Econometrics with Pre-Trained Embeddings for Unstructured Data

Yuya Shimizu*

This version: September 27, 2026
(First 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, pre-trained models are typically trained on different datasets and for different tasks, making it unclear when they can be used reliably for the target task. Second, the embedding function is subject to an identification problem, complicating the analysis of its estimation error and the effect of that error on the target task. We provide sufficient conditions to overcome these difficulties. A key condition, which we call transferability, governs the convergence rate we derive. This rate depends on three components: target estimation error with the pre-trained embeddings treated as standard covariates, source estimation error adjusted for the strength of transferability, and approximation error of the target model class using embeddings as inputs. We also extend this analysis for high-dimensional embeddings. To assess transferability, we develop a computationally feasible bootstrap test that does not require re-estimating the embeddings and nuisance functions. Our theory applies to a wide range of double machine learning applications, including partially linear regression with unstructured controls, price elasticity estimation in demand models accounting for product quality measured by images and text, missing-data imputation using unstructured data, and average treatment effect estimation with unstructured confounders. As an empirical application, we estimate the labor supply elasticity on Amazon Mechanical Turk, an online labor market platform, using job-description embeddings as controls.

*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, Hidehiko Ichimura, Hiroaki Kaido, Tetsuya Kaji, Hiroyuki Kasahara, Lorenzo Magnolfi, Ashesh Rambachan, Kensuke Sakamoto, Suproteem Sarkar, Jing Tao, Keyon Vafa, Kenneth West, Kohei Yata, and seminar participants at Wisconsin, ESIF-AIML, and the AIE Summer Conference for helpful comments and suggestions. I thank the authors of Ahrens et al. (2026), especially Achim Ahrens, for sharing their code and insights. This work is supported by the Richard E. Stockwell Dissertation Fellowship, the Richard A. Meese Dissertation Fellowship, and the summer research fellowship from Wisconsin.

1 Introduction

Empirical economists increasingly use pretrained deep-learning models to extract embeddings from images and text and then include these embeddings as covariates in downstream economic analyses.[1] Embedding functions map unstructured inputs into vectors in a smaller but potentially high-dimensional latent space that preserve relevant information from the original data. These models, however, are generally trained on a source dataset to perform a source task, both of which typically differ from their counterparts in the target economic application. It therefore remains unclear when the resulting embeddings can be used reliably in the target task. From an econometric perspective, this workflow constitutes a two-step estimation problem (Pagan, 1984; Murphy and Topel, 1985). It also raises an identification problem: embedding functions are generally not uniquely identified, complicating the analysis of their convergence rates and downstream effects. This paper provides sufficient conditions under which this workflow yields valid inference in econometric analyses and a computationally feasible bootstrap test to assess a key condition, which we call transferability.

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 can provide valuable information 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

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

[2] The following is an incomplete list of applications that use embeddings as covariates. Jean, Burke, Xie, Alampay Davis, Lobell, and Ermon (2016) use satellite image embeddings to predict local poverty status. Avivi (2025) uses patent-text embeddings to control for application quality when studying gender bias in patent examination. Dube, Jacobs, Naidu, and Suri (2020) estimate the elasticity of task vacancy-filling

training 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 best-approximating 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

speed with respect to rewards, using task-description embeddings as controls. Bach, Chernozhukov, Klaassen, Spindler, Teichert-Kluge, and Vijaykumar (2024); Bajari et al. (2025); 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. Related applied work uses product embeddings not primarily as controls, but to measure similarity in product space (Han and Lee, 2025; Magnolfi, McClure, and Sorensen, 2025). In macroeconomics, Casella, Fernández-Villaverde, Hansen, Oishi, and Shin (2026) propose a framework to incorporate textual information into structural macroeconomic models using topic models and text embeddings.

[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).

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 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.

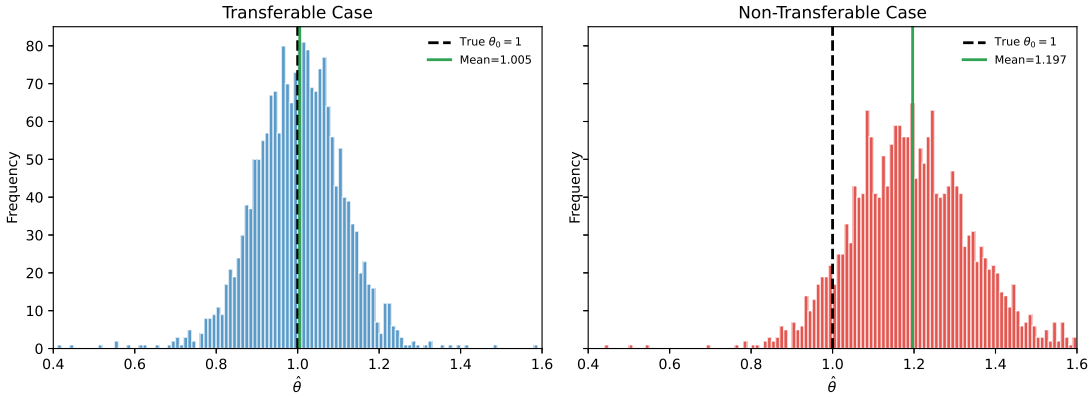


Figure 1: Histogram of $\widehat{\theta}$ for the in-span and out-of-span designs.

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 transferability 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.

We also develop a computationally feasible multiplier bootstrap test for transferability. The test compares DML estimators across embeddings pre-trained for the same source task using different pre-training procedures, such as different random seeds for the initial values. It computes critical values from weighted sums of estimated influence functions; thus, bootstrap draws do not require re-estimating embeddings or nuisance functions. Under the regularity conditions, rejection provides evidence against transferability.

As an empirical illustration, we apply the framework to Amazon Mechanical Turk, an online labor market platform, using the data of Dube et al. (2020). We use pre-trained embeddings of job descriptions as controls to estimate the elasticity of task-filling speed with respect to posted rewards. We compare alternative text representations and nuisance learners and use the bootstrap test to assess robustness to embeddings trained with different random seeds. The empirical application motivates a new empirical practice: assessing and reporting the robustness of DML estimates across embeddings pre-trained for the same source task with different random seeds.

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); Battaglia, Christensen, Hansen, and Sacher (2025); Carlson and Dell (2025); Duan and Pelger (2026); Ludwig, Mullainathan, and Rambachan (2026); Zhang, Xue, Yu, and Tan (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); Carlson (2025); Modarressi, Spiess, and Venugopal (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

[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).

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 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 best predictor 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 an $n^{-1/4}$ 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.

[5]Such reparameterization is not available for the general DML setup we consider in this paper, which includes Examples 2.1 to 2.4.

Structure of the Paper Section 2 introduces the DML framework, motivating applications, and the transfer learning setup. Section 3 develops the general convergence theory and characterizes task diversity. Section 4 derives convergence rates under lower-level conditions, including results for high-dimensional embeddings. Section 5 develops the multiplier bootstrap test for transferability. Section 6 presents simulation results, and Section 7 applies the framework to the Amazon Mechanical Turk data. Section 8 concludes. The supplementary appendix contains proofs, additional theoretical results, an extension to multimodal transfer learning, and further simulation and empirical results.

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. Let $\widehat{f}_\gamma$ and $\widehat{f}_\alpha$ denote the estimated target heads. For each nuisance component $\diamond \in \{\gamma, \alpha\}$, let $\Lambda_\diamond$ be the identity map when the target head estimates a conditional mean by squared loss, and let $\Lambda_\diamond(u) = \Lambda(u) := 1/(1 + \exp(-u))$ when the head estimates log odds by binary logistic loss. The nuisance estimators entering the DML score are

$$\widehat{\gamma} = \Lambda_\gamma \circ \widehat{f}_\gamma \circ \breve{h}, \qquad \widehat{\alpha} = \Lambda_\alpha \circ \widehat{f}_\alpha \circ \breve{h}.$$

Here, $\Lambda_\gamma \circ \widehat{f}_\gamma \circ \breve{h}$ is the composite function defined as $\Lambda_\gamma \circ \widehat{f}_\gamma \circ \breve{h}(z) = \Lambda_\gamma(\widehat{f}_\gamma(\breve{h}(z)))$. For vector-valued nuisance functions, the prediction maps are applied componentwise. The DML estimator $\widehat{\theta}$ is the solution of the sample moment equation

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

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) a sufficiently fast convergence rate of the nuisance parameters:[7]

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

where $\|\cdot\|_{P,2}$ is the $L^2(P)$ norm for the distribution of W_i, and the o_P notation is with respect

[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 Assumptions 3.3(b) and 3.4(c) in Chernozhukov et al. (2018) and the discussion after that.

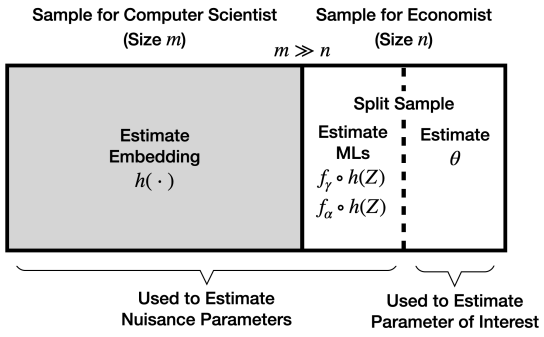


Figure 2: Sample Splitting with Transfer Learning

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 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 2.1. 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 2.1 (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

$$\widehat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\alpha}(Z_i))^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\alpha}(Z_i))\,(Y_i - \widehat{\gamma}(Z_i))\right). \tag{4}$$

Example 2.2 (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 coefficient. 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 2.1 to estimate the price coefficient 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 2.3 (Imputation with Unstructured Data)**.** *Suppose that an economist observes a sample of size n, $\{(Y_i, D_iX_i, D_i, Z_i)\}_{i=1}^n$, 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 fully 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 the moments involving X_i. We impose the missing at random assumption, $(X_i, Y_i) \perp\!\!\!\perp D_i \mid Z_i$ (Rubin, 1976). Let $\alpha^*(Z_i) = P[D_i = 1 \mid Z_i]$ be the propensity score and assume that $\alpha^*(Z_i)$ is bounded away from zero. The parameter θ^* is estimated by solving the Neyman orthogonal moment equation $\mathbb{E}[\psi(W_i; \theta, \gamma^*, \alpha^*)] = 0$, where $W_i = (Y_i, D_iX_i, D_i, Z_i)$,*

[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 nuisance parameters (κ^*, α^*). Chernozhukov et al. (2018) show that both scores are Neyman orthogonal.

$\psi(W_i;\theta,\gamma,\alpha) = (D_i/\alpha(Z_i))(Y_iX_i - X_i^2\theta - (\gamma_1(Z_i) - \gamma_2(Z_i)\theta)) + \gamma_1(Z_i) - \gamma_2(Z_i)\theta$, $\gamma_1^*(Z_i) = \mathbb{E}[X_iY_i|Z_i]$, $\gamma_2^*(Z_i) = \mathbb{E}[X_i^2|Z_i]$, *and* $\gamma^* = (\gamma_1^*, \gamma_2^*)$. *When the propensity is trained by binary logistic loss, the score uses* $\widehat{\alpha} = \Lambda \circ \widehat{f}_\alpha \circ \breve{h}$ *on the probability scale.*

Example 2.4 (Average Treatment Effect with Unstructured Confounders)**.** *Let* $D_i \in \{0,1\}$ *denote treatment,* $Y_i(d)$ *the potential outcome under treatment status* $d \in \{0,1\}$*, and* Z_i *the unstructured or ultra-high-dimensional confounders. The average treatment effect (ATE) is* $\theta^* = \mathbb{E}[Y_i(1) - Y_i(0)]$*. Assume that the observed outcome satisfies* $Y_i = D_iY_i(1) + (1-D_i)Y_i(0)$*, and impose unconfoundedness,* $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid Z_i$ *(Rosenbaum and Rubin, 1983). Let* $\alpha^*(Z_i) = P[D_i = 1 \mid Z_i]$ *be the propensity score and assume the overlap condition:* $\varepsilon \le \alpha^*(Z_i) \le 1-\varepsilon$ *almost surely for some fixed* $\varepsilon \in (0, 1/2)$*. Let* $\gamma^*(d, Z_i) = \mathbb{E}[Y_i \mid D_i = d, Z_i] = \mathbb{E}[Y_i(d) \mid Z_i]$*. The ATE* $\theta^* = \mathbb{E}[\gamma^*(1, Z_i) - \gamma^*(0, Z_i)]$ *is estimated using DML with the Neyman orthogonal moment equation* $\mathbb{E}[\psi(W_i;\theta,\gamma^*,\alpha^*)] = 0$*, where* $W_i = (Y_i, D_i, Z_i)$ *and* $\psi(W_i;\theta,\gamma,\alpha) = (D_i(Y_i - \gamma(1,Z_i))/(\alpha(Z_i)) - (1-D_i)(Y_i - \gamma(0,Z_i))/(1-\alpha(Z_i))) + (\gamma(1,Z_i) - \gamma(0,Z_i) - \theta)$*. If the propensity is trained by binary logistic loss, its fitted log odds are converted to* $\widehat{\alpha} = \Lambda \circ \widehat{f}_\alpha \circ \breve{h}$ *before evaluating this score.*

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 best-approximating embedding function h_m (Definition 3.1). 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 $Y \in \mathcal{Y}$

[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.

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:

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

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 VGG19's final hidden fully connected layer after activation) 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 4.1 and 4.2.

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 2.5 (DNN Embeddings for Ultra-High-Dimensional Data)**.** *Consider the setting*

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

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.

An autoencoder is a special case of the DNN in Example 2.5, 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 2.6 (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 2.5. The embedding function h is the output of the final hidden fully connected layer after activation in VGG19 and the flattened output of the final average-pooling layer in ResNet50. In both cases, the remaining classifier is affine, and the source head g returns its logit contrasts relative to a baseline class, as in Section 2.3. The covariate Z is an image, and the CNN can extract high-dimensional embeddings of dimension 4,096 for VGG19 and 2,048 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 2.7 (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 2.5. BERT is pre-trained on two tasks: masked language modeling (MLM) and next sentence prediction (NSP). The last transformer block returns a contextualized state for each token in the input text. Masked language modeling predicts the original token at randomly selected positions, including selected tokens that remain unchanged in the corrupted input. Next sentence prediction uses the state of the special token* `[CLS]`

[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.

to predict whether the second sentence follows the first, in which the special token `[SEP]` *separates the sentences. Both tasks use cross-entropy loss, with approximately 30,000 token labels for masked language modeling and two labels for next sentence prediction. For a text-level embedding of dimension 1,024 in BERT LARGE, we consider either the transformed* `[CLS]` *state or the mean of the transformed MLM states over positions eligible for token prediction, excluding* `[CLS]`*,* `[SEP]`*, and* `[PAD]`*.*

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

Example 2.8 (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 section and the next present our main theoretical results on transfer learning for DML. For this purpose, we first introduce the best predictors for the source and target tasks. Suppose that the source and target tasks admit best predictors satisfying

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

and

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

respectively. For example, suppose that the source loss function ℓ_{so} is the multinomial logit loss, $P_{\mathrm{so}}(S = t \mid Z) > 0$ 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 $a_{\mathrm{so},T+1}^*(Z) = 0$ and suitable moment conditions, the best source predictor is almost surely unique and given by the log-odds of the class probabilities, $a_{\mathrm{so},t}^*(Z) = \log(P_{\mathrm{so}}(S = t \mid Z)/P_{\mathrm{so}}(S = T+1 \mid Z))$ for $t = 1, \ldots, T$. For the target task, if the loss is the squared loss $\ell_{\mathrm{ta}}(a_{\mathrm{ta}}(Z), Y) = (1/2) \times (Y - a_{\mathrm{ta}}(Z))^2$ and suitable moment conditions are satisfied, the best predictor is almost surely unique and given by the conditional expectation of the outcome as $a_{\mathrm{ta}}^*(Z) = \mathbb{E}_{\mathrm{ta}}[Y \mid Z]$, and if the loss is the binary logit loss $\ell_{\mathrm{ta}}(a_{\mathrm{ta}}(Z), Y) = -Y \log(\exp(a_{\mathrm{ta}}(Z))/(1+\exp(a_{\mathrm{ta}}(Z)))) - (1-Y)\log(1/(1+\exp(a_{\mathrm{ta}}(Z))))$ and suitable moment conditions are satisfied, the best predictor is almost surely unique and given by the log-odds of the conditional probability as $a_{\mathrm{ta}}^*(Z) = \log(P_{\mathrm{ta}}(Y = 1 \mid Z)/P_{\mathrm{ta}}(Y =$

$0 \mid Z)$). In DML applications, these two loss functions are the most common for the target task.[14]

In this section, we develop a general convergence rate theory for the target model $\widehat{f} \circ \check{h}$ by relating it to the source task performance of the pre-trained model $\check{g} \circ \check{h}$ relative to the best predictor a^*_{so}. In Section 3.1, we introduce two key concepts: a best-approximating 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 Best-Approximating Embedding Function

We first introduce the concept of a best-approximating 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 $\mathcal{F}_n^{\text{sub}}$ 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 3.1 (Approximation Error)**.** For sequences of functions $(g_m, h_m) \in \mathcal{R}_m$ and $f_n \in \mathcal{F}_n^{\text{sub}}(h_m)$, we define the approximation errors for the source and target tasks as

$$\|g_m \circ h_m - a^*_{\text{so}}\|_{P_{\text{so}},2} =: \epsilon_{\text{so},m}, \tag{10}$$

$$\|f_n \circ h_m - a^*_{\text{ta}}\|_{P_{\text{ta}},2} =: \epsilon_{\text{ta},n}. \tag{11}$$

Remark 3.1. 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 best 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 best embedding function and g^* and f^* as the best source and target predictors, respectively. Although it is common to assume the existence of the best embedding function in the transfer

[14] For binary logistic loss, the probability nuisance is $\Lambda \circ a^*_{\text{ta}}$, and its estimator is $\Lambda \circ \widehat{f} \circ \check{h}$. Since $0 < \Lambda'(u) = \Lambda(u)(1 - \Lambda(u)) \leq 1/4$ for every $u \in \mathbb{R}$,

$$\left\|\Lambda \circ \widehat{f} \circ \check{h} - \Lambda \circ a^*_{\text{ta}}\right\|_{P_{\text{ta}},2} \leq \frac{1}{4} \left\|\widehat{f} \circ \check{h} - a^*_{\text{ta}}\right\|_{P_{\text{ta}},2}. \tag{9}$$

Thus, the convergence rates below for the fitted log odds also bound the probability-scale nuisance error required by DML.

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 best embedding function, but the source and target models can be well approximated by some embedding function. ■

Under the assumptions of Theorem 3.1, target prediction consistency follows when the target estimation error, the source estimation error adjusted for transferability, and the target approximation error all decay to zero. The approximation errors depend on the richness of the function classes $\mathcal{G}_m$ and $\mathcal{F}_n^{\text{sub}}$ and how well h_m captures the common features of the source and target tasks. Suppose that both tasks use squared loss, $\mathcal{G}_m$ and $\mathcal{F}_n^{\text{sub}}$ are unrestricted linear head classes, and the outcomes and coordinates of h_m are square-integrable under their respective task distributions. Then $g_m \circ h_m$ and $f_n \circ h_m$ are the $L^2(P_{\text{so}})$ and $L^2(P_{\text{ta}})$ projections of a_{so}^* and a_{ta}^* onto the span of the constant and the coordinates of h_m, respectively, with the source projection interpreted coordinatewise. This parallels series regression (Newey, 1997), with learned embeddings replacing fixed basis functions. Note that constrained or nonlinear head classes need not yield this projection.

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 3.2 for another example of partial identification of the embedding function for middle hidden layers

of a deep neural network.

Definition 3.2 (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

$$\mathcal{H}_{\text{so},m}(\rho_{\text{so},m}) \subseteq \mathcal{H}_{\text{ta},m}(\rho_{\text{ta},n}). \tag{12}$$

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 3.3 (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 $f_n^{\bullet} \in \arg\min_{f \in \mathcal{F}_n^{\text{sub}}} \|f \circ \check{h} - f_n \circ h_m\|_{P_{\text{ta}},2}$ be the best approximation of the population target model $f_n \circ h_m$ by the estimated embedding function $\check{h}$. Formally, $f_n^{\bullet}$ depends on $\check{h}$, but we suppress the dependence for notational simplicity.

Assumption 3.1 (Convergence Rate of Models)**.**

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

$$\left\| \check{g} \circ \check{h} - g_m \circ h_m \right\|_{P_{\text{so}},2} = O_{P_{\text{so}}}(\delta_{\text{so},m}). \tag{13}$$

(ii) (Target Model given $\check{h}$) For every sequence of estimated embedding functions $\check{h} \in \mathcal{H}_m$ satisfying (13), there exists some sequence $r_{\mathrm{ta},n} \geq 0$ such that

$$\left\| \widehat{f} \circ \check{h} - f_n^{\bullet} \circ \check{h} \right\|_{P_{\mathrm{ta},2}} = O_P(r_{\mathrm{ta},n}). \tag{14}$$

The first part of this assumption states that the pre-trained model $\check{g} \circ \check{h}$ converges to the population source model $g_m \circ h_m$ at some rate $\delta_{\mathrm{so},m}$. The second part states that the target model $\widehat{f} \circ \check{h}$ converges to the population target model $f_n^{\bullet} \circ \check{h}$ at some rate $r_{\mathrm{ta},n}$ 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 3.2 (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 3.1.*

Assumption 3.2 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 3.1 (Convergence Rate of Target Model)**.** *Suppose that Assumptions 3.1 and 3.2 hold. Then, for every sequence of estimated embedding functions $\check{h} \in \mathcal{H}_m$ satisfying (13), we have*

$$\left\| \widehat{f} \circ \check{h} - a_{\mathrm{ta}}^* \right\|_{P_{\mathrm{ta},2}} = O_P\left(r_{\mathrm{ta},n} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n} \right). \tag{15}$$

This theorem states that the convergence rate of the target model $\widehat{f} \circ \check{h}$ to the best predictor 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 a best-approximating 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. In particular, under the theorem's assumptions, $r_{\text{ta},n} + \delta_{\text{so},m}/\nu_n + \epsilon_{\text{ta},n} = o(1)$ suffices for consistency of the target predictor.

Remark 3.2. 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 best-approximating 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.[15] Second, applied work often ignores nonidentification of h_m and the associated task diversity condition. Without considering the nonidentification issue, the convergence point of machine learning models in the target task can depend on a specific optimization algorithm including the choice of initialization and the order of observations in the optimization steps used by the computer scientist to estimate the pre-trained model in the source task, which is not desirable for the economist.

The ν_n-transferability condition ensures that, for any nonnegative sequence $\rho_m = O(\delta_{\text{so},m})$ and all sufficiently large n, the source task near-identified set $\mathcal{H}_{\text{so},m}(\rho_m)$ is included in the target task near-identified set $\mathcal{H}_{\text{ta},m}(\rho_m/\nu_n)$. While the sequence $\check{h}$ does not converge in probability to one specific function due to nonidentification, as $\check{h}$ eventually belongs to the source task near-identified set $\mathcal{H}_{\text{so},m}(\rho_m)$ with high probability, the ν_n-transferability condition ensures that it also eventually belongs to the target task near-identified set $\mathcal{H}_{\text{ta},m}(\rho_m/\nu_n)$ with high probability (Figure 3). Together with Assumption 3.1, this ensures that the target model $\widehat{f} \circ \check{h}$ attains the rate in Theorem 3.1 for every pre-trained embedding sequence satisfying (13), regardless of the optimization algorithm used to obtain it. ■

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.

[15] Our rate is pointwise with respect to the sequence of estimated embedding functions $\check{h} \in \mathcal{H}_m$ satisfying (13). 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.

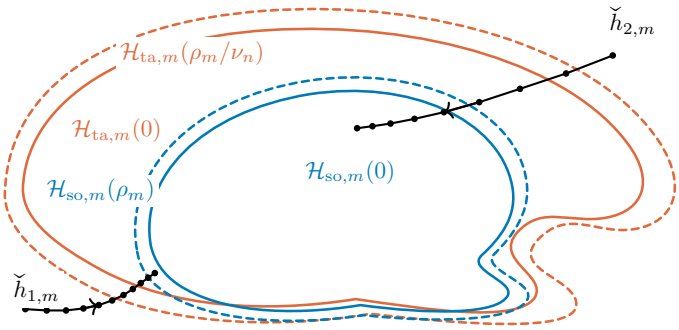


Figure 3: Transferability and embedding nonidentification.

Notes: The figure illustrates the concept of transferability and embedding nonidentification. The black dots represent the sequences $\{\check{h}_{1,m}\}_m$ and $\{\check{h}_{2,m}\}_m$ converging to two different functions, the blue sets represent the source task identified set (solid line) and near-identified set (dashed line), and the orange sets represent the target task identified set (solid line) and near-identified set (dashed line).

Theorem 3.2 (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 3.3. The ν_n-transferability condition in Assumption 3.2 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 (13). 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 (13), 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 3.2. ■

3.3 Closer Look at Task Diversity

In this subsection, we discuss sufficient and necessary conditions for task diversity in Definition 3.2.

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 3.3 (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 a common dominating measure.*

Assumption 3.3 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 3.3 (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 3.3, 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 3.3, 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. Holding the remaining quantities fixed, a smaller L_Γ gives a tighter sufficient bound.

In the next example, we verify the sufficient condition in Theorem 3.3 when both the computer scientist and the economist use linear models.

Example 3.1 (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, $L_\Gamma = \|\beta'_{\mathrm{ta},n} B^+_{\mathrm{so},m}\|_2 \le \|\beta_{\mathrm{ta},n}\|_2 / \sigma^+_{\min}(B_{\mathrm{so},m})$ in both cases.[17] Thus, by Theorem 3.3, 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 3.1 shows that, when the target class contains all linear heads with intercepts, the row-span condition is sufficient for transferability with $\varrho_n = 0$. Under the regularity conditions in Theorem 3.4, 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

[16] More generally, by allowing some approximation error $\varrho_n \ge 0$, we can instead assume that there exists $b \in \mathbb{R}^T$ such that $\|(\beta'_{\mathrm{ta},n} - b' B_{\mathrm{so},m}) h_m(Z)\|_{P_{\mathrm{ta}},2} \le \varrho_n$. 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 E.16 and E.17.

($T \geq K$). The next corollary states sufficient conditions for general Lipschitz target models, which include the linear target case as a special case.

Corollary 3.1 (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 $g_m \circ h_m = \alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m$, where $\alpha_{\mathrm{so},m} \in \mathbb{R}^T$ and $B_{\mathrm{so},m} \in \mathbb{R}^{T\times K}$. Consider any Lipschitz target model class $\mathcal{F}_n^{\mathrm{sub}}$ with Lipschitz constant $L_{\mathcal{F}}$. Suppose that, for every $\rho_{\mathrm{so},m} \geq 0$, every $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$, and every corresponding best linear source head*

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

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

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

Assume Assumption 3.3 holds. If $B_{\mathrm{so},m}$ has full column rank, then g_m is $(\rho_{\mathrm{so},m}, \rho_{\mathrm{ta},n})$-diverse over f_n with respect to h_m for every $\rho_{\mathrm{so},m} \geq 0$, where

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

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 $\nu_n = \underline{w}_n \sigma_{\min}(B_{\mathrm{so},m})/L_{\mathcal{F}}$ is larger if the source task has a larger minimum singular value $\sigma_{\min}(B_{\mathrm{so},m})$, 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 $\underline{w}_n$. Figure 4 illustrates the minimum singular value of the source task coefficient matrix $B_{\mathrm{so},m}$, writing $\beta_t = \beta_{\mathrm{so},m,t}$ for $t = 1, 2, 3$.

(a) Diversified tasks ($\sigma_{\min}(B_{\text{so},m}) \approx 0.70711$) (b) Non-diversified tasks ($\sigma_{\min}(B_{\text{so},m}) \approx 0.05467$)

Figure 4: Minimum singular values of the source coefficient matrices $B_{\text{so},m} = (\beta_1, \beta_2, \beta_3)'$.

Notes: In both panels, $\beta_1 = (\cos 15°, \sin 15°, 0)'$ and $\beta_2 = (\cos 75°, \sin 75°, 0)'$. The third coefficient is $\beta_3 = (1, 1, 2)'/\sqrt{6}$ in panel (a) and $\beta_3 = (10, 10, 1)'/\sqrt{201}$ in panel (b). All coefficient vectors have unit norm, and both matrices have full rank. The coefficients in panel (b) are nearly linearly dependent. Holding $L_{\mathcal{F}}$ and $\underline{w}_n$ fixed, the larger minimum singular value in panel (a) yields a larger transferability measure ν_n.

3.3.2 Necessary Condition

The following theorems provide necessary conditions for task diversity. We start with the linear case as in Example 3.1. 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 3.4 (Necessary Condition for Linear Models)**.** *Suppose that the population source and target heads are linear as in Example 3.1, so that* $g_m \circ h_m = \alpha_{\text{so},m} + B_{\text{so},m} h_m$ *and* $f_n \circ h_m = \alpha_{\text{ta},n} + \beta'_{\text{ta},n} h_m$*. 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* $f \in \mathcal{F}_n^{\text{sub}}$ *is a linear target head* $v \mapsto \alpha + \beta' v$*, where* $\alpha \in \mathbb{R}$ *and* $\beta \in \mathbb{R}^K$*. Assume also that* $\mathbb{E}_{\text{ta}}[(1, h_m(Z)')'(1, h_m(Z)')]$ *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* $\beta'_{\text{ta},n} = b' B_{\text{so},m}$*, and there exists an affine function* $\Gamma(x) = \alpha_{\text{ta},n} + \beta'_{\text{ta},n} B^{+}_{\text{so},m}(x - \alpha_{\text{so},m})$ *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 of h in the proof of Theorem 3.4 is possible. Note that we do not put any restriction on the function class $\mathcal{F}_n^{\text{sub}}$.

Theorem 3.5 (Necessary Condition for Transferability in General Models)**.** *Assume that $f_n \circ h_m \in L^2(P_{\text{ta}})$, where P_{ta} in this norm denotes the target marginal distribution of Z. We also assume that there exists $h_0 \in \mathcal{H}_{\text{so},m}(0)$ such that $h_0 \neq h_m$ and $f \circ h_0 \in L^2(P_{\text{ta}})$ for every $f \in \mathcal{F}_n^{\text{sub}}$. 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

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

where the right-hand side is the P_{ta}-completion defined in Lemma E.2. 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 3.5 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 $\mathcal{H}_{\text{so},m}(0) \subseteq \mathcal{H}_{\text{ta},m}(0)$, 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 assumptions of Theorem 3.5, including the sigma-field inclusion, measurable factorization $f_n \circ h_m(Z) = \Gamma \circ g_m \circ h_m(Z)$, P_{ta}-a.s., is necessary for $(0,0)$-diversity. Lipschitz continuity of Γ and the target-class representation condition remain additional sufficient restrictions in Theorem 3.3; their necessity is not established by Theorem 3.5. 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 3.2 (Counterexample for Nonlinear Pre-training Model)**.** *Suppose that $h_m(Z) \in \mathbb{R}_+$ and both tasks use the mean squared loss. Suppose that $\mathcal{F}_n^{\text{sub}}$ is a set of linear models as in*

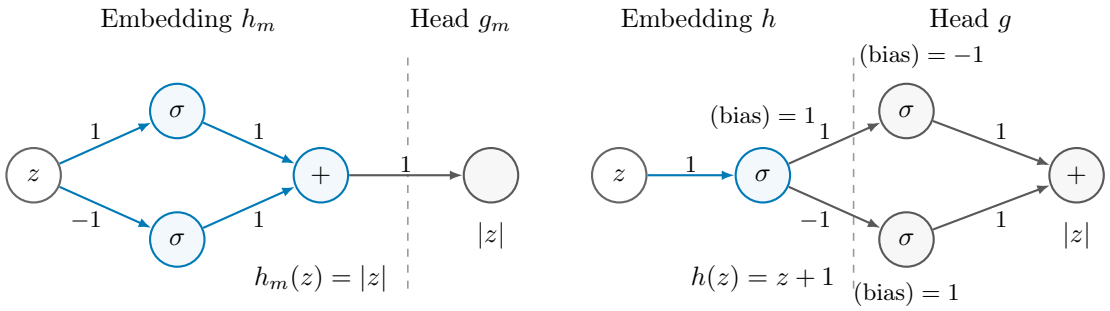
Example 3.1, but $\mathcal{G}_m$ includes any shallow neural networks with one hidden layer with ReLU activation of width 2. Suppose that $Z \sim$ Unif $[-1,1]$, $h_m(Z) = |Z|$, $f_n(\cdot) = g_m(\cdot) = (\cdot)$ are identity maps and $T = 1$. If we choose $h(Z) = Z + 1$, then we have

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

since $g \in \mathcal{G}_m$ can approximate the function $g_m \circ h_m(Z) = |Z|$, for example, by $g(x) = \text{ReLU}(x-1) + \text{ReLU}(-(x-1))$ (see Figure 5). On the other hand,

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

since $f \in \mathcal{F}_n^{\text{sub}}$ 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.



