying

$$f_{n,h} \in \mathcal{F}_n^{\mathrm{sub}}(h), \qquad h \in \mathcal{H}_{\mathbf{m}}, \tag{70}$$

and $f_{n,h_{\mathbf{m}}} = f_n$. Given such a family, define

$$D_{r,n}(h_r \mid h_{-r}) := \|f_{n,(h_r,h_{-r})} \circ (h_r, h_{-r}) - f_{n,(h_{r,m_r},h_{-r})} \circ (h_{r,m_r}, h_{-r})\|_{P_{\mathrm{ta}},2}$$

and the target-side neighborhood induced by this family

$$\mathcal{H}_{\mathrm{ta},r,n}(\rho \mid h_{-r}) := \{h_r \in \mathcal{H}_{r,m_r} : D_{r,n}(h_r \mid h_{-r}) \leq \rho\}.$$

Relative to this family, we say that the modality-r source task g_{r,m_r} is $(\rho_{\mathrm{so},r,m_r}, \rho_{\mathrm{ta},r,n})$-diverse over f_n with respect to h_{r,m_r} if

$$\mathcal{H}_{\mathrm{so},r,m_r}(\rho_{\mathrm{so},r,m_r}) \subseteq \mathcal{H}_{\mathrm{ta},r,n}(\rho_{\mathrm{ta},r,n} \mid h_{-r})$$

holds for every $h_{-r} \in \prod_{s \neq r} \mathcal{H}_{s,m_s}$. Let $\nu_{n,r} > 0$ for every n and $r = 1, \ldots, R$. Given a sequence of jointly selected families, we say that, relative to this sequence, the modality-r source task g_{r,m_r} is $\nu_{n,r}$-transferable to f_n with respect to h_{r,m_r} at rate $\delta_{\mathrm{so},r,m_r}$ if, for every sequence $\rho_{\mathrm{so},r,m_r} = O(\delta_{\mathrm{so},r,m_r})$, the $(\rho_{\mathrm{so},r,m_r}, \rho_{\mathrm{so},r,m_r}/\nu_{n,r})$-diversity condition holds for all sufficiently large n. We say that these modality-wise diversity conditions hold jointly if there exists a single jointly selected family relative to which they hold simultaneously for every $r = 1, \ldots, R$ and every $h_{-r} \in \prod_{s \neq r} \mathcal{H}_{s,m_s}$. We say that modality-wise transferability holds jointly if there exists a single sequence of jointly selected families relative to which the preceding transferability condition holds simultaneously for every $r = 1, \ldots, R$.

This definition is the modality-wise analogue of Definition 2. The joint condition requires a common selected population target model family at each n such that inclusion in each modality-specific source-side approximate identified set implies inclusion in the corresponding target-side neighborhood for every configuration of the other embeddings. This common selection permits the modalities to be replaced sequentially by their estimated embeddings.

Theorem 9 (Convergence Rate for Multimodal Transfer Learning)**.** *Fix* $(g_{r,m_r}, h_{r,m_r}) \in \mathcal{R}_{r,m_r}$ *for* $r = 1, \ldots, R$ *and* $f_n \in \mathcal{F}_n^{\mathrm{sub}}(h_{\mathbf{m}})$*. Define*

$$\|f_n \circ h_{\mathbf{m}} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} =: \epsilon_{\mathrm{ta},n}.$$

Given $\breve{h}$*, define the* $L^2(P_{\mathrm{ta}})$*-best approximation to* $f_n \circ h_{\mathbf{m}}$ *by*

$$f^{\bullet}_{n,\breve{h}} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ \breve{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2}.$$

For each modality $r = 1, \ldots, R$*, suppose that* $\nu_{n,r} > 0$ *for every* n *and that*

$$\|\breve{g}_r \circ \breve{h}_r - g_{r,m_r} \circ h_{r,m_r}\|_{P_{\mathrm{so},r},2} = O_{P_{\mathrm{so},r}}(\delta_{\mathrm{so},r,m_r}).$$

Suppose also that modality-wise transferability holds jointly in the sense of Definition 6; that is, there exists a common sequence of jointly selected population target model families relative to which g_{r,m_r} *is* $\nu_{n,r}$*-transferable to* f_n *with respect to* h_{r,m_r} *at rate* $\delta_{\mathrm{so},r,m_r}$ *for every* $r = 1, \ldots, R$*. Suppose further that*

$$\|\widehat{f} \circ \breve{h} - f^{\bullet}_{n,\breve{h}} \circ \breve{h}\|_{P_{\mathrm{ta}},2} = O_P(r_{\mathrm{ta},n}).$$

