THE UNIVERSITY OF CHICAGO

MACHINE LEARNING FOR ECONOMETRICS: THEORETICAL FOUNDATIONS AND EMPIRICAL INSIGHTS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
ZHOUYU SHEN

CHICAGO, ILLINOIS
JUNE 2025

To everyone who has guided, encouraged, and supported me throughout this journey

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I wish to express my deepest gratitude to my advisor, Prof. Dacheng Xiu. His guidance has been instrumental not only in shaping my academic journey but also in supporting me personally. Prof. Xiu's passion for academia has profoundly inspired my own aspirations in the field. I am also deeply appreciative of my committee members: Prof. Chao Gao, Prof. Christian Hansen, and Prof. Tengyuan Liang. Their invaluable insights and unwavering support have significantly enriched my PhD experience.

My time at Booth has been further enhanced by the exceptional support from our PhD office, for which I am truly thankful. I am fortunate to have met a number of inspiring friends at Booth, especially Rui Da, Dake Zhang, Boxin Zhao, Wenxuan Guo, Yifei Chen, Cong Zhang, Ming Gao, Chang Deng, and many others. The friendships we have formed have been a great source of joy throughout my PhD, and I am deeply grateful for their companionship.

Finally, my heartfelt thanks go to my family. To my parents, Xuchun Zhou and Haojie Shen, whose encouragement of my curiosity and unwavering support have been the bedrock of my academic pursuits. To my girlfriend, Ruxin Shi, whose steadfast support and companionship have been my anchor throughout this journey.

ABSTRACT

Machine learning techniques have been increasingly adopted in economics and finance for forecasting, policy evaluation, and structural modeling. Despite their empirical success, a key question remains: can machine learning methods—originally developed for tasks in computer science—effectively address the unique challenges posed by economic data, such as weak signals, cross-sectional dependence, and temporal dependence?

This dissertation explores the theoretical foundations and empirical relevance of modern machine learning methods in econometric applications. Chapter I examines the performance of several popular machine learning estimators in high-dimensional regressions with low signal-to-noise ratios. We show that Ridge regression and ℓ_2-regularized neural networks can effectively exploit weak and possibly nonlinear signals, while Lasso may fail to outperform a zero benchmark due to its reliance on sparsity. Both theoretical analysis and empirical studies across multiple economic datasets highlight that signal weakness, rather than lack of sparsity, may be the main bottleneck in prediction.

Chapter II develops non-asymptotic guarantees for deep autoencoders within a high-dimensional nonlinear factor model. We show that autoencoders can consistently recover latent structures, with approximation errors diminishing as dimensionality and sample size increase. These results justify the use of autoencoders in tasks such as asset return prediction and structured matrix completion for causal analysis.

Chapter III studies the application of Recurrent Neural Networks (RNNs) in modeling nonlinear time series. Under a nonlinear autoregressive and moving-average model with exogenous variables, we establish finite-sample error bounds for RNN-based predictions. These bounds reveal the role of function smoothness, input dimensionality, and process dependence in shaping prediction accuracy. Our analysis shows that RNNs are particularly effective in capturing nonlinear dynamics in noise, offering advantages over traditional nonparametric methods.

Together, these essays contribute to a deeper understanding of the conditions under which machine learning methods are both theoretically sound and practically useful for econometric modeling and economic prediction.

CHAPTER 1

CAN MACHINES LEARN WEAK SIGNALS?

1.1 Introduction

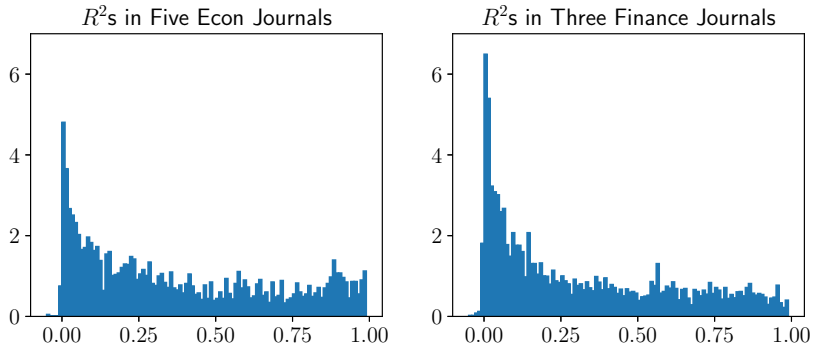
In regression analysis, covariates with non-zero coefficients are identified as true signals, while those with zero coefficients are considered false signals. In a population model, this distinction is clear-cut, resembling a "black and white" scenario. However, in finite samples, the presence of minuscule non-zero coefficients introduces a "gray" area, blurring the boundary between true and false signals.[1] This gray area represents weak signals—covariates that, individually, exert negligible influence on the outcome variable.

The investigation of weak signals holds tangible implications for economic and financial decision-making. Often, it is the collective impact of these weak signals that drives the outcomes in these fields. Supporting this, Figure 1.1 provides empirical evidence by presenting R^2 values gathered from a compendium of Economics and Finance journal articles published in 2022. The 25% quantiles of these R^2 values stand at 9.7% for economics and 5.8% for finance, suggesting that models in these disciplines frequently rely on covariates with modest explanatory power. Furthermore, Figure 1.1 is based exclusively on published studies, which are likely biased toward higher R^2 values due to selection effects. This suggests that the presence of weak signals may be even more widespread than the data here indicates.

The decision to incorporate weak signals into a regression model carries the risk of overfitting, potentially undermining predictive accuracy. Overfitting arises when the increased variance from estimating the coefficients of these weak signals outweighs the bias reduction gained from their inclusion. This trade-off becomes even more pronounced in high-dimensional settings, where the limited sample size relative to the number of covariates amplifies prediction errors.

1. The comparison of regression coefficients' magnitudes is meaningful only when predictor variables are normalized, an assumption implicitly adopted in the subsequent discussion.

Figure 1.1: Histograms of R^2s in Selected Economics and Finance Journals



Note: The histograms depict R^2s manually collected from published papers in a selection of Economics and Finance journals in 2022. This collection comprises data from five Economics journals (left) and three Finance journals (right). Specifically, there are a total of 411 papers published in the American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Studies, resulting in $8,129$ R^2 observations. In addition, there are 380 papers from the Journal of Finance, Journal of Financial Economics, and Review of Financial Studies, contributing $12,198$ R^2 observations.

Machine learning methods, renowned for their focus on variable selection and dimension reduction, have proven effective in mitigating overfitting and detecting true signals from false ones, particularly when the true signals are strong. These methods employ regularization techniques, such as penalizing the ℓ_1 or ℓ_2 norms of model parameters, to achieve this objective. However, a pivotal question emerges: Can machines learn weak signals, or in other words, can they surpass the naive zero predictor? The zero predictor, which disregards all covariates and always predicts zero, serves as a passive baseline in the context of weak signals. Surpassing this baseline indicates that a method has successfully extracted valuable signals. Conversely, failing to exceed it highlights a limitation in its learning capacity. In view of these considerations, we focus on evaluating the relative performance of regularized predictors, Ridge and Lasso, against the zero predictor in high-dimensional regressions, where the prediction target is driven by predictors that exhibit weak correlations with it.

In scenarios with sufficiently strong signals, both Lasso and Ridge are expected to outperform zero by effectively capturing and utilizing at least some of these signals. Hence,

accurately defining the notion of "weak" signals is crucial at the outset of our investigation. This definition serves a dual purpose: it prevents the scenario from defaulting to trivial comparisons akin to strong signal cases and ensures practical relevance to finite sample scenarios in economics and finance. We characterize a weak signal scenario as one in which the zero predictor achieves the minimal Bayes prediction risk asymptotically. This setting turns out to encompass a wide class of data generating processes (DGPs), and it approximates a finite sample reality in which zero serves as a competitive benchmark.

In the defined weak signal scenarios, conventional error-bound analyses are insufficient in distinguishing the performance of different predictors. Considering the optimal Lasso and Ridge—each tuned to minimize prediction error—both can perform no worse than the zero predictor. Intuitively, as the tuning parameters for Ridge and Lasso approach infinity, Ridge converges to zero, while Lasso effectively becomes equivalent to zero. Consequently, all three predictors—optimal Lasso, optimal Ridge, and zero—can achieve the optimal Bayes prediction error, making them indistinguishable.

However, a more refined analysis reveals that, across a broad range of tuning parameter choices, Ridge outperforms zero. In contrast, the optimal Lasso effectively reduces to zero, meaning Lasso does not exceed zero's performance regardless of its tuning parameter choice. This conclusion is based on a precise error approach, which enables us to zoom in and explicitly characterize the *relative* prediction errors of Ridge and Lasso compared to the zero predictor. These results highlight an important distinction between shrinkage and selection methods. Shrinkage methods like Ridge are more effective in environments with more homogenous signal strength. On the other hand, selection methods like Lasso are preferable in scenarios where there is a clear distinction between true and false signals. In weak signal contexts where this distinction is blurred, the advantage of Lasso tends to wane.

Our study further demonstrates the effectiveness of the cross-validation (CV) algorithm in selecting the optimal tuning parameter for Ridge, even in weak signal contexts. This under-

scores CV's robustness as a model-tuning tool in these settings. Moreover, the out-of-sample R^2 from the optimal Ridge—a metric frequently used for assessing predictor performance on unseen data—proves a relevant indicator of the signal-to-noise ratio in the DGP, despite a notable gap between its asymptotic limit and the population R^2.

In the final aspect of our theoretical analysis, we expand our framework to include models featuring a mix of signal strengths. This section specifically addresses scenarios in which a benchmark model contains potentially strong signals. Our focus then shifts to evaluating the benefits of leveraging the predictive power of the remaining weak signals. To this end, we derive ordinary least squares (OLS) residuals, from which the impact of potentially strong covariates in the benchmark model has been removed. Consistent with our earlier findings, applying Ridge regression to these residuals, using the remaining covariates, enhances predictive performance compared to a baseline predictor that ignores these additional covariates.

Our simulation analysis corroborates our theoretical findings: Ridge surpasses zero, which in turn edges out Lasso, especially in DGPs characterized by low R^2 values. When exploring more sophisticated machine learning techniques, we observe that Random Forest (RF), which produces dense models by including nearly all variables, outperforms the zero predictor. The latter, in turn, surpasses Gradient Boosted Regression Trees (GBRT), as GBRT, similar to Lasso, tends to generate sparse models. Furthermore, Neural Networks (NNs), when paired with the ℓ_2-norm regularization, can deliver superior predictions. In contrast, applying an ℓ_1-penalty in these networks fails to achieve comparable results.

Our empirical analysis spans six datasets across macroeconomics, microeconomics, and finance. Five align with Giannone et al. [2022], and one is sourced from Gu et al. [2020]. Our finance examples delve into predicting market returns using financial and economic indicators, as well as firm-level return prediction based on their specific characteristics. In the macroeconomic context, we examine time-series predictions of industrial production using macroeconomic indicators, and a cross-country GDP growth prediction, utilizing socio-

economic, institutional, and geographical factors. Our microeconomic studies focus on crime rate predictions and pro-plaintiff appellate decisions in takings law rulings.

The relevance of weak signals in datasets is contingent on the choice of benchmark models. For instance, when compared to a benchmark with an intercept alone, weak signals are revealed in four out of six datasets. Further benchmarking against covariates informed by economic theory reveals weak signals across all datasets, making them particularly well-suited for the application of our asymptotic theory. Drawing from their empirical analysis of these datasets, Giannone et al. [2022] argue that sparsity may be an illusion, as optimal predictive models often rely on a large number of covariates. Our collective theoretical and empirical evidence points to signal weakness as a key factor in the underperformance of Lasso. As our results suggest, even in cases where the majority of signals have zero coefficients in the true DGP, Ridge may still outperform Lasso if the true signals are weak. Their comparative performance thus does not necessarily offer insights into the sparsity of the DGP itself.

In light of these findings, we recommend a cautious approach to employing Lasso in economic and financial settings. Despite its popularity as a modern counterpart to OLS, Lasso's effectiveness may be compromised in scenarios characterized by weak signals. Our study complements the findings of Kolesár et al. [2024], who highlight issues with sparsity-based estimators, such as their lack of invariance to reparametrization and sensitivity to normalizations that are otherwise innocuous to OLS.

Our paper is closely related to the literature on the theoretical performance of Ridge and Lasso, with two main threads being particularly relevant. The first focuses on error-bound analysis. For Ridge, Hoerl and Kennard [1970] show that the prediction error decreases at a rate of p/n, where p is the number of covariates and n the sample size, with its magnitude tied to the eigen-structure of the design matrix. For Lasso, the prediction error vanishes if $s \log p/n \to 0$, where s is the number of non-zero parameters (see, e.g., Wainwright [2019]). However, we consider an asymptotic setting where these error bounds fail to distinguish Ridge

and Lasso from the zero predictor, as their leading-order prediction errors are identical. This motivates a more granular, higher-order analysis of prediction errors.

The second, more recent strand of research focuses on determining the precise probability limit of the prediction error for Ridge and Lasso.[2] Bayati and Montanari [2012] employ approximate message passing algorithms to link them with Lasso and derive its error limit. Alternatively, Thrampoulidis et al. [2015] use the Convex Gaussian Minimax Theory (CGMT) to simplify Lasso's optimization problem, enabling precise error derivation. For Ridge, Dicker [2016] provides analogous insights into its prediction error. However, these precise error analyses often rely on stringent parametric assumptions, such as independently Gaussian-distributed design matrix elements. Dobriban and Wager [2018] extend Dicker [2016]'s work by accommodating dependent covariates and non-Gaussian predictors, leveraging universality results from random matrix theory.

This chapter is organized as follows. Section 1.2 presents the main theoretical results regarding Ridge and Lasso. Section 1.3 conducts simulations to illuminate our theoretical predictions while also expanding the analysis to assess the performance of advanced machine learning methods under weak signals. Lastly, Section 1.4 provides empirical results. The appendix contains mathematical proofs of main results, while the appendix provides additional theoretical results, technical lemmas, and their proofs.

Notation: For any $x \in \mathbb{R}$, we refer to $\max(x, 0)$ as x_+. For any vector x, $\|x\|_0$, $\|x\|_1$, $\|x\|$ and $\|x\|_\infty$ represent its ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms, respectively. For a real matrix A, we use $\|A\|$ and $\|A\|_{\mathrm{F}}$ to denote its spectral norm (or ℓ_2 norm), and the Frobenius norm, that is, $\sqrt{\lambda_{\max}(A^\top A)}$, and $\sqrt{\mathrm{Tr}(A^\top A)}$, respectively. In the case where A is a $p \times p$ matrix, $\lambda_i(A)$ denotes its i-th largest eigenvalue, for $1 \le i \le p$. We use the notation $x_n \lesssim y_n$ when

2. The precise error analysis has provided valuable insights into various machine learning methods. For example, Liang and Sur [2022] examine the properties of minimum ℓ_1-norm interpolation and boosting in linear models. Miolane and Montanari [2021] explore cross-validation for Lasso, while Hastie et al. [2022] investigate minimum ℓ_2-norm interpolation, shedding light on the double-descent phenomenon in neural networks. Regarding variable selection, Su et al. [2017] study the false discovery rate of the Lasso path, and Wang et al. [2020] compare the variable selection properties of bridge estimators.

there exists a constant C such that $x_n \leq Cy_n$ holds for sufficiently large n. Similarly, we use $x_n \lesssim_{\mathrm{P}} y_n$ to denote $x_n = O_{\mathrm{P}}(y_n)$. If $x_n \lesssim y_n$ and $y_n \lesssim x_n$, we write $x_n \asymp y_n$ for short. Similarly, we use $x_n \asymp_{\mathrm{P}} y_n$ if $x_n \lesssim_{\mathrm{P}} y_n$ and $y_n \lesssim_{\mathrm{P}} x_n$.

1.2 Theoretical Results

1.2.1 Model Setup

We start with the following linear regression model:

$$y = X\beta_0 + \varepsilon, \tag{1.1}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta_0 \in \mathbb{R}^p$ and $\varepsilon \in \mathbb{R}^n$. Throughout our discussion, X, β_0, and ε are treated as random variables, mutually independent of one another. Central to our analysis is the calculation of the probability limit of the prediction errors, which necessities assuming a (prior) probability distribution on the coefficients. This setting aligns with the literature on precise errors, which also connects our analysis of prediction error with Bayes risk.

Our objective is to investigate the predictive performance of machine learning techniques in the presence of weak signals.[3] To accomplish this, we focus on a high-dimensional regression setting characterized by an increasing number of predictors, that is, $p \to \infty$. In such a context, regularization techniques become not just relevant but often necessary. Moreover, our specific focus is on situations where the signals are weak, characterized by the condition: $\|\beta_0\|^2 \asymp_{\mathrm{P}} \tau \to 0$. The choice to use $\|\beta_0\|$ as the metric for characterizing weak signals is due to its close relationship with the widely-adopted R^2 metric in regression analysis, which provides a familiar and intuitive understanding of signal strength. Our investigation then progresses to an asymptotic analysis under these two conditions. We will detail the specific

3. While Donoho and Jin [2004] and Hall and Jin [2010] study variable selection in rare and weak signals, we focus on prediction in asymptotic settings where identifying nonzero coefficients is infeasible.

requirements for p, τ, and sample size n after introducing the baseline predictor.

Now, we proceed to present the assumptions governing the DGP of X:

Assumption 1. *The covariates $X \in \mathbb{R}^{n\times p}$ are generated as $X = \Sigma_1^{1/2} Z \Sigma_2^{1/2}$ for an $n \times p$ matrix Z with i.i.d. standard Gaussian entries, deterministic $n \times n$ and $p \times p$ positive definite matrices Σ_1 and Σ_2. In addition, there exist positive constants c_1, C_1, c_2, C_2 such that $c_1 \leq \lambda_i(\Sigma_1) \leq C_1$, $i = 1, 2, \ldots, n$ and $c_2 \leq \lambda_i(\Sigma_2) \leq C_2$, $i = 1, 2, \ldots, p$.*

This assumption accommodates time series dependencies via Σ_1 and cross-sectional correlations via Σ_2. The eigenvalue constraints serve two purposes: upper bounds limit excessive dependencies, while lower bounds prevent multicollinearity and ensure observations are not linearly dependent across time. Moreover, since we focus on prediction rather than variable selection, the dependence structure among X does not adversely impact Lasso's predictive performance (see Section 7.4 of Wainwright [2019]), even though its variable selection properties are sensitive to strong dependence among covariates.

While the Gaussian assumption for X is integral to our use of Gordon's inequality (Gordon [1988]) for Gaussian processes in the proof, it does raise concerns regarding the robustness of our findings when this assumption is not met in real-world scenarios. Our simulation results (not included for reasons of space) indicate that the Gaussian assumption appears non-essential and our asymptotic theory approximates finite sample behavior even with relatively small sizes, typically a few hundred observations. This observation aligns with similar findings in random matrix theory, where asymptotic properties initially derived for Gaussian ensembles were subsequently shown to extend to a wider spectrum of random matrices—a phenomenon referred to as the universality. Notably, when $\Sigma_1 = \mathbb{I}$, Dobriban and Wager [2018] use random matrix theory to bypass the Gaussian assumption. However, their technique appears only applicable to Ridge. Our objective is to compare Ridge and Lasso under a unified framework, which necessitates the Gaussian assumption.

Next, we specify the assumption regarding ε:

Assumption 2. *Let $\varepsilon = \Sigma_\varepsilon^{1/2} z$, where z comprises i.i.d. variables with mean zero, variance one and finite fourth moment and Σ_ε is a positive definite matrix satisfying $c_\varepsilon \leq \lambda_i(\Sigma_\varepsilon) \leq C_\varepsilon$, $i = 1, 2, \ldots, n$, for some fixed positive constants c_ε and C_ε.*

This assumption allows for autocorrelations and heteroscedasticity. If Σ_ε is a diagonal matrix, its spectral norm is evidently bounded under the condition that each entry of ε has finite variance. Moreover, if ε follows a stationary process characterized by exponentially decaying autocorrelations, it can be shown that the spectral norm of Σ_ε remains bounded.

Under Assumptions 1 and 2, it follows that $\|X\beta_0\| \asymp_{\mathrm{P}} \sqrt{n}\,\|\beta_0\|$ and $\|\varepsilon\| \asymp_{\mathrm{P}} \sqrt{n}$. This indicates that the magnitude of each entry in X and ε neither explode nor vanish asymptotically. Consequently, the magnitude of the signal-to-noise ratio (or R^2) is entirely dictated by $\|\beta_0\|$. Next, we impose an assumption that governs the properties of β_0:

Assumption 3. *The vector $b_0 = \sqrt{p\tau^{-1}}\beta_0$ comprises i.i.d. random variables, each following a prior probability distribution F belonging to the class $\mathcal{F}$. The class $\mathcal{F}$ is defined such that any included random variable can be represented as $q^{-1/2} b_1 b_2$, where b_1 and b_2 are independent, b_1 follows a binomial distribution $B(1, q)$, and b_2 is a sub-exponential random variable with a mean of zero and a variance denoted as σ_β^2.*

Without loss of generality, we use the term $\sqrt{p\tau^{-1}}$ as the normalization factor, ensuring that $\|\beta_0\|^2 \asymp_{\mathrm{P}} \tau$. This normalization facilitates a clearer interpretation of our results. While the i.i.d. assumption may seem strong, it offers greater transparency by simplifying more complex technical assumptions necessary to derive essential probability bounds. In particular, this assumption allows for important classes of models, such as a spike-and-slab prior for b_0, extensively studied by Giannone et al. [2022] to examine the empirical relevance of sparsity in economic datasets. Each element of b_0 follows a mixed distribution, such as when $q = 1$, with b_2 modeled by $(1 - \upsilon)\psi_0 + \upsilon\psi_1$, where υ, a fixed constant within $[0, 1]$, modulates the mix between the spike (ψ_0) and slab (ψ_1) components of the prior. More generally, the formulation $q^{-1/2} b_1 b_2$ accommodates a spike-and-slab model with more

extreme sparsity ($q \to 0$) through the component $q^{-1/2}b_1$. This scaling, $q^{-1/2}$, ensures the variance of $q^{-1/2}b_1b_2$ remains finite and non-vanishing. Essentially, q dictates the sparsity of β_0: when $\mathrm{P}(b_2 = 0) = 0$, $\|\beta_0\|_0 \asymp_{\mathrm{P}} pq$. In scenarios with strong signals ($\tau = 1$), a DGP with q nearing zero typically favors Lasso, whereas a q closer to one suggests a preference for Ridge. Therefore, this framework does not inherently privilege either.

The underlying assumptions that justify Ridge and Lasso are notably distinct, particularly in the context of error-bound analysis. For instance, the analysis of Lasso often requires the approximate sparsity condition and the restricted eigenvalue condition (see, e.g., Belloni et al. [2013b] and Bickel et al. [2009]). On the other hand, the convergence rate of Ridge's prediction error requires intricate conditions on the eigenvalue structure of the design matrix, as discussed in Tsigler and Bartlett [2023]. In contrast, our analysis here compares the asymptotic properties of different estimators within a common framework.

1.2.2 Predictors

We now turn our attention to the discussion of the predictors. In scenarios involving weak signals, characterized by $\|\beta_0\| \to 0$, a straightforward and natural baseline predictor emerges, that is, the naive zero predictor. As an estimator, zero is clearly consistent in terms of the ℓ_2-loss of the estimation error, because it reduces to $\|\beta_0\|$, which vanishes in this context.

The zero predictor serves as a passive benchmark for a scenario where no learning occurs. To surpass its performance, any alternative predictor must harness some signals. This indicates the alternative predictor's capability to successfully identify and leverage weak signals. Therefore, to address the earlier question of whether machines can learn weak signals, we need to compare the machine learning method's performance with that of the zero predictor. Only if they can do so can they outperform the naive zero predictor.

In our study, we consider Ridge and Lasso as contenders that leverage machine learning techniques. These methods are widely used benchmarks in practice, owing to their simplicity

and universality. An in-depth analysis of these predictors provides valuable insights into their specific regularization techniques, which can be extended to more advanced models.

The Ridge estimator, denoted as $\hat{\beta}_r$, is the solution to the following optimization problem:

$$\hat{\beta}_r(\lambda_n) := \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2, \tag{1.2}$$

where λ_n is its tuning parameter governing the strength of the regularization. In contrast, the Lasso estimator, denoted as $\hat{\beta}_l$, is defined as:

$$\hat{\beta}_l(\lambda_n) := \arg\min_{\beta} \frac{1}{n}\|y - X\beta\|^2 + \frac{\lambda_n}{\sqrt{n}}\|\beta\|_1. \tag{1.3}$$

By convention, and without loss of generality, the terms involving penalties are typically scaled by p/n in the case of Ridge and by $1/\sqrt{n}$ in the case of Lasso.

In addition, our theoretical results also encompass the OLS estimator and the Ridgeless estimator, both of which correspond to special cases of Ridge when the tuning parameter λ_n is set to zero. When $p > n$, the least squares problem has an infinite number of solutions. Among these solutions, the Ridgeless estimator can be regarded as a minimum-norm interpolating linear predictor, as noted by Bartlett et al. [2020]:

$$\hat{\beta}_r(0) = \arg\min_{\beta} \|\beta\|, \quad \text{s.t.} \quad X\beta = y. \tag{1.4}$$

With β_0 estimated by some $\hat{\beta}$, it is straightforward to construct corresponding predictor, $(x^{\text{new}})^\top\hat{\beta}$, for a new data point x^{new}.

1.2.3 Bayes Risk

Now, we proceed to define the metric by which we assess various predictors. For any predictor, our primary interest is its Bayes prediction risk. This risk is related to the expected

squared prediction error evaluated at a new, independent data point $(x^{\text{new}}, y^{\text{new}})$. In the case of a linear predictor, $\hat{y}^{\text{new}} = (x^{\text{new}})^\top \hat{\beta}$, we can write the prediction error explicitly as:

$$\mathbb{E}_F(y^{\text{new}} - \hat{y}^{\text{new}})^2 = \sigma_\varepsilon^2 + \mathbb{E}_F\left[(x^{\text{new}})^\top(\hat{\beta} - \beta_0)\right]^2 = \sigma_\varepsilon^2 + \mathbb{E}_F\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2, \qquad (1.5)$$

where the subscript in the expectation operator $\mathbb{E}_F(\cdot)$ emphasizes the fact that the expectation is taken with respect to the prior distribution of $b_0 = \sqrt{p\tau^{-1}}\beta_0$. Obviously, it is the second term in (1.5) that governs the relative predictive performance. Barring $\|\Sigma_2\|$, the prediction risk is closely tied to the estimation error of $\hat{\beta}$, i.e., $\|\hat{\beta} - \beta_0\|$.

Formally, we define the Bayes prediction risk associated with an estimator $\hat{\beta}$ as

$$\mathcal{R}(\hat{\beta}, F) := \mathbb{E}_F\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2.$$

The predictor that minimizes this Bayes risk is termed the Bayes predictor. In our framework, it is straightforward to show (see Chapter 4 of Berger [1985]) that the Bayes predictor corresponds with the predictor that is derived from the posterior mean of β_0, which is represented as $\mathbb{E}_F(\beta_0|X, y)$. We denote its Bayes risk as $\mathcal{R}(F)$.[4]

Under strong signals, i.e., $\tau = 1$, and suppose that Σ_1, Σ_2 and Σ_ε are identity matrices, $p/n \to c_0 \in \mathbb{R}^+$, significant progress has been made in understanding the asymptotic behavior of Bayes risk. For Ridge regression, notable studies by Dicker [2016] and Dobriban and Wager [2018] have derived the asymptotic limit of the Bayes risk.[5] Similarly, in the case of Lasso, several studies, such as those conducted by Bayati and Montanari [2012] and

4. There exists an extensive body of literature focused on empirical Bayes methods, which explores feasible approaches for implementing $\mathbb{E}_F(\beta_0|X, y)$, when F is unknown, see, e.g., Robbins [1964] and Efron [2012].

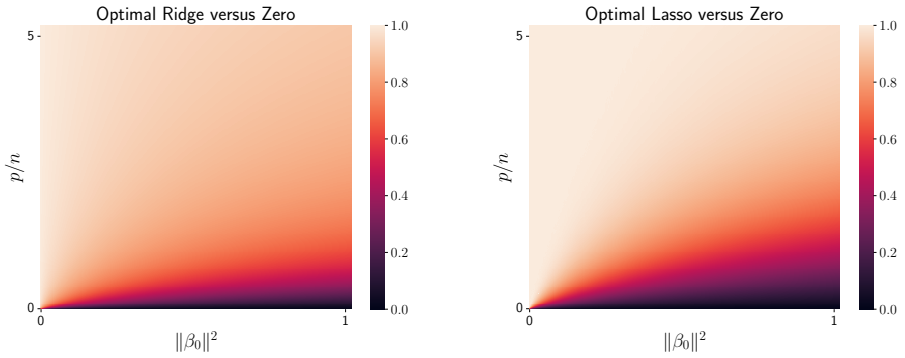
5. The exact form of the limit is given by

$$\lim_{n\to\infty} \mathcal{R}(\hat{\beta}_r(\lambda_n), F) = c_0 m(-\lambda, c_0) + \lambda(\lambda\sigma_\beta^2 - c_0)m'(-\lambda, c_0), \qquad (1.6)$$

where $\lambda = \lim_{n\to\infty} c_0\lambda_n$ and $m(-\lambda, c_0) = \big(-(1 - c_0 + \lambda) + \sqrt{(1 - c_0 + \lambda)^2 + 4c_0\lambda}\big)/2c_0\lambda$.

Thrampoulidis et al. [2018], have established its asymptotic Bayes risk limit.[6]

Figure 1.2: Comparison of Prediction Errors: Optimal Ridge and Lasso vs. the Zero Estimator



Note: The left panel illustrates the ratio of prediction error between the optimal Ridge and the baseline zero estimator. Conversely, the right panel presents a similar comparison for the optimal Lasso estimator against the zero estimator. Both axes, y and x, depict a range of p/n ratios and $\|\beta_0\|^2$ values, corresponding to data generated in accordance with the model described in (1.1), with Σ_1, Σ_2, and Σ_ε in Assumptions 1 and 2 set as identity matrices. We set b_0 as a Dirac-spike and a Gaussian slab with $q = 1/5$. In this context of strong signals, the prediction errors for both optimal Ridge and Lasso are calculated using tuning parameters that are optimally selected to minimize the expected prediction errors' probability limits, as given by (1.6) and (1.7).

In Figure 1.2, we present two heatmaps, with the y and x axes representing various values of p/n and $\|\beta_0\|^2 = \tau\sigma_\beta^2$. The left heatmap illustrates the ratio of Bayes risk between optimal Ridge and the zero estimator, while the right heatmap represents the ratio of optimal Lasso against the zero estimator. For both Ridge and Lasso, their optimal tuning parameters are selected by minimizing the probability limits of their Bayes risk given by (1.6) and (1.7), respectively. A prediction error ratio below 1 suggests that zero is outperformed.

The heatmaps, as anticipated, clearly demonstrate that both optimal Ridge and Lasso

6. The limit in the case of Lasso can be explicitly written as follows: $\lim_{n\to\infty} \mathcal{R}(\hat{\beta}_l(\lambda_n), F) = (\alpha^*)^2$, where

$$\alpha^* = \arg\min_{\alpha\geq 0} \left\{ \inf_{\tau_g>0} \sup_{\substack{\beta\geq 0 \\ \tau_h>0}} \frac{\beta\tau_g}{2} + \frac{1}{c_0} L\left(\alpha, \frac{\tau_g}{\beta}\right) - \frac{\alpha\tau_h}{2} - \frac{\alpha\beta^2}{2\tau_h} + \lambda G\left(\frac{\alpha\beta}{\tau_h}, \frac{\alpha\lambda}{\tau_h}\right) \right\}. \tag{1.7}$$

Here $\lambda = \lim_{n\to\infty} \lambda_n/\sqrt{c_0}$, $L(c,\tau) := \mathbb{E}[e_{x^2}(cZ+\varepsilon,\tau) - \varepsilon^2]$, and $G(c,\tau) := \mathbb{E}[e_{|x|}(cZ+X_b,\tau) - |X_b|]$, with $\mathrm{e}_f(y,\tau) := \min_v (y-v)^2/2\tau + f(v)$, where random variables Z, ε, and X_b follow a standard normal distribution ($Z \sim \mathcal{N}(0,1)$), the distribution of the noise, and the distribution F, respectively.

surpass the performance of the zero estimator. This superiority is particularly pronounced in scenarios involving strong signals and relatively lower dimensions. Notably, the disparity between these estimators becomes less pronounced as the norm of $\|\beta_0\|$ approaches zero and the ratio p/n increases, indicating a shift towards scenarios characterized by weaker signals. The existing result on precise error analysis is primarily built upon the assumption of strong signals, where $\tau = 1$. To discern the performance of various estimators under weak signal conditions, a more intricate analysis in the limiting case ($\tau \to 0$ and $p/n \not\to 0$) is necessary.

1.2.4 Zero's Optimality and Relative Prediction Error

Figure 1.2 also indicates that our attention should be directed towards a regime where the zero estimator exhibits meaningful competitiveness. Otherwise, we may question the appropriateness of our definition of "weak" signals if some machine learning approaches can obviously outperform it by a wide margin.

One might be tempted to define weak signals as instances where the signal strength falls below a certain "detection boundary," thereby becoming indiscernible through hypothesis testing. Our focus is on prediction, rather than signal detection. This distinction is key because, even when signals are undetectable by hypothesis testing, their collective contribution to prediction can still outperform the zero predictor. The zero predictor serves as a natural benchmark for demonstrating the capacity of machine learning to utilize weak signals.

To motivate our concept of weak signals, we analyze a regime where the zero predictor achieves certain notion of optimality, indicated by its Bayes risk being identical to that of the Bayes predictor. This scenario is delineated more precisely by the assumption below:

Assumption 4. $n^{-1}p \to c_0 \in (0, \infty]$, $n^{-1}\tau p(\log p)^4 \to 0$, $n\tau p^{-2/3}(\log p)^{-4} \to \infty$, *and* $n^{-1}pq\tau^{-1}(\log p)^{-4} \to \infty$.

Assumption 4 covers a wide spectrum of signal strengths and counts, while simultaneously imposing constraints to prevent an excessively large p/n ratio and overly rapid vanishing of

τ. The first two constraints imply $\tau \to 0$, while the third imposes a lower bound on τ. Together, these constraints require that τ is bounded below by $n^{-1/3}$.

The final constraint addresses cases of extreme sparsity in β_0. It becomes redundant when q does not vanish, as it is already implied by the first two constraints. Collectively, these conditions imply that $(pq) \log p/n$ is bounded below by τ, and consequently by $n^{-1/3}$ (up to a logarithmic factor). Importantly, these constraints do not exclude the sparsity assumptions commonly adopted in the literature for Lasso: $\|\beta_0\|_0 \log p/n \to 0$, where $\|\beta_0\|_0 \asymp_{\mathrm{P}} pq$.

These constraints serve to exclude edge cases where the relative performance of different estimators cannot be conclusively determined using our proof technique. In Section 1.2.9, we investigate scenarios outside the scope of these constraints (e.g., lower bound of $n^{-1/3}$), such as cases of extreme sparsity with only one true signal in the DGP. By employing an alternative proof method that leverages the closed-form solution of Lasso in a special case, we demonstrate that these constraints are not necessary for arriving at our conclusions.

To facilitate the discussion of the optimal estimator in our context, we refer to the definition provided by Robbins [1964].

Definition 1. *We say $\hat{\beta}$ is asymptotically optimal relative to F, if it satisfies*

$$\lim_{n\to\infty} \frac{\mathcal{R}(\hat{\beta}, F)}{\mathcal{R}(F)} = 1.^{7}$$

Theorem 1. *Suppose that Assumptions 1–4 hold. Furthermore, assume that the error term ε in Assumption 2 follows a Gaussian distribution.*[8] *Under these conditions, the zero estimator is asymptotically optimal relative to any distribution $F \in \mathcal{F}$.*

This theorem demonstrates that, unlike the Bayes predictor, which relies on prior knowl-

7. The definition of asymptotic optimality is provided in terms of a ratio to accommodate more general scenarios where $\mathcal{R}(F)$ varies with the sample size n.

8. The Gaussian assumption on ε is only used to facilitate considerations of optimality, which is a standard assumption in the empirical Bayes literature, e.g., Jiang and Zhang [2009]. While this assumption motivates our characterization of weak signal regimes, it is not utilized in follow-up analysis of the estimators.

edge F and is thus impractical, the zero predictor achieves the same asymptotic optimal prediction risk without requiring such information, making it both feasible and optimal.

The zero estimator can be considered as a special case of both Ridge and Lasso when a sufficiently large tuning parameter is chosen. Given this perspective, and as implied from Theorem 1, the relative Bayes risk of the optimal Ridge and Lasso compared to the zero estimator asymptotically approaches one. This suggests that under weak signal conditions, merely comparing their Bayes risk ratios is insufficient to distinguish among these estimators.

As such, we shift our attention to the relative prediction error between any estimator $\hat{\beta}$ and the zero estimator, defined as follows, in absolute difference rather than their ratio, in the spirit of Bayesian regret. To ensure a meaningful scale in the limit, we multiply the relative error by $pn^{-1}\tau^{-2}$, and adopt the following metric for comparison:

$$\Delta(\hat{\beta}) = pn^{-1}\tau^{-2}(\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2). \tag{1.8}$$

Based on this definition, if $\Delta(\hat{\beta}) > 0$ holds with probability approaching one, it indicates that the estimator $\hat{\beta}$ exhibits inferior prediction performance compared to zero. Conversely, if $\Delta(\hat{\beta}) < 0$ holds with probability approaching one, $\hat{\beta}$ outperforms zero.

Before we proceed to present our main results in the following section, we need to provide technical conditions governing the limiting behavior of Σ_1, Σ_2, and Σ_ε:

Assumption 5. *For matrices Σ_1, Σ_2 and Σ_ε:*

$$\frac{1}{n}\operatorname{Tr}(\Sigma_1) = 1 + O(n^{-1/2}), \quad \frac{1}{p}\operatorname{Tr}(\Sigma_2) = \sigma_x^2 + O(p^{-1/2}), \quad \frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon) = \sigma_\varepsilon^2 + O(n^{-1/2}).$$

Additionally, there exist constants θ_1 to θ_4 such that

$$\frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon\Sigma_1) = \sigma_\varepsilon^2\theta_1 + o(n\tau/p),$$
$$\frac{1}{p}\operatorname{Tr}(\Sigma_2^2) = \sigma_x^4\theta_2 + o(1), \quad \frac{1}{n}\operatorname{Tr}(\Sigma_\varepsilon\Sigma_1^2) = \sigma_\varepsilon^2\theta_3 + o(n/p), \quad \frac{1}{n}\operatorname{Tr}(\Sigma_1^2) = \theta_4 + o(n/p).$$

As Σ_1, Σ_2, and Σ_ε are positive definite, all of these constants θ_i, $i = 1, 2, 3$, and 4, are positive. The condition concerning Σ_2 can be verified through a more primitive condition: the existence of the limit of Σ_2's empirical spectral distribution (see Dobriban and Wager [2018]). In situations where all three matrices reduce to identity matrices, a common setting in the literature on precise error analysis, Assumption 5 holds trivially.

1.2.5 Analysis of the Ridge Predictor

In this section, we present the results of Ridge in the context of weak signals. We begin by presenting the relative error of Ridge for any tuning parameter value:

Theorem 2. *Assuming that Assumptions 1−5 hold, and setting $\lambda_n = \tau^{-1}\lambda$, we establish:*

$$\Delta(\hat{\beta}_r(\lambda_n)) \xrightarrow{\mathrm{P}} \alpha^* := 2\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2\theta_1}{2\lambda^2} - \frac{\sigma_\beta^2}{\lambda}\right).$$

This theorem yields several important findings. First, by minimizing α^* with respect to λ, we can determine the optimal tuning parameter value: $\lambda_n^{opt} = \tau^{-1}\sigma_\varepsilon^2\theta_1/\sigma_\beta^2$. Furthermore, with the optimal tuning parameter in place, α^* is negative, indicating that Ridge can effectively learn weak signals when the tuning parameter is chosen appropriately.

Second, when we set $\lambda \to \infty$ (equivalently, $\tau\lambda_n \to \infty$), the value of α^* converges to zero. This outcome is expected, as the use of a large tuning parameter makes the predictor's performance increasingly resemble that of the zero predictor. Nevertheless, it is noteworthy that $\alpha^* \to 0^-$; in other words, as λ increases, Ridge consistently outperforms the zero

predictor until it gradually becomes indistinguishable from it in the limit.

Third, as $\lambda \to 0$, in which case $\lambda_n = o(\tau^{-1})$, α^* approaches positive infinity. This indicates that Ridge's performance deteriorates to the point where the Ridgeless predictor (corresponding to $\lambda = 0$) is surpassed by the zero predictor. This is a significant departure from the strong signal setup in which Ridgeless can still outperform the zero predictor, as shown by Hastie et al. [2022]. Notably, Ridgeless, defined by $\lambda = 0$, is not devoid of regularization; it applies implicit regularization by minimizing the ℓ_2-norm of the interpolator. This form of regularization allows Ridgeless to effectively control variance when the number of predictors p exceeds the sample size n, ensuring desirable performance in strong signal scenarios. However, under weak signals this implicit regularization fails to adequately control variance, causing $\|\hat{\beta}\|$ to be much larger than $\|\beta_0\|$, leading to poor performance of $\hat{\beta}$.

Furthermore, given that Ridgeless is the interpolator that minimizes $\|\hat{\beta}\|$, all linear interpolators, including, for instance, the one that minimizes the ℓ_1-norm, result in even larger $\|\hat{\beta}\|$. Therefore, these interpolators also fail to outperform zero in contexts with weak signals.

Figure 1.3 provides an illustrative example of the relationship between the relative error of Ridge and the tuning parameter λ, showcasing the theoretical insights we have discussed. Corollary 1 below summarizes the result on the Ridgeless estimator:

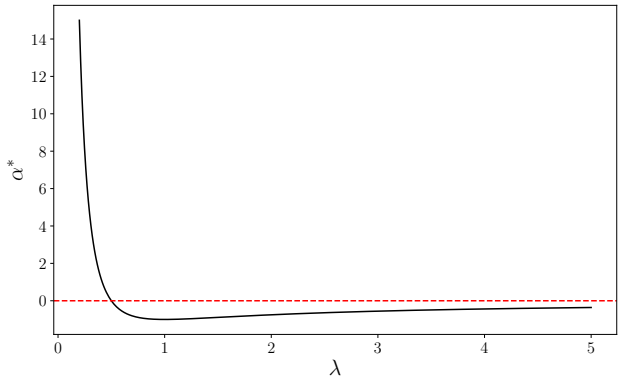
Corollary 1. *Under the same assumptions as in Theorem 2, the Ridgeless estimator, defined by* (1.4)*, satisfies:*

$$\Delta(\hat{\beta}_r(0)) \xrightarrow{\mathrm{P}} \infty.$$

Given Ridge's ability to effectively learn weak signals with an appropriately tuned parameter, the data-dependent selection of this parameter becomes crucial. A paradigmatic approach for this purpose is K-fold CV.

In cases where the signals are strong, Hastie et al. [2022] demonstrate the effectiveness of CV for Ridge. Specifically, the cross-validated tuning parameter converges in probability to the optimal value within a pre-specified interval. The fact that this optimal value lies in

Figure 1.3: Ridge vs. Zero Predictor's Relative Precise Error



Note: In this plot, the black curve represents the probability limit of $\Delta(\hat{\beta}_r(\lambda_n))$, denoted as α^*, as a function of the tuning parameter λ, defined in Theorem 2, in the context of weak signals. To create this plot, we set all parameters $(c_0, \theta_1, \theta_2, \sigma_x, \sigma_\beta, \sigma_\varepsilon)$ to one for simplicity.

some known interval simplifies the derivation of the theoretical properties of CV. In scenarios with weak signals, however, the optimal tuning parameter tends to diverge as the sample size increases. The rate of divergence depends on the unknown strength of the weak signal, τ. As we show next, CV remains a valid and useful tool in this case. To narrow our focus to the matter of weak signals without delving into a complicated CV procedure, we consider the case where both Σ_ε and Σ_1 are identity matrices. This assumption of no temporal dependence in the data facilitates a more straightforward CV procedure for i.i.d. data.

To determine the optimal tuning parameter using K-fold CV, denoted as $\hat{\lambda}^{K-CV}$, the rows of the design matrix X are partitioned into K distinct subsets, labeled as $X_{(1)}, \cdots, X_{(K)}$. For each $i \in \{1, \cdots, K\}$, the submatrix $X_{(-i)}$ is formed by excluding the rows corresponding to $X_{(i)}$. Similarly, the associated subvectors are $y_{(i)}$, $\varepsilon_{(i)}$, $y_{(-i)}$, and $\varepsilon_{(-i)}$. The Ridge estimator for each fold, $\hat{\beta}_r^i(\lambda_n)$, is defined as the solution to the optimization problem for $i = 1, \cdots, K$:

$$\hat{\beta}_r^i(\lambda_n) = \arg\min_{\beta} \left\{ \frac{1}{n} \|y_{(-i)} - X_{(-i)}\beta\|^2 + \frac{p\lambda_n}{n} \|\beta\|^2 \right\}.$$

Consequently, the tuning parameter selected by K-fold CV is given by

$$\hat{\lambda}_n^{K-CV} = \underset{\lambda_n \in [\epsilon, \infty)}{\arg\min} \frac{1}{n} \sum_{i=1}^{K} \|y_{(i)} - X_{(i)} \hat{\beta}_r^i(\lambda_n)\|^2,$$

where $\epsilon > 0$ is an arbitrary small constant. The following theorem provides its justification:

Theorem 3. *Under the same assumptions as in Theorem 2, if we also assume that* $\Sigma_1 = \mathbb{I}$, $\Sigma_\varepsilon = \sigma_\varepsilon^2 \mathbb{I}$, ε *follows a sub-exponential distribution, and that* $q^{-1}\tau^{-1}n^{-1/2}\log(p) \to 0$ *and* $q^{-1/2}\tau^{-3/2}n^{-1/2}\log(p) \to 0$, *then we can establish that:*

$$\tau\hat{\lambda}_n^{K-CV} \xrightarrow{\mathrm{P}} \lambda^{opt} = \sigma_\varepsilon^2/\sigma_\beta^2 \quad \textit{and} \quad \Delta(\hat{\beta}_r(\hat{\lambda}_n^{K-CV})) - \Delta(\hat{\beta}_r(\lambda_n^{opt})) \xrightarrow{\mathrm{P}} 0.$$

This theorem justifies the use of $\hat{\lambda}_n^{K-CV}$ as an approximation for the optimal tuning parameter $\lambda_n^{opt} = \lambda^{opt}/\tau$ ($\theta_1 = 1$ in this case) for Ridge. Importantly, this result does not require prior knowledge of τ, making the CV approach directly applicable in practical scenarios. The additional constraints on q become relevant only when q vanishes; otherwise, they naturally follow from Assumption 4. These conditions ensure uniform convergence across the spectrum of tuning parameter values, a prerequisite for the results of Theorem 3.