(a) Reference network

(b) Equivalent source network

Figure 5: Equivalent Neural Networks with Different Embeddings

Notes: Both networks satisfy $g_m(h_m(z)) = g(h(z)) = \sigma(z) + \sigma(-z) = |z|$ for $z \in [-1,1]$. $\sigma(u) = \max\{u, 0\}$ is ReLU and $+$ denotes a weighted sum. Edge labels give weights and biases are zero unless shown.

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 3.1. 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 4.1 (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 E.3 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 4.1 (Convergence Rate of Target Model)**.** *Let*

$$r_{\mathrm{ta},n} = \sqrt{\frac{\mathsf{V}(\mathcal{F}_n)\log(n)}{n}}. \tag{16}$$

For every estimated embedding function $\breve{h}$ *satisfying Assumption 3.1 (i), let* $\widehat{f} \in \mathcal{F}_n$ *be an estimated target model satisfying*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{ta}}(\widehat{f} \circ \breve{h}(Z_i), Y_i) \leq \min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{ta}}(f \circ \breve{h}(Z_i), Y_i) \\ + O_P(r_{\mathrm{ta},n}^2 + (\delta_{\mathrm{so},m}/\nu_n)^2 + \epsilon_{\mathrm{ta},n}^2).\end{aligned} \tag{17}$$

Suppose that $\|f\|_\infty \leq M_F < \infty$ *for every* $f \in \mathcal{F}_n$*. Then, under Assumption 3.1 (i), Assumptions 3.2, 3.3 and 4.1, we have*

$$\left\| \widehat{f} \circ \breve{h} - a_{\mathrm{ta}}^* \right\|_{P_{\mathrm{ta}},2} = O_P\left(r_{\mathrm{ta},n} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}\right). \tag{18}$$

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 best predictor a_{ta}^* is governed by the VC dimension of $\mathcal{F}_n$ and the target sample size n.

Remark 4.1 (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 4.1. 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 4.2 (Conditions on Source Task)**.**

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

$$\tau_m^2 \left\| g \circ h - a_{\mathrm{so}}^* \right\|_{P_{\mathrm{so}},2}^2 \lesssim \mathbb{E}_{\mathrm{so}}[\ell_{\mathrm{so}}(g \circ h(Z), S)] - \mathbb{E}_{\mathrm{so}}[\ell_{\mathrm{so}}(a_{\mathrm{so}}^*(Z), S)] \lesssim \left\| g \circ h - a_{\mathrm{so}}^* \right\|_{P_{\mathrm{so}},2}^2 \tag{19}$$

   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 E.4 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 3.1 (i) to the VC dimensions of the function classes $\mathcal{H}_m$ and $\mathcal{G}_m$.

Theorem 4.2 (Convergence Rate of Source Model)**.** *Let*

$$r_{\mathrm{so},m} = \sqrt{\frac{\sum_{k=1}^{K} \mathsf{V}(\mathcal{H}_{m,k}) \log\left(m L_{\mathcal{G},m} K\right) + \sum_{t=1}^{T} \mathsf{V}(\mathcal{G}_{m,t}) \log\left(mT\right)}{\tau_m^2 m}}. \tag{20}$$

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

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \ell_{\mathrm{so}}(\breve{g} \circ \breve{h}(Z_i), S_i) \leq & \min_{g \in \mathcal{G}_m, h \in \mathcal{H}_m} \frac{1}{m} \sum_{i=1}^{m} \ell_{\mathrm{so}}(g \circ h(Z_i), S_i) \\ & + O_{P_{\mathrm{so}}}(r_{\mathrm{so},m}^2 + \epsilon_{\mathrm{so},m}^2). \end{aligned} \tag{21}$$

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 4.2, we have*

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

$$\left\|\breve{g} \circ \breve{h} - a^*_{\mathrm{so}}\right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}\left(r_{\mathrm{so},m}/\tau_m + \epsilon_{\mathrm{so},m}/\tau_m\right). \tag{23}$$

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 best predictor 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 4.1 and 4.2 together imply that, under the transferability assumption, the convergence rate of the target model $\widehat{f} \circ \breve{h}$ to the best predictor 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. To use the primitive source bound, set $\delta_{\mathrm{so},m} = (r_{\mathrm{so},m} + \epsilon_{\mathrm{so},m})/\tau_m$ and impose the target theorem's assumptions at this rate. Under the assumptions of both theorems, $r_{\mathrm{ta},n} + (r_{\mathrm{so},m} + \epsilon_{\mathrm{so},m})/(\tau_m \nu_n) + \epsilon_{\mathrm{ta},n} = o(1)$ suffices for target prediction consistency. 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 ReLU DNNs with W_n parameters 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 best predictor 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 4.1 is directly applicable to derive the convergence rate of the target model $\widehat{f} \circ \breve{h}$ to the best predictor 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 4.1 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 3.1 and $h_m(Z) \in \mathbb{R}^K$ is the last hidden layer of the source model. In this case, we obtain a convergence rate for the projected estimated embedding toward an affine transformation of a best-approximating embedding function h_m. Recall that the population source model is $g_m \circ h_m(Z) = \alpha_{\mathrm{so},m} + B_{\mathrm{so},m} h_m(Z)$ for some vector $\alpha_{\mathrm{so},m} \in \mathbb{R}^T$, some matrix $B_{\mathrm{so},m} \in \mathbb{R}^{T \times K}$, and an embedding function $h_m(Z) \in \mathbb{R}^K$. The pre-trained source model is $\breve{g} \circ \breve{h}(Z) = \breve{\alpha}_{\mathrm{so}} + \breve{B}_{\mathrm{so}} \breve{h}(Z)$ for some vector $\breve{\alpha}_{\mathrm{so}} \in \mathbb{R}^T$, some matrix $\breve{B}_{\mathrm{so}} \in \mathbb{R}^{T \times K}$, and an embedding function $\breve{h}(Z) \in \mathbb{R}^K$ estimated from the source task. Define a transformed embedding function $\breve{h}^{\mathrm{proj}}$ as the projection of $\breve{h}$ onto the row space of $\breve{B}_{\mathrm{so}}$ as follows:

$$\breve{h}^{\mathrm{proj}} := \breve{B}^{+}_{\mathrm{so}} \breve{B}_{\mathrm{so}} \breve{h}. \tag{24}$$

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

Lemma 4.1 (Convergence of the projected embedding)**.** *Suppose that the population and pre-trained source heads are linear, as defined in Example 3.1, and that Assumption 3.1 (i)*

holds. If, for some deterministic sequence $\underline{\sigma}_m > 0$,

$$\|\breve{B}_{\text{so}}^{+}\|_{\text{sp}} = O_{P_{\text{so}}}(\underline{\sigma}_m^{-1}), \tag{25}$$

then

$$\left\|\breve{h}^{\text{proj}} - (\breve{q} + \breve{Q}h_m)\right\|_{P_{\text{so}},2} = O_{P_{\text{so}}}(\delta_{\text{so},m}/\underline{\sigma}_m). \tag{26}$$

In particular, the condition (25) 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 4.1 controls the distance between the projected embedding $\breve{h}^{\text{proj}}$ and $\breve{q} + \breve{Q}h_m$. The matrix $\breve{Q}$ describes the linear part of this relation for the component of $\breve{h}$ in the row space of $\breve{B}_{\text{so}}$. 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, when $B_{\text{so},m}$ or $\breve{B}_{\text{so}}$ is rank deficient, this affine relation applies to projected embeddings. Components in the null space of the corresponding source-head matrix remain unidentified by the source output, and source-equivalent embeddings

[20] The event $\{\text{rank}(\breve{B}_{\text{so}}) \geq 1\}$ is only used to ensure that the smallest positive 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. Target predictions are preserved if the target class contains $v \mapsto f(v + c)$ for each relevant head f and shift c. For an affine head $f(v) = \alpha + \beta'v$, this requires an intercept whose permitted range includes $\alpha + \beta'c$. Applying the target convergence results also requires the translated heads to satisfy their parameter and regularity bounds.

need not be affine transformations of one another.

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, $\mathcal{H}_{\mathrm{so},m}(0)$ includes invertible linear transformations of a best-approximating embedding function h_m. Theorem 3.2 suggests that, to ensure $\|\widehat{f}_{n,h_m^\dagger} \circ h_m^\dagger - a_{\mathrm{ta}}^*\|_{P_{\mathrm{ta}},2} = o_{P_{\mathrm{ta}}}(1)$ for any sequence $h_m^\dagger \in \mathcal{H}_{\mathrm{so},m}(0)$, it is necessary that $\mathcal{H}_{\mathrm{so},m}(0) \subseteq \mathcal{H}_{\mathrm{ta},m}(\rho_{\mathrm{ta},n})$ for some sequence $\rho_{\mathrm{ta},n} = o(1)$ 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 $f \in \mathcal{F}_n^{\mathrm{sub}}$ such that $\|f_n \circ h_m - f \circ (q + Qh_m)\|_{P_{\mathrm{ta}},2}$ 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 $\check{h}$ from the source task. To use LASSO, the existing theory typically requires an approximate-sparsity assumption. On the other hand, since $\mathcal{H}_{\mathrm{so},m}(0)$ includes invertible linear transformations of a best-approximating 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 $h \in \mathcal{H}_{\mathrm{so},m}(0)$. 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 (25) 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 transformation in Appendix D normalizes the positive singular values of a nonzero source-head matrix if one wants to ensure the conditions on the singular value bounds.

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

$$g_{\alpha,B} : v \mapsto \alpha + Bv. \tag{27}$$

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 E.17 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

$$\mathcal{H}^{\mathrm{proj}}_m := \left\{B^+Bh : h \in \mathcal{H}_m, (\alpha, B) \in \Theta_{\mathrm{so},m} \text{ for some } \alpha \in \mathbb{R}^T\right\}. \tag{28}$$

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

$$\widehat{f}^{\mathrm{ridge}}_n(v) := \widehat{\alpha}^{\mathrm{ridge}}_n + \widehat{\beta}^{\mathrm{ridge}'}_n v \in \underset{\alpha+\beta' v\in\mathcal{F}_n}{\arg\min} \left\{\frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{ta}}\left(\alpha + \beta'\breve{h}^{\mathrm{proj}}(Z_i), Y_i\right) + \lambda_n^2\|\beta\|_2^2\right\}. \tag{29}$$

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

Assumption 4.3 (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

$$\mathcal{N}_{h,n}(\lambda) := \mathrm{tr}\left(\Sigma_{h,n}(\Sigma_{h,n} + \lambda^2 I_K)^{-1}\right) = \sum_{j=1}^{K} \frac{\mu_j(\Sigma_{h,n})}{\mu_j(\Sigma_{h,n}) + \lambda^2}, \tag{30}$$

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

Theorem 4.3 (Convergence Rate of Ridge Target Model)**.** *Suppose that Assumption 3.1 (i) and Assumptions 3.3 and 4.3 hold. We also assume Assumption 4.1 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 (29). Then,*

$$\begin{aligned}
&\left\|\widehat{f}_n^{\mathrm{ridge}} \circ \breve{h}^{\mathrm{proj}} - a_{\mathrm{ta}}^*\right\|_{P_{\mathrm{ta}},2} \\
=&O_P\left(\sqrt{\frac{1 + \mathcal{N}_{h_m,n}(\lambda_n)}{n}} + \lambda_n + \frac{\delta_{\mathrm{so},m}/\underline{w}_n + (\delta_{\mathrm{so},m}/\underline{w}_n)^{1/2}\,\mathrm{tr}(\Sigma_{h_m,n})^{1/4}}{\sqrt{n}\lambda_n} + \frac{\delta_{\mathrm{so},m}}{\underline{w}_n} + \epsilon_{\mathrm{ta},n}\right).
\end{aligned} \tag{31}$$

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.

5 Test for Transferability

As discussed in Remark 3.2 and illustrated in Figure 3, ν_n-transferability enables accurate estimation of target nuisance functions using pre-trained embeddings even when the source task only partially identifies the embedding. In this section, we develop a formal test using an

[22] Another line of research in deep learning theory on convergence rates makes use of functional shape assumptions on the function of interest (Kohler and Langer, 2021; Bhattacharya, Fan, and Mukherjee, 2024) or data structure assumptions such as approximate manifold structure (Jiao, Shen, Lin, and Huang, 2023; Schulte et al., 2025). For a VC-based estimation term of the form $\sqrt{\mathsf{V}(\mathcal{F}_n)\log(n)/n}$, raw input dimension may grow provided the resulting complexity grows sufficiently slowly: $\mathsf{V}(\mathcal{F}_n)\log(n) = o(n)$ makes this term vanish, and $\mathsf{V}(\mathcal{F}_n)\log(n) = o(n^{1/2})$ makes it $o(n^{-1/4})$. The other approximation and transfer terms must satisfy their corresponding rate requirements.

implication of transferability: under the rate and regularity conditions below, DML estimators based on different embeddings in the source near-identified set should yield similar estimates of the target parameter. We propose a computationally feasible procedure based on the multiplier bootstrap and show its uniform validity under the null hypothesis and its uniform consistency under well-separated alternatives.

The test uses several embeddings pre-trained for the same source task, for example by training the same model architecture with different random seeds in optimization. We split the target sample into three parts, using one part to fit the nuisance functions for each embedding and the other two to estimate the parameter of interest. We compare estimators across embeddings and estimation samples to assess their sensitivity to the choice of pre-trained embedding. Using separate estimation samples prevents the leading sampling randomness from canceling in these comparisons. We reject the transferability null when the maximum standardized difference exceeds a critical value computed by a multiplier bootstrap, without refitting the nuisance functions or embeddings. Under the regularity conditions below, rejection provides evidence against transferability.

5.1 Multiplier Bootstrap

Consider the setup in Section 2.1, with a scalar parameter of interest $\theta^* \in \Theta \subset \mathbb{R}$ and a scalar Neyman orthogonal score $\psi(W; \theta, \gamma, \alpha)$ satisfying $\mathbb{E}_{\text{ta}}[\psi(W; \theta^*, \gamma^*, \alpha^*)] = 0$. The nuisance functions γ^* and α^* may each have a fixed number of components. In that case, the transferability and rate conditions below apply to each component. While we focus on scalar nuisance functions and scalar θ for simplicity, the results extend to vector-valued nuisance functions by replacing the scalar norm with an appropriate vector norm.

Suppose that $R \geq 2$ pre-trained models $(\breve{g}_j, \breve{h}_j) \in \mathcal{G}_m \times \mathcal{H}_m$, $j = 1, \ldots, R$, are available for the same source task. The number of models R is deterministic and may diverge as $n \to \infty$. The models may share a source sample and differ in their pre-training randomness. Their collection is independent of the target sample. Such pre-trained variants are available for both text and image embeddings.[23]

Fix the source distribution P_{so}, function classes $\mathcal{G}_m$ and $\mathcal{H}_m$, and sequences $(g_m, h_m) \in \mathcal{R}_m$ and $\delta_{\text{so},m}$. For each $\diamond \in \{\gamma, \alpha\}$ and each target distribution P_{ta}, let $a^*_\diamond$ denote the best target predictor on the scale of the training loss, so that $\diamond^* = \Lambda_\diamond \circ a^*_\diamond$ with the prediction maps defined in Section 2.1. Let $\mathcal{F}^{\text{sub}}_{\diamond,n}$ denote the corresponding target head class and let $f_{\diamond,n} \in \mathcal{F}^{\text{sub}}_{\diamond,n}(h_m)$ be a population target head. For positive sequences of transferability strengths $\nu_{\diamond,n}$, $\diamond \in \{\gamma, \alpha\}$, the null requires that g_m be $\nu_{\diamond,n}$-transferable to $f_{\diamond,n}$ with respect to h_m at rate $\delta_{\text{so},m}$, in the

[23] Examples include BERT, GPT-2, and Pythia for text, and ResNet-50, Swin-Tiny, and ViT for images.

sense of Definition 3.3. Equivalently,

$$H_{0,n}: \quad \sup_{h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})} \min_{f \in \mathcal{F}^{\mathrm{sub}}_{\diamond,n}} \|f \circ h - f_{\diamond,n} \circ h_m\|_{P_{\mathrm{ta}},2} \leq \rho_{\mathrm{so},m}/\nu_{\diamond,n}, \tag{32}$$

for each $\diamond \in \{\gamma, \alpha\}$ and every nonnegative sequence $\rho_{\mathrm{so},m} = O(\delta_{\mathrm{so},m})$, for all sufficiently large n. Here, $\mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$ is the source task near-identified set defined in Section 3.1.2. While the null restricts all embeddings in this set, the observed pre-trained embeddings provide a way to test an implication of this restriction.

For each n, let $\mathcal{P}_{0,\mathrm{ta},n}$ be a class of target distributions satisfying the null hypothesis $H_{0,n}$ for the fixed source problem P_{so}. Define the corresponding class of joint source and target laws by $\mathcal{P}_{0,n} = \{P_{\mathrm{so}} \times P_{\mathrm{ta}} : P_{\mathrm{ta}} \in \mathcal{P}_{0,\mathrm{ta},n}\}$. Target population objects and deterministic target rates may depend on P_{ta}, but this dependence is suppressed in the notation for simplicity.

In our test, we partition the target sample independently of the data into a nuisance-training sample $\mathcal{I}_3$ and two disjoint estimation samples $\mathcal{I}_1$ and $\mathcal{I}_2$. Let $n_s = |\mathcal{I}_s|$ for $s \in \{1, 2, 3\}$, with all three sample proportions converging to positive constants. For each $\diamond \in \{\gamma, \alpha\}$ and $j = 1, \ldots, R$, fit the target head $\widehat{f}_{\diamond,j}$ using $\mathcal{I}_3$ and embedding $\check{h}_j$. Write $\widehat{\gamma}_j = \Lambda_\gamma \circ \widehat{f}_{\gamma,j} \circ \check{h}_j$ and $\widehat{\alpha}_j = \Lambda_\alpha \circ \widehat{f}_{\alpha,j} \circ \check{h}_j$. For $j = 1, \ldots, R$ and $s \in \{1, 2\}$, define the sample moment by $\widehat{M}_{j,s}(\theta) = n_s^{-1} \sum_{i \in \mathcal{I}_s} \psi(W_i; \theta, \widehat{\gamma}_j, \widehat{\alpha}_j)$ and let $\widehat{\theta}_{j,s}$ solve $\widehat{M}_{j,s}(\widehat{\theta}_{j,s}) = 0$.

Algorithm 1 constructs the test statistic as the largest absolute standardized difference between estimators based on different embeddings and the two estimation samples $\mathcal{I}_1$ and $\mathcal{I}_2$. Let the population score derivative be

$$J = \frac{\partial}{\partial \theta} \mathbb{E}_{\mathrm{ta}}[\psi(W; \theta, \gamma^*, \alpha^*)]\Big|_{\theta=\theta^*}.$$

Define the population influence function and its variance by $\phi(W) = -J^{-1}\psi(W; \theta^*, \gamma^*, \alpha^*)$ and $V = \mathbb{E}_{\mathrm{ta}}[\phi(W)^2]$. Under $H_{0,n}$ in (32) and Assumption 5.1 below, the estimators have the expansion

$$\widehat{\theta}_{j,s} - \theta^* = \frac{1}{n_s} \sum_{i \in \mathcal{I}_s} \phi(W_i) + o_P(n^{-1/2}), \tag{33}$$

uniformly over $j = 1, \ldots, R$, $s \in \{1, 2\}$, and $P \in \mathcal{P}_{0,n}$. The centering at θ^* and the influence function do not depend on the embedding. The leading terms in (33) would therefore cancel if two estimators used the same estimation sample. Using separate estimation samples $\mathcal{I}_1$ and $\mathcal{I}_2$ avoids this cancellation and gives the leading difference positive variance $V(n_1^{-1} + n_2^{-1})$ conditional on the partition.

For each bootstrap draw, we approximate the sampling variation in these leading terms

using weighted sums of the estimated influence functions, with independent standard normal multipliers across observations. We use the same multiplier for an observation across all embeddings to preserve dependence across comparisons. The embeddings, fitted nuisance functions, and sample partition remain fixed throughout the bootstrap; only the multipliers vary. Thus, the bootstrap requires no re-estimation of the nuisance functions or embeddings.

5.2 Uniform Size Validity of the Test

We now state sufficient conditions for validity of this procedure uniformly over $P \in \mathcal{P}_{0,n}$. For $P = P_{\mathrm{so}} \times P_{\mathrm{ta}} \in \mathcal{P}_{0,n}$, let $\Pr_P$ denote probability under this law, including the randomness of both samples, the partition, pre-training randomness, and bootstrap draws when applicable. For the asymptotic statements below, set $\widehat{\phi}_{ij,s} = q_{ij,s} = 0$ whenever $\widehat{J}_{j,s} = 0$, and define $\widehat{\sigma}_{jk}$ by (36) using these extended weights. The algorithm still stops without returning a test decision on these events.

For each $\diamond \in \{\gamma, \alpha\}$, let $r_{\diamond,n}$ be a common rate for estimating the target head across embeddings, and let $\epsilon_{\diamond,n} = \|f_{\diamond,n} \circ h_m - a^*_\diamond\|_{P_{\mathrm{ta}},2}$ denote the approximation error of the population target model on the training scale. Since each $\Lambda_\diamond$ has Lipschitz constant at most one, these rates also bound the corresponding nuisance errors.[24] For each pre-trained embedding function $\breve{h}_j$, let

$$f^\bullet_{\diamond,j,n} \in \operatorname*{arg\,min}_{f \in \mathcal{F}^{\mathrm{sub}}_{\diamond,n}} \left\| f \circ \breve{h}_j - f_{\diamond,n} \circ h_m \right\|_{P_{\mathrm{ta}},2}$$

be a target head that best approximates $f_{\diamond,n} \circ h_m$ when paired with $\breve{h}_j$. We write the score estimation error at θ^* as $\Delta\psi_j(W) := \psi(W; \theta^*, \widehat{\gamma}_j, \widehat{\alpha}_j) - \psi(W; \theta^*, \gamma^*, \alpha^*)$.

Assumption 5.1 (Conditions for the Transferability Test)**.**

(i) (Sampling and Sample Splitting) *The number of pre-trained embedding functions $R \geq 2$ is deterministic and may diverge as $n \to \infty$. For each $P_{\mathrm{ta}} \in \mathcal{P}_{0,\mathrm{ta},n}$, the target observations are i.i.d. from P_{ta}. The pre-trained models and auxiliary pre-training randomness are jointly independent of the target observations. The partition $(\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3)$ is independent of all these quantities. Its sample sizes satisfy $n_1, n_2 \geq 2$ and $n_3 \geq 1$, with $n_s/n \xrightarrow{P} \pi_s \in (0,1)$ uniformly over $P \in \mathcal{P}_{0,n}$ for $s \in \{1,2,3\}$. The limits π_s do not depend on P. Each fitted nuisance function depends only on the pre-trained embedding functions, observations in $\mathcal{I}_3$, and auxiliary pre-training randomness.*

[24] See (9).

Algorithm 1: Multiplier bootstrap test of transferability for a DML parameter

Input: An i.i.d. target sample $\{W_i\}_{i=1}^n$; Neyman orthogonal score $\psi(W;\theta,\gamma,\alpha)$; pre-trained embeddings $\breve{h}_1,\dots,\breve{h}_R$; significance level $\mathsf{a}\in(0,1)$; number of bootstrap draws B.

1. **Split the target sample.** Randomly select a nuisance-training sample $\mathcal{I}_3$, then partition the remaining observations into $\mathcal{I}_1$ and $\mathcal{I}_2$. Draw the partition independently of the observed data and use it for every embedding. Let $n_s=|\mathcal{I}_s|$ for $s\in\{1,2\}$, with $n_1,n_2\geq 2$.

2. **Estimate nuisance functions on $\mathcal{I}_3$.** For each j, use only $\mathcal{I}_3$ to fit the target heads, including preprocessing and tuning. Set $\widehat{\gamma}_j=\Lambda_\gamma\circ\widehat{f}_{\gamma,j}\circ\breve{h}_j$ and $\widehat{\alpha}_j=\Lambda_\alpha\circ\widehat{f}_{\alpha,j}\circ\breve{h}_j$, converting fitted log odds to probabilities before evaluating the DML score.

3. **Compute estimators and influence functions on $\mathcal{I}_1$ and $\mathcal{I}_2$.** For each j and $s\in\{1,2\}$, solve the estimating equation with respect to θ:

$$\frac{1}{n_s}\sum_{i\in\mathcal{I}_s}\psi(W_i;\theta,\widehat{\gamma}_j,\widehat{\alpha}_j)=0,$$

and denote the solution by $\widehat{\theta}_{j,s}$. Compute the estimator of the derivative of the score with respect to θ:

$$\widehat{J}_{j,s}=\frac{1}{n_s}\sum_{i\in\mathcal{I}_s}\partial_\theta\psi(W_i;\widehat{\theta}_{j,s},\widehat{\gamma}_j,\widehat{\alpha}_j). \tag{34}$$

Require $\widehat{J}_{j,s}\neq 0$ for every j,s; otherwise, stop without returning a test decision. Compute

$$\widehat{\phi}_{ij,s}=-\widehat{J}_{j,s}^{-1}\psi(W_i;\widehat{\theta}_{j,s},\widehat{\gamma}_j,\widehat{\alpha}_j),\quad q_{ij,s}=\frac{\widehat{\phi}_{ij,s}}{\sqrt{n_s(n_s-1)}}. \tag{35}$$

Here, $\sum_{i\in\mathcal{I}_s}q_{ij,s}^2$ estimates the variance of $\widehat{\theta}_{j,s}$.

4. **Form the test statistic.** For every $1\leq j<k\leq R$, calculate

$$\widehat{\sigma}_{jk}^2=\sum_{i\in\mathcal{I}_1}q_{ij,1}^2+\sum_{i\in\mathcal{I}_2}q_{ik,2}^2. \tag{36}$$

Require every $\widehat{\sigma}_{jk}>0$; otherwise, stop without returning a test decision. Then compute

$$\mathcal{T}=\max_{1\leq j<k\leq R}\frac{|\widehat{\theta}_{j,1}-\widehat{\theta}_{k,2}|}{\widehat{\sigma}_{jk}}. \tag{37}$$

5. **Generate the bootstrap distribution.** For each $b=1,\dots,B$, draw independent $\xi_i^{(b)}\sim N(0,1)$ for $i\in\mathcal{I}_1\cup\mathcal{I}_2$, independently across draws and of the data. Use the same multiplier for observation i across all embeddings and compute bootstrap test statistics

$$\mathcal{T}_b^{(\xi)}=\max_{1\leq j<k\leq R}\frac{\left|\sum_{i\in\mathcal{I}_1}\xi_i^{(b)}q_{ij,1}-\sum_{i\in\mathcal{I}_2}\xi_i^{(b)}q_{ik,2}\right|}{\widehat{\sigma}_{jk}}. \tag{38}$$

6. **Return the decision and p-value.** Let $c_{1-\mathsf{a}}^{(\xi)}$ be the $(1-\mathsf{a})$-quantile of $\{\mathcal{T}_b^{(\xi)}\}_{b=1}^B$, using the inverse empirical distribution function. Reject the transferability null $H_{0,n}$ in (32) if $\mathcal{T}>c_{1-\mathsf{a}}^{(\xi)}$. Report the bootstrap p-value $B^{-1}\sum_{b=1}^B\mathbb{1}\{\mathcal{T}_b^{(\xi)}\geq\mathcal{T}\}$.

(ii) (Source and Target Convergence) *The convergence bounds in Assumption 3.1 hold uniformly over the embedding functions. Specifically, the source models satisfy*

$$\max_{1\le j\le R}\left\|\breve{g}_j\circ\breve{h}_j - g_m\circ h_m\right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}). \tag{39}$$

For each $\diamond\in\{\gamma,\alpha\}$, the target fits based on $\mathcal{I}_3$ satisfy

$$\max_{1\le j\le R}\left\|\widehat{f}_{\diamond,j}\circ\breve{h}_j - f^{\bullet}_{\diamond,j,n}\circ\breve{h}_j\right\|_{P_{\mathrm{ta}},2} = O_P(r_{\diamond,n}), \tag{40}$$

uniformly over $P\in\mathcal{P}_{0,n}$. The deterministic rates satisfy

$$\max_{\diamond\in\{\gamma,\alpha\}}(r_{\diamond,n}+\delta_{\mathrm{so},m}/\nu_{\diamond,n}+\epsilon_{\diamond,n}) = o(n^{-1/4}), \tag{41}$$

uniformly over $P_{\mathrm{ta}}\in\mathcal{P}_{0,\mathrm{ta},n}$.

(iii) (Neyman Orthogonality and Score Regularity) *For each $P_{\mathrm{ta}}\in\mathcal{P}_{0,\mathrm{ta},n}$, the parameter θ^* is an interior point of Θ and $\mathbb{E}_{\mathrm{ta}}[\psi(W;\theta^*,\gamma^*,\alpha^*)]=0$. There exist constants $0<\underline{J}\le\overline{J}<\infty$ and $0<\underline{V}\le\overline{V}<\infty$ such that $\underline{J}\le|J|\le\overline{J}$ and $\underline{V}\le V\le\overline{V}$ uniformly over $P_{\mathrm{ta}}\in\mathcal{P}_{0,\mathrm{ta},n}$. For some fixed $\zeta>0$, the population influence function satisfies the moment condition*

$$\sup_{n\ge1}\sup_{P_{\mathrm{ta}}\in\mathcal{P}_{0,\mathrm{ta},n}}\mathbb{E}_{\mathrm{ta}}\left[\left|\frac{\phi(W)}{\sqrt{V}}\right|^{2+\zeta}\right]<\infty. \tag{42}$$

For each n and P_{ta}, there is a deterministic set $\mathcal{N}_n=\mathcal{N}_n(P_{\mathrm{ta}})$ of nuisance pairs containing (γ^*,α^*) such that

$$\inf_{P\in\mathcal{P}_{0,n}}\Pr_P\left((\widehat{\gamma}_j,\widehat{\alpha}_j)\in\mathcal{N}_n \text{ for all } j=1,\dots,R\right)\longrightarrow 1. \tag{43}$$

For every $(\gamma,\alpha)\in\mathcal{N}_n$, the path $(\gamma^*+r(\gamma-\gamma^*),\alpha^*+r(\alpha-\alpha^*))$ remains in the domain of the function $\psi(W;\theta^*,\cdot,\cdot)$ for $r\in[0,1]$. The population moment along this path at θ^* is continuous on $[0,1]$ and twice continuously differentiable on $(0,1)$. The score satisfies Neyman orthogonality:

$$\frac{\partial}{\partial r}\mathbb{E}_{\mathrm{ta}}\left[\psi\left(W;\theta^*,\gamma^*+r(\gamma-\gamma^*),\alpha^*+r(\alpha-\alpha^*)\right)\right]\Big|_{r=0}=0. \tag{44}$$

The second derivative satisfies

$$\sup_{(\gamma,\alpha)\in\mathcal{N}_n}\sup_{r\in(0,1)}\left|\frac{\partial^2}{\partial r^2}\mathbb{E}_{\mathrm{ta}}\left[\psi\left(W;\theta^*,\gamma^*+r(\gamma-\gamma^*),\alpha^*+r(\alpha-\alpha^*)\right)\right]\right|=o(n^{-1/2}), \tag{45}$$

uniformly over $P_{\text{ta}} \in \mathcal{P}_{0,\text{ta},n}$. *The pointwise maximum of the score estimation errors satisfies*

$$\sqrt{\log R}\left\|\max_{1\le j\le R}|\Delta\psi_j(W)|\right\|_{P_{\text{ta}},2} = o_P(1), \tag{46}$$

uniformly over $P \in \mathcal{P}_{0,n}$.