Then,

$$\|\widehat{f} \circ \breve{h} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} = O_P\left(r_{\mathrm{ta},n} + \sum_{r=1}^{R} \frac{\delta_{\mathrm{so},r,m_r}}{\nu_{n,r}} + \epsilon_{\mathrm{ta},n}\right).$$

Proof. Fix a common sequence of jointly selected population target model families witnessing simultaneous transferability for every modality. For $r = 0, \dots, R$, define the intermediate multimodal embeddings

$$\widetilde{\check{h}}^{(r)} := (\check{h}_1, \dots, \check{h}_r, h_{r+1,m_{r+1}}, \dots, h_{R,m_R}).$$

Thus, $\widetilde{\check{h}}^{(0)} = h_{\mathbf{m}}$ and $\widetilde{\check{h}}^{(R)} = \check{h}$. The triangle inequality yields

$$\begin{aligned}\|\widehat{f} \circ \check{h} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \le \|\widehat{f} \circ \check{h} - f^{\bullet}_{n,\check{h}} \circ \check{h}\|_{P_{\mathrm{ta}},2} \\ + \|f^{\bullet}_{n,\check{h}} \circ \check{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2} \\ + \|f_n \circ h_{\mathbf{m}} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2}.\end{aligned}$$

The first term on the right-hand side is $O_P(r_{\mathrm{ta},n})$, and the last term equals $\epsilon_{\mathrm{ta},n}$. By the definition of $f^{\bullet}_{n,\check{h}}$, the fact that $f_{n,\check{h}} \in \mathcal{F}^{\mathrm{sub}}_n(\check{h}) \subseteq \mathcal{F}^{\mathrm{sub}}_n$, and the triangle inequality,

$$\begin{aligned}\|f^{\bullet}_{n,\check{h}} \circ \check{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2} &\le \|f_{n,\check{h}} \circ \check{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2} \\ &\le \sum_{r=1}^{R} \|f_{n,\widetilde{\check{h}}^{(r)}} \circ \widetilde{\check{h}}^{(r)} - f_{n,\widetilde{\check{h}}^{(r-1)}} \circ \widetilde{\check{h}}^{(r-1)}\|_{P_{\mathrm{ta}},2},\end{aligned}$$

where the second inequality uses $f_{n,\widetilde{\check{h}}^{(0)}} = f_n$ and $f_{n,\widetilde{\check{h}}^{(R)}} = f_{n,\check{h}}$.

For each modality r, $\check{g}_r \in \mathcal{G}_{r,m_r}$ and the source-rate condition imply

$$\begin{aligned}\min_{g_r \in \mathcal{G}_{r,m_r}} \|g_r \circ \check{h}_r - g_{r,m_r} \circ h_{r,m_r}\|_{P_{\mathrm{so},r},2} &\le \|\check{g}_r \circ \check{h}_r - g_{r,m_r} \circ h_{r,m_r}\|_{P_{\mathrm{so},r},2} \\ &= O_{P_{\mathrm{so},r}}(\delta_{\mathrm{so},r,m_r}).\end{aligned}$$

Each source-rate event depends only on the modality-r source sample, so the same probability bound holds under the joint law P. Fix $\varepsilon > 0$. For each r, the source-rate condition gives a constant $M_{\varepsilon,r} < \infty$ such that for all sufficiently large n and corresponding m_r,

$$P\left(\check{h}_r \notin \mathcal{H}_{\mathrm{so},r,m_r}(M_{\varepsilon,r}\delta_{\mathrm{so},r,m_r})\right) \le \frac{\varepsilon}{R}.$$

Let $M_\varepsilon = \max_{1 \le r \le R} M_{\varepsilon,r}$. Since R is fixed, the union bound implies that the event

$$\mathcal{E}_n := \bigcap_{r=1}^{R} \left\{\check{h}_r \in \mathcal{H}_{\mathrm{so},r,m_r}(M_\varepsilon \delta_{\mathrm{so},r,m_r})\right\}$$

satisfies

$$P(\mathcal{E}_n) \geq 1 - \sum_{r=1}^{R} \frac{\varepsilon}{R} = 1 - \varepsilon$$

for all sufficiently large n and corresponding m_r.