With our analysis of Ridge concluded, we will now turn our attention to Lasso.

1.2.6 Analysis of the Lasso Predictor

Unlike Ridge, the analysis of Lasso is more intricate, primarily because it lacks a closed-form formula. In the special case where Σ_1, Σ_2, and Σ_ε are identity matrices, several studies, including Bayati and Montanari [2012] and Thrampoulidis et al. [2018], have established Lasso's precise error given by (1.7). Additionally, based on (1.7), Wang et al. [2020] conducted a small-signal Taylor expansion of α^* with respect to σ_β^2, which affects α^* through the prior distribution F. They concluded that the optimal Lasso estimator fails to outperform

optimal Ridge.[9] In the general case we consider, pinpointing the exact precise error appears a daunting task. Instead, we seek probability bounds of the limit. This turns out sufficient for us to conclude that Lasso cannot outperform zero for all values of its tuning parameter in the context of weak signals. The next theorem summarizes our main findings:

Theorem 4. *Assume that Assumptions 1−5 are satisfied and the tuning parameter λ_n is chosen such that the following equation holds for some $C_\lambda > 0$:*

$$pn^{-2}\tau^{-2}\mathbb{E}_{U\sim\mathcal{N}(0,\Sigma_2)}\left\|(2\sigma_\varepsilon\sqrt{\theta_1}|U| - \lambda_n)_+\right\|^2 = C_\lambda.^{10} \tag{1.9}$$

Then, with probability approaching one, we have $c_\alpha \leq \Delta(\hat{\beta}_l(\lambda_n)) \leq C_\alpha$, where c_α and C_α are the solutions to the following equation in terms of x:

$$x - \sqrt{\frac{2C_\lambda}{c_2}x} = -\frac{C_\lambda}{100C_2}, \tag{1.10}$$

where c_2 and C_2 are constants defined in Assumption 1.

Equation (1.9) implicitly determines the rate at which λ_n diverges to infinity. For any fixed $C_\lambda > 0$, we can solve for the tuning parameter λ_n from (1.9), and derive the upper and lower bounds, C_α and c_α, from equation (1.10). Furthermore, equation (1.10) directly implies that C_α and c_α are non-negative, indicating that Lasso does not outperform the zero predictor for any given tuning parameter value in the context of weak signals.

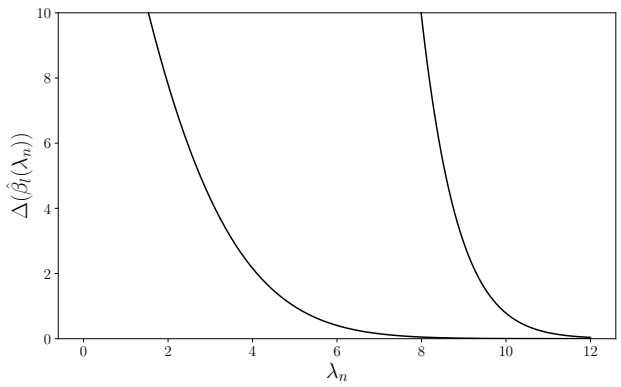
Moreover, as λ_n approaches zero, C_λ diverges to infinity, leading to a simultaneous divergence of both c_α and C_α. This suggests that Lasso behaves increasingly worse compared to the zero predictor. Conversely, a larger tuning parameter λ_n causes C_λ to converge to zero from the positive side. As a result, both c_α and C_α converge to zero while remaining

9. Their analysis does not address the scenario of Lasso with an arbitrary tuning parameter, nor does it elucidate its relative performance compared to zero.

10. When applied to a vector, $|\cdot|$ and $(\cdot)_+$ represent element-wise operations.

non-negative. This implies that Lasso's performance improves but remains inferior to the zero predictor, until they become equivalent in the limit. Figure 1.4 illustrates upper and lower bounds for the relative error of Lasso versus the zero predictor across different λ_n values.

Figure 1.4: Lasso vs. Zero Predictor: Relative Precise Error Bounds



Note: In this plot, the black curves represent the lower and upper probability bounds on $\Delta(\hat{\beta}_l(\lambda_n))$, i.e., c_α and C_α, as a function of the tuning parameter λ_n in the context of weak signals. In this setup, we fix $n = p = 2,000$. We assign the elements of b_0 to follow a standard Gaussian distribution and set $\Sigma_2 = \mathbb{I}$. As in Figure 1.3, we set all parameters $(c_0, \theta_1, \theta_2, \sigma_x, \sigma_\beta, \sigma_\varepsilon)$ to one. Finally, we select $\tau = 0.001$, which results in a population R^2 around 0.1%.

Lasso struggles with weak signals because it has difficulty distinguishing true signals from spurious ones. Its inability to select true weak signals does not significantly impact its performance compared to the zero predictor, which ignores these signals entirely. The main issue with Lasso is its ineffectiveness in penalizing irrelevant signals. While increasing the tuning parameter might help, our theory indicates that Lasso only reaches the zero predictor's effectiveness when the penalty is large enough to essentially disregard all signals.

Given Lasso's limitations with weak signals, the elastic net—merging the ℓ_1 and ℓ_2 norms in its penalty—also underperforms compared to Ridge, in settings involving weak signals.[11]

In predictive regressions with many covariates, Lasso is often viewed as the successor to

11. Although a formal justification for this observation can be provided in the setting of $\Sigma_2 = \mathbb{I}$, we omit it here due to space constraints.

OLS. However, our analysis reveals a critical caveat: Lasso, regardless of tuning parameter choice, underperforms a zero predictor in weak-signal settings. This has important implications for economics and finance, where large-scale regressions with low signal-to-noise ratios are increasingly common. Instead, our results favor Ridge in such scenarios, emphasizing the need to assess signal strength when selecting regularization techniques.

1.2.7 Assessing Signal-to-Noise Ratio

In line with this perspective, we delve into the assessment of the signal-to-noise ratio, a measure that can offer insights into the viability of different machine learning techniques. Our preceding analyses provide indirect guidance in this regard. Specifically, if Lasso underperforms the zero predictor, it implies a potential issue with the data's signal strength.

A more conventional and direct approach to evaluating the signal-to-noise ratio is through R^2. However, in-sample R^2 is prone to overfitting, and as such, out-of-sample R^2 is commonly used in machine learning. This metric essentially involves the comparison of mean-squared errors between two predictors. For our specific application, we have chosen to use zero as the reference predictor and define this metric for a given estimator $\hat{\beta}$ as follows:

$$R^2_{\text{oos}}(\hat{\beta}) = 1 - \frac{\sum_{i \in \text{OOS}} (y_i - X_i \hat{\beta})^2}{\sum_{i \in \text{OOS}} y_i^2}, \tag{1.11}$$

where "OOS" represents the out-of-sample data.

Since a model's predictive performance hinges on the signal-to-noise ratio, this metric is a natural choice for evaluating the signal strength inherent in the DGP. In strong-signals scenarios, the out-of-sample R^2 consistently estimates the population R^2, irrespective of the specific estimator $\hat{\beta}$, provided it is consistent with respect to β_0, i.e., $\|\hat{\beta} - \beta_0\| = o_{\text{P}}(1)$. However, in weak-signal settings, the outcome critically depends on the estimator chosen, beyond the signal-to-noise ratio itself. As an illustration, while both Lasso and Ridge are

consistent as their prediction errors asymptotically diminish, the out-of-sample R^2 for Lasso can become non-positive, indicating either no improvement or underperformance relative to the zero predictor, as demonstrated in Theorem 4.

In the context of weak signals, the following proposition provides theoretical support for the relevance of optimal Ridge's R^2_{oos} in assessing the signal-to-noise ratio in the data.

Proposition 1. *Under the same assumptions as Theorem 3, and assuming that the out-of-sample data follows the same DGP as the in-sample data, if $n_{oos}p^{-2}n^2\tau^2 \to \infty$, where n_{oos} is the size of the out-of-sample data, then for the optimal Ridge predictor, it holds that*

$$R^2_{oos}(\hat{\beta}_r(\lambda_n^{opt})) = p^{-1}n\theta_2(R^2)^2\left(1 + o_{\text{P}}(1)\right),$$

where R^2 denotes the population R-squared, given by $\tau\sigma_x^2\sigma_\beta^2/(\tau\sigma_x^2\sigma_\beta^2 + \sigma_\varepsilon^2)$ in this context.

Interestingly, when n_{oos} is sufficiently large, to the extent that the estimation error in R^2_{oos} does not mask the performance differential between the optimal Ridge and the zero predictor, R^2_{oos} is approximately proportional to $(R^2)^2$. While it does not exactly mirror R^2, R^2_{oos} still serves as an indicator of signal strength. The reason for their discrepancy is that in the weak signal case, the numerator of the R^2_{oos}—which reflects the relative prediction error between the two predictors—decreases more rapidly than the numerator of R^2. Therefore, the numerator of R^2_{oos} only provides a higher-order characterization of signal strength.

1.2.8 Mixed Signal Strengths and Alternative Benchmarks

In the preceding sections, our analysis primarily focuses on scenarios where all signals are weak, leading us to consider the zero predictor as our natural benchmark. This section, however, expands our analysis to include models where potentially strong signals serve as

benchmarks. Consider another DGP:

$$y = W\gamma_0 + X\beta_0 + \varepsilon, \tag{1.12}$$

where $W \in \mathbb{R}^{n\times d}$ represents a predefined set of covariates. These covariates include potentially strong signals and form the basis of the benchmark model. We allow the dimension d to increase to ∞, however, it does so at a slower rate compared to n, ensuring that OLS of y against W remains a viable method for estimation.

In many cases, W could simply be a vector of ones, allowing us to remain agnostic about the magnitude of the regression's intercept. In our empirical analysis, W can be motivated from economic theory, whose impact on the response variable is of central interest. Alternatively, W can encompass lagged values of y, thereby facilitating the inclusion of temporal dependence in the benchmark model. This setup is particular relevant when using an autoregressive model as a benchmark for forecasting economic variables. Exploring the possibility of a data-driven selection of W is an intriguing direction for future research.

Building on these considerations, our focus now shifts to evaluating and comparing the performance against the OLS benchmark with covariates in W. In this context, the OLS benchmark predictor, $\hat{y}_b^{new}$, for a new observation (w^{new}, x^{new}) is defined as follows:

$$\hat{y}_b^{new} = (w^{new})^\top \hat{\gamma}, \quad \text{where} \quad \hat{\gamma} = (W^\top W)^{-1} W^\top y. \tag{1.13}$$

The inclusion of W leads us to explore the following Ridge estimator with regularization only imposed on coefficients of X:

$$\begin{aligned}\hat{\beta}(\lambda_n) &:= \arg\min_{\beta} \left\{ \min_{\gamma} \left(\frac{1}{n}\|y - W\gamma - X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2 \right) \right\} \\ &= \arg\min_{\beta} \left\{ \frac{1}{n}\|\mathcal{M}_W y - \mathcal{M}_W X\beta\|^2 + \frac{p\lambda_n}{n}\|\beta\|^2 \right\}, \end{aligned} \tag{1.14}$$

where $\mathcal{M}_W = \mathbb{I} - W(W^\top W)^{-1}W^\top$. Consequently, the estimator for γ is thus given by

$$\hat{\gamma}(\lambda_n) = (W^\top W)^{-1}W^\top(y - X\hat{\beta}(\lambda_n)). \tag{1.15}$$

The construction for Lasso is similar. Utilizing the estimated parameters $(\hat{\beta}(\lambda_n), \hat{\gamma}(\lambda_n))$ we are able to formulate a predictor for y as

$$\hat{y}^{new} = (w^{new})^\top\hat{\gamma}(\lambda_n) + (x^{new})^\top\hat{\beta}(\lambda_n) = \hat{y}_b^{new} + (\hat{x}^{new})^\top\hat{\beta}(\lambda_n), \tag{1.16}$$

where $\hat{x}^{new} = x^{new} - X^\top W(W^\top W)^{-1}w^{new}$.

Notably, equation (1.16) illuminates the role of the second term, $(\hat{x}^{new})^\top\hat{\beta}(\lambda_n)$, in capturing the contribution of weak signals relative to the OLS benchmark, $\hat{y}_b^{new}$. Moreover, a comparison with the previously analyzed zero-benchmark scenario reveals a key distinction: the modification of the response and covariates in equation (1.14). Here, the approach involves regressing $\mathcal{M}_W y$ on $\mathcal{M}_W X$, effectively predicting the residuals of the benchmark model using covariates adjusted to eliminate dependence on W. While our earlier conclusions are likely still valid, the inclusion of generated covariates introduces an additional layer of statistical error that warrants careful examination. The forthcoming theorem will demonstrate that this extra error does not compromise our prior conclusions. Given this context and our prior comparative analysis, we focus on the optimal Ridge in this scenario for reason of space.

Theorem 5. *Suppose that $X = W\eta_0 + U$. Assume that the triplet $(U, \beta_0, \varepsilon)$ follows the same distribution as $(X, \beta_0, \varepsilon)$ in Theorem 2. Additionally, the matrix W is independent from U, β_0, and ε. Each covariate within W is assumed to have a finite second moment. Furthermore, we assume that $d = o(n^2p^{-1}\tau)$, and the eigenvalues of $n^{-1}W^\top W$ are lower bounded by some positive constant. Given these assumptions, the predictor, $\hat{y}^{new}$, as defined in* (1.16) *and based on the Ridge estimator from* (1.14) *with $\lambda_n = \tau^{-1}\lambda$, and the benchmark*

predictor $\hat{y}_b^{new}$ *from* (1.13)*, satisfy the following:*

$$pn^{-1}\tau^2\Big(\mathbb{E}_F\left[(\hat{y}^{new}-y^{new})^2 \mid \mathcal{I}\right]-\mathbb{E}_F\left[(\hat{y}_b^{new}-y^{new})^2 \mid \mathcal{I}\right]\Big)\xrightarrow{\mathrm{P}}\alpha^*, \tag{1.17}$$

where $\mathcal{I}$ *denotes the information set generated by* $(W, X, y, \gamma_0, \beta_0)$*,* α^* *is defined in Theorem 2, and the tuple* $(y^{new}, w^{new}, x^{new})$ *satisfies* (1.12)*.*

This result shows that a Ridge-augmented model outperforms the benchmark model alone, aligning with our earlier conclusion that Ridge can effectively leverage weak signals.

1.2.9 Ridge vs. Lasso in Extremely Sparse Settings

This section clarifies that our asymptotic restrictions are mainly technical. Even in extreme sparsity beyond our main analysis, Lasso fails to outperform zero when signals are sufficiently weak, while Ridge retains a non-negligible probability of learning them. However, the setting is stylized, as the proof relies on a distinct approach using Lasso's closed-form solution.

Consider the Gaussian sequence model where $y = \beta_0 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \mathbb{I})$. For this model, $X = \mathbb{I}$ and $n = p$. We let $\beta_0 = (\sqrt{n\tau}, 0, \ldots, 0)^\top$, so that $s = \|\beta_0\|_0 = 1$ and $R^2 = \frac{\|X\beta_0\|^2}{\|X\beta_0+\varepsilon\|^2} \asymp \tau \to 0$. This represents the most extreme form of sparsity, with only one true signal. Moreover, we relax the $n^{-1/3}$ lower bound restriction on τ.

Proposition 2. *Assume* $\tau \leq n^{-1}\log(n)/100$*. There exists* $n_0 > 0$ *such that* $n > n_0$*, Lasso satisfies*

$$\mathrm{P}\left(\|\hat{\beta}_l(\lambda)-\beta_0\|^2-\|\beta_0\|^2 \geq 0, \forall\lambda \geq 0\right) \geq 1-n^{-1/2}, \tag{1.18}$$

while for Ridge,

$$\mathrm{P}\left(\exists\lambda > 1 \ \textit{s.t.} \ \|\hat{\beta}_r(\lambda)-\beta_0\|^2-\|\beta_0\|^2 < 0\right) \geq 0.5. \tag{1.19}$$

This result highlights that in extreme sparsity, Ridge can still learn weak signals, whereas Lasso remains ineffective regardless of tuning.

1.3 Monte Carlo Simulations

In this section, we conduct simulation experiments to assess the finite sample relevance of our asymptotic theory. We begin by examining Ridge and Lasso in a linear model setup.

1.3.1 Ridge and Lasso for Linear Models

We now detail the DGP specified by (1.1) for the first simulation exercise. We set $(\Sigma_1)_{ij} = 2^{-|i-j|}$ for $1 \leq i, j \leq n$. We construct Σ_ε as a diagonal matrix with i.i.d. entries sampled from the uniform distribution $U(0.5, 1.5)$. The eigenvalues of Σ_2 are also simulated from $U(0.5, 1.5)$, with corresponding eigenvectors from a randomly generated orthogonal matrix. These components form Σ_2, which, along with other matrices, remains fixed throughout the simulations. By direct calculations, we have $\theta_1 = 1$, $\theta_2 = 13/12$, and $\theta_3 = \theta_4 = 5/3$.

We experiment with $n = 500$ and $p = 300$, dimensions that align closely with the first microeconomic example studied below. For each simulated sample, we construct β_0 as $\sqrt{p^{-1}\tau}b_0$, with b_0 drawn from a spike-and-slab distribution: $(1-q)\delta_0 + q\mathcal{N}(0, q^{-1}\sigma_\beta^2)$. Here, δ_0 represents the Dirac delta function, and we set $\sigma_\beta^2 = 1$. The error term ε is sampled from $\mathcal{N}(0,1)$, while the design matrix X is drawn from $\mathcal{N}(0,1)$, then transformed via pre-multiplication by $\Sigma_1^{1/2}$ and post-multiplication by $\Sigma_2^{1/2}$. We consider two cases for the sparsity parameter, $q = 0.2$ and $q = 0.8$. To represent weak and strong signal scenarios, we calibrate two values of τ to achieve $R^2 = 5\%$ and 50%, respectively. The parameters q and τ (through R^2) are varied as they are critical to the asymptotic performance, as highlighted in Assumption 4. A total of $1,000$ Monte Carlo repetitions are conducted.

For each simulated sample, we compute the relative prediction error, $\Delta(\hat{\beta}(\hat{\lambda}_n^{K-CV}))$, as defined in equation (1.8), with the tuning parameter for each method selected via 10-fold

CV. The histograms of these errors are presented in Figure 1.5. Additionally, Table 1.1 provides summary statistics, including quantiles and the proportion of values classified as "zeros" (where "zero" is defined as being within machine precision).

The histograms reveal notable differences in the performance of Ridge and Lasso when R^2 is low, while showing that both methods outperform the zero predictor when R^2 is high. Ridge displays a clear probability mass on the negative side of the error distribution, even in weak signal settings, and its performance is largely unaffected by changes in sparsity levels.

In contrast, Lasso struggles to capture weak signals, as evidenced by a substantial probability mass on the positive side of the y-axis. While increasing sparsity (lowering q) leads to modest improvement in Lasso's performance, it still lags behind Ridge overall. This pattern is supported by Table 1.1, which presents quantiles of these distributions. In finite samples with high sparsity, some probability mass for Lasso can fall on the negative side. Nevertheless, Lasso also shows a heavier probability mass at zero compared to Ridge, consistent with the theoretical prediction that optimal Lasso collapses to zero in weak-signal settings.[12]

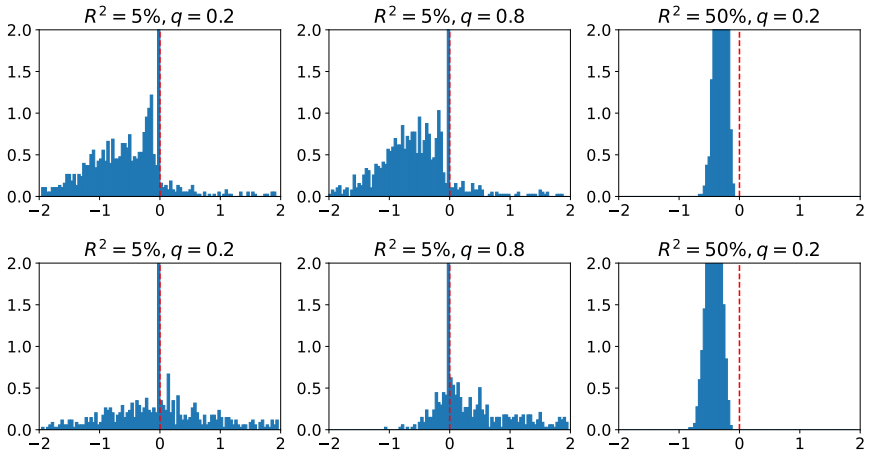
Table 1.1: Summary Statistics for Ridge and Lasso in Linear DGPs

		Lasso				Ridge			
q	R^2 (%)	Q1	Q2	Q3	#Zero	Q1	Q2	Q3	#Zero
0.2	5%	-0.127	0.000	0.521	360	-0.992	-0.501	-0.129	97
0.8	5%	0.000	0.000	0.975	400	-0.918	-0.541	-0.169	88
0.2	50%	-0.508	-0.414	-0.331	0	-0.375	-0.300	-0.233	0

Note: The table illustrate the summary statistics (quantiles and the percentage of zeros) of relative prediction error $\Delta(\hat{\beta}(\hat{\lambda}_n^{K-CV}))$, for Ridge and Lasso, based on 1,000 Monte Carlo samples.

12. For reason of space, we omit further experiments with extreme sparsity ($q = 0.1, 0.05, 0.02$), where Lasso's performance further improves, yet reducing the signal strength to $R^2 = 2\%$ negates this advantage, rendering it ineffective against zero. Ridge continues to outperform both the zero predictor and Lasso. We also omit simulation results with fixed tuning parameters that validate our theoretical findings independently of tuning parameter selection.

Figure 1.5: Simulation Results for Ridge and Lasso in Linear DGPs



Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_r(\hat{\lambda}_n^{K-CV}))$ (top) and $\Delta(\hat{\beta}_l(\hat{\lambda}_n^{K-CV}))$ (bottom) following equation (1.8) across $1,000$ Monte Carlo samples. We analyze three setups of (R^2, q).

1.3.2 Advanced Machine Learning Methods for Nonlinear Models

In this section, we extend our investigation to nonlinear machine learning methodologies, including RF, GBRT, and NNs, through simulation experiments. While providing a precise theoretical analysis of errors for these algorithms remains challenging—and this part therefore involves some degree of speculation—we draw on insights from linear models to interpret and contextualize our simulation findings.

We simulate the following DGP, expressed explicitly in element-wise form:

$$y_i = \sum_{j=1}^{p} \beta_{0,j} f(Z_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.20}$$

where y_i denotes the ith observation of the response variable, $\beta_{0,j}$ represents the coefficient associated with a function $f(\cdot)$ of the predictor variable Z_{ij}. We adopt the following procedure for simulating this model: Z_{ij} is generated by applying an inverse transform to X_{ij}, previously simulated in Section 1.3.1. Specifically, $Z_{ij} = f^{-1}(X_{ij})$, where X is constructed

using the same DGP as before.[13] Additionally, both the coefficients β_0 and the error term ε_i are drawn from the same baseline DGP. This approach guarantees the replication of the exact simulation results observed when regressing y on X. However, the focus now shifts to predicting y using nonlinear models of Z without prior knowledge of $f(\cdot)$. The effective signal-to-noise ratio decreases due to the added complexity of learning an unknown function.

1.3.2.1 Simulations with Tree Algorithms

Tree algorithms are essential in machine learning for handling complex DGPs with discrete variables, nonlinearities, and intricate interactions. However, single tree models often underperform, prompting the use of ensemble methods to enhance predictions.

Two popular ensemble techniques are RF and GBRT. RF uses bagging, where multiple trees are trained independently on bootstrap samples, and their predictions are averaged to improve performance. In contrast, GBRT employs boosting, iteratively fitting residuals from prior trees to build a strong ensemble from weak learners.

Since trees are invariant to monotonic transformations, it suffices to report their prediction results for the linear DGP, as these are identical to those for the nonlinear DGP under consideration. However, tree methods may underperform Ridge and Lasso, partly due to an additional approximation error from using piecewise constant functions to approximate the linear DGP. Therefore, the primary focus here should not be on comparing tree methods with linear models but rather on assessing the effectiveness of different ensemble techniques in capturing weak signals and comparing their performance to the zero predictor.

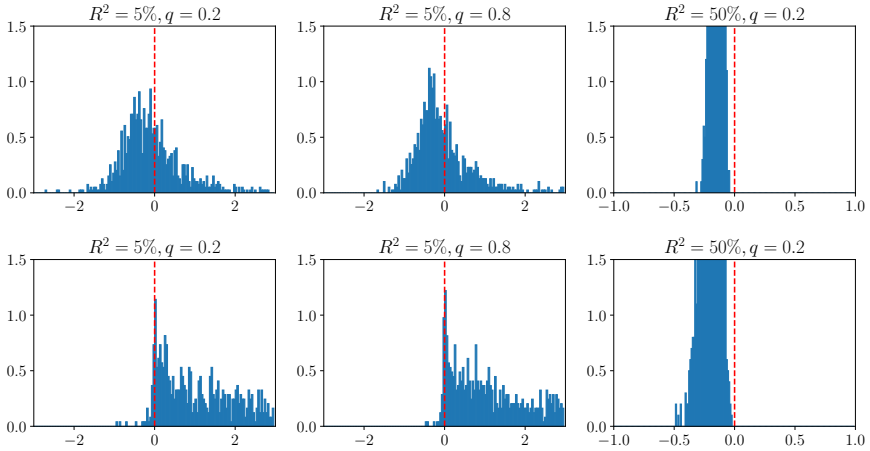
We repeat the experiments from Section 1.3.1, this time generating an additional set of n_{oos} out-of-sample observations to evaluate predictive performance.[14] As illustrated in Figure 1.6, RF demonstrates its ability to learn weak signals when $R^2 = 5\%$, with more than

13. We present results for $f(x) = \tan(x)$. The results are nearly identical to those obtained with cubic or hyperbolic sine functions, and are therefore omitted for brevity.

14. We set $n_{oos} \approx [\tau^{-3}]$ to meet the assumption in Proposition 1.

half of the probability mass located to the left of the y-axis. In contrast, GBRT struggles at this signal strength level. Sparsity does not appear to significantly impact either method. Nonetheless, both methods markedly outperform the zero predictor as R^2 increases to 50%.

Figure 1.6: Simulation Results for RF and GBRT in Linear DGPs



Note: The histograms depict the relative prediction error, $pn^{-1}\tau^{-2}n_{oos}^{-1}\sum_{i\in\text{OOS}}((y_i-\hat{y}_i)^2-y_i^2)$, from 1,000 Monte Carlo samples. We consider RF and GBRT in settings of $n = 500$, $q = 300$, $n_{oos} = 10,000$, across three (R^2, q) configurations. The red dashed line marks the y axis for reference.

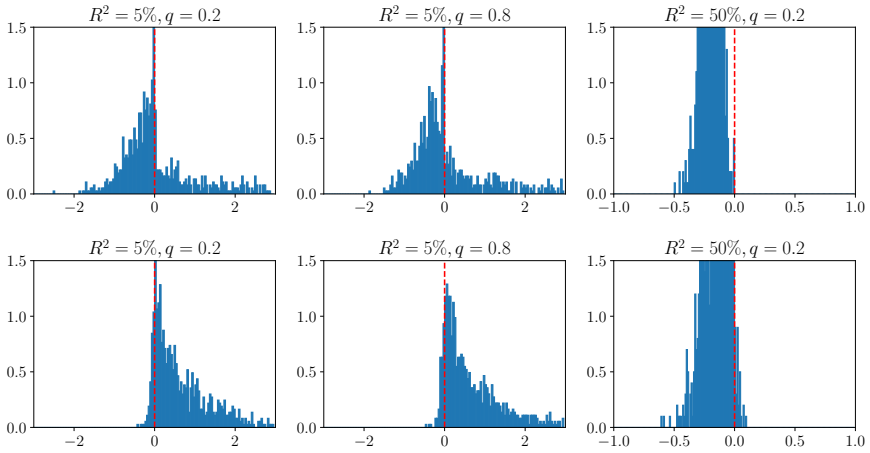
A possible explanation for GBRT's performance pattern could be an inherent ℓ_1-like regularization in its boosting approach. This conjecture draws support from Efron et al. [2004], which demonstrates a parallel between boosting and the Lasso path in linear regressions. To substantiate this conjecture, we examine the number of active variables (those with a non-zero importance score) from both tree methods under the benchmark case ($q = 0.2$, $R^2 = 5\%$). For RF, the average count of active variables is 300. This mirrors that of Ridge. Conversely, GBRT demonstrates a significantly lower count of active variables, with an average of 27.9, aligning more with the variable selection feature of Lasso. Based on our theoretical findings that Ridge outperforms Lasso in weak signal scenarios, we can infer that RF is more adept than GBRT in settings characterized by low signal strengths.

1.3.2.2 Simulations with Neural Networks

Next, we study fully-connected feed-forward NNs. Still, we fix $n = 500$ and $p = 300$. Such parameters lead to an input layer in the NN configured with 300 neurons. Our chosen architecture features a single hidden layer, which includes 16 neurons.

The training process of these NNs often incorporates a sophisticated mix of optimization and regularization techniques, such as dropout and batch normalization, crucial for enhancing performance. To isolate and assess the impact of ℓ_1 and ℓ_2 regularization, we use plain stochastic gradient descent (SGD), deliberately avoiding other optimization techniques to minimize interference, albeit at the expense of not fully exploiting the NN's potential.[15]

Figure 1.7: Simulation Results for NNs in Nonlinear DGPs



Note: The histograms show the relative prediction error, $pn^{-1}\tau^{-2}n_{oos}^{-1}\sum_{i\in\text{OOS}}((y_i - \hat{y}_i)^2 - y_i^2)$, across $1,000$ Monte Carlo samples, using ℓ_2 (top) and ℓ_1 (bottom) penalties. The settings include $n = 500$, $q = 300$, $n_{oos} = 10,000$ for three (R^2, q) configurations. The red dashed line marks the y axis for reference.

The histograms in Figure 1.7 display the relative prediction errors of NNs. We observe that ℓ_2 regularization effectively leverages weak signals, as indicated by most of the probability mass in their histograms being on the negative side of the y-axis when $R^2 = 5\%$. In contrast, ℓ_1 regularization shows a notable decline in performance, which is consistent with

15. We avoid early stopping, as our simulations (not reported due to space constraints) reveal it functions similarly to ℓ_2-regularization by shrinking parameter values towards their initial, smaller magnitudes.

our theoretical insights from linear models.

1.4 Empirical Analysis of Six Economic Datasets

In this section, we demonstrate the practical relevance of our theoretical insights by applying seven machine learning methods—Ridge, Lasso, OLS/Ridgeless, RF, GBRT, NNs with both ℓ_1 and ℓ_2 penalties—across six datasets from finance, macroeconomics, and microeconomics, with two datasets per field. We use five datasets similar to those in Giannone et al. [2022], updated with the latest data when feasible, and introduce an updated dataset from Gu et al. [2020] for our second finance example. Our empirical strategy differs notably from Giannone et al. [2022], who estimate a parametric model using a Spike-and-Slab prior within a Bayesian framework. In contrast, our study, more aligned with Gu et al. [2020], focuses on a comparative analysis of these methods.

At the outset of each empirical exercise, we face a variety of decisions regarding our implementation strategy. These include setting the in-sample and out-of-sample periods, choosing between a rolling window or an expanding window approach, selecting a CV procedure, and deciding on covariate normalization.[16] We follow the frameworks established by Giannone et al. [2022] and Gu et al. [2020] to limit degrees of freedom, thereby improving the robustness, comparability, and reproducibility of our findings. In applying each machine learning method, choosing the right grid for tuning parameters is crucial, balancing performance optimization with computational efficiency. Finer and wider grids can potentially improve performance but also increase computational demands. Details on our model configuration and tuning parameter selection are provided in Appendix 1.8.

Below we summarize the empirical findings from six distinct datasets, each analyzed and reported separately. The primary summary statistics, R^2_{oos}s, are collected in Table 1.2.

16. Normalizing covariates is essential before using machine learning methods, as it standardizes their scales, aids in regularization, and improves the convergence of optimization algorithms. To prevent forward-looking bias, normalization is performed using each covariate's in-sample mean and standard deviation.

1.4.1 Finance 1: Market Equity Premium

In the first analysis, we focus on predicting market equity returns using a dataset of financial and macroeconomic indicators compiled by Welch and Goyal [2007].[17] This dataset comprises 16 predictors and includes 74 annual observations, covering a period from 1948 to 2021. Despite Welch and Goyal [2007] reporting a consistently negative R^2_{oos} for this dataset, many other studies, such as Campbell and Thompson [2007], have developed forecasting strategies resulting in economically meaningful R^2_{oos} values, which translate into significant economic gains through simple market timing strategies.

Table 1.2: Out-of-sample R-squared Values in Empirical Studies

	Ridge	Lasso	OLS/Ridgeless	RF	GBRT	NN(ℓ_2)	NN(ℓ_1)
Finance 1	0.80	−12.19	−81.08	1.30	−14.21	1.41	−10.31
Finance 2	0.19	0.10	−1.25	0.10	−0.30	0.26	0.14
Macro 1a	15.29	15.40	−1375	25.04	16.44	16.94	19.09
Macro 1b	3.49	3.69	−2939	9.08	1.11	7.09	5.39
Macro 2	6.58	−14.58	−837	9.03	1.28	4.00	1.92
	(4.83)	(43.74)	(854)	(9.92)	(14.04)	(18.42)	(13.36)
Micro 1	0.48	−1.01	−13198	−1.75	−5.07	0.49	−6.77
	(0.84)	(2.01)	(12479)	(2.70)	(6.60)	(0.27)	(17.87)
Micro 2a	26.27	20.37	−12729	27.69	16.44	23.87	23.37
	(7.50)	(6.41)	(9213)	(6.15)	(3.40)	(10.07)	(10.09)
Micro 2b	1.89	-3.43	-14724	2.86	-6.45	1.11	-1.73
	(3.09)	(5.25)	(10506)	(3.67)	(6.83)	(2.20)	(5.09)

Note: This table reports R^2_{oos} values, presented in percentages, for Ridge, Lasso, OLS/Ridgeless, RF, GBRT, and NNs with respective ℓ_1 and ℓ_2 penalties, across six empirical studies spanning Finance, Macroeconomics, Microeconomics. For the first example in Macroeconomics and the second example in Microeconomics, two benchmark models are considered. Standard deviations are provided in parentheses when applicable.

We revisit this exercise, by following the empirical framework of Giannone et al. [2022], and evaluate R^2_{oos} based on 57 different predictions using annually expanding windows.[18] The R^2_{oos} results are summarized in Table 1.2 and align with our theoretical predictions. Specifically, Ridge records an R^2_{oos} of 0.80%, significantly outperforming Lasso's -12.19%. With the smallest sample size being 17—just sufficient to run OLS with 16 predictors—OLS produces a highly negative R^2_{oos} of -81.08%. Our RF attains an R^2_{oos} of 1.30% while GBRT

17. The data, sourced from Amit Goyal's website, was processed using Giannone et al. [2022]'s methodology.

18. In this example, $R^2_{\text{oos}} = 1 - \sum_{t=1965}^{2021}(y_t - \hat{y}_t)^2 / \sum_{t=1965}^{2021}(y_t - \bar{y}_t)^2$, where $\bar{y}_t = \sum_{s=1948}^{t-1} y_t/(t-1948)$.

show performance similar to Lasso, with an R^2_{oos} of -14.21%. The NN with an ℓ_2 penalty, NN(ℓ_2), achieves the highest R^2_{oos} of 1.41%, whereas NN(ℓ_1) has -10.31%.

1.4.2 Finance 2: Cross-Section of Expected Returns

In our second analysis, we build upon the predictors utilized by Gu et al. [2020] for predicting monthly individual equity returns. Our dataset spans from March 1957 to December 2021, incorporating a total of 920 covariates and averaging over 6,200 stocks per month. Following Gu et al. [2020], we employ an expanding-window procedure annually.

Table 1.2 compares the model performance in terms of R^2_{oos}, where the zero predictor serves as the benchmark.[19] In this case, NN(ℓ_2) emerges as the leading model, closely followed by Ridge, achieving R^2_{oos} of 0.26% and 0.19%, respectively. These models dominate NN(ℓ_1) and Lasso, which records an R^2_{oos} of 0.14% and 0.10%. The performance of the OLS continues to be underwhelming in this exercise. For the tree-based models, RF outperforms GBRT, achieving an R^2_{oos} of 0.10%, while GBRT yields a negative R^2_{oos} of -0.30%.

To illustrate the economic significance of these relatively low R^2_{oos}s, we adopt a stock selection portfolio strategy. This strategy involves going long on the top 10% and shorting the bottom 10% of stocks, sorted based on their predicted returns for the upcoming month, with equal weighting applied to each stock every month. Over 35 out-of-sample years, NN(ℓ_2) achieves the highest Sharpe ratio at 2.13, followed closely by Ridge at 1.64. NN(ℓ_1) also performs well, with a Sharpe ratio of 1.55. By contrast, GBRT exhibits the least impressive performance, with the lowest Sharpe ratio of 0.80.

1.4.3 Macro 1: Macroeconomic Forecasting

The prediction of US macroeconomic activity using a wide range of predictors has been a topic of significant interest since its exploration by Stock and Watson [2002]. In our current

19. In this pooled regression setting, we define $R^2_{\text{oos}} = 1 - \sum_{i,t\in\text{OOS}}(y_{i,t} - \hat{y}_{i,t})^2 / \sum_{i,t\in\text{OOS}} y^2_{i,t}$.

study, we utilize the FRED-MD dataset, compiled by McCracken and Ng [2016], to forecast the monthly growth rate of US industrial production (IP). This dataset includes 119 potential predictors and extends from February 1960 to December 2019. Our evaluation methodology aligns with the prediction procedure outlined by Giannone et al. [2022].

We begin by training all models using data from February 1960 to December 1974 and then evaluate their performance on data from the subsequent year. This process is repeated 45 times, using a similar expanding-window approach on an annual basis. We follow the guidelines outlined by McCracken and Ng [2016] for covariate transformation and data quality management, including outlier removal and imputation of missing data.

We start with a benchmark predictor that only uses an intercept term. In this scenario, all machine learning methods significantly outperform this benchmark, achieving R^2_{oos} values ranging from 15.29% to 25.04%.[20] In contrast, OLS overfits the data, resulting in a negative R^2_{oos} of -14. This result indicates the existence of strong signals within the covariates. The benchmark model's lack of competitiveness aligns with our expectations, given the temporal dependence prevalent in macroeconomic time series.

We thereby propose an alternative benchmark that incorporates lagged values of IP growth. Within each training sample, we fit an AR model to the IP growth, selecting the order based on the AIC. The residuals from this model then serve as our prediction target. As discussed in Section 1.2.8, this approach combines the predictions from the AR model with those from our machine learning methods, yielding a hybrid out-of-sample forecast.[21] Consequently, the new benchmark becomes the direct use of AR model predictions. In this alternative setup, the comparison of R^2_{oos} values reveals a pattern somewhat associated with

20. In this case, the definition of R^2_{oos} is similar to how it is defined in the Finance 1 case.

21. In implementing advanced machine learning methods alongside a linear component $W\gamma$, we adopt a methodology that parallels the one used in Eq. (1.14). This approach is based on a DGP assumption that $\mathcal{M}_W y$ is a general function of $\mathcal{M}_W X$. This assumption plays a critical role in streamlining the implementation of these machine learning methods, ensuring that the results are directly comparable to those obtained in linear settings. However, it is generally not equivalent to the assumption that $y - W\gamma$ is a function of X.

scenarios of weak signal strength: NN(ℓ_2) and RF emerge as the top performers, achieving R^2_{oos} values of 7.09% and 9.08%, respectively. Following closely are NN(ℓ_1) at 5.39%, while GBRT lags with a considerably lower R^2_{oos} of 1.11%. In this example, linear models, specifically Ridge and Lasso, demonstrate comparable performance, achieving R^2_{oos} values of 3.49% and 3.69%, respectively. This suggests that, for this particular case, the challenges associated with signal weakness are primarily evident in nonlinear features.

1.4.4 Macro 2: Economic Growth Across Countries

Next, we explore a dataset originally compiled by Barro and Lee [1994], which includes 60 socio-economic, institutional, and geographical covariates across 90 countries. This dataset is utilized for predicting long-term economic growth, specifically measured by the growth rate of GDP per capita from 1960 to 1985. A pivotal aspect of this analysis involves testing a key prediction of the classical Solow-Swan-Ramsey growth model, which concerns the effect of an initial (lagged) GDP per capita level on subsequent growth rates. By incorporating the logarithm of each country's GDP per capita in 1960 alongside a constant term, our prediction model includes a total of 62 potential covariates.

Belloni et al. [2013b] implement the Square-root-Lasso technique in their regression, anticipating sparsity among the control variables. This methodology results in a remarkable sparse model, characterized by the inclusion of a singular control variable: the log of the black market premium, a measure of trade openness. In contrast, Giannone et al. [2022] employ a Bayesian approach with a spike-and-slab prior, concluding that a dense model, which includes all covariates, yields the best log-predictive score.

In our predictive analysis, we adopt the same empirical strategy outlined by Giannone et al. [2022]. We begin by randomly selecting half of the data samples for model estimation and then proceed to assess the performance of these models using the remaining samples. This process is repeated 100 times. The average out-of-sample R^2_{oos} from these 100 repeti-

tions, in comparison to a benchmark model that includes only the intercept, is presented in Table 1.2, accompanied by their standard deviations, provided in parentheses.[22][23]

Our empirical findings align with Giannone et al. [2022], indicating that dense models, specifically Ridge, RF, and NN(ℓ_2), exhibit superior performance compared to their sparse counterparts, such as Lasso, GBRT, and NN(ℓ_1). The limited sample size appears to disadvantage complex NN models, rendering them less effective than the simpler Ridge regression. RF demonstrates strong performance, achieving an R^2_{oos} of 9.03%, although it concurrently introduces a twofold increase in the variability of R^2_{oos} values compared to those based on Ridge. The variance of Lasso is pronounced, driven by a handful of extreme values; excluding these, its R^2_{oos} improves but remains notably low at -0.56%.

1.4.5 Micro 1: Crime Rates across US States

Our first microeconomic case revisits the study by Donohue and Levitt [2001], which investigates the impact of abortion legalization post-Roe vs. Wade on the decline in crime rates. They analyze the change in log per-capita murder rates from 1986 to 1997 across 48 states, totaling 576 observations. Belloni et al. [2013a] expand this analysis by including a broader set of 284 control variables, such as interactions and higher-order terms, to mitigate potential confounders. They apply Lasso to select control variables, finding none selected.[24] Similarly, Giannone et al. [2022] observe from their Bayesian analysis of this regression that the posterior density is concentrated on very low probability values of the slab component, which suggests that the regression model is sparse with high likelihood.

We employ the same benchmark model and sample splitting strategy outlined by Gi-

22. Here $R^2_{\text{oos}} = 1 - \sum_{i\in\text{OOS}} (y_i - \hat{y}_i)^2 / \sum_{i\in\text{OOS}} (y_i - \bar{y})^2$, where $\bar{y}$ is in-sample average of y_i.

23. We may also consider a benchmark model with GDP per capita in 1960 included, as predicted by theory. Interestingly, mandating this variable's inclusion reduces predictive performance in all models, resulting in a negative R^2_{oos} compared to a model with just an intercept.

24. Belloni et al. [2013a] proposes a double-Lasso estimator to make inference on the effect of abortion on murder rate. Part of their procedure involves a Lasso regression of murder rate on control variables. Notably, they use differences as the dependent variable, but observe no substantial changes when using levels instead.

annone et al. [2022], yielding 104 distinct training and evaluation samples. We report the mean and standard deviation of R^2_{oos}s in Table 1.2. Our findings reveal that NN(ℓ_2) exhibits a slight edge over Ridge, attaining a R^2_{oos} of 0.49%, compared to Ridge's 0.48%. Apart from these two models, all other models tested demonstrate negative R^2_{oos} values. Together, these results indicate an absence of strong signals in the data. We interpret the scarcity of significant signals reported in the literature as a result of their inherent weakness. Empirical evidence does not definitively categorize the underlying DGP as either dense or sparse—rather, it may simply be that no individual signals are particularly strong. Although a null model might initially seem appropriate, our findings indicate that while individual signals may be weak, their combined predictive power should not be overlooked.

1.4.6 Micro 2: Eminent Domain and Economic Outcomes

In our final study, we consider a causal analysis pertinent to eminent domain. Previous research by Chen and Yeh [2012], and subsequently Belloni et al. [2012], employ instrumental variable regressions to understand the impact of eminent domain decisions on economic outcomes. Differing from their broader focus, we closely follow Giannone et al. [2022], focusing on the first stage of predicting pro-plaintiff decisions in takings law cases based on judicial panel characteristics, using a dataset of 312 observations with 138 covariates across 100 unique training and evaluation sets.

We initially consider a benchmark model with only an intercept. In this setting, all machine learning methods successfully identify predictive signals and outperform this benchmark, as evidenced by large R^2_{oos}s. RF performs best with an R^2_{oos} of 27.69%, while other models also perform well, except for GBRT. However, the scenario shifts markedly when the benchmark is expanded to include a dummy variable for the absence of cases in a given circuit-year and the number of takings appellate decisions, for a total of three covariates. Against this expanded benchmark, the incremental predictive power of the remaining covari-

ates drops sharply. Ridge's R^2_{oos} falls to 1.89%, RF to 2.86%, and NN(ℓ_2) to 1.11%, while all other methods yield negative R^2_{oos} values.