(iv) (Estimation on Each Subsample) *The estimators satisfy* $\widehat{M}_{j,s}(\widehat{\theta}_{j,s}) = 0$ *almost surely for every* $j = 1,\dots,R$ *and* $s \in \{1,2\}$, *and* $\max_{1\le j\le R, s\in\{1,2\}} |\widehat{\theta}_{j,s} - \theta^*| = o_P(1)$ *uniformly over* $P \in \mathcal{P}_{0,n}$.

(v) (Influence-Function Estimation) *For each* n *and* $P \in \mathcal{P}_{0,n}$, *there exists a deterministic open interval* $\mathcal{U}_n(P)$ *containing* θ^*. *For every* $j = 1,\dots,R$, *the fitted score* $\psi(W;\theta,\widehat{\gamma}_j,\widehat{\alpha}_j)$ *is well defined and continuously differentiable in* θ *on* $\mathcal{U}_n(P)$, *almost surely. The intervals satisfy*

$$\inf_{P\in\mathcal{P}_{0,n}} \Pr_P\left(\widehat{\theta}_{j,s} \in \mathcal{U}_n(P) \text{ for all } j = 1,\dots,R,\ s\in\{1,2\}\right) \longrightarrow 1. \tag{47}$$

The sample derivatives satisfy

$$\max_{1\le j\le R, s\in\{1,2\}} \sup_{\theta\in\mathcal{U}_n(P)} \left|\frac{1}{n_s}\sum_{i\in\mathcal{I}_s} \partial_\theta\psi(W_i;\theta,\widehat{\gamma}_j,\widehat{\alpha}_j) - J\right| = o_P(1), \tag{48}$$

uniformly over $P \in \mathcal{P}_{0,n}$. *For* $q_{ij,s}$ *in (35), the variance estimators for the first embedding function satisfy* $n_s\sum_{i\in\mathcal{I}_s} q_{i1,s}^2 \xrightarrow{P} V$ *for* $s \in \{1,2\}$, *uniformly over* $P \in \mathcal{P}_{0,n}$. *Differences in* $q_{ij,s}$ *across embedding functions satisfy* $\sqrt{\log R}\max_{1\le j\le R, s\in\{1,2\}} \sqrt{n_s\sum_{i\in\mathcal{I}_s}(q_{ij,s} - q_{i1,s})^2} = o_P(1)$ *uniformly over* $P \in \mathcal{P}_{0,n}$.

(vi) (Uniformity of the Transferability Bound) *For every fixed* $M > 0$, *there exists* $N_M < \infty$, *independent of* P_{ta}, *such that, for all* $n \ge N_M$ *and each* $\diamond \in \{\gamma,\alpha\}$,

$$\sup_{h\in\mathcal{H}_{\text{so},m}(M\delta_{\text{so},m})} \min_{f\in\mathcal{F}^{\text{sub}}_{\diamond,n}} \|f\circ h - f_{\diamond,n}\circ h_m\|_{P_{\text{ta}},2} \le M\delta_{\text{so},m}/\nu_{\diamond,n}, \tag{49}$$

uniformly over $P_{\text{ta}} \in \mathcal{P}_{0,\text{ta},n}$.

Assumption 5.1(iii) adapts the conditions for nonlinear scores in Chernozhukov et al. (2018): the moment condition in Assumption 3.3(a), smoothness in 3.3(b), the Jacobian bounds in 3.3(c), and orthogonality in 3.3(d). From Assumption 3.4, it uses (a) jointly over embedding functions, only the moment requirement of (b) at the true score, some components of (c), and (d). The score error bound accommodates growing R. Assumption 5.1(v) requires the estimator of the derivative of the score to converge to J uniformly over j, s and $\theta \in \mathcal{U}_n(P)$.

The mean value theorem then yields uniform linearization of the sample moment functions. Evaluating the derivative bound at the estimators also gives uniform consistency of $\widehat{J}_{j,s}$ in (34). The intervals need not shrink with n. Part (v) also requires consistency of the variance estimators for the first embedding function and a bound on the differences in $q_{ij,s}$ across embedding functions. Together, these variance and weight conditions imply consistency of the variance estimators uniformly over embedding functions, estimation samples, and $P \in \mathcal{P}_{0,n}$. Assumption 5.1(vi) requires the bound in (49) to hold beyond a common threshold N_M for all target distributions. For any fixed target distribution satisfying the null, this bound holds for all sufficiently large n. The common threshold may depend on M and the fixed source problem, but not on P_{ta}.

In the following theorem, we write $\Pr^{(\xi)}$ for probability conditional on all observed data, the partition, and pre-training randomness. For asymptotic statements, we set $\mathcal{T}$, $c_{1-\mathsf{a}}^{(\xi)}$, and all $\mathcal{T}_b^{(\xi)}$ to zero if the algorithm stops without a test decision.

Theorem 5.1 (Uniform Asymptotic Size of the Transferability Test)**.** *For the fixed source distribution, suppose that $H_{0,n}$ in (32) and Assumption 5.1 hold for the sequence of classes $\mathcal{P}_{0,n}$. Fix $\mathsf{a} \in (0,1)$ and let B be any deterministic sequence of positive integers tending to infinity as $n \to \infty$. Compute $\mathcal{T}$, $\{\mathcal{T}_b^{(\xi)}\}_{b=1}^B$, and $c_{1-\mathsf{a}}^{(\xi)}$ using Algorithm 1, with the zero convention stated above. The standard normal multipliers are independent across observations and bootstrap draws, and independent of the data, partition, and pre-training randomness. Then the following properties hold.*

(i) The probability that the algorithm stops without a decision tends to zero uniformly over $P \in \mathcal{P}_{0,n}$.

(ii) For each $b \in \{1, \ldots, B\}$, the conditional bootstrap distribution satisfies

$$\sup_{t \in \mathbb{R}} \left| \Pr^{(\xi)}(\mathcal{T}_b^{(\xi)} \leq t) - \Pr_P(\mathcal{T} \leq t) \right| = o_P(1),$$

uniformly over $P \in \mathcal{P}_{0,n}$.

(iii) The rejection probability satisfies

$$\sup_{P \in \mathcal{P}_{0,n}} \left| \Pr_P\left(\mathcal{T} > c_{1-\mathsf{a}}^{(\xi)}\right) - \mathsf{a} \right| \longrightarrow 0 \qquad \text{as } n, B \to \infty,$$

where the probability includes the randomness of the source and target samples, sample partition, training procedures, and bootstrap draws.

5.3 Uniform Consistency under Alternatives

The test compares the estimated θ induced by the fitted nuisance functions for the observed embeddings. Failure of transferability need not make these estimators differ since the available embedding functions may omit the same relevant information or induce the same bias in the parameter of interest. We establish uniform consistency on a subclass of alternatives under which at least two embedding functions yield sufficiently separated solutions of their population moment equations. That is, under such alternatives, where the population moment solutions associated with different embedding functions are well separated, the test detects the lack of transferability with probability approaching one uniformly over this subclass.

For each embedding function j, let $\theta^{\dagger}_{j,n} \in \Theta$ be a measurable solution of

$$\mathbb{E}_{\mathrm{ta}}[\psi(W; \theta^{\dagger}_{j,n}, \widehat{\gamma}_j, \widehat{\alpha}_j)] = 0. \tag{50}$$

The expectation holds the fitted nuisance functions fixed; thus, these roots may depend on the source sample, the nuisance-training sample, the partition, and pre-training randomness. Under an alternative, these solutions need not converge in probability to θ^*, and the derivative estimators in Algorithm 1 need not converge in probability to the common derivative J.

Let $\mathcal{P}_{\mathrm{sep},n} \subseteq \mathcal{P}^{c}_{0,n}$ be a subclass of joint laws $P = P_{\mathrm{so}} \times P_{\mathrm{ta}}$ under alternatives to (32), with the same fixed source problem as above. The complement is taken within the class of joint laws with the fixed source problem. For the uniform consistency results, we focus on this subclass of alternatives satisfying certain separation conditions.

Assumption 5.2 (Conditions for Uniform Consistency of the Transferability Test)**.** *For some deterministic estimation error rate $r_{\theta,n} > 0$ and separation sequence $\Delta_n > 0$, common across $\mathcal{P}_{\mathrm{sep},n}$, the following conditions hold:*

(i) There exist constants $c_J, c_\psi > 0$ such that

$$\inf_{P \in \mathcal{P}_{\mathrm{sep},n}} \Pr_P \left(\min_{\substack{1 \le j \le R \\ s \in \{1,2\}}} |\widehat{J}_{j,s}| \ge c_J, \quad \min_{\substack{1 \le j \le R \\ s \in \{1,2\}}} \frac{1}{n_s} \sum_{i \in \mathcal{I}_s} \psi(W_i; \widehat{\theta}_{j,s}, \widehat{\gamma}_j, \widehat{\alpha}_j)^2 \ge c_\psi \right) \longrightarrow 1, \tag{51}$$

where the derivative estimators are defined in (34).

(ii) Uniformly over $P \in \mathcal{P}_{\mathrm{sep},n}$,

$$\max_{1 \le j \le R, s \in \{1,2\}} |\widehat{\theta}_{j,s} - \theta^{\dagger}_{j,n}| = O_P(r_{\theta,n}), \qquad \sqrt{n} \max_{j<k} \widehat{\sigma}_{jk} = O_P(1), \tag{52}$$

where $\widehat{\sigma}^2_{jk}$ is defined in (36).

(iii) The population solutions satisfy

$$\inf_{P\in\mathcal{P}_{\mathrm{sep},n}} \Pr_P\left(\max_{j<k}|\theta^\dagger_{j,n}-\theta^\dagger_{k,n}|\geq\Delta_n\right)\longrightarrow 1, \tag{53}$$

with $r_{\theta,n}=o(\Delta_n)$ *and* $\sqrt{\log R/n}=o(\Delta_n)$.

Assumption 5.2(ii) and (iii) require that the separation between the population solutions is sufficiently larger than the estimation error of the DML estimators. The next result establishes the uniform consistency of the transferability test under these conditions.

Theorem 5.2 (Uniform Consistency of the Transferability Test)**.** *Suppose that the sampling conditions in Assumption 5.1(i) hold uniformly over* $\mathcal{P}_{\mathrm{sep},n}$*, the source convergence bound in (39) holds under* P_{so}*, and the solutions in (50) exist. Suppose also that Assumption 5.2 holds. Fix* $\mathsf{a}\in(0,1)$ *and let* $B\geq 1$ *be any deterministic sequence of integers. Compute the statistic and empirical bootstrap critical value using Algorithm 1, with the same zero convention as in Theorem 5.1. Then,*

$$\inf_{P\in\mathcal{P}_{\mathrm{sep},n}} \Pr_P\left(\mathcal{T}>c^{(\xi)}_{1-\mathsf{a}}\right)\longrightarrow 1. \tag{54}$$

6 Simulations

We study the finite-sample performance of DML with pre-trained embeddings in two designs. The linear design compares ordinary least squares, cross-validated LASSO, and cross-validated ridge as target nuisance learners over 5,000 Monte Carlo replications. The nonlinear design compares nine nonlinear nuisance learners and evaluates the transferability test over 100 Monte Carlo replications per scenario. The author is currently working on extending the simulations to additional scenarios and larger numbers of Monte Carlo replications. Both designs have $m=20{,}000$ source observations and a true target coefficient $\theta^*=0.5$. Each replication draws independent source and target samples and estimates the source embeddings. The oracle benchmark supplies the population embedding h^* to the target learner, yielding the performance of the ordinary DML procedure with known embeddings. We write $[h^*(z)]_j$ for the jth coordinate of $h^*(z)$. We use the last hidden layer after the activation and the layer normalization as embeddings, as we do for the BERT embedding in the empirical application.

All specifications estimate the partially linear score in Example 2.1, using three-fold cross-fitting repeated five times to reduce the variability of the cross-fitted estimates due to random fold assignments. For each embedding and learner, we report the median coefficient across the five repetitions and estimate its variance by the median of the repetition-specific variances augmented by the squared deviation of each repetition's coefficient from the median coefficient.

The median aggregation is recommended over the mean aggregation and theoretically justified in Chernozhukov et al. (2018, Corollary 3.3). Confidence intervals use normal critical values at the 95% level. All target learners' tuning and validation take place within the corresponding outer training fold. We report bias and coverage for each target learner.

6.1 Linear Design with High-Dimensional Embeddings

The population embedding has 200 coordinates and is given by[25]

$$h^*(z) = s_0^{-1}\,\mathrm{LayerNorm}(\mathrm{ReLU}(z + \mathbf{1})),$$

where $s_0 \approx 1.0025$ is a scaling factor explained below. Source coordinates are independent uniform draws on $Z \sim [-\sqrt{3}, \sqrt{3}]$, whereas target coordinates are independent uniform draws on $Z \sim [-\sqrt{2}, \sqrt{2}]$ and they are mutually independent. We stack these independent Z and construct the 200-dimensional vectors to form source and target inputs $Z_j^{\mathrm{so}} \in \mathbb{R}^{200}$ and $Z_i^{\mathrm{ta}} \in \mathbb{R}^{200}$, respectively. We set s_0 so that $([h^*(Z)]_1 - [h^*(Z)]_2)/\sqrt{2}$ and $([h^*(Z)]_3 - [h^*(Z)]_4)/\sqrt{2}$ each have variance approximately one under the source distribution.

The source response is $S_j = B' h^*(Z_j^{\mathrm{so}}) + \eta_j$, where $\eta_j \sim N(0, I_{200})$ is independent noise. The fixed source coefficient matrix $B \in \mathbb{R}^{200\times 200}$ has rank 198. The columns of B are scaled so that the first column has Euclidean norm 10 and each of the other 199 columns is orthogonal to the first and has Euclidean norm $\sqrt{3}$.

The target sample has $n = 2{,}000$ observations and is generated via

$$X_i = 0.3\left\{\frac{[h^*(Z_i)]_1 - [h^*(Z_i)]_2}{\sqrt{2}} + \frac{[h^*(Z_i)]_3 - [h^*(Z_i)]_4}{\sqrt{2}}\right\} + V_i,$$
$$Y_i = 0.5X_i + 2\left\{\frac{[h^*(Z_i)]_3 - [h^*(Z_i)]_4}{\sqrt{2}} - \frac{[h^*(Z_i)]_1 - [h^*(Z_i)]_2}{\sqrt{2}}\right\} + U_i,$$

where $V_i \sim N(0,1)$ and $U_i \sim N(0, 0.25)$ are independent of each other and of all covariates and source observations. The coefficient vectors of both target contrasts lie in the column space of B, so the reported design satisfies the linear span condition for transferability. They are also sparse, involving only a few coordinates of the source representation.

The learned source model has one dense hidden layer of width 200, followed by ReLU,

[25] For $v \in \mathbb{R}^K$, let $\overline{v} = K^{-1}\sum_{k=1}^K v_k$. For $k = 1, \ldots, K$, we use

$$[\mathrm{ReLU}(v)]_k = \max\{v_k, 0\},$$
$$[\mathrm{LayerNorm}(v)]_k = \frac{v_k - \overline{v}}{\sqrt{K^{-1}\sum_{k=1}^K (v_k - \overline{v})^2 + 10^{-5}}}.$$

Here, layer normalization centers and rescales the coordinates within each observation.

LayerNorm, and a linear source head. Source training uses an 80/10/10 training, validation, and test split. We compare three initializations and select the epoch and initialization with the highest average validation R^2 across source outcomes. We then refit the selected specification on the combined training and validation observations. The target nuisance regressions use either this learned embedding or h^*.

Table 1 reports the results. With the population embedding, coverage is 93.66% for OLS, 92.14% for LASSO, and 95.42% for ridge. With learned embeddings, their coverages are 93.42%, 83.86%, and 94.78%, respectively. Thus, the coverage loss from replacing the population embedding by the learned embedding is substantially larger for LASSO in this design. LASSO also has the largest absolute bias among the three learners under either representation. This pattern is consistent with the sensitivity of coordinatewise sparsity to the representation, as discussed in Section 4.2.2.

Table 1: Linear design: bias and interval coverage

	Oracle		Learned	
Learner	Bias	Coverage (%)	Bias	Coverage (%)
OLS	0.0000	93.66	−0.0005	93.42
CV-Lasso	−0.0062	92.14	−0.0149	83.86
CV-Ridge	−0.0005	95.42	−0.0025	94.78

Notes: Each cell uses 5,000 successful Monte Carlo replications; there are no failed fits. The target and source sample sizes are $n = 2{,}000$ and $m = 20{,}000$, with $\theta^* = 0.5$. Bias is relative to θ^*. Coverage is for nominal 95% normal intervals.

6.2 Nonlinear Design

Next, we consider a nonlinear design. We first compare the coverage between the scenarios with and without the transferability condition being satisfied. Then, we evaluate the size and power of our test for transferability.

6.2.1 Coverage

Source and target inputs are independent draws from $Z \sim N(0, I_{30})$. The population embedding is $h^*(z) = \text{LayerNorm}(\text{GELU}(z))$, where $\text{GELU}(z_j) = z_j \Phi(z_j)$ and Φ is the standard normal distribution function. Using the map Λ defined in Section 2.1, the target equations are

$$X_i = [h^*(Z_i)]_1 + \frac{1}{4}\Lambda([h^*(Z_i)]_3) + V_i,$$
$$Y_i = 0.5X_i + \Lambda([h^*(Z_i)]_1) + \frac{1}{4}[h^*(Z_i)]_3 + U_i,$$

where V_i and U_i are independent standard normal innovations, independent of the covariates and source sample. This functional form is the one used in the simulation study of Chernozhukov et al. (2018). For $c(z) = ([h^*(z)]_1, [h^*(z)]_3, [h^*(z)]_2, [h^*(z)]_4, [h^*(z)]_5, [h^*(z)]_6)'$, the source response is

$$S_j = Q\,\mathrm{diag}(1-\rho, 1-\rho, 1, 1, 1, 1)c(Z_j^{\mathrm{so}}) + \eta_j, \qquad \eta_j \sim N(0, I_{20}),$$

where Q is a fixed 20×6 matrix with orthonormal columns, generated by a reduced QR decomposition of a standard Gaussian matrix, and source noise is independent of all other draws. We consider $\rho \in \{0, 1\}$ and $n \in \{5{,}000, 10{,}000\}$. At $\rho = 0$, the source task retains the direct loadings on the target-relevant components and the transferability condition is satisfied. At $\rho = 1$, the source task removes the direct loadings on the two target-relevant components and the transferability condition is violated.

The trained source network has one dense hidden layer of width 30, followed by GELU (Gaussian Error Linear Unit) activation, affine LayerNorm, and a linear head with 20 outputs. Training minimizes unstandardized source mean squared error on an 80% training split, selects the best checkpoint using a 10% validation split, and reserves the remaining 10% for diagnostics; there is no refit after checkpoint selection. Source and target draws are paired across the two scenarios. The source samples and fitted embeddings are identical across target sample sizes within a scenario and replication.

The target learners are three random forests (RF 1–3), three gradient-boosted tree models (XGB 1–3), and three neural networks (NN 1–3). The forests use 200 trees, search all features, and have minimum leaf sizes 100, 30, and 1. The boosted trees have maximum depth three and at most 200 rounds; their minimum child weights are 10% and 3% of the inner fitting sample size, rounded up, and 1, respectively. The neural networks use ReLU hidden layers of widths (16), (32, 32), and (32, 32, 32), with at most 600 epochs and early stopping after 60 epochs without validation improvement.

Table 2 reports bias and coverage. For learned embeddings, each successful replication contributes one estimate and interval, computed using the first source embedding in the fixed candidate order and the full-sample cross-fitting procedure described above. The same candidate position is used in every replication. At $\rho = 0$, coverage with learned embeddings ranges from 93% to 96% at $n = 5{,}000$ and from 91% to 99% at $n = 10{,}000$. For the neural networks, oracle coverage ranges from 94% to 95% at the smaller sample size and from 97% to 99% at the larger sample size; the corresponding ranges with learned embeddings are 94%–95% and 91%–94%. At $\rho = 1$, coverage deteriorates sharply, ranging from 0% to 4% at $n = 5{,}000$

and equaling 0% for every learner at $n = 10{,}000$. All of these learners have positive Monte Carlo bias at $\rho = 1$.

Table 2: Nonlinear design: bias and interval coverage

	$n = 5{,}000$					
	Oracle $\rho = 0$		Learned $\rho = 0$		Learned $\rho = 1$	
Learner	Bias	Coverage (%)	Bias	Coverage (%)	Bias	Coverage (%)
RF 1	−0.0006	92.00	0.0061	94.00	0.0716	0.00
RF 2	−0.0035	93.00	0.0022	93.00	0.0668	0.00
RF 3	−0.0081	89.00	−0.0033	96.00	0.0612	0.00
XGB 1	0.0007	94.00	0.0037	93.00	0.0571	0.00
XGB 2	−0.0030	93.00	0.0007	94.00	0.0535	2.00
XGB 3	−0.0040	92.00	−0.0003	95.00	0.0524	4.00
NN 1	−0.0043	94.00	0.0021	94.00	0.0627	0.00
NN 2	−0.0046	94.00	0.0027	95.00	0.0632	0.00
NN 3	−0.0036	95.00	0.0034	94.00	0.0632	0.00
	$n = 10{,}000$					
	Oracle $\rho = 0$		Learned $\rho = 0$		Learned $\rho = 1$	
Learner	Bias	Coverage (%)	Bias	Coverage (%)	Bias	Coverage (%)
RF 1	−0.0014	99.00	0.0053	94.00	0.0703	0.00
RF 2	−0.0033	97.00	0.0024	97.00	0.0663	0.00
RF 3	−0.0071	90.00	−0.0017	99.00	0.0613	0.00
XGB 1	0.0012	97.00	0.0047	94.00	0.0594	0.00
XGB 2	−0.0019	99.00	0.0022	97.00	0.0555	0.00
XGB 3	−0.0024	98.00	0.0015	96.00	0.0540	0.00
NN 1	−0.0022	99.00	0.0042	94.00	0.0655	0.00
NN 2	−0.0012	98.00	0.0049	94.00	0.0670	0.00
NN 3	0.0005	97.00	0.0065	91.00	0.0682	0.00

Notes: Bias is the mean estimation error relative to $\theta^* = 0.5$. Coverage is for nominal 95% normal intervals. Intervals use three-fold cross-fitting repeated five times, with median aggregation across splits.

6.2.2 Transferability Test

For each ρ and replication, we train 25 source embeddings on the same source sample with different initialization and training-order seeds. We use a nominal 5% rejection threshold. Within each held-out fold, observations are split into two estimation blocks that share nuisance fits from the training complement. After pooling held-out observations within each block across folds, we form the maximum statistic over all 300 pairs of embeddings (all possible pairs of the 25 source embeddings) in each of the two block orientations, using 500 multiplier bootstrap draws per repetition and orientation. While Theorem 5.1 assumes a single nuisance-training block independent of both estimation blocks, we encourage implementation of the cross-fitting procedure for the efficient use of the data, even though its theoretical size validity requires a

separate argument. The five cross-fitting repetitions and two test sample orientations yield ten bootstrap p-values. We compare the grid-harmonic merger with the corrected upper-median rule to aggregate the ten bootstrap p-values into a single p-value for each replication.

For any positive integer J, let $p_1, \ldots, p_J \in [0,1]$ denote the input p-values and write $\ell_J = \sum_{j=1}^{J} 1/j$. For positive inputs, the grid-harmonic p-value is the smallest $t \in (0,1]$ satisfying

$$\sum_{i=1}^{J} \frac{\mathbf{1}\{\ell_J p_i \le t\}}{\lceil J\ell_J p_i/t \rceil} \ge 1,$$

and is one if no such threshold exists. We set it to zero if any input p-value is zero. Vovk, Wang, and Wang (2022, Theorem 7.1) show that the grid-harmonic merging function is admissible among symmetric p-merging functions and, when J is not prime, among all p-merging functions. Here admissibility means that no other rule in the relevant class returns weakly smaller p-values for every input vector and strictly smaller p-values for at least one input vector.

For the corrected upper-median rule, order the inputs as $p_{(1)} \le \cdots \le p_{(J)}$ and set $k_J = \lfloor J/2 \rfloor + 1$. The corrected upper-median p-value justified by Meinshausen, Meier, and Bühlmann (2009, Theorem 3.1) is

$$p_{\text{med}} = \min\{1, (J/k_J) p_{(k_J)}\}.$$

Here $p_{(k_J)}$ is the middle order statistic for odd J and the larger of the two middle order statistics for even J. Rejection at level $\alpha \in (0,1)$ requires at least k_J inputs to be at most $\alpha k_J/J$. For ten inputs, the corrected upper-median rule is $\min\{1, (10/6)p_{(6)}\}$, where $p_{(6)}$ is the sixth order statistic.

Table 3 reports rejection frequencies in both scenarios. For the grid-harmonic merger, rejection at $\rho = 0$ ranges from 4% to 7% at $n = 5{,}000$ and from 0% to 3% at $n = 10{,}000$. Rejection at $\rho = 1$ ranges from 23% to 36% at the smaller sample size and from 44% to 71% at the larger sample size. The corrected upper-median rule has lower rejection frequencies: at $\rho = 0$, every frequency is zero; at $\rho = 1$, frequencies range from 5% to 11% and from 23% to 51% at the two sample sizes. In these designs, the grid-harmonic merger rejects more frequently under non-transferability, and rejection frequencies under non-transferability increase with the target sample size for both merging rules.

Table 3: Nonlinear design: rejection rates of transferability tests

	$n = 5{,}000$				$n = 10{,}000$			
	$\rho = 0$		$\rho = 1$		$\rho = 0$		$\rho = 1$	
Learner	Median	Harmonic	Median	Harmonic	Median	Harmonic	Median	Harmonic
RF 1	0.00	4.00	6.00	23.00	0.00	0.00	23.00	44.00
RF 2	0.00	5.00	5.00	29.00	0.00	1.00	36.00	50.00
RF 3	0.00	6.00	6.00	32.00	0.00	1.00	32.00	50.00
XGB 1	0.00	5.00	9.00	31.00	0.00	3.00	33.00	55.00
XGB 2	0.00	5.00	11.00	35.00	0.00	1.00	48.00	64.00
XGB 3	0.00	7.00	10.00	36.00	0.00	2.00	51.00	71.00
NN 1	0.00	7.00	8.00	28.00	0.00	1.00	38.00	53.00
NN 2	0.00	7.00	8.00	31.00	0.00	1.00	37.00	54.00
NN 3	0.00	7.00	11.00	32.00	0.00	1.00	44.00	55.00

Notes: Entries are rejection percentages at the 5% level. Median and harmonic denote the corrected upper-median rule and grid-harmonic merger, respectively. Both methods combine the same ten p-values from five cross-fitting repetitions and two swap test orientations, using 500 bootstrap draws for each.

7 Empirical Application: Monopsony on an Online Labor Market Platform

We study the use of pre-trained text embeddings as controls in the online labor-market application of Dube et al. (2020). The empirical question is how the speed at which workers accept a task responds to its posted reward, holding task characteristics fixed. Task descriptions contain information about these characteristics, but converting that information into suitable controls requires choices about both the text representation and the nuisance learner. The dataset is revisited by Ahrens et al. (2026) with another text embedding function; however, our focus is on illustrating our transferability theory, comparing different pre-trained text embeddings as well as different machine learners in the downstream analysis, applying the test for transferability in Section 5 to pre-trained BERT representations, and suggesting a new aggregation robust to the pre-trained embeddings.

7.1 Background and Text as Controls

The analysis uses the sample of Amazon Mechanical Turk postings, comprising $n = 258{,}352$ task groups posted by 24,923 requesters. We treat each requester as a cluster in the analysis to account for potential within-requester correlation. Let Y_i be log posting duration in minutes, X_i be log posted reward in cents, and Z_i collect the job description text and structured controls. The identifying specification is the partially linear model

$$Y_i = X_i\theta^* + \kappa^*(Z_i) + U_i, \qquad \mathbb{E}[U_i \mid X_i, Z_i] = 0, \tag{55}$$

where κ^* is an unknown nonparametric function that captures the relationship between task characteristics and duration. Since the outcome is log duration and the main regressor is log posted reward, the labor supply elasticity of filling speed is $-\theta^*$ under the monopsony model considered in Dube et al. (2020). To control for unstructured heterogeneity among the job postings, they include both structured controls and unstructured text-derived controls in the specification.

The first control specification, labeled Dube, contains 611 numerically distinct variables: 12 non-text controls, 110 hand-engineered text controls, and 489 learned text controls consisting of supervised n-grams, topic-model features, and Doc2Vec coordinates. This specification provides a comparison with an application-specific construction of text controls.[26] The remaining six specifications replace all text-derived Dube controls with frozen BERT embeddings, while retaining the same 12 non-text controls. We use Small, Base, and Large models with embedding dimensions 512, 768, and 1,024, respectively, and two pooling rules for each size. The CLS rule uses the final encoder state at the `[CLS]` position. The mean rule first applies the pre-trained masked-language-modeling transformation to each token in sentences and then averages these transformed vectors in the last hidden layer, excluding padding and special tokens. Both rules use the task title followed by its description, truncated to at most 160 WordPiece tokens, and the BERT parameters are not fine-tuned on the application sample. The resulting BERT specifications contain 524, 780, or 1,036 controls. To visualize the embeddings, we provide figures in Appendix J.3 illustrating the relationships between the embeddings and task characteristics.

We emphasise that while Dube controls depend on an application-specific selection of text features by researchers, the BERT-based controls provide a flexible representation of the textual information in job postings without requiring manual feature engineering.

7.2 Estimation and Comparison of Specifications

We implement the orthogonal score in Example 2.1 by predicting Y_i and X_i from the chosen controls and regressing the out-of-fold outcome residual on the corresponding reward residual. For each control specification, we use twelve nuisance learners: ordinary least squares, cross-validated lasso and ridge, three random forests (RF 1–3), three gradient-boosted tree models (XGB 1–3), and three neural networks (NN 1–3). The same learner specification is used for both nuisance regressions, with separate fitted models. RF 1–3 use 600 trees and minimum

[26] The n-gram features are selected by Dube et al. (2020). They first select the n-gram features and then treat them as fixed in the DML estimation. We collect all selected n-gram features and include them as fixed controls in our analysis after eliminating numerical duplicates.

leaf sizes 100, 30, and 1. XGB 1–3 use trees of maximum depth six and minimum child weights 2,000, 500, and 1, with up to 800 boosting rounds and early stopping. The neural networks have three hidden layers of width ten, three of width thirty, or five of width fifty, respectively, with ReLU activations. Thus, learners labeled with 1 correspond to the least flexible models, and those labeled with 3 correspond to the most flexible models.

The results use three-fold cross-fitting with five repetitions clustered at the requester level.[27] All postings by a requester belong to the same fold, and learner-specific tuning and validation splits within a fold also preserve requester groups. Each repetition produces a coefficient using all out-of-fold residuals and a requester-clustered variance estimator. We report the median coefficient across repetitions and take the median of the repetition variances augmented by the squared deviation of each repetition's coefficient from that median. Reported standard errors are the square roots of these aggregated variances. Prediction performance is summarized by out-of-fold R_Y^2 and R_X^2 for duration and reward, averaged across repetitions.