The embeddings $\widetilde{h}^{(r)}$ and $\widetilde{h}^{(r-1)}$ differ only in their r-th components. Therefore, on $\mathcal{E}_n$, modality-wise transferability applies with the remaining components fixed at $(\check{h}_1, \ldots, \check{h}_{r-1}, h_{r+1,m_{r+1}}, \ldots, h_{R,m_R})$ and yields

$$\begin{aligned} D_{r,n}\left(\check{h}_r \mid \left(\check{h}_1, \ldots, \check{h}_{r-1}, h_{r+1,m_{r+1}}, \ldots, h_{R,m_R}\right)\right) &= \|f_{n,\widetilde{h}^{(r)}} \circ \widetilde{h}^{(r)} - f_{n,\widetilde{h}^{(r-1)}} \circ \widetilde{h}^{(r-1)}\|_{P_{\mathrm{ta}},2} \\ &\leq \frac{M_\varepsilon \delta_{\mathrm{so},r,m_r}}{\nu_{n,r}}. \end{aligned}$$

It follows that

$$\sum_{r=1}^{R} \|f_{n,\widetilde{h}^{(r)}} \circ \widetilde{h}^{(r)} - f_{n,\widetilde{h}^{(r-1)}} \circ \widetilde{h}^{(r-1)}\|_{P_{\mathrm{ta}},2} = O_P\left(\sum_{r=1}^{R} \frac{\delta_{\mathrm{so},r,m_r}}{\nu_{n,r}}\right).$$

Combining these bounds proves the result. □

The population target models $f_{n,h}$ used in modality-wise transferability are target-risk minimizers, whereas $f^{\bullet}_{n,\check{h}}$ in the target-stage condition is the $L^2(P_{\mathrm{ta}})$-best approximation to $f_n \circ h_{\mathbf{m}}$. The best-approximation property links these objects in the proof.

The theorem shows that, under modality-wise transferability, the representation error accumulates additively across modalities. In the image-text special case with $R = 2$, the bound becomes

$$\|\widehat{f} \circ \check{h} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} = O_P\left(r_{\mathrm{ta},n} + \frac{\delta_{\mathrm{so,image},m_{\mathrm{image}}}}{\nu_{n,\mathrm{image}}} + \frac{\delta_{\mathrm{so,text},m_{\mathrm{text}}}}{\nu_{n,\mathrm{text}}} + \epsilon_{\mathrm{ta},n}\right).$$

Thus, the convergence rate of the target estimator depends on the sum of the source rates for each modality, adjusted by the corresponding transferability coefficients.

I Simulation

This section reports the Monte Carlo design for the partially linear model in Example 1. The goal is to illustrate how the downstream DML estimator behaves when the target nuisance functions are aligned with the source-task representation and when they contain a component that is not identified from the source task.

I.1 Design

In each replication, the target sample satisfies

$$Y_i = D_i\theta_0 + h^*(Z_i^{\mathrm{ta}})'\beta_0 + U_i,$$
$$D_i = h^*(Z_i^{\mathrm{ta}})'\delta_0 + V_i,$$

where $\theta_0 = 1$, $Z_i^{\mathrm{ta}} \sim N(0, I_{50})$, $U_i \sim N(0, \sigma_U^2)$, and $V_i \sim N(0, \sigma_V^2)$ are mutually independent. The target nuisance functions are therefore

$$\alpha_0(z) = \mathbb{E}[D_i \mid Z_i^{\mathrm{ta}} = z] = h^*(z)'\delta_0,$$
$$\gamma_0(z) = \mathbb{E}[Y_i - \theta_0 D_i \mid Z_i^{\mathrm{ta}} = z] = h^*(z)'\beta_0.$$

The common representation map $h^* : \mathbb{R}^{50} \to \mathbb{R}^{10}$ is a fixed random ReLU network:

$$x^{(0)}(z) = z,$$
$$x^{(\ell)}(z) = \mathrm{ReLU}(x^{(\ell-1)}(z)W_\ell + b_\ell), \quad \ell = 1, \dots, 5,$$
$$h^*(z) = x^{(5)}(z),$$

with layer dimensions $(d_0, d_1, d_2, d_3, d_4, d_5) = (50, 64, 64, 64, 64, 10)$, $W_\ell \in \mathbb{R}^{d_{\ell-1} \times d_\ell}$, and $b_\ell \in \mathbb{R}^{d_\ell}$. Each entry of W_ℓ is drawn from $N(0, 1/d_{\ell-1})$ and each entry of b_ℓ is drawn from $N(0, 1)$ once per replication and then held fixed.