1.5 Conclusion

In this chapter, we scrutinize the performance of machine learning techniques in contexts characterized by low signal-to-noise ratios, a situation frequently observed in economics and finance. Our theoretical analysis indicates that while Lasso is often considered a modern alternative to traditional ordinary least squares, its application in these areas should be approached cautiously, primarily due to its lessened effectiveness with weak signals.

Our research complements and expands upon the arguments made by Giannone et al. [2022], who cast doubt on the prevalence of sparsity in economic datasets. We take this debate further by showing that it is signal weakness, not necessarily the absence of sparsity, that more significantly contributes to the observed limitations of Lasso in economic applications. Furthermore, the lack of significant variables in empirical studies may be attributed more to signal weakness than to the sparse nature of the underlying DGP.

Our analysis also reveals a marked difference in the performance of Ridge regression. Notably, Ridge demonstrates superior resilience and effectiveness in these environments. Our theoretical findings are further substantiated by simulation studies encompassing a range of advanced machine learning techniques, including trees and neural networks. These experiments consistently reveal that algorithms designed to exploit sparsity tend to underperform in environments where signals are inherently weak. Broadly, our findings emphasize the importance of a nuanced, context-sensitive application of machine learning techniques, adapting to the distinctive data characteristics encountered across various domains.

1.6 Mathematical Proofs

1.6.1 Proof of Theorem 1

Proof. Throughout the proof, we employ the shorthand notation "w.a.p.1" to denote "with probability approaching one." For two random variables X and Y, we write $X \perp Y$ when they are independent and $X \overset{d}{=} Y$ when they have the same distribution. For convenience, we omit the subscript F from the expectation operator $\mathbb{E}_F(\cdot)$.

Our objective is to demonstrate that $\mathbb{E}\|\Sigma_2^{1/2}(\mathbb{E}(\beta_0|X,y)-\beta_0)\|^2/\mathbb{E}\|\Sigma_2^{1/2}\beta_0\|^2 \to 1$, as $n \to \infty$. This can be shown by proving $\mathbb{E}\|\Sigma_2^{1/2}\mathbb{E}(\beta_0|X,y)\|^2 = o(\mathbb{E}\|\Sigma_2^{1/2}\beta_0\|^2)$. Given that the eigenvalues of Σ_2 are bounded away from zero and positive infinity, it suffices to establish that $\mathbb{E}\|\mathbb{E}(\beta_0|X,y)\|^2 = o(\tau)$. Therefore, we need to prove for all $1 \le i \le p$, $\mathbb{E}(\mathbb{E}(\beta_{0,i}|X,y))^2 = o(p^{-1}\tau)$, or, equivalently, $\mathbb{E}(\mathbb{E}(b_{0,i}|X,y))^2 = o(1)$.

By the inequality $\mathbb{E}(\mathbb{E}(A|\mathcal{F})^2) \le \mathbb{E}(\mathbb{E}(A|\mathcal{G})^2)$ for $\mathcal{F} \subset \mathcal{G}$, and that β_0 is i.i.d., we have

$$\mathbb{E}(\mathbb{E}(b_{0,i}|X,y))^2 \le \mathbb{E}(\mathbb{E}(b_{0,i}|X,y,\beta_{0,-i}))^2 = \mathbb{E}(\mathbb{E}(b_{0,i}|X_{\cdot,i},\beta_{0,i}\Sigma_\varepsilon^{-1/2}X_{\cdot,i}+z))^2,$$

where z is defined in Assumption 2, $X_{\cdot,i}$ represents the i-th column of X, and $\beta_{0,-i}$ denotes the subvector of β without the ith entry. Denote the information set generated by $\{X_{\cdot,i},\beta_{0,i}\Sigma_\varepsilon^{-1/2}X_{\cdot,i}+z\}$ as $\mathcal{G}_i$. By Assumption 3, $b_{0,i}$ can be written as $q^{-1/2}b_{1i}b_{2i}$ where $b_{1i} \sim B(1,q)$ and b_{2i} is a sub-exponential random variable with mean zero and variance σ_β^2, whose distribution function is denoted by F_{b_2}. For any $M_1 < 0$, find M_2 (a function of M_1) such that $\mathbb{E}b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]} = q^{1/2}\mathbb{E}b_{2i}\mathbf{1}_{b_{2i}\in[M_1,M_2]} = 0$. This is always feasible because $\mathbb{E}b_{2i} = 0$. By Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}(\mathbb{E}(b_{0,i}|\mathcal{G}_i))^2 \le& 3\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i)\big)^2 + 3\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}>M_2}|\mathcal{G}_i)\big)^2 \\ &+ 3\mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}<M_1}|\mathcal{G}_i)\big)^2 =: 3S_{1n}+3S_{2n}+3S_{3n}. \qquad (1.21)\end{aligned}$$

Lemma 13 proves that for any given M_1, $\lim_{n\to\infty} S_{1n} = 0$. Observe that

$$S_{2n} = \mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}>M_2}|\mathcal{G}_i)\big)^2 \leq \mathbb{E}b_{0,i}^2\mathbf{1}_{q^{1/2}b_{0,i}>M_2}.$$

As a result, (1.21) implies

$$\lim_{n\to\infty} \mathbb{E}(\mathbb{E}(b_{0,i}|\mathcal{G}_i))^2 \leq 3\mathbb{E}b_{2i}^2\mathbf{1}_{b_{2i}>M_2} + 3\mathbb{E}b_{2i}^2\mathbf{1}_{b_{2i}<M_1}.$$

Since b_{2i} has finite variance, the right-hand-side of the above inequality can be arbitrarily small by letting $M_1 \to -\infty$, which completes the proof. □

1.6.2 Proof of Theorem 2

Proof. For ease of notation, we let $\hat{\beta} := \hat{\beta}_r(\lambda_n)$ and $c_n := p/n$. Additionally, define $\delta_1^* := 2\sqrt{\sigma_\varepsilon^2\theta_1}$, $\delta_2^* := (2\lambda\sigma_x^2\sigma_\beta^2\theta_4 - 4\sigma_\varepsilon^2\sigma_x^2\theta_3)/\delta_1^*\lambda$, $\mu(\sigma_x\sigma_\beta, \delta_1^*, \delta_2) := (\delta_1^*\delta_2 - 2\sigma_x^2\sigma_\beta^2\theta_4)/4\sigma_\varepsilon^2\theta_3$, and $C_n^\phi := c_n\tau^{-1}\sigma_x^2\sigma_\beta^2 - c_n\tau^{-1}\sigma_x^2(\delta_1^*)^2(4\lambda)^{-1} + c_n\sigma_\varepsilon^2\sigma_x^4\theta_3\lambda^{-2} - c_n\sigma_x^4\sigma_\beta^2\theta_4\lambda^{-1}$.

We first show that it is sufficient to establish that

$$c_n\tau^{-3/2}(\|\Sigma_2^{1/2}(\hat{\beta}-\beta_0)\| - \|\Sigma_2^{1/2}\beta_0\|) \xrightarrow{\mathrm{P}} \alpha_2^* := \theta_2\sigma_x^3\left(\frac{\sigma_\varepsilon^2\theta_1}{2\lambda^2\sigma_\beta} - \frac{\sigma_\beta}{\lambda}\right). \tag{1.22}$$

This is because Eq. (1.22) implies that $\|\Sigma_2^{1/2}(\hat{\beta}-\beta_0)\|^2 = \|\Sigma_2^{1/2}\beta_0\|^2 + 2c_n^{-1}\tau^{3/2}\alpha_2^*\|\Sigma_2^{1/2}\beta_0\| + o_{\mathrm{P}}(c_n^{-1}\tau^2)$. Additionally, using Lemma 2 and $q^{-1/2}p^{-1/2} = o(1)$ by Assumption 4, we have

$$\|\Sigma_2^{1/2}\beta_0\| = \tau^{1/2}\sigma_x\sigma_\beta + O_{\mathrm{P}}(q^{-1/2}p^{-1/2}\tau^{1/2}). \tag{1.23}$$

The above two equations together yield the desired result of the theorem. To prove Eq.

(1.22), by incorporating (1.23) and $c_n q^{-1/2} p^{-1/2} \tau^{-1} = o(1)$ by Assumption 4, it reduces to

$$c_n \tau^{-1} (\tau^{-1/2} \|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\| - \sigma_x \sigma_\beta) \xrightarrow{\mathrm{P}} \alpha_2^*. \tag{1.24}$$

Set $w = \tau^{-3/2}\Sigma_2^{1/2}(\beta - \beta_0)$ and $\hat{w} = \tau^{-3/2}\Sigma_2^{1/2}(\hat{\beta} - \beta_0)$. By Eq. (1.2), we have

$$\hat{w} = \arg\min_w \frac{c_n}{n} \left\| \tau^{1/2}\Sigma_1^{1/2} Z w - \tau^{-1}\varepsilon \right\|^2 + c_n^2 \lambda \left\| \Sigma_2^{-1/2} w + \tau^{-3/2}\beta_0 \right\|^2 - \frac{c_n \tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi, \tag{1.25}$$

where subtracting $c_n \tau^{-2}\|\varepsilon\|^2/n$ and C_n^ϕ from the objective function does not alter the solution. Using the definition of $\hat{w}$, proving (1.24) is equivalent to proving $c_n\|\hat{w}\| - c_n\tau^{-1}\sigma_x\sigma_\beta \xrightarrow{\mathrm{P}} \alpha_2^*$. Equivalently, we need to prove for all $\epsilon > 0$, w.p.a.1,

$$\alpha_2^* - \epsilon \le c_n\|\hat{w}\| - c_n\tau^{-1}\sigma_x\sigma_\beta \le \alpha_2^* + \epsilon. \tag{1.26}$$

Note that Eq. (1.25) is a high-dimensional optimization problem. By employing CGMT, Lemma 14 connects it to a scalar optimization problem below:

$$\begin{aligned}
&\min_{c_n|\alpha - \tau^{-1}\sigma_x\sigma_\beta| \le K_\alpha} \max_{\substack{\gamma > 0 \\ 0 \le \delta \le 4\tau^{-1}\sqrt{C_1 C_\varepsilon}}} -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top \\
&\times (\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{c_n\gamma}{2} \\
&+ \frac{c_n^2\lambda^2}{4}\left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0 + \frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right)^\top \left(\frac{\lambda}{4}\Sigma_2 + \frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1} \\
&\times \left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0 + \frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right) - \frac{c_n\tau\alpha^2\delta^2}{2\gamma n}\|h\|^2 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi,
\end{aligned} \tag{1.27}$$

where K_α is a sufficiently large fixed constant, μ_n is defined in Eq. (1.53), and $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ are standard Gaussian vectors independent of the other random variables. Denote the objective function as $Q_n(\alpha, \delta, \gamma)$. Let us define $\gamma = \tau^{-1}\gamma_1$, $\delta = \tau^{-1}\delta_1^* + \delta_2^* + c_n^{-1/2}\delta_3$ and $\alpha =$

$\tau^{-1}\sigma_x\sigma_\beta + c_n^{-1}\alpha_2$. Subsequently, we present the modified objective function $\tilde{Q}_n(\alpha_2, \delta_3, \gamma_1)$:

$$\tilde{Q}_n(\alpha_2, \delta_3, \gamma_1) := Q_n(\tau^{-1}\sigma_x\sigma_\beta + c_n^{-1}\alpha_2, \tau^{-1}\delta_1^* + \delta_2^* + c_n^{-1/2}\delta_3, \tau^{-1}\gamma_1). \tag{1.28}$$

By Lemma 14, $Q_n(\alpha, \delta, \gamma)$ is convex with respect to α and jointly concave with respect to (δ, γ). Therefore, it is evident that $\tilde{Q}_n$ remains convex with respect to α_2 and jointly concave with respect to (δ_3, γ_1). Lemma 14 implies that if the following claims hold

$$\begin{aligned}
&\text{(i) } \forall \text{ compact } A \subset [-K_\alpha, K_\alpha],\ \min_{\alpha_2 \in A} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n(\alpha_2, \delta_3, \gamma_1) \xrightarrow{\text{P}} \min_{\alpha_2 \in A} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in \mathbb{R}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1), \\
&\text{(ii) } \min_{\alpha_2 \in [-K_\alpha, K_\alpha]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1) < \min_{\alpha_2 \in [-K_\alpha, \alpha_2^* - \epsilon] \cup [\alpha_2^* + \epsilon, K_\alpha]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in \mathbb{R}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1)
\end{aligned} \tag{1.29}$$

where $\tilde{Q}(\alpha_2, \delta_3, \gamma_1) := -\frac{\delta_3^2\theta_1}{4\theta_3} + 2\sigma_x\sigma_\beta\alpha_2 - \frac{\gamma_1^2}{4\sigma_x^2\sigma_\beta^2\lambda}\theta_2 - \frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta} + \frac{(\delta_1^*)^2\gamma_1}{8\lambda^2\sigma_x^2\sigma_\beta^2}\theta_2$ and $K_{\delta_3} := [-c_n^{1/2}(\tau^{-1}\delta_1^* + \delta_2^*), 4c_n^{1/2}\tau^{-1}\sqrt{C_1 C_\varepsilon} - c_n^{1/2}(\tau^{-1}\delta_1^* + \delta_2^*)]$, then Eq. (1.26) holds. The above two conditions are verified in Lemma 18, thereby completing the proof.

□

1.6.3 Proof of Theorem 3

Proof. For convenience, we define the shorthand notation $\hat{\beta}^i_{\lambda_n} := \hat{\beta}^i_r(\lambda_n)$. Also, we define $\hat{R}^{K-CV}(\lambda_n) := \frac{1}{n}\sum_{i=1}^K \|y_{(i)} - X_{(i)}\hat{\beta}^i_r(\lambda_n)\|^2$. By Lemmas 2, 3 and 6, w.p.a.1, we have

$$n^{-1}\|X_{(-i)}^\top X_{(-i)}\| \le n^{-1}C_2\|Z_{(-i)}^\top Z_{(-i)}\| \le C_2(1 + \sqrt{c_n})^2, \quad i = 1, \ldots, K, \tag{1.30}$$

$$n^{-1}\|\varepsilon\|^2 \le 2\sigma_\varepsilon^2 \quad \text{and} \quad n^{-1}\|y\|^2 \le 2\sigma_\varepsilon^2. \tag{1.31}$$

Based on these inequalities, Lemma 19 proves

$$\inf_{\lambda\in[\epsilon,\tilde{c}\tau^{-1}]} pn^{-1}\tau^{-2}\left\{\hat{R}^{K-CV}(\lambda)-\frac{1}{n}\|\varepsilon\|^2-\|\Sigma_2^{1/2}\beta_0\|^2\right\}>0, \tag{1.32}$$

w.p.a.1 for some constant $\tilde{c}>0$. Additionally, for any fixed $\lambda>0$, Lemma 19 also proves

$$pn^{-1}\tau^{-2}\left\{\hat{R}^{K-CV}(\tau^{-1}\lambda)-\frac{1}{n}\|\varepsilon\|^2-\|\Sigma_2^{1/2}\beta_0\|^2\right\}\xrightarrow{\mathrm{P}}\frac{2(K-1)}{K}\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2}{2\lambda^2}-\frac{\sigma_\beta^2}{\lambda}\right). \tag{1.33}$$

Using (1.33) and λ^{opt}'s definition, $pn^{-1}\tau^{-2}\left\{\hat{R}^{K-CV}(\tau^{-1}\lambda^{opt})-\frac{1}{n}\|\varepsilon\|^2-\|\Sigma_2^{1/2}\beta_0\|^2\right\}<0$. Together with Eq. (1.32), the minimizer of $\hat{R}^{K-CV}(\lambda)$ must satisfy $\hat{\lambda}_n^{K-CV}\geq\tilde{c}\tau^{-1}$, w.a.p.1, so that, $\hat{\lambda}_n^{K-CV}=\arg\min_{\lambda_n\in[\tilde{c}\tau^{-1},\infty)}\hat{R}^{K-CV}(\lambda_n)$. Moreover, it also implies that $\tau^{-1}\lambda^{opt}\notin[\epsilon,\tilde{c}\tau^{-1}]$, that is, $\lambda^{opt}\geq\tilde{c}$.

Next, we re-parametrize the above optimization problem: $\widetilde{\mu}=\arg\min_{\mu\in[0,\tilde{c}^{-1}]}\widetilde{R}(\mu)$, where $\widetilde{R}(\mu):=\hat{R}^{K-CV}(\tau^{-1}\mu^{-1})$, and we extend the domain of $\widetilde{R}(\cdot)$ to include 0: $\widetilde{R}(0):=\lim_{\mu\to0}\widetilde{R}(\mu)=\|y\|^2/n$. Lemma 20 implies that $pn^{-1}\tau^{-2}\widetilde{R}(\mu)$ satisfies stochastic equicontinuity. Using this fact and Theorem 1 of Newey [1991], the convergence of

$$pn^{-1}\tau^{-2}\left\{\widetilde{R}(\mu)-\frac{1}{n}\|\varepsilon\|^2-\|\Sigma_2^{1/2}\beta_0\|^2\right\}\xrightarrow{\mathrm{P}}\frac{2(K-1)}{K}\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2\mu^2}{2}-\sigma_\beta^2\mu\right)$$

holds uniformly over the interval $[0,\tilde{c}^{-1}]$. Since $(\lambda^{opt})^{-1}$ is a unique minimizer of the right-hand-side and is nonzero, it follows that $\widetilde{\mu}\xrightarrow{\mathrm{P}}(\lambda^{opt})^{-1}$ and $\widetilde{\mu}=\tau^{-1}(\hat{\lambda}_n^{K-CV})^{-1}$, w.a.p.1, which implies that $\hat{\lambda}_n^{K-CV}/\lambda_n^{opt}\xrightarrow{\mathrm{P}}1$.

Now we prove $\Delta(\hat{\beta}_r(\hat{\lambda}_n^{K-CV}))-\Delta(\hat{\beta}_r(\lambda_n^{opt}))\xrightarrow{\mathrm{P}}0$. For notational simplicity, we write $\hat{\beta}_r(\hat{\lambda}_n^{K-CV})$ as $\hat{\beta}_{cv}$ and $\hat{\beta}_r(\lambda_n^{opt})$ as $\hat{\beta}_{opt}$. We need to prove $c_n\tau^{-2}(\|\Sigma_2^{1/2}(\hat{\beta}_{cv}-\beta_0)\|^2-$

$\|\Sigma_2^{1/2}(\hat{\beta}_{opt}-\beta_0)\|^2)=o_{\mathrm{P}}(1)$. By direct calculation, we have

$$
\begin{aligned}
&c_n\tau^{-2}\big(\|\Sigma_2^{1/2}(\hat{\beta}_{cv}-\beta_0)\|^2-\|\Sigma_2^{1/2}(\hat{\beta}_{opt}-\beta_0)\|^2\big)\\
=&c_n\tau^{-2}\big(\hat{\beta}_{cv}^\top\Sigma_2\hat{\beta}_{cv}-\hat{\beta}_{opt}^\top\Sigma_2\hat{\beta}_{opt}+2(\hat{\beta}_{opt}-\hat{\beta}_{cv})^\top\Sigma_2\beta_0\big) \qquad (1.34)\\
=&c_n\tau^{-2}\big((\hat{\beta}_{cv}-\hat{\beta}_{opt})^\top\Sigma_2\hat{\beta}_{cv}+\hat{\beta}_{opt}^\top\Sigma_2(\hat{\beta}_{cv}-\hat{\beta}_{opt})+2(\hat{\beta}_{opt}-\hat{\beta}_{cv})^\top\Sigma_2\beta_0\big)=:A_1+A_2+A_3.
\end{aligned}
$$

Lemma 21 proves that these terms converge to zero w.p.a.1, which completes the proof. □

1.6.4 Proof of Theorem 4

Proof. For ease of notation, we let $\hat{\beta}:=\hat{\beta}_l(\lambda_n)$. We adopt the same notation δ_1^* and $\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)$ as used in the proof of Theorem 2. Moreover, we define $\delta_2^*=2\sigma_x^2\sigma_\beta^2\theta_4/\delta_1^*$ and $C_n^\phi=c_n\tau^{-1}\sigma_x^2\sigma_\beta^2$, respectively. Similar to Theorem 2, it is essential to establish that the following inequality holds w.a.p.1 for any sufficiently small $\epsilon>0$:

$$
\frac{c_\alpha}{2\sigma_\beta}+\epsilon\le c_n\tau^{-1}(\tau^{-1/2}\|\Sigma_2^{1/2}(\hat{\beta}-\beta_0)\|-\sigma_x\sigma_\beta)\le\frac{C_\alpha}{2\sigma_\beta}-\epsilon. \qquad (1.35)
$$

Define $w=\tau^{-3/2}\Sigma_2^{1/2}(\beta-\beta_0)$. Using this, we can rewrite (1.3) as the following problem:

$$
\arg\min_w\frac{c_n}{n}\|\tau^{1/2}\Sigma_1^{1/2}Zw-\tau^{-1}\varepsilon\|^2+\frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}\|\Sigma_2^{-1/2}w+\tau^{-3/2}\beta_0\|_1-\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2-C_n^\phi.
$$

By CGMT, Lemma 22 establishes its connection with the following optimization problem:

$$
\begin{aligned}
&\min_{\alpha\in K_\alpha}\max_{\substack{\gamma>0\\ 0\le\delta\le4\tau^{-1}\sqrt{C_1C_\varepsilon}}}-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}\\
&\times(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)-\frac{c_n\gamma}{2}+\frac{c_n\gamma\tau^{-3}}{2\alpha^2}\beta_0\Sigma_2\beta_0+\frac{c_n\tau^{-1}\delta}{\sqrt{n}}h^\top\Sigma_2^{1/2}\beta_0-C_n^\phi \qquad (1.36)\\
&-\min_{\|v\|_\infty\le1}\frac{c_n\alpha^2}{2\gamma}\Big\|n^{-1/2}\tau^{-1/2}\lambda_n\Sigma_2^{-1/2}v-n^{-1/2}\tau^{1/2}\delta h-\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2^{1/2}\beta_0\Big\|^2-\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2,
\end{aligned}
$$

where $K_\alpha := \{\alpha | c_n\alpha - c_n\tau^{-1}\sigma_x\sigma_\beta \in [c_\alpha/4\sigma_\beta, C_\alpha/\sigma_\beta]\}$, μ_n is defined in Eq. (1.53), and $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ are standard Gaussian vectors independent of the other random variables. Denote the objective function as $Q_n(\alpha, \delta, \gamma)$. Similar to Theorem 2, we define $\gamma = \tau^{-1}\gamma_1$, $\delta = \tau^{-1}\delta_1^* + \delta_2^* + c_n^{-1/2}\delta_3$, and $\alpha = \tau^{-1}\sigma_x\sigma_\beta + c_n^{-1}\alpha_2$, obtaining the modified objective function:

$$\tilde{Q}_n(\alpha_2, \delta_3, \gamma_1) := Q_n(\tau^{-1}\sigma_x\sigma_\beta + c_n^{-1}\alpha_2, \tau^{-1}\delta_1^* + \delta_2^* + c_n^{-1/2}\delta_3, \tau^{-1}\gamma_1). \tag{1.37}$$

Note that $\delta_3 \in K_{\delta_3} := [-c_n^{1/2}(\tau^{-1}\delta_1^* + \delta_2^*), 4c_n^{1/2}\tau^{-1}\sqrt{C_1 C_\varepsilon} - c_n^{1/2}(\tau^{-1}\delta_1^* + \delta_2^*)]$. Lemma 22 implies that if the following inequalities

$$\begin{aligned}
&\min_{\alpha_2 \in [\frac{c_\alpha}{4\sigma_\beta}, \frac{C_\alpha}{\sigma_\beta}]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n(\alpha_2, \delta_3, \gamma_1) < -\frac{C_\lambda}{8C_2} + \eta, \\
&\min_{\alpha_2 \in [\frac{c_\alpha}{4\sigma_\beta}, \frac{c_\alpha}{2\sigma_\beta} + \epsilon] \cup [\frac{C_\alpha}{2\sigma_\beta} - \epsilon, \frac{C_\alpha}{\sigma_\beta}]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n(\alpha_2, \delta_3, \gamma_1) > -\frac{C_\lambda}{100C_2} - \eta,
\end{aligned} \tag{1.38}$$

hold for sufficiently small $\epsilon > 0$ and $\eta > 0$, then Eq. (1.35) holds. Finally, Lemma 24 verifies Eq. (1.38), thereby completing the proof. □

1.6.5 Proof of Proposition 1

Proof. Since the out-of-sample data are mutually independent, Lemmas 2 and 3 lead to

$$\sum_{i \in \mathrm{OOS}} y_i^2 = n_{\mathrm{oos}}(\sigma_\varepsilon^2 + \tau\sigma_x^2\sigma_\beta^2) + o_{\mathrm{P}}(n_{\mathrm{oos}}\tau)$$

and

$$\sum_{i \in \mathrm{OOS}} y_i^2 - (y_i - X_i\hat{\beta}_r(\lambda_n^{opt}))^2 = n_{\mathrm{oos}}p^{-1}n\tau^2\theta_2\sigma_x^4\sigma_\beta^4\sigma_\varepsilon^{-2}(1 + o_{\mathrm{P}}(1)),$$

where we use $\Delta(\hat{\beta}_r(\lambda_n^{opt})) = -\theta_2\sigma_x^4\sigma_\beta^4\sigma_\varepsilon^{-2} + o_{\mathrm{P}}(1)$ by Theorem 2 and $n_{\mathrm{oos}}p^{-2}n^2\tau^2 \to \infty$. The estimates above offer the key components for deriving the limit of R^2_{oos}:

$$R^2_{\mathrm{oos}} = \left(\sum_{i\in\mathrm{OOS}} y_i^2 - (y_i - X_i\hat{\beta}_r(\lambda_n^{opt}))^2\right)\left(\sum_{i\in\mathrm{OOS}} y_i^2\right)^{-1} = p^{-1}n\theta_2(R^2)^2(1+o_{\mathrm{P}}(1)). \quad \square$$

1.6.6 Proof of Theorem 5

Proof. For convenience, let $\hat{\beta} := \hat{\beta}_r(\lambda_n)$. We write the prediction error of the benchmark as:

$$\begin{aligned} y^{new} - \hat{y}_b^{new} =& (w^{new})^\top\gamma_0 + (x^{new})^\top\beta_0 + \varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top(W\gamma_0 + X\beta_0 + \varepsilon) \\ =& ((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)\beta_0 + (\varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top\varepsilon). \end{aligned}$$

Similarly, for the Ridge estimator, we have $y^{new} - \hat{y}^{new}$ equals

$$((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)(\beta_0 - \hat{\beta}) + (\varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top\varepsilon).$$

As a result, with simple algebra we can rewrite $\mathbb{E}\left[(y^{new} - \hat{y}^{new})^2|\mathcal{I}\right] - \mathbb{E}\left[(y^{new} - \hat{y}_b^{new})^2|\mathcal{I}\right]$ as

$$\begin{aligned} &\mathbb{E}\left[\left((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)(\beta_0 - \hat{\beta})\right)^2|\mathcal{I}\right] \\ &\quad - \mathbb{E}\left[\left((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)\beta_0\right)^2|\mathcal{I}\right] \\ &\quad - 2\mathbb{E}\left[((u^{new})^\top - (w^{new})^\top(W^\top W)^{-1}W^\top U)\hat{\beta}(\varepsilon^{new} - (w^{new})^\top(W^\top W)^{-1}W^\top\varepsilon)|\mathcal{I}\right] \\ :=& S_1 - S_2 - S_3. \end{aligned}$$

Below we analyze S_1 to S_3 one by one. With respect to S_1, we note that

$$\hat{\beta} = n^{-1}(n^{-1}X^\top\mathcal{M}_W X + n^{-1}p\tau^{-1}\lambda\mathbb{I})^{-1}X^\top\mathcal{M}_W(U\beta_0 + \varepsilon).$$

Define $R_X = (n^{-1}X^\top \mathcal{M}_W X + n^{-1}p\tau^{-1}\lambda\mathbb{I})^{-1}$. By direct calculations, we have

$$\begin{aligned}
&\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1}W^\top U\hat{\beta})^2 | \mathcal{I}\right] \asymp \|(W^\top W)^{-1}W^\top U\hat{\beta}\|^2 \\
\leq & 2n^{-2}\|(W^\top W)^{-1}W^\top U R_X X^\top \mathcal{M}_W U\beta_0\|^2 + 2n^{-2}\|(W^\top W)^{-1}W^\top U R_X X^\top \mathcal{M}_W \varepsilon\|^2.
\end{aligned} \tag{1.39}$$

For the second term in (1.39), we first note that for any constant $\lambda > 0$,

$$\begin{aligned}
&\|R_X X^\top \mathcal{M}_W X R_X\| = \|R_U U^\top \mathcal{M}_W U R_U\| = \frac{n\lambda_1(n^{-1}U^\top \mathcal{M}_W U)}{(\lambda_1(n^{-1}U^\top \mathcal{M}_W U) + n^{-1}p\tau^{-1}\lambda)^2} \\
&\asymp_{\mathrm{P}} p^{-1}n^2\tau^2,
\end{aligned}$$

since $\|n^{-1}U^\top \mathcal{M}_W U\| \lesssim_{\mathrm{P}} n^{-1}\|U\|^2 \lesssim_{\mathrm{P}} 1 + c_n = o(n^{-1}p\tau^{-1})$ by Lemma 6. Therefore, by Lemma 2 and using inequality $\mathrm{Tr}(AB) \leq \|A\|\,\mathrm{Tr}(B)$, for any $A = A^\top$ and $B \geq 0$, we have

$$\begin{aligned}
&n^{-2}\|(W^\top W)^{-1}W^\top U R_X X^\top \mathcal{M}_W \varepsilon\|^2 \\
\asymp_{\mathrm{P}} & n^{-2}\,\mathrm{Tr}((W^\top W)^{-1}W^\top U R_X X^\top \mathcal{M}_W X R_X U^\top W(W^\top W)^{-1}) \\
= & n^{-2}\,\mathrm{Tr}(R_X X^\top \mathcal{M}_W X R_X U^\top W(W^\top W)^{-2}W^\top U) \\
\leq & n^{-2}\|R_X X^\top \mathcal{M}_W X R_X\|\,\mathrm{Tr}(U^\top W(W^\top W)^{-2}W^\top U) \\
\lesssim_{\mathrm{P}} & p^{-1}\tau^2\,\mathrm{Tr}(U^\top W(W^\top W)^{-2}W^\top U) \leq p^{-1}\tau^2\|U^\top U\|\,\mathrm{Tr}((W^\top W)^{-1}) = o_{\mathrm{P}}(p^{-1}n\tau^3).
\end{aligned}$$

Similarly, we can prove that the first term in (1.39) is of order $o_{\mathrm{P}}(p^{-1}n\tau^3)$. Therefore, we have

$$\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1}W^\top U\hat{\beta})^2 | \mathcal{I}\right] = o_{\mathrm{P}}(p^{-1}n\tau^3). \tag{1.40}$$

In addition, using the independence of w^{new} with W, U, $\mathcal{I}$, and β_0, the facts that w^{new} has bounded variance, $\|(W^\top W)^{-1}W^\top \Sigma_1^{1/2}x\|^2 \asymp_{\mathrm{P}} \|(W^\top W)^{-1}W^\top \Sigma_1^{1/2}\|_F^2$ by Lemma 2,

$\mathrm{Tr}(W^\top W)^{-1} = o_{\mathrm{P}}(p^{-1}n\tau)$, and that $\|\Sigma_2^{1/2}\beta_0\|^2 \asymp_{\mathrm{P}} \tau$,

$$\begin{aligned}
&\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1}W^\top U\beta_0)^2|\mathcal{I}\right] \asymp \|(W^\top W)^{-1}W^\top U\beta_0\|^2 \\
&= \|(W^\top W)^{-1}W^\top \Sigma_1^{1/2} Z \Sigma_2^{1/2}\beta_0\|^2 \asymp_{\mathrm{P}} \|\Sigma_2^{1/2}\beta_0\|^2 \|(W^\top W)^{-1}W^\top \Sigma_1^{1/2}\|_{\mathrm{F}}^2 \\
&\asymp_{\mathrm{P}} o_{\mathrm{P}}(p^{-1}n\tau^2). \qquad (1.41)
\end{aligned}$$

With (1.40) and (1.41), $\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1}W^\top U(\beta_0 - \hat{\beta}))^2|\mathcal{I}\right] = o_{\mathrm{P}}(p^{-1}n\tau^2)$. In addition, since u^{new} is independent of $\mathcal{I}$ and w^{new}, we have

$$\mathbb{E}[(u^{new})^\top((\beta_0 - \hat{\beta})((w^{new})^\top (W^\top W)^{-1}W^\top U(\beta_0 - \hat{\beta})|\mathcal{I}] = 0.$$

Therefore, we conclude

$$S_1 = \mathbb{E}\left[((u^{new})^\top(\beta_0 - \hat{\beta}))^2|\mathcal{I}\right] + o_{\mathrm{P}}(p^{-1}n\tau^2) = \|\Sigma_2^{1/2}(\beta_0 - \hat{\beta})\|^2 + o_{\mathrm{P}}(p^{-1}n\tau^2).$$

By applying a similar argument, with the use of the independence of w^{new} with W, U, $\mathcal{I}$, and β_0, as well as the fact that w^{new} has bounded variance, it can be shown that

$$S_2 = \mathbb{E}\left[((u^{new})^\top\beta_0)^2|\mathcal{I}\right] + o_{\mathrm{P}}(p^{-1}n\tau^2) = \|\Sigma_2^{1/2}\beta_0\|^2 + o_{\mathrm{P}}(p^{-1}n\tau^2).$$

Finally we bound S_3. Since u^{new} and ε^{new} are mean zero, mutually independent, and independent of $\mathcal{I}$, along with Eq. (1.40), Lemma 2 and Cauchy-Schwartz inequality, we have

$$\begin{aligned}
|S_3| &= \left|2\mathbb{E}\left[(w^{new})^\top (W^\top W)^{-1}W^\top U\hat{\beta}(w^{new})^\top (W^\top W)^{-1}W^\top \varepsilon|\mathcal{I}\right]\right| \\
&\leq 2\left(\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1}W^\top U\hat{\beta})^2|\mathcal{I}\right]\right)^{1/2}\left(\mathbb{E}\left[((w^{new})^\top (W^\top W)^{-1}W^\top \varepsilon)^2|\mathcal{I}\right]\right)^{1/2} \\
&\asymp_{\mathrm{P}} o_{\mathrm{P}}(p^{-1/2}n^{1/2}\tau^{3/2})(\mathrm{Tr}(W(W^\top W)^{-2}W^\top))^{1/2} = o_{\mathrm{P}}(p^{-1}n\tau^2).
\end{aligned}$$

In sum, we have $S_1 - S_2 - S_3 = \|\Sigma_2^{1/2}(\beta_0 - \hat{\beta})\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 + o_{\mathrm{P}}(p^{-1}n\tau^2)$. Next we prove,

$$pn^{-1}\tau^{-2}(\|\Sigma_2^{1/2}(\hat{\beta} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2) \xrightarrow{\mathrm{P}} \alpha^*. \tag{1.42}$$

By Theorem 2, $\widetilde{\beta} := n^{-1}\widetilde{R}_U U^\top (U\beta_0 + \varepsilon)$ satisfies (1.42) with $\hat{\beta}$ being replaced by $\widetilde{\beta}$, where $\widetilde{R}_U := (n^{-1}U^\top U + n^{-1}p\tau^{-1}\lambda\mathbb{I})^{-1}$. Given that $\|\Sigma_2^{1/2}(\beta_0 - \hat{\beta})\|^2 = \|\Sigma_2^{1/2}(\beta_0 - \widetilde{\beta}) + \Sigma_2^{1/2}(\hat{\beta} - \widetilde{\beta})\|^2$ and that $\|\Sigma_2^{1/2}(\beta_0 - \widetilde{\beta})\| \asymp_{\mathrm{P}} \|\Sigma_2^{1/2}\beta_0\| \asymp_{\mathrm{P}} \tau^{1/2}$, it is easy to verify that (1.42) follows from

$$\|\hat{\beta} - \widetilde{\beta}\|^2 = o(n^2 p^{-2}\tau^3), \tag{1.43}$$

which is given by Lemma 26. □

1.6.7 Proof of Proposition 2

Proof. Let $\tau_0 = \sqrt{n\tau} \le \sqrt{\log n}/10$ and $\lambda_0 = \sqrt{n}\lambda/2$. Note that when $X = \mathbb{I}$, Lasso has a closed form solution $\hat{\beta}_l(\lambda) = ((|\varepsilon_1 + \tau_0| - \lambda_0)_+ \mathrm{sgn}(\varepsilon_1 + \tau_0), (|\varepsilon_2| - \lambda_0)_+ \mathrm{sgn}(\varepsilon_2), \ldots, (|\varepsilon_n| - \lambda_0)_+ \mathrm{sgn}(\varepsilon_n))^\top$. Therefore, $\|\hat{\beta}_l(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 = ((|\varepsilon_1 + \tau_0| - \lambda_0)_+ - \tau_0)^2 - \tau_0^2 + \sum_{i=2}^{n}((|\varepsilon_i| - \lambda_0)_+)^2$. Assume for now that $\max_{1\le i\le n}|\varepsilon_i| \ge |\varepsilon_1| + 2\tau_0$. Let $i_0 = \arg\max|\varepsilon_i|$, then we have

$$\|\hat{\beta}_l(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 \ge ((|\varepsilon_1 + \tau_0| - \lambda_0)_+ - \tau_0)^2 - \tau_0^2 + ((|\varepsilon_{i_0}| - \lambda_0)_+)^2 := Q$$

Consider the following three cases for ε_1: (i) $\varepsilon_1 \in [-\lambda_0 - \tau_0, \lambda_0 - \tau_0]$: $Q = 0 + ((|\varepsilon_{i_0}| - \lambda_0)_+)^2 \ge 0$; (ii) $\varepsilon_1 \in (-\infty, -\lambda_0 - \tau_0)$: $Q \ge -\tau_0^2 + ((|\varepsilon_1| + 2\tau_0 - \lambda_0)_+)^2 \ge -\tau_0^2 + 9\tau_0^2 \ge 0$;(iii) $\varepsilon_1 \in (\lambda_0 - \tau_0, \infty)$: $Q \ge -\tau_0^2 + ((|\varepsilon_1| + 2\tau_0 - \lambda_0)_+)^2 \ge -\tau_0^2 + \tau_0^2 = 0$. Therefore, under the event that $\max_{1\le i\le n}|\varepsilon_i| \ge |\varepsilon_1| + 2\tau_0$, it holds that $\|\hat{\beta}_l(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 \ge 0$. Now we evaluate

the probability of this event. Note that

$$\mathbb{P}(\max_{1\le i\le n}|\varepsilon_i|\le u)=(\mathbb{P}(|\varepsilon_i|\le u))^n=\left(\operatorname{erf}\left(\frac{u}{\sqrt{2}}\right)\right)^n\le\exp\left\{-\frac{n}{2}\exp\left\{-\frac{2}{\pi}u^2\right\}\right\},$$

where $\operatorname{erf}(\cdot)$ represents the Gauss error function. The last inequality uses the fact that $(\operatorname{erf}(x))^2\le 1-\exp(-4x^2/\pi)$ and $1+x\le e^x$. Reparametrizing u in terms of δ by solving $\delta=\exp\left\{-\frac{n}{2}\exp\left\{-\frac{2}{\pi}u^2\right\}\right\}$, we obtain that, with probability at least $1-\delta$:

$$\max_{1\le i\le n}|\varepsilon_i|\ge\sqrt{\frac{\pi}{2}\log\frac{n}{2}-\frac{\pi}{2}\log\log\frac{1}{\delta}}. \tag{1.44}$$

Choosing $\delta=n^{-1}$, the event $\mathcal{C}=\left\{\max_{1\le i\le n}|\varepsilon_i|\ge\sqrt{\frac{\pi}{2}\log\frac{n}{2}-\frac{\pi}{2}\log\log n}\right\}$ happens with probability at least $1-n^{-1}$. Setting $u=\sqrt{\frac{\pi}{2}\log\frac{n}{2}-\frac{\pi}{2}\log\log n}-2\tau_0$. There exists $n_0\in\mathbb{N}$, when $n\ge n_0$, $u\ge\sqrt{1.7\log n}-0.2\sqrt{\log n}\ge\sqrt{\log n}$. By Mills' inequalities, when $n\ge n_0$,

$$\begin{aligned}&\mathbb{P}\left(\max_{1\le i\le n}|\varepsilon_i|-2\tau_0\ge|\varepsilon_1|\Big|\mathcal{C}\right)\ge\mathbb{P}\left(|\varepsilon_1|\le\sqrt{\frac{\pi}{2}\log\frac{n}{2}-\frac{\pi}{2}\log\log n}-2\tau_0\right)\\ =&\mathbb{P}\left(|\varepsilon_1|\le u\right)\ge 1-\sqrt{\frac{2}{\pi}}\frac{\exp\left(-u^2/2\right)}{u}\ge 1-\sqrt{\frac{2}{\pi\log n}}n^{-1/2}.\end{aligned}$$

There exists $n_1\ge 1$, when $n\ge n_1$, $(1-\sqrt{\frac{2}{\pi\log n}}n^{-1/2})(1-\frac{1}{n})\ge 1-n^{-1/2}$. Therefore,

$$\mathbb{P}\left(\max_{1\le i\le n}|\varepsilon_i|-2\tau_0\ge|\varepsilon_1|\right)\ge\mathbb{P}(\mathcal{C})\cdot\mathbb{P}\left(\max_{1\le i\le n}|\varepsilon_i|-2\tau_0\ge|\varepsilon_1|\Big|\mathcal{C}\right)\ge 1-n^{-1/2},$$

for $n\ge\max(n_0,n_1)$, which implies Eq. (1.18). For Ridge, since $\hat{\beta}_r(\lambda)=(1+n\lambda)^{-1}y$, we have

$$\|\hat{\beta}_r(\lambda)-\beta_0\|^2-\|\beta_0\|^2=\frac{-2n\lambda\varepsilon_1\tau_0-(1+2n\lambda)\tau_0^2+\sum_{i=1}^n\varepsilon_i^2}{(1+n\lambda)^2}.$$

Observe that with probability equal to 0.5, $\varepsilon_1\tau_0 \geq 0$. Under this event, $\|\hat{\beta}_r(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 \leq \frac{-(1+2n\lambda)\tau_0^2+\sum_{i=1}^n \varepsilon_i^2}{(1+n\lambda)^2}$. Therefore, $\|\hat{\beta}_r(\lambda) - \beta_0\|^2 - \|\beta_0\|^2 < 0$ as long as $\lambda > (2n)^{-1}(-1 + \tau_0^{-2}\sum_{i=1}^n \varepsilon_i^2)$. □

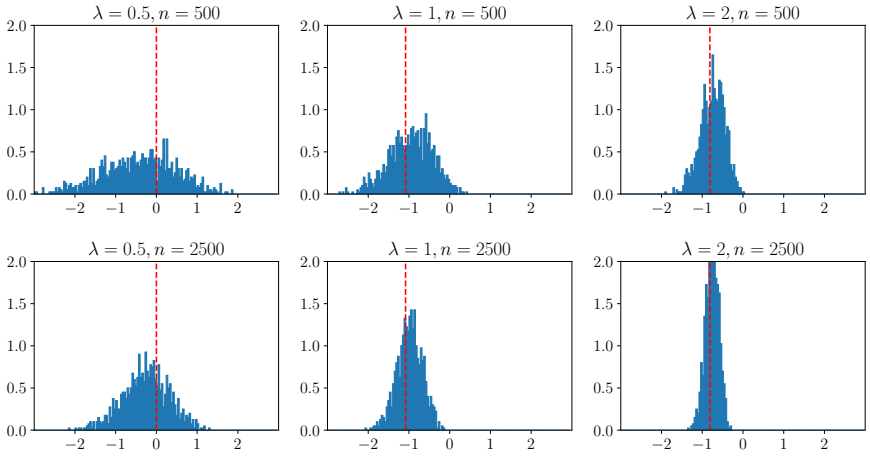
1.7 Supplemental Simulation Results

1.7.1 Additional Simulations with Fixed Tunings

In the simulation for the main paper, cross-validation is applied for Ridge and Lasso. In this section, we verify our theories with manually selected λ_n. We also experiment with two sample sizes, $n = 500$ and $n = 2,500$, while maintaining $p/n = 3/5$. We fix $q = 0.2$ and $R^2 = 5\%$. In the case of Ridge regression, we set λ as 0.5, 1 and 2, where $\lambda = 1$ corresponding to the optimal tuning. The histograms of relative prediction error are presented in Figure 1.8.