Table 4 reports all combinations of controls and nuisance learners. The XGBoost specifications attain the highest outcome out-of-fold R_Y^2 among the reported learners for every control specification, with R_Y^2 ranging from 0.8421 to 0.8621. Their reward R_X^2 values are lower and vary more across control specifications, with R_X^2 ranging from 0.2944 to 0.4109. XGBoost also attains the highest R_X^2 for six of the seven control specifications; the exception is BERT Large mean, for which ridge performs best. The choice of text embeddings matters more for reward prediction than for duration prediction within the XGBoost specifications. Mean pooling consistently yields higher R_Y^2 and generally higher R_X^2 than CLS pooling, while increasing BERT size from Base to Large lowers R_X^2 for all three XGBoost learners under both pooling rules. BERT Base mean embeddings, used in the subsequent robustness analysis, attain the highest R_X^2 within each XGBoost learner: 0.4011, 0.4109, and 0.3847 for XGB 1–3, respectively, compared with 0.3562, 0.3348, and 0.3434 using the Dube controls.

These patterns highlight the importance of both the choice of text embeddings and the selection of the nuisance learner for accurate reward prediction. As our theory suggests the tradeoff of the embedding size between the approximation errors and model complexity, the choice of text embeddings should balance these two considerations to optimize prediction performance. In practice, we recommend comparing multiple embedding sizes and pooling strategies based on the out-of-fold prediction performance measured by R^2. Moreover, BERT

[27] Appendix J.1 reports the Dube specification with individual-level cross-fitting. While Dube et al. (2020) use the cluster-robust variance for requester-level clustering as we do here, they use individual-level folds for cross-fitting. Potentially, the very small p-values in Appendix J.1 can be explained by the over-fitting of the learners due to within-cluster dependence. To mitigate this concern, we use cross-fitting with cluster-level folds in our analysis.

Table 4: DML estimates by embedding and nuisance learner

Panel A: Linear and random-forest learners						
	OLS	CV-Lasso	CV-Ridge	RF 1	RF 2	RF 3
Dube (611 features)						
Estimate (s.e.)	0.0235 (1.1945)	−0.2648 (0.1751)	−0.2087 (1.4814)	0.0379 (0.2028)	0.0285 (0.2101)	0.1085 (0.2571)
p-value	0.9843	0.1304	0.8880	0.8519	0.8922	0.6731
R^2_Y / R^2_D	−16.5183 / −15.3010	−15.9878 / 0.3218	−21.8118 / −18.9989	0.0355 / 0.1083	0.0470 / 0.1311	0.0883 / 0.2178
BERT Small Mean (524 features)						
Estimate (s.e.)	0.0109 (0.2010)	−0.1795 (0.1425)	−0.1500 (0.1539)	0.0403 (0.1980)	0.0583 (0.2047)	0.0721 (0.2017)
p-value	0.9568	0.2079	0.3297	0.8386	0.7757	0.7208
R^2_Y / R^2_D	0.0064 / 0.2589	0.1191 / 0.2723	0.1177 / 0.3500	0.0913 / 0.1505	0.0948 / 0.1681	0.1403 / 0.2073
BERT Small CLS (524 features)						
Estimate (s.e.)	−0.1545 (0.1784)	−0.1606 (0.1386)	−0.1799 (0.1437)	0.0978 (0.2254)	0.1018 (0.2289)	0.1175 (0.2355)
p-value	0.3863	0.2464	0.2104	0.6643	0.6566	0.6178
R^2_Y / R^2_D	0.1214 / 0.2449	0.1000 / 0.2800	0.1393 / 0.3520	0.0332 / 0.1562	0.0370 / 0.1783	0.0952 / 0.2154
BERT Base Mean (780 features)						
Estimate (s.e.)	−0.0386 (0.1801)	−0.1403 (0.1657)	−0.1774 (0.1589)	0.1134 (0.2231)	0.1170 (0.2281)	0.1309 (0.2234)
p-value	0.8303	0.3972	0.2642	0.6112	0.6081	0.5579
R^2_Y / R^2_D	0.0622 / 0.1775	0.0764 / 0.2465	0.1376 / 0.3093	0.0493 / 0.1513	0.0512 / 0.1752	0.1020 / 0.2099
BERT Base CLS (780 features)						
Estimate (s.e.)	−0.1058 (0.1383)	−0.1254 (0.1815)	−0.1141 (0.1525)	0.1325 (0.2485)	0.1430 (0.2548)	0.1736 (0.2536)
p-value	0.4446	0.4897	0.4543	0.5939	0.5745	0.4936
R^2_Y / R^2_D	0.1082 / 0.2657	0.0821 / 0.2648	0.1266 / 0.3052	−0.0121 / 0.1258	−0.0099 / 0.1438	0.0433 / 0.1756
BERT Large Mean (1,036 features)						
Estimate (s.e.)	−0.0848 (0.1379)	−0.1790 (0.1687)	−0.1956 (0.1455)	0.0636 (0.2195)	0.0710 (0.2272)	0.0946 (0.2206)
p-value	0.5388	0.2885	0.1789	0.7720	0.7547	0.6680
R^2_Y / R^2_D	−0.0239 / 0.2993	0.0871 / 0.2754	0.1231 / 0.3376	0.0637 / 0.1432	0.0662 / 0.1607	0.1137 / 0.1951
BERT Large CLS (1,036 features)						
Estimate (s.e.)	0.0248 (0.1866)	−0.1665 (0.1598)	0.0137 (0.2146)	0.1672 (0.2629)	0.1842 (0.2674)	0.2022 (0.2639)
p-value	0.8943	0.2974	0.9492	0.5247	0.4909	0.4435
R^2_Y / R^2_D	0.0272 / 0.2288	0.0097 / 0.2469	0.0541 / 0.2446	−0.0069 / 0.0689	−0.0030 / 0.0843	0.0516 / 0.1160
Panel B: XGBoost and neural-network learners						
	XGB 1	XGB 2	XGB 3	NN 1	NN 2	NN 3
Dube (611 features)						
Estimate (s.e.)	−0.0695 (0.0460)	−0.0422 (0.0338)	−0.0262 (0.0361)	1.0597 (0.7457)	−0.0764 (1.3586)	0.7116 (0.6657)
p-value	0.1307	0.2116	0.4687	0.1553	0.9551	0.2851
R^2_Y / R^2_D	0.8441 / 0.3562	0.8539 / 0.3348	0.8566 / 0.3434	−157.8101 / −68.2209	−138.9729 / −85.0758	−103.6923 / −88.4491
BERT Small Mean (524 features)						
Estimate (s.e.)	−0.0263 (0.0217)	−0.0285 (0.0204)	−0.0263 (0.0135)	−0.2543 (0.1082)	−0.0920 (0.1658)	−0.1024 (0.1391)
p-value	0.2257	0.1627	0.0520	0.0187	0.5791	0.4613
R^2_Y / R^2_D	0.8508 / 0.3509	0.8596 / 0.3683	0.8621 / 0.3677	0.3904 / 0.3335	0.2605 / 0.2858	0.2299 / 0.3165
BERT Small CLS (524 features)						
Estimate (s.e.)	−0.0422 (0.0271)	−0.0376 (0.0250)	−0.0430 (0.0212)	−0.1729 (0.1320)	−0.1670 (0.1710)	−0.1181 (0.1450)
p-value	0.1198	0.1331	0.0424	0.1904	0.3286	0.4155
R^2_Y / R^2_D	0.8421 / 0.3475	0.8567 / 0.3602	0.8572 / 0.3457	0.2968 / 0.3145	0.1840 / 0.3201	0.1768 / 0.3372
BERT Base Mean (780 features)						
Estimate (s.e.)	−0.0369 (0.0199)	−0.0374 (0.0175)	−0.0334 (0.0216)	−0.0571 (0.1384)	−0.0396 (0.1887)	−0.0287 (0.1537)
p-value	0.0635	0.0326	0.1225	0.6800	0.8339	0.8519
R^2_Y / R^2_D	0.8467 / 0.4011	0.8566 / 0.4109	0.8605 / 0.3847	0.2608 / 0.2938	0.2423 / 0.2819	0.2185 / 0.3107
BERT Base CLS (780 features)						
Estimate (s.e.)	−0.0444 (0.0330)	−0.0422 (0.0214)	−0.0355 (0.0221)	0.0138 (0.1690)	0.0981 (0.2110)	0.1057 (0.2120)
p-value	0.1780	0.0492	0.1077	0.9348	0.6419	0.6181
R^2_Y / R^2_D	0.8430 / 0.3522	0.8530 / 0.3897	0.8575 / 0.3466	0.2778 / 0.1560	0.1966 / 0.2298	0.1613 / 0.1600
BERT Large Mean (1,036 features)						
Estimate (s.e.)	−0.0376 (0.0317)	−0.0419 (0.0253)	−0.0255 (0.0184)	−0.1140 (0.1142)	−0.1645 (0.2040)	−0.0450 (0.1447)
p-value	0.2360	0.0978	0.1661	0.3183	0.4201	0.7559
R^2_Y / R^2_D	0.8459 / 0.3293	0.8589 / 0.3281	0.8615 / 0.3175	0.3068 / 0.3094	0.2174 / 0.2702	0.2111 / 0.3017
BERT Large CLS (1,036 features)						
Estimate (s.e.)	−0.0451 (0.0399)	−0.0274 (0.0254)	−0.0321 (0.0312)	0.0471 (0.2022)	0.1126 (0.2679)	0.1571 (0.2075)
p-value	0.2584	0.2814	0.3028	0.8160	0.6743	0.4489
R^2_Y / R^2_D	0.8422 / 0.2944	0.8539 / 0.3145	0.8555 / 0.3205	0.1998 / 0.2629	0.1360 / 0.1884	0.0906 / 0.2166

Notes: Requester-level cross-fitting uses 3 folds and 5 repetitions; 258,352 observations and 24,923 requesters. Estimates are median coefficients on log reward in the log-duration equation. Parentheses report requester-clustered standard errors with split-deviation inflation. Two-sided normal p-values test a zero coefficient. The final row lists out-of-sample outcome and treatment R^2 scores, averaged across repetitions. Dube uses all 611 numerically unique controls. BERT embeddings include 12 non-text controls.

embeddings with mean pooling tend to perform better than those with CLS pooling for reward prediction, which is also consistent with our intuition discussed in Appendix C.2. Finally, we emphasize that the BERT embeddings do not require the hand-engineered features used in the Dube controls for accurate reward prediction. By using the modern text representations provided by BERT, we can achieve high prediction performance without relying on additional manually crafted features.

The XGBoost coefficients are comparatively stable across text representations. Across the six BERT specifications and three XGBoost learners, the duration coefficient ranges from -0.0451 to -0.0255, corresponding to labor supply elasticities between 0.0255 and 0.0451. With BERT Base mean embeddings, XGB 1–3 give coefficients of -0.0369, -0.0374, and -0.0334, with standard errors 0.0199, 0.0175, and 0.0216. Their coefficient p-values are 0.0635, 0.0326, and 0.1225, respectively. In the subsequent analysis, we focus on the XGBoost specifications with BERT Base mean embeddings.[28]

7.3 Robustness to Pre-Trained Embeddings

Next, we use a transferability diagnostic to assess the sensitivity of the estimated duration coefficient to different pre-trained embeddings. We compare the original BERT Base model with the 25 MultiBERT reproductions of Sellam et al. (2022), which use different training random seeds governing the initial values and the order of observations used in the optimization process. All 26 embeddings, including the original BERT embeddings, use the same architecture, mean pooling, and 12 non-text controls. We apply the sensitivity diagnostic separately to XGB 1–3.

We adapt the bootstrap in Section 5 to requester-clustered cross-fitting and pool contributions across folds. This implementation is not covered by Theorem 5.1, which assumes i.i.d. target observations and a single nuisance-training block independent of both estimation blocks. Extensions to pooled cross-fitting and to clustered observations require separate justification and are left to future work. Within each held-out fold, requesters are split into estimation blocks $\mathcal{I}_1$ and $\mathcal{I}_2$, sharing nuisance models trained on the complement $\mathcal{I}_3$. After pooling block residuals across folds, each repetition tests the maximum absolute studentized contrast between embedding j on $\mathcal{I}_1$ and k on $\mathcal{I}_2$ for $j < k$, and repeats the test with the block assignments reversed. Influence contributions are summed within requesters, with one Gaussian multiplier per requester shared across embeddings, to preserve dependence among contributions from the same requester in the bootstrap. We use 2,000 bootstrap draws per DML cross-fitting repetition and testing orientation, yielding ten p-values from five cross-fitting repetitions and

[28] Formally, we also need to consider the model uncertainty among both learners and text representations, but for simplicity, we simply concentrate on this particular combination.

two testing orientations. For each learner, we aggregate these same ten p-values using the grid-harmonic and corrected upper-median rules defined in Section 6.2.2.

Table 5 reports the resulting aggregated p-values for XGB 1–3. The grid-harmonic merger yields larger p-values for XGB 1 and XGB 2 but a smaller value for XGB 3. Neither diagnostic rejects at the nominal 5% or 10% threshold for any learner.

Table 5: Transferability-test p-values

Learner	Median	Harmonic
XGB 1	0.6817	1.0000
XGB 2	0.4875	0.8567
XGB 3	0.6550	0.6078

Notes: Test for transferability compares BERT Original and 25 MultiBERT seeds. Both mergers use the same 10 bootstrap p-values from 5 cross-fitting repetitions and both testing orientations with swapped blocks. The corrected upper median is $\min\{1, (10/6)p_{(6)}\}$, where $p_{(6)}$ is the order statistic of rank 6. The grid-harmonic merger follows Vovk et al. (2022).

Given non-rejection under both aggregation methods, we suggest reporting the aggregated coefficients and their corresponding confidence intervals to make them more robust to embedding-specific variations. If the test for transferability had rejected, such aggregation might be misleading, and we would report the range of coefficients across embeddings instead. Figure 6 displays the XGB 2 coefficients and pointwise 90% and 95% normal confidence intervals. The corresponding figures for XGB 1 and XGB 3 appear in Appendix J.2. The estimates use the full-sample cross-fitting procedure described above. Aggregate is the median of all 130 embedding-by-repetition coefficients, including the original BERT model. Figure 6 shows that BERT Original's coefficient is generally close to the median of all embeddings; however, its confidence intervals are on the lower side compared to those of other embeddings. Consequently, if the researcher reports BERT Original's coefficient and confidence intervals only, they might overstate the evidence for its effect. By considering the aggregated median coefficient and its confidence intervals, the researcher can provide a more robust summary of the embedding-specific variations.

The aggregated median coefficient in Figure 6 is negative, corresponding to a small positive estimate of the labor supply elasticity, $-\theta^*$. Both reported confidence intervals for the aggregate include zero. Under the monopsony model, these estimates are consistent with inelastic labor supply with respect to the posted reward after controlling for job characteristics, including text, and hence with potential employer monopsony power in this online labor market.

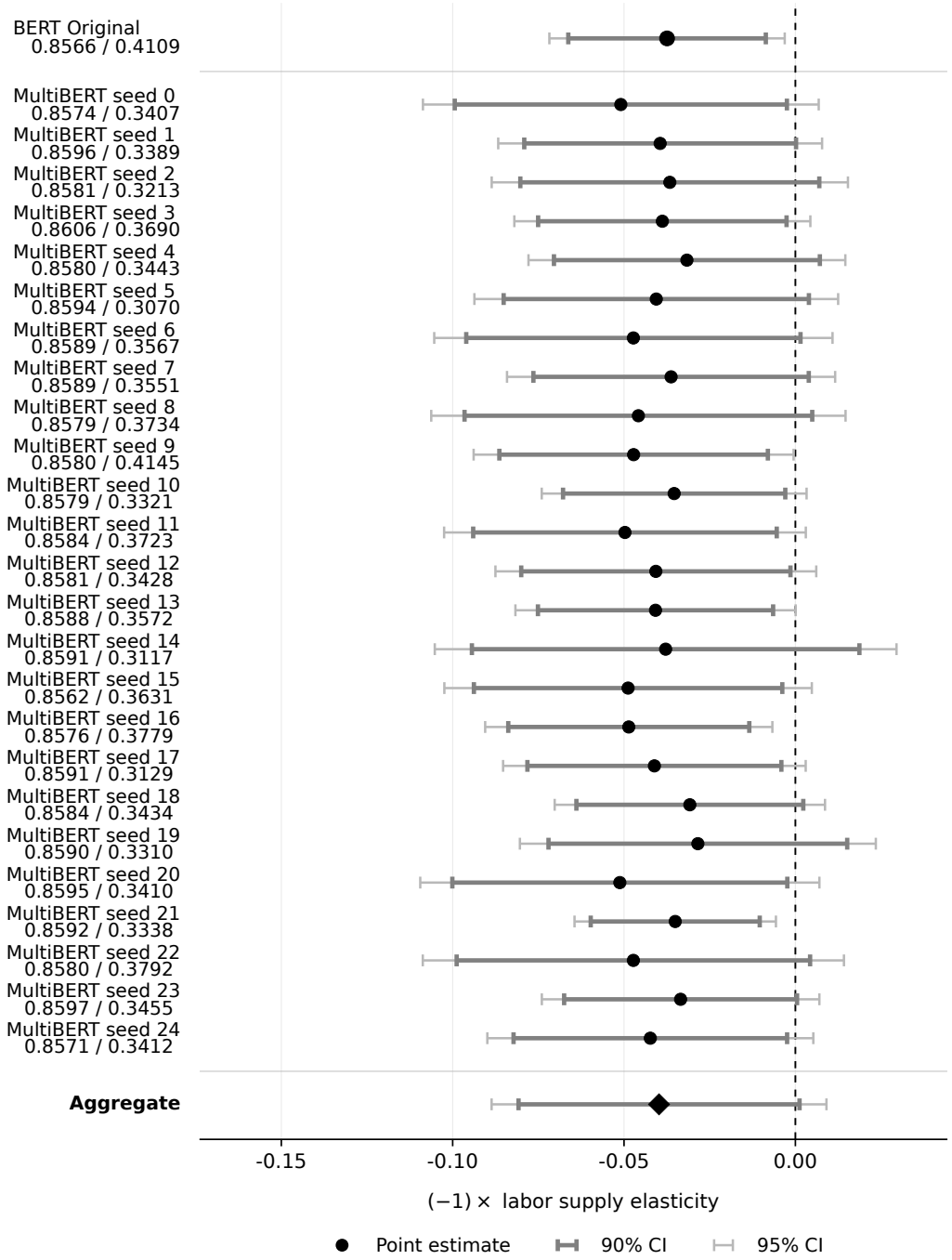


Figure 6: Coefficient robustness to 26 BERT embeddings (Base BERT with Mean pooling) with XGB 2 learners

Notes: Round points show duration coefficients; gray and light-gray bars show pointwise 90% and 95% intervals using requester-clustered standard errors, respectively. Two numbers under the label report duration/reward out-of-fold R^2 averaged across cross-fitting repetitions. The diamond point shows the median coefficient across all embeddings and repetitions (130 estimated values) with the corresponding confidence intervals.

8 Conclusion

This paper establishes sufficient conditions for valid DML inference with pre-trained embeddings, accounting for source estimation error and partial identification of the embedding function. The convergence bounds depend on three components: target estimation error with the pre-trained embeddings treated as standard covariates, source estimation error adjusted for the strength of transferability, and approximation error of the target model class using embeddings as inputs. We also develop a computationally feasible multiplier bootstrap test to detect violations of transferability by comparing DML estimators across embeddings trained for the same source task with different random seeds. Given the partial identification of the embedding function, our theory suggests a new empirical practice: assessing and reporting the robustness of DML estimates across embeddings pre-trained for the same source task with different random seeds.

References

Ahrens, A., V. Chernozhukov, C. Hansen, D. Kozbur, M. E. Schaffer, and T. Wiemann (2026): "An Introduction to Double/Debiased Machine Learning," Journal of Economic Literature. [Cited on pages 1 and 52]

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-powered inference," Science, 382, 669–674. [Cited on page 5]

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 3, 6, and 32]

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 3, 10, and 32]

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 31 and 32]

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 10]

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 37]

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]

Casella, S., J. Fernández-Villaverde, S. Hansen, R. Oishi, and M. Shin (2026): "Structural Estimation with Unstructured Data," Tech. rep., National Bureau of Economic Research. [Cited on page 3]

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 9]

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, 8, 9, 10, 43, 47, 49, and 109]

Christensen, T. and G. Compiani (2026): "From Unstructured Data to Demand Counterfactuals: Theory and Practice," arXiv preprint arXiv:2601.05374. [Cited on page 6]

Compiani, G., I. Morozov, and S. Seiler (2025): "Demand estimation with text and image data," arXiv preprint arXiv:2503.20711. [Cited on pages 3 and 10]

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 14 and 73]

Duan, J. and M. Pelger (2026): "Inference with AI-Generated Covariates," Tech. rep., National Bureau of Economic Research. [Cited on page 5]

Dube, A., J. Jacobs, S. Naidu, and S. Suri (2020): "Monopsony in online labor markets," American Economic Review: Insights, 2, 33–46. [Cited on pages 2, 5, 52, 53, 54, 141, 142, and 145]

Dudley, R. M. (2002): Real analysis and probability, Cambridge University Press. [Cited on page 117]

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 8, 70, and 71]

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep neural networks for estimation and inference," Econometrica, 89, 181–213. [Cited on pages 12, 28, 37, 84, and 117]

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 29]

Folland, G. B. (1999): Real Analysis: Modern Techniques and Their Applications, John Wiley & Sons, 2 ed. [Cited on page 117]

Fong, C. and M. Tyler (2021): "Machine learning predictions as regression covariates," Political Analysis, 29, 467–484. [Cited on page 5]

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 17]

Hahn, J. and G. Ridder (2013): "Asymptotic variance of semiparametric estimators with generated regressors," Econometrica, 81, 315–340. [Cited on page 71]

Han, S. and K. Lee (2025): "Copyright and Competition: Estimating Supply and Demand with Unstructured Data," arXiv preprint arXiv:2501.16120. [Cited on page 3]

Harville, D. A. (2008): Matrix Algebra From a Statistician's Perspective, Springer Science & Business Media. [Cited on pages 33 and 131]

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 37]

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 14 and 72]

Hinton, G. E. and R. R. Salakhutdinov (2006): "Reducing the dimensionality of data with neural networks," Science, 313, 504–507. [Cited on page 14]

Horn, R. and C. R. Johnson (1994): Topics in matrix analysis, Cambridge University Press Cambridge, UK. [Cited on page 131]

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 37]

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 22]

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 3]

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 37]

Kolesár, M., U. K. Müller, and S. T. Roelsgaard (2026): "The fragility of sparsity," arXiv preprint arXiv:2311.02299. [Cited on page 34]

Kosorok, M. R. (2008): Introduction to empirical processes and semiparametric inference, Springer. [Cited on pages 114 and 115]

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 29]

Ledoux, M. and M. Talagrand (1991): Probability in Banach Spaces: isoperimetry and processes, vol. 23, Springer. [Cited on page 130]

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 3]

Mammen, E., C. Rothe, and M. Schienle (2016): "Semiparametric estimation with generated covariates," Econometric Theory, 32, 1140–1177. [Cited on page 71]

McInnes, L., J. Healy, and J. Melville (2018): "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426. [Cited on page 142]

Meinshausen, N., L. Meier, and P. Bühlmann (2009): "p-Values for High-Dimensional Regression," Journal of the American Statistical Association, 104, 1671–1681. [Cited on page 51]

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 17]

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 10]

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 11]

Rubin, D. B. (1976): "Inference and missing data," Biometrika, 63, 581–592. [Cited on page 10]

Schulte, R., D. Rügamer, and T. Nagler (2025): "Adjustment for confounding using pre-trained representations," arXiv preprint arXiv:2506.14329. [Cited on pages 6, 33, and 37]

Sellam, T., S. Yadlowsky, I. Tenney, J. Wei, N. Saphra, A. D'Amour, T. Linzen, J. Bastings, I. R. Turc, J. Eisenstein, D. Das, and E. Pavlick (2022): "The MultiBERTs: BERT Reproductions for Robustness Analysis," in International Conference on Learning Representations. [Cited on page 56]

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 14 and 72]

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 37]

Stewart, G. W. and J.-g. Sun (1990): Matrix Perturbation Theory, Computer Science and Scientific Computing, Boston: Academic Press. [Cited on pages 90 and 131]

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 6, 31, 69, and 85]

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 6]

van der Vaart, A. and J. A. Wellner (2023): Weak Convergence and Empirical Processes: With Applications to Statistics, Springer Nature. [Cited on pages 28, 87, 88, 108, 113, 121, 122, and 125]

Van der Vaart, A. W. (1998): Asymptotic statistics, vol. 3, Cambridge University Press. [Cited on page 115]

Vovk, V., B. Wang, and R. Wang (2022): "Admissible ways of merging p-values under arbitrary dependence," The Annals of Statistics, 50, 351–375. [Cited on pages 51 and 57]

Wainwright, M. J. (2019): High-dimensional statistics: A non-asymptotic viewpoint, vol. 48, Cambridge University Press. [Cited on pages 34 and 116]

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, 31, and 69]

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 37]

Xu, Z. and A. Tewari (2021): "Representation learning beyond linear prediction functions," Advances in Neural Information Processing Systems, 34, 4792–4804. [Cited on page 69]

Yarotsky, D. (2017): "Error bounds for approximations with deep ReLU networks," Neural Networks, 94, 103–114. [Cited on page 32]

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 3.2 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 3.3 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 3.3 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 3.4 and 3.5 are new to the best of our knowledge. For linear heads, Theorem 3.4 establishes the necessity of the row-span condition for $(0, 0)$-diversity under its stated class and regularity assumptions. For general heads, Theorem 3.5 establishes necessary measurable factorization through the source prediction for $(0, 0)$-diversity under its assumptions, including the sigma-field inclusion.

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 4.1 (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 Double Machine Learning with Generated Regressors

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 $\widehat{\eta} = (\widehat{\gamma}, \widehat{\alpha})$ 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 2.1) with squared-loss nuisance training, so that Λ_γ and Λ_α are identity maps. The same logic applies to the other applications with closed-form Neyman orthogonal scores and nuisance functions taking conditional expectation forms, including Examples 2.2 to 2.4.

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 2.1 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 A.1 (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,

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

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).[29] 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 ($\epsilon_{\mathrm{ta},n} > 0$), 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 $H_{L_{\mathrm{NN}}} = T$. 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

[29] 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.

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

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

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 input to the final affine classification layer, with dropout disabled when extracting embeddings. For VGG19 (Simonyan and Zisserman, 2015), we can use the final hidden fully connected layer after activation, with $K = 4{,}096$. For ResNet50 (He et al., 2016), the flattened final average-pooling output has $K = 2{,}048$ coordinates and feeds directly into the final affine classifier. The source head g_m consists of the baseline logit contrasts of the final

affine classifier, so both models have the form $g_m \circ h_m$ with an affine source head. Economists can use the frozen embeddings $\breve{h}(Z_i)$ as regressors in the target task.

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 $T + 1 \approx 30{,}000$, so there are T log-odds relative to a baseline token, as in Section 2.3. 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 $\mathcal{M}_{\text{pred}}(Z) \subseteq \{1, \ldots, M\}$ be the positions eligible for token prediction, excluding `[CLS]`, `[SEP]`, and `[PAD]`. Let $\mathcal{M} \subseteq \mathcal{M}_{\text{pred}}(Z)$ be the random subset selected for prediction and let $\widetilde{Z}$ be the corrupted sequence used as input, where we randomly replace the tokens at positions in $\mathcal{M}$ with a special `[MASK]` token or random tokens. Selected positions contribute to the loss even when their tokens remain unchanged. The MLM head produces raw class logits through a shallow neural network

$$\widetilde{g}_{\text{MLM},m} : \mathbb{R}^K \to \mathbb{R}^{T+1},$$

applied to each contextualized token state $u_m^{(L_{\text{tr}})}(\widetilde{Z}, p)$ for $p \in \mathcal{M}$. This head applies a dense layer with an activation function, followed by layer normalization, before the final affine vocabulary projection. Let

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

collect the MLM and NSP labels for one sequence. If S_p denotes the original token at a selected position p, then, for sequences with a fixed number $|\mathcal{M}| = k \geq 1$ of selected positions, the per-sequence averaged MLM loss is

$$\ell_{\mathrm{MLM}}(\widetilde{Z}, S^{\mathrm{BERT}}) = -\frac{1}{k} \sum_{p \in \mathcal{M}} \log \Pr(S_p \mid \widetilde{Z}, p).$$

For fixed k, averaging this loss across sequences gives the same weighting as normalizing by the total number of selected positions in a batch, as in the original implementation. For NSP, the head produces two raw logits

$$\widetilde{g}_{\mathrm{NSP},m} : \mathbb{R}^K \to \mathbb{R}^2,$$

from the `[CLS]` state $u_m^{(L_{\mathrm{tr}})}(\widetilde{Z}, \texttt{[CLS]})$, including a dense layer with a hyperbolic tangent activation (the pooler) before the final affine classifier. If $S_{\mathrm{NSP}} \in \{0, 1\}$ indicates whether span B truly follows span A, and

$$\pi_{\mathrm{NSP}}(\widetilde{Z}) = \texttt{softmax}(\widetilde{g}_{\mathrm{NSP},m}(u_m^{(L_{\mathrm{tr}})}(\widetilde{Z}, \texttt{[CLS]}))),$$

then the NSP loss is

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

With this normalization, the combined pre-training loss is

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

The source outputs g_m in our theory are obtained by subtracting a baseline logit within each classification task, leaving T MLM contrasts at each selected position and one NSP contrast. To apply the results for affine source heads, we absorb the task-specific transformations into the embeddings. Let $u_{\mathrm{NSP},m}$ denote the dense layer with a hyperbolic tangent activation in the NSP pooler, and let $u_{\mathrm{MLM},m}$ denote the dense layer, activation, and layer normalization before the MLM vocabulary projection. With dropout disabled, one text-level embedding is the input to the final NSP classifier,

$$\breve{h}(Z) = u_{\mathrm{NSP},m} \circ u_m^{(L_{\mathrm{tr}})}(Z, \texttt{[CLS]}).$$

For the MLM-based analysis, we instead use the mean of the transformed states over eligible

positions,

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

where $|\mathcal{M}_{\text{pred}}(Z)| \geq 1$. This average includes all eligible positions, whether or not they were selected in a particular pre-training draw. Its relation to the selected-position source objective is established in Appendix C.2. The economist treats $\breve{h}(Z)$ as the generated regressor in the target sample and estimates the downstream model $f \circ \breve{h}$.[30]

Across fully connected networks, CNNs, and transformers, a shared encoder is learned in the source task, task-specific heads predict the source labels, and an estimated embedding $\breve{h}$ is carried to the target sample. Applying the transfer-learning framework requires matching the chosen embedding to the source outputs and verifying that source prediction error controls the outputs used in the transferability argument.

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 3.3 and Corollary 3.1 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 matrix of slope contrasts has enough rank, and the target class satisfies the required functional class condition, the sufficient conditions follow from Corollary 3.1. When this contrast 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.

Throughout this appendix, source outputs are log-odds relative to a baseline category, as in Section 2.3. To make the normalization explicit, consider raw logits $\widetilde{g}(v) = \widetilde{W}v + \widetilde{b}$ for $T+1$ classes, where $\widetilde{W} \in \mathbb{R}^{(T+1)\times K}$ and $\widetilde{b} \in \mathbb{R}^{T+1}$. With class $T+1$ as the baseline, define $W \in \mathbb{R}^{T\times K}$ and $b \in \mathbb{R}^T$ by

$$\begin{aligned} W_{t,\cdot} &= \widetilde{W}_{t,\cdot} - \widetilde{W}_{T+1,\cdot}, \qquad b_t = \widetilde{b}_t - \widetilde{b}_{T+1}, \\ g_t(v) &= \widetilde{g}_t(v) - \widetilde{g}_{T+1}(v) = W_{t,\cdot}v + b_t, \end{aligned}$$

for $t = 1, \ldots, T$, where $W_{t,\cdot}$ denotes the t-th row of W. Softmax probabilities identify the logit differences $g_t(v)$ and are invariant to adding a common affine function to all raw logits.