The source sample is generated from

$$Z_j^{\mathrm{so}} \sim N(0, I_{50}),$$
$$S_j = h^*(Z_j^{\mathrm{so}})'B + \varepsilon_j^{\mathrm{so}},$$
$$\varepsilon_j^{\mathrm{so}} \sim N(0, \sigma_{\mathrm{so}}^2 I_T),$$

where $T = 100$ and $B = UA$ with $U \in \mathbb{R}^{10 \times 5}$ having orthonormal columns and $A \in \mathbb{R}^{5 \times 100}$ Gaussian. Hence, the source task only identifies the five-dimensional subspace $\mathrm{span}(U) \subset \mathbb{R}^{10}$.

To compare transferable and non-transferable designs, let $a_\beta, a_\delta \in \mathbb{R}^5$ be Gaussian vectors and let $q \in \mathbb{R}^{10}$ satisfy $q \perp \mathrm{span}(U)$ and $\|q\|_2 = 1$. The in-span design sets

$$\beta_{\mathrm{in}} = \frac{Ua_\beta}{\|Ua_\beta\|_2}, \quad \delta_{\mathrm{in}} = \frac{Ua_\delta}{\|Ua_\delta\|_2},$$

whereas the out-of-span design sets

$$\beta_{\text{out}} = 2q, \quad \delta_{\text{out}} = 2q.$$

I.2 Estimation

For each replication, the source stage estimates a five-layer ReLU representation with the same architecture as the true h^* using the mean squared error on the source sample. The reported run uses $R = 2000$ replications, target sample size $n = 500$, source sample size $m = 2000$, $T = 100$ source outputs, $\sigma_{\text{so}} = \sigma_U = 1$, $\sigma_V = 0.5$, five-fold cross-fitting, 20 source-training epochs, batch size 128, and learning rate 10^{-3}.

Given the learned embedding $\check{h}$, each cross-fitting split estimates linear projections of Y_i, D_i, and $Y_i - \theta_0 D_i$ on $\check{h}(Z_i)$:

$$\begin{aligned}
\widehat{\ell}^{(-k)}(z) &= \widehat{a}_\ell^{(-k)} + \widehat{b}_\ell^{(-k)\prime}\check{h}(z), \\
\widehat{\alpha}^{(-k)}(z) &= \widehat{a}_\alpha^{(-k)} + \widehat{b}_\alpha^{(-k)\prime}\check{h}(z), \\
\widehat{\gamma}^{(-k)}(z) &= \widehat{a}_\gamma^{(-k)} + \widehat{b}_\gamma^{(-k)\prime}\check{h}(z).
\end{aligned}$$

The partially linear DML estimator is then

$$\begin{aligned}
\widetilde{Y}_i &= Y_i - \widehat{\ell}^{(-k(i))}(Z_i), \\
\widetilde{D}_i &= D_i - \widehat{\alpha}^{(-k(i))}(Z_i), \\
\widehat{\theta} &= \frac{\sum_{i=1}^n \widetilde{D}_i \widetilde{Y}_i}{\sum_{i=1}^n \widetilde{D}_i^2}.
\end{aligned}$$

The estimator $\widehat{\gamma}^{(-k)}$ is not used in the final ratio; it is recorded only to measure the prediction error for the nuisance function γ_0.

I.3 Monte Carlo Results

Table 1 shows that the in-span design is nearly centered at the truth, while the out-of-span design produces a sizable upward shift in the distribution of $\widehat{\theta}$. The mean of $\widehat{\theta}$ increases from 1.005 in the in-span design to 1.197 in the out-of-span design; thus, the difference in Monte Carlo means is approximately 0.191. This shift is accompanied by much larger nuisance prediction errors in the out-of-span design, as shown in Table 2. In particular, the mean squared prediction error for γ_0 increases from 0.110 to 10.579, and the corresponding error for α_0 increases from 0.044 to 2.677. These patterns are consistent with the theory. When

the target nuisance functions contain a component outside the source-identified span, transfer learning does not recover that direction accurately enough for the downstream DML step.

Case	Mean	SD	Median
In-span	1.005	0.212	1.004
Out-of-span	1.197	0.228	1.190

Table 1: Empirical distribution of $\widehat{\theta}$ across 2000 replications.

Case	Mean PE for γ_0	SD PE for γ_0	Mean PE for α_0	SD PE for α_0
In-span	0.110	0.887	0.044	0.372
Out-of-span	10.579	464.783	2.677	115.744

Table 2: Cross-fitted nuisance prediction errors from the saved Monte Carlo run.