Several noteworthy observations can be made from these histograms. First, across all plots, the probability mass is concentrated around the red vertical line. As the sample size increases from 500 to 2,500 (and dimension increases from 300 to 1,500), the histograms become increasingly concentrated. This aligns with our theory, which predicts that the relative prediction error converges in probability to the limit α^* as the sample size grows. Second, the value of α^* corresponding to the optimal tuning parameter $\lambda = 1$ is the smallest. This is because the optimal Ridge estimator achieves the smallest prediction error. Moreover, almost all the probability mass corresponding to the optimal Ridge estimator is situated on the negative side of the x-axis, indicating that this estimator outperforms the zero estimator with high probability. Third, when $\lambda = 0.5$, it results in the worst performance, with a large portion of the probability mass on the positive side of zero. In contrast, for $\lambda = 2$, α^* gets closer to zero, and the variance of the relative prediction error decreases. This behavior is due to the increasing amount of penalization, which ultimately drives the estimator towards

Figure 1.8: Simulation Results for Ridge with Fixed Tuning Parameters



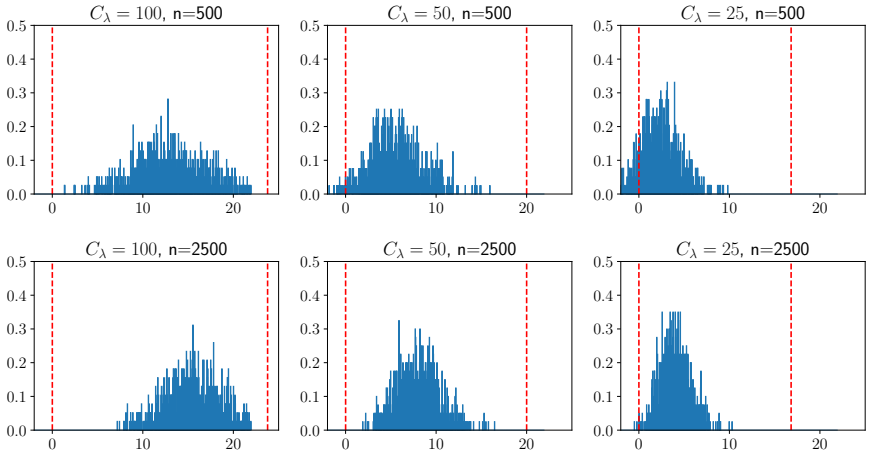
Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_r(\lambda_n))$ following equation (1.8) across $1,000$ Monte Carlo samples. We consider two different sample sizes ($n = 500$ and $n = 2,500$) and examine three different values of λ, where $\lambda = 0.5, 1$, and 2. Notably, $\lambda = 1$ represents the optimal tuning parameter. The red dashed line indicates the values of α^*.

zero, and in turn, α^* towards zero as well.

In contrast to the results obtained for Ridge regression, our theoretical framework does not provide a precise error limit for Lasso. Instead, Theorem 4 offers high probability bounds on relative prediction errors. Figure 1.9 displays histograms of these errors for various tuning parameters and sample sizes, accompanied by two red vertical lines in each plot representing the lower and upper bounds, c_α and C_α.

These plots yield several interesting findings. First, as the sample size increases, we observe that the probability mass becomes more concentrated and largely falls within the intervals defined by the bounds. Second, regardless of the tuning parameter values, Lasso consistently underperforms the zero estimator in almost all samples when the sample size is large. Third, as the tuning parameter increases (indicated by a decrease in C_λ), both the lower and upper bounds approach zero. This behavior is a consequence of the increased regularization, which, in turn, steers the estimator closer to zero. In the end, Lasso becomes identical to the zero estimator.

Figure 1.9: Simulation Results for Lasso with Fixed Tuning Parameters



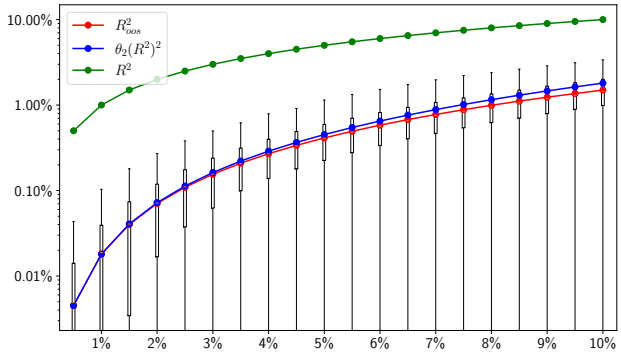
Note: The histograms depict the relative prediction error $\Delta(\hat{\beta}_l(\lambda_n))$ following equation (1.8) across $1,000$ Monte Carlo samples. We consider two different sample sizes ($n = 500$ and $n = 2,500$) and examine three different values of C_λ. The two dashed lines in each figure indicate the values of c_α and C_α that are solutions to (1.10).

1.7.2 Out-of-sample R^2

Continuing our investigation in the main text, we conduct an experiment to analyze R^2_{oos} based on the optimal Ridge. Proposition 1 describes the expected asymptotic behavior of R^2_{oos}. To empirically test this, we implement the optimal Ridge, setting $\lambda = 1$, on a training dataset comprising $n = 500$ observations. We then calculate R^2_{oos} based on predictions for a separate test dataset of size $n_{\text{oos}} = 10,000$. The comparative analysis between the population R^2, the empirically estimated R^2_{oos}, and the theoretically derived limit of R^2_{oos} is illustrated in Figure 1.10. For a clearer visual presentation, we apply a logarithmic transformation to the y-axis. We vary τ to compare against a range of population R^2 values from 0.5% to 10% on the x-axis. The red line represents the average R^2_{oos} over $1,000$ Monte Carlo simulations. Additionally, we draw boxplots to describe the distributions of R^2_{oos} across these simulations. The theoretical limit, expressed as $p^{-1}n\theta_2(R^2)^2$, is traced by the blue line, and the green line illustrates the population R^2, which would align with a 45-degree line on a standard scale. Notably, in this weak signal setting, the population R^2 significantly

surpasses the empirically achievable R^2_{oos}. Furthermore, the close alignment between the red and blue lines, particularly for scenarios with small R^2 values, substantiates our theoretical predictions.

Figure 1.10: Out-of-Sample R^2 for Optimal Ridge in Linear DGPs



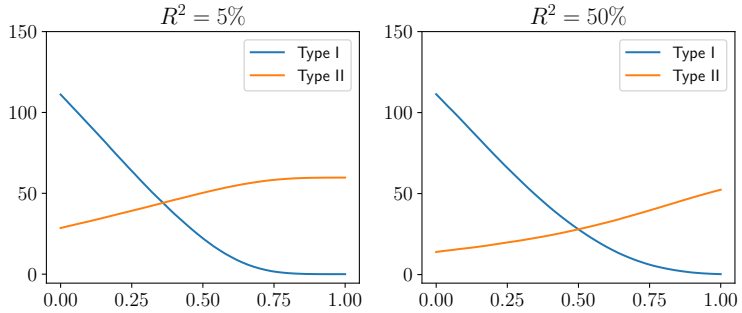
Note: The figure presents boxplots showing the distributions of R^2_{oos} for optimal Ridge regression ($\lambda = 1$) over $1,000$ Monte Carlo repetitions, with $n = 500$, $p = 300$, $q = 0.2$, and $n_{oos} = 10,000$. We explore a range of population R^2 values, from 0.5% to 10% in increments of 0.5% by adjusting τ. The plot features red, blue, and green lines to represent the average R^2_{oos} over Monte Carlo samples, the theoretical limit as given by Proposition 1, and the population R^2. In this plot, we employ a logarithmic scale for the y-axis. Without the logarithmic transformation, the green line would align with a 45-degree line. Additionally, the lower boundaries of the boxplots surpass the axis limits in instances where the R^2_{oos} values are negative.

1.7.3 Why Lasso Fails?

A plausible explanation for the Lasso's suboptimal performance with weak signals is its difficulty in distinguishing between genuine and spurious signals. The failure to identify genuine weak signals has a minor impact on Lasso's performance relative to the zero estimator, which does not utilize any true signals. Hence, the primary challenge for the Lasso lies in its failure to adequately filter out irrelevant signals. This issue could be addressed with a sufficiently large tuning parameter. However, our theory indicates that only when the penalty is so substantial that the Lasso effectively becomes equivalent to the zero estimator does it apply an adequate penalty.

To empirically explore this issue, we quantify Type I and Type II errors in simulations of Lasso's selection relative to its tuning parameter λ_n. The findings are presented in Figure 1.11. Considering our previous discussion, Type I errors represent a significant cost for Lasso. Indeed, a considerable portion of the variables selected by Lasso are incorrectly deemed genuine when λ_n is small. As λ_n increases, Type I errors decrease, enhancing Lasso's performance. Meanwhile, Type II errors persist and eventually converge to the number of non-zero betas in the DGP.

Figure 1.11: Lasso's Type I and Type II Errors



Note: The plots compare average Type I and Type II errors of Lasso using the linear DGP following equation (1.1) over $1,000$ Monte Carlo samples. Two population R^2 are considered: $R^2 = 5\%$ (left panel) and $R^2 = 50\%$ (right panel). The horizontal axis of each plot represents the logarithm of λ_n, spanning a range from 0 to 1. The vertical axis measures the count of errors incurred while testing the null hypothesis $\mathbb{H}_{0,i} : \beta_i = 0$.

1.7.4 Robustness Check

In our subsequent series of experiments, we intentionally deviate from the assumptions originally established during the development of our theoretical framework. This deviation is aimed at evaluating the robustness and generalizability of our theoretical predictions beyond their premises and initial parameters. To facilitate this evaluation, we introduce specific modifications to the baseline configuration along three key dimensions: First, we adjust (R^2, q), exploring more extreme sparsity levels and reducing signal strength accordingly compared to the settings in the main text. Second, we increase the ratio p/n to 2 by increasing p

while maintaining n, making it more challenging for both Ridge and Lasso to capture the underlying signals. Third, we modify the distribution of Z from standard Gaussian to a t-distribution characterized by four degrees of freedom, and with a mean of zero and a variance of one. In addition, we introduce heteroscedasticity into the error distribution, following the configuration outlined by Giannone et al. [2022]. The error term's variance is defined by the function $\sigma^2 \exp(\alpha X_i^\top \delta / \sqrt{\sum_{i=1}^n (X_i^\top \delta)^2 / n})$ with $\alpha = 0.5$. Here, X_i represents the i-th row of X. σ serves as a scaling parameter to standardize the variance and match $\sigma_\varepsilon^2 = 1$. The vector δ is a $p \times 1$ vector with zero elements in the same positions as the zero elements of β_0, while non-zero elements are drawn from a standard Gaussian distribution.

Table 1.3 compare the summary statistics for various cases under consideration. In Case I, when q is small, the performance of the Lasso estimator improves relative to the baseline scenario (reproduced from Table 1.1 for ease of comparison). This improvement is evident at $R^2 = 5\%$ for all levels of q, as the Q1 values become negative, indicating that Lasso surpasses zero in predictive accuracy for a larger proportion of Monte Carlo repetitions. However, as R^2 is further reduced to 2%, Lasso once again becomes falls below the performance of zero. In contrast, Ridge's performance remains largely unaffected by changes in sparsity levels. As expected, its performance deteriorates in finite samples as the signal strength weakens (i.e., as R^2 decreases). Nonetheless, Ridge continues to outperform Lasso, although its relative advantage over the zero estimator diminishes. The theoretical support for these observations is discussed in Appendix 1.2.9. In Case II, we observe the increased ratio of p/n does not affect our conclusion. Case III demonstrate the robustness of our theoretical findings, as it aligns closely with the baseline scenario despite variations in distributional assumptions.

1.8 Choice of Tuning Parameters

In this section, we discuss the selection of tuning parameters for implementing machine learning methods. The selection process aims to balance performance and cost while ensuring

Table 1.3: Robustness Analysis of Ridge and Lasso in Alternative DGPs

			Lasso				Ridge			
	q	R^2 (%)	Q1	Q2	Q3	#Zero	Q1	Q2	Q3	#Zero
	0.20	5%	-0.127	0.000	0.521	360	-0.992	-0.501	-0.129	97
	0.10	5%	-0.871	0.000	0.187	327	-0.981	-0.475	-0.077	113
	0.10	2%	0.000	0.000	3.435	493	-0.622	0.000	0.440	237
Case I	0.05	5%	-2.688	-0.305	0.000	255	-1.037	-0.387	0.000	130
	0.05	2%	0.000	0.000	2.948	473	-0.642	0.000	0.426	238
	0.02	5%	-6.542	-2.050	0.000	215	-1.304	-0.230	0.000	149
	0.02	2%	0.000	0.000	1.695	432	-0.605	0.000	0.625	254
Case II	0.20	5%	0.000	0.000	3.228	470	-0.768	-0.416	0.000	183
Case III	0.20	5%	0.000	0.000	0.591	392	-0.848	-0.384	0.000	129

Note: The table illustrate the summary statistics of relative prediction error $\Delta(\hat{\beta}(\hat{\lambda}_n^{K-CV}))$ for Ridge and Lasso based on 1,000 Monte Carlo samples. We explore several distinct DGPs, each involving the alteration of a specific condition. In Case I, we try a series of different values of R^2 and q. In Case II, we adjust n/p to 0.5. In Case III, we introduce t-distributed covariates with heterogeneous variance of ε. The benchmark DGP adheres to the following specifications: $n = 500$, $p = 300$, $p/n = 3/5$, and complies with Assumptions 1 and 2. 10-fold cross-validation is used throughout these experiments.

fair method comparisons.

For Ridge and Lasso, each with one tuning parameter, we use the glmnet package, which optimizes it via ten-fold CV. The grid is adaptively selected for efficiency. Our implementation of RF involves three tuning parameters: the depth of each individual tree, the number of randomly selected features used in each tree split, and the proportion of sample data used for bootstrap.[25] [26] For GBRT, we tune tree depth, number of trees, and learning rate.[27] In the case of NNs, we adhere to a uniform architectural choice across our analyses, featuring a single hidden layer. The number of neurons in this hidden layer is approximately equal to the square root of the total number of neurons in the input layer, aligning the architecture with the complexity and dimensions of the dataset. By not tuning the NN architecture extensively, we streamline the model selection process while retaining adequate complexity for

25. This procedure is known as subbagging, which helps address weak signals. In linear regressions, LeJeune et al. [2020] show that the asymptotic risk of subbagging least squares matches that of Ridge regression.

26. In our simulations, tree depth varies from 5 to 20, selected features from 10 to 300, and bootstrap sample proportion from 0.1 to 0.2. The RF ensemble size is fixed at 5,000 trees, as 10,000 offers no significant improvement.

27. The learning rate varies from 0.001 to 0.5, tree depth from 1 to 6, and the maximum number of trees is 100, though training usually stops earlier, reflecting GBRT's preference for shallower and fewer trees.

effective learning. For the remaining tuning parameters in trees and NNs, we select suitable ranges based on model performance from the cross-validation step. A critical element in selecting our grid is to ensure that the optimal tuning parameters are situated within the median range of the grid. The details regarding model configuration and tuning parameters for empirical studies are provided in Table 1.4.[28]

1.9 Technical Lemmas and Their Proofs

For completeness, the following section introduces a collection of lemmas, including proofs for some. We start with the Convex Gaussian Min-max Theorem (CGMT), a pivotal theorem to our proof. For a detailed exposition of its proof, we direct readers to the work of Thrampoulidis et al. [2015]. The CGMT pertains to the following optimization problems:

$$\Phi(G) := \min_{w\in\mathcal{S}_w}\max_{u\in\mathcal{S}_u} u^\top Gw + \psi(w,u), \text{ and } \phi(g,h) := \min_{w\in\mathcal{S}_w}\max_{u\in\mathcal{S}_u} \|w\|g^\top u - \|u\|h^\top w + \psi(w,u),$$

where $G \in \mathbb{R}^{m\times n}, g \in \mathbb{R}^m, h \in \mathbb{R}^n, \mathcal{S}_w \subset \mathbb{R}^n, \mathcal{S}_u \subset \mathbb{R}^m$, and $\psi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$.

Lemma 1 (CGMT)**.** *Suppose that $\mathcal{S}_w$ and $\mathcal{S}_u$ are compact sets, ψ is continuous on $\mathcal{S}_w \times \mathcal{S}_u$, and the entries of G, g, and h are i.i.d. Gaussian. Then we have $\mathrm{P}(\Phi(G) < c) \leq 2\mathrm{P}(\phi(g,h) \leq c)$, $\forall c \in \mathbb{R}$. Moreover, if $\mathcal{S}_w$ and $\mathcal{S}_u$ are convex sets, and ψ is convex-concave on $\mathcal{S}_\mathrm{w} \times \mathcal{S}_u$, then $\mathrm{P}(\Phi(G) > c) \leq 2\mathrm{P}(\phi(g,h) \geq c)$, $\forall c \in \mathbb{R}$.*

Lemma 2 (Lemma B.26 from Bai and Silverstein [2009])**.** *Let $x = (x_1, \cdots, x_n)^\top$ be a random vector of i.i.d. entries. Assume that $\mathbb{E}x_i = 0$, $\mathbb{E}x_i^2 = 1$, and $\mathbb{E}x_i^4 \leq v_4$. Then, for any $A \in \mathbb{R}^{n\times n}$, it holds that $x^\top Ax - \mathrm{Tr}(A) = O_\mathrm{P}\big(\sqrt{v_4\,\mathrm{Tr}(AA^\top)}\big)$.*

28. We follow the same approach for the empirical analysis, except for Finance 2. Due to its scale and computational constraints, we use two-fold CV and a narrower grid: $\log(\lambda)$ ranges from 6 to 7 for Ridge and from -3.5 to -2.5 for Lasso. We ensure that the optimal tuning parameters fall within the central range of these specified grids.

Table 1.4: Model Configuration for Machine Learning Methods

	RF	GBRT	NN(ℓ_2)	NN(ℓ_1)
Finance 1	depth=1∼20 #trees=500 #features=1∼15 %samples=0.05∼1	depth=1∼5 #trees=1∼10 lr ∈ {0.01,0.02, 0.05,0.1,0.2,0.5,1}	architecture∼{16,4,1} batch size=16 (lr,epochs)={(0.1,5), (0.01,50), (0.0025,200)} $\log(\lambda) \in [-2, 1]$	architecture∼{16,4,1} batch size=16 (lr,epochs)={(0.4,1), (0.08,5), (0.02,20)} $\log(\lambda) \in [-2, 1]$
Finance 2	depth=2∼12 #trees=500 #features ∈ {1, 2,3,5} %samples=0.5∼1	depth=1∼6 #trees=10∼400 lr ∈ {0.0001,0.001, 0.01,0.02,0.05}	architecture∼{920,32,1} batch size=10000 (lr,epochs)={(0.5,2), (0.1,10), (0.067,15)} $\log(\lambda) \in [-4, 0]$	architecture∼{920,32,1} batch size=10000 (lr,epochs)={(0.5,2), (0.2,5), (0.067,15), (0.05,20),(0.04,25)} $\log(\lambda) \in [-5, -3]$
Macro 1	depth=5∼50 #trees=500 #features=1∼60 %samples=0.5∼1	depth=1∼5 #trees=1∼600 lr ∈ {0.005,0.01, 0.02,0.05,0.1, 0.2,0.5}	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.008,10), (0.004,20), (0.002,40), (0.0008,100), (0.0005,160)} $\log(\lambda) \in [-2, 2]$	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.05,2), (0.02,5),(0.01,10), (0.005,20), (0.002,50)} $\log(\lambda) \in [-10.5, 1.5]$
Macro 1b	depth=5∼50 #trees=500 #features=5∼100 %samples=0.5∼1	depth=1∼5 #trees=1∼200 lr ∈ {0.01,0.02, 0.05,0.1,0.2,0.5}	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.024,75), (0.012,150), (0.006,300) (0.003,600)} $\log(\lambda) \in [-1, -0.5]$	architecture∼{119,8,1} batch size=16 (lr,epochs)={(0.004,25), (0.002,50),(0.001,100), (0.0005,200)} $\log(\lambda) \in [2, 3]$
Macro 2	depth=1∼30 #trees=500 #features=1∼60 %samples=0.5∼1	depth=1∼10 #trees=1∼500 lr ∈ {0.01,0.02, 0.05,0.1,0.2,0.5}	architecture∼{61,8,1} batch size=16 (lr,epochs)={(0.02,50), (0.005,200), (0.00125,800)} $\log(\lambda) \in [-3, -3]$	architecture∼{61,8,1} batch size=16 (lr,epochs)={(0.05,20), (0.02,50), (0.002,500)} $\log(\lambda) \in [-7, 4]$
Micro 1	depth=1∼20 #trees=500 #features=1∼20 %samples=0.005∼0.5	depth=1∼5 #trees=1∼20 lr ∈ {10^{-15},10^{-14}, ...,0.05,0.1,0.2,0.5}	architecture∼{297,16,1} batch size=16 (lr,epochs)={(0.1,1), (0.01,10), (0.001,100), (0.0001,1000), (0.00005,2000)} $\log(\lambda) \in [-11, -7]$	architecture∼{297,16,1} batch size=16 (lr,epochs)={(0.4,1), (0.2,2), (0.08,5), (0.04,10), (0.02,20)} $\log(\lambda) \in [-8, 0]$
Micro 2	depth=5∼30 #trees=500 #features=1∼30 %samples=0.5∼1.0	depth=1∼6 #trees=1∼30 lr ∈ {0.05,0.1, 0.15,. . . ,1}	architecture∼{217,16,1} batch size=16 (lr,epochs)={(0.1,1),(0.02,5), (0.01,10),(0.005,20)} $\log(\lambda) \in [-10, -6]$	architecture∼{217,16,1} batch size=16 (lr,epochs)={(0.1,1),(0.01,10), (0.001,100), (0.0001,1000)} $\log(\lambda) \in [-12, -9]$
Micro 2b	depth=1∼50 #trees=500 #features=1∼3 %samples=0.5∼1	depth=1∼10 #trees=1∼50 lr ∈ {10^{-10},10^{-9}, . . . ,0.1,0.2,0.5,1}	architecture∼{215,16,1} batch size=16 (lr,epochs)={(0.04,5), (0.02,10), (0.01,20)} $\log(\lambda) \in [0, 2]$	architecture∼{215,16,1} batch size=16 (lr,epochs)={(0.01,50), (0.005,100), (0.0025,200)} $\log(\lambda) \in [0, 2]$

Note: The table reports the range of tuning parameters for RF, GBRT, and NNs, as well as the architecture of NNs applied across six datasets. For RF, we fix the number of trees at #trees= 500, and tune three other parameters: the depth of the tree (depth), the number of features (#features), and the ratio of bootstrapped samples (%samples) within a predefined grid. In the case of GBRT, we tune depth and #trees, and the learning rate (lr). For NNs, we adopt a fixed model architecture, denoted by the number of neurons in each layer indicated in brackets. Additionally, we fix the batch size for SGD and focus on jointly tuning the learning rate (lr) and the number of epochs (epochs), as well as the ℓ_1- or ℓ_2-penalty parameter ($\log(\lambda)$).

Lemma 3. *Let $x = (x_1, \cdots, x_n)^\top$ and $y = (y_1, \cdots, y_m)^\top$ be two independent random vectors with i.i.d. entries. Assume that each element has a mean of zero and a variance of one. Then, for any $A \in \mathbb{R}^{n\times m}$, it holds that $x^\top Ay = O_{\mathrm{P}}\big(\sqrt{\mathrm{Tr}(AA^\top)}\big)$.*

Proof. The conclusion follows from the fact that $\mathbb{E}(x^\top Ay)^2 = \mathrm{Tr}(AA^\top)$. □

The lemma below pertains to the Neumann series. See Meyer [2000] for a detailed proof.

Lemma 4. *If A is a square matrix with $\|A\| < 1$, then $\mathbb{I} - A$ is nonsingular and $(\mathbb{I} - A)^{-1} = \sum_{k=0}^{\infty} A^k$. As a consequence, $\|(\mathbb{I}-A)^{-1} - \sum_{k=0}^{\ell} A^k\| \le \sum_{k=\ell+1}^{\infty} \|A\|^k = \|A\|^{\ell+1}/(1-\|A\|)$.*

Lemma 5. *Assume $x = (x_1, \cdots, x_n)^\top$ and $y = (y_1, \cdots, y_p)^\top$ are two independent random vectors with i.i.d. sub-exponential random variables with their sub-exponential norm bounded by K. Then for any $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times p}$, there exists a constant $c > 0$ such that*

$$\mathrm{P}\left(|x^\top Ax - \mathbb{E}x^\top Ax| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|A\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|A\|^{1/2}}\right\}\right), \tag{1.45}$$

$$\mathrm{P}\left(|x^\top By| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|B\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|B\|^{1/2}}\right\}\right). \tag{1.46}$$

Proof. Inequality (1.45) is given by Proposition 1.1 presented in Götze et al. [2021] for the case of symmetric A. To extend it for the asymmetric case, we use the identity that $x^\top Ax = x^\top(A + A^\top)x/2$, which allows us to apply (1.45) to $(A + A^\top)/2$. Using triangle inequalities, we have $\left\|(A + A^\top)/2\right\|_{\mathrm{F}}^2 \le \|A\|_{\mathrm{F}}^2$ and $\left\|(A + A^\top)/2\right\|^{1/2} \le \|A\|^{1/2}$. Thus, (1.45) holds for asymmetric A. For (1.46), let $z = (x^\top, y^\top)^\top$ and $C = \begin{pmatrix} 0_{n\times n} & B \\ 0_{p\times n} & 0_{p\times p} \end{pmatrix}$. By (1.45), we obtain

$$\begin{aligned}\mathrm{P}\left(|x^\top By| \ge t\right) = \mathrm{P}\left(|z^\top Cz| \ge t\right) &\le 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|C\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|C\|^{1/2}}\right\}\right) \\ &= 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|B\|_{\mathrm{F}}^2}, \frac{t^{1/2}}{K\|B\|^{1/2}}\right\}\right).\end{aligned}$$

□

The next lemma is established in Bai and Silverstein [2009] and Chen and Pan [2012].

Lemma 6. *Suppose Z is an $n \times p$ matrix with i.i.d. Gaussian entries. Then for any positive constant $\epsilon > 0$, it holds that $n^{-1}Z^\top Z \leq (1+\epsilon)(1+\sqrt{c_n})^2$, w.p.a.1, for $c_n = p/n \in [0, \infty]$.*

Lemma 7 (Convexity)**.** *Let $O \subseteq \mathbb{R}^d$ be open and convex and D be a dense subset of O. For $\theta \in O$, both $M_n(\theta)$ and $M(\theta)$ are convex in θ. If $M_n(\theta) \xrightarrow{\mathrm{P}} M(\theta)$, for any $\theta \in D$, then $\sup_{\theta \in K} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0$, for any compact subset $K \subset O$.*

This lemma has been shown by Lemma 7.75 of Liese and Miescke [2008]. Next, we present a min-convergence theorem for functions defined on an open set $(0, \infty)$, as shown by Lemma 10 of Thrampoulidis et al. [2018].

Lemma 8. *Consider a sequence of proper, convex stochastic functions $M_n : \mathbb{R}^+ \to \mathbb{R}$, and a deterministic function $M : \mathbb{R}^+ \to \mathbb{R}$, satisfying (a) $M_n(x) \xrightarrow{\mathrm{P}} M(x)$, $\forall x > 0$; (b) there exists $z > 0$ such that $M(x) > \inf_{y>0} M(y)$, $\forall x \geq z$. Then we have $\inf_{x>0} M_n(x) \xrightarrow{\mathrm{P}} \inf_{x>0} M(x)$.*

Relatedly, we introduce a lemma for functions on a diverging sequence of closed sets.

Lemma 9. *Consider a sequence of closed intervals $\{[x_n, y_n]\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} x_n = -\infty$ and $\lim_{n\to\infty} y_n = +\infty$. Additionally, let there be a sequence of proper random and convex functions $M_n : [x_n, y_n] \to \mathbb{R}$, and a convex, continuous, and deterministic function $M : \mathbb{R} \to \mathbb{R}$ that satisfy: (a) $M_n(x) \xrightarrow{\mathrm{P}} M(x)$ for every $x \in \mathbb{R}$; (b) there exists $z > 0$ such that $M(x) > \inf_{y\in\mathbb{R}} M(y)$ holds for all $|x| \geq z$. Then it holds that $\inf_{x\in[x_n,y_n]} M_n(x) \xrightarrow{\mathrm{P}} \inf_{x\in\mathbb{R}} M(x)$.*

Proof. For n sufficiently large, $z \in [x_n, y_n]$. Assume $x^* \in [-z, z]$ minimizes $M(x)$. Condition (b) in fact implies that $x^\star \in (-z, z)$ and that $M(x^\star) = \inf_{x\in\mathbb{R}} M(x)$. Consider the event $\inf_{|x|>z, x\in[x_n,y_n]} M_n(x) < M_n(x^*)$. Under this event, there exists $|z_n| > z$ and $z_n \in [x_n, y_n]$ such that $M_n(z_n) < M_n(x^*)$. The geometry implies that there exists $\theta_n \in (0, 1)$, such

that either $z_n\theta_n + x^\star(1-\theta_n) = z$ or $z_n\theta_n + x^\star(1-\theta_n) = -z$ holds. Using convexity, we have $\min(M_n(z), M_n(-z)) \leq \theta_n M_n(z_n) + (1-\theta_n)M_n(x^*) < M_n(x^*)$. By taking limits on both sides, we have $\min(M(z), M(-z)) \leq M(x^*)$, which contradicts condition (b). Therefore, w.p.a.1, we have $\inf_{|x|>z, x\in[x_n,y_n]} M_n(x) \geq M_n(x^*)$. Furthermore, by Lemma 7, for all arbitrarily small $\epsilon > 0$, w.p.a.1, $\sup_{|x|\leq z} |M_n(x) - M(x)| < \epsilon$. In addition, by definition, there exists a sequence of z_n, such that $|z_n| \leq z$ and $\inf_{|x|\leq z} M_n(x) \geq M_n(z_n) - \epsilon$. Combining these two inequalities with the fact that $M(x^\star)$ minimizes M on $\mathbb{R}$ leads to $\inf_{|x|\leq z} M_n(x) \geq M_n(z_n) - \epsilon \geq M(z_n) - 2\epsilon \geq M(x^*) - 2\epsilon$, w.p.a.1. On the other hand, $\inf_{|x|\leq z} M_n(x) \leq M_n(x^*) \xrightarrow{\text{P}} M(x^*)$. Since ϵ is arbitrary, we have $\inf_{|x|\leq z} M_n(x) \xrightarrow{\text{P}} M(x^*)$. Along with $\inf_{|x|>z, x\in[x_n,y_n]} M_n(x) \geq M_n(x^*)$ and $M_n(x^\star) \xrightarrow{\text{P}} M(x^*)$ by (a), we have

$$\inf_{x\in[x_n,y_n]} M_n(x) = \min\Big(\inf_{|x|\leq z} M_n(x), \inf_{|x|>z, x\in[x_n,y_n]} M_n(x)\Big) \xrightarrow{\text{P}} M(x^*). \qquad \square$$

Lemma 10. *Suppose X is a standard Gaussian random variable, then for $x > 0$,*

$$\frac{\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{x^2}{2}\right)\left(2x^{-3} - 12x^{-5} - 15x^{-7}\right) \leq \mathbb{E}(|X| - x)_+^2 \leq \frac{\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{x^2}{2}\right)\left(2x^{-3} + 3x^{-5}\right).$$

Proof. With integration by parts, we find

$$\mathbb{E}(|X| - x)_+^2 = \sqrt{\frac{2}{\pi}} \int_x^\infty (t-x)^2 \exp\left(-\frac{t^2}{2}\right) dt = \sqrt{\frac{2}{\pi}}\left(-x\exp\left(-\frac{x^2}{2}\right) + (x^2+1) f_G(x)\right),$$

where $f_G(x) = \int_x^\infty \exp(-t^2/2)dt$. Lemma 10 then follows from the tail inequality:

$$\exp\left(-\frac{x^2}{2}\right)\left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7}\right) \leq f_G(x) \leq \exp\left(-\frac{x^2}{2}\right)\left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}\right). \qquad \square$$

Lemma 11. *Given that X is a standard Gaussian random variable, the following inequalities*

hold when $x > 0$ and x is sufficiently large:

$$\mathbb{E}|X|(|X|-x)_+^2 \leq 2x\mathbb{E}(|X|-x)_+^2 \quad \textit{and} \quad \mathbb{E}X^2(|X|-x)_+^2 \leq 2x^2\mathbb{E}(|X|-x)_+^2.$$

Proof. The proof is analogous to that of Lemma 10 and is therefore omitted. □

Definition 2. *A centered random variable X belongs to the sub-exponential class $SE\left(\nu^2, \alpha\right)$ with $\nu > 0$ and $\alpha > 0$, if $\mathbb{E}e^{\lambda X} \leq e^{\frac{\lambda^2\nu^2}{2}}$, for all λ such that $|\lambda| < \alpha^{-1}$.*

Lemma 12. *Let $\{x_k\}_{k=1}^{\infty}$ be a sequence of diverging positive numbers. Then as $p \to \infty$, we have w.p.a.1, $\|\Sigma_2 b_0\|_\infty < x_p q^{-1/2}\log(p)$ and $\|\Sigma_2^{1/2}h\|_\infty < x_p\sqrt{\log(p)}$, where Σ_2 and b_0 are defined in Assumptions 1 and 3, respectively, and $h \in \mathbb{R}^p$ is a standard Gaussian vector.*

Proof. We only present the proof for the first inequality, noting that the proof for the second inequality follows similarly. By definition, there exist $b_{1i} \sim B(1,q)$ and a sub-exponential random variable b_{2i} such that $b_{0,i} = q^{-1/2}b_{1i}b_{2i}$. Note that $b_{1i}b_{2i}$ is still sub-exponential. Without loss of generality, assume $q^{1/2}b_{0,i} = b_{1i}b_{2i} \in \mathrm{SE}(1,1)$.

Write the (i,j)-th element of Σ_2 as $\Sigma_{2,ij}$. By the properties of sub-exponential variables, we have $\left(\Sigma_2 q^{1/2}b_0\right)_i \in \mathrm{SE}\left(\sum_{j=1}^p \Sigma_{2,ij}^2, \max_j|\Sigma_{2,ij}|\right)$. Given that $\sum_{j=1}^p \Sigma_{2,ij}^2 = (\Sigma_2^2)_{i,i} \leq \lambda_1(\Sigma_2^2) = C_2^2$ and $\max_j|\Sigma_{2,ij}| \leq C_2$, we conclude that $\left(\Sigma_2 q^{1/2}b_0\right)_i \in \mathrm{SE}(C_2^2, C_2)$. The tail bound of sub-exponential variables yields $\mathrm{P}\left(|(\Sigma_2 q^{1/2}b_0)_i| > x_p\log(p)\right) \leq 2\exp\left(-\frac{x_p\log(p)}{2C_2}\right)$. The conclusion follows by union bound inequality and $2p\exp\left(-\frac{x_p\log(p)}{2C_2}\right) \to 0$. □

Lemma 13. *For any given M_1, it holds that $\lim_{n\to\infty} S_{1n} = 0$ for S_{1n} defined in Eq. (1.21).*

Proof. Write $\tilde{x}_k = (\Sigma_\varepsilon^{-1/2}X_{\cdot,i})_k$ and $\tilde{y}_k = \beta_{0,i}\tilde{x}_k + z_k$ for $k = 1,\ldots,n$. By definition, we

have

$$\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i) = \frac{\int b\mathbf{1}_{q^{1/2}b\in[M_1,M_2]}\exp\left(-\sum_{k=1}^{n}\left(\tilde{y}_k - p^{-1/2}\tau^{1/2}\tilde{x}_k b\right)^2/2\right)dF(b)}{\int \exp\left(-\sum_{k=1}^{n}\left(\tilde{y}_k - p^{-1/2}\tau^{1/2}\tilde{x}_k b\right)^2/2\right)dF(b)}$$
$$= \frac{\int b\mathbf{1}_{q^{1/2}b\in[M_1,M_2]}\exp\left(-p^{-1}\tau b^2\sum_{k=1}^{n}\tilde{x}_k^2/2 + bp^{-1/2}\tau^{1/2}\sum_{k=1}^{n}\tilde{y}_k\tilde{x}_k\right)dF(b)}{\int \exp\left(-p^{-1}\tau b^2\sum_{k=1}^{n}\tilde{x}_k^2/2 + bp^{-1/2}\tau^{1/2}\sum_{k=1}^{n}\tilde{y}_k\tilde{x}_k\right)dF(b)} := \frac{Q_{1n}}{Q_{2n}},$$

where F is the distribution function of $b_{0,i}$. By the facts that $\int b\mathbf{1}_{q^{1/2}b\in[M_1,M_2]}dF(b) = 0$ and $dF(b) = (1-q)\delta_0 + qdF_{b_2}(q^{1/2}b)$, we have $|Q_{1n}|$ equals

$$\left|\int qb\mathbf{1}_{q^{1/2}b\in[M_1,M_2]}\left[\exp\left(-p^{-1}\tau b^2\sum_{k=1}^{n}\tilde{x}_k^2/2 + bp^{-1/2}\tau^{1/2}\sum_{k=1}^{n}\tilde{y}_k\tilde{x}_k\right) - 1\right]dF_{b_2}(q^{1/2}b)\right|$$
$$\leq q^{1/2}\tilde{M}\int\left|\exp\left(-p^{-1}\tau q^{-1}\tilde{b}^2\sum_{k=1}^{n}\tilde{x}_k^2/2 + \tilde{b}q^{-1/2}p^{-1/2}\tau^{1/2}\sum_{k=1}^{n}\tilde{y}_k\tilde{x}_k\right) - 1\right|dF_{b_2}(\tilde{b}),$$

where $\tilde{M} := \max(|M_1|, |M_2|)$. Define the event

$$A_n := \left\{\left|p^{-1/2}\tau^{1/2}\sum_{k=1}^{n}\tilde{y}_k\tilde{x}_k\right| \leq \tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p) \text{ and } p^{-1}\tau\sum_{k=1}^{n}\tilde{x}_k^2 \leq \tilde{C}p^{-1}n\tau\right\}, \tag{1.47}$$

where $\tilde{C} := 5C_1C_2c_\varepsilon^{-1}$. Under this event, we observe that

$$\left|\exp\left(-p^{-1}\tau q^{-1}\tilde{b}^2\sum_{k=1}^{n}\tilde{x}_k^2/2 + \tilde{b}q^{-1/2}p^{-1/2}\tau^{1/2}\sum_{k=1}^{n}\tilde{y}_k\tilde{x}_k\right) - 1\right|$$
$$\leq \exp\left(\tilde{C}|\tilde{b}|p^{-1/2}\tau^{1/2}n^{1/2}q^{-1/2}\log^2(p)\right)$$
$$- \exp\left(-\tilde{C}\tilde{b}^2p^{-1}n\tau q^{-1} - \tilde{C}|\tilde{b}|p^{-1/2}\tau^{1/2}n^{1/2}q^{-1/2}\log^2(p)\right).$$

Since $p^{-1/2}\tau^{1/2}n^{1/2}q^{-1/2}\log^2(p) \to 0$ and $p^{-1}n\tau q^{-1} \to 0$ by Assumption 4, and given that F_{b_2} follows a sub-exponential distribution, the integral of both terms on the right-hand-side converges to zero as $n \to \infty$. Therefore, for any $\epsilon > 0$, there exists n_0 such that for all $n > n_0$, we have $|Q_{1n}| \leq \epsilon$ under the event A_n. Similarly, it can be proven that there exists n_1 such that for all $n > n_1$, we obtain $Q_{2n} \geq 1/2$ under the event A_n.

Next we analyze the event A_n. We start with the second inequality in A_n. Note that

$$\begin{aligned}\sum_{k=1}^{n} \tilde{x}_k^2 =& X_{\cdot,i}^\top \Sigma_\varepsilon^{-1} X_{\cdot,i} \leq c_\varepsilon^{-1} X_{\cdot,i}^\top X_{\cdot,i} = c_\varepsilon^{-1} e_i^\top \Sigma_2^{1/2} Z^\top \Sigma_1 Z \Sigma_2^{1/2} e_i \\ \leq& c_\varepsilon^{-1} C_1 \|Z\Sigma_2^{1/2} e_i\|^2 \overset{d}{=} c_\varepsilon^{-1} C_1 \|\Sigma_2^{1/2} e_i\|^2 \chi^2(n) \leq c_\varepsilon^{-1} C_1 C_2 \chi^2(n),\end{aligned}$$

where e_i is the i-th standard basis vector. By Lemma 5, with probability at least $1 - 2\exp(-cp)$ for some fixed constant $c > 0$, $\chi^2(n) \leq 5n$, which implies the second inequality.

For the first inequality in A_n, using the second inequality, we observe that,

$$\begin{aligned}p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n} \tilde{x}_k \tilde{y}_k\Big| &= p^{-1/2}\tau^{1/2}\Big|\beta_{0,i}\sum_{k=1}^{n} \tilde{x}_k^2 + \sum_{k=1}^{n} \tilde{x}_k z_k\Big| \\ &\leq \tilde{C}p^{-1}q^{-1/2}\tau n|b_{2i}| + p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n} \tilde{x}_k z_k\Big|.\end{aligned}$$

Using the property of a sub-exponential random variable, for some constant $c > 0$, with probability at least $1 - 2\exp(-c\log^2(p))$, we have $|b_2| \leq \log^2(p)/2$, which implies

$$\tilde{C}p^{-1}q^{-1/2}\tau n|b_{2i}| \leq \tilde{C}p^{-1}q^{-1/2}\tau n\log^2(p)/2 = o(\tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p)/2)$$

by Assumption 4. In addition, by Lemma 5, with probability at least $1 - 2\exp(-c\log^2(p))$, we have $\Big|\sum_{k=1}^{n} \tilde{x}_k z_k\Big| \leq \tilde{C}n^{1/2}\log^2(p)/2$, which implies

$$p^{-1/2}\tau^{1/2}\Big|\sum_{k=1}^{n} \tilde{x}_k z_k\Big| \leq \tilde{C}p^{-1/2}\tau^{1/2}n^{1/2}\log^2(p)/2.$$

In sum, using the facts that $\max(\exp(-cp), \exp(-c\log^2(p))) = o(p^{-1})$, we conclude that with probability at least $1 - p^{-1}$ as $p \to \infty$, A_n holds. Hence we have

$$\begin{aligned}\lim_{n\to\infty} S_{1n} &= \lim_{n\to\infty} \mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i)\big)^2\mathbf{1}_{A_n} + \lim_{n\to\infty} \mathbb{E}\big(\mathbb{E}(b_{0,i}\mathbf{1}_{q^{1/2}b_{0,i}\in[M_1,M_2]}|\mathcal{G}_i)\big)^2\mathbf{1}_{A_n^c} \\ &\leq 4\epsilon^2 + \lim_{n\to\infty} q^{-1}\tilde{M}^2\mathrm{P}(A_n^c) \leq 4\epsilon^2 + \lim_{n\to\infty} p^{-1}q^{-1}\tilde{M}^2 = 4\epsilon^2.\end{aligned}$$

The conclusion then follows from the arbitrariness of ϵ. □

Lemma 14. *The objective function in Eq. (1.27) is convex with respect to α and jointly concave with respect to (δ, γ). Additionally, as long as Eq. (1.29) holds, we have Eq. (1.26).*

Proof. By Lemma 15, it suffices to prove Eq. (1.26) holds for $\hat{w}^B$, which equals

$$\arg\min_{w\in S_w^n} \frac{c_n}{n}\left\|\tau^{1/2}\Sigma_1^{1/2}Zw - \tau^{-1}\varepsilon\right\|^2 + c_n^2\lambda\left\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\right\|^2 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi, \tag{1.48}$$

and $S_w^n = \{w\big|c_n\tau^{-1}\sigma_x\sigma_\beta - K_\alpha \leq c_n\|w\| \leq c_n\tau^{-1}\sigma_x\sigma_\beta + K_\alpha\}$ for some sufficiently large K_α. With slight abuse of notation, we refer to the optimal solution $\hat{w}$ instead of using $\hat{w}^B$.

Note that for any vector x, $\|x\|^2 = \max_u \sqrt{n}u^\top x - n\|u\|^2/4$, where its argmax is $2x/\sqrt{n}$, and similarly $\|x\|^2 = \max_v v^\top x - \|v\|^2/4$. Applying these equalities to $\|\tau^{1/2}\Sigma_1^{1/2}Zw - \tau^{-1}\varepsilon\|^2$ and $\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\|^2$, setting $\tilde{u} = \Sigma_1^{1/2}u$, and $\tilde{v} = \Sigma_2^{-1/2}v$, we can rewrite (1.48) as

$$\begin{aligned}\min_{w\in S_w^n}\max_{\tilde{u},\tilde{v}} &\frac{c_n\tau^{1/2}}{\sqrt{n}}\tilde{u}^\top Zw - \frac{c_n\tau^{-1}}{\sqrt{n}}\tilde{u}^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}\tilde{u}\|^2}{4} + c_n^2\lambda\tilde{v}^\top w + c_n^2\lambda\tau^{-3/2}\tilde{v}^\top\Sigma_2^{1/2}\beta_0 \\ &- \frac{c_n^2\lambda\|\Sigma_2^{1/2}\tilde{v}\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.\end{aligned} \tag{1.49}$$

To simplify notation and without ambiguity, we continue using u and v in place of $\tilde{u}$ and $\tilde{v}$.

For a given w, the argmax of Eq. (1.49), denoted by $\hat{u}$, is equal to $\frac{2}{\sqrt{n}}(\tau^{1/2}\Sigma_1 Zw - \tau^{-1}\Sigma_1^{1/2}\varepsilon)$. Given the definition of S_w^n and Assumptions 1 and 2, we have $\|w\| \leq \tau^{-1}\sigma_x\sigma_\beta +$

$c_n^{-1}K_\alpha$, $\|\Sigma_1\| \leq C_1$, $\|\Sigma_\varepsilon\| \leq C_\varepsilon$. Furthermore, w.p.a. 1, $\|z\| \leq \sqrt{2n}$ by the law of large numbers, which implies $\|\varepsilon\| \leq \sqrt{2C_\varepsilon n}$. Together with Lemma 6 and that $\tau c_n \to 0$ by Assumption 4, we have the following upper bound for $\|\hat{u}\|$ as n is large enough: $\|\hat{u}\| \leq \frac{2\tau^{1/2}}{\sqrt{n}}\|\Sigma_1 Z w\| + \frac{2}{\sqrt{n}}\|\tau^{-1}\Sigma_1^{1/2}\varepsilon\| \leq 4\tau^{-1}\sqrt{C_1 C_\varepsilon}$. Let $S_u^n = \{u \big| \|u\| \leq 4\tau^{-1}\sqrt{C_1 C_\varepsilon}\}$. Based on the above result, w.a.p.1, the following optimization problem is equivalent to (1.49):

$$\min_{w\in S_w^n} \max_{\substack{u\in S_u^n \\ v}} \frac{c_n\tau^{1/2}}{\sqrt{n}} u^\top Z w - \frac{c_n\tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} + c_n^2\lambda v^\top w + c_n^2\lambda\tau^{-3/2} v^\top \Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi. \tag{1.50}$$

Next, we need introduce an auxiliary problem for the purpose of applying CGMT:

$$\begin{aligned} \phi(g,h) =& \max_{\substack{0\leq\delta\leq 4\tau^{-1}\sqrt{C_1C_\varepsilon} \\ v}} \min_{w\in S_w^n} \max_{\|u\|=\delta} \mathcal{R}_n(w,v,u), \quad \text{where} \\ \mathcal{R}_n(w,v,u) =& \frac{c_n\tau^{1/2}}{\sqrt{n}}\|w\| g^\top u - \frac{c_n\tau^{1/2}}{\sqrt{n}}\delta h^\top w - \frac{c_n\tau^{-1}}{\sqrt{n}} u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} \\ &+ c_n^2\lambda v^\top w + c_n^2\lambda\tau^{-3/2} v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi, \end{aligned} \tag{1.51}$$

and $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ are standard Gaussian vectors, independent of the other random variables. Similarly, let $\tilde{\mathcal{S}}_n := \{w \big| |c_n\|w\| - c_n\tau^{-1}\sigma_x\sigma_\beta - \alpha_2^*| < \epsilon\}$, define $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$ as the optimal value of an analogous optimization problem to (1.51), with w restricted to $S_w^n \cap \tilde{\mathcal{S}}_n^c$.