[30] 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 $\breve{h}$.

All rank, singular-value, row-span, and pseudo-inverse conditions below therefore concern W, and all intercept adjustments concern b. Candidate source heads use the same baseline normalization. Full column rank of $\widetilde{W}$ alone is insufficient.

C.1 Transferability Condition for Pre-trained CNN

When the CNN embedding is the input to the final affine classifier, the normalized source head has the form

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

where $v \in \mathbb{R}^K$ is the chosen embedding, $W_{\text{cnn},m} \in \mathbb{R}^{T \times K}$, and $b_{\text{cnn},m} \in \mathbb{R}^T$. Here $W_{\text{cnn},m}$ and $b_{\text{cnn},m}$ are the slope and intercept contrasts of the raw classifier with $T+1$ classes, defined relative to the baseline class as above. This decomposition applies to VGG19's final hidden fully connected layer after activation and ResNet50's flattened final average-pooling output, as specified in Appendix B. For their 1,000-class ImageNet versions, $T = 999$ is smaller than both embedding dimensions, so the full-column-rank condition below cannot hold for these embeddings.

Suppose that the target model class $\mathcal{F}_n = \mathcal{F}_n^{\text{sub}}$ is Lipschitz with constant $L_{\mathcal{F}}$ and that Assumption 3.3 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 3.1, the CNN source task is $(\rho_{\text{so},m}, \rho_{\text{ta},n})$-diverse over f_n with

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

provided that the functional class requirement in Corollary 3.1 is satisfied. Specifically, for every embedding function h and corresponding best normalized 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 embedding function 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 3.4 hold, $(0,0)$-diversity can hold only if the target coefficient belongs to the row span of $W_{\mathrm{cnn},m}$. More generally, Theorem 3.5 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. For the population analysis in this subsection, we change the notation of Appendix B: $u^{(\ell)}_m$, $u_{\mathrm{NSP},m}$, $u_{\mathrm{MLM},m}$, and the corresponding source heads now denote population reference maps, with h_m denoting the resulting reference embedding. Fitted maps carry check accents.

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

$$h_m(Z) = u^{(L_{\mathrm{tr}})}_m(Z, \texttt{[CLS]}).$$

BERT's NSP head includes the nonlinear pooler described in Appendix B, so it is nonlinear in this raw `[CLS]` state. To apply Corollary 3.1, the pooler must be absorbed into the embedding, or a head linear in the chosen embedding must be assumed. For a K-dimensional input v to the final linear NSP classifier, let $\widetilde{W}_{\mathrm{NSP},m} \in \mathbb{R}^{2\times K}$ and $\widetilde{b}_{\mathrm{NSP},m} \in \mathbb{R}^2$ be its raw slope matrix and intercept vector, with rows indexed by the labels 0 and 1. Taking class 0 as the baseline gives

$$W_{\mathrm{NSP},m} = \widetilde{W}_{\mathrm{NSP},m,1,\cdot} - \widetilde{W}_{\mathrm{NSP},m,0,\cdot},$$
$$b_{\mathrm{NSP},m} = \widetilde{b}_{\mathrm{NSP},m,1} - \widetilde{b}_{\mathrm{NSP},m,0},$$

and the normalized source head is

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

where $W_{\mathrm{NSP},m} \in \mathbb{R}^{1\times K}$ and $b_{\mathrm{NSP},m} \in \mathbb{R}$. The full-rank sufficient condition in Corollary 3.1 can then hold only if $\mathrm{rank}(W_{\mathrm{NSP},m}) = K$, which requires $K \leq 1$. For empirically relevant transformer embeddings such as BERT, K is much larger than 1. Therefore, next sentence prediction alone cannot generically identify a high-dimensional representation. Under the conditions of Theorem 3.5, transferability based only on this binary source task requires the target function to depend on the chosen representation through the scalar NSP log-odds, including the pooler when the raw `[CLS]` state is used.

Mean-pooled token embedding. For MLM, we pool only positions eligible for prediction and include the MLM-specific transformation in the token states. Let

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

denote the state after BERT's dense activation and layer normalization, before the final vocabulary projection. Recall that $\mathcal{M}_{\mathrm{pred}}(Z)$ excludes the positions of `[CLS]`, `[SEP]`, and padding. The embedding is

$$h_m(Z) = \frac{1}{|\mathcal{M}_{\mathrm{pred}}(Z)|} \sum_{p\in\mathcal{M}_{\mathrm{pred}}(Z)} u_m(Z,p).$$

Thus, every eligible position enters the pool, including positions not selected in a particular corruption draw. Special tokens remain in the encoder input but their states do not enter this

average. The normalized MLM head is

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

where $W_{\mathrm{MLM},m} \in \mathbb{R}^{T \times K}$ and $b_{\mathrm{MLM},m} \in \mathbb{R}^{T}$ are the baseline slope and intercept contrasts of the vocabulary projection. In BERT, the raw vocabulary projection is tied to the input token-embedding matrix. Full column rank of $W_{\mathrm{MLM},m}$ requires $K \leq T$. Since the same linear head is applied at every position,

$$g_{\mathrm{MLM},m} \circ h_m(Z) = \frac{1}{|\mathcal{M}_{\mathrm{pred}}(Z)|} \sum_{p \in \mathcal{M}_{\mathrm{pred}}(Z)} g_{\mathrm{MLM},m}(u_m(Z,p)).$$

This identity does not by itself relate the MLM objective to the pooled output, because the objective evaluates only selected positions. We establish that relation before applying the full-rank result.

From selected-token predictions to raw pooled predictions. Let P_{raw} denote the law of the raw source sequence before corruption. We consider ordinary WordPiece masking, for which we select a prescribed number of positions without replacement, subject to a prediction cap. For the following result, assume that this number is a fixed integer $k \geq 1$ and $|\mathcal{M}_{\mathrm{pred}}(Z)| \geq k$, P_{raw}-a.s. Conditional on Z, let $\mathcal{M}$ be a uniformly drawn k-element subset of $\mathcal{M}_{\mathrm{pred}}(Z)$. Independently across selected positions, replace the token by `[MASK]` with probability 0.8, draw a uniform vocabulary replacement with probability 0.1, and retain the original token with probability 0.1. Positions outside $\mathcal{M}$ are unchanged. With

$$q = 0.1 + \frac{0.1}{T+1},$$

the resulting corrupted input satisfies

$$\Pr(\widetilde{Z} = Z \mid Z, \mathcal{M}) = q^k. \tag{56}$$

The second term in q accounts for a random replacement matching the original token.

For a candidate token-state map u and shared affine head $g(v) = Wv + b$, define its pooled

embedding and its token-level difference from the reference model by

$$h(Z) = \frac{1}{|\mathcal{M}_{\text{pred}}(Z)|} \sum_{p \in \mathcal{M}_{\text{pred}}(Z)} u(Z, p),$$
$$\Delta_p(Z) = g(u(Z, p)) - g_{\text{MLM},m}(u_m(Z, p)),$$

and define the population discrepancy at selected prediction positions by

$$D_{\text{MLM}}(g, u)^2 = \mathbb{E}\left[\frac{1}{k} \sum_{p \in \mathcal{M}} \|\Delta_p(\widetilde{Z})\|_2^2\right], \tag{57}$$

where the expectation integrates over $Z \sim P_{\text{raw}}$, the selected subset, and the corruption draws.

Lemma C.1 (MLM Prediction Bound for Eligible-Token Pooling)**.** *Under the selection and corruption conditions above, every candidate pair (g, u) with a shared affine head and $D_{\text{MLM}}(g, u) < \infty$ satisfies*

$$\|g \circ h - g_{\text{MLM},m} \circ h_m\|_{P_{\text{raw}},2} \le q^{-k/2} D_{\text{MLM}}(g, u). \tag{58}$$

Proof. Conditional on Z and $\mathcal{M}$, the prediction maps on the event $\{\widetilde{Z} = Z\}$ equal their values at the raw input. Hence (56) gives

$$\mathbb{E}\left[\mathbb{1}\{\widetilde{Z} = Z\} \frac{1}{k} \sum_{p \in \mathcal{M}} \|\Delta_p(\widetilde{Z})\|_2^2 \mid Z, \mathcal{M}\right]$$
$$= q^k \frac{1}{k} \sum_{p \in \mathcal{M}} \|\Delta_p(Z)\|_2^2.$$

Since $\Pr(p \in \mathcal{M} \mid Z) = k/|\mathcal{M}_{\text{pred}}(Z)|$ for every $p \in \mathcal{M}_{\text{pred}}(Z)$,

$$\mathbb{E}\left[\frac{1}{k} \sum_{p \in \mathcal{M}} \|\Delta_p(Z)\|_2^2 \mid Z\right] = \frac{1}{|\mathcal{M}_{\text{pred}}(Z)|} \sum_{p \in \mathcal{M}_{\text{pred}}(Z)} \|\Delta_p(Z)\|_2^2.$$

Dropping the nonnegative contribution from $\{\widetilde{Z} \neq Z\}$ in (57) and taking expectations therefore yields

$$D_{\text{MLM}}(g, u)^2 \ge q^k \mathbb{E}_{\text{raw}}\left[\frac{1}{|\mathcal{M}_{\text{pred}}(Z)|} \sum_{p \in \mathcal{M}_{\text{pred}}(Z)} \|\Delta_p(Z)\|_2^2\right].$$

The shared affine heads imply

$$g \circ h(Z) - g_{\mathrm{MLM},m} \circ h_m(Z) = \frac{1}{|\mathcal{M}_{\mathrm{pred}}(Z)|} \sum_{p \in \mathcal{M}_{\mathrm{pred}}(Z)} \Delta_p(Z).$$

Jensen's inequality now gives

$$\begin{aligned} \|g \circ h - g_{\mathrm{MLM},m} \circ h_m\|_{P_{\mathrm{raw}},2}^2 &\le \mathbb{E}_{\mathrm{raw}} \left[\frac{1}{|\mathcal{M}_{\mathrm{pred}}(Z)|} \sum_{p \in \mathcal{M}_{\mathrm{pred}}(Z)} \|\Delta_p(Z)\|_2^2 \right] \\ &\le q^{-k} D_{\mathrm{MLM}}(g,u)^2. \end{aligned}$$

Taking square roots proves (58). □

Implication for transferability. Suppose that $\mathcal{F}_n^{\mathrm{sub}}$ consists of $L_{\mathcal{F}}$-Lipschitz functions and fix $f_n \in \mathcal{F}_n^{\mathrm{sub}}(h_m)$. Assume that $W_{\mathrm{MLM},m}$ has full column rank and that Assumption 3.3 holds with $P_{\mathrm{so}} = P_{\mathrm{raw}}$, so the density of the raw source sequence relative to the target density is at least $\underline{w}_n^2$. Suppose also that the class of pooled embeddings and the affine source-head class satisfy the target functional class condition in Corollary 3.1, with source prediction distances measured under P_{raw}. In particular, for every pooled embedding h and its best affine head $v \mapsto W^\bullet v + b^\bullet$ approximating $g_{\mathrm{MLM},m} \circ h_m$ under P_{raw}, the target class must contain a function agreeing P_{ta}-a.s. on $h(Z)$ with

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

For any candidate pair with $D_{\mathrm{MLM}}(g,u) \le \rho_{\mathrm{so},m}$, Lemma C.1 places its pooled embedding in the raw-source near-identified set at radius $q^{-k/2}\rho_{\mathrm{so},m}$. Applying Corollary 3.1 at this enlarged radius gives

$$\min_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - f_n \circ h_m\|_{P_{\mathrm{ta}},2} \le \frac{L_{\mathcal{F}}}{\underline{w}_n \sigma_{\min}(W_{\mathrm{MLM},m}) q^{k/2}} \rho_{\mathrm{so},m}. \tag{59}$$

Here $\rho_{\mathrm{so},m}$ bounds selected-token prediction error, and $\underline{w}_n$ compares raw source and target sequences. The factor $q^{-k/2}$ already accounts for corruption. Let $\breve{u}$ denote the fitted counterpart of the transformed token-state map u_m, let $\breve{g}$ denote its fitted affine MLM head, and let $\breve{h}$ denote the mean of $\breve{u}(Z,p)$ over eligible positions, as in Appendix B. If these fitted maps satisfy $D_{\mathrm{MLM}}(\breve{g},\breve{u}) = O_P(\delta_{\mathrm{so},m})$, their pooled embedding satisfies

$$\min_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ \breve{h} - f_n \circ h_m\|_{P_{\mathrm{ta}},2} = O_P\left(\frac{L_{\mathcal{F}} \delta_{\mathrm{so},m}}{\underline{w}_n \sigma_{\min}(W_{\mathrm{MLM},m}) q^{k/2}}\right).$$

This selected-token convergence condition must be established separately. Relating cross-entropy regret to (57) requires curvature and approximation conditions for the selected-token logit contrasts, as in Assumptions 3.1 and 4.2, with sequences as the sampling units.

This bound is very 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 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. A sharper comparison requires additional restrictions on how corruption changes the token-state functions.

D Source Head Transformation for Ridge Regression

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

$$\begin{aligned} R_B &:= (B'B)^{1/2} + I_K - B^+B, \\ B^{\mathrm{N}} &:= BR_B^{-1}, \\ h_B^{\mathrm{N}} &:= R_B h. \end{aligned} \tag{60}$$

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 $(B^{\mathrm{N}})^+ = (B^{\mathrm{N}})'$.[31] 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 $\overline{\sigma} = \underline{\sigma} = 1$.

[31] 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.

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 3.1 (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

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

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 E.1 (Transferability Decomposition)**.** *Let P denote the distribution of Z, and suppose that $\mathcal{F}_n^{\mathrm{sub}}$ and $\mathcal{H}_m$ are classes of measurable functions. Fix measurable functions f_n and h_m such that $f_n \circ h_m \in L^2(P)$. Then, for every $h \in \mathcal{H}_m$ such that $f \circ h \in L^2(P)$ for every $f \in \mathcal{F}_n^{\mathrm{sub}}$,*

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

Remark E.1. 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 3.3 ensures that both terms are small when $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$. Our necessary conditions in Theorems 3.4 and 3.5 make the first term equal to zero for some $h \in \mathcal{H}_{\mathrm{so},m}(0)$. Example 3.2 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 E.2 (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 E.3 (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 E.4 (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 E.2. 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 E.5 (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 E.6 (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 E.1 (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 E.7 (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,

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

Lemma E.8 (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 E.9 (Covering Number for VC Function Classes, van der Vaart and Wellner, 2023, Theorem 2.6.7)**.** *For a nonempty VC-class of functions with measurable envelope function F and $r \geq 1$, every probability measure Q with $\|F\|_{Q,r} > 0$ satisfies*

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

for $0 < \varepsilon < 1$.

Lemma E.10 (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,

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \sup_{f \in \mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f_{n,h}^*\right) \leq \delta} M_n(f, h) - M_n\left(f_{n,h}^*, h\right) \lesssim_\varepsilon -\tau_n^2 \delta^2, \tag{61}$$

and

$$\sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{f \in \mathcal{F}_n : d_{n,h}\left(f, f_{n,h}^*\right) < \delta} \sqrt{n}\left|\left(\mathbb{M}_n - M_n\right)(f, h) - \left(\mathbb{M}_n - M_n\right)\left(f_{n,h}^*, h\right)\right|\right] \lesssim_\varepsilon \phi_n(\delta) \tag{62}$$

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*

$$\phi_n\left(\delta_n/\tau_n\right) \lesssim_\varepsilon \sqrt{n}\delta_n^2, \quad \delta_n^2 \gtrsim_\varepsilon \sup_{h\in\mathcal{H}_{m,\varepsilon}} M_n\left(f_{n,h}^*, h\right) - M_n\left(f_{n,h}, h\right), \quad \delta_n \gtrsim_\varepsilon \underline{\delta}_n. \tag{63}$$

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

Remark E.3. 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 E.11 (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* $\breve{h} \in \mathcal{H}_m$ *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* $\limsup_m P(\breve{h} \notin \mathcal{H}_{m,\varepsilon}) \leq \varepsilon$. *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,

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \sup_{f\in\mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f_{n,h}^*\right) \leq \delta} M_n(f,h) - M_n\left(f_{n,h}^*, h\right) \lesssim_\varepsilon -\tau_n^2\delta^2, \tag{64}$$

and

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}_n : d_{n,h}\left(f, f_{n,h}^*\right) < \delta \\ \mathcal{J}_n(f) < \tau_n\delta/\lambda_n}} \sqrt{n}\left|\left(\mathbb{M}_n - M_n\right)(f,h) - \left(\mathbb{M}_n - M_n\right)\left(f_{n,h}^*, h\right)\right|\right] \lesssim_\varepsilon \phi_n(\delta) \tag{65}$$

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*

$$\phi_n\left(\delta_n/\tau_n\right) \lesssim_\varepsilon \sqrt{n}\delta_n^2, \quad \delta_n^2 \gtrsim_\varepsilon \sup_{h\in\mathcal{H}_{m,\varepsilon}} M_n\left(f_{n,h}^*, h\right) - M_n\left(f_{n,h}, h\right), \quad \delta_n \gtrsim_\varepsilon \underline{\delta}_n,$$
$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \lambda_n \mathcal{J}_n\left(f_{n,h}\right) < \delta_n, \quad \sup_{h\in\mathcal{H}_{m,\varepsilon}} d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) < \delta_n/\tau_n. \tag{66}$$

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*

$$\mathbb{M}_n\left(\widehat{f}_{n,\check{h}}, \check{h}\right) - \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) \geq \mathbb{M}_n\left(f_{n,\check{h}}, \check{h}\right) - \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(f_{n,\check{h}}\right) - O_P\left(\delta_n^2\right), \tag{67}$$

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 E.12 (Curvature Condition for Approximate Maximizer)**.** *Let* $d_{n,h}(\cdot,\cdot)$ *be a pseudo-metric on a set containing* $\mathcal{F}_n \cup \mathcal{F}_n^*$ *for every fixed* $h \in \mathcal{H}_m$. *Distinct heads may have zero distance when their compositions with* h *agree almost surely. 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* $\delta > 0$,

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \sup_{f\in\mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f_{n,h}^*\right) \leq \delta} M_n(f,h) - M_n\left(f_{n,h}^*, h\right) \lesssim_\varepsilon -\tau_n^2\delta^2 \tag{68}$$

and

$$d_{n,h}\left(f_{n,h}, f_{n,h}^*\right)^2 \gtrsim_\varepsilon M_n\left(f_{n,h}^*, h\right) - M_n\left(f_{n,h}, h\right) \tag{69}$$

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,

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \sup_{f\in\mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f_{n,h}\right) \leq \delta} M_n(f,h) - M_n\left(f_{n,h}, h\right) \lesssim_\varepsilon -\tau_n^2\delta^2. \tag{70}$$

Lemma E.13 (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 E.14 (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*

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

Lemma E.15. *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 E.16. *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*

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

Lemma E.17. *For every matrix* $A \in \mathbb{R}^{T \times K}$ *with* $\text{rank}(A) \geq 1$, *the spectral norm of its Moore-Penrose inverse satisfies*

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

Lemma E.18 (Bound on Effective Degrees of Freedom)**.** *Let* $\mathcal{N}_A(\lambda)$ *be the effective degrees of freedom defined as* $\mathcal{N}_A(\lambda) = \text{tr}\left(A(A + \lambda^2 I)^{-1}\right)$ *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 E.19 (Transformation on Effective Degrees of Freedom)**.** *Let* $\mathcal{N}_A(\lambda)$ *be the effective degrees of freedom defined as in Lemma E.18. Let* $Q \in \mathbb{R}^{K \times K}$ *be another matrix. Then,*

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

F Proofs for Main Results

F.1 Proof for Theorem 3.1

Proof. By the triangle inequality, we have

$$\begin{aligned}\left\|\widehat{f}\circ\check{h}-a^*_{\mathrm{ta}}\right\|_{P_{\mathrm{ta},2}} &\le \left\|\widehat{f}\circ\check{h}-f^\bullet_n\circ\check{h}\right\|_{P_{\mathrm{ta},2}}+\left\|f^\bullet_n\circ\check{h}-f_n\circ h_m\right\|_{P_{\mathrm{ta},2}}+\|f_n\circ h_m-a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta},2}}\\ &=: \mathrm{(I)}+\mathrm{(II)}+\mathrm{(III)}. \end{aligned} \tag{71}$$

We bound each term separately.

By (14) in Assumption 3.1 (ii),

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

Next, we bound the second term (II). By (13) in Assumption 3.1 (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 3.2 and (12) 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

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

By Definition 3.1,

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

Combining the three bounds yields

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

which proves the claim. □

F.2 Proof for Theorem 3.2

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

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

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

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

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

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

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

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

We claim that $\sup_{h \in \mathcal{H}_{\text{so},m}(0)} A_n(h) = o(1)$. Otherwise, there would exist $\varepsilon > 0$, a subsequence $\{n_j\}$, and deterministic functions $h^\dagger_{m_{n_j}} \in \mathcal{H}_{\text{so},m_{n_j}}(0)$ such that $A_{n_j}(h^\dagger_{m_{n_j}}) \ge \varepsilon$ 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 3.3

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 $f \in \mathcal{F}_n^{\mathrm{sub}}$,

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

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

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

where the second inequality follows from the existence of $f \in \mathcal{F}_n^{\mathrm{sub}}$ such that $\|f \circ h - \Gamma \circ g_m^\bullet \circ h\|_{P_{\mathrm{ta}},2} \leq \varrho_n$. The third inequality follows from the Lipschitz continuity of Γ, and the last inequality follows from Assumption 3.3 and the definition of $\mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$. □

F.4 Proof for Corollary 3.1

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 $L_\Gamma \leq L_{\mathcal{F}} \|B^+_{\mathrm{so},m}\|_{\mathrm{sp}}$ since

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

for every $x_1, x_2 \in \mathbb{R}^T$, where the last inequality follows from Lemma E.16.

Fix $\rho_{\mathrm{so},m} \geq 0$ and $h \in \mathcal{H}_{\mathrm{so},m}(\rho_{\mathrm{so},m})$, and let $(\alpha^\bullet_{\mathrm{so},m}, B^\bullet_{\mathrm{so},m})$ be the minimizer in the statement of the corollary. Define $g^\bullet_m(v) = \alpha^\bullet_{\mathrm{so},m} + B^\bullet_{\mathrm{so},m} v$. By the assumption, there exists $f \in \mathcal{F}^{\mathrm{sub}}_n$ such that

$$\begin{aligned}f \circ h &= f_n\left(B^+_{\mathrm{so},m}(\alpha^\bullet_{\mathrm{so},m} - \alpha_{\mathrm{so},m} + B^\bullet_{\mathrm{so},m} h)\right) \\ &= \Gamma \circ g^\bullet_m \circ h,\end{aligned}$$

P_{ta}-a.s. Therefore, by Theorem 3.3, 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 E.17 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 3.4

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

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

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

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

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 3.5

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 $P_{\mathrm{ta}}(f_n \circ h_m(Z) \neq \Gamma_0 \circ h_0(Z)) > 0$. The assumed square integrability

of $f_n \circ h_m$ and $f \circ h_0$ under P_{ta} allows us to apply Lemma E.1 with $P = P_{\mathrm{ta}}$ and $h = h_0$:

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

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 E.2, 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 4.1

Proof. By Theorem 3.1, it suffices to verify (14) 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

$$f_{n,h}^{\bullet} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - f_n \circ h_m\|_{P_{\mathrm{ta}},2},$$

and retain the shorthand $f_n^{\bullet} = f_{n,\check{h}}^{\bullet}$. By definition and the inclusion $\mathcal{F}_n^{\mathrm{sub}} \subseteq \mathcal{F}_n$, we have

$$f_{n,h}^{\bullet} \in \mathcal{F}_n^{\mathrm{sub}} \subseteq \mathcal{F}_n.$$

Moreover, $\widehat{f} \in \mathcal{F}_n$ by the statement of the theorem. If $\mathsf{V}(\mathcal{F}_n) = 0$, the nonempty class $\mathcal{F}_n$ contains only one pointwise function. Then $\widehat{f} = f_{n,h}^{\bullet}$ for every h, so the target estimation error is zero and the general theorem gives the result. We therefore consider $\mathsf{V}(\mathcal{F}_n) \geq 1$ and enlarge M_F to at least one if necessary. By (13) in Assumption 3.1 (i), for every $\varepsilon > 0$, there exists a constant $M_\varepsilon < \infty$ such that

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

We will apply Lemma E.10 with $\mathcal{H}_{m,\varepsilon} = \mathcal{H}_{\mathrm{so},m}(M_\varepsilon \delta_{\mathrm{so},m})$, $\mathcal{F}_n$ as given in the theorem, $f^*_{n,h} = f_{n,h} = f^\bullet_{n,h}$, $\widehat{f}_{n,h} = \widehat{f}$, $P = P_{\mathrm{ta}}$, $M_n(f,h) = -\mathbb{E}_{\mathrm{ta}}[\ell_{\mathrm{ta}}(f \circ h(Z), Y)]$, $\mathbb{M}_n(f,h) = -\frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{ta}}(f \circ h(Z_i), Y_i)$, $d_{n,h}(f,f') = \|f \circ h - f' \circ h\|_{P_{\mathrm{ta}},2}$, and $\tau_n = 1$, where the notation on the left-hand side is that of Lemma E.10.[32]

By Lemma E.8, the pointwise measurability condition in Lemma E.10 is satisfied. Note that $f^\bullet_{n,h}$ and $\widehat{f}$ depend on h through the minimization problem given h.

Since $f^\bullet_n \circ \breve{h}$ need not equal a^*_{ta}, we cannot directly apply the curvature condition in Assumption 4.1 to $f^\bullet_n \circ \breve{h}$ to verify (61). For this purpose, we first apply Lemma E.12 with $\mathcal{H}_{m,\varepsilon} = \mathcal{H}_{\mathrm{so},m}(M_\varepsilon \delta_{\mathrm{so},m})$, $\mathcal{F}_n$ as given in the theorem, $f_{n,h} = f^\bullet_{n,h}$ and $f^*_{n,h} = a^*_{\mathrm{ta}}$. We can bound

$$\begin{aligned}
&\sup_{h \in \mathcal{H}_{m,\varepsilon}} \|f^\bullet_{n,h} \circ h - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \\
&\quad \leq \sup_{h \in \mathcal{H}_{\mathrm{ta},m}(M_\varepsilon \delta_{\mathrm{so},m}/\nu_n)} \|f^\bullet_{n,h} \circ h - f_n \circ h_m\|_{P_{\mathrm{ta}},2} + \|f_n \circ h_m - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \\
&\quad \leq M_\varepsilon \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}, \qquad (72)
\end{aligned}$$

where the first inequality follows from the triangle inequality and the inclusion $\mathcal{H}_{m,\varepsilon} \subseteq \mathcal{H}_{\mathrm{ta},m}(M_\varepsilon \delta_{\mathrm{so},m}/\nu_n)$, which follows from the definition of $\mathcal{H}_{m,\varepsilon}$ and Assumption 3.2. The second inequality follows from $f^\bullet_{n,h} \in \arg\min_{f \in \mathcal{F}_n^{\mathrm{sub}}} \|f \circ h - f_n \circ h_m\|_{P_{\mathrm{ta}},2}$, the definition of $\mathcal{H}_{\mathrm{ta},m}(\cdot)$, and Definition 3.1. If we set $\underline{\delta}_n = \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}$, we obtain the conditions in Lemma E.12 with $\tau_n = 1$ by the curvature condition in Assumption 4.1. Thus, Lemma E.12 implies that there exists a constant $C_{\varepsilon,0} > 0$ such that for every $\delta > C_{\varepsilon,0}(\delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n})$, (61) in Lemma E.10 holds with $f_{n,h} = f^*_{n,h} = f^\bullet_{n,h}$ and $\tau_n = 1$.

Next, we verify (62). By Lemma E.7 with $P = P_{\mathrm{ta}}$, $\ell = \ell_{\mathrm{ta}}$, $C_\ell = C_{\ell,\mathrm{ta}}$, and $M = M_F$,

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

[32] To apply Lemma E.12 below, extend these definitions to the best target predictor a^*_{ta} by setting, for every $f \in \mathcal{F}_n$ and $h \in \mathcal{H}_m$, $M_n(a^*_{\mathrm{ta}}, h) = -\mathbb{E}_{\mathrm{ta}}[\ell_{\mathrm{ta}}(a^*_{\mathrm{ta}}(Z), Y)]$, $\mathbb{M}_n(a^*_{\mathrm{ta}}, h) = -\frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{ta}}(a^*_{\mathrm{ta}}(Z_i), Y_i)$, $d_{n,h}(f, a^*_{\mathrm{ta}}) = d_{n,h}(a^*_{\mathrm{ta}}, f) = \|f \circ h - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2}$, and $d_{n,h}(a^*_{\mathrm{ta}}, a^*_{\mathrm{ta}}) = 0$.

For $0 < \delta \le M_F$, Lemma E.9 gives

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

For $\delta > M_F$, split the entropy integral at one and use monotonicity of covering numbers to obtain $J(\delta/M_F, \mathcal{F}_n|M_F, L_2) \lesssim (\delta/M_F)\sqrt{\mathsf{V}(\mathcal{F}_n)}$. For $n \ge 2$, define the following modulus on every nonnegative radius:

$$\phi_n(\delta) := \delta\sqrt{n}r_{\mathrm{ta},n} + M_F\sqrt{n}r_{\mathrm{ta},n}^2.$$

For $\delta \ge M_F/\sqrt{n}$, the entropy bounds above and $\log(en) \lesssim \log(n)$ show that the preceding empirical-process expectation is bounded by a constant multiple of $\phi_n(\delta)$, uniformly in h. For $0 < \delta < M_F/\sqrt{n}$, nesting of the localized classes bounds that expectation by its value at $M_F/\sqrt{n}$. Since $\mathsf{V}(\mathcal{F}_n) \ge 1$,

$$\phi_n(M_F/\sqrt{n}) \lesssim M_F\sqrt{n}r_{\mathrm{ta},n}^2 \le \phi_n(\delta).$$

Thus (62) holds for every positive radius. The function ϕ_n is increasing, and

$$\frac{\phi_n(\delta)}{\delta} = \sqrt{n}r_{\mathrm{ta},n} + \frac{M_F\sqrt{n}r_{\mathrm{ta},n}^2}{\delta}$$

is nonincreasing for $\delta > 0$, so we may take $\xi = 1$ in Lemma E.10. Moreover, $\phi_n(\delta) \lesssim \sqrt{n}\delta^2$ whenever $\delta \ge r_{\mathrm{ta},n}$. Taking $\delta_n = r_{\mathrm{ta},n} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}$ also satisfies the approximation and optimization requirements in (63) and the approximate maximization condition, by (17) and $f_{n,h} = f_{n,h}^*$. By Lemma E.10, we obtain the following convergence rate using the condition (63):

$$\left\|\widehat{f}\circ\breve{h} - f_n^\bullet\circ\breve{h}\right\|_{P_{\mathrm{ta}},2} = O_P\left(\sqrt{\frac{\mathsf{V}(\mathcal{F}_n)\log(n)}{n}} + \delta_{\mathrm{so},m}/\nu_n + \epsilon_{\mathrm{ta},n}\right). \tag{73}$$

Thus, Assumption 3.1 (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 3.1 yields (18), since adding the transfer and approximation terms once more does not change the resulting stochastic order. This completes the proof. □

F.8 Proof for Theorem 4.2

Proof. If all coordinate classes $\mathcal{H}_{m,k}$ and $\mathcal{G}_{m,t}$ have VC dimension zero, the composite class $\mathcal{G}_m \circ \mathcal{H}_m$ contains only one pointwise function. Then $\breve{g} \circ \breve{h} = g_m \circ h_m$, and the error relative to a^*_{so} equals $\epsilon_{\mathrm{so},m}$. Both conclusions follow since $\tau_m \le 1$. We therefore consider the case in which at least one coordinate VC dimension is positive. As in the proof of Theorem 4.1, we will apply Lemma E.10, 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 $P = P_{\mathrm{so}}$. We define the objective functions for joint optimization over g and h, rather than optimization over f for fixed h. For every $a, \widetilde{a} \in (\mathcal{G}_m \circ \mathcal{H}_m) \cup \{a^*_{\mathrm{so}}\}$, define $M_m(a) = -\mathbb{E}_{\mathrm{so}}[\ell_{\mathrm{so}}(a(Z), S)]$, $\mathbb{M}_m(a) = -\frac{1}{m}\sum_{j=1}^m \ell_{\mathrm{so}}(a(Z_j), S_j)$, and $d_m(a, \widetilde{a}) = \|a - \widetilde{a}\|_{P_{\mathrm{so}},2}$.

The curvature condition (61) is verified as in the proof of Theorem 4.1, by applying Lemma E.12 with $f_m = g_m \circ h_m$, $f^*_m = a^*_{\mathrm{so}}$, and $\underline{\delta}_m = \epsilon_{\mathrm{so},m}$. That is, there exists a constant $C > 0$ such that for every $\delta > \tau_m^{-1} C \epsilon_{\mathrm{so},m}$, (61) holds.

The constant function M_G is an envelope for $\mathcal{G}_m \circ \mathcal{H}_m$ (enlarge it so that $M_G \ge \max\{1, M_H\}$ if necessary). By Lemmas E.5, E.6 and E.9 and the uniform Lipschitz condition in Assumption 4.2 (iii), for $0 < \varepsilon \le 1$ and every discrete probability measure Q,

$$1 + \log N\left(\varepsilon M_G, \mathcal{G}_m \circ \mathcal{H}_m, \|\cdot\|_{Q,2}\right) \lesssim \sum_{k=1}^{K} \mathsf{V}(\mathcal{H}_{m,k}) \log\left(\frac{e L_{\mathcal{G},m} K}{\varepsilon}\right) + \sum_{t=1}^{T} \mathsf{V}(\mathcal{G}_{m,t}) \log\left(\frac{eT}{\varepsilon}\right).$$

The positive coordinate dimension ensures that the right-hand side also bounds the constant one. Integrating as in the target proof, and using monotonicity of covering numbers beyond radius one, gives, for every $\delta > 0$,

$$J\left(\delta, \mathcal{G}_m \circ \mathcal{H}_m | M_G, L_2\right) \lesssim \delta \left[\sum_{k=1}^{K} \mathsf{V}(\mathcal{H}_{m,k}) \log\left(\frac{e L_{\mathcal{G},m} K}{\delta \wedge 1}\right) + \sum_{t=1}^{T} \mathsf{V}(\mathcal{G}_{m,t}) \log\left(\frac{eT}{\delta \wedge 1}\right)\right]^{1/2}.$$

For $m \ge 2$, define

$$\phi_m(\delta) := \delta \tau_m \sqrt{m} r_{\mathrm{so},m} + M_G \tau_m^2 \sqrt{m} r^2_{\mathrm{so},m}, \qquad \delta \ge 0.$$

For $\delta \ge M_G/\sqrt{m}$, apply Lemma E.7 with $\mathcal{F}_n = \mathcal{G}_m \circ \mathcal{H}_m$, $C_\ell = C_{\ell,\mathrm{so}}$, and $M = M_G$. The entropy bound and (20) imply (62) with this ϕ_m, since the logarithms are bounded by constant multiples of $\log(m L_{\mathcal{G},m} K)$ and $\log(mT)$. For a smaller positive radius, nesting bounds the

empirical-process expectation by its value at $M_G/\sqrt{m}$. Since at least one coordinate VC dimension is positive,

$$\phi_m(M_G/\sqrt{m}) \lesssim M_G\tau_m^2\sqrt{m}r_{\mathrm{so},m}^2 \le \phi_m(\delta).$$

Hence the modulus bound holds at every positive radius. The function ϕ_m is increasing and $\phi_m(\delta)/\delta$ is nonincreasing, so the rate lemma applies with $\xi = 1$. For $\delta \ge r_{\mathrm{so},m}$,

$$\phi_m(\delta/\tau_m) = \delta\sqrt{m}r_{\mathrm{so},m} + M_G\tau_m^2\sqrt{m}r_{\mathrm{so},m}^2 \lesssim \sqrt{m}\delta^2.$$

Thus $\delta_m = r_{\mathrm{so},m} + \epsilon_{\mathrm{so},m}$ satisfies the rate and approximation conditions in (63); the approximate maximization condition follows from (21). By Lemma E.10, we obtain the following convergence rate using the condition (63):

$$\left\|\check{g}\circ\check{h} - g_m\circ h_m\right\|_{P_{\mathrm{so}},2} = O_{P_{\mathrm{so}}}\left(r_{\mathrm{so},m}/\tau_m + \epsilon_{\mathrm{so},m}/\tau_m\right). \tag{74}$$

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 (23). This completes the proof. □

F.9 Proof for Lemma 4.1

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

$$\begin{aligned}\left\|\check{h}^{\mathrm{proj}} - (\check{q} + \check{Q}h_m)\right\|_{P_{\mathrm{so}},2} &= \left\|\check{B}^+_{\mathrm{so}}\left(\check{\alpha}_{\mathrm{so}} + \check{B}_{\mathrm{so}}\check{h} - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m}h_m)\right)\right\|_{P_{\mathrm{so}},2} \\ &\le \|\check{B}^+_{\mathrm{so}}\|_{\mathrm{sp}}\left\|\check{\alpha}_{\mathrm{so}} + \check{B}_{\mathrm{so}}\check{h} - (\alpha_{\mathrm{so},m} + B_{\mathrm{so},m}h_m)\right\|_{P_{\mathrm{so}},2}.\end{aligned} \tag{75}$$

The inequality follows from Lemma E.16. By Assumption 3.1 (i), the second factor on the right-hand side of (75) is $O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m})$. Combining this rate with $\|\check{B}^+_{\mathrm{so}}\|_{\mathrm{sp}} = O_{P_{\mathrm{so}}}(\underline{\sigma}_m^{-1})$ proves the stated result. Finally, on the event in the sufficient condition, Lemma E.17 gives $\|\check{B}^+_{\mathrm{so}}\|_{\mathrm{sp}} \le \underline{\sigma}_m^{-1}$, which verifies the condition in (25). □

F.10 Proof for Theorem 4.3

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

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

By definition, $(\alpha_{\mathrm{so},m,h}^{\bullet}, B_{\mathrm{so},m,h}^{\bullet}) \in \Theta_{\mathrm{so},m}$. Indeed, Assumption 4.3 implies, uniformly over $h \in \mathcal{H}_m^{\mathrm{proj}}$,

$$\begin{aligned}
|\alpha_{\mathrm{ta},n,h}^{\bullet}| &\le |\alpha_{\mathrm{ta},n}| + \|b'_n B_{\mathrm{so},m} B_{\mathrm{so},m}^{+}\|_2 \|\alpha_{\mathrm{so},m,h}^{\bullet} - \alpha_{\mathrm{so},m}\|_2 \le |\alpha_{\mathrm{ta},n}| + 2C_b C_{\alpha,\mathrm{so}} \le C_\alpha, \\
\|\beta_{\mathrm{ta},n,h}^{\bullet}\|_2 &\le \|b'_n B_{\mathrm{so},m} B_{\mathrm{so},m}^{+}\|_2 \|B_{\mathrm{so},m,h}^{\bullet}\|_{\mathrm{sp}} \le C_b \overline{\sigma} \le C_\beta.
\end{aligned} \tag{76}$$

Hence, $f_{n,h}^{\bullet} \in \mathcal{F}_n$ for every $h \in \mathcal{H}_m^{\mathrm{proj}}$. For each fixed $h \in \mathcal{H}_m^{\mathrm{proj}}$, 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 $\widehat{f}_{n,\check{h}^{\mathrm{proj}}} = \widehat{f}_n^{\mathrm{ridge}}$, and let $(\widehat{\alpha}_{\mathrm{ta},n,h}, \widehat{\beta}_{\mathrm{ta},n,h})$ denote the corresponding parameters.

As in (71), we decompose the estimation error of the target model into three terms.

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

By Definition 3.1, $\mathrm{(III)} = \epsilon_{\mathrm{ta},n}$. 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 $h \in \mathcal{H}_m^{\mathrm{proj}}$,

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

where the inequality follows from Assumption 3.3 and Lemma E.16. 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 4.3 (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 3.1 (i) gives

$$\begin{aligned}
\|g_{\alpha^{\bullet}_{\mathrm{so},m,\breve{h}^{\mathrm{proj}}}, B^{\bullet}_{\mathrm{so},m,\breve{h}^{\mathrm{proj}}}} \circ \breve{h}^{\mathrm{proj}} - g_m \circ h_m\|_{P_{\mathrm{so}},2} &\le \|\breve{g} \circ \breve{h}^{\mathrm{proj}} - g_m \circ h_m\|_{P_{\mathrm{so}},2} \\
&= \|\breve{g} \circ \breve{h} - g_m \circ h_m\|_{P_{\mathrm{so}},2} \\
&= O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}). \qquad (77)
\end{aligned}$$

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

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 $\|\breve{Q}\|_{\mathrm{sp}} \le \overline{\sigma}/\underline{\sigma}$, so $(\breve{q}, \breve{Q})$ is feasible. It also implies $\|\breve{B}_{\mathrm{so}}^{+}\|_{\mathrm{sp}} \le \underline{\sigma}^{-1}$. The minimizing property of $(q_{\breve{h}^{\mathrm{proj}}}, Q_{\breve{h}^{\mathrm{proj}}})$ and Lemma 4.1, applied with $\underline{\sigma}_m = \underline{\sigma}$, give

$$\begin{aligned}
\|\breve{h}^{\mathrm{proj}} - q_{\breve{h}^{\mathrm{proj}}} - Q_{\breve{h}^{\mathrm{proj}}} h_m\|_{P_{\mathrm{so}},2} &\le \|\breve{h}^{\mathrm{proj}} - (\breve{q} + \breve{Q}h_m)\|_{P_{\mathrm{so}},2} \\
&= O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}/\underline{\sigma}) \\
&= O_{P_{\mathrm{so}}}(\delta_{\mathrm{so},m}), \qquad (78)
\end{aligned}$$

where the last equality uses that $\underline{\sigma}$ is fixed. For every $\varepsilon > 0$, (77) and (78) therefore give a constant $M_\varepsilon < \infty$ such that

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

where

$$\begin{aligned}
\mathcal{H}_{m,\varepsilon} := \big\{ h \in \mathcal{H}_m^{\mathrm{proj}} : &\|g_{\alpha^{\bullet}_{\mathrm{so},m,h}, B^{\bullet}_{\mathrm{so},m,h}} \circ h - g_m \circ h_m\|_{P_{\mathrm{so}},2} \le M_\varepsilon \delta_{\mathrm{so},m}, \\
&\|h - (q_h + Q_h h_m)\|_{P_{\mathrm{so}},2} \le M_\varepsilon \delta_{\mathrm{so},m} \big\}. \qquad (79)
\end{aligned}$$

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 E.11 are pointwise measurable. We apply Lemma E.11 with $\mathcal{H}_m = \mathcal{H}^{\mathrm{proj}}_m$, $\breve{h} = \breve{h}^{\mathrm{proj}}$, $\mathcal{F}^*_n = \mathcal{F}_n$, $f_{n,h} = f^*_{n,h} = f^\bullet_{n,h}$, $\tau_n = 1$, the estimator family $\widehat{f}_{n,h}$ defined above, $\mathcal{J}_n(f_{\alpha,\beta}) = \|\beta\|_2$, and $\widehat{\lambda}_{n,h} = \lambda_n$.[33] 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 $f^\bullet_{n,h}$. The transported-head bound above and Definition 3.1 give

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

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

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

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 4.1. Then,

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

where the last inequality follows by increasing $C_{\varepsilon,0}$ if necessary. This proves (64).

The sample ridge estimator satisfies

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

[33] The event in Assumption 4.3 (i) has probability approaching one. For the formal application of Lemma E.11, redefine $\breve{h}^{\mathrm{proj}}$ on the complement of this event as $B^+_{\mathrm{so},m} B_{\mathrm{so},m} h_m \in \mathcal{H}^{\mathrm{proj}}_m$ 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.

so (67) 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 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 4.1. Symmetrization and Lemma E.13, applied before enlarging the parameter set, yield

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}_{\mathrm{ta}} \left[\sup_{\substack{f\in\mathcal{F}_n : d_{n,h}(f, f^{\bullet}_{n,h}) < \delta \\ \mathcal{J}_n(f) < \delta/\lambda_n}} \sqrt{n} \left| (\mathbb{M}_n - M_n)(f,h) - (\mathbb{M}_n - M_n)(f^{\bullet}_{n,h}, h) \right| \right] \\
&\quad \le 4C_{\ell,\mathrm{ta}} \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}_{\xi,\mathrm{ta}} \left[\sup_{\substack{f_{\alpha,\beta}\in\mathcal{F}_n : \|\Delta\alpha + \Delta\beta' h\|_{P_{\mathrm{ta}},2} < \delta \\ \lambda_n \|\beta\|_2 < \delta}} |\mathfrak{R}_n(\Delta\alpha + \Delta\beta' h)| \right] \\
&\quad \le 4C_{\ell,\mathrm{ta}} \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}_{\xi,\mathrm{ta}} \left[\sup_{\substack{(\Delta\alpha,\Delta\beta) : \|\Delta\alpha + \Delta\beta' h\|_{P_{\mathrm{ta}},2} < \delta \\ \lambda_n \|\Delta\beta + \beta^{\bullet}_{\mathrm{ta},n,h}\|_2 < \delta}} |\mathfrak{R}_n(\Delta\alpha + \Delta\beta' h)| \right].
\end{aligned} \tag{80}$$

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, (76) 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,

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

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

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

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 3.3 and (79),

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

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 $\|h_m - \mathbb{E}_{\mathrm{ta}}[h_m]\|_{P_{\mathrm{ta}},2} = \mathrm{tr}(\Sigma_{h_m,n})^{1/2}$, the nuclear-norm bounds in Lemmas E.14 and E.15 and the Cauchy-Schwarz inequality give

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

Applying Lemmas E.18 and E.19 and absorbing the constants depending on $\overline{\sigma}, \underline{\sigma}$ now yields

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

Combining these bounds with (80) verifies (65) 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 E.11 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 (66) hold, and Lemma E.11 gives

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

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

F.11 Proof for Theorem 5.1

Proof. **Step 1: Joint asymptotic linearity.** We first establish (33). By the source bound in (39), for every $\varepsilon > 0$ there is a constant $M > 0$, depending on ε but not on n, such that

$$\Pr_P\left(\max_{1 \le j \le R} \left\|\check{g}_j \circ \check{h}_j - g_m \circ h_m\right\|_{P_{\mathrm{so},2}} \le M\delta_{\mathrm{so},m}\right) \ge 1 - \varepsilon$$

for all sufficiently large n. Since $\check{g}_j \in \mathcal{G}_m$, the minimum source approximation error for $\check{h}_j$ satisfies

$$\min_{g \in \mathcal{G}_m} \left\|g \circ \check{h}_j - g_m \circ h_m\right\|_{P_{\mathrm{so},2}} \le \left\|\check{g}_j \circ \check{h}_j - g_m \circ h_m\right\|_{P_{\mathrm{so},2}}.$$

Thus, on the event in the probability bound, every $\check{h}_j$ belongs to $\mathcal{H}_{\mathrm{so},m}(M\delta_{\mathrm{so},m})$ by the definition of the source near-identified set. The maximum in the source bound controls this membership simultaneously over all embeddings. This event depends only on the source

sample and pre-training randomness, whose law is the same for every $P \in \mathcal{P}_{0,n}$. Hence, the probability bound holds uniformly over $P \in \mathcal{P}_{0,n}$. On this event, the triangle inequality and the definition of $f^{\bullet}_{\diamond,j,n}$ give

$$\begin{aligned}\max_{1\le j\le R}\max_{\diamond\in\{\gamma,\alpha\}}\left\|\widehat{f}_{\diamond,j}\circ\check{h}_j - a^*_\diamond\right\|_{P_{\mathrm{ta}},2} \le{}& \max_{1\le j\le R}\max_{\diamond\in\{\gamma,\alpha\}}\left\|\widehat{f}_{\diamond,j}\circ\check{h}_j - f^{\bullet}_{\diamond,j,n}\circ\check{h}_j\right\|_{P_{\mathrm{ta}},2}\\ &+ \max_{\diamond\in\{\gamma,\alpha\}}\sup_{h\in\mathcal{H}_{\mathrm{so},m}(M\delta_{\mathrm{so},m})}\min_{f\in\mathcal{F}^{\mathrm{sub}}_{\diamond,n}}\|f\circ h - f_{\diamond,n}\circ h_m\|_{P_{\mathrm{ta}},2}\\ &+ \max_{\diamond\in\{\gamma,\alpha\}}\epsilon_{\diamond,n},\end{aligned}$$

where the maxima range over $j = 1,\dots,R$ and $\diamond \in \{\gamma,\alpha\}$. For each fixed M and all $n \ge N_M$, (49) bounds the middle term by $\max_{\diamond\in\{\gamma,\alpha\}}(M\delta_{\mathrm{so},m}/\nu_{\diamond,n})$, uniformly over $P_{\mathrm{ta}} \in \mathcal{P}_{0,\mathrm{ta},n}$. By (40) and (41), the right-hand side is therefore $o_P(n^{-1/4})$ uniformly over $P \in \mathcal{P}_{0,n}$. Each $\Lambda_\diamond$ has Lipschitz constant at most one and $\diamond^* = \Lambda_\diamond \circ a^*_\diamond$. Since ε can be arbitrarily small, applying these prediction maps yields the nuisance-scale bound

$$\max_{1\le j\le R}\max_{\diamond\in\{\gamma,\alpha\}}\left\|\Lambda_\diamond\circ\widehat{f}_{\diamond,j}\circ\check{h}_j - \diamond^*\right\|_{P_{\mathrm{ta}},2} = o_P(n^{-1/4}), \tag{81}$$

uniformly over $P \in \mathcal{P}_{0,n}$.

We next bound the effect of estimation error of the nuisance functions on the sample moment at θ^*. We first bound the population mean of the score error $\Delta\psi_j(W)$ and then the deviation of its sample average from the population mean. For the former, we use Assumption 5.1(iii). By (43), all fitted nuisance pairs belong to $\mathcal{N}_n$ with probability approaching one uniformly over $P \in \mathcal{P}_{0,n}$. On this event, for each j, a second-order Taylor expansion around zero, together with the population moment condition and Neyman orthogonality in (44), gives, for some $r \in (0,1)$,

$$\begin{aligned}&\mathbb{E}_{\mathrm{ta}}[\psi(W;\theta^*,\widehat{\gamma}_j,\widehat{\alpha}_j)]\\ &= \frac{1}{2}\frac{\partial^2}{\partial r^2}\mathbb{E}_{\mathrm{ta}}\left[\psi\left(W;\theta^*,\gamma^* + r(\widehat{\gamma}_j-\gamma^*),\alpha^* + r(\widehat{\alpha}_j-\alpha^*)\right)\right].\end{aligned}$$

By (45), each fitted population moment is therefore bounded in absolute value by $o(n^{-1/2})$, with the bound uniform over $P_{\mathrm{ta}} \in \mathcal{P}_{0,\mathrm{ta},n}$ and all nuisance pairs in $\mathcal{N}_n$. Hence, this bound applies simultaneously to all fitted population moments on the event in (43). Since that event has probability approaching one uniformly over $P \in \mathcal{P}_{0,n}$, and the true score has population

mean zero, we obtain

$$\max_j |\mathbb{E}_{\text{ta}}[\Delta\psi_j(W)]| = \max_j |\mathbb{E}_{\text{ta}}[\psi(W;\theta^*,\widehat{\gamma}_j,\widehat{\alpha}_j)]| = o_P(n^{-1/2}), \tag{82}$$

uniformly over $P \in \mathcal{P}_{0,n}$.

To analyze the sampling variation in the score errors, we first condition on everything used to construct the nuisance estimators, including the pre-trained models, the partition, the observations in $\mathcal{I}_3$, and the auxiliary pre-training randomness. By Assumption 5.1(i), the functions $\Delta\psi_j$ and sample sizes are then fixed, while observations in each estimation sample remain i.i.d. from P_{ta}. For each $s \in \{1,2\}$, the symmetrization inequality in Lemma 2.3.1 and the maximal inequality in Lemma 2.2.2 of van der Vaart and Wellner (2023), applied with the Orlicz norm given by $x \mapsto e^{x^2} - 1$, bound the conditional expectation of

$$\max_{1\le j\le R} \left| \frac{1}{\sqrt{n_s}} \sum_{i\in\mathcal{I}_s} \{\Delta\psi_j(W_i) - \mathbb{E}_{\text{ta}}[\Delta\psi_j(W)]\} \right| \tag{83}$$

by a universal constant times

$$\sqrt{\log R} \left\| \max_{1\le j\le R} |\Delta\psi_j(W)| \right\|_{P_{\text{ta}},2}.$$

To obtain this bound, let ξ_i, $i \in \mathcal{I}_s$, be independent Rademacher variables, independent of all other randomness. All expectations below remain conditional on the quantities used to construct the nuisance estimators. Write $\mathbb{E}_\xi$ for expectation over the Rademacher variables, $\mathbb{E}_{\text{ta}}$ for expectation over the target observations with the nuisance functions fixed, and $\mathbb{E}_{\xi,\text{ta}}$ for expectation over both. Symmetrization gives

$$\begin{aligned} &\mathbb{E}_{\text{ta}}\left[\max_{1\le j\le R} \left| \frac{1}{\sqrt{n_s}} \sum_{i\in\mathcal{I}_s} \{\Delta\psi_j(W_i) - \mathbb{E}_{\text{ta}}[\Delta\psi_j(W)]\} \right| \right] \\ &\le 2\mathbb{E}_{\xi,\text{ta}}\left[\max_{1\le j\le R} \left| \frac{1}{\sqrt{n_s}} \sum_{i\in\mathcal{I}_s} \xi_i \Delta\psi_j(W_i) \right| \right]. \end{aligned} \tag{84}$$

Denote the Orlicz norm given by $x \mapsto e^{x^2} - 1$ under the conditional law of the signs, with the estimation observations fixed, by $\|\cdot\|_{e^{x^2}-1}$. Hoeffding's inequality (van der Vaart and Wellner, 2023, Lemma 2.2.8) implies

$$\left\| \frac{1}{\sqrt{n_s}} \sum_{i\in\mathcal{I}_s} \xi_i \Delta\psi_j(W_i) \right\|_{e^{x^2}-1} \le \sqrt{6} \left(\frac{1}{n_s} \sum_{i\in\mathcal{I}_s} |\Delta\psi_j(W_i)|^2 \right)^{1/2}. \tag{85}$$

Since the Orlicz norm bounds the expected absolute value up to a universal constant, we obtain

$$\begin{aligned}
&\mathbb{E}_\xi\left[\max_{1\le j\le R}\left|\frac{1}{\sqrt{n_s}}\sum_{i\in\mathcal{I}_s}\xi_i\Delta\psi_j(W_i)\right|\right]\\
&\lesssim\left\|\max_{1\le j\le R}\left|\frac{1}{\sqrt{n_s}}\sum_{i\in\mathcal{I}_s}\xi_i\Delta\psi_j(W_i)\right|\right\|_{e^{x^2}-1}\\
&\lesssim\sqrt{\log(1+2R)}\max_{1\le j\le R}\left\|\frac{1}{\sqrt{n_s}}\sum_{i\in\mathcal{I}_s}\xi_i\Delta\psi_j(W_i)\right\|_{e^{x^2}-1}\\
&\lesssim\sqrt{\log R}\left(\frac{1}{n_s}\sum_{i\in\mathcal{I}_s}\max_{1\le j\le R}|\Delta\psi_j(W_i)|^2\right)^{1/2},
\end{aligned}\tag{86}$$

where the second inequality follows from the maximal inequality applied to the $2R$ signed sums and the last inequality follows from (85), $R\ge 2$, and bounding each squared score error by the pointwise maximum across embeddings. Taking expectations over the estimation observations and applying Jensen's inequality gives

$$\begin{aligned}
&\mathbb{E}_{\mathrm{ta}}\left[\left(\frac{1}{n_s}\sum_{i\in\mathcal{I}_s}\max_{1\le j\le R}|\Delta\psi_j(W_i)|^2\right)^{1/2}\right]\\
&\le\left(\mathbb{E}_{\mathrm{ta}}\left[\frac{1}{n_s}\sum_{i\in\mathcal{I}_s}\max_{1\le j\le R}|\Delta\psi_j(W_i)|^2\right]\right)^{1/2}\\
&=\left\|\max_{1\le j\le R}|\Delta\psi_j(W)|\right\|_{P_{\mathrm{ta}},2}.
\end{aligned}\tag{87}$$

Combining (84), (86) and (87) gives the stated conditional expectation bound:

$$\begin{aligned}
&\mathbb{E}_{\mathrm{ta}}\left[\max_{1\le j\le R}\left|\frac{1}{\sqrt{n_s}}\sum_{i\in\mathcal{I}_s}\{\Delta\psi_j(W_i)-\mathbb{E}_{\mathrm{ta}}[\Delta\psi_j(W)]\}\right|\right]\\
&\lesssim\sqrt{\log R}\left\|\max_{1\le j\le R}|\Delta\psi_j(W)|\right\|_{P_{\mathrm{ta}},2}.
\end{aligned}\tag{88}$$

By (46), the right-hand side of (88) is $o_P(1)$ uniformly over $P\in\mathcal{P}_{0,n}$.

Conditional Markov's inequality and Lemma 6.1(a) of Chernozhukov et al. (2018) therefore imply that (83) is $o_P(1)$ unconditionally. This conclusion is uniform over $P\in\mathcal{P}_{0,n}$ since the conditional expectation bound holds along every sequence of laws drawn from these classes, and the lemma applies along every such sequence. Since there are only two estimation samples

and $n_s/n \xrightarrow{P} \pi_s > 0$ uniformly, the largest centered score average is $o_P(n^{-1/2})$ uniformly over j, s and $P \in \mathcal{P}_{0,n}$. Adding the population bias in (82) gives

$$\widehat{M}_{j,s}(\theta^*) := n_s^{-1} \sum_{i \in \mathcal{I}_s} \psi(W_i; \theta^*, \widehat{\gamma}_j, \widehat{\alpha}_j) = \frac{1}{n_s} \sum_{i \in \mathcal{I}_s} \psi(W_i; \theta^*, \gamma^*, \alpha^*) + o_P(n^{-1/2}), \tag{89}$$

uniformly over j, s and $P \in \mathcal{P}_{0,n}$. Conditional on the partition, the leading term in (89) has mean zero and variance $J^2 V/n_s \leq \overline{J}^2 \overline{V}/n_s$ by Assumption 5.1(i) and (iii). Chebyshev's inequality and the sample proportion condition therefore imply that this average is $O_P(n^{-1/2})$ uniformly over s and $P \in \mathcal{P}_{0,n}$. By (47), all estimators lie in $\mathcal{U}_n(P)$ with probability approaching one uniformly over $P \in \mathcal{P}_{0,n}$. On this event, the segment between θ^* and each $\widehat{\theta}_{j,s}$ lies in $\mathcal{U}_n(P)$. The sample estimating equations in Assumption 5.1(iv), the mean value theorem, and (48) give

$$-\widehat{M}_{j,s}(\theta^*) = (J + o_P(1))(\widehat{\theta}_{j,s} - \theta^*),$$

uniformly over j, s and $P \in \mathcal{P}_{0,n}$. Evaluating the derivative bound in (48) at each $\widehat{\theta}_{j,s}$ gives $\max_{j,s} |\widehat{J}_{j,s} - J| = o_P(1)$ uniformly over $P \in \mathcal{P}_{0,n}$. The preceding moment expansion therefore yields (33), uniformly over j, s and $P \in \mathcal{P}_{0,n}$:

$$\begin{aligned} \widehat{\theta}_{j,s} - \theta^* &= -(J + o_P(1))^{-1} \widehat{M}_{j,s}(\theta^*) \\ &= \frac{1}{n_s} \sum_{i \in \mathcal{I}_s} \phi(W_i) + o_P(n^{-1/2}). \end{aligned}$$

Step 2: Limiting distribution of the statistic. Define

$$a_n := \max_{j,s} \left\{ n_s \sum_{i \in \mathcal{I}_s} (q_{ij,s} - q_{i1,s})^2 \right\}^{1/2}. \tag{90}$$

By Assumption 5.1(v), $a_n \sqrt{\log R} = o_P(1)$ and hence $a_n = o_P(1)$, both uniformly over $P \in \mathcal{P}_{0,n}$. The reverse triangle inequality gives, uniformly over $P \in \mathcal{P}_{0,n}$,

$$\max_{j,s} \left| \sqrt{n_s \sum_{i \in \mathcal{I}_s} q_{ij,s}^2} - \sqrt{n_s \sum_{i \in \mathcal{I}_s} q_{i1,s}^2} \right| \leq a_n = o_P(1). \tag{91}$$

Combining (91), consistency of the variance estimators for the first embedding function in

Assumption 5.1(v), and the upper bound $V \leq \overline{V}$ gives, uniformly over $P \in \mathcal{P}_{0,n}$,

$$\max_{j,s} \left| n_s \sum_{i \in \mathcal{I}_s} q_{ij,s}^2 - V \right| = o_P(1). \tag{92}$$

By (92), the definition of $\widehat{\sigma}_{jk}$ in (36), and the common lower bound $V \geq \underline{V} > 0$, we obtain, uniformly over $P \in \mathcal{P}_{0,n}$,

$$\max_{j<k} \left| \frac{\widehat{\sigma}_{jk}^2}{V(n_1^{-1} + n_2^{-1})} - 1 \right| = o_P(1). \tag{93}$$

The bounds $V \geq \underline{V} > 0$ and $|J| \geq \underline{J} > 0$, (93), and consistency of the derivative estimators imply that the probability that any $\widehat{\sigma}_{jk}$ or $\widehat{J}_{j,s}$ equals zero tends to zero uniformly over $P \in \mathcal{P}_{0,n}$. These are the stopping conditions in Algorithm 1, and therefore the probability that the algorithm stops tends to zero uniformly over $P \in \mathcal{P}_{0,n}$. We have shown Theorem 5.1(i).

Combining (33) and (93) with the sample proportion condition in Assumption 5.1(i) gives, uniformly over $j < k$ and $P \in \mathcal{P}_{0,n}$,

$$\frac{\widehat{\theta}_{j,1} - \widehat{\theta}_{k,2}}{\widehat{\sigma}_{jk}} = \frac{n_1^{-1} \sum_{i \in \mathcal{I}_1} \phi(W_i) - n_2^{-1} \sum_{i \in \mathcal{I}_2} \phi(W_i)}{\sqrt{V(n_1^{-1} + n_2^{-1})}} + o_P(1). \tag{94}$$

By Assumption 5.1(i), conditional on the partition, the common leading term in (94) is a sum of independent random variables with population mean zero and total variance one. To establish asymptotic normality, we verify Lyapunov's condition. On the event that $n_s/n \geq \pi_s/2$ for $s \in \{1, 2\}$, we have

$$\begin{aligned} &\sum_{s=1}^{2} \sum_{i \in \mathcal{I}_s} \mathbb{E}\left[\left| \frac{\phi(W_i)}{n_s \sqrt{V(n_1^{-1} + n_2^{-1})}} \right|^{2+\zeta} \mid \mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3 \right] \\ &= \frac{n_1^{-1-\zeta} + n_2^{-1-\zeta}}{(n_1^{-1} + n_2^{-1})^{1+\zeta/2}} \mathbb{E}_{\mathrm{ta}}\left[\left| \frac{\phi(W)}{\sqrt{V}} \right|^{2+\zeta} \right] = O(n^{-\zeta/2}), \end{aligned}$$

uniformly over $P_{\mathrm{ta}} \in \mathcal{P}_{0,\mathrm{ta},n}$ by (42). The sample proportion condition implies that this event has probability approaching one uniformly over $P \in \mathcal{P}_{0,n}$. Note that the bound for the Lyapunov condition is uniform over partitions satisfying these sample-size bounds and over sequences of target distributions. Applying Lyapunov's central limit theorem conditionally on the partition and then averaging over the partition therefore gives convergence of the leading term to a standard normal variable uniformly over $P \in \mathcal{P}_{0,n}$.