Lemma 16 characterizes the limiting behavior of the optimal solution to (1.50), $\hat{w}$, and in turn, proves the desired (1.26), under conditions pertaining to the optimization problem (1.51). Therefore, we only need show that conditions outlined in Lemma 16 hold as long as (1.29) holds. That is, under (1.29), we need to prove the existence of the constants $\bar{\phi} < \bar{\phi}_{\tilde{\mathcal{S}}_n^c}$ such that for all $\eta > 0$, w.p.a.1 in the limit of $n \to \infty$, $\phi(g,h) < \bar{\phi} + \eta$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h) > \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \eta$.

Let $\bar{u} = u/\delta$, maximizing part of $\mathcal{R}_n(w, v, u)$ over u simplifies to the following problem:

$$\begin{aligned}
&\max_{\|u\|=\delta} \frac{c_n \tau^{1/2}}{\sqrt{n}} \|w\| g^\top u - \frac{c_n \tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2} \varepsilon - \frac{c_n \|\Sigma_1^{-1/2} u\|^2}{4} \\
&= \max_{\|\bar{u}\|=1} \frac{c_n \delta}{\sqrt{n}} (\tau^{1/2} \|w\| g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)^\top \bar{u} - \frac{c_n \delta^2}{4} \bar{u}^\top \Sigma_1^{-1} \bar{u}.
\end{aligned}$$

The latter is a quadratic programming problem, which has been studied in Gander et al. [1989]. The optimal value associated with this problem is given by:

$$-\frac{c_n \delta^2}{4} \mu_n(\alpha, \delta) + \frac{c_n}{n} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)^\top (\Sigma_1^{-1} - \mu_n(\alpha, \delta) \mathbb{I})^{-1} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon) \tag{1.52}$$

where $\alpha := \|w\|$ and $\mu_n(\alpha, \delta)$ is the solution to

$$\frac{1}{n} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)^\top (\Sigma_1^{-1} - \mu_n(\alpha, \delta) \mathbb{I})^{-2} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon) - \frac{\delta^2}{4} = 0, \tag{1.53}$$

under the condition $\Sigma_1^{-1} - \mu_n(\alpha, \delta) \mathbb{I}$ is positive semidefinite. With this, Eq. (1.51) equals:

$$\begin{aligned}
\max_{\substack{0 \le \delta \le 4\tau^{-1} \sqrt{C_1 C_\varepsilon} \\ v}} \min_{w \in S_w^n} &- \frac{c_n \delta^2}{4} \mu_n(\alpha, \delta) + \frac{c_n}{n} (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon)^\top (\Sigma_1^{-1} - \mu_n(\alpha, \delta) \mathbb{I})^{-1} \\
&\times (\tau^{1/2} \alpha g - \tau^{-1} \Sigma_1^{-1/2} \varepsilon) - \frac{c_n \tau^{1/2}}{\sqrt{n}} \delta h^\top w + c_n^2 \lambda v^\top w \\
&+ c_n^2 \lambda \tau^{-3/2} v^\top \Sigma_2^{1/2} \beta_0 - \frac{c_n^2 \lambda \|\Sigma_2^{1/2} v\|^2}{4} - \frac{c_n \tau^{-2}}{n} \|\varepsilon\|^2 - C_n^\phi.
\end{aligned}$$

Solving the inside minimization problem with respect to w/α while fixing α leads to

$$\max_{0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}} \min_{|c_n\alpha - c_n\tau^{-1}\sigma_x\sigma_\beta|\le K_\alpha} -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top$$
$$(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - c_n\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|\alpha \tag{1.54}$$
$$+ c_n^2\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.$$

By Lemma 17, the objective function of the above optimization is convex in α and jointly concave in (δ, v). As a result, we can switch the order of min and max by Corollary 3.3 in Sion [1958]. Also, note that for any vector x, $\|x\| = \min_{\gamma>0}\frac{1}{2\gamma}\|x\|^2 + \frac{\gamma}{2}$. Applying this equation to $\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|\alpha$, Eq. (1.54) becomes

$$\min_{c_n|\alpha-\tau^{-1}\sigma_x\sigma_\beta|\le K_\alpha} \max_{\substack{\gamma>0\\ 0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}}} \max_v -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top$$
$$\times(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) - \frac{c_n\gamma}{2} - \frac{c_n\alpha^2}{2\gamma}\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|^2$$
$$+ c_n^2\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \frac{c_n^2\lambda\|\Sigma_2^{1/2}v\|^2}{4} - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.$$

Note that the objective function above is jointly concave in (δ,γ,v). To see why this is true, it is sufficient to prove that $-\frac{\alpha^2}{2\gamma}\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|^2$ is jointly concave in (δ,γ,v), which follows by Lemma 13 in Thrampoulidis et al. [2018]. Consequently, after solving the first maximization problem over v, the resulting function remains jointly concave in (δ,γ). Maximizing over v is again a standard quadratic programming problem, which leads to Eq. (1.27). Thus, we conclude that (1.27) is convex with respect to α and jointly concave with respect to (δ,γ).

For any compact set A, define $\phi_A(g,h) := \min_{\alpha_2\in A}\max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}} \tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$. Based on

the above argument and condition (i) in (1.29), we can deduce

$$\phi(g,h) = \phi_{[-K_\alpha,K_\alpha]}(g,h) \xrightarrow{\text{P}} \min_{\alpha_2\in[-K_\alpha,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1),$$

$$\begin{aligned}\phi_{\tilde{\mathcal{S}}_n^c}(g,h) &= \min\{\phi_{[-K_\alpha,\alpha_2^*-\epsilon]}(g,h), \phi_{[\alpha_2^*+\epsilon,K_\alpha]}(g,h)\}\\ &\xrightarrow{\text{P}} \min_{\alpha_2\in[-K_\alpha,\alpha_2^*-\epsilon]\cup[\alpha_2^*+\epsilon,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1).\end{aligned}$$

Together with condition (ii), the conditions in Lemma 16 are satisfied by letting $\bar{\phi} = \min_{\alpha_2\in[-K_\alpha,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1)$ and $\bar{\phi}_{\tilde{\mathcal{S}}_n^c} = \min_{\alpha_2\in[-K_\alpha,\alpha_2^*-\epsilon]\cup[\alpha_2^*+\epsilon,K_\alpha]} \max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2,\delta_3,\gamma_1)$. □

We now introduce two lemmas whose proofs follow the same reasoning as those of Lemma 5 and Lemma 7 in Thrampoulidis et al. [2018], and are therefore omitted here.

Lemma 15. *Under the conditions of Theorem 2, define* $S_w^n := \{w \big| c_n\tau^{-1}\sigma_x\sigma_\beta - K_\alpha \le c_n\|w\| \le c_n\tau^{-1}\sigma_x\sigma_\beta + K_\alpha\}$ *for some* K_α *such that* $|\alpha_2^*| < K_\alpha$. *If the solution* $\hat{w}^B$ *to*

$$\arg\min_{w\in S_w^n} \frac{c_n}{n}\left\|\tau^{1/2}\Sigma_1^{1/2}Zw - \tau^{-1}\varepsilon\right\|^2 + c_n^2\lambda\left\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\right\|^2 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi$$

satisfies $c_n\|\hat{w}^B\| - c_n\tau^{-1}\sigma_x\sigma_\beta \to \alpha_2^*$, *then the same holds true for* $\hat{w}$ *of Eq. (1.25).*

Lemma 16. *Let* $\hat{w}$ *denote an optimal solution of Eq. (1.50). Regarding* $\phi(g,h)$ *and* $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$, *as introduced and discussed in relation to Eq. (1.51), suppose there are constants* $\bar{\phi}$ *and* $\bar{\phi}_{\tilde{\mathcal{S}}_n^c}$ *with* $\bar{\phi} < \bar{\phi}_{\tilde{\mathcal{S}}_n^c}$, *such that for all* $\eta > 0$, *the following hold w.a.p.1 as* $n \to \infty$: *(a)* $\phi(g,h) < \bar{\phi} + \eta$, *(b)* $\phi_{\tilde{\mathcal{S}}_n^c}(g,h) > \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \eta$. *Under these conditions, we have* $\hat{w} \in \tilde{\mathcal{S}}_n$ *w.p.a.1.*

Lemma 17. *The objective function of Eq. (1.54) is convex in* α *and jointly concave in* (δ, v).

Proof. First, we prove the objective function is convex in $\alpha = \|w\|$. Revisiting the term $f(\alpha,u) := \frac{c_n\tau^{1/2}}{\sqrt{n}}\alpha g^\top u - \frac{c_n\tau^{-1}}{\sqrt{n}}u^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4}$ in Eq. (1.51), we observe that it is convex in α. After maximizing over the direction of u, the term remains convex

in α since $\max_{\|u\|=\delta} f(\theta\alpha_1 + (1-\theta)\alpha_2, u) \leq \max_{\|u\|=\delta}\{\theta f(\alpha_1, u) + (1-\theta) f(\alpha_2, u)\} \leq \theta \max_{\|u\|=\delta} f(\alpha_1, u) + (1-\theta)\max_{\|u\|=\delta} f(\alpha_2, u)$, for $\theta \in (0,1)$. Note that from Eq. (1.52), $\max_{\|u\|=\delta} f(\alpha, u) = -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)$, which yield the first two terms in Eq. (1.54). Meanwhile, the term $-\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|\alpha$ is also convex in α. Consequently, we deduce that the objective function of Eq. (1.54) is convex in α.

Next, we demonstrate that this function is jointly concave in (δ, v). It is easy to verify that $-\|c_n\lambda v - n^{-1/2}\tau^{1/2}\delta h\|\alpha$ is jointly concave in (δ, v), since $\alpha \geq 0$. Moreover, $\lambda\tau^{-3/2}v^\top\Sigma_2^{1/2}\beta_0 - \lambda\|\Sigma_2^{1/2}v\|^2/4$ is concave in v. Therefore, it suffices to prove

$$-\frac{\delta^2}{4}\mu_n(\alpha,\delta) + \frac{1}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) \tag{1.55}$$

is concave in δ. Let the eigenvalues and normalized eigenvectors of Σ_1 be $\{(\lambda_i, v_i)\}_{i=1}^n$, and let $w_i = (\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top v_i$, for $i = 1, 2, \ldots, n$. Then (1.55) equals $-\frac{\delta^2}{4}\mu_n(\alpha,\delta) + \frac{1}{n}\sum_{i=1}^n \frac{1}{1/\lambda_i - \mu_n(\alpha,\delta)} w_i^2$. The first order derivative of this equation with respect to δ is

$$-\frac{\delta}{2}\mu_n(\alpha,\delta) - \frac{\delta^2}{4}\partial_\delta\mu_n(\alpha,\delta) + \frac{\partial_\delta\mu_n(\alpha,\delta)}{n}\sum_{i=1}^n \frac{1}{(1/\lambda_i - \mu_n(\alpha,\delta))^2} w_i^2 = -\frac{\delta}{2}\mu_n(\alpha,\delta), \tag{1.56}$$

where the last equation follows from the definition of the function $\mu_n(\alpha,\delta)$:

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{(1/\lambda_i - \mu_n(\alpha,\delta))^2} w_i^2 = \frac{\delta^2}{4}. \tag{1.57}$$

Further, the second-order derivative with respect to δ can be calculated as: $-\frac{1}{2}\mu_n(\alpha,\delta) - \frac{\delta}{2}\partial_\delta\mu_n(\alpha,\delta)$. By the chain rule of differentiation, $\partial_\delta\mu_n(\alpha,\delta)$ is the reciprocal of $\partial\mu_n\delta$. The

latter can be calculated directly using the definition of μ_n via Eq. (1.57):

$$\partial_{\mu_n}\delta = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^2}w_i^2\right)^{-1/2}\cdot\frac{2}{n}\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^3}w_i^2.$$

With this, we can write the second-order derivative as follows:

$$-\frac{1}{2}\mu_n(\alpha,\delta)-\frac{\delta}{2}\partial_\delta\mu_n(\alpha,\delta) = -\frac{1}{2}\mu_n-\frac{1}{2}\cdot\frac{\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^2}w_i^2}{\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^3}w_i^2} = -\frac{1}{2}\cdot\frac{\sum_{i=1}^{n}\frac{1/\lambda_i}{(1/\lambda_i-\mu_n)^3}w_i^2}{\sum_{i=1}^{n}\frac{1}{(1/\lambda_i-\mu_n)^3}w_i^2}.$$

Since $\Sigma_1^{-1}-\mu_n\mathbb{I}$ is positive semidefinite, the right-hand-side is no larger than zero, which concludes the proof. □

Lemma 18. *For $\tilde{Q}_n = \tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$ in Eq. (1.28), Eq. (1.29) holds.*

Proof. The notation below is defined in the proof of Theorem 2. Let $\delta_2 = \delta_2^* + c_n^{-1/2}\delta_3$. First, we demonstrate that $c_n\tau^{-1}\mu_n(\alpha,\delta) - c_n\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2) \xrightarrow{\mathrm{P}} 0$. Let $f(x) := \frac{1}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-x\mathbb{I})^{-2}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)$. Recall that $\mu_n(\alpha,\delta)$ is the solution to $f(x)=\delta^2/4$. Note that $f(x)$ exhibits a monotonic increase in x when $x \le 1/C_1$. Therefore, it suffices to show that, given any arbitrarily small $\epsilon>0$, w.p.a.1, the following inequalities hold: $c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon) - c_n\delta^2\tau/4 > c_+ > 0$ and $c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2) - \tau c_n^{-1}\epsilon) - c_n\delta^2\tau/4 < c_- < 0$, for some constants c_+ and c_-.

By Lemmas 2 and 3, we can deduce the following equations:

$$
\begin{aligned}
&\frac{c_n\tau^{-1}}{n}\varepsilon^\top\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\varepsilon\\
&\quad-\frac{c_n\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]\\
&=O_{\mathrm{P}}(c_n\tau^{-1}n^{-1/2}),\\
&\frac{c_n\tau^2}{n}\alpha^2 g^\top\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}g\\
&\quad-\frac{c_n\alpha^2\tau^2}{n}\operatorname{Tr}\left[\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\right]=O_{\mathrm{P}}(c_n n^{-1/2}),\\
&\frac{c_n\tau^{1/2}\alpha}{n}\varepsilon\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}g=O_{\mathrm{P}}(c_n\tau^{-1/2}n^{-1/2}).
\end{aligned}
$$

Therefore, using the definition of $f(\cdot)$ we can deduce that:

$$
\begin{aligned}
&c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon)\\
&\quad-\frac{c_n\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]\\
&\quad-\frac{c_n\alpha^2\tau^2}{n}\operatorname{Tr}\left[\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\right]=O_{\mathrm{P}}(c_n\tau^{-1}n^{-1/2})=o_{\mathrm{P}}(1).
\end{aligned}
\tag{1.58}
$$

Note that for sufficiently small x such that $x\|\Sigma_1\|<1$,

$$
\begin{aligned}
&\frac{\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-x\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}-\Sigma_\varepsilon^{1/2}\Sigma_1^{1/2}\left(\mathbb{I}+2x\Sigma_1\right)\Sigma_1^{1/2}\Sigma_\varepsilon^{1/2}\right]\\
=&\frac{\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{1/2}\left(\mathbb{I}-x\Sigma_1\right)^{-2}\Sigma_1^{1/2}\Sigma_\varepsilon^{1/2}-\Sigma_\varepsilon^{1/2}\Sigma_1^{1/2}\left(\mathbb{I}+2x\Sigma_1\right)\Sigma_1^{1/2}\Sigma_\varepsilon^{1/2}\right]\\
\le&\tau^{-1}C_1C_\varepsilon\|\left(\mathbb{I}-x\Sigma_1\right)^{-2}-\left(\mathbb{I}+2x\Sigma_1\right)\|\lesssim\tau^{-1}x^2,
\end{aligned}
$$

where we apply Lemma 4 in the last inequality. As a consequence, we have:

$$\begin{aligned}
&\frac{\tau^{-1}}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]\\
=&\frac{1}{n}\operatorname{Tr}\left[\Sigma_\varepsilon^{1/2}\Sigma_1^{-1/2}\left(\tau^{-1}\Sigma_1^2+2\left(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon\right)\Sigma_1^3\right)\Sigma_1^{-1/2}\Sigma_\varepsilon^{1/2}\right]+O(\tau)\\
=&\tau^{-1}\sigma_\varepsilon^2\theta_1+2(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon)\sigma_\varepsilon^2\theta_3+O(\tau)+o(c_n^{-1}),
\end{aligned}\tag{1.59}$$

where the last equation follows by Assumption 5. By the same argument, it follows that:

$$\frac{\alpha^2\tau^2}{n}\operatorname{Tr}\left[\left(\Sigma_1^{-1}-\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\mathbb{I}-\tau c_n^{-1}\epsilon\mathbb{I}\right)^{-2}\right]=\sigma_x^2\sigma_\beta^2\theta_4+O(\tau)+o(c_n^{-1}).$$

Applying the above estimates to the left-hand-side of (1.58), we can deduce that:

$$\begin{aligned}
&c_n\tau f(\tau\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+\tau c_n^{-1}\epsilon)-\frac{c_n\delta^2\tau}{4}\\
=&2c_n(\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+c_n^{-1}\epsilon)\sigma_\varepsilon^2\theta_3+c_n\sigma_x^2\sigma_\beta^2\theta_4-\frac{c_n\delta_1^*\delta_2}{2}+o_{\mathrm{P}}(1).
\end{aligned}$$

By the definition of $\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)$, the right-hand side of the above equation is positive w.p.a.1. The proof of the other inequality is similar. Hence, we have proved

$$c_n\tau^{-1}\mu_n(\alpha,\delta)-c_n\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)\xrightarrow{\mathrm{P}}0.\tag{1.60}$$

Next, to analyze $\tilde{Q}_n$, we first investigate the limiting behavior of:

$$-\frac{\delta^2}{4}\mu_n(\alpha,\delta)+\frac{1}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right).$$

By (1.60), we have $\|\mu_n(\alpha,\delta)\Sigma_1\| = O_{\mathrm{P}}(\tau)$. Applying Lemma 4 again, we deduce:

$$\begin{aligned}
&\| \left(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I}\right)^{-2} - \Sigma_1^2 - 2\mu_n(\alpha,\delta)\Sigma_1^3 - 3\mu_n^2(\alpha,\delta)\Sigma_1^4\| \lesssim_{\mathrm{P}} \tau^3 \\
&\| \left(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I}\right)^{-1} - \Sigma_1 - \mu_n(\alpha,\delta)\Sigma_1^2 - \mu_n^2(\alpha,\delta)\Sigma_1^3\| \lesssim_{\mathrm{P}} \tau^3.
\end{aligned}$$

Furthermore, by the fact that $\|\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\| = O_{\mathrm{P}}(n\tau^{-1})$ and Eq. (1.53), we have

$$\begin{aligned}
\frac{\delta^2}{4}\mu_n(\alpha,\delta) = &\frac{1}{n}(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\mu_n(\alpha,\delta)\Sigma_1^2 + 2\mu_n^2(\alpha,\delta)\Sigma_1^3)(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon) \\
&+ O_{\mathrm{P}}(\tau).
\end{aligned}$$

With a similar approach, we have

$$\begin{aligned}
&\frac{1}{n}\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top \left(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right) \\
&= \frac{1}{n}\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top \left(\Sigma_1 + \mu_n(\alpha,\delta)\Sigma_1^2 + \mu_n^2(\alpha,\delta)\Sigma_1^3\right)\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right) \\
&+ O_{\mathrm{P}}(\tau).
\end{aligned}$$

As a consequence, based on Lemmas 2 and 3, as well as the definition of α_2 and the fact that $c_n\tau^{-1}\mu_n(\alpha,\delta) - c_n\mu(\sigma_x\sigma_\beta,\delta_1^*,\delta_2) \xrightarrow{\mathrm{P}} 0$, we have:

$$\begin{aligned}
&-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta) + \frac{c_n}{n}\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top \left(\Sigma_1^{-1} - \mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right) \\
&= \frac{c_n}{n}\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top (\Sigma_1 - \mu_n^2(\alpha,\delta)\Sigma_1^3)\left(\tau^{1/2}\alpha g - \tau^{-1}\Sigma_1^{-1/2}\varepsilon\right) + O_{\mathrm{P}}(c_n\tau) \\
&= c_n\tau^{-1}\sigma_x^2\sigma_\beta^2 - c_n\sigma_\varepsilon^2\theta_3\mu^2(\sigma_x\sigma_\beta,\delta_1^*,\delta_2) + 2\sigma_x\sigma_\beta\alpha_2 + \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 + o_{\mathrm{P}}(1). \qquad (1.61)
\end{aligned}$$

Finally, we examine the remainder term that contributes to $\tilde{Q}_n$:

$$\frac{c_n^2\lambda^2}{4}\left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0+\frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right)^\top\left(\frac{\lambda}{4}\Sigma_2+\frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}\left(\tau^{-3/2}\Sigma_2^{1/2}\beta_0+\frac{\alpha^2\delta\tau^{1/2}}{\sqrt{n}\gamma}h\right)$$
$$-\frac{c_n\tau\alpha^2\delta^2}{2\gamma n}\|h\|^2.$$

Using Lemmas 2-3, $p^{1/2}\tau^{-1}n^{-1}q^{-1/2}=o(1)$ by Assumption 4, and the assumptions on Σ_2, this term converges in probability to:

$$\frac{c_n^2\lambda^2\tau^{-2}\sigma_\beta^2}{4p}\operatorname{Tr}\left[\Sigma_2^{1/2}\left(\frac{\lambda}{4}\Sigma_2+\frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}\Sigma_2^{1/2}\right]$$
$$+c_n\tau\operatorname{Tr}\left[\frac{c_n\lambda^2\alpha^4\delta^2}{4n\gamma_h^2}\left(\frac{\lambda}{4}\Sigma_2+\frac{c_n\alpha^2\lambda^2}{2\gamma}\mathbb{I}\right)^{-1}-\frac{\alpha^2\delta^2}{2\gamma n}\mathbb{I}\right]$$
$$=\frac{c_n\gamma n_1}{2}\tau^{-1}-\frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}-\tau^{-1}\frac{(\delta_1^*)^2\sigma_x^2c_n}{4\lambda}-\frac{c_n\sigma_x^2\delta_1^*\delta_2}{2\lambda}+\frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2+o_n(1),$$

where we apply Lemma 4 and the same argument in proving Eq. (1.59). Combining this estimate with (1.61) we conclude that

$$\tilde{Q}_n=-\frac{\delta_3^2\theta_1}{4\theta_3}+2\sigma_x\sigma_\beta\alpha_2-\frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}+\frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2+o_{\mathrm{P}}(1).$$

We now proceed to establish Claims (i) to (ii) in Eq. (1.29). Fix $\alpha_2\in A$ and $\gamma_1>0$, and observe that $\lim_{\delta_3\to\pm\infty}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)\to-\infty$. By the concave version of Lemma 9, we conclude that $\max_{\delta_3\in K_{\delta_3}}\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)\xrightarrow{\mathrm{P}}\max_{\delta_3\in\mathbb{R}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)$. Since $\tilde{Q}_n$ is jointly concave in (δ_3,γ_1), after maximizing with respect to δ_3, the function should remain concave in γ_1. Moreover, consider the following equation: $\max_{\delta_3\in\mathbb{R}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)=2\sigma_x\sigma_\beta\alpha_2-\frac{\gamma_1^2}{4\sigma_\beta^2\lambda}\theta_2-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}+\frac{(\delta_1^*)^2\gamma_1\sigma_x^2}{8\lambda^2\sigma_\beta^2}\theta_2$. As a result, $\lim_{\gamma_1\to\infty}\max_{\delta_3\in\mathbb{R}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)\to-\infty$. By Lemma 8, we conclude that $\max_{\substack{\gamma_1>0\\ \delta_3\in K_{\delta_3}}}\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)\xrightarrow{\mathrm{P}}\max_{\substack{\gamma_1>0\\ \delta_3\in\mathbb{R}}}\tilde{Q}(\alpha_2,\delta_3,\gamma_1)$. Since $\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$ is convex in α_2, it should retain its convexity in α_2 after being maximized with respect to δ_2

and γ_1. Since the above equation holds for any $\alpha_2 \in A$, by Lemma 7, we conclude that Claim (i) holds.

The first-order condition implies a unique solution: $\alpha_2^* := \arg\min_{\alpha_2} \max_{\substack{\gamma_1>0 \\ \delta_3\in\mathbb{R}}} \tilde{Q}(\alpha_2, \delta_3, \gamma_1)$, which is given by $\theta_2\sigma_x^3 \left(\frac{\sigma_\varepsilon^2\theta_1}{2\lambda^2\sigma_\beta} - \frac{\sigma_\beta}{\lambda} \right)$. Thus, Claim (ii) holds true, concluding the proof. □

Lemma 19. *Under the conditions of Theorem 3, there exists a constant $\tilde{c} > 0$ that depends solely on fixed constants, such that w.p.a.1, inequality (1.32) holds. In addition, as $n \to \infty$, for any given fixed $\lambda > 0$, Eq. (1.33) holds.*

Proof. Under the condition that $\lambda_n \geq \epsilon$, using (1.30) and (1.31), we have, w.p.a.1,

$$\begin{aligned}
&\|\hat{\beta}^i_{\lambda_n}\|^2 = \left\| \frac{1}{n} \left(\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n\lambda_n \mathbb{I} \right)^{-1} X_{(-i)} y_{(-i)} \right\|^2 \\
&\leq \frac{\|X_{(-i)} y_{(-i)}\|^2}{c_n^2\epsilon^2 n^2} \leq \frac{2C_2\sigma_\varepsilon^2(1+\sqrt{c_n})^2}{c_n^2\epsilon^2}, \\
&\|\hat{\beta}^i_{\lambda_1} - \hat{\beta}^i_{\lambda_2}\|^2 \\
&= c_n^2(\lambda_1-\lambda_2)^2 \left\| \frac{1}{n} \left(\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n\lambda_2 \right)^{-1} \left(\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n\lambda_1 \right)^{-1} X_{(-i)} y_{(-i)} \right\|^2 \\
&\quad \leq \frac{(\lambda_1-\lambda_2)^2}{n^2 c_n^2 \lambda_1^2 \lambda_2^2} \|X_{(-i)} y_{(-i)}\|^2 \leq \frac{2C_2\sigma_\varepsilon^2(1+\sqrt{c_n})^2(\lambda_1-\lambda_2)^2}{c_n^2\epsilon^4}. \qquad (1.62)
\end{aligned}$$

With the inequalities above and triangle inequalities, we obtain, w.a.p.1,

$$\begin{aligned}
&|\hat{R}^{K-CV}(\lambda_1) - \hat{R}^{K-CV}(\lambda_2)| = \frac{1}{n} \left| \sum_{i=1}^K \left(\|y_{(i)} - X_{(i)}\hat{\beta}^i_{\lambda_1}\|^2 - \|y_{(i)} - X_{(i)}\hat{\beta}^i_{\lambda_2}\|^2 \right) \right| \\
&\leq \frac{2}{n} \sum_{i=1}^K \|X_{(i)}\| \|\hat{\beta}^i_{\lambda_1} - \hat{\beta}^i_{\lambda_2}\| \left(\|y_{(i)}\| + \frac{2C_2^{1/2}\sigma_\varepsilon(1+\sqrt{c_n})}{c_n\epsilon} \|X_{(i)}\| \right) \leq \tilde{C}|\lambda_1 - \lambda_2|, \qquad (1.63)
\end{aligned}$$

where $\tilde{C}$ is some fixed constant.

Let us fix a constant $\tilde{c}$ such that the inequality $\frac{c_2^2\sigma_\varepsilon^2}{2K\tilde{c}^2} - 4C_2^2 \frac{(1+\sqrt{c_n})^2}{c_n\tilde{c}} > 100$ remains true as $n, p \to \infty$. This is possible because $(1+\sqrt{c_n})^2/c_n$ is bounded as $n, p \to \infty$. In addition, let

$S := \{\lambda_j = \epsilon + p^{-9}(j-1) : 1 \le j \le 1 + [p^9(\tilde{c}\tau^{-1} - \epsilon)]\}$. Given $\tau^{-1} = o(p)$, the cardinality of the set satisfies $|S| \le p^{10}$. By definition, for any $\lambda \in [\epsilon, \tilde{c}\tau^{-1}]$, there exists a $\lambda_{j^*} \in S$ such that $|\lambda - \lambda_{j^*}| \le p^{-9}$. By Eq. (1.63), we have $|\hat{R}^{K-CV}(\lambda) - \hat{R}^{K-CV}(\lambda_{j^*})| \le \tilde{C}|\lambda - \lambda_{j^*}| \le \tilde{C}p^{-9}$. Therefore, if we show that

$$\inf_{\lambda_j \in S} \left\{ \hat{R}^{K-CV}(\lambda_j) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 \right\} > np^{-1}\tau^2 \tag{1.64}$$

holds w.p.a.1, we have

$$\inf_{\lambda \in [\epsilon, \tilde{c}\tau^{-1}]} \left\{ \hat{R}^{K-CV}(\lambda_j) - \frac{1}{n}\|\varepsilon\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 \right\} > np^{-1}\tau^2 - \tilde{C}p^{-9} > \frac{np^{-1}\tau^2}{2}, \tag{1.65}$$

which implies Eq. (1.32). By the definition of $\hat{R}^{K-CV}$, it is easy to verify that we need prove:

$$\inf_{\lambda_j \in S} \left\{ n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - 2n^{-1}K\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0) - \|\Sigma_2^{1/2}\beta_0\|^2 \right\} > np^{-1}\tau^2$$

holds w.p.a.1 for all $i = 1, \ldots, K$. By the independence of $Z_{(i)}$ and $\hat{\beta}^i_{\lambda_j}$, the first term on the left-hand-side is distributed as: $n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 \stackrel{d}{=} n^{-1}K\chi^2\left(K^{-1}n\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2$, where $\chi^2\left(K^{-1}n\right)$ denotes a Chi-squared random variable with $K^{-1}n$ degrees of freedom. Consequently, we can deduce that:

$$\begin{aligned}
&\mathrm{P}\left(\left| n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 \right| \ge \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 \right) \\
&= \mathrm{P}\left(\left| n^{-1}K\chi^2\left(K^{-1}n\right) - 1 \right| \ge \frac{\log(p)}{\sqrt{n}} \right) \le 2\exp(-\tilde{c}_1 \log^2(p)),
\end{aligned}$$

where the last step uses Lemma 5, and $\tilde{c}_1$ is a fixed positive constant. Analogously, we have:

$$\mathrm{P}\left(\left| n^{-1}K\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0) \right| \ge \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\| \right) \le 2\exp(-\tilde{c}_2 \log^2(p)),$$

with $\tilde{c}_2$ being another fixed positive constant. For simplicity, we consolidate the constants $\tilde{c}_1$ and $\tilde{c}_2$ into a unified constant denoted as $\tilde{c}_1$. By the union bound inequality, we have that with probability exceeding $1 - 4p^{10}\exp(-\tilde{c}_1 \log^2(p))$, the following relation holds:

$$\begin{aligned}
&\inf_{\lambda_j \in S}\left\{ n^{-1}K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - 2n^{-1}K\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0) - \|\Sigma_2^{1/2}\beta_0\|^2 \right\} \\
&\quad \geq \inf_{\lambda_j \in S}\left\{ \left(1 - \frac{\log(p)}{\sqrt{n}}\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\| - \|\Sigma_2^{1/2}\beta_0\|^2 \right\}.
\end{aligned}$$

Assume for now that $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 \geq 50np^{-1}\tau^2$ holds. In this scenario, $\left(1 - \frac{\log(p)}{\sqrt{n}}\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|$ is monotonically increasing in $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|$ since $\log(p)/\sqrt{n} = o(\tau)$, hence it achieves its minimum when $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2 = 50np^{-1}\tau^2$. As a result, it can be shown that $\left(1 - \frac{\log(p)}{\sqrt{n}}\right)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \frac{\log(p)}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\| - \|\Sigma_2^{1/2}\beta_0\|^2 \geq np^{-1}\tau^2$. Therefore, we only need to prove $\inf_{\lambda_j \in S}\{\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2\} \geq 50np^{-1}\tau^2$ holds w.p.a.1.

We now establish a uniform lower bound for $\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j} - \beta_0)\|^2 - \|\Sigma_2^{1/2}\beta_0\|^2$, which can be written as: $\|\Sigma_2^{1/2}\hat{\beta}^i_{\lambda_j}\|^2 - 2\beta_0^\top\Sigma_2\hat{\beta}^i_{\lambda_j}$. By direct calculation, we have for each i,

$$\begin{aligned}
\|\Sigma_2^{1/2}\hat{\beta}^i_{\lambda_j}\|^2 &\geq c_2\|\hat{\beta}^i_{\lambda_j}\|^2 = c_2\left\|\frac{1}{n}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1}X_{(-i)}y_{(-i)}\right\|^2 \\
&\geq \frac{c_2}{n^2}\left\|\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right\|^{-2}\left\|X_{(-i)}y_{(-i)}\right\|^2 \\
&\geq \frac{c_2}{n^2}\left(C_2(1+\sqrt{c_n})^2 + c_n\lambda_j\right)^{-2}\|X_{(-i)}^\top y_{(-i)}\|^2.
\end{aligned}$$

Further, by Lemmas 2 and 3, we have

$$\|X_{(-i)}^\top y_{(-i)}\|^2 = \sigma_\varepsilon^2\operatorname{Tr}(X_{(-i)}X_{(-i)}^\top) + p^{-1}\tau\sigma_\beta^2\operatorname{Tr}(X_{(-i)}^\top X_{(-i)}X_{(-i)}^\top X_{(-i)}) + o_{\mathrm{P}}(n^{-1/2}).$$

By the fact that $\lambda_{\min}(A)\operatorname{Tr}(B) \leq \operatorname{Tr}(AB) \leq \lambda_{\max}(A)\operatorname{Tr}(B)$ when A, B are positive semidefinite, we have $c_2\operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top) \leq \operatorname{Tr}(X_{(-i)}X_{(-i)}^\top) \leq C_2\operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top)$, which, along with

$(np)^{-1}\operatorname{Tr}(Z_{(-i)}Z_{(-i)}^\top) \xrightarrow{\mathrm{P}} (K-1)/K$ and Eq. (1.30), imply that

$$p^{-1}\tau \operatorname{Tr}(X_{(-i)}^\top X_{(-i)} X_{(-i)}^\top X_{(-i)}) \le p^{-1}\tau \|X_{(-i)}^\top X_{(-i)}\| \operatorname{Tr}(X_{(-i)}^\top X_{(-i)}) \lesssim_{\mathrm{P}} \tau pn = o_{\mathrm{P}}(np).$$

Therefore, w.p.a.1, we obtain

$$\frac{c_2\sigma_\varepsilon^2 pn}{2K} \le \|X_{(-i)}^\top y_{(-i)}\|^2 \le 2C_2\sigma_\varepsilon^2 pn. \tag{1.66}$$

Consequently, uniformly over $\lambda_j \in S$, we deduce:

$$\|\Sigma_2^{1/2}\hat\beta_{\lambda_j}^i\|^2 \ge \frac{c_2^2\sigma_\varepsilon^2 p}{2nK}\left(C_2(1+\sqrt{c_n})^2 + c_n\lambda_j\right)^{-2}. \tag{1.67}$$

On the other hand, we have

$$\begin{aligned}|\beta_0^\top \Sigma_2 \hat\beta_{\lambda_j}^i| \le{}& \frac{1}{n}\left|\beta_0^\top\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\beta_0\right| \\ &+ \frac{1}{n}\left|\varepsilon_{(-i)}^\top X_{(-i)}\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1}\Sigma_2\beta_0\right| =: L_1 + L_2. \end{aligned}\tag{1.68}$$

To bound L_1, note that $\operatorname{Tr}(AB) \le \|AB\|\operatorname{rank}(AB) \le \|A\|\,\|B\|\operatorname{rank}(B)$, which implies

$$\begin{aligned}&\frac{1}{n}\operatorname{Tr}\left(\Sigma_2\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1} X_{(-i)}^\top X_{(-i)}\right) \\ &\le \|\Sigma_2\|\left\|\left(\frac{1}{n}X_{(-i)}^\top X_{(-i)} + c_n\lambda_j\mathbb{I}\right)^{-1}\frac{1}{n}X_{(-i)}^\top X_{(-i)}\right\|\operatorname{rank}(X_{(-i)}^\top X_{(-i)}) \le \frac{nC_2^2(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n\lambda_j},\end{aligned}$$

where the last inequality uses $\lambda_1((A+\mathbb{I})^{-1}A) = (\lambda_1(A)+1)^{-1}\lambda_1(A)$ and $n^{-1}\|X_{(-i)}^\top X_{(-i)}\| \le C_2(1+\sqrt{c_n})^2$. In addition, by Lemma 5 and the fact that the sub-exponential norm of $b_{0,i}$

is of order $O(q^{-1/2})$, we have, with probability exceeding $1 - 2p^{10}\exp(-\tilde{c}_1 \log^2(p))$,

$$\sup_{\lambda_j \in S} \left| L_1 - \frac{p^{-1}\tau}{n} \operatorname{Tr}\left(\Sigma_2 \left(\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n \lambda_j \mathbb{I} \right)^{-1} X_{(-i)}^\top X_{(-i)} \right) \right| \leq q^{-1} p^{-1} \tau n^{1/2} \log(p).$$

Combining the above two inequalities, we have

$$L_1 \leq n p^{-1} \tau C_2^2 \frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n \lambda_j} + q^{-1} p^{-1} \tau n^{1/2} \log(p), \quad \forall \lambda_j \in S.$$

To bound L_2 in (1.68), by definition, we have

$$L_2 = |n^{-1} p^{-1/2} \tau^{1/2} q^{-1/2} z_{(-i)}^\top X_{(-i)} (\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n \lambda_j \mathbb{I})^{-1} \Sigma_2 (\sqrt{q} b_0)|.$$

Using the facts that $\lambda_{\min}(n^{-1} X_{(-i)}^\top X_{(-i)} + c_n \lambda_j \mathbb{I}) \geq c_n \lambda_j \geq c_n \epsilon$, $\|A\|_{\mathrm{F}} \leq \sqrt{\operatorname{rank}(A)}\, \|A\|$, and Eq. (1.30), we have

$$\begin{aligned}
&\big\| X_{(-i)} \big(\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n \lambda_j \mathbb{I} \big)^{-1} \Sigma_2 \big\| \leq C_2 c_n^{-1} \epsilon^{-1} \| X_{(-i)} \| \lesssim n p^{-1/2}, \\
&\big\| X_{(-i)} \big(\frac{1}{n} X_{(-i)}^\top X_{(-i)} + c_n \lambda_j \mathbb{I} \big)^{-1} \Sigma_2 \big\|_{\mathrm{F}}^2 \lesssim \operatorname{rank}(X_{(-i)}) n^2 p^{-1} \lesssim n^3 p^{-1}.
\end{aligned}$$

Therefore, by Lemma 5 and the fact that $\sqrt{q} b_{0,i}$ has bounded sub-exponential norm, it holds that, for some constant $\tilde{c}_1$, $\mathrm{P}\left(L_2 > q^{-1/2} n^{1/2} \tau^{1/2} p^{-1} \log(p) \right) \leq 2\exp(-\tilde{c}_1 \log^2(p))$. As a consequence, with probability at least $1 - 2p^{10}\exp(-\tilde{c}_1 \log^2(p))$, we have $\sup_{\lambda_j \in S} L_2 \leq q^{-1/2} n^{1/2} \tau^{1/2} p^{-1} \log(p)$. Therefore, taking bounds for L_1 and L_2 together, we have, w.p.a.1,

$$|\beta_0^\top \Sigma_2 \hat{\beta}_{\lambda_j}^i| \leq 2 n p^{-1} \tau C_2^2 \frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2 + c_n \lambda_j}, \tag{1.69}$$

for each $\lambda_j \in S$. In the last inequality, we use $p^{-1} q^{-1} \tau n^{1/2} \log(p), q^{-1/2} n^{1/2} \tau^{1/2} p^{-1} \log(p) =$

$o(np^{-1}\tau^2)$ by the assumptions of Theorem 3. With (1.67) and (1.69), we have

$$\begin{aligned}&\|\Sigma_2^{1/2}(\hat{\beta}^i_{\lambda_j}-\beta_0)\|^2-\|\Sigma_2^{1/2}\beta_0\|^2=\|\Sigma_2^{1/2}\hat{\beta}^i_{\lambda_j}\|^2-2\beta_0^\top\Sigma_2\hat{\beta}^i_{\lambda_j}\\ \geq&\frac{c_2^2\sigma_\varepsilon^2 p}{2nK}\left(C_2(1+\sqrt{c_n})^2+c_n\lambda_j\right)^{-2}-4np^{-1}\tau C_2^2\frac{(1+\sqrt{c_n})^2}{C_2(1+\sqrt{c_n})^2+c_n\lambda_j}.\end{aligned}$$

This inequality holds w.p.a. 1 as $n,p\to\infty$ uniformly for all $\lambda_j\in S$. Given our initial choice for $\tilde{c}$, it is easy to check that the right-hand side exceeds $50np^{-1}\tau^2$, which implies Eq. (1.32).

To prove Eq. (1.33), note that

$$\frac{1}{n}\|y_{(i)}-X_{(i)}\hat{\beta}^i_{\tau^{-1}\lambda}\|^2-\frac{1}{n}\|\varepsilon_{(i)}\|^2=\frac{1}{n}\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2+\frac{2}{n}\varepsilon_{(i)}^\top Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0).$$

By the facts $Z_{(i)}\perp\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0$ and $n^{-1}\chi^2(K^{-1}n)=K^{-1}+O_p(n^{-1/2})$, we have

$$\begin{aligned}\frac{1}{n}\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2\overset{d}{=}&\frac{1}{n}\chi^2(K^{-1}n)\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2\\ =&\frac{1}{K}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2+O_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2\right).\end{aligned}$$

Additionally, by Theorem 2, we deduce:

$$\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2-\|\Sigma_2^{1/2}\beta_0\|^2=\frac{2(K-1)}{K}np^{-1}\tau^2\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2}{2\lambda^2}-\frac{\sigma_\beta^2}{\lambda}\right)+o_{\mathrm{P}}(\tau^2np^{-1}).$$

Hence, using the fact that $\|\Sigma_2^{1/2}\beta_0\|^2\asymp_{\mathrm{P}}\tau$, we derive the following equation:

$$\frac{1}{n}\sum_{i=1}^K\|Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)\|^2-\|\Sigma_2^{1/2}\beta_0\|^2=\frac{2(K-1)}{K}np^{-1}\tau^2\theta_2\sigma_x^4\left(\frac{\sigma_\varepsilon^2}{2\lambda^2}-\frac{\sigma_\beta^2}{\lambda}\right)+o_{\mathrm{P}}(\tau^2np^{-1}).$$

Thus, to prove Eq. (1.33), it remains to show that: $\frac{2}{n}\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda}-\beta_0)=o_{\mathrm{P}}(\tau^2np^{-1})$.