On the event that the algorithm does not stop, $\mathcal{T}$ defined in (37) differs from the absolute value of this leading term by at most $o_P(1)$ uniformly over $P \in \mathcal{P}_{0,n}$, as shown in (94). By the continuous mapping theorem and Slutsky's lemma, together with the uniformly vanishing stopping probability, $\mathcal{T}$ converges in distribution to the absolute value of a standard normal random variable: $\sup_{P\in\mathcal{P}_{0,n}} |\mathrm{Pr}_P(\mathcal{T} \le t) - (2\Phi(t) - 1)| \longrightarrow 0$ for every fixed $t \ge 0$, where Φ denotes the standard normal distribution function. To obtain uniformity in t, we take a finite grid of points starting at zero, independent of n and P, such that each interval between consecutive points and the region above the last point has small probability under the limiting distribution of $\mathcal{T}$. By monotonicity of distribution functions, the supremum error over $t \ge 0$ is bounded by the largest error at the grid points plus an arbitrarily small error. Since convergence at each grid point is uniform over P and the grid is finite, we obtain

$$\sup_{P\in\mathcal{P}_{0,n}} \sup_{t\ge 0} |\mathrm{Pr}_P(\mathcal{T} \le t) - (2\Phi(t) - 1)| \longrightarrow 0. \tag{95}$$

Step 3: Conditional bootstrap distribution. Define the stacked vectors

$$v_{jk} := \begin{pmatrix} (q_{ij,1})_{i\in\mathcal{I}_1} \\ (-q_{ik,2})_{i\in\mathcal{I}_2} \end{pmatrix}, \qquad v_0 := v_{11}.$$

For the Euclidean norm $\|\cdot\|_2$, we have $\|v_{jk}\|_2 = \widehat{\sigma}_{jk}$ for $j < k$ by the definition in (36). By (92) and (93), the squared norms of all v_{jk}, $j < k$, and v_0, divided by $V(n_1^{-1} + n_2^{-1})$, converge to one uniformly in probability. Thus, with probability approaching one uniformly, the algorithm does not stop and these norms are positive. On this event, the reverse triangle inequality and (90) give uniformly over $P \in \mathcal{P}_{0,n}$,

$$\begin{aligned} \max_{j<k} \left\| \frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2} \right\|_2 &\le 2 \max_{j<k} \frac{\|v_{jk} - v_0\|_2}{\|v_0\|_2} \\ &\le \frac{2a_n\sqrt{n_1^{-1} + n_2^{-1}}}{\|v_0\|_2} = O_P(a_n). \end{aligned} \tag{96}$$

Fix $b \in \{1, \dots, B\}$ and define

$$\xi^{(b)} := \begin{pmatrix} (\xi_i^{(b)})_{i\in\mathcal{I}_1} \\ (\xi_i^{(b)})_{i\in\mathcal{I}_2} \end{pmatrix}.$$

Its coordinates are independent standard normal variables, independent of the data, partition,

pre-training randomness, and other bootstrap draws. Define

$$Z_b^{(\xi)} := \frac{v_0^\top \xi^{(b)}}{\|v_0\|_2},$$

setting $Z_b^{(\xi)} = 0$ when $\|v_0\|_2 = 0$. On the event above, $Z_b^{(\xi)}$ is conditionally standard normal and (38) gives $\mathcal{T}_b^{(\xi)} = \max_{j<k} |v_{jk}^\top \xi^{(b)}|/\|v_{jk}\|_2$. Thus,

$$\left|\mathcal{T}_b^{(\xi)} - |Z_b^{(\xi)}|\right| \le \max_{j<k} \left| \left(\frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2} \right)^\top \xi^{(b)} \right|. \tag{97}$$

Conditional on all observed data, the partition, and pre-training randomness, the vectors v_{jk} and v_0 are fixed, while the coordinates of $\xi^{(b)}$ remain independent standard normal variables. All expectations in the following calculations are taken under this conditional law. Each variable $(v_{jk}/\|v_{jk}\|_2 - v_0/\|v_0\|_2)^\top \xi^{(b)}$ on the right-hand side of (97) is therefore mean-zero Gaussian, with variance

$$\mathbb{E}^{(\xi)} \left[\left\{ \left(\frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2} \right)^\top \xi^{(b)} \right\}^2 \right] = \left\| \frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2} \right\|_2^2.$$

Since $|x| = \max\{x, -x\}$, the maximum of the absolute values of the $R(R-1)/2$ linear forms equals the maximum over $R(R-1)$ signed forms. For a mean-zero Gaussian variable, the Orlicz norm associated with $x \mapsto e^{x^2} - 1$ is a universal constant times its standard deviation. This norm also bounds the expected absolute value up to a universal constant. Applying Lemma 2.2.2 of van der Vaart and Wellner (2023) to the signed forms therefore gives

$$\begin{aligned}
&\mathbb{E}^{(\xi)} \left[\max_{j<k} \left| \left(\frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2} \right)^\top \xi^{(b)} \right| \right] \\
&\lesssim \sqrt{\log(1 + R(R-1))} \max_{j<k} \left\| \frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2} \right\|_2 \\
&\lesssim \sqrt{\log R} \max_{j<k} \left\| \frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2} \right\|_2.
\end{aligned}$$

Conditional Markov's inequality, (97), and the preceding expectation bound now give, for

every fixed $\varepsilon > 0$,

$$\begin{aligned}
\Pr^{(\xi)}\left(\left|\mathcal{T}_b^{(\xi)} - |Z_b^{(\xi)}|\right| > \varepsilon\right) &\leq \frac{1}{\varepsilon}\mathbb{E}^{(\xi)}\left[\left|\mathcal{T}_b^{(\xi)} - |Z_b^{(\xi)}|\right|\right] \\
&\leq \frac{1}{\varepsilon}\mathbb{E}^{(\xi)}\left[\max_{j<k}\left|\left(\frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2}\right)^\top \xi^{(b)}\right|\right] \\
&\lesssim \frac{\sqrt{\log R}}{\varepsilon}\max_{j<k}\left\|\frac{v_{jk}}{\|v_{jk}\|_2} - \frac{v_0}{\|v_0\|_2}\right\|_2 \\
&= O_P\left(\frac{a_n\sqrt{\log R}}{\varepsilon}\right) = o_P(1), \qquad (98)
\end{aligned}$$

where the O_P bound follows from (96), and the final equality uses $a_n\sqrt{\log R} = o_P(1)$ from Assumption 5.1(v) and the fact that ε is fixed. Both statements hold uniformly over $P \in \mathcal{P}_{0,n}$. These calculations apply on the event above, but its complement has uniformly vanishing probability and hence does not affect convergence in probability. Thus, the conditional probability on the left-hand side of (98) converges to zero in probability unconditionally and uniformly over $P \in \mathcal{P}_{0,n}$.

Along any sequence of laws whose nth member belongs to $\mathcal{P}_{0,n}$, (98), the conditional standard normality of $Z_b^{(\xi)}$ on the event above, and the uniformly vanishing probability of its complement imply convergence in probability of the conditional expectations of all bounded Lipschitz functions of $\mathcal{T}_b^{(\xi)}$ to their expectations under the absolute standard normal law. Since the limiting distribution function is continuous, Lemma 10.11(i) of Kosorok (2008) gives

$$\sup_{t\geq 0}\left|\Pr^{(\xi)}(\mathcal{T}_b^{(\xi)} \leq t) - (2\Phi(t) - 1)\right| = o_P(1). \qquad (99)$$

This convergence is uniform over $P \in \mathcal{P}_{0,n}$ because the sequence of laws is arbitrary. Combining (95) and (99) by the triangle inequality yields

$$\sup_{t\in\mathbb{R}}\left|\Pr^{(\xi)}(\mathcal{T}_b^{(\xi)} \leq t) - \Pr_P(\mathcal{T} \leq t)\right| = o_P(1),$$

uniformly over $P \in \mathcal{P}_{0,n}$. We have shown Theorem 5.1(ii).

Step 4: Empirical critical value and rejection probability. Finally, we show Theorem 5.1(iii). Define the empirical bootstrap distribution function by

$$\widehat{F}_B^{(\xi)}(t) := \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\{\mathcal{T}_b^{(\xi)} \leq t\}. \qquad (100)$$

Conditional on the data, partition, and pre-training randomness, the bootstrap statistics are

i.i.d. by Algorithm 1. The Dvoretzky-Kiefer-Wolfowitz inequality (Theorem 11.6 of Kosorok (2008)) therefore gives, for every $\varepsilon > 0$,

$$\Pr^{(\xi)}\left(\sup_{t\in\mathbb{R}}\left|\widehat{F}_B^{(\xi)}(t) - \Pr^{(\xi)}(\mathcal{T}_b^{(\xi)} \le t)\right| > \varepsilon\right) \le 2\exp(-2B\varepsilon^2). \tag{101}$$

By (99) and (101) and $B \to \infty$,

$$\sup_{t\ge 0}\left|\widehat{F}_B^{(\xi)}(t) - (2\Phi(t) - 1)\right| = o_P(1),$$

uniformly over $P \in \mathcal{P}_{0,n}$, with no restriction on the relative growth of n and B. Since the distribution function $2\Phi(t) - 1$ is strictly increasing, applying Lemma 21.2 of Van der Vaart (1998) along every sequence of laws whose nth member belongs to $\mathcal{P}_{0,n}$ therefore gives, uniformly over $P \in \mathcal{P}_{0,n}$,

$$c_{1-\mathsf{a}}^{(\xi)} \xrightarrow{P} \Phi^{-1}(1 - \mathsf{a}/2). \tag{102}$$

Since the distribution function $2\Phi(t) - 1$ is continuous, combining (95) and (102) by Slutsky's lemma, as in the proof of Lemma 23.3 of Van der Vaart (1998), yields the uniform size conclusion by the same sequence argument:

$$\sup_{P\in\mathcal{P}_{0,n}}\left|\Pr_P\left(\mathcal{T} > c_{1-\mathsf{a}}^{(\xi)}\right) - \mathsf{a}\right| \longrightarrow 0.$$

This completes the proof of Theorem 5.1. □

F.12 Proof for Theorem 5.2

Proof. On the event in (51), every $\widehat{J}_{j,s}$ is nonzero, and (35) gives

$$\sum_{i\in\mathcal{I}_s} q_{ij,s}^2 = \frac{\sum_{i\in\mathcal{I}_s}\psi(W_i;\widehat{\theta}_{j,s},\widehat{\gamma}_j,\widehat{\alpha}_j)^2}{n_s(n_s-1)\widehat{J}_{j,s}^2} \ge \frac{c_\psi}{(n_s-1)\widehat{J}_{j,s}^2} > 0.$$

Thus, every $\widehat{\sigma}_{jk}$ is positive by (36). Assumption 5.2(i) therefore implies that the probability that the algorithm stops without a test decision tends to zero uniformly over $P \in \mathcal{P}_{\mathrm{sep},n}$.

Conditional on the data, partition, and pre-training randomness, and on the event that the algorithm does not stop, each standardized difference

$$\frac{\sum_{i\in\mathcal{I}_1}\xi_i^{(b)} q_{ij,1} - \sum_{i\in\mathcal{I}_2}\xi_i^{(b)} q_{ik,2}}{\widehat{\sigma}_{jk}}$$

in (38) has a standard normal distribution for every $1 \le j < k \le R$ and $b = 1, \dots, B$. The Gaussian tail bound (Wainwright, 2019, Example 2.1) and the union bound over $R(R-1)/2$ terms therefore give, for every $t \ge 0$,

$$\Pr^{(\xi)}(\mathcal{T}_b^{(\xi)} > t) \le R(R-1)\exp(-t^2/2). \tag{103}$$

If the empirical $(1-\mathsf{a})$-quantile exceeds t, then more than a fraction a of the bootstrap draws exceed t. Thus, conditional Markov's inequality and the bound in (103) give

$$\begin{aligned}\Pr^{(\xi)}(c_{1-\mathsf{a}}^{(\xi)} > t) &\le \frac{1}{\mathsf{a}} \Pr^{(\xi)}(\mathcal{T}_b^{(\xi)} > t) \\ &\le \frac{R(R-1)}{\mathsf{a}} \exp(-t^2/2).\end{aligned} \tag{104}$$

This bound also holds on the stopping event by the zero convention. Taking $t = K\sqrt{\log R}$ for $K > 2$, the right-hand side is bounded by $\mathsf{a}^{-1}2^{2-K^2/2}$, since $R \ge 2$. Taking unconditional probabilities and then letting $K \to \infty$, we have $c_{1-\mathsf{a}}^{(\xi)} = O_P(\sqrt{\log R})$ uniformly over $P \in \mathcal{P}_{\mathrm{sep},n}$ and over all $B \ge 1$.

The triangle inequality gives

$$\max_{j<k} |\widehat{\theta}_{j,1} - \widehat{\theta}_{k,2}| \ge \max_{j<k} |\theta_{j,n}^{\dagger} - \theta_{k,n}^{\dagger}| - 2 \max_{j,s} |\widehat{\theta}_{j,s} - \theta_{j,n}^{\dagger}|.$$

By (52) and (53) and $r_{\theta,n} = o(\Delta_n)$, the right-hand side is at least $\Delta_n/2$ with probability approaching one uniformly over $P \in \mathcal{P}_{\mathrm{sep},n}$. On this event, if the algorithm does not stop,

$$\frac{\mathcal{T}}{\sqrt{\log R}} \ge \frac{\Delta_n}{2\sqrt{\log R}\max_{j<k}\widehat{\sigma}_{jk}} = \frac{\sqrt{n}\Delta_n}{\sqrt{\log R}} \frac{1}{2\sqrt{n}\max_{j<k}\widehat{\sigma}_{jk}}.$$

By Assumption 5.2(ii) and (iii), the first factor on the right-hand side diverges, and the denominator of the second is bounded in probability uniformly over $P \in \mathcal{P}_{\mathrm{sep},n}$. Since the algorithm does not stop with probability approaching one uniformly, the inequality above shows that $\mathcal{T}/\sqrt{\log R}$ diverges in probability uniformly over $P \in \mathcal{P}_{\mathrm{sep},n}$. Consequently, for every fixed $K > 0$, $\sup_{P \in \mathcal{P}_{\mathrm{sep},n}} \Pr_P(\mathcal{T} \le K\sqrt{\log R}) \to 0$. Finally,

$$\Pr_P(\mathcal{T} \le c_{1-\mathsf{a}}^{(\xi)}) \le \Pr_P(\mathcal{T} \le K\sqrt{\log R}) + \Pr_P(c_{1-\mathsf{a}}^{(\xi)} > K\sqrt{\log R}).$$

Taking the supremum over $P \in \mathcal{P}_{\mathrm{sep},n}$, letting $n \to \infty$, and then letting $K \to \infty$ completes the proof. □

G Proofs for the Appendix

G.1 Proof for Lemma E.1

Proof. The assumed square integrability of the composed predictors implies that their conditional expectations and the products in the following expansion are integrable:

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

where the last equality follows because the conditional mean of $f_n \circ h_m(Z) - \mathbb{E}[f_n \circ h_m(Z) \mid h(Z)]$ given $h(Z)$ is zero, so the cross term vanishes by the law of iterated expectation. □

G.2 Proof for Lemma E.2

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 $j = 1, \ldots, q$, 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, $\Gamma = (\Gamma_1, \ldots, \Gamma_q)' : \mathcal{U} \to \mathbb{R}^q$ is measurable and satisfies $\widetilde{V} = \Gamma \circ U$. Since $V = \widetilde{V}$, P-a.s., the result follows. □

G.3 Proof for Lemma E.3

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

G.4 Proof for Lemma E.4

Proof. The assumed Euclidean-norm bounds imply that

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

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

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

Write $p(u) = (p_1(u), \ldots, p_T(u))'$. For $y = 1, \ldots, T$, 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 $y = 1, \ldots, T$,

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

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 E.5

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 E.6

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 E.7

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,

$$\begin{aligned}\left\|\left(m_{f,h}-m_{f^*_{n,h},h}\right)-\left(m_{\widetilde{f},h}-m_{f^*_{n,h},h}\right)\right\|_{Q,2} &= \left\|m_{f,h}-m_{\widetilde{f},h}\right\|_{Q,2}\\ &\le C_\ell\left\|f\circ h-\widetilde{f}\circ h\right\|_{Q,2}\\ &= C_\ell d_{Q,h}(f,\widetilde{f}). \end{aligned} \tag{105}$$

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

$$\begin{aligned}N\left(\varepsilon/(2C_\ell M),\widetilde{\mathcal{M}}_{n,h}(\delta),L_2(Q)\right) &= N\left(\varepsilon,\mathcal{M}_{n,h}(\delta),L_2(Q)\right)\\ &\le N\left(\varepsilon/C_\ell,B_{n,h}(\delta),d_{Q,h}\right)\\ &\le N\left(\varepsilon/C_\ell,\mathcal{F}_n,d_{Q,h}\right), \end{aligned} \tag{106}$$

where the first inequality follows from (105) 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 $f\in B_{n,h}(\delta)$,

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

Hence, $\widetilde{\mathcal{M}}_{n,h}(\delta)$ has an 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

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

Because $B_{n,h}(\delta)$ is pointwise measurable by Lemma E.8, 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

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

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

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

where the first equality follows from the definition of the uniform entropy integral, the first inequality follows from (106), 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

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

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

G.8 Proof for Lemma E.8

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 E.9

Proof. For $\mathsf{V}(\mathcal{F}) \ge 1$, apply Theorem 2.6.7 in van der Vaart and Wellner (2023). If $\mathsf{V}(\mathcal{F}) = 0$, all functions in the nonempty class agree pointwise: two different values at any point would allow their subgraphs to shatter a singleton. Hence the covering number is one, and the displayed bound also holds. □

G.10 Proof for Lemma E.10

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 (61), 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 (63), 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

$$M_n(f,h) - M_n\left(f_{n,h}, h\right) \lesssim_\varepsilon -2^{2j}\delta_n^2 \quad \text{for every } h \in \mathcal{H}_{m,\varepsilon} \text{ and } f \in S_{n,h,j}. \tag{107}$$

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 (107),

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

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

$$\begin{aligned}
&\sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{f \in S_{n,h,j}} \left|\mathbb{G}_n(f,h) - \mathbb{G}_n\left(f_{n,h},h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2) \\
&\lesssim_\varepsilon \sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{f \in \mathcal{F}_n : d_{n,h}\left(f, f_{n,h}^*\right) \leq 2^j\tau_n^{-1}\delta_n} \left|\mathbb{G}_n(f,h) - \mathbb{G}_n\left(f_{n,h}^*,h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2) \qquad (108) \\
&\quad + \sup_{h \in \mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\left|\mathbb{G}_n\left(f_{n,h}^*,h\right) - \mathbb{G}_n\left(f_{n,h},h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2), \qquad (109)
\end{aligned}$$

where the inequality follows from the triangle inequality and the definition of $S_{n,h,j}$. By the modulus bound in (62), the first term (108) 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 (109), 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 (61) 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 (63), 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 (62) to bound the second term (109) 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.[34] 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 (63), 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 E.11

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

[34] 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 (62). 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

$$\begin{aligned}
&\limsup_{n\to\infty} P\Bigg(\tau_n^2 d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right)^2 + \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) > 2^{2J}\delta_n^2\,, \\
&\qquad \tau_n^2 d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right)^2 + \widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(\widehat{f}_{n,\check{h}}\right) > 2^{2J}\widehat{\lambda}_{n,\check{h}}^2 \mathcal{J}_n^2\left(f_{n,\check{h}}\right), \\
&\qquad\qquad d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right) \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n, \widehat{\lambda}_{n,\check{h}} \geq \lambda_n\Bigg) \\
&\leq \limsup_{n\to\infty} P\left((\widehat{f}_{n,\check{h}}, \widehat{\lambda}_{n,\check{h}}) \in \bigcup_{j>J} S_{n,\check{h},j}, d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right) \geq \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n, \widehat{\lambda}_{n,\check{h}} \geq \lambda_n\right) \\
&\leq \varepsilon. \qquad (110)
\end{aligned}$$

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 $\mathcal{H}_{m,\varepsilon}$.[35]

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 (64) with $\delta = d_{n,h}\left(f, f_{n,h}^*\right)$ to get

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

[35] Let $R_n \geq 0$ denote the optimization-error remainder, with $R_n = O_P(\delta_n^2)$, as in the proof of Lemma E.10. 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 $\{R_n \leq 2^J\delta_n^2, \check{h} \in \mathcal{H}_{m,\varepsilon}\}$.

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

$$\begin{aligned}
&M_n(f,h) - M_n\left(f_{n,h},h\right) - \lambda^2\mathcal{J}_n^2(f) + \lambda^2\mathcal{J}_n^2\left(f_{n,h}\right) \\
&\quad = \left(M_n(f,h) - M_n\left(f_{n,h}^*,h\right)\right) + \left(M_n\left(f_{n,h}^*,h\right) - M_n\left(f_{n,h},h\right)\right) - \lambda^2\mathcal{J}_n^2(f) + \lambda^2\mathcal{J}_n^2\left(f_{n,h}\right) \\
&\quad \le -\tau_n^2 C_{\varepsilon,1} d_{n,h}\left(f, f_{n,h}^*\right)^2 + C_{\varepsilon,2}\delta_n^2 - \lambda^2\mathcal{J}_n^2(f) + \lambda^2\mathcal{J}_n^2\left(f_{n,h}\right) \\
&\quad \le -(\tau_n^2 d_{n,h}\left(f, f_{n,h}^*\right)^2 + \lambda^2\mathcal{J}_n^2(f))(\min\{C_{\varepsilon,1},1\} - 2^{-2J}) + C_{\varepsilon,2}\delta_n^2 \\
&\quad \le -(2^{2j-2}(\min\{C_{\varepsilon,1},1\} - 2^{-2J}) - C_{\varepsilon,2})\delta_n^2 \\
&\quad \lesssim_\varepsilon -2^{2j}\delta_n^2, \qquad (111)
\end{aligned}$$

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 (66), 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 $2^{2j-2}\delta_n^2 \le \tau_n^2 d_{n,h}\left(f, f_{n,h}^*\right)^2 + \lambda^2\mathcal{J}_n^2(f)$ 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 $d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right) \ge \tau_n^{-1} C_{\varepsilon,0}\underline{\delta}_n$ and $\widehat{\lambda}_{n,\check{h}} \ge \lambda_n$. Then,

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

where the first inequality follows from the approximate-maximization property of $(\widehat{f}_{n,h}, \widehat{\lambda}_{n,h})$ in (67) and the second inequality follows from (111). Here, for large enough J,

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

Also, $(\widehat{f}_{n,\check{h}}, \widehat{\lambda}_{n,\check{h}}) \in S_{n,\check{h},j}$ with $\widehat{\lambda}_{n,\check{h}} \ge \lambda_n$ implies that $d_{n,\check{h}}\left(\widehat{f}_{n,\check{h}}, f_{n,\check{h}}^*\right) \le 2^j\tau_n^{-1}\delta_n$ and

$\mathcal{J}_n\left(\widehat{f}_{n,\check{h}}\right) < 2^j\delta_n/\lambda_n$. Thus, by the set inclusion, we have

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

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

$$
\begin{aligned}
&\sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}_n: d_{n,h}\left(f,f^*_{n,h}\right)\leq 2^j\tau_n^{-1}\delta_n, \\ \mathcal{J}_n(f)<2^j\delta_n/\lambda_n}} \left|\mathbb{G}_n\left(f,h\right) - \mathbb{G}_n\left(f_{n,h},h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2) \\
&\leq \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}_n: d_{n,h}\left(f,f^*_{n,h}\right)\leq 2^j\tau_n^{-1}\delta_n, \\ \mathcal{J}_n(f)<2^j\delta_n/\lambda_n}} \left|\mathbb{G}_n\left(f,h\right) - \mathbb{G}_n\left(f^*_{n,h},h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2) \qquad (112) \\
&\quad + \sup_{h\in\mathcal{H}_{m,\varepsilon}} \mathbb{E}\left[\left|\mathbb{G}_n\left(f^*_{n,h},h\right) - \mathbb{G}_n\left(f_{n,h},h\right)\right|\right] / (\sqrt{n}2^{2j}\delta_n^2). \qquad (113)
\end{aligned}
$$

By the modulus bound in (65), the first term (112) 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 (113), 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 (66). Apply (65) to bound the second term (113) 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 E.10, using the properties of

ϕ_n in (66) and taking J large enough. We obtain (110), 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 $\widehat{\lambda}^2_{n,\check{h}}\mathcal{J}^2_n\left(\widehat{f}_{n,\check{h}}\right) \geq 0$. 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 E.12

Proof. Pick an arbitrary $\varepsilon > 0$. By the assumptions, there exist constants $C_{\varepsilon,1} > 0$, $C_{\varepsilon,2} > 0$, and $C_{\varepsilon,3} > 0$ such that

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} d_{n,h}\left(f_{n,h}, f^*_{n,h}\right) \leq C_{\varepsilon,1}\underline{\delta}_n, \tag{114}$$

$$\sup_{h\in\mathcal{H}_{m,\varepsilon}} \sup_{f\in\mathcal{F}_n : \delta/2 < d_{n,h}\left(f, f^*_{n,h}\right) \leq \delta} M_n(f,h) - M_n\left(f^*_{n,h}, h\right) \leq -C_{\varepsilon,2}\tau_n^2\delta^2, \tag{115}$$

and

$$M_n\left(f^*_{n,h}, h\right) - M_n\left(f_{n,h}, h\right) \leq C_{\varepsilon,3} d_{n,h}\left(f_{n,h}, f^*_{n,h}\right)^2 \tag{116}$$

for every $h \in \mathcal{H}_{m,\varepsilon}$. Set

$$C_{\varepsilon,0} := C_{\varepsilon,1}\max\left\{4, \sqrt{32C_{\varepsilon,3}/C_{\varepsilon,2}}\right\},$$

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 $\delta/2 < d_{n,h}(f, f_{n,h}) \leq \delta$. By the choice of $C_{\varepsilon,0}$, (114), and $\tau_n \leq 1$, we have

$$d_{n,h}(f, f_{n,h}) > \delta/2 \geq 2d_{n,h}\left(f_{n,h}, f^*_{n,h}\right), \tag{117}$$

$$d_{n,h}(f, f_{n,h}) > \delta/2 \geq \sqrt{8C_{\varepsilon,3}/C_{\varepsilon,2}}\,\tau_n^{-1} d_{n,h}\left(f_{n,h}, f^*_{n,h}\right). \tag{118}$$

The reverse triangle inequality and (117) imply

$$d_{n,h}\left(f, f_{n,h}^*\right) \geq d_{n,h}\left(f, f_{n,h}\right) - d_{n,h}\left(f_{n,h}, f_{n,h}^*\right) \geq \frac{1}{2} d_{n,h}\left(f, f_{n,h}\right) > 0. \tag{119}$$

Adding and subtracting $M_n\left(f_{n,h}^*, h\right)$, applying (115) with shell radius $d_{n,h}\left(f, f_{n,h}^*\right)$, and using (116), 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 (119), the third inequality follows from (118), 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 E.13

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

$$\mathbb{E}\left[\sup_{u \in \mathcal{U}} \left|\sum_{i=1}^n \xi_i \phi_{0,i}(u_i)\right|\right] \leq 2 \cdot \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left|\sum_{i=1}^n \xi_i u_i\right|\right].$$

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,

$$\begin{aligned} \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left|\sum_{i=1}^n \xi_i \phi_i(u_i)\right|\right] &= L_\phi \cdot \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left|\sum_{i=1}^n \xi_i \phi_{0,i}(u_i)\right|\right] \\ &\leq 2L_\phi \cdot \mathbb{E}\left[\sup_{u \in \mathcal{U}} \left|\sum_{i=1}^n \xi_i u_i\right|\right]. \end{aligned}$$

This completes the proof. □

G.14 Proof for Lemma E.14

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 E.15

Proof. By the singular value decomposition, we can write $A = \sigma uv'$ for some $\sigma \geq 0$ 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 E.16

Proof. By the definition of the $L_2(P)$ norm,

$$\|Aa\|_{P,2}^2 = \int \|Aa(v)\|_2^2 \, dP(v) \leq \int \|A\|_{\text{sp}}^2 \|a(v)\|_2^2 \, dP(v) = \|A\|_{\text{sp}}^2 \|a\|_{P,2}^2,$$

where the inequality follows from $\|Aa(v)\|_2 = \|Aa(v)\|_F$ and Lemma E.14. This gives the desired inequality after taking square roots of both sides. □

G.17 Proof for Lemma E.17

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 E.18

Proof. By the definition of $\mathcal{N}_A(\lambda)$, we have

$$\begin{aligned}
&\mathcal{N}_{\widetilde{A}}(\lambda) - \mathcal{N}_A(\lambda) \\
&\quad = \lambda^2 \,\text{tr}\left((A + \lambda^2 I)^{-1} - (\widetilde{A} + \lambda^2 I)^{-1}\right) \\
&\quad = \lambda^2 \,\text{tr}\left((\widetilde{A} + \lambda^2 I)^{-1}(\widetilde{A} - A)(A + \lambda^2 I)^{-1}\right),
\end{aligned}$$

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:

$$\begin{aligned}
&\lambda^2 \left|\operatorname{tr}\left((\widetilde{A} + \lambda^2 I)^{-1}(\widetilde{A} - A)(A + \lambda^2 I)^{-1}\right)\right| \\
&\quad \le \lambda^2 \|(\widetilde{A} + \lambda^2 I)^{-1}(\widetilde{A} - A)(A + \lambda^2 I)^{-1}\|_* \\
&\quad \le \lambda^2 \|(\widetilde{A} + \lambda^2 I)^{-1}\|_{\mathrm{sp}} \|\widetilde{A} - A\|_* \|(A + \lambda^2 I)^{-1}\|_{\mathrm{sp}} \\
&\quad \le \|A - \widetilde{A}\|_*/\lambda^2,
\end{aligned}$$

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 E.14 twice, and the last inequality follows from $\|(A + \lambda^2 I)^{-1}\|_{\mathrm{sp}} \le 1/\lambda^2$ for any positive semi-definite matrix A. □

G.19 Proof for Lemma E.19

Proof. Define $Q^* = Q/\sqrt{\max\{1, \|Q\|_{\mathrm{sp}}^2\}}$. Then $\|Q^*\|_{\mathrm{sp}} \le 1$ and

$$QAQ' = \max\{1, \|Q\|_{\mathrm{sp}}^2\} Q^* A Q^{*'}.$$

Since $\|Q^*\|_{\mathrm{sp}} \le 1$, we have $Q^{*\prime} Q^* \preceq I$. Therefore, for every $x \in \mathbb{R}^K$,

$$x' A^{1/2} Q^{*\prime} Q^* A^{1/2} x \le x' A x.$$

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

$$\mathcal{N}_{Q^*AQ^{*\prime}}(\lambda) = \operatorname{tr}\left(I - \lambda^2(A^{1/2}Q^{*\prime}Q^*A^{1/2} + \lambda^2 I)^{-1}\right) \leq \operatorname{tr}\left(I - \lambda^2(A + \lambda^2 I)^{-1}\right) = \mathcal{N}_A(\lambda).$$

Finally, let $\mu_1, \ldots, \mu_K$ denote the eigenvalues of $Q^*AQ^{*\prime}$. Using the definition of Q^*, we obtain

$$\begin{aligned}
\mathcal{N}_{QAQ'}(\lambda) &= \mathcal{N}_{\max\{1,\|Q\|_{\mathrm{sp}}^2\}Q^*AQ^{*\prime}}(\lambda) \\
&= \sum_{j=1}^{K} \frac{\max\{1, \|Q\|_{\mathrm{sp}}^2\}\mu_j}{\max\{1, \|Q\|_{\mathrm{sp}}^2\}\mu_j + \lambda^2} \\
&\leq \max\{1, \|Q\|_{\mathrm{sp}}^2\} \sum_{j=1}^{K} \frac{\mu_j}{\mu_j + \lambda^2} \\
&= \max\{1, \|Q\|_{\mathrm{sp}}^2\}\mathcal{N}_{Q^*AQ^{*\prime}}(\lambda) \\
&\leq \max\{1, \|Q\|_{\mathrm{sp}}^2\}\mathcal{N}_A(\lambda).
\end{aligned}$$

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 $r = 1, \ldots, R$ 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

$$(Y_i, Z_{0,i}, Z_{1,i}, \ldots, Z_{R,i}) \sim_{i.i.d.} P_{\mathrm{ta}}, \qquad i = 1, \ldots, n,$$

where $Z_{0,i}$ denotes the structured covariates and $Z_{r,i} \in \mathcal{Z}_r$ denotes the unstructured input for modality r. Write $Z_i = (Z_{0,i}, Z_{1,i}, \ldots, Z_{R,i})$. The structured covariates may be low- or high-dimensional. For each modality r, the computer scientist observes a source sample

$$(S_{r,j}, Z_{r,j}) \sim_{i.i.d.} P_{\mathrm{so},r}, \qquad j = 1, \ldots, m_r,$$

where the modality-specific source samples and the target sample are mutually independent. Let P denote the joint distribution of these samples, and let $\mathbb{E}_{\mathrm{so},r}$ denote expectation under $P_{\mathrm{so},r}$.

Let $\ell_{\mathrm{so},r}$ 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

$$\mathcal{R}_{r,m_r} = \operatorname*{arg\,min}_{g_r \in \mathcal{G}_{r,m_r}, h_r \in \mathcal{H}_{r,m_r}} \mathbb{E}_{\mathrm{so},r}[\ell_{\mathrm{so},r}(g_r \circ h_r(Z_r), S_r)]. \tag{120}$$

For each modality r, let

$$(\breve{g}_r, \breve{h}_r) \in \operatorname*{arg\,min}_{g_r \in \mathcal{G}_{r,m_r}, h_r \in \mathcal{H}_{r,m_r}} \frac{1}{m_r} \sum_{j=1}^{m_r} \ell_{\mathrm{so},r}(g_r \circ h_r(Z_{r,j}), S_{r,j}). \tag{121}$$

Thus, $\breve{h}_r$ is the pre-trained embedding estimated from the modality-r source sample.

Define $\mathbf{m} = (m_1, \ldots, m_R)$ 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

$$\mathcal{F}_n^{\mathrm{sub}}(h) = \operatorname*{arg\,min}_{f \in \mathcal{F}_n^{\mathrm{sub}}} \mathbb{E}_{\mathrm{ta}}[\ell_{\mathrm{ta}}((f \circ h)(Z), Y)]. \tag{122}$$

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 H.1 (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

$$f_{n,h} \in \mathcal{F}_n^{\mathrm{sub}}(h), \qquad h \in \mathcal{H}_{\mathbf{m}}, \tag{123}$$

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

$$\mathcal{H}_{\mathrm{ta},r,n}(\rho \mid h_{-r}) := \{h_r \in \mathcal{H}_{r,m_r} : D_{r,n}(h_r \mid h_{-r}) \le \rho\}.$$

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 3.2. 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 H.1 (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

$$f^{\bullet}_{n,\breve{h}} \in \operatorname*{arg\,min}_{f \in \mathcal{F}^{\mathrm{sub}}_n} \|f \circ \breve{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2}.$$

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 H.1; 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 $r = 0, \ldots, R$, define the intermediate multimodal embeddings

$$\widetilde{h}^{(r)} := (\breve{h}_1, \ldots, \breve{h}_r, h_{r+1,m_{r+1}}, \ldots, h_{R,m_R}).$$

Thus, $\widetilde{h}^{(0)} = h_{\mathbf{m}}$ and $\widetilde{h}^{(R)} = \breve{h}$. The triangle inequality yields

$$\begin{aligned}\|\widehat{f} \circ \breve{h} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} \le{}& \|\widehat{f} \circ \breve{h} - f^{\bullet}_{n,\breve{h}} \circ \breve{h}\|_{P_{\mathrm{ta}},2} \\ &+ \|f^{\bullet}_{n,\breve{h}} \circ \breve{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2} \\ &+ \|f_n \circ h_{\mathbf{m}} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2}.\end{aligned}$$

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 $f^{\bullet}_{n,\breve{h}}$, the fact that $f_{n,\breve{h}} \in \mathcal{F}^{\mathrm{sub}}_n(\breve{h}) \subseteq \mathcal{F}^{\mathrm{sub}}_n$, and the triangle inequality,

$$\begin{aligned}\|f^{\bullet}_{n,\breve{h}} \circ \breve{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2} &\le \|f_{n,\breve{h}} \circ \breve{h} - f_n \circ h_{\mathbf{m}}\|_{P_{\mathrm{ta}},2} \\ &\le \sum_{r=1}^{R} \|f_{n,\widetilde{h}^{(r)}} \circ \widetilde{h}^{(r)} - f_{n,\widetilde{h}^{(r-1)}} \circ \widetilde{h}^{(r-1)}\|_{P_{\mathrm{ta}},2},\end{aligned}$$

where the second inequality uses $f_{n,\widetilde{h}^{(0)}} = f_n$ and $f_{n,\widetilde{h}^{(R)}} = f_{n,\breve{h}}$.

For each modality r, $\breve{g}_r \in \mathcal{G}_{r,m_r}$ and the source-rate condition imply

$$\begin{aligned}\min_{g_r \in \mathcal{G}_{r,m_r}} \|g_r \circ \breve{h}_r - g_{r,m_r} \circ h_{r,m_r}\|_{P_{\mathrm{so},r},2} &\le \|\breve{g}_r \circ \breve{h}_r - g_{r,m_r} \circ h_{r,m_r}\|_{P_{\mathrm{so},r},2} \\ &= O_{P_{\mathrm{so},r}}(\delta_{\mathrm{so},r,m_r}).\end{aligned}$$

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,

$$P\left(\breve{h}_r \notin \mathcal{H}_{\mathrm{so},r,m_r}(M_{\varepsilon,r}\delta_{\mathrm{so},r,m_r})\right) \le \frac{\varepsilon}{R}.$$

Let $M_\varepsilon = \max_{1 \le r \le R} M_{\varepsilon,r}$. Since R is fixed, the union bound implies that the event

$$\mathcal{E}_n := \bigcap_{r=1}^{R} \left\{\breve{h}_r \in \mathcal{H}_{\mathrm{so},r,m_r}(M_\varepsilon \delta_{\mathrm{so},r,m_r})\right\}$$

satisfies

$$P(\mathcal{E}_n) \ge 1 - \sum_{r=1}^{R} \frac{\varepsilon}{R} = 1 - \varepsilon$$

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 $(\breve{h}_1, \ldots, \breve{h}_{r-1}, h_{r+1,m_{r+1}}, \ldots, h_{R,m_R})$ and yields

$$\begin{aligned}D_{r,n}\left(\breve{h}_r \mid \left(\breve{h}_1, \ldots, \breve{h}_{r-1}, h_{r+1,m_{r+1}}, \ldots, h_{R,m_R}\right)\right) &= \|f_{n,\widetilde{h}^{(r)}} \circ \widetilde{h}^{(r)} - f_{n,\widetilde{h}^{(r-1)}} \circ \widetilde{h}^{(r-1)}\|_{P_{\mathrm{ta}},2} \\ &\le \frac{M_\varepsilon \delta_{\mathrm{so},r,m_r}}{\nu_{n,r}}.\end{aligned}$$

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 $f^{\bullet}_{n,\breve{h}}$ 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

$$\|\widehat{f} \circ \breve{h} - a^*_{\mathrm{ta}}\|_{P_{\mathrm{ta}},2} = O_P\left(r_{\mathrm{ta},n} + \frac{\delta_{\mathrm{so,image},m_{\mathrm{image}}}}{\nu_{n,\mathrm{image}}} + \frac{\delta_{\mathrm{so,text},m_{\mathrm{text}}}}{\nu_{n,\mathrm{text}}} + \epsilon_{\mathrm{ta},n}\right).$$

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 Additional Details and Results for Simulations

This section reports the Monte Carlo design reported in the introduction for the partially linear model in Example 2.1. 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

$$\begin{aligned} Y_i &= D_i\theta_0 + h^*(Z_i^{\mathrm{ta}})'\beta_0 + U_i, \\ D_i &= h^*(Z_i^{\mathrm{ta}})'\delta_0 + V_i, \end{aligned}$$

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

$$\begin{aligned} \alpha_0(z) &= \mathbb{E}[D_i \mid Z_i^{\mathrm{ta}} = z] = h^*(z)'\delta_0, \\ \gamma_0(z) &= \mathbb{E}[Y_i - \theta_0 D_i \mid Z_i^{\mathrm{ta}} = z] = h^*(z)'\beta_0. \end{aligned}$$

The common representation map $h^* : \mathbb{R}^{50} \to \mathbb{R}^{10}$ is a fixed random ReLU network:

$$\begin{aligned} x^{(0)}(z) &= z, \\ x^{(\ell)}(z) &= \mathrm{ReLU}(x^{(\ell-1)}(z)W_\ell + b_\ell), \quad \ell = 1, \ldots, 5, \\ h^*(z) &= x^{(5)}(z), \end{aligned}$$

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

$$\begin{aligned} Z_j^{\mathrm{so}} &\sim N(0, I_{50}), \\ S_j &= h^*(Z_j^{\mathrm{so}})' B + \varepsilon_j^{\mathrm{so}}, \\ \varepsilon_j^{\mathrm{so}} &\sim N(0, \sigma_{\mathrm{so}}^2 I_T), \end{aligned}$$

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

$$\beta_{\mathrm{in}} = \frac{U a_\beta}{\|U a_\beta\|_2}, \quad \delta_{\mathrm{in}} = \frac{U a_\delta}{\|U a_\delta\|_2},$$

whereas the out-of-span design sets

$$\beta_{\mathrm{out}} = 2q, \quad \delta_{\mathrm{out}} = 2q.$$

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, $\sigma_{\mathrm{so}} = \sigma_U = 1$, $\sigma_V = 0.5$, five-fold cross-fitting, 20 source-training epochs, batch size 128, and learning rate 10^{-3}.

Given the learned embedding $\breve{h}$, each cross-fitting split estimates linear projections of Y_i, D_i, and $Y_i - \theta_0 D_i$ on $\breve{h}(Z_i)$:

$$\begin{aligned} \widehat{\ell}^{(-k)}(z) &= \widehat{a}_\ell^{(-k)} + \widehat{b}_\ell^{(-k)\prime} \breve{h}(z), \\ \widehat{\alpha}^{(-k)}(z) &= \widehat{a}_\alpha^{(-k)} + \widehat{b}_\alpha^{(-k)\prime} \breve{h}(z), \\ \widehat{\gamma}^{(-k)}(z) &= \widehat{a}_\gamma^{(-k)} + \widehat{b}_\gamma^{(-k)\prime} \breve{h}(z). \end{aligned}$$

The partially linear DML estimator is then

$$\widetilde{Y}_i = Y_i - \widehat{\ell}^{(-k(i))}(Z_i),$$
$$\widetilde{D}_i = D_i - \widehat{\alpha}^{(-k(i))}(Z_i),$$
$$\widehat{\theta} = \frac{\sum_{i=1}^{n} \widetilde{D}_i \widetilde{Y}_i}{\sum_{i=1}^{n} \widetilde{D}_i^2}.$$

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 6 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 7. 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 6: 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 7: Cross-fitted nuisance prediction errors from the saved Monte Carlo run.

J Additional Details for Empirical Application

J.1 Individual-Level Cross-Fitting for Dube's Variables

Table 8 repeats the Dube specification with three individual-level folds and five repetitions. Both the outer cross-fitting folds and the inner tuning or validation samples are split at the individual level, as in Dube et al. (2020), rather than at the requester level as in the main results. Standard errors remain requester-clustered and use repetition-variance aggregation as in the main results. The sample and the 611 controls are unchanged.

The individual-level split produces OLS, lasso, and ridge duration coefficients of −0.3557, −0.3478, and −0.3538, with standard errors close to 0.052. These differ substantially from the requester-level results in Table 4. The XGB 1–3 coefficients, −0.0417, −0.0354, and −0.0343, are closer to their requester-level counterparts, but their standard errors fall to approximately 0.004. For these learners, reward R^2_X values rise from 0.33–0.36 under requester-level splitting to 0.75–0.77 under individual-level splitting. This comparison illustrates that clustering the reported standard errors and choosing the unit of cross-fitting are distinct decisions. Potentially, the very small p-values in Table 8 can be explained by the over-fitting of the learners due to within-cluster dependence. To mitigate this concern, we use cross-fitting with cluster-level folds in our analysis reported in Table 4.

Table 8: Dube controls with individual-row cross-fitting

Panel A: Linear and random-forest learners						
	OLS	CV-Lasso	CV-Ridge	RF 1	RF 2	RF 3
Dube (611 features)						
Estimate (s.e.)	−0.3557 (0.0517)	−0.3478 (0.0521)	−0.3538 (0.0518)	−0.1833 (0.0838)	−0.2494 (0.0669)	−0.3692 (0.0382)
p-value	5.935×10^{-12}	2.395×10^{-11}	8.243×10^{-12}	0.0287	0.0002	4.713×10^{-22}
R^2_Y / R^2_D	−3.0073 / −2.6939	0.0960 / −2.5199	−3.0079 / −2.6945	0.2294 / 0.2817	0.2783 / 0.3523	0.4693 / 0.6629
Panel B: XGBoost and neural-network learners						
	XGB 1	XGB 2	XGB 3	NN 1	NN 2	NN 3
Dube (611 features)						
Estimate (s.e.)	−0.0417 (0.0040)	−0.0354 (0.0043)	−0.0343 (0.0040)	−0.2320 (0.0291)	−0.2877 (0.0639)	−0.2708 (0.0338)
p-value	4.542×10^{-25}	2.076×10^{-16}	7.302×10^{-18}	1.574×10^{-15}	6.620×10^{-6}	1.184×10^{-15}
R^2_Y / R^2_D	0.8866 / 0.7471	0.8917 / 0.7562	0.8944 / 0.7704	0.5868 / −1.4819	−17.0267 / −3.7225	−4.0260 / −3.4895

Notes: Individual-level cross-fitting uses 3 folds and 5 repetitions; 258,352 observations and 24,923 requesters. Estimates are median coefficients on log reward in the log-duration equation. Parentheses report requester-clustered standard errors with split-deviation inflation. Two-sided normal p-values test a zero coefficient. The final row lists out-of-sample outcome and treatment R^2 scores, averaged across repetitions.

J.2 Additional Results on Robustness to Pre-Trained Embeddings

Figures 7 and 8 display the coefficients and confidence intervals for XGB 1 and XGB 3. These specifications use the same embeddings and estimation procedure as the XGB 2 results in

Figure 6.

J.3 Visualization of Task-Description Embeddings

We visualize the task-description embeddings from the empirical application in Section 7 using Uniform Manifold Approximation and Projection (UMAP), a nonlinear dimension-reduction method developed by McInnes, Healy, and Melville (2018). UMAP constructs a weighted graph of nearby observations in the original embedding space and finds a low-dimensional representation that approximates these neighborhood relationships. We use two dimensions for visualization; the axes have no direct economic interpretation, and distances between separated groups need not reflect distances in the original embedding space. It is known that UMAP can reveal meaningful low-dimensional structures in high-dimensional data, especially for the local neighborhood relationships.

Figures 9 to 11 display the UMAP projection of task-description embeddings constructed from the BERT mean, colored by task category, log posted reward, and log duration. These visualizations describe how embeddings capture task characteristics hand-engineered by Dube et al. (2020), and they are useful controls in subsequent empirical analyses using the log reward as the main regressor and log duration as the outcome.

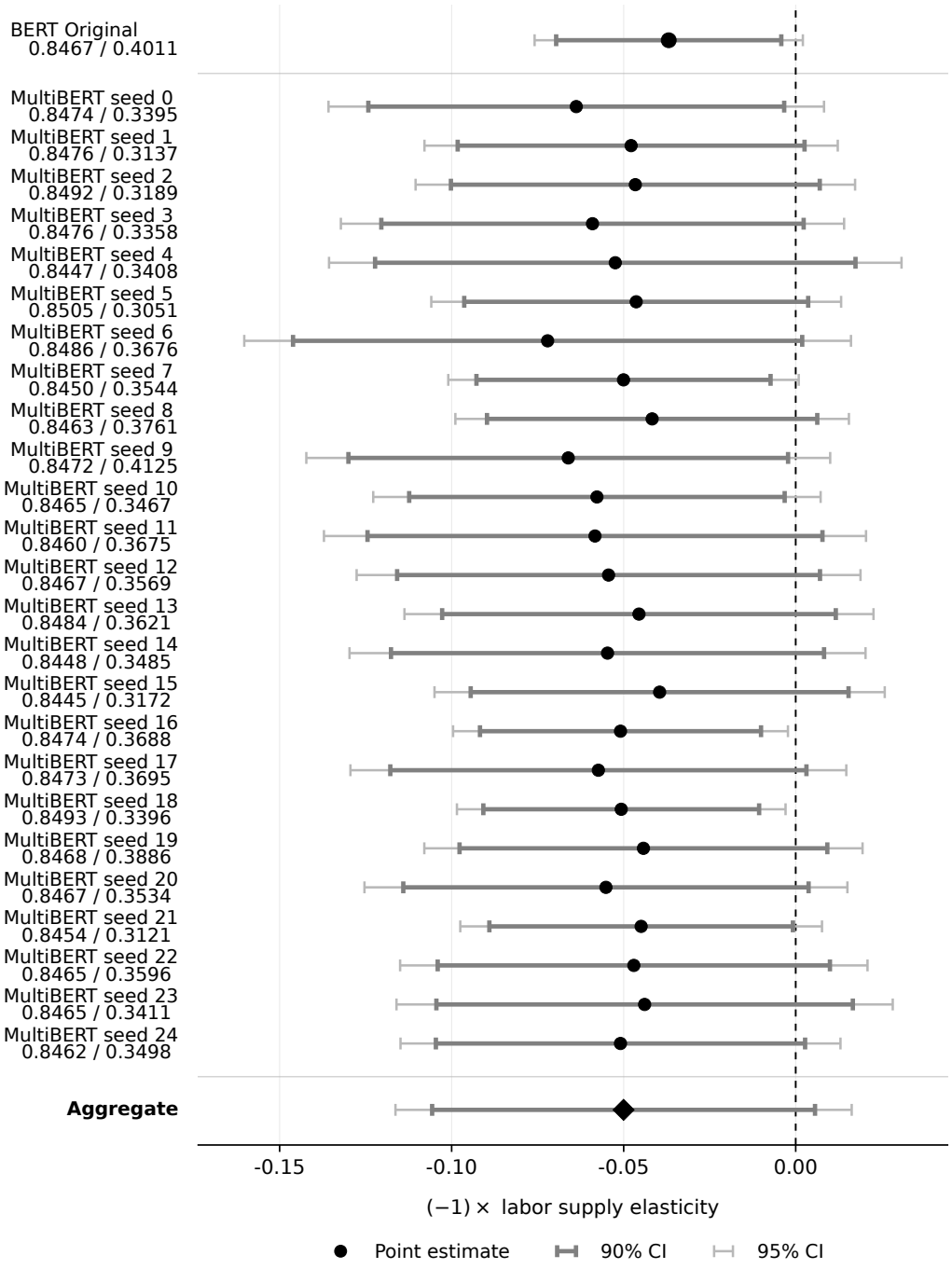


Figure 7: Coefficient robustness to 26 BERT embeddings (Base BERT with Mean pooling) with XGB 1 learners

Notes: Same as Figure 6.

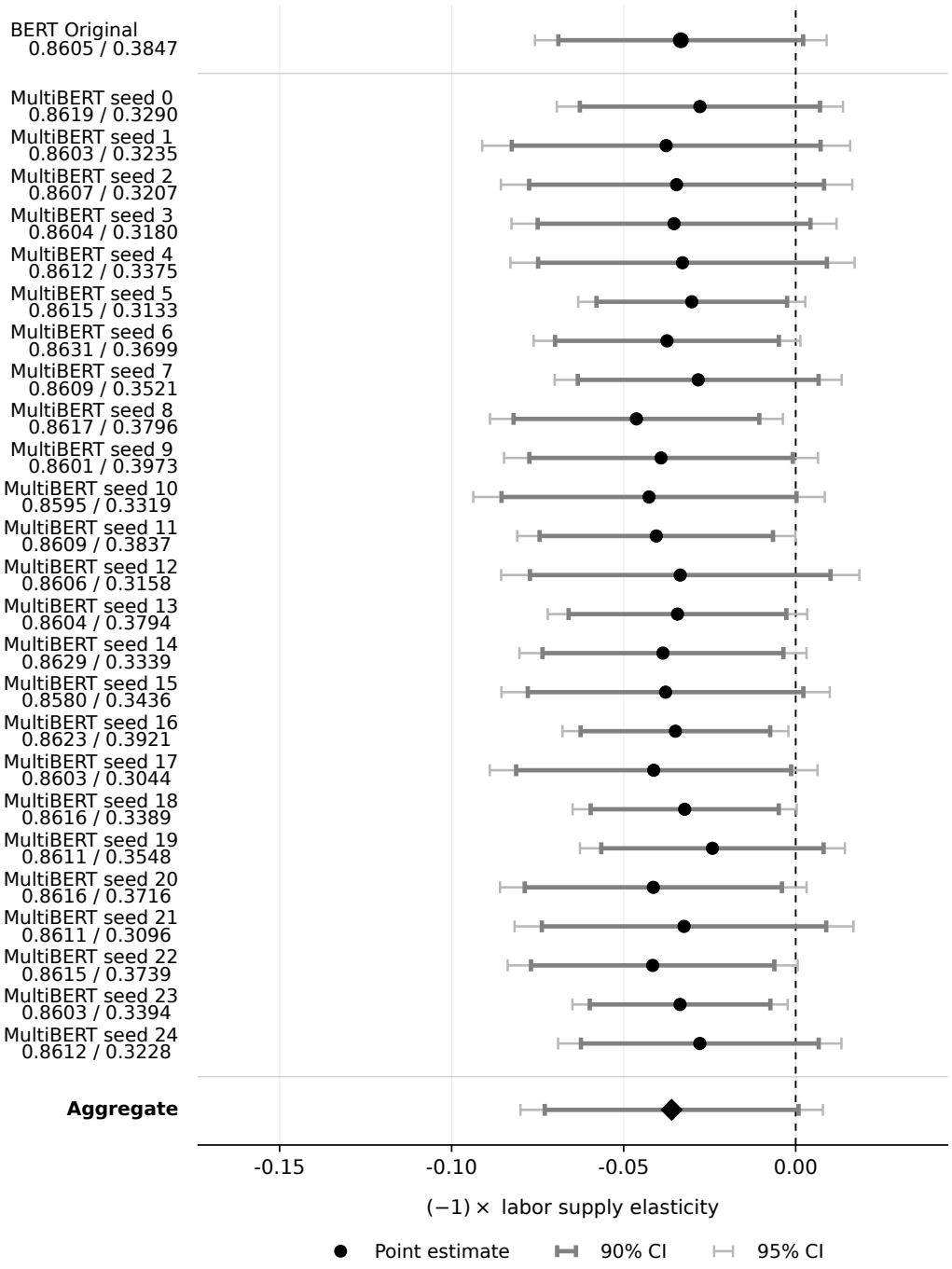


Figure 8: Coefficient robustness to 26 BERT embeddings (Base BERT with Mean pooling) with XGB 3 learners

Notes: Same as Figure 6.

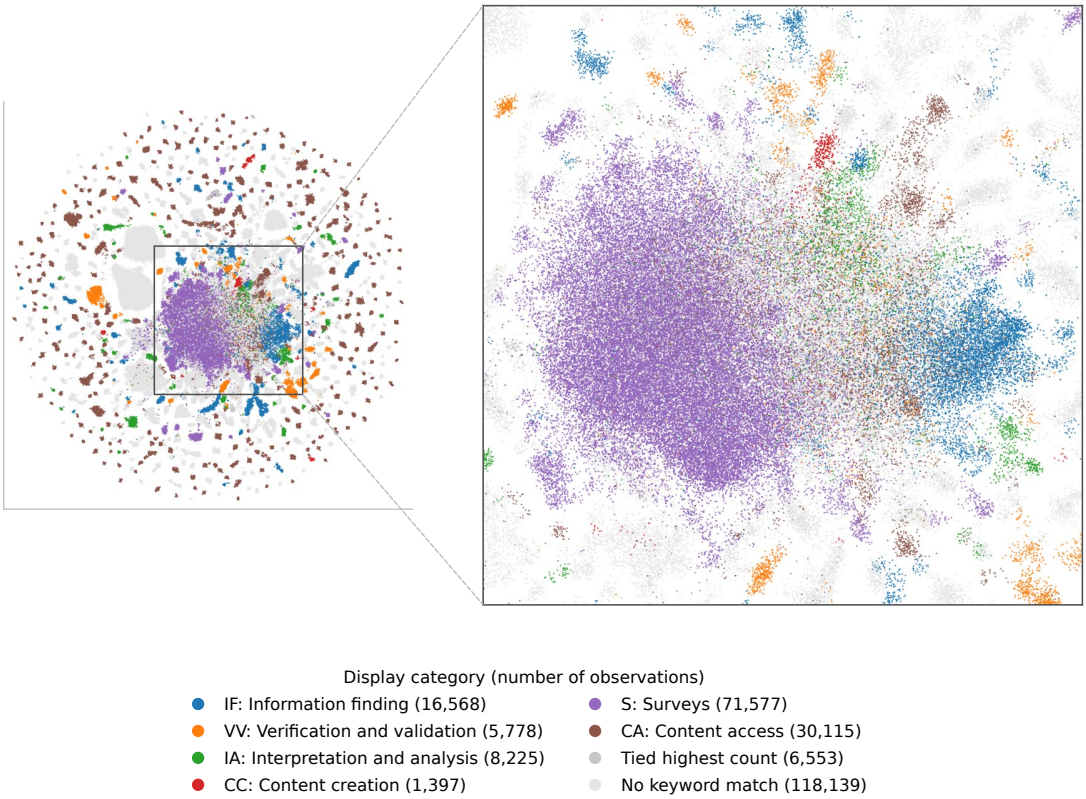


Figure 9: UMAP projection of task-description embeddings, colored by task category used by Dube et al. (2020)

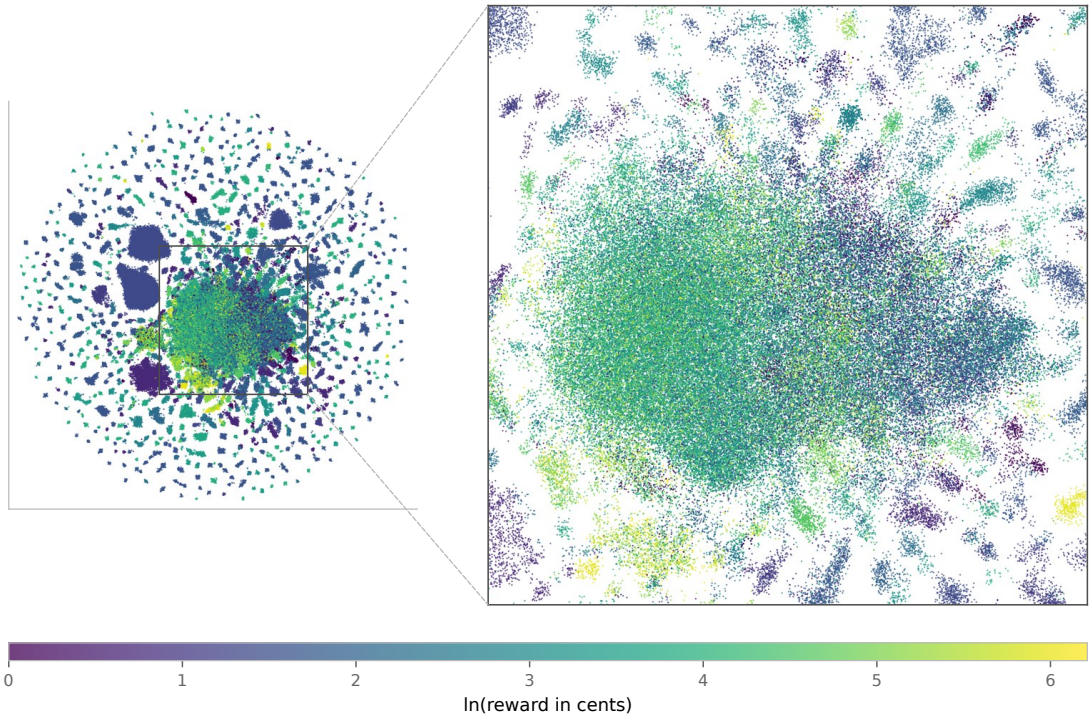


Figure 10: UMAP projection of task-description embeddings, colored by log reward

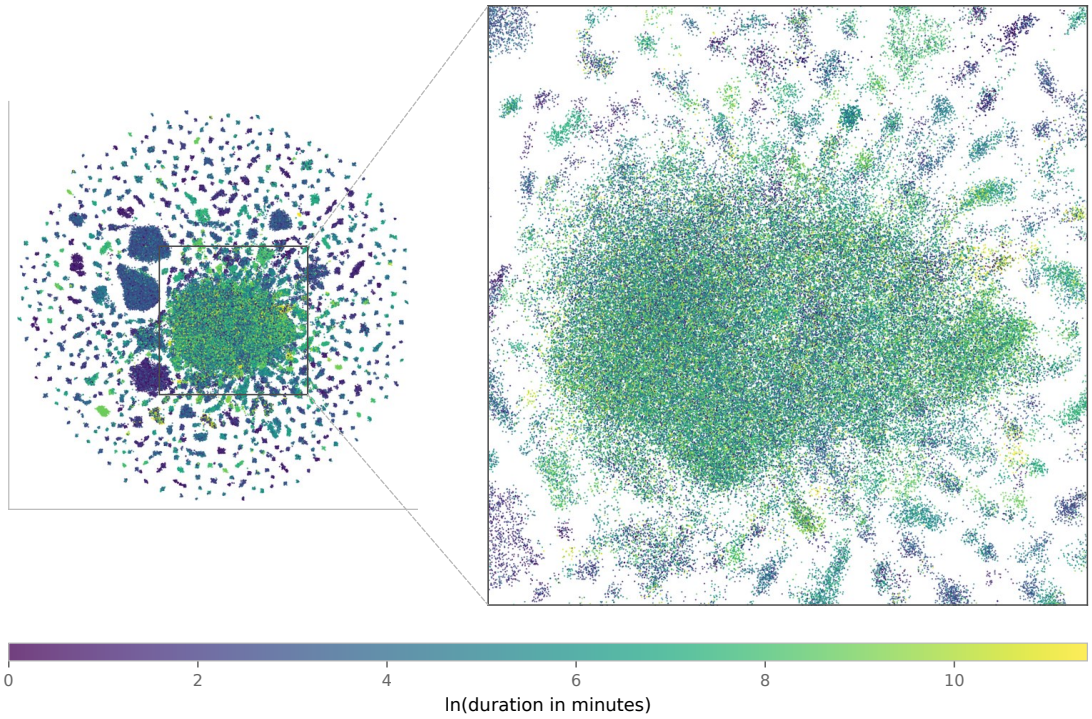


Figure 11: UMAP projection of task-description embeddings, colored by log duration