Given that $\varepsilon_{(i)} \perp Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0)$ and $n^{-3/2}\tau^{-3/2}p \to 0$ by Assumption 4, we have

$$\frac{2}{n}\varepsilon_{(i)}Z_{(i)}\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0) \overset{d}{=} \frac{2}{n}\|\Sigma_2^{1/2}(\hat{\beta}^i_{\tau^{-1}\lambda} - \beta_0)\|\varepsilon^\top_{(i)}x = O_{\mathrm{P}}(n^{-1/2}\tau^{1/2}) = o_{\mathrm{P}}(\tau^2 np^{-1}),$$

where x is a standard Gaussian vector independent of $\varepsilon_{(i)}$. This concludes the proof. □

Lemma 20. *There exists a constant $\tilde{C}_1$ such that, w.p.a.1, uniformly for $\mu_1, \mu_2 \in [0, \tilde{c}^{-1}]$,*

$$pn^{-1}\tau^{-2}|\widetilde{R}^{K-CV}(\mu_1) - \widetilde{R}^{K-CV}(\mu_2)| \le \tilde{C}_1|\mu_1 - \mu_2| + o_{\mathrm{P}}\left(pn^{-1}\tau^{-2}\right).$$

Proof. By the Woodbury identity, we deduce that

$$\begin{aligned}
&\left(\frac{1}{n}X^\top_{(-i)}X_{(-i)} + c_n\tau^{-1}\mu^{-1}\mathbb{I}\right)^{-1} - c_n^{-1}\tau\mu\mathbb{I} \\
&= -\frac{c_n^{-2}\tau^2\mu^2}{n}X^\top_{(-i)}\left(\mathbb{I} + \frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X^\top_{(-i)}\right)^{-1}X_{(-i)}.
\end{aligned}$$

Hence, we arrive at:

$$\begin{aligned}
&\sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}} c_n\tau^{-3}\log^{-1}(p)\left\|\left(\frac{1}{n}X^\top_{(-i)}X_{(-i)} + c_n\tau^{-1}\mu^{-1}\mathbb{I}\right)^{-1} - c_n^{-1}\tau\mu\mathbb{I} + \frac{c_n^{-2}\tau^2\mu^2}{n}X^\top_{(-i)}X_{(-i)}\right\| \\
&= \sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}} c_n^{-1}\mu^2\tau^{-1}\log^{-1}(p)\left\|\frac{1}{n}X^\top_{(-i)}\left[\left(\mathbb{I} + \frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X^\top_{(-i)}\right)^{-1} - \mathbb{I}\right]X_{(-i)}\right\| \\
&\le \sup_{\substack{1\le i\le K\\ \mu\in[0,\tilde{c}^{-1}]}} \mu^3 c_n^{-2}\log^{-1}(p)\left\|\frac{1}{n}X^\top_{(-i)}X_{(-i)}\right\|^2 \xrightarrow{\mathrm{P}} 0. \qquad (1.70)
\end{aligned}$$

The last inequality is a consequence of Eq. (1.30) and the fact that

$$\begin{aligned}
\left\|\left(\mathbb{I} + \frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X^\top_{(-i)}\right)^{-1} - \mathbb{I}\right\| &\le \left\|\left(\mathbb{I} + \frac{c_n^{-1}\tau\mu}{n}X_{(-i)}X^\top_{(-i)}\right)^{-1}\right\| \cdot c_n^{-1}\tau\mu\left\|\frac{1}{n}X^\top_{(-i)}X_{(-i)}\right\| \\
&\le c_n^{-1}\tau\mu\left\|\frac{1}{n}X^\top_{(-i)}X_{(-i)}\right\|.
\end{aligned}$$

On the other hand, by direct calculation we have that $\widetilde{R}^{K-CV}(\mu_1) - \widetilde{R}^{K-CV}(\mu_2)$ equals:

$$\begin{aligned}
&\sum_{i=1}^{K}\left(\frac{1}{n}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}\|^2 - \frac{1}{n}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}}\|^2\right) - \frac{2}{n}y_{(i)}^\top X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}} - \hat{\beta}^i_{\tau^{-1}\mu_2^{-1}}) \\
=:&\sum_{i=1}^{K} W_{1i}(\mu_1,\mu_2) - W_{2i}(\mu_1,\mu_2).
\end{aligned}$$

We next investigate $W_{1i}(\mu_1,\mu_2)$ and $W_{2i}(\mu_1,\mu_2)$ separately. For $W_{1i}(\mu_1,\mu_2)$, we have

$$\begin{aligned}
W_{1i}(\mu_1,\mu_2) \le{}& \frac{1}{n}\left\|X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}} - \hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\right\| \cdot \|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}\| \\
&+ \frac{1}{n}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}}\| \cdot \left\|X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}} - \hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\right\|.
\end{aligned}$$

Define $\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}} = \frac{1}{n}\left[c_n^{-1}\tau\mu_1\mathbb{I} - \frac{c_n^{-2}\tau^2\mu_1^2}{n}X_{(-i)}^\top X_{(-i)}\right]X_{(-i)}^\top y_{(-i)}$. Observe that

$$\begin{aligned}
&\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\left\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}} - X_{(i)}\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}\right\| \\
&= O_{\mathrm{P}}(\tau^3\log(p)) = o_{\mathrm{P}}(c_n^{-1/2}\tau),
\end{aligned} \tag{1.71}$$

where we use Eq. (1.70), Eq. (1.30), and Eq. (1.66). Additionally, it is easy to verify that

$$\begin{aligned}
&\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\left\|\frac{1}{n}X_{(i)}\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}\right\| \\
&= \sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\left\|\frac{1}{n}X_{(i)}\left[c_n^{-1}\tau\mu_1\mathbb{I} - \frac{c_n^{-2}\tau^2\mu_1^2}{n}X_{(-i)}^\top X_{(-i)}\right]X_{(-i)}^\top y_{(-i)}\right\| \\
&\le \frac{c_n^{-1}\tau\tilde{c}^{-1}}{n\sqrt{n}}\left\|X_{(i)}X_{(-i)}^\top y_{(-i)}\right\| + \frac{c_n^{-2}\tau^2\tilde{c}^{-2}}{n^2\sqrt{n}}\left\|X_{(i)}X_{(-i)}^\top X_{(-i)}X_{(-i)}^\top y_{(-i)}\right\|.
\end{aligned}$$

For the first term, by Eq. (1.66), it equals

$$\frac{c_n^{-1}\tau\tilde{c}^{-1}}{n\sqrt{n}}\left\|Z_{(i)}\Sigma_2^{1/2}X_{(-i)}^\top y_{(-i)}\right\| \overset{d}{=} \frac{c_n^{-1}\tau\tilde{c}^{-1}}{n\sqrt{n}}\sqrt{\chi^2(n/K)}\left\|\Sigma_2^{1/2}X_{(-i)}^\top y_{(-i)}\right\| \le \frac{\tilde{C}_1}{2}c_n^{-1/2}\tau,$$

w.p.a.1 for some constant $\tilde{C}_1$ that only depends on fixed constants. The second term can be bounded in the same way. Therefore, we have

$$\sup_{\mu_1\in[0,\tilde{c}^{-1}]}\frac{1}{\sqrt{n}}\|X_{(i)}\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}\|\leq\tilde{C}_1c_n^{-1/2}\tau+o_{\mathrm{P}}(c_n^{-1/2}\tau). \tag{1.72}$$

Analogously, we can prove that $\frac{1}{\sqrt{n}}\|X_{(i)}(\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\|\leq\tilde{C}_1|\mu_1-\mu_2|c_n^{-1/2}\tau+o_{\mathrm{P}}(c_n^{-1/2}\tau)$ holds uniformly for $\mu_1,\mu_2\in[0,\tilde{c}^{-1}]$, where $\tilde{C}_1$ is a fixed constant that may vary from line to line. In light of this, we deduce that: $\sup_{1\leq i\leq K}W_{1i}(\mu_1,\mu_2)\leq\tilde{C}_1^2c_n^{-1}\tau^2|\mu_1-\mu_2|+o_{\mathrm{P}}(c_n^{-1}\tau^2)$ holds w.p.a.1 uniformly for $\mu_1,\mu_2\in[0,\tilde{c}^{-1}]$.

To bound $W_{2i}(\mu_1,\mu_2)$, we first define $\tilde{W}_{2i}(\mu_1,\mu_2)=\frac{2}{n}y_{(i)}^\top X_{(i)}(\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\tilde{\beta}^i_{\tau^{-1}\mu_2^{-1}})$. By Eq. (1.71), it holds that

$$\begin{aligned}&\sup_{\mu_1,\mu_2\in[0,\tilde{c}^{-1}]}|\tilde{W}_{2i}(\mu_1,\mu_2)-W_{2i}(\mu_1,\mu_2)|\leq\frac{2}{n}\|y_{(i)}^\top\|\|X_{(i)}(\tilde{\beta}^i_{\tau^{-1}\mu_1^{-1}}-\hat{\beta}^i_{\tau^{-1}\mu_1^{-1}})\|\\&+\frac{2}{n}\|y_{(i)}^\top\|\|X_{(i)}(\tilde{\beta}^i_{\tau^{-1}\mu_1^{-2}}-\hat{\beta}^i_{\tau^{-1}\mu_2^{-1}})\|=O_{\mathrm{P}}(\tau^3\log(p))=o_{\mathrm{P}}(c_n^{-1}\tau^2).\end{aligned}$$

Moreover, employing a similar argument to that used in proving Eq. (1.72), we have

$$\begin{aligned}&\sup_{\mu_1,\mu_2\in[0,\tilde{c}^{-1}]}|\tilde{W}_{2i}(\mu_1,\mu_2)|\\&=\sup_{\mu_1,\mu_2\in[0,\tilde{c}^{-1}]}\left|\frac{2}{n}y_{(i)}^\top X_{(i)}\frac{1}{n}\left[c_n^{-1}\tau(\mu_1-\mu_2)\mathbb{I}-\frac{c_n^{-2}\tau^2(\mu_1^2-\mu_2^2)}{n}X_{(-i)}^\top X_{(-i)}\right]X_{(-i)}^\top y_{(-i)}\right|\\&\lesssim|\mu_1-\mu_2|\frac{c_n^{-1}\tau}{n^2}\left|y_{(i)}^\top X_{(i)}X_{(-i)}^\top y_{(-i)}\right|+|\mu_1-\mu_2|\frac{c_n^{-2}\tau^2}{n^3}\left|y_{(i)}^\top X_{(i)}X_{(-i)}^\top X_{(-i)}X_{(-i)}^\top y_{(-i)}\right|.\end{aligned}$$

For the first term, by Lemmas 2 and 3, it is easy to verify that

$$\begin{aligned}\frac{c_n^{-1}\tau}{n^2}\left|y_{(i)}^\top X_{(i)}X_{(-i)}^\top y_{(-i)}\right|&\leq\frac{c_n^{-1}\tau}{n^2}\left|\varepsilon_{(i)}^\top X_{(i)}X_{(-i)}^\top\varepsilon_{(-i)}\right|+\frac{c_n^{-1}\tau}{n^2}\left|\varepsilon_{(i)}^\top X_{(i)}X_{(-i)}^\top X_{(-i)}\beta_0\right|\\&+\frac{c_n^{-1}\tau}{n^2}\left|\beta_0^\top X_{(i)}^\top X_{(i)}X_{(-i)}^\top\varepsilon_{(-i)}\right|+\frac{c_n^{-1}\tau}{n^2}\left|\beta_0^\top X_{(i)}^\top X_{(i)}X_{(-i)}^\top X_{(-i)}\beta_0\right|\leq\tilde{C}_1c_n^{-1}\tau^2,\end{aligned}$$

for some constant $\tilde{C}_1$ w.p.a.1. The second term can be shown analogously. As a result, we have $\sup_{1\le i\le K} W_{2i}(\mu_1,\mu_2) \le \tilde{C}_1 c_n^{-1}\tau^2|\mu_1-\mu_2| + o_{\mathrm{P}}(c_n^{-1}\tau^2)$ w.p.a.1, uniformly for $\mu_1,\mu_2 \in [0,\tilde{c}^{-1}]$. Combining the bounds for $W_{1i}(\mu_1,\mu_2)$ and $W_{2i}(\mu_1,\mu_2)$ concludes the proof. □

Lemma 21. *A_1 to A_3 defined in (1.34) converge to zero w.p.a. 1.*

Proof. For A_1, by using the same argument as Eq. (1.62) and noting that $\lambda_n^{opt} \asymp \tau^{-1}$, $\hat{\lambda}_n^{K-CV} \asymp_{\mathrm{P}} \tau^{-1}$ and $\tau(\lambda_n^{opt} - \hat{\lambda}_n^{K-CV}) = o_{\mathrm{P}}(1)$, we have

$$|A_1| \lesssim c_n\tau^{-2}\|\hat{\beta}_{cv}-\hat{\beta}_{opt}\|\|\hat{\beta}_{cv}\| \lesssim_{\mathrm{P}} c_n\tau^{-2}\cdot c_n^{-1/2}|\lambda_n^{opt}-\hat{\lambda}_n^{K-CV}|\tau^2\cdot c_n^{-1/2}\tau = o_{\mathrm{P}}(1).$$

Similarly, $A_2 = o_{\mathrm{P}}(1)$. To prove $A_3 = o_{\mathrm{P}}(1)$, define $\tilde{\beta}(\lambda_n) = \frac{1}{n}\left[c_n^{-1}\lambda_n^{-1}\mathbb{I} - \frac{c_n^{-2}\lambda_n^{-2}}{n}X^\top X\right]X^\top y$ and write $\tilde{\beta}(\hat{\lambda}_n^{K-CV})$ as $\tilde{\beta}_{cv}$ and $\tilde{\beta}(\lambda_n^{opt})$ as $\tilde{\beta}_{opt}$ for simplicity. By using a similar result as Eq. (1.70) as well as the fact $n^{-1}\|X^\top y\| \lesssim n^{-1}\|X\|\|y\| \lesssim_{\mathrm{P}} c_n^{1/2}$ according to Lemma 6, we have

$$\begin{aligned}
&c_n\tau^{-2}|(\hat{\beta}_{opt}-\tilde{\beta}_{opt})^\top\Sigma_2\beta_0| \lesssim c_n\tau^{-2}\|\hat{\beta}_{opt}-\tilde{\beta}_{opt}\|\|\beta_0\| \\
\lesssim& c_n\tau^{-2}\left\|\left(\frac{1}{n}X^\top X + c_n\lambda_n^{opt}\mathbb{I}\right)^{-1} - c_n^{-1}(\lambda_n^{opt})^{-1}\mathbb{I} + \frac{c_n^{-2}(\lambda_n^{opt})^{-2}}{n}X^\top X\right\|\left\|\frac{1}{n}X^\top y\right\|\cdot\|\beta_0\| \\
=& o_{\mathrm{P}}(c_n\tau^{-2}\cdot c_n^{-1}\tau^3\log(p)\cdot c_n^{1/2}\cdot\tau^{1/2}) = o_{\mathrm{P}}(c_n^{1/2}\tau^{3/2}\log(p)) = o_{\mathrm{P}}(1),
\end{aligned}$$

where the last equation holds by Assumption 4. Similarly, $c_n\tau^{-2}|(\hat{\beta}_{cv}-\tilde{\beta}_{cv})^\top\Sigma_2\beta_0| = o_{\mathrm{P}}(1)$. Therefore, $A_3 = o_{\mathrm{P}}(1)$ follows by $c_n\tau^{-2}(\tilde{\beta}_{opt}-\tilde{\beta}_{cv})^\top\Sigma_2\beta_0 = o_{\mathrm{P}}(1)$. Note that

$$\begin{aligned}
c_n\tau^{-2}(\tilde{\beta}_{opt}-\tilde{\beta}_{cv})^\top\Sigma_2\beta_0 &= n^{-1}\tau^{-2}\big((\hat{\lambda}_n^{K-CV})^{-1} - (\lambda_n^{opt})^{-1}\big)\beta_0^\top\Sigma_2X^\top y \\
- n^{-2}c_n^{-1}\tau^{-2}\big((\hat{\lambda}_n^{K-CV})^{-2} - (\lambda_n^{opt})^{-2}\big)\beta_0^\top\Sigma_2X^\top XX^\top y &=: B_1 + B_2.
\end{aligned}$$

By Lemma 2 and Lemma 3, we have

$$B_1 = o_{\mathrm{P}}(\tau^{-1}\big((\hat{\lambda}_n^{K-CV})^{-1} - (\lambda_n^{opt})^{-1}\big)) = o_{\mathrm{P}}(1).$$

The same argument proves $B_2 = o_{\mathrm{P}}(1)$, leading to $A_3 = o_{\mathrm{P}}(1)$, which concludes the proof.

□

Lemma 22. *The objecive function in (1.36) is convex with respect to α and jointly concave with respect to (δ, γ). Additionally, as long as Eq. (1.38) holds, we have Eq. (1.35).*

Proof. Define $S_w^n = \{w \big| c_n\tau^{-1}\sigma_x\sigma_\beta + c_\alpha/4\sigma_\beta \leq c_n\|w\| \leq c_n\tau^{-1}\sigma_x\sigma_\beta + C_\alpha/\sigma_\beta\}$. Analogous to the result proved by Lemma 15, if the solution $\hat{w}^B$ to the following problem

$$\min_{w\in S_w^n} \frac{c_n}{n}\|\tau^{1/2}\Sigma_1^{1/2}Z\tilde{w} - \tau^{-1}\varepsilon\|^2 + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\|_1 - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi \tag{1.73}$$

satisfies $c_n\|\hat{w}^B\| - c_n\tau^{-1}\sigma_x\sigma_\beta \in [c_\alpha/2\sigma_\beta + \epsilon, C_\alpha/2\sigma_\beta - \epsilon]$ w.a.p.1, then the same holds true for $\hat{w}$, which leads to the desired result, (1.35). In light of this, without ambiguity we now directly focus on (1.73), and refer to $\hat{w}^B$ as $\hat{w}$ for ease of notation.

Note that for any vector x, it holds that $\|x\|^2 = \max_u \sqrt{n}u^\top x - n\|u\|^2/4$, and $\|x\|_1 = \max_{\|v\|_\infty\leq 1} v^\top x$. By applying these equations to $\|\tau^{1/2}\Sigma_1^{1/2}Z\tilde{w} - \tau^{-1}\varepsilon\|^2$ and $\|\Sigma_2^{-1/2}w + \tau^{-3/2}\beta_0\|_1$, and letting $\tilde{u} := \Sigma_1^{1/2}u$, the problem (1.73) can be reformulated as:

$$\begin{aligned}\min_{w\in S_w^n} \max_{\substack{\tilde{u}\\ \|v\|_\infty\leq 1}} \frac{c_n\tau^{1/2}}{\sqrt{n}}\tilde{u}^\top Zw - \frac{c_n\tau^{-1}}{\sqrt{n}}\tilde{u}^\top\Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}\tilde{u}\|^2}{4} + \frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}}v^\top\beta_0 \\ + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}}v^\top\Sigma_2^{-1/2}w - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi.\end{aligned} \tag{1.74}$$

For convenience, we shall continue to employ u in place of $\tilde{u}$ throughout the remainder of

the proof. Let $S_u^n = \{u \big| \|u\| \leq 4\tau^{-1}\sqrt{C_1 C_\varepsilon}\}$. Similar to the proof of Lemma 14, w.a.p.1,

$$\min_{w\in S_w^n} \max_{\substack{u\in S_u^n \\ \|v\|_\infty \leq 1}} \frac{c_n\tau^{1/2}}{\sqrt{n}} u^\top Z w - \frac{c_n\tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} + \frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}} v^\top \beta_0 + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}} v^\top \Sigma_2^{-1/2} w - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi \tag{1.75}$$

is equivalent to Eq. (1.74). Next, we construct an auxiliary optimization problem:

$$\begin{aligned} \phi(g,h) &= \max_{\substack{0\leq\delta\leq 4\tau^{-1}\sqrt{C_1C_\varepsilon} \\ \|v\|_\infty\leq 1}} \min_{w\in S_w^n} \max_{\|u\|=\delta} \mathcal{R}_n(w,v,u), \quad \text{where} \\ \mathcal{R}_n(w,v,u) &= \frac{c_n\tau^{1/2}}{\sqrt{n}}\|w\| g^\top u - \frac{c_n\tau^{1/2}}{\sqrt{n}}\|u\| h^\top w - \frac{c_n\tau^{-1}}{\sqrt{n}} u^\top \Sigma_1^{-1/2}\varepsilon - \frac{c_n\|\Sigma_1^{-1/2}u\|^2}{4} \\ &\quad + \frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}} v^\top \beta_0 + \frac{c_n\tau^{-1/2}\lambda_n}{\sqrt{n}} v^\top \Sigma_2^{-1/2} w - \frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2 - C_n^\phi, \end{aligned} \tag{1.76}$$

and both $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^p$ are standard Gaussian vectors, independent of all other random variables. Moreover, let $\tilde{\mathcal{S}}_n := \{w \big| c_\alpha/2\sigma_\beta + \epsilon < c_n\|w\| - c_n\tau^{-1}\sigma_x\sigma_\beta < C_\alpha/2\sigma_\beta - \epsilon\}$, define $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$ as the optimal value of the optimization problem (1.76), with $w \in S_w^n \cap \tilde{\mathcal{S}}_n^c$.

Lemma 23 characterizes the limiting behavior of the optimal solution to (1.75), $\hat{w}$, and in turn, proves the desired (1.35), under conditions pertaining to the optimization problem (1.76). Therefore, we only need show that conditions outlined in Lemma 23 hold as long as (1.38) holds. That is, under (1.38), we need to prove the existence of the constants $\bar{\phi} < \bar{\phi}_{\tilde{\mathcal{S}}_n^c}$ such that for all $\eta > 0$, w.a.p.1, $\phi(g,h) < \bar{\phi} + \eta$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h) > \bar{\phi}_{\tilde{\mathcal{S}}_n^c} - \eta$.

Following the same argument as in the proof of Lemma 14, after maximizing over the

direction of u and minimizing over the direction of w, Eq. (1.76) becomes equivalent to:

$$\max_{\substack{0\le\delta\le 4\tau^{-1}\sqrt{C_1C_\varepsilon}\\ \|v\|_\infty\le 1}} \min_{\alpha\in K_\alpha} -\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}$$
$$\times(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)-c_n\left\|n^{-1/2}\tau^{1/2}\delta h-n^{-1/2}\tau^{-1/2}\lambda_n\Sigma_2^{-1/2}v\right\|\alpha$$
$$+\frac{c_n\tau^{-2}\lambda_n}{\sqrt{n}}v^\top\beta_0-\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2-C_n^\phi,$$

where $K_\alpha=\{\alpha|c_n\alpha-c_n\tau^{-1}\sigma_x\sigma_\beta\in[c_\alpha/4\sigma_\beta,C_\alpha/\sigma_\beta]\}$. By Lemma 17, the objective function of the above optimization problem is convex in α and jointly concave in (δ,v). Consequently, we can interchange the order of min and max by applying Corollary 3.3 in Sion [1958]. Applying $\|x\|=\min_{\gamma>0}\frac{1}{2\gamma}\|x\|^2+\frac{\gamma}{2}$ to $\left\|n^{-1/2}\tau^{1/2}\delta h^\top-n^{-1/2}\tau^{-1/2}\lambda_n v^\top\Sigma_2^{-1/2}\right\|\alpha$, and By completing the square for terms associated with v, we can rewrite this problem as (1.36). As a consequence, we conclude that (1.36) is convex with respect to α and jointly concave with respect to (δ,γ).

Finally, based on the above argument, we conclude that under (1.38), for all $\eta>0$, w.a.p.1, $\phi(g,h)<\bar\phi+\eta$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)>\bar\phi_{\tilde{\mathcal{S}}_n^c}-\eta$ by choosing $\bar\phi=-\frac{C_\lambda}{8C_2}$ and $\bar\phi_{\tilde{\mathcal{S}}_n^c}=-\frac{C_\lambda}{100C_2}$, thereby verifying conditions outlined in Lemma 23. □

Next, we introduce a lemma that resembles Lemma 16.

Lemma 23. *Let $\hat w$ denote an optimal solution of Eq. (1.75). Regarding $\phi(g,h)$ and $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)$, as introduced and discussed in relation to Eq. (1.76), suppose there are constants $\bar\phi$ and $\bar\phi_{\tilde{\mathcal{S}}_n^c}$ with $\bar\phi<\bar\phi_{\tilde{\mathcal{S}}_n^c}$, such that for all $\eta>0$, the following hold w.a.p.1 as $n\to\infty$: (a) $\phi(g,h)<\bar\phi+\eta$, (b) $\phi_{\tilde{\mathcal{S}}_n^c}(g,h)>\bar\phi_{\tilde{\mathcal{S}}_n^c}-\eta$. Under these conditions, we have $\hat w\in\tilde{\mathcal{S}}_n$ w.p.a.1.*

Lemma 24. *There exists some sufficiently small $\epsilon>0$, such that for any $\eta>0$, w.p.a.1, the inequalities in (1.38) hold.*

Proof. By Eq. (1.61) in Lemma 18, we have the following result:

$$
\begin{aligned}
&-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)^\top\left(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I}\right)^{-1}\left(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon\right)\\
&\quad=c_n\tau^{-1}\sigma_x^2\sigma_\beta^2-c_n\sigma_\varepsilon^2\theta_3\mu^2(\sigma_x\sigma_\beta,\delta_1^*,\delta_2)+2\sigma_x\sigma_\beta\alpha_2+\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2+o_{\mathrm{P}}(1).
\end{aligned}
$$

Additionally, by Lemmas 2-3 and $p^{1/2}\tau^{-1}n^{-1}q^{-1/2}=o(1)$ by Assumption 4, we deduce that $-\frac{c_n\gamma}{2}+\frac{c_n\gamma\tau^{-3}}{2\alpha^2}\beta_0\Sigma_2\beta_0+\frac{c_n\tau^{-1}\delta}{\sqrt{n}}h^\top\Sigma_2^{1/2}\beta_0\xrightarrow{\mathrm{P}}-\frac{\gamma_1\alpha_2}{\sigma_x\sigma_\beta}$. In the sequel, we examine the asymptotic behavior of the remaining term in $\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$:

$$
\min_{\|v\|_\infty\le 1}\left\{\frac{c_n\alpha^2}{2\gamma}\left\|n^{-1/2}\tau^{-1/2}\lambda_n\Sigma_2^{-1/2}v-n^{-1/2}\tau^{1/2}\delta h-\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2^{1/2}\beta_0\right\|^2\right\}. \tag{1.77}
$$

By using $\|\Sigma_2^{-1}\|\le c_2^{-1}$, we see (1.77) is upper bounded by

$$
\begin{aligned}
&\frac{c_n\alpha^2}{2\gamma c_2}\min_{\|v\|_\infty\le 1}\left\{\left\|n^{-1/2}\tau^{-1/2}\lambda_n v-n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h-\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0\right\|^2\right\}\\
=&\frac{c_n\alpha^2}{2\gamma c_2}\left\|\left(\left|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h+\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0\right|-n^{-1/2}\tau^{-1/2}\lambda_n\right)_+\right\|^2.
\end{aligned}
$$

Similarly, with $\lambda_{\min}\Sigma_2^{-1}\ge C_2^{-1}$, (1.77) is lower bounded by $\frac{c_n\alpha^2}{2\gamma C_2}\left\|\left(|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h+\frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0|-n^{-1/2}\tau^{-1/2}\lambda_n\right)_+\right\|^2$. Together with Lemma 25, we deduce that, w.p.a.1, (1.77) lies in $\left[\frac{\sigma_x^2\sigma_\beta^2}{4\gamma_1C_2}C_\lambda,\frac{\sigma_x^2\sigma_\beta^2}{\gamma_1c_2}C_\lambda\right]$.

Recall that $\tilde{Q}_n(\alpha_2,\delta_3,\gamma_1)$ is defined in (1.37). We introduce $\tilde{Q}_n^{\mathrm{upper}}(\alpha_2,\delta_3,\gamma_1)$, defined as:

$$
\begin{aligned}
&-\frac{c_n\delta^2}{4}\mu_n(\alpha,\delta)+\frac{c_n}{n}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)^\top(\Sigma_1^{-1}-\mu_n(\alpha,\delta)\mathbb{I})^{-1}(\tau^{1/2}\alpha g-\tau^{-1}\Sigma_1^{-1/2}\varepsilon)\\
&\quad-\frac{c_n\gamma}{2}-\frac{\sigma_x^2\sigma_\beta^2}{4\gamma_1C_2}C_\lambda+\frac{c_n\gamma\tau^{-3}}{2\alpha^2}\beta_0\Sigma_2\beta_0+\frac{c_n\tau^{-1}\delta}{\sqrt{n}}h^\top\Sigma_2^{1/2}\beta_0-\frac{c_n\tau^{-2}}{n}\|\varepsilon\|^2-C_n^\phi.
\end{aligned}
$$

Similarly, we define $\tilde{Q}_n^{\mathrm{lower}}$ with the term $-\frac{\sigma_x^2\sigma_\beta^2}{4\gamma_1C_2}C_\lambda$ in $\tilde{Q}_n^{\mathrm{upper}}$ being replaced by $-\frac{\sigma_x^2\sigma_\beta^2}{\gamma_1c_2}C_\lambda$.

Consequently, $\tilde{Q}_n^{\text{lower}} \le \tilde{Q}_n \le \tilde{Q}_n^{\text{upper}}$. Note also that $\tilde{Q}_n^{\text{lower}}(\alpha_2, \delta_2, \gamma_1)$ and $\tilde{Q}_n^{\text{upper}}(\alpha_2, \delta_3, \gamma_1)$ maintain their convexity in α_2 and joint concavity in (δ_3, γ_1). By employing a similar line of reasoning as presented in Lemma 18, alongside the definitions of c_α and C_α, it becomes evident that there exists a sufficiently small $\epsilon > 0$ such that

$$\min_{\alpha_2 \in [\frac{c_\alpha}{4\sigma_\beta}, \frac{C_\alpha}{\sigma_\beta}]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n^{\text{upper}} \xrightarrow{\text{P}} -\frac{C_\lambda}{8C_2},$$

$$\text{and} \quad \min_{\alpha_2 \in [\frac{c_\alpha}{4\sigma_\beta}, \frac{c_\alpha}{2\sigma_\beta}+\epsilon] \cup [\frac{C_\alpha}{2\sigma_\beta}-\epsilon, \frac{C_\alpha}{\sigma_\beta}]} \max_{\substack{\gamma_1 > 0 \\ \delta_3 \in K_{\delta_3}}} \tilde{Q}_n^{\text{lower}} \xrightarrow{\text{P}} -\frac{C_\lambda}{100C_2}.$$

These results immediately yield the desired inequalities. □

Lemma 25. *For any $\alpha_2, \delta_3 \in \mathbb{R}$ and $\gamma_1 > 0$, w.p.a.1, we have*

$$\frac{C_\lambda}{2} \le c_n \tau^{-1} \left\| \left(\left| n^{-1/2} \tau^{1/2} \Sigma_2^{1/2} \delta h + \frac{\gamma}{\alpha^2} \tau^{-3/2} \Sigma_2 \beta_0 \right| - n^{-1/2} \tau^{-1/2} \lambda_n \right)_+ \right\|^2 \le 2C_\lambda. \quad (1.78)$$

Proof. We first establish the following:

$$c_n n^{-1} \tau^{-2} \left\{ \left\| \left(\left| \Sigma_2^{1/2} \delta_1^* h \right| - \lambda_n \right)_+ \right\|^2 - \mathbb{E} \left\| \left(\left| \Sigma_2^{1/2} \delta_1^* h \right| - \lambda_n \right)_+ \right\|^2 \right\} \xrightarrow{\text{P}} 0. \quad (1.79)$$

Let $\tilde{h} := \Sigma_2^{1/2} h \sim \mathcal{N}(0, \Sigma_2)$. Let us denote the (i,j)-th element of Σ_2 as $\Sigma_{2,ij}$, thus we have $\tilde{h}_j | \tilde{h}_i \overset{d}{=} \Sigma_{2,ij} \Sigma_{2,ii}^{-1} \tilde{h}_i + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1} \Sigma_{2,ij}^2} g_1$, where g_1 is a standard Gaussian random variable independent of $\tilde{h}_i$. Consequently, $\text{Cov}\left((|\delta_1^* \tilde{h}_i| - \lambda_n)_+^2, (|\delta_1^* \tilde{h}_j| - \lambda_n)_+^2 \right)$ equals

$$\begin{aligned} &\mathbb{E}\left\{ (|\delta_1^* \tilde{h}_i| - \lambda_n)_+^2 \mathbb{E}\left[(|\delta_1^* \tilde{h}_j| - \lambda_n)_+^2 - \mathbb{E}(|\delta_1^* \tilde{h}_j| - \lambda_n)_+^2 | \tilde{h}_i \right] \right\} \\ =&\mathbb{E}\left\{ (|\delta_1^* \tilde{h}_i| - \lambda_n)_+^2 \mathbb{E}\left[\eta(\tilde{h}_i)^2 - \eta(g_2)^2 \middle| \tilde{h}_i \right] \right\}, \end{aligned}$$

where $\eta(x) := \left(\left| \delta_1^* \left(\Sigma_{2,ij} \Sigma_{2,ii}^{-1} x + \sqrt{\Sigma_{2,jj} - \Sigma_{2,ii}^{-1} \Sigma_{2,ij}^2} g_1 \right) \right| - \lambda_n \right)_+$ and $g_2 \sim \mathcal{N}(0, \Sigma_{2,ii})$ is independent of both g_1 and $\tilde{h}_i$. In addition, note that $\left| \eta(\tilde{h}_i)^2 - \eta(g_2)^2 \right| \le |\delta_1^* \Sigma_{2,ij} \Sigma_{2,ii}^{-1} (\tilde{h}_i -$

$g_2)| \cdot \left|\eta(\tilde{h}_i) + \eta(g_2)\right|$. Applying the Cauchy-Schwarz inequality to the above inequality yields

$$\begin{aligned}
&\mathbb{E}\Big[\eta(\tilde{h}_i)^2 - \eta(g_2)^2\Big|\tilde{h}_i\Big] \lesssim \left(\mathbb{E}\left(|\Sigma_{2,ij}\Sigma_{2,ii}^{-1}(\tilde{h}_i - g_2)|^2\big|\tilde{h}_i\right)\right)^{1/2}\left(\mathbb{E}\left(\eta(g_2)^2 + \eta(\tilde{h}_i)^2\Big|\tilde{h}_i\right)\right)^{1/2} \\
&\lesssim |\Sigma_{2,ij}\Sigma_{2,ii}^{-1}|\sqrt{\Sigma_{2,ii} + \tilde{h}_i^2}\sqrt{\Sigma_{2,ij}^2\Sigma_{2,ii}^{-2}\tilde{h}_i^2 + \mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y| - \lambda_n)_+^2} \\
&\lesssim |\Sigma_{2,ij}|(1 + |\tilde{h}_i|)\left(|\Sigma_{2,ij}||\tilde{h}_i| + \sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y| - \lambda_n)_+^2}\right),
\end{aligned}$$

where the last step is due to $c_2 \le \Sigma_{2,ii} \le C_2$. Therefore, by Lemma 11, we have

$$\begin{aligned}
&\mathrm{Cov}\left((|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2, (|\delta_1^*\tilde{h}_j| - \lambda_n)_+^2\right) \\
\lesssim&|\Sigma_{2,ij}|(\lambda_n + 1)\sqrt{\mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y| - \lambda_n)_+^2}\mathbb{E}(|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2 \\
&+ \Sigma_{2,ij}^2(\lambda_n^2 + \lambda_n)\mathbb{E}(|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2.
\end{aligned}$$

Further, by Lemma 10 and Eq. (1.9), we have $\lambda_n = o(\log(p))$. The above inequality leads to:

$$\begin{aligned}
&\mathrm{Var}\left(c_n n^{-1}\tau^{-2}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right| - \lambda_n\right)_+\right\|^2\right) \\
&+ c_n^2 n^{-2}\tau^{-4}(\log(p))^2\sum_{i=1}^p \mathbb{E}(|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2\sum_{j=1}^p \Sigma_{2,ij}^2 \\
&\le c_n^2 n^{-2}\tau^{-4}\Bigg\{\log(p)C_2\sum_{i=1}^p \mathbb{E}(|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2\left(\sum_{j=1}^p \mathbb{E}_{Y\sim\mathcal{N}(0,1)}(|\delta_1^*\Sigma_{2,jj}Y| - \lambda_n)_+^2\right)^{1/2} \\
&+ (\log(p))^2 C_2^2\sum_{i=1}^p \mathbb{E}(|\delta_1^*\tilde{h}_i| - \lambda_n)_+^2\Bigg\} \\
&= O\left(c_n^{1/2}n^{-1/2}\tau^{-1}\log(p) + c_n n^{-1}\tau^{-2}\log^2(p)\right) = o_n(1),
\end{aligned}$$

where we use $\sum_{j=1}^p \Sigma_{2,ij}^2 \le C_2^2$ and Cauchy–Schwartz inequality in the second step. This

leads to Eq. (1.79). Using the same approach, we can prove

$$\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right| - \lambda_n + \log^{-1}(p)\right)_+\right\|^2 - \mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right| - \lambda_n + \log^{-1}(p)\right)_+\right\|^2 = o_{\mathrm{P}}(c_n^{-1} n\tau^2),$$
$$\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right| - \lambda_n - \log^{-1}(p)\right)_+\right\|^2 - \mathbb{E}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right| - \lambda_n - \log^{-1}(p)\right)_+\right\|^2 = o_{\mathrm{P}}(c_n^{-1} n\tau^2).$$

Now we are ready to establish Eq. (1.78). Note that w.p.a.1, we have

$$\begin{aligned} & c_n\tau^{-1}\left\|\left(\left|n^{-1/2}\tau^{1/2}\Sigma_2^{1/2}\delta h + \frac{\gamma}{\alpha^2}\tau^{-3/2}\Sigma_2\beta_0\right| - n^{-1/2}\tau^{-1/2}\lambda_n\right)_+\right\|^2 \\ \leq & 2c_n n^{-1}\tau^{-2}\left\|\left(\left|\Sigma_2^{1/2}\delta_1^* h\right| - \lambda_n + \log^{-1}(p)\right)_+\right\|^2, \end{aligned}$$

where the inequality is given by Lemma 12 and the facts that $\tau\log^4(p) = o(1)$ and that $n^{1/2}\tau^{1/2}p^{-1/2}q^{-1/2}\log^2(p) = o(1)$ by Assumption 4. Similarly, the left-hand-side is lower bounded by $\frac{1}{2}c_n n^{-1}\tau^{-2}\mathbb{E}\big\|\big(\big|\Sigma_2^{1/2}\delta_1^* h\big| - \lambda_n - \log^{-1}(p)\big)_+\big\|^2$ w.p.a.1. Finally, by Lemma 10 and the fact that $\lambda_n = o(\log(p))$, it is not hard to verify that $\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h| - \lambda_n - \log^{-1}(p)\big)_+\big\|^2 = \mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h| - \lambda_n\big)_+\big\|^2\big(1+o(1)\big)$ and $\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h| - \lambda_n + \log^{-1}(p)\big)_+\big\|^2 = \mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h| - \lambda_n\big)_+\big\|^2\big(1+o(1)\big)$. Together with the definition that

$$C_\lambda = \lim_{n\to\infty} pn^{-2}\tau^{-2}\mathbb{E}\big\|\big(|\Sigma_2^{1/2}\delta_1^* h| - \lambda_n\big)_+\big\|^2$$

given by Eq. (1.9), we conclude the proof. □

Lemma 26. *Eq. (1.43) defined in the proof of Theorem 5 holds as $n \to \infty$.*

Proof. Note that $\|\hat{\beta} - \widetilde{\beta}\|^2$ is upper bounded by

$$\begin{aligned}
&\frac{2}{n^2}\|R_U U^\top \mathcal{M}_W(U\beta_0 + \epsilon) - R_U U^\top (U\beta_0 + \epsilon)\|^2 + \frac{2}{n^2}\|(R_U - \widetilde{R}_U)U^\top (U\beta_0 + \epsilon)\|^2 \\
&\leq \frac{4}{n^2}\|R_U U^\top \mathcal{M}_W \varepsilon - R_U U^\top \varepsilon\|^2 + \frac{4}{n^2}\|R_U U^\top \mathcal{M}_W U\beta_0 - R_U U^\top U\beta_0\|^2 \\
&+ \frac{4}{n^2}\|(R_U - \widetilde{R}_U)U^\top \varepsilon\|^2 + \frac{4}{n^2}\|(R_U - \widetilde{R}_U)U^\top U\beta_0\|^2.
\end{aligned}$$

For the first term, using $\|R_U\| \lesssim_{\mathrm{P}} np^{-1}\tau$, $\|U^\top U\| \lesssim_{\mathrm{P}} p$, and Lemma 2, we have

$$\begin{aligned}
&\frac{4}{n^2}\|R_U U^\top \mathcal{M}_W \varepsilon - R_U U^\top \varepsilon\|^2 \asymp_{\mathrm{P}} \frac{1}{n^2}\operatorname{Tr}((\mathcal{M}_W U - U)R_U^2(U^\top \mathcal{M}_W - U^\top)) \\
=&\frac{1}{n^2}\operatorname{Tr}(U^\top W(W^\top W)^{-1}W^\top U R_U^2) \lesssim_{\mathrm{P}} p^{-1}\tau^2 \operatorname{rank}(W).
\end{aligned}$$

Similarly, it can be shown that the second term is of order $O_{\mathrm{P}}(p^{-1}\tau^2 \operatorname{rank}(W))$. In addition, by Lemma 2, and using the fact that $\operatorname{Tr}(AB) \leq \|A\| \operatorname{Tr}(B)$ and $\|\widetilde{R}_U\| \lesssim_{\mathrm{P}} np^{-1}\tau$, we have

$$\begin{aligned}
&\frac{4}{n^2}\|(R_U - \widetilde{R}_U)U^\top \varepsilon\|^2 \asymp_{\mathrm{P}} \frac{1}{n^2}\operatorname{Tr}(U(R_U - \widetilde{R}_U)^2 U^\top) \leq \frac{1}{n^2}\|U^\top U\| \operatorname{Tr}((R_U - \widetilde{R}_U)^2) \\
\lesssim_{\mathrm{P}}&\frac{p}{n^2}\operatorname{Tr}((R_U(\widetilde{R}_U^{-1} - R_U^{-1})\widetilde{R}_U)^2) = \frac{p}{n^4}\operatorname{Tr}((R_U U^\top W(W^\top W)^{-1}W^\top U\widetilde{R}_U)^2) \\
\leq&\frac{p}{n^4}\|R_U\|^2\|\widetilde{R}_U\|^2\|U^\top U\|^2 \operatorname{Tr}(W(W^\top W)^{-1}W^\top) \lesssim_{\mathrm{P}} p^{-1}\tau^4 \operatorname{rank}(W).
\end{aligned}$$

Similarly, the final term is of order $O_{\mathrm{P}}(p^{-1}\tau^4 \operatorname{rank}(W))$. To sum up, we have $\|\hat{\beta} - \widetilde{\beta}\|^2 = O(p^{-1}\tau^2 \operatorname{rank}(W)) = o(n^2 p^{-2}\tau^3)$, since $\operatorname{rank}(W) = o(n^2 p^{-1}\tau)$. □

CHAPTER 2

DEEP AUTOENCODERS FOR NONLINEAR FACTOR MODELS: THEORY AND APPLICATIONS

2.1 Introduction

Autoencoders (AEs) are specialized neural network (NN) models designed to replicate their inputs at their outputs, playing a fundamental role in unsupervised machine learning. These networks, which gained traction since the 1980s (e.g., LeCun [1987]), have a canonical architecture with two main components: an encoder, which compresses the inputs into a lower-dimensional representation known as features, codes, embeddings, or factors, and a decoder, which reconstructs the inputs from this compressed form.

We are particularly drawn to AEs due to their close connection with linear factor models and their capability in conducting nonlinear dimensionality reduction. It is well-documented (e.g., Baldi and Hornik [1989]) that a single-layer AE with linear activations is equivalent to Principal Component Analysis (PCA), with the number of neurons in that layer corresponding to the number of components. PCA has been extensively studied, both theoretically and empirically, demonstrating its effectiveness in estimating linear factor models widely used in macroeconomics [Stock and Watson, 1999, Bai and Ng, 2002] and finance [Chamberlain and Rothschild, 1983, Connor and Korajczyk, 1986]. Beyond factor analysis, PCA has also been applied in matrix completion (Bai and Ng [2021]) and causal inference (Agarwal and Singh [2024]). The success of PCA motivates extending theoretical guarantees to AEs, particularly for data-generating processes (DGPs) with nonlinear low-dimensional structures.

Despite the broad application of AEs in machine learning, there has been scant research providing theoretical justifications. To address this gap, our paper positions AEs as estimators for nonlinear factor models, setting the stage for a systematic investigation into their statistical properties. Within this framework, we explore several pivotal questions to

advance our understanding of deep and nonlinear AEs. These include whether AEs can effectively identify and extract common structures in the input data, and if so, what statistical error bounds characterize this recovery. We also examine how architectural parameters of AEs, such as depth, width, and the number of neurons, affect AEs' statistical performance. Furthermore, we investigate whether AEs can recover hidden low-dimensional embeddings inherent in nonlinear factor models. Addressing these questions theoretically lays the groundwork for broader methodological and applied uses of AEs across disciplines.

Our contributions are threefold. First, on model, we introduce a variant of the canonical AE architecture, termed Disjoint Output AEs. This architecture retains the fully connected encoder, with its output—the bottleneck layer—serving as the intermediary between the encoder and decoder. The bottleneck layer consists of multiple neurons, whose number corresponds to the dimensionality of the embeddings, akin to the number of factors in a linear factor model. The decoder departs from the standard fully connected design by using separate subnetworks to map the entire embedding to each individual output, thereby inducing a sparse-link structure. While this architecture is a special case of the fully connected decoder and does not expand the general approximation capacity of AEs, its structured sparsity enables us to address theoretical challenges more tractably. In particular, this architecture preserves the AE's capacity to approximate a broad class of nonlinear factor model DGPs, yielding desirable approximation errors—defined as the difference between the true nonlinear function and its optimal approximation within the AE class. Moreover, the reduced number of weight parameters accelerates training and lowers estimation error, defined as the difference between this optimal approximation and the estimated AE.

Second, on theory, we establish non-asymptotic guarantees for AE's convergence rates in approximating the common components of input data. Our analysis accommodates both increasing depth of AE architecture and width of any layer—including the embedding dimension. We demonstrate that AE's convergence rate is driven by two key components: the

approximation and estimation errors in recovering latent factors via the encoder, and the corresponding errors in reconstructing the input through the decoder.

The encoder's estimation error in recovering latent factors decreases with increasing input dimensionality, reflecting a form of the "blessings of dimensionality" observed in linear factor models. Importantly, the AE maintains an optimal convergence rate as long as the bottleneck layer's width is bounded—even if it exceeds the true dimensionality of the factors in the DGP. For the encoder's approximation error, we analyze two scenarios: in the first, the encoder is over-parameterized, achieving zero training error but risking divergence out of sample; in the second, the encoder is appropriately parameterized, achieving diminishing approximation errors in both in sample and out of sample, under a pervasiveness condition analogous to that in linear factor models.

The other source of error arises from the decoder and concerns the approximation and estimation of the mapping from latent factors to observables. As sample size grows, this component vanishes at the optimal nonparametric regression rate—as if the latent embeddings were directly observed. This is conceptually similar to error bound in nonparametric regression, as in our case the "loadings" are high-dimensional nonlinear functions.

Finally, our theoretical results further extend to the convergence rate of the factors themselves, which are identifiable up to invertible functional transformations. We show that AEs recover these latent factors at the same convergence rate, modulo such transformations, completing the theoretical foundation for nonlinear factor recovery using AEs.

Third, on empirical methodology, we extend our results to supervised AEs (SAEs), demonstrating their applicability in factor-augmented regressions and structured matrix completion, thereby broadening the AE framework's relevance in economics. We further illustrate the empirical potential of AEs and SAEs through three distinct applications: forecasting a panel of macroeconomic indicators; predicting the cross-section of factor returns; and conducting causal analysis with corrupted covariates. In all three cases, AEs significantly

outperform PCA due to their superior ability to extract nonlinear factors.

Our work contributes to the growing literature on nonlinear factor models, which—despite early appearances in psychology, notably by McDonald [1962]—have long posed significant challenges due to their inherent complexity. Traditional approaches often rely on parametric methods and operate under restrictive assumptions. More recently, building on the insight that certain nonlinear factor models can be effectively approximated by low-rank matrices, Agarwal et al. [2021] show that PCA remains robust in recovering common components even in nonlinear settings, while Freeman and Weidner [2023] apply this idea to estimate panel regressions with two-way unobserved heterogeneity featuring a nonlinear factor structure. Extending this line of inquiry, Feng [2023] considers a broader class of models originally proposed by Amemiya and Yalcin [2001], and introduces a local PCA method inspired by local regression techniques. This approach achieves a convergence rate consistent with our theoretical results. However, while this approach is effective at noise reduction, it falls short of constructing the underlying factors as the locally estimated components lack integration into a cohesive time series of factors. By contrast, the encoder of AEs provides a global solution for factor construction, integrating information across the entire dataset to form cohesive factors. Additionally, AEs are versatile, extending to various NN architectures that can capture complex data types, including text and images, making them broadly applicable across diverse domains.

Our paper adds to the growing body of literature on the theoretical properties of deep neural networks (DNNs). This literature builds on the early work of Barron [1993] and Chen and White [1999], which establish and refine nonparametric approximation rates for single-layer sigmoid neural networks. Chen and Shen [1998a] provide a general theory on the convergence rate of sieve extremum estimates for time series data, incorporating single-layer sigmoid neural networks as a special case. Yarotsky [2017] explores the optimal approximation errors in deep ReLU networks for a class of smooth functions. Building on this, Schmidt-Hieber

[2020] and Farrell et al. [2021] derive the optimal error rate for sparse DNNs, while Kohler and Langer [2021] demonstrate that fully connected DNNs can also achieve this optimal rate. Additionally, works by Bauer and Kohler [2019], Nakada and Imaizumi [2020], and Jiao et al. [2023] highlight the potential of DNNs to overcome the curse of dimensionality, particularly when the data exhibits intrinsic low-dimensionality or when the target function has a compositional structure. Unlike these studies, which focus on supervised learning problems for feedforward DNNs, our work investigates the theoretical properties of a specific class of DNNs—AEs—in unsupervised learning settings.

This chapter is organized as follows: Section 2.2 presents the nonlinear factor model and introduces Disjoint Output AEs. Section 2.3 presents the main theoretical results on the statistical properties of these AEs and extends the analysis to SAEs. Section 2.4 conducts simulations to validate our theoretical predictions and provide additional comparisons with Kernel PCA. The appendix and online appendix present empirical results that illustrate the practical applications of AEs, and include detailed mathematical proofs.

Notation: We denote the set of positive integers by $\mathbb{N}_+$. Furthermore, we represent the set that includes both zero and all positive integers, $\{0\} \cup \mathbb{N}_+$, by $\mathbb{N}_0$. For $n \in \mathbb{N}_+$, we use $[n]$ to denote the set $\{1, \ldots, n\}$. For a vector $x = (x_1, \ldots, x_d)^\top$, as usual, we define $\|x\|_p = (\sum_{i=1}^d |x_i|^p)^{\frac{1}{p}}$, $\|x\|_0 = \sum_{i=1}^d \mathbb{I}(x_i \neq 0)$, and $\|x\|_\infty = \max_i |x_i|$. Additionally, let $\mathcal{P}_n^d$ be the linear span of all monomials of the form $\prod_{k=1}^d x_k^{r_k}$ for some $r_1, \ldots, r_d \in \mathbb{N}_+$, where $r_1 + \cdots + r_d = n$. For a matrix $A \in \mathbb{R}^{n \times m}$, we use $\|A\|$, $\|A\|_0$, and $\|A\|_\infty$ to represent its spectral norm, the number of non-zero entries, and the maximum absolute value of its entries, respectively. Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, we define $\|f\|_\infty = \sup_{x \in \mathcal{D}} \|f(x)\|_\infty$, where $\mathcal{D}$ represents the domain of f. We say $f \in \mathcal{C}^\beta$ if it has β continuous derivatives on its domain. Furthermore, we define the norm $\|\cdot\|_{\mathcal{C}^\beta}$ of the smooth function space $\mathcal{C}^\beta$ by $\|f\|_{\mathcal{C}^\beta} := \max\left\{\|D^\alpha f\|_\infty : \|\alpha\|_1 \leq \beta, \alpha \in \mathbb{N}_0^d\right\}$. For a set $\mathcal{F}$, we use $|\mathcal{F}|$ to indicate the number of elements within the set. We use the notation $x_n \lesssim y_n$ when there exists a constant

C such that $x_n \leq Cy_n$ holds for sufficiently large n. If $x_n \lesssim y_n$ and $y_n \lesssim x_n$, we write $x_n \asymp y_n$ for short. For two random variables X and Y, we write $X \leq_d Y$ if Y stochastically dominates X and $X \overset{d}{=} Y$ if X has the same distribution as Y. For a sub-Gaussian variable X, its sub-Gaussian norm is defined as $\|X\|_{\psi_2} = \inf\left\{c > 0 : \mathrm{E}\left[\exp\left(X^2/c^2\right)\right] \leq 2\right\}$.

2.2 Model Setup

We begin by introducing the underlying DGP that will be used to characterize the statistical properties of AEs.

2.2.1 Nonlinear Factor Model

We model the observed data, X_{it}, for $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, based on a framework initially proposed by Amemiya and Yalcin [2001]:

$$X_{it} = X_{it}^{\star} + U_{it} = \varphi_i^{\star}(F_t^{\star}) + U_{it}, \tag{2.1}$$

where the common component, denoted by $X_{it}^{\star}$, is governed by a potentially nonlinear factor loading function $\varphi_i^{\star}(\cdot)$ of a K-dimensional vector of unknown factors, $F_t^{\star}$. The term U_{it} represents the noise component. We use X_t, $X_t^{\star}$, $\varphi^{\star}(F_t^{\star})$, and U_t to denote $N \times 1$ vectors containing the values for each entry of these variables at time t. Similarly, X, $X^{\star}$, and U represent their $N \times T$ matrix versions, and $F^{\star}$ denotes the $K \times T$ matrix of factors.

This model nests the linear factor model as a special case, where $\varphi_i^{\star}(x) = \Lambda_i^{\top} x$, and Λ_i is the $K \times 1$ column vector corresponding to the ith row of some $K \times N$ factor exposure matrix, Λ. Other examples include the polynomial factor model, where $\varphi_i^{\star}(x)$ is a multivariate polynomial function of x, and the additive nonlinear factor model, where $\varphi_i^{\star}(x) = \Lambda_i^{\top}\varphi(x)$ for some nonlinear function $\varphi(\cdot)$. This framework also encompasses generalized linear latent variable models, where $\varphi_i^{\star}(x) = \varphi(\Lambda_i^{\top} x)$, and the generalized factor model, in which $\varphi_i^{\star}(x) =$

$\varphi(\Lambda_i, x)$ allows for flexible interactions between factor loadings and latent factors.

Next, we make assumptions regarding the bounds on various norms of U_t and $F_t^\star$, as well as smoothness conditions on $\varphi_i^\star(\cdot)$. A fixed constant $B > 0$ is introduced to jointly bound relevant quantities in the assumptions below.

Assumption 6 (Boundedness)**.** *$F_t^\star$ is supported on a compact set $\mathcal{B} \subset [-B, B]^K$ almost surely for all t. Conditional on $F^\star$, $\operatorname{vec}(U) \overset{d}{=} \Sigma^{1/2} \operatorname{vec}(Z)$, where $Z \in \mathbb{R}^{N \times T}$ contains independent, zero-mean sub-Gaussian entries with sub-Gaussian norm bounded by σ_z^2. Moreover, the matrix Σ is positive semi-definite with bounded spectral norm.*

The bound on the support of $F_t^\star$ may seem stronger than what is typically assumed in the literature on linear factor models. However, this assumption is common in nonparametric literature (see, e.g., Chen and White [1999]) and is particularly important for developing theoretical results on NNs (e.g., Farrell et al. [2021]). In practice, this assumption is nearly harmless, though it excludes distributions supported on the entire real line.

This assumption permits heteroscedastic noise, allowing Σ to depend on $F^\star$. Additionally, $F_t^\star$ and U_t may follow specific time series models, such as autoregressive processes. Nevertheless, the imposed zero-mean condition on Z also implies strict exogeneity: the factors $F_i^\star$ are uncorrelated with the past, present, and future values of the idiosyncratic components U_j for all i and j. While this condition may seem restrictive, it serves primarily as a technical device for proofs and is unlikely to pose issues in empirical applications.

The bound on Σ's spectral norm rules out strong time-series and cross-sectional dependence among the entries of U. Nevertheless, it accommodates the model $U = \Sigma_1^{1/2} Z \Sigma_2^{1/2}$, where Σ_1 and Σ_2 are positive semi-definite matrices with bounded spectral norms. This "sandwich" structure has been employed by Onatski [2005] and Ahn and Horenstein [2013a] in their analysis of linear factor models.

Next, we impose a smoothness assumption on $\varphi_i^\star(\cdot)$. As is standard in the nonparametric literature (see, e.g., Stone [1982]), the degree of smoothness of the true underlying function

plays a crucial role in determining how accurately it can be approximated by NNs.

Definition 3 (Hölder Ball)**.** *Let $\beta > 0$ and Ω be a subset of some Euclidean space. The Hölder ball $\mathcal{H}^{\beta}(\Omega, C)$ is the set of scalar functions $f : \Omega \to \mathbb{R}$ such that*

$$\max_{|\alpha| \leq \lfloor \beta \rfloor} \sup_{x \in \Omega} |D^{\alpha} f(x)| + \max_{|\alpha| = \lfloor \beta \rfloor} \sup_{x \neq x'} \frac{|D^{\alpha} f(x) - D^{\alpha} f(x')|}{\|x - x'\|^{\beta - \lfloor \beta \rfloor}} \leq C,$$

where $\lfloor \beta \rfloor$ represents the largest integer strictly smaller than β. For a vector-valued function $f = (f_1, \ldots, f_d)^{\top} : \Omega \to \mathbb{R}^d$, we say $f \in (\mathcal{H}^{\beta}(\Omega, C))^d$ if $f_i \in \mathcal{H}^{\beta}(\Omega, C)$.

Assumption 7 (Smoothness)**.** *There exist $\beta \in \mathbb{N}_+$ such that $\varphi^{\star}(\cdot) \in (\mathcal{H}^{\beta}(\mathcal{B}, B))^N$.*

If the factors $F_t^{\star}$ were observable, estimating $\varphi_i^{\star}(\cdot)$ would reduce to a standard (supervised) nonparametric regression, approachable with methods such as sieve estimators. However, with unobserved factors, the problem is both unsupervised and nonparametric, requiring a different approach. We now proceed to construct AEs to jointly estimate the latent factors and the nonlinear functions.

2.2.2 Autoencoders and Their Architecture

To lay the foundation for constructing AEs, we first introduce the key building blocks of NNs. We begin with the rectified linear unit (ReLU) activation function, defined as $\sigma(x) = \max(x, 0)$. Given $v = (v_1, \ldots, v_r) \in \mathbb{R}^r$, we define the shifted activation function $\sigma_v : \mathbb{R}^r \to \mathbb{R}^r$ as follows:

$$\sigma_v \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} \sigma(y_1 - v_1) \\ \vdots \\ \sigma(y_r - v_r) \end{pmatrix}.$$

An NN's architecture, denoted by (d, w, n), consists of the depth $d \in \mathbb{N}$, representing the number of hidden layers, and the width $w \in \mathbb{N}$, which bounds the number of neurons per

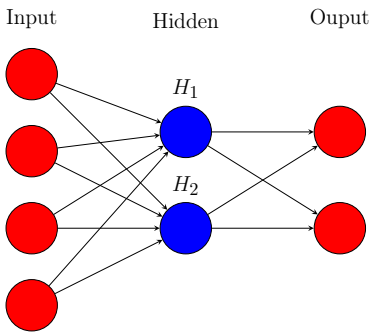


Figure 2.1: Illustration of Neural Network

Note: A fully connected neural network with architecture $d = 1$, $w = 2$, and $n = (4, 2, 2)$. Input and output neurons are in red; hidden-layer neurons are in blue.

hidden layer. The width vector $n = (n_0, \ldots, n_{d+1}) \in \mathbb{N}^{d+2}$ satisfies $n_i \leq w$ for $i = 1, \ldots, d$, where n_0 and n_{d+1} denote the input and output dimensions, respectively. An NN with architecture (d, w, n) defines a function

$$f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{d+1}}, x \to f(x) = W_d \sigma_{v_d} W_{d-1} \sigma_{v_{d-1}} \cdots W_1 \sigma_{v_1} W_0 x, \tag{2.2}$$

where each W_i is an $n_{i+1} \times n_i$ weight matrix, and σ_{v_i} is the shifted activation function with shift vector $v_i \in \mathbb{R}^{n_i}$.[1] This notation refers to a fully connected NN; if some connections are dropped, the corresponding weights in W_i are set to zero. We refer to NNs with possibly diverging depth as DNNs. Figure 2.1 illustrates an NN with architecture $(1, 2, (4, 2, 2))$.

We follow a setup similar to Schmidt-Hieber [2020], bounding the parameters of the DNN to ensure well-behaved networks. Specifically, we define the class of DNNs as:

$$\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C, B) := \left\{ f \text{ of the form (2.2)} : \max_{j=0,\ldots,d} \|W_j\|_\infty \vee |v_j|_\infty \leq C,\ \|f\|_\infty \leq B \right\},$$

where B is specified in Assumption 6.[2]

1. We omit parentheses in $\sigma_{v_i}(\cdot)$ when clear from context.

2. For simplicity, we omit the dependence of $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C, B)$ on the width vector n.

The first condition bounds the weight matrices and shift vectors by C. While Schmidt-Hieber [2020] fix $C = 1$, we let $C = T^{5\beta+5}$, allowing it to scale with the sample size. This improves DNNs' approximation accuracy without inflating the estimation error.

The second condition imposes a uniform bound on the network outputs, independent of the sample size or architecture. That is, we restrict attention to DNNs with bounded outputs when used as estimators. This aligns with the boundedness of the true function $\varphi_i^\star(\cdot)$ in the DGP, as implied by Assumption 7. In practice, this constraint acts not as a strong restriction but as a mild form of regularization.

Having introduced the basic architecture of a DNN, we are now ready to define a special class of DNNs known as AEs. The primary objective of an AE is to reconstruct the input from their compressed form, thereby facilitating dimensionality reduction.

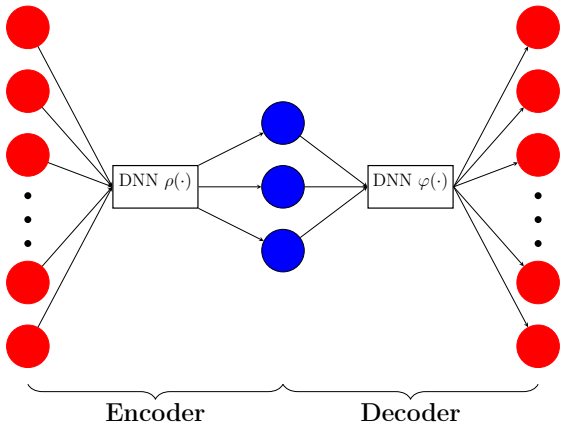


Figure 2.2: Illustration of Canonical Autoencoder Architecture

Note: This graph illustrates a canonical autoencoder architecture. The input is passed through a series of hidden layers that compress the data into a central bottleneck layer—the narrowest part of the network—shown here with three neurons in blue. The network then reconstructs the input through additional layers, producing an output of the same dimension as the input.

Specifically, given an input x, an AE first applies an *encoder*, a DNN denoted by $\rho(\cdot)$, which maps the input into a low-dimensional *bottleneck*, represented by a small number of neurons in a hidden layer—this being the narrowest layer of the AE. The encoded data at the bottleneck layer is then passed through a second DNN, called the *decoder*, denoted by

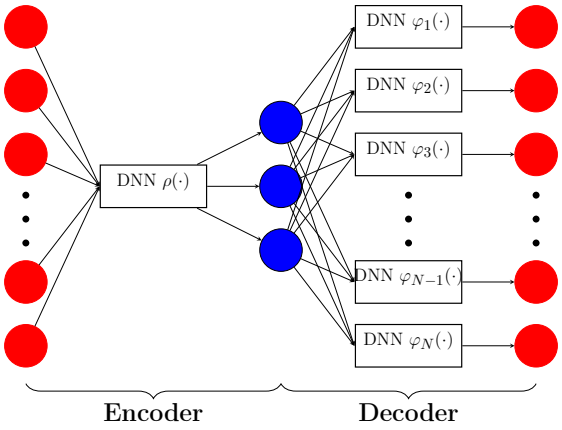


Figure 2.3: Illustration of Disjoint Output Autoencoder

Note: This graph illustrates a disjoint output autoencoder architecture. The encoder retains the same structure as the canonical autoencoder's encoder shown in Figure 2.2. The decoder, however, consists of multiple separate DNNs, each mapping the bottleneck layer to a single output.

$\varphi(\cdot)$, which reconstructs an approximation of the original input x at the output. Figure 2.2 illustrates the canonical architecture of a vanilla AE.

Unlike standard DNNs used in supervised learning to predict an external target, AEs are trained to reproduce the input itself. The input and output dimensions are the same, and no additional variables are required, making AEs a natural tool for unsupervised learning.

In this paper, we shift our focus from the canonical form of AE to what we refer to as a *disjoint output AE*, with its architecture depicted in Figure 2.3. This variant retains the conventional encoder but replaces the decoder with a collection of independent DNNs, each mapping from the bottleneck to a single output neuron.

More specifically, the encoder, a single DNN denoted as $\rho(\cdot) \in \mathcal{F}_N^{K_1}(d_1, w_1, T^{5\beta+5}, B)$, produces a K_1-dimensional bottleneck layer. For each of the N outputs, a separate DNN is employed, where each $\varphi_i(\cdot) \in \mathcal{F}_{K_1}^1(d_2, w_2, T^{5\beta+5}, B)$ is specific to the ith output neuron for $i \in [N]$. As a result, for an input x, the i^{th}-output of the disjoint output AE is given by

$\varphi_i \circ \rho(x)$. Formally, we define the disjoint output AE function class as follows:

$$\mathcal{F}_{AE}^{K_1} := \Big\{(\varphi_1, \ldots, \varphi_N) \circ \rho : \rho \in \mathcal{F}_N^{K_1}(d_1, w_1, T^{5\beta+5}, B), \\ \varphi_1, \ldots, \varphi_N \in \mathcal{F}_{K_1}^1(d_2, w_2, T^{5\beta+5}, B)\Big\}. \tag{2.3}$$

Consider the canonical form of an AE with a decoder of the same width as that of the disjoint output AE, denoted as $\varphi \circ \rho(x)$, where the encoder is $\rho(\cdot) \in \mathcal{F}_N^{K_1}(d_1, w_1, T^{5\beta+5}, B)$, and the decoder $\varphi(\cdot) \in \mathcal{F}_{K_1}^N(d_2, Nw_2, T^{5\beta+5}, B)$ is fully connected. The total number of weight parameters in this decoder is of order $O((Nw_2)^2 d_2)$. In contrast, the decoder $(\varphi_1, \ldots, \varphi_N)$ of the disjoint output AE in $\mathcal{F}_{AE}^{K_1}$ contains at most $N \times O((w_2)^2 d_2)$ parameters. This design introduces sparsity in the decoder's weights, reducing the parameter count by a factor of N compared to a canonical AE with the same width. While sparsity can also be induced in a fully connected decoder through ℓ_1-regularization during training, the disjoint output AE achieves this reduction structurally, without the need for regularization.

However, this parameter reduction introduces a structural constraint: the disjoint output decoder requires a wider architecture with a potentially excessive number of neurons. For example, if each subnetwork $\varphi_i(\cdot)$ shares the same architecture and includes at least one hidden layer with at least one neuron, then the combined hidden layer across all N subnetworks must contain at least N neurons—making it wider than the output layer. This rigidity stems from the absence of parameter sharing in the disjoint design. In contrast, a fully connected decoder may, in principle, choose a width arbitrarily smaller than N, thanks to parameter sharing across output dimensions.

As we will explain later, despite its architectural rigidity, the disjoint output AE delivers strong approximation performance. It is flexible enough to capture complex, low-dimensional nonlinear structures in the input while substantially reducing the number of weight parameters. This reduction also helps control model complexity, leading to favorable statistical estimation properties. While fully connected decoders may achieve similar

approximation and estimation with fewer parameters—potentially mitigating error through parameter sharing—we focus on the disjoint architecture for its analytical tractability.[3]

The final step in constructing an AE is to specify a loss function that quantifies the reconstruction error. This is typically done by minimizing the following objective:

$$(\widehat{\varphi}_1, \ldots, \widehat{\varphi}_N) \circ \widehat{\rho} = \underset{(\varphi_1,\ldots,\varphi_N)\circ\rho \in \mathcal{F}_{AE}^{K_1}}{\arg\min} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(\varphi_i(\rho(X_t)) - X_{it}\right)^2 . \tag{2.4}$$

This optimization problem is highly non-convex, necessitating the use of more sophisticated algorithms for solving it. We will explore the optimization perspective in greater detail in the simulation and empirical analysis sections.

Before training an AE, it is necessary to select a key hyperparameter: the number of neurons in the bottleneck layer, K_1. The overall network architecture must also be designed, including decisions about its depth and width. These crucial considerations will be discussed after we present the theoretical analysis, which we turn to now.

2.3 Main Theoretical Results

In this section, we present a non-asymptotic analysis of the statistical properties of the disjoint output AEs. Throughout, we assume that all constants are independent of N and T, but may depend on the latent dimension K and the smoothness parameter β, which we do not explicitly track. Our analysis begins with the reconstructed data, $\widehat{X}_{it} := \widehat{\varphi}_i(\widehat{\rho}(X_t))$. where $\widehat{\rho}(\cdot)$ and $\widehat{\varphi}_i(\cdot)$ for $1 \leq i \leq N$ are defined in equation (2.4).

3. This tractability arises from two considerations: first, the reduction in approximation error due to parameter sharing in a fully connected decoder remains analytically intractable; second, the impact of parameter sharing on out-of-sample estimation error is not well understood. Leveraging a disjoint output structure eliminates the need to analyze the complexities associated with parameter sharing.

2.3.1 Recovery of the Common Components

Naturally, the population counterpart of $\widehat{X}_{it}$ is $X^{\star}_{it} = \varphi^{\star}_i(F^{\star}_t)$ in the DGP, as specified in equation (2.1). Accordingly, we focus on the mean squared error aggregated across the entire panel: $(NT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it} - X^{\star}_{it}\right)^2$.[4]

Theorem 6. *Consider the AE class $\mathcal{F}_{AE}^{K_1}$ with $K \leq K_1 \leq \min(w_1, w_2)$, $d_2 \asymp \log(T)$, and $w_2 \asymp T^{\frac{K}{2(2\beta+K)}}$. Suppose Assumptions 6-7 hold. Then, with probability at least $1 - C\exp(-cT)$, for $\min(N,T)$ large enough, we have:*

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it} - X^{\star}_{it}\right)^2 \lesssim \left(T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1 + T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F^{\star}_t\|^2\right)\log^4(T),$$

where c and C are constants independent of N, T, and $\mathcal{F}_N^K := \mathcal{F}_N^K(d_1, w_1, T^{5\beta+5}, B)$.

We first discuss the assumptions. The conditions require $K_1 \geq K$, meaning the bottleneck layer must have at least as many neurons as the number of factors specified in the DGP. However, the notion of "number of factors" in a nonlinear factor model is somewhat ambiguous—a point we revisit in detail in Section 2.3.2. The condition $K_1 \leq \min(w_1, w_2)$ ensures the bottleneck is among the narrowest layers in the network, allowing for ties. On the decoder side, both d_2 and w_2 must increase with T to guarantee sufficient flexibility. Finally, the results only require that N and T exceed a fixed constant, so the non-asymptotic bound directly implies consistency under the Frobenius norm as N and T increase.

As is standard in nonparametrics, this result hinges on appropriately selecting the dimension of the bottleneck layer, K_1, and the width of the decoder, w_2, based on the smoothness parameter β and the number of factors K in the true DGP. In practice, these values are typically unknown, so conservative choices are often made. This entails assuming the true functions are smoother than specified and deliberately using a relatively large number of

4. While deriving a uniform error bound across i and t would also be meaningful, such results are, to the best of our knowledge, not yet established for NN settings.

factors—exceeding what theory or economists would typically propose or expect.

Next, we make a few observations regarding the result on the error bound. There are three terms on the right-hand side of the inequality. The first term in the error, $T^{-\frac{2\beta}{2\beta+K}}$, matches the minimax rate in nonparametric regression of X_t on $F_t^\star$, as if these factors were known. For a nonparametric supervised learning task, this rate can be achieved by estimators other than NNs; examples include spline methods by Speckman [1985] or sieve estimators by Newey [1997] and Chen and Shen [1998a], among others.

The second term, $N^{-1}K_1$, arises from estimation error associated with unknown factors. Intuitively, since there are TK_1 unknown values to estimate and NT observed values in total, the ratio of unknowns per observation yields this rate. Our assumptions permit the number of neurons, K_1, in the bottleneck layer to grow, and the results show that while selecting more factors than necessary slows convergence, the effect is only linear in K_1, indicating that the model remains reasonably robust to overestimating the number of factors.

The third term reflects the error from approximating the unknown factors using the encoder. An important observation is that the estimation error for the encoder does not impact the overall convergence rate—only its approximation error is relevant. In light of this, an over-parameterized encoder can be employed such that $\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^T\|\rho(X_t)-F_t^\star\|^2=0$, eliminating the approximation error and thereby improving the overall convergence rate.

However, while an over-parameterized encoder can achieve strong in-sample performance, it may result in poor expected out-of-sample performance, as both estimation and approximation errors can arise and become substantial.[5] A more meaningful strategy is to balance these two error sources out-of-sample when designing the encoder, even if this leads to a larger approximation error in-sample. To achieve this, we need a more explicit characteriza-

5. Recent literature indicates that under certain conditions on the data, overfitting can be benign; see, for example, Bartlett et al. [2019b], Tsigler and Bartlett [2024], and Hastie et al. [2022]. However, these analyses are mainly focused on linear settings for supervised learning problems, and a comprehensive theoretical framework for NNs remains unavailable. We leave the investigation of benign overfitting in unsupervised learning problems for future research.

tion of approximation error. To this end, we introduce the following notation to characterize the trade-off between estimation and approximation error out-of-sample.

Definition 4. *For any positive sequence L_N diverging with N, define*

$$\epsilon_N := \inf_{\substack{\|W\|_0 \leq LK, \|W\varphi^\star\|_\infty \leq B \\ \rho(\cdot) \in \mathcal{H}^\beta(W\varphi^\star(\mathcal{B}), B)}} \left\{ \sup_{x \in \mathcal{B}} \|\rho(W\varphi^\star(x)) - x\|^2 + \|W\|^2 \right\},$$

where W is any $K \times N$ matrix, and $\rho(\cdot)$ is a mapping from $W\varphi^\star(\mathcal{B})$ to $\mathbb{R}^K$.

By construction, the term ϵ_L is decreasing in L and explicitly depends on the high-dimensional function $\varphi^\star$.

To illustrate the behavior of ϵ_L, we provide two concrete examples, both leading to an error $\epsilon_L \asymp L^{-1}$. First we consider a setting as in Bonhomme et al. [2022], where $\varphi^\star(x) = \alpha(\lambda_i, x)$ with scalar arguments $\lambda_i, x \in \mathbb{R}$. This covers widely used fixed-effect model where $\alpha(\lambda_i, x) = \lambda_i + x$. Suppose α is strictly increasing in its second argument, analogous to the injectivity condition in Bonhomme et al. [2022, footnote 6]. By setting $W^\star = (L^{-1}1_L^\top, 0_{N-L}^\top)^\top$, the function $W^\star\varphi^\star(x)$ remains strictly monotone in x, ensuring the existence of an inverse function $\rho^\star$ mapping $W^\star\varphi^\star(x)$ back to x. Hence, in this scenario, we obtain $\epsilon_L \asymp L^{-1}$.

The second example is a nonlinear generalization of linear factor models, let $\varphi^\star(x) = \exp(\Lambda^\top F_t)$, where Λ consists of bounded i.i.d. random variables. Constructing

$$\widetilde{\Lambda} = (\Lambda_1^\top, \ldots, \Lambda_L^\top, 0_K^\top, \ldots, 0_K^\top)$$

ensures invertibility of $\widetilde{\Lambda}^\top\widetilde{\Lambda}$. Setting $W^\star := L^{-1}\widetilde{\Lambda}^\top$, the function $W^\star \exp(\widetilde{\Lambda}x)$ has a positive definite Jacobian, guaranteeing invertibility via the inverse function theorem. Thus, there exists a smooth inverse function $\rho^\star(\cdot)$ satisfying $\rho^\star(W^\star \exp(\widetilde{\Lambda}x)) = x$, again yielding $\epsilon_L \asymp L^{-1}$.

Theorem 7. *Under the conditions of Theorem 6, , if* $d_1 \asymp \log(T)$, $w_1 \asymp T^{\frac{K}{2(2\beta+K)}}$, *and* $s \asymp L + T^{\frac{K}{2\beta+K}} \log T$, *and that the data is i.i.d., then for a new data point* X_{T+1}, *we have*

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left(\widehat{\varphi}_i \circ \widehat{\rho}(X_{T+1}) - X^{\star}_{T+1}\right)^2 \leq \left(T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1 + \epsilon_L + T^{-1}L\right)\log^4(NT).$$

Theorem 7 reveals that, under the assumption of i.i.d. data, the expected out-of-sample loss includes not only the in-sample approximation error but also an additional term, $T^{-1}s$, which reflects the estimation error in the encoder and worsens as its complexity, measured by its number of parameters, s, increases.[6]

That said, implementing ℓ_0-regularization to induce sparsity in NNs is known to be challenging, as noted in Schmidt-Hieber [2020] and Farrell et al. [2021]. Consequently, while the optimal rate is theoretically attainable, it may be difficult to achieve in practice. A common empirical workaround is to incorporate a pruning step during training, which we adopt and further elaborate on in our implementation.

2.3.2 Recovery of the Factors

In this section, we provide the theoretical justification for the potential of AEs to recover underlying factors in a nonlinear DGP. An important motivation for using AEs stems from dimensionality reduction, which aims to extract a lower-dimensional set of features that effectively capture the underlying structure of the data. By reducing the dimensionality, we can highlight the most informative aspects of the data, making patterns more interpretable and facilitating downstream tasks such as clustering, visualization, and forecasting.

We begin by noting that for any injective mapping $\mu : \mathbb{R}^K \to \mathbb{R}^K$, the DGP in equation

6. The i.i.d. assumption is imposed for technical simplicity; the trade-off itself remains valid beyond this setting.

(2.1) can be equivalently expressed as

$$X_{it} = \varphi_i^\star \circ \mu^{-1} \circ \mu(F_t^\star) + U_{it}.$$

This formulation implies that $\mu(F_t^\star)$ can also serve as valid factors, meaning that the original factors $F_t^\star$ are only identified up to a nonlinear invertible transformation. This ambiguity is similar to the linear case, where factors can only be identified up to an invertible matrix transformation.

The next theorem presents the convergence rate on factor recovery:

Theorem 8. *Under the same conditions as in Theorem 7, there exists a function* $\mu : \mathbb{R}^{K_1} \to \mathbb{R}^K$ *such that, with probability at least* $1-C\exp(-cT)-C\exp(-cL)$*, for* N *and* T *sufficiently large,*

$$\frac{1}{T}\sum_{t=1}^{T} \|\mu(\widehat{F}_t) - F_t^\star\|^2 \lesssim (L^{-1}K_1 + NL^{-1}T^{-\frac{2\beta}{2\beta+K}} + NL^{-2} + NL^{-1}\epsilon)\log^4(T).$$

The result indicates that when the factors are strong, meaning $L = N$, the convergence rate aligns with that in Theorem 7, and the estimator achieves consistency. However, if the factors are extremely weak—affecting only a small number of variables in X, such as when $L \asymp 1$—the right-hand side fails to converge to zero, and consistent estimation of the factors may not even be possible.

It is insightful to compare our result with that of Bai and Ng [2023b], which examines the convergence of factor estimates using PCA when the true factors are weak. Using our notation, their result suggests an error rate of $L^{-1} + (NL^{-1}T^{-1})^2$ in the linear case. In contrast, our rate, when $\epsilon \asymp L^{-1}$, $L^{-1} + (NL^{-1}T^{-1}) + NL^{-2}$, is not sharp due to the challenges associated with analyzing the outputs of a specific hidden layer in a DNN.

Our result does not require the number of neurons in the bottleneck layer to match the number of factors in the DGP. Instead, it only assumes the existence of a DGP specified in

(2.1), where the number of factors K is less than or equal to our chosen K_1. Identifying the number of factors is a critical problem in (linear) factor analysis. However, determining the true number of factors in nonlinear models is significantly more challenging because of the inherent ambiguity, even at the population level.

For context, consider a linear factor model given by $X_t = \Lambda F_t^\star + U_t$, where $F_t^\star \in \mathbb{R}^K$. Under the conditions that the minimal eigenvalues of $N^{-1}\Lambda^\top\Lambda$ and $\mathrm{Cov}(F_t^\star)$ are asymptotically bounded from below, and that the eigenvalues of $\mathrm{Cov}(U_t)$ are bounded from above, the number of factors K can be readily identified, as demonstrated by Chamberlain and Rothschild [1983]. Bai and Ng [2002] also provide a consistent procedure for recovering K.

In nonlinear factor models, however, the notion of the number of factors becomes ambiguous. According to Theorem 2 from Schmidt-Hieber [2021], which extends the Kolmogorov–Arnold representation theorem, any nonlinear multivariate function can be reduced to a single-variable form. Specifically, there exists a function $\psi : [-B, B]^K \to \mathbb{R}$, such that for any function $\varphi_i^\star : [-B, B]^K \to \mathbb{R}$, a corresponding function $\widetilde{\varphi}_i^\star : \mathbb{R} \to \mathbb{R}$ exists, satisfying

$$\varphi_i^\star(x_1, \ldots, x_K) = \widetilde{\varphi}_i^\star \left(3 \sum_{j=1}^{K} 3^{-j} \psi(x_j) \right).$$

This implies that the original multi-factor model can be reconstructed using a single factor, denoted as $\widetilde{F}_t \in \mathbb{R}$, where $\widetilde{F}_t$ is defined by $3\sum_{j=1}^{K} 3^{-j}\psi(F_{jt}^\star)$. Consequently, the function $\varphi_i^\star(F_t^\star)$ can be rewritten as $\widetilde{\varphi}_i^\star(\widetilde{F}_t)$. Although this representation reduces the number of factors, there are, to the best of our knowledge, no existing results concerning the smoothness properties of $\widetilde{\varphi}_i^\star(\cdot)$ following this transformation. It is likely that the function becomes less smooth in general. Otherwise, this would provide a means to surpass the minimax rate, which depends on K/β. Consequently, we cannot use this result to justify always adopting a single neuron in the bottleneck layer, since our approach relies on the smoothness of the nonlinear functions.

In some cases, however, it may be possible to reformulate the original K-factor model in a way that introduces additional factors while yielding smoother $\varphi_i^\star(\cdot)$. To illustrate, consider a single factor model defined as $\varphi_i^\star(F_t^\star) = \Lambda_i F_t^\star + \psi(F_t^\star)$, where $\psi(\cdot)$ is a β-smooth function. By treating $\psi(F_t^\star)$ as an additional factor, the model becomes a two-factor linear model with an effectively infinite degree of smoothness. This modification enables a convergence rate of $N^{-1} + T^{-1}$, which is faster than the rate $N^{-1} + T^{-\frac{2\beta}{2\beta+1}}$ achievable by fitting a single-factor nonlinear model.

Given these considerations, the conventional notion of the number of factors in a nonlinear factor model becomes ambiguous and may need to be defined alongside the smoothness of the factor loading functions. In practice, we treat it as one of the hyperparameters in the architecture of AEs, determining it through a model selection procedure. This approach makes intuitive sense and aligns with common practice, though providing a formal justification is left for future work.

2.3.3 Extensions and Applications

In this section, we delve into an extension of AEs that pave the way for a wider range of applications, enhancing their flexibility and utility.

2.3.3.1 Factor-Augmented Prediction

When dealing with large datasets, we often encounter scenarios where we not only wish to reconstruct or compress the original data but also make accurate predictions for related outcomes. Standard AEs excel at capturing latent structures in data through unsupervised learning, but they do not directly incorporate information that could enhance predictive performance for external targets. This is where supervised extensions of AEs come into play. By integrating predictive tasks into the AE framework, we can leverage shared low-dimensional structure to improve both reconstruction and prediction, making the model

more versatile and powerful for applications where joint modeling of input and output data is beneficial.

Consider the scenario where we also observe a large panel $Y_t \in \mathbb{R}^M$ that we wish to predict, in addition to reconstructing X_t. We assume that Y_{t+1} depends on the same latent factors, $F_t^\star$, modeled as:

$$Y_{i,t+1} = \phi_i^\star(F_t^\star) + V_{i,t+1}, \quad \text{for} \quad t = 1, 2, \ldots, T. \tag{2.5}$$

To train this model, one approach is to perform a nonparametric regression of $Y_{i,t+1}$ on the factors extracted from AEs pre-trained solely with X_t. Alternatively, a more cohesive method involves jointly fitting X_t and Y_{t+1} while training the AE. This strategy originates from Le et al. [2018], who introduce a supervised autoencoder (SAE) architecture. In this setting, X_t remains the input, and the model is trained to simultaneously construct X_t and predict Y_{t+1}, leveraging the shared latent structure. Figure 2.4 presents the architecture of the SAE.

This SAE approach differs from factor-augmented regressions (as in Bernanke and Boivin [2003]) in two key aspects. First, the SAE framework is nonparametric, enabling flexible and complex nonlinear relationships between the target and the factors. Second, the high dimensionality of the target in SAEs imposes meaningful supervision on the factor construction process, enhancing the model's predictive power. In contrast, traditional factor-augmented regressions typically use low-dimensional targets, rely on linear DGPs, and construct factors without incorporating information from the target.

Suppose we observe the sequences X_t and Y_{t+1}, for $t = 1, 2, \ldots, T-1$, and our goal is

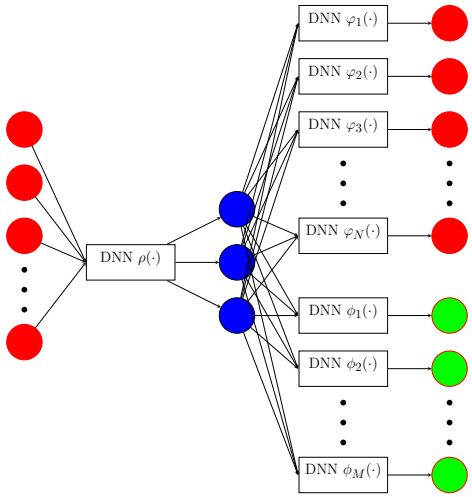


Figure 2.4: Illustration of Disjoint Output Supervised Autoencoder

Note: This graph illustrates a disjoint output supervised autoencoder architecture. The encoder retains the same structure as the canonical autoencoder's encoder shown in Figure 2.2. The decoder, however, consists of multiple separate DNNs, each responsible for mapping the bottleneck layer to a single output. The portion of the output highlighted in red is aimed at reconstructing the input, using the same architecture as shown in Figure 2.3. The other portion, marked in green, is designed to predict additional target variables, while supervising the factor construction process.

to predict Y_{T+1}. To achieve this, we first train an SAE using the following loss function:

$$\begin{aligned}
&(\widehat{\varphi}_1, \ldots, \widehat{\varphi}_N, \widehat{\phi}_1, \ldots, \widehat{\phi}_M) \circ \widehat{\rho} \\
&= \underset{(\varphi_1,\ldots,\varphi_N,\phi_1,\ldots,\phi_M)\circ\rho\in\mathcal{F}_{SAE}^{K_1}}{\arg\min} \sum_{t=1}^{T-1} \left(\sum_{i=1}^{N} \left(\varphi_i(\rho(X_t)) - X_{it}\right)^2 + \sum_{j=1}^{M} \left(\phi_j(\rho(X_t)) - Y_{j,t+1}\right)^2 \right).
\end{aligned} \tag{2.6}$$

The supervised AE function class $\mathcal{F}_{SAE}^{K_1}$ is defined as:

$$\begin{aligned}
\mathcal{F}_{SAE}^{K_1} := \Big\{ &(\varphi_1, \ldots, \varphi_N, \phi_1, \ldots, \phi_M) \circ \rho : \rho \in \mathcal{F}_N^{K_1}(d_1, w_1, T^{5\beta+5}, B), \\
&\varphi_i \in \mathcal{F}_{K_1}^1(d_2, w_2, T^{5\beta+5}, B), i \in [N], \phi_j \in \mathcal{F}_{K_1}^1(d_2, w_2, T^{5\beta+5}, B), j \in [M] \Big\}.
\end{aligned}$$

The predictor for $Y_{j,T+1}$, denoted by $\widehat{Y}_{j,T+1}$, is thus given by $\widehat{\phi}_j(\widehat{\rho}(X_T))$. To understand its

statistical performance, we establish the following corollary:

Corollary 2. *Assume $\phi_i^\star \in \mathcal{H}^\beta(\Omega, B)$ for $i = 1, \ldots, M$. In addition, we assume that conditional on $F^\star$, $(U_t^\top, V_{t+1}^\top)^\top \in \mathbb{R}^{N+M}$ is i.i.d. and takes the form $\overline{\Sigma}^{1/2}\overline{Z}_t$, where $\overline{Z}_t \in \mathbb{R}^{N+M}$ consists of independent sub-Gaussian random variables with sub-Gaussian norm bounded by σ_z^2 and $\overline{\Sigma} \in \mathbb{R}^{N+M}$ is positive semi-definite with bounded spectral norm. Under the same conditions as in Theorem 7, as $\min(N, M, T)$ becomes sufficiently large, we have:*

$$\begin{aligned}&\frac{1}{N+M}\left(\sum_{i=1}^{N}\mathbb{E}\big(\widehat{X}_{i,T} - X_{i,T}^\star\big)^2 + \sum_{j=1}^{M}\mathbb{E}\big(\widehat{Y}_{j,T+1} - \phi_j^\star(F_T^\star)\big)^2\right)\\ &\lesssim ((N+M)^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + L^{-1} + \epsilon + T^{-1}L)\log^4((N+M)T). \qquad (2.7)\end{aligned}$$

Consequently, when $N \lesssim M$, we obtain:

$$\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}\big(\widehat{Y}_{j,T+1} - \phi_j^\star(F_T^\star))\big)^2 \lesssim (M^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + L^{-1} + \epsilon + T^{-1}L)\log^4(MT). \quad (2.8)$$

The intuition behind equation (2.7) is that the SAE model can be regarded as a special case of the AE, where the input is (X_t, Y_{t+1}) and the weights for Y_{t+1} in the encoder are fixed to zero. Compared to the AE discussed in Theorem 7, the SAE's encoder approximation and estimation errors remain $L^{-1} + T^{-1}L$, as the true factors can be effectively approximated using X_t alone, without requiring additional parameters to be estimated required for the encoder. The decoder's error maintains the same structure as in Theorem 7, except that N is replaced by $N + M$, leading to the final convergence rate.

Equation (2.8) in this corollary follows directly from equation (2.7). It implies that when the dimension of X is of a smaller or comparable order to the dimension of Y, the prediction error for $\widehat{Y}_{T+1}$ vanishes as the sample size increases, allowing the model to effectively leverage information from Y for accurate predictions. Conversely, if M is significantly smaller than N, the contribution of Y_t during training diminishes, potentially compromising the prediction

accuracy for $\widehat{Y}_{T+1}$.

2.3.3.2 Matrix Completion

AEs provide a powerful framework for tackling problems involving nonlinear matrix completion and missing data. In many real-world applications, such as recommendation systems, financial time series, or large-scale survey data, we encounter datasets that are both high-dimensional and partially observed, sometimes with missing values scattered throughout. Traditional matrix completion methods, such as those proposed by Candes and Recht [2009], Cai et al. [2010], Keshavan et al. [2010], Recht [2011], and Mazumder et al. [2010], typically rely on low-rank (or equivalently, linearity) assumptions, which may struggle to capture the complex, nonlinear relationships inherent in such data.

AEs offer a natural solution to this challenge by learning a compact, latent representation of the observed data that can encode nonlinear structures effectively. By training the AE to reconstruct the original matrix from its latent representation, the model can infer and fill in missing entries in a way that captures intricate patterns. This capability is crucial not only for matrix completion but also for broader applications, such as causal inference.

In the estimation of causal effects within a panel setting, counterfactual outcomes—representing what would have occurred under alternative treatment conditions—can be regarded as missing data. AEs can assist in this context by imputing these unobserved counterfactuals, offering a novel and effective approach to estimating causal relationships (see, e.g., Athey et al. [2017], Bai and Ng [2021]). Importantly, missing data in these scenarios are often not random. In many instances, the data can be structured so that all missing entries are concentrated within a sub-block of a matrix. Therefore, we also explore approaches for addressing and imputing this type of structured missing data.

Specifically, we first consider a scenario with random missing data, where we observe an

incomplete version of the matrix X, denoted by $\tilde{X}$, which we assume follows the DGP:

$$\tilde{X}_{it} = \begin{cases} 0, & \text{with probability } 1 - \pi_i, \\ X_{it}, & \text{with probability } \pi_i, \end{cases} \tag{2.9}$$

where X follows the nonlinear factor model described in (2.1), and $\pi_i \in (0, 1)$ represents the probability of observing an entry for the ith variable. This probabilistic framework captures scenarios where the data is missing at random and the degree of missingness is heterogeneous.

Using $\tilde{X}_t$ to train an AE, we derive the following result regarding the output $\widehat{X}_t$:

Corollary 3. *Assume* $\min_{1 \le i \le N} \pi_i$ *is lower bounded by some positive constant* ε, U_{it} *are independent of each other, and* $\tilde{X}_{it}$ *is missing at random, satisfying* (2.9). *Under the conditions specified in Theorem 6, with probability at least* $1 - C\exp(-cT^{K/(2\beta+K)})$, *for sufficiently large* $\min(N, T)$, *we have:*

$$\begin{aligned} &\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it}/\widehat{\pi}_i - X_{it}^{\star}\right)^2 \\ &\lesssim \left(T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1 + T^{-1}\inf_{\rho \in \mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(\tilde{X}_t) - F_t^{\star}\|^2\right)\log^4(T), \end{aligned}$$

where c *and* C *are constants independent of* N *and* T *and* $\widehat{\pi}_i = \max(\varepsilon, T^{-1}\sum_{t=1}^{T} 1_{\{\tilde{X}_{i,t} \neq 0\}})$.

The theorem indicates that consistent recovery of the expected values of the missing entries is achievable under the Frobenius norm. Given that matrix completion is primarily an in-sample exercise, utilizing an over-parameterized encoder can be beneficial for attaining a desirable error rate, as discussed previously.

Next, we consider the scenario of structured missing data, for which we employ SAEs. Suppose we observe $X \in \mathbb{R}^{N \times T}$. For each entry, we use $(X_{it}(0), X_{it}(1))$ to denote its untreated and treated values. We assume that $X_t(0) = (X_t^{(1)}(0), X_t^{(2)}(0))$ follows the nonlinear

factor model described in (2.1), where $X_t^{(1)}(0) \in \mathbb{R}^{N_1}$, $X_t^{(2)}(0) \in \mathbb{R}^{N_2}$, and $N_1 + N_2 = N$. Here $X_t^{(1)}$ represents the control group and $X_t^{(2)}$ represents the treatment group. We assume there exists a non-random $T_0 > 0$ where $X^{(2)}$ received the treatment after T_0. The objective is to assess how the treatment impacts the treated units and periods by imputing the potential outcomes for $X^{(2)}$ from $T_0 + 1$ to T, and comparing them with the actual observed outcomes. This analysis assumes that potential outcomes at each point depend only on the contemporaneous treatment status of each unit and do not rely on past treatments or the treatments of other units. Therefore, the problem can be seen as a block matrix completion problem.

The matrix below illustrates the separation between observed and missing data segments:

$$X = \left(\begin{array}{cccc|cccc} X_{1,1}^{(1)} & \cdots & \cdots & X_{1,T_0}^{(1)} & X_{1,T_0+1}^{(1)} & \cdots & \cdots & X_{1,T}^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{N_1,1}^{(1)} & \cdots & \cdots & X_{N_1,T_0}^{(1)} & X_{N_1,T_0+1}^{(1)} & \cdots & \cdots & X_{N_1,T}^{(1)} \\ \hline X_{1,1}^{(2)} & \cdots & \cdots & X_{1,T_0}^{(2)} & * & \cdots & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{N_2,1}^{(2)} & \cdots & \cdots & X_{N_2,T_0}^{(2)} & * & \cdots & \cdots & * \end{array}\right). \tag{2.10}$$

As discussed by Athey et al. [2017], this missing data pattern encompasses two significant special cases in the causal inference literature. The first is the unconfoundedness literature, such as Imbens and Rubin [2015], which typically focuses on scenarios involving a single treated period. The second is the synthetic control literature, including Abadie et al. [2010] and Abadie [2021], which centers on settings with a single treated unit.

To impute these missing values, we train an SAE using $X_t^{(1)}$ as the input and X_t as the output, with the portion corresponding to $X_t^{(2)}$ serving as the supervised target. This approach leverages the "in-sample" data available up to T_0 to learn a low-dimensional structure that helps predict $X_t^{(2)}$ for $t > T_0$.

The following corollary becomes a straightforward application of Corollary 2.

Corollary 4. *Under the same conditions as in Theorem 7 for $X(0)$, as $\min(N_1, N_2, T_0)$ becomes sufficiently large, and $N_1 \lesssim N_2$, we have:*

$$\begin{aligned}&\frac{1}{N_2(T-T_0)}\sum_{t=T_0+1}^{T}\sum_{i=1}^{N_2}\mathbb{E}\big(\widehat{X}_{i,t}^{(2)}-X_{i,t}^{(2),\star}\big)^2\\ \lesssim&(N_2^{-1}K_1+T_0^{-\frac{2\beta}{2\beta+K}}+L^{-1}+\epsilon+T_0^{-1}L)\log^4(NT_0).\end{aligned}$$

To ensure valid causal inference regarding the average treatment effect

$$\tau=\frac{1}{N_2(T-T_0)}\sum_{t=T_0+1}^{T}\iota^\top(X_t^{(2)}(1)-X_t^{(2)}(0)),$$

we propose the a debiased estimator. We partition the sample before T_0 into two subsets: D_1 and D_2, which serve as the training and calibration sets, respectively. The training set D_1 is used to fit the SAE model, and we generate predictions for the calibration set, denoted as $\widehat{X}_t^{(2)}$ for $t \in D_2$.

The debiased point estimator for average treat effect is given by

$$\begin{aligned}\hat{\tau}:=&\frac{1}{N_2(T-T_0)}\sum_{t=T_0}^{T}\iota^\top(X_t^{(2)}-\widehat{X}_t^{(2)})-\frac{1}{N_2|D_2|}\sum_{t\in D_2}\iota^\top(X_t^{(2)}-\widehat{X}_t^{(2)})\\ =&\frac{1}{N_2(T-T_0)}\sum_{t=T_0+1}^{T}\iota^\top(X_t^{(2)}(1)-\widehat{X}_t^{(2)})-\frac{1}{N_2|D_2|}\sum_{t\in D_2}\iota^\top(X_t^{(2)}-\widehat{X}_t^{(2)})\\ =&\tau+\frac{1}{N_2(T-T_0)}\sum_{t=T_0+1}^{T}\iota^\top(X_t^{(2)}(0)-\widehat{X}_t^{(2)})-\frac{1}{N_2|D_2|}\sum_{t\in D_2}\iota^\top(X_t^{(2)}(0)-\widehat{X}_t^{(2)}).\end{aligned}$$

Under the i.i.d. assumption for $X_t(0)$, we observe that $\hat{\tau}$ is unbiased for τ and it is a sum of

two independent terms. Thus, we can construct 95% confidence intervals for τ as

$$\hat{\tau} \pm 1.96\sqrt{((T-T_0)^{-1} + |D_2|^{-1})\hat{\sigma}^2},$$

where $\hat{\sigma}^2$ is the sample variance of $N_2^{-1}\iota^\top(X_t^{(2)}(0) - \widehat{X}_t^{(2)})$ for $t \in D_2$. This confidence interval is asymptotically valid when $\min(|D_1|, |D_2|, T - T_0) \to \infty$.

2.3.3.3 Panel Regression With Unobserved Heterogeneity

Modeling panel data with unobserved heterogeneity is a central yet challenging problem in econometrics. A general specification takes the form

$$Y_{it} = X_{it}\theta + \alpha_{it} + V_{it},$$

where α_{it} captures unobserved heterogeneity, and V_{it} is an idiosyncratic error term. For notational simplicity, we assume $X_{it} \in \mathbb{R}$ is scalar, although our discussion readily extends to vector-valued regressors.

A common approach assumes additive fixed effects, where $\alpha_{it} = \alpha_i + \beta_t$. Bai [2009] extend this to a factor structure by modeling $\alpha_{it} = \Lambda_i^\top F_t$, where both Λ_i and F_t are unobserved. However, such specifications may fail to capture richer forms of individual heterogeneity, especially in the presence of nonlinear dynamics. To address this, more recent literature considers discrete heterogeneity. For example, Keane and Wolpin [1997] introduce discrete-type random effects, while Bonhomme and Manresa [2015] propose a grouped fixed effects framework with $\alpha_{it} = \alpha_{g_i t}$, where g_i denotes the (latent) group membership. Bonhomme et al. [2022] and Freeman and Weidner [2023] further generalize this structure by allowing for continuous heterogeneity of the form $\alpha_{it} = \alpha^\star(\Lambda_i, F_t)$, effectively providing a nonlinear extension of Bai [2009].

Building on these developments, we consider an even more flexible model that allows for

heterogeneous effects through individual-specific functions of a common latent factor process:

$$Y_{it} = X_{it}\theta + \alpha_i^\star(F_t^\star) + V_{it},$$
$$X_{it} = \varphi_i^\star(F_t^\star) + U_{it},$$

where $F_t^\star \in [-B, B]^K$ is an unobserved common factor, and both $\varphi_i^\star$ and $\alpha_i^\star$ belong to $\mathcal{H}^\beta$. For simplicity, we assume U_{it} and V_{it} are i.i.d. sub-Gaussian random variables conditional on $F^\star$.

In this general nonlinear panel model, classical estimators—such as fixed effects, factor models, or grouped estimators—are generally inconsistent. To address this, we propose a novel double Autoencoders (DAE) approach that yields consistent estimation of the structural parameter θ.

Our estimation strategy proceeds as follows. We apply autoencoders (AEs) separately to X and Y to estimate the nonlinear components $\hat{X}_{it}$ and $\hat{Y}_{it}$, which serve as approximations to $\varphi_i^\star(F_t^\star)$ and $\varphi_i^\star(F_t^\star)\theta + \alpha_i^\star(F_t^\star)$, respectively. We then regress the residualized dependent variable $Y_{it} - \hat{Y}_{it}$ on the residualized regressor $X_{it} - \hat{X}_{it}$ to obtain the estimator $\hat{\theta}$. These residuals correspond to noisy realizations of $U_{it}\theta + V_{it}$ and U_{it}, respectively. The following proposition characterizes the asymptotic expansions of $\hat{\theta}$:

Proposition 3. *Let $\hat{\theta}$ be the estimator obtained via the DAE. Suppose that AEs architecture follows the form of Theorem 6 with over-parametrized encoders. Then,*

$$\hat{\theta} - \theta = O_{\mathrm{P}}(T^{-\frac{\beta}{2\beta+K}} \log^2 T) + O_{\mathrm{P}}(N^{-1/2} K_1^{1/2} \log^2 T)$$

This result establishes the consistency of the double AE estimator. While deriving the asymptotic distribution of $\hat{\theta} - \theta$ would be of significant interest, to the best of our knowledge, such results are not available in the existing literature, even under the simplified model $\alpha_{it} = \alpha^\star(\Lambda_i, F_t)$. We therefore leave the derivation of the asymptotic distribution of $\hat{\theta}$ for

future research.

2.4 Monte Carlo Simulations

In this section, we conduct simulation experiments to validate our theoretical predictions and examine practical considerations in the design and training of AEs and SAEs. This analysis bridges theoretical insights with empirical applications, exploring the strengths and limitations of these methods in finite sample settings.

2.4.1 Simulation Setup

We begin by introducing the DGPs used in our simulations. Specifically, we consider four distinct DGPs by (2.1), comprising one linear factor model and three nonlinear factor models.

The baseline linear model is defined as $\varphi_i^\star(x) = C\Lambda_i^\top x$, serving as a starting point for comparison with more complex models. The first nonlinear model is a generalized linear latent factor model, specified as $\varphi_i^\star(x) = C\exp(\Lambda_i^\top x)$. By introducing exponential transformations to the baseline, this model incorporates nonlinearity while retaining an inherently low-dimensional linear structure. The next model, a generalized nonlinear factor model, , fully abandons the linear framework. It is expressed as $\varphi_i^\star(x) = C\exp(-\|\Lambda_i - x\|^2)$, posing significant challenges for methods that may benefit from linearity. The final model is a polynomial factor model, expressed as $\varphi_i^\star(x) = C_1\Lambda_{1i}^\top x + C_2 x^\top \Lambda_{2i} x$. This model blends linear patterns with higher-order nonlinearities and can produce negative values, making it particularly relevant for calibrating empirical datasets.

Across all Monte Carlo repetitions, we employ five factors by setting $K = 5$, and each element of $F_t^\star \in \mathbb{R}^5$ follows a uniform distribution $\mathbb{U}(-2, 2)$. The elements of $\Lambda_i, \Lambda_{1i}, \Lambda_{2i}$ are i.i.d. from $\mathbb{U}(-1, 1)$, and the noise U_t is normally distributed as $\mathcal{N}(0, \mathbb{I}_N)$. We calibrate C, C_1, and C_2 to ensure that the unconditional variance of $\varphi_i^\star(F_t^\star)$ is normalized to 1 for each model and that the linear and nonlinear components of the polynomial factor model

contribute equally to the total variance.

We benchmark AEs' performance against PCA. While PCA is traditionally associated with linear factor analysis, recent research has demonstrated its effectiveness in approximating certain nonlinear models.[7] We report results using the number of factors ranging from 1 to 19 in increments of two to illustrate its impact on PCA performance.

Turning to the design of AEs, we adhere to the principle of parsimony, experimenting with simple yet non-trivial structures to validate our theoretical results. We first address the bottleneck layer. We vary the number of neurons in this layer, K_1, in the same way as with PCA, ranging from 1 to 19, and compare the results of different architectures. For the encoder, we fix a single hidden layer with 20 neurons to ensure it is wider than the bottleneck. We experiment with five decoder architectures of varying complexity. The first architecture (AE1) serves as the benchmark, consisting of a single hidden layer in the decoder with four neurons. We compare it with two more complex AEs. AE2 includes two hidden layers, with two neurons in the first layer and four neurons in the second. AE3, in contrast, has a single hidden layer with eight neurons. These three AEs feature disjoint outputs, and we illustrate AE3's architecture in Figure 2.5. In addition, we include two AEs with fully connected decoders. AE4 builds on AE3 by employing a fully connected decoder, making it the most complex among all models. The final model, AE5, also uses a fully connected decoder but matches the number of neurons in the decoder to those in the encoder, resulting in a symmetric AE. The total number of parameters in AE5's decoder is comparable to that

7. According to Udell and Townsend [2019], a matrix with a small spectral norm can be effectively approximated by a low-rank matrix. For a matrix $X \in \mathbb{R}^{m\times n}$ where $m \geq n$ and $0 < \epsilon < 1$, the approximation error is bounded as:

$$\inf_{\text{rank}(Y)\leq r} \|X - Y\|_\infty \leq \epsilon \|X\|_2, \text{ with } r = \lceil 72 \log(2n+1)/\epsilon^2 \rceil.$$

Additionally, Agarwal et al. [2021] demonstrate that the matrix $X_{it} := \varphi(\Lambda_i, F_t^\star)$ admits a low-rank approximation, where for any $\delta > 0$,

$$\inf_{\text{rank}(Y)\leq r} \|X - Y\|_\infty \lesssim \delta^\beta, \text{ with } r \asymp \delta^{-K},$$

where K is the dimension of $F_t^\star$, and β is given by Assumption 7.

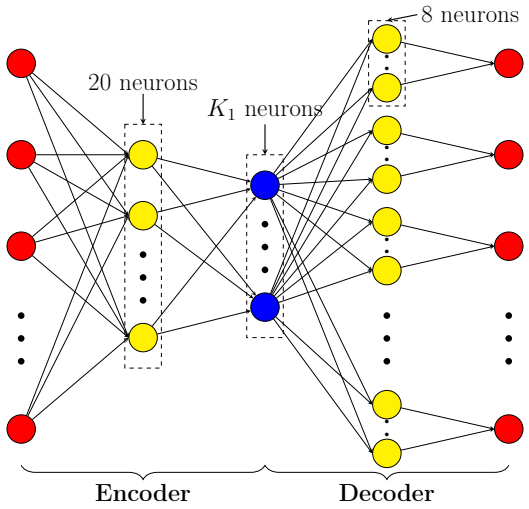


Figure 2.5: Illustration of AE3's Architecture

Note: This graph illustrates the architecture of AE3 used in our numerical analysis. The encoder has one hidden layer with 20 neurons, while the decoder consists of multiple separate neural nets, each containing a hidden layer with 8 neurons. We vary the number of neurons, K_1, in the bottleneck layer.

of AE1, with their relative complexity depending on the specific choices of K_1 and N.

We address the training details separately based on different AE applications, starting with using AEs for extracting common components.

2.4.2 Finite-Sample Recovery of Common Components with Autoencoders

To train the aforementioned AEs in this scenario, we use stochastic gradient descent with the Adam optimizer [Kingma and Ba, 2014] to minimize the loss defined in (2.4). The batch size is fixed at $T/10$, ensuring each epoch consists of ten gradient descent updates. In simulations for AEs, the only parameter we tune is the learning rate in the optimization algorithm, selected from $\{0.005, 0.01, 0.05, 0.1\}$ using a training-validation scheme. To mitigate the effects of random initialization and stochastic batch selection during training, we adopt an ensemble approach, averaging the outputs of 10 independently trained models, consistent with standard practices in the literature.

For each tuning parameter (e.g., learning rate), we train the AEs using the first 4/5 of

the sample (training sample) and apply the inferred encoders and decoders to construct an estimate of $X^\star$ in the remaining sample (validation sample). This approach allows us to calculate the validation loss and determine the optimal tuning. When training, we apply early stopping if the validation loss does not decrease for 50 consecutive epochs. Finally, we refit the model using the entire sample with the selected tuning parameter.

We report the mean squared error (MSE) given by:

$$N^{-1}T^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N}(\widehat{X}_{it}-\varphi_i^\star(F_t^\star))^2,$$

with a variety of choices $(N,T)=(50,500)$, $(200,500)$, and $(200,50)$. The comparative performance, as illustrated in Figure 2.6, highlights notable differences in how PCA and AEs respond to variations in sample size and the number of factors.

In the baseline linear case, PCA achieves optimal MSE when the number of factors is set to five (the true value) and is clearly the best performer when the sample size N is larger than T. In this linear setting, AE1 through AE3 exhibit similar performance. These AEs tend to achieve lower MSEs than PCA when the number of factors is below five; however, their performance, like that of PCA, begins to deteriorate as the number of factors increases beyond five. This decline is due to the addition of extra factors, which introduces noise into the estimation rather than improving accuracy. AE4 follows a U-shaped pattern similar to the other AEs but demonstrates significantly worse performance. Its greater complexity makes it more prone to overfitting, resulting in higher MSEs compared to its disjoint output benchmark model, AE3. In contrast, AE5 limits complexity by reducing the number of neurons in the decoder layer, which restricts the number of weight parameters despite being fully connected. This design enables AE5 to achieve performance comparable to the other AEs while notably outperforming PCA when an excessive number of factors is allowed, effectively mitigating the impact of overfitting.

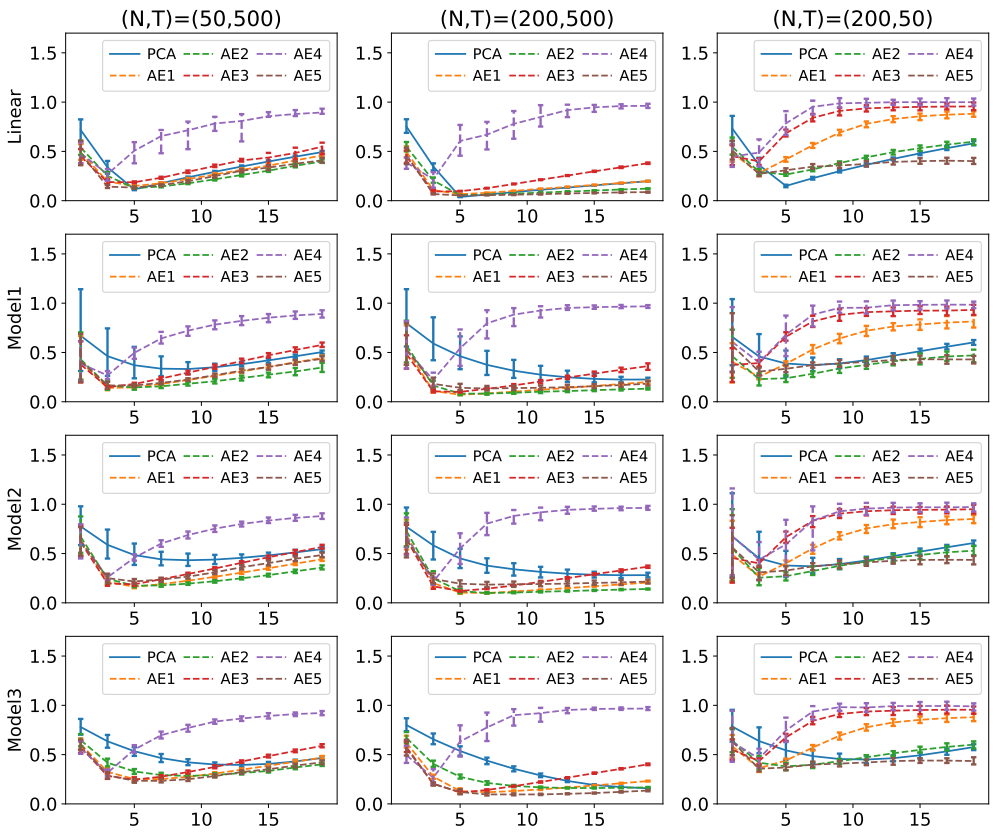


Figure 2.6: Mean Squared Error Comparison in Extracting Common Components

Note: This figure shows boxplots of the mean squared errors in extracting common components, $X^{\star}$, for each specified number of factors shown on the x-axis. It compares the performance of five different AEs and PCA applied to four DGPs, evaluated across various choices of N and T, based on 100 Monte Carlo repetitions.

The results for all nonlinear DGPs reveal a similar pattern. As the number of factors increases, PCA's performance continues to improve, suggesting that nonlinearity compels PCA to extract more linear factors to better approximate the model. In all cases, PCA struggles to match the performance of the nonlinear methods, even with up to 20 factors in the case where $N = 200$ and $T = 500$, while these nonlinear methods deliver superior results with just five factors. When N is small relative to T, AE1, AE2, AE3, and AE5 deliver comparable performance and significantly outperform AE4, which begins to overfit as its complexity increases. Conversely, when N is large relative to T, AE1 and AE5 tend to dominate, maintaining superior performance. Meanwhile, AE2 and AE3 become more comparable to AE4, indicating that their relative advantage diminishes as the ratio between sample size and dimensionality shifts.

2.4.3 Simulation Comparison with Kernel PCA and Local PCA

We conduct simulations to compare AEs with Kernel PCA (KPCA) and Local PCA (LPCA) for Model3. For KPCA, we apply three different kernels: polynomial, radial basis function (RBF), and sigmoid. We utilize the same procedure to select their hyper-parameters. For LPCA, we follow the same setup as Feng [2023], taking the pseudo-max distance for determining nearest neighbors and set the number of nearest neighbors $cn^{2/3}$ for c = 0.5, 1 and 1.5. Due to the similarity in results across AE1 to AE3, only the performance of AE1 is shown. The findings indicate that KPCA's performance is highly sensitive to the choice of kernel. Among the tested kernels, the polynomial kernel outperforms PCA but remains inferior to AE1. In contrast, the RBF kernel show poor performance, with MSE values even higher than those of PCA.

These results highlight that while KPCA introduces nonlinearity through kernel functions, it lacks the adaptive ability of AEs to learn optimal representations directly from the data. This dependency on choosing the right kernel can significantly impact performance.

In contrast, AEs, with their deep, flexible architectures, adapt effectively to complex data structures and consistently outperform KPCA. This highlights the robustness and versatility of AEs in capturing nonlinear relationships. Regarding LPCAs, they are capable of adapting to nonlinearities, resulting in a lower MSE with fewer factors compared to PCA. However, their minimum MSE remains higher than that of AEs, and they are more prone to overfitting when an over-specified number of factors is used, especially in comparison to AEs.

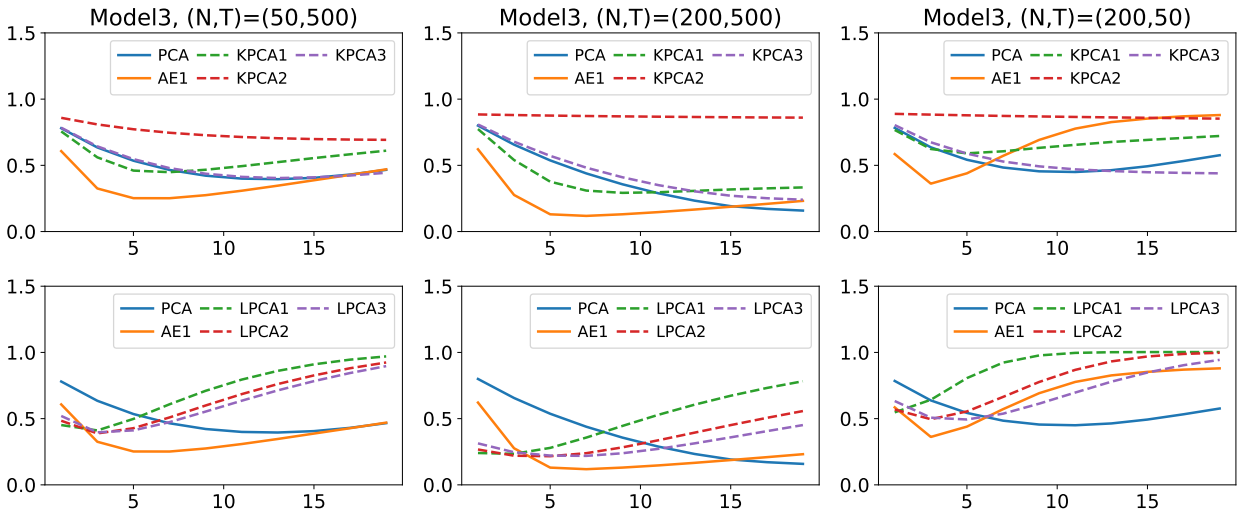


Figure 2.7: Mean Squared Error Comparison in Simulations with Kernel PCA and Local PCA

Note: This figure shows boxplots of the mean squared errors for each specified number of factors, comparing three different KPCA and LPCA methods applied to Model3 across various choices of N and T, based on 100 Monte Carlo repetitions. Here KPCA1 represents polynomial kernel, KPCA2 represents RDF kernel, and KPCA3 represents sigmoid kernel. LPCA1 to LPCA3 correspond to LPCA with the number of nearest neighbors $cn^{2/3}$ for c = 0.5, 1 and 1.5.

2.4.4 Finite-Sample Predictive Performance with Supervised Autoencoders

Next, we evaluate the performance of SAEs, focusing on a prediction setting. We adopt a polynomial factor model as the DGP for X_{it}. Additionally, we simulate target variables Y_{it} based on (2.5), where $\phi_i^\star(x)$ also follows a polynomial factor model but with a distinct set of randomly generated parameters (e.g., Λ_{1i} and Λ_{2i}). We adjust C_1 and C_2 in the DGP to

calibrate the signal-to-noise ratio (SNR) for X and Y based on our empirical application on return predictions. Specifically, the SNR of X is set to 1, whereas for Y, the SNR is set to 1%, aligning with the low-SNR in practice. We fix $N = 50$, while varying the dimension of Y_{it}, M. The in-sample size is set at $T = 200$, matching that of our empirical study.

The decoders of our SAEs connecting to Y retain the same structure as the AEs connecting to X, with SAE1 through SAE3 employing disjoint output decoders and SAE4 and SAE5 using fully connected decoders. As discussed earlier, a sparse encoder can enhance model's out-of-sample performance. Sparsity is introduced here via a pruning procedure. After training SAEs with a fully connected encoder, weights below a specified pruning threshold are set to zero. The threshold is determined by sorting all weights and truncating those below a specified percentile, i.e., the pruning ratio. We vary the pruning ratio from 0% to 95% in increments of 5%. Since pruning is applied after training, adjusting the pruning ratio is computationally efficient.[8] The pruning ratio and learning rate are selected jointly during the validation step.

After training and validation, we evaluate the selected model's out-of-sample performance using a separate testing sample of 100 observations. For each target variable, we calculate its out-of-sample R^2 against its in-sample mean, and then report the average over all such R^2s in Table 2.1. For comparison, we include principal component regression (PCR) as a linear benchmark. In PCR, each variable Y_i is regressed on the principal components of X_{it} using the entire in-sample dataset. Predictions are then made out-of-sample using the in-sample estimates of principal component weights and regression coefficients.

The results highlight the advantages of our SAE approach compared to PCR, which consistently yields negative out-of-sample R^2 values regardless of the number of factors. This underperformance can be attributed to PCR's lack of supervision when recovering factors and its reliance on linear approximations, which is inadequate for capturing the nonlinear

8. For alternative approaches to pruning a NN, see LeCun et al. [1989], Hassibi and Stork [1992], and Frankle and Carbin [2018].

$M = 25$						$M = 50$					
K_1	1	3	5	7	9	K_1	1	3	5	7	9
PCR	-0.323	-0.922	-1.731	-2.45	-3.38	PCR	-0.297	-1.049	-1.869	-2.692	-3.585
SAE1	0.245	0.258	0.257	0.261	0.271	SAE1	0.302	0.328	0.341	0.342	0.354
SAE2	0.252	0.246	0.258	0.247	0.267	SAE2	0.299	0.318	0.314	0.317	0.317
SAE3	0.176	0.232	0.26	0.246	0.249	SAE3	0.264	0.316	0.337	0.347	0.345
SAE4	-0.146	-0.039	-0.038	0.021	0.025	SAE4	-0.087	-0.048	-0.032	0.009	0.025
SAE5	-0.023	0.061	0.082	0.093	0.135	SAE5	-0.038	0.204	0.207	0.191	0.204
$M = 100$						$M = 200$					
PCR	-0.327	-1.087	-1.924	-2.762	-3.674	PCR	-0.31	-1.098	-1.885	-2.739	-3.561
SAE1	0.343	0.375	0.382	0.386	0.382	SAE1	0.374	0.39	0.389	0.392	0.393
SAE2	0.364	0.368	0.371	0.373	0.371	SAE2	0.386	0.389	0.39	0.388	0.39
SAE3	0.315	0.368	0.378	0.379	0.381	SAE3	0.339	0.384	0.388	0.392	0.392
SAE4	-0.08	-0.052	-0.021	-0.031	0.005	SAE4	-0.091	-0.051	-0.042	-0.038	-0.012
SAE5	0.008	0.237	0.272	0.266	0.258	SAE5	0.058	0.242	0.261	0.283	0.279

Table 2.1: Simulation Results for Supervised Autoencoders

Note: This table reports the out-of-sample R^2 (in percentages) of predicted values across various predictive methods, including the benchmark Principal Component Regression (PCR) method and multiple SAE architectures of increasing complexity. K_1 corresponds to the number of factors in PCA and the number of neurons in the bottleneck layers of the SAEs. The dimension of Y, M, is varied while the dimension of X is fixed at $N = 50$. The sample size is fixed at $T = 200$.

relationships needed to predict the target variables. SAE4 also struggles to achieve positive R^2 values due to its propensity for overfitting in out-of-sample scenarios. In contrast, SAE5 mitigates the curse of complexity, yielding positive R^2 values. Notably, SAE1 through SAE3 outperform SAE5 by delivering higher R^2 values, demonstrating their ability to extract nonlinear factors effectively while maintaining sufficient complexity control to avoid overfitting. As M (the dimension of Y_{it}) increases, prediction target becomes increasingly important in training, leading to improved predictive performance. While the number of neurons in the decoder layer grows as M increases, the total number of weights per target remains fixed for SAE1, SAE2, SAE3, and SAE5. However, for SAE4, the total number of weights per target scales linearly with M, contributing to its progressively worse performance.

Our stylized simulation experiments on AEs and SAEs have thus far validated our theoretical predictions. While advancing architectures, optimization algorithms, and parameter-

tuning schemes are important and active areas in machine learning research, our primary objective is not to identify the optimal model. Instead, we aim to validate our theoretical insights and identify a useful model for empirical analysis.

2.5 Conclusion

This paper develops theoretical insights into a nonparametric unsupervised learning problem—the application of deep AEs within nonlinear factor models—demonstrating their effectiveness in extracting latent common components from high-dimensional inputs. By extending this framework to include SAEs, we pave the way for broader and more versatile applications in economics. These methods offer a systematic approach to nonlinear dimensionality reduction and enhance predictive accuracy by leveraging underlying low-dimensional structures.

While this work primarily focuses on estimation and prediction using a specific class of AEs and SAEs, it opens several promising directions for future research. A natural extension involves formal inference, particularly in the context of causal analysis, to better quantify uncertainty in causal effects. Expanding the framework to include more general AE structures beyond the disjoint output class could enhance their flexibility and applicability across diverse settings. Furthermore, developing principled methodologies for model selection, architecture optimization, and regularization strategies would improve both interpretability and computational efficiency. Finally, pursuing theoretical analyses to understand the integration of AEs with other machine learning frameworks, such as generative models or variational inference, could unlock opportunities for novel and distinct economic applications.

2.6 Mathematical Proofs

2.6.1 Proof of Theorem 6

Proof. For simplicity, denote $\mathcal{F}_N^{K_1}(d_1, w_1, T^{5\beta+5}, B)$ and $\mathcal{F}_{K_1}^1(d_2, w_2, T^{5\beta+5}, B)$ by $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. We further define a function space of DNNs which are potentially unbounded: $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C) := \left\{ f \text{ of the form (2.2)} : \max_{j=0,\ldots,d} \left\|W_j\right\|_\infty \vee \left|v_j\right|_\infty \leq C \right\}$.

Throughout the proof, c and C denote fixed constants, which may vary from line to line.

2.6.1.1 Main Inequality

By the definition of $(\widehat{\varphi}_1, \ldots, \widehat{\varphi}_N) \circ \widehat{\rho}$, for any $(\varphi_1, \ldots, \varphi_N) \circ \rho \in \mathcal{F}_{AE}^{K_1}$, it holds that

$$\sum_{t=1}^{T}\sum_{i=1}^{N} \left(X_{it} - \widehat{\varphi}_i(\widehat{\rho}(X_t))\right)^2 \leq \sum_{t=1}^{T}\sum_{i=1}^{N} \left(X_{it} - \varphi_i(\rho(X_t))\right)^2.$$

As a consequence, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{N} \left(X_{it}^\star - \widehat{\varphi}_i(\widehat{\rho}(X_t))\right)^2 = \sum_{t=1}^{T}\sum_{i=1}^{N} \Big[\left(X_{it} - \widehat{\varphi}_i(\widehat{\rho}(X_t))\right)^2 - U_{it}^2 + 2\widehat{X}_{it}U_{it} - 2X_{it}^\star U_{it}\Big] \\
\leq &\sum_{t=1}^{T}\sum_{i=1}^{N} \Big[\left(X_{it} - \varphi_i(\rho(X_t))\right)^2 - U_{it}^2 + 2\widehat{X}_{it}U_{it} - 2X_{it}^\star U_{it}\Big] \qquad (2.11) \\
= &\sum_{t=1}^{T}\sum_{i=1}^{N} \Big[\left(X_{it}^\star - \varphi_i(\rho(X_t))\right)^2 + 2U_{it}\Big(\left(X_{it}^\star - \varphi_i(\rho(X_t))\right) + \left(\widehat{X}_{it} - X_{it}^\star\right)\Big)\Big].
\end{aligned}$$

Below we establish three inequalities which are denoted by (I) to (III).

(I) Assume $(\varphi_1, \ldots, \varphi_N) \circ \rho \in \mathcal{F}_{AE}^{K_1}$. As long as $\varphi_1, \ldots, \varphi_N$ are deterministic, with

probability at least $1 - C\exp(-cT)$,

$$\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^{\star} - \varphi_i(\rho(X_t))\big) \lesssim T^{1/2}K_1^{1/2}\log(T)\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^{\star} - \varphi_i(\rho(X_t))\big)^2\Big)^{1/2} + K_1^{1/2}T^{1/2}N^{1/2}\log(T) + N.$$

(II): With probability at least $1 - C\exp(-cT)$, we have

$$\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X_{it}^{\star}\big)U_{it}\right| \lesssim (NT^{\frac{K}{2\beta+K}} + TK_1)^{1/2}\log^2(T) \times\left(\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X_{it}^{\star}\big)^2\right)^{1/2} + N^{1/2}\right) + N.$$

(III): There exists $(\varphi_1^{\dagger},\ldots,\varphi_N^{\dagger})\circ\rho^{\dagger} \in \mathcal{F}_{AE}^{K_1}$ such that $\varphi_1^{\dagger},\ldots,\varphi_N^{\dagger}$ are deterministic and

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^{\star} - \varphi_i^{\dagger}(\rho^{\dagger}(X_t))\big)^2 \lesssim T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F_t^{\star}\|^2.$$

2.6.1.2 Proof of (I)

Write $\varepsilon_T = K_1^{-1}T^{-1-(5\beta+5)(d_2+1)}w_2^{-d_2}$ and define the set

$$\mathcal{S} = \{x \in \mathbb{R}^{K_1} : \forall 1 \le i \le K_1,\ \exists 1 \le j(\in\mathbb{N}) \le 2B\varepsilon_T^{-1},\ \text{s.t. } x_i = -B + j\varepsilon_T\}. \tag{2.12}$$

By definition, we see that for any $F \in [-B,B]^{K_1}$, there exists a $\bar{F} \in \mathcal{S}$ such that $\|F - \bar{F}\|_{\infty} \le \varepsilon_T$. Write the set as $\{\bar{F}_k, 1 \le k \le |\mathcal{S}|\}$. It is not hard to verify that

$$\log|\mathcal{S}| \le K_1\log(2BK_1T^{1+(5\beta+5)(d_2+1)}w_2^{d_2}) \lesssim K_1\log^2(T). \tag{2.13}$$

For $1 \le t \le T$, assume $\|\rho(X_t) - \bar{F}_{k_t^\star}\|_\infty \le \varepsilon_T$. As a consequence, by Lemma 31,

$$|\varphi_i(\rho(X_t)) - \varphi_i(\bar{F}_{k_t^\star})| \le T^{-1}. \tag{2.14}$$

Together with Lemma 28, with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\rho(X_t))\big) \lesssim \sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t^\star})\big) + N \\
&\le \max_{\substack{k_t\in[|\mathcal{S}|]\\ 1\le t\le T}} \frac{\Big|\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} \Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t^\star})\big)^2\Big)^{1/2} + N \qquad (2.15)\\
&\lesssim \max_{\substack{k_t\in[|\mathcal{S}|]\\ 1\le t\le T}} \frac{\Big|\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} \left(\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\rho(X_t))\big)^2\Big)^{1/2} + N^{1/2}\right) \\
&\quad + N,
\end{aligned}$$

where we use Eq. (2.14) and Lemma 28 in the first inequality and Eq. (2.14) in the last inequality. For any $u \ge 0$, by Lemma 27 and the property of subGaussian distribution, as well as the fact that $\varphi_1, \ldots, \varphi_N$, $\bar{F}_{k_1}, \ldots, \bar{F}_{k_T}$, and $\{X_{it}^\star\}_{i,t}$ are deterministic given $\{F_t^\star\}_{t=1}^T$,

$$\mathrm{P}\left(\frac{\Big|\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} > u \,\middle|\, \{F_t^\star\}_{t=1}^T\right) \le 2\exp\left(-\frac{u^2}{2\sigma_u^2}\right).$$

Therefore, by union bound inequality and the arbitrary of $\{F_t^\star\}_{t=1}^T$ in the above inequality,

$$\mathrm{P}\left(\max_{\substack{k_t\in[|\mathcal{S}|]\\ 1\le t\le T}} \frac{\Big|\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)\Big|}{\Big(\sum_{t=1}^{T}\sum_{i=1}^{N}\big(X_{it}^\star - \varphi_i(\bar{F}_{k_t})\big)^2\Big)^{1/2}} > u\right) \le 2|\mathcal{S}|^T \exp\left(-\frac{u^2}{2\sigma_u^2}\right).$$

By Eq. (2.13), Eq. (2.15), and setting $u = 2\sigma_u(T\log|\mathcal{S}|)^{1/2}$, we complete the proof for (I).

2.6.1.3 Proof of (II)

To prove this assertion, we rely on Lemma 32. The lemma states that for $\delta = T^{-1}$, there exists a function class $\mathcal{F}_{2,\delta} \subset \mathcal{F}_2$ with

$$\log|\mathcal{F}_{2,\delta}| \lesssim (w_2^2 d_2 + K_1 w_2)\log\left(8\delta^{-1}BT^{(5\beta+5)(d_2+1)}(1+w_2)^{d_2}K_1 d_2\right) \lesssim T^{\frac{K}{2\beta+K}}\log^4(T) \tag{2.16}$$

such that for any $\varphi \in \mathcal{F}_2$, there exists $\overline{\varphi} \in \mathcal{F}_{2,\delta}$ meeting the requirement $|\overline{\varphi}(x)-\varphi(x)| \leq T^{-1}$, $\forall \|x\|_\infty \leq B$. Since $\|\varphi\|_\infty \leq B$, we have $|\overline{\varphi}(x)| \leq T^{-1}+B$. Hence we can assume all the functions in $\mathcal{F}_{2,\delta}$ are bounded by $2B$ for simplicity, i.e., $\mathcal{F}_{2,\delta} \in \mathcal{F}_{K_1}^1(d_2, w_2, T^{5\beta+5}, 2B)$. Write $\mathcal{F}_{2,\delta} = \{\overline{\varphi}_j, 1 \leq j \leq |\mathcal{F}_{2,\delta}|\}$. By definition, there exists $\ell_i^\star \in [|\mathcal{F}_{2,\delta}|]$ such that $|\widehat{\varphi}_i(x) - \overline{\varphi}_{\ell_i^\star}(x)| \leq T^{-1}$, $\forall \|x\|_\infty \leq B$. Additionally, recall the definition of $\mathcal{S}$ in Eq. (2.12), there exists $k_t^\star \in [|\mathcal{S}|]$ such that $\|\overline{F}_{k_t^\star} - \widehat{\rho}(X_t)\|_\infty \leq \varepsilon_T$. Since $\widehat{\varphi}_i \in \mathcal{F}_{K_1}^1(d_2, w_2, T^{5\beta+5}, B)$, by Lemma 31, we have

$$|\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \overline{\varphi}_{\ell_i^\star}(\overline{F}_{k_t^\star})| \leq |\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \widehat{\varphi}_i(\overline{F}_{k_t^\star})| + |\widehat{\varphi}_i(\overline{F}_{k_t^\star}) - \overline{\varphi}_{\ell_i^\star}(\overline{F}_{k_t^\star})| \leq 2T^{-1}.$$

Together with Lemma 27, with probability at least $1 - C\exp(-cT)$ we have

$$\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X_{it}^{\star}\right)U_{it}\right| \lesssim \left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i^{\star}}(\overline{F}_{k_t^{\star}}) - X_{it}^{\star}\right)U_{it}\right| + N \tag{2.17}$$

$$\leq \max_{\substack{k_t\in[|\mathcal{S}|],1\leq t\leq T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],1\leq i\leq N}} \frac{\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^{\star}\right)U_{it}\right|}{\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^{\star}\right)^2\right)^{1/2}} \left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i^{\star}}(\overline{F}_{k_t^{\star}}) - X_{it}^{\star}\right)^2\right)^{1/2} + N$$

$$\lesssim \max_{\substack{k_t\in[|\mathcal{S}|],1\leq t\leq T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],1\leq i\leq N}} \frac{\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^{\star}\right)U_{it}\right|}{\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^{\star}\right)^2\right)^{1/2}}$$
$$\times \left(\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - X_{it}^{\star}\right)^2\right)^{1/2} + N^{1/2}\right) + N.$$

By Lemma 27, the property of subGaussian distribution, as well as the union bound inequality and the arbitrary of $\{F_t^{\star}\}_{t=1}^{T}$, we obtain

$$\mathrm{P}\left(\max_{\substack{k_t\in[|\mathcal{S}|],1\leq t\leq T\\ \ell_i\in[|\mathcal{F}_{2,\delta}|],1\leq i\leq N}} \frac{\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^{\star}\right)U_{it}\right|}{\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\overline{\varphi}_{\ell_i}(\overline{F}_{k_t}) - X_{it}^{\star}\right)^2\right)^{1/2}} > u\right) \leq 2|\mathcal{S}|^T \cdot |\mathcal{F}_{2,\delta}|^N \exp\left(-\frac{u^2}{2\sigma_u^2}\right).$$

The conclusion follows by Eqs. (2.13), (2.16) and (2.17) with $u = 2\sigma_u(T\log|\mathcal{S}| + N\log|\mathcal{F}_{2,\delta}|)^{1/2}$.

2.6.1.4 Proof of (III)

It is sufficient to construct $(\varphi_1^{\dagger},\ldots,\varphi_N^{\dagger})\circ\rho^{\dagger} \in \mathcal{F}_{AE}^{K}$ satisfying the inequality. If $K_1 > K$, we can introduce useless neurons into the bottleneck with zero weights and biases, ensuring that the output of the network remains unchanged.

Assume $\rho^{\dagger}(\cdot) := \arg\min_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F_t^{\star}\|^2$. According to Proposition 4, there

exist deterministic $\varphi_i^\dagger \in \mathcal{F}_1^K(d_2, w_2, T^{5\beta+5}, B)$ such that:

$$|\varphi_i^\dagger(x) - \varphi_i^\star(x)| \lesssim T^{-\frac{\beta}{2\beta+K}}, \quad \forall x \in [-B, B]^K \text{ and } i = 1, \ldots, N. \tag{2.18}$$

By the triangle inequality, the following inequality holds:

$$\begin{aligned}\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\varphi_i^\star(F_t^\star) - \varphi_i^\dagger(\rho^\dagger(X_t))\right)^2 \leq \frac{2}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\varphi_i^\star(F_t^\star) - \varphi_i^\star(\rho^\dagger(X_t))\right)^2 \\ + \frac{2}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\varphi_i^\star(\rho^\dagger(X_t))) - \varphi_i^\dagger(\rho^\dagger(X_t)))\right)^2 =: A_1 + A_2.\end{aligned}$$

By Assumptions 7, we confirm:

$$|\varphi_i^\star(x) - \varphi_i^\star(y)| \lesssim \|x - y\|_2 \quad \forall i = 1, \ldots, N. \tag{2.19}$$

Therefore, $A_1 \lesssim \frac{1}{T}\sum_{t=1}^{T}\|F_t^\star - \rho^\dagger(X_t)\|^2 = \inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F_t^\star\|^2$. For A_2, by Eq. (2.18), we have $A_2 \lesssim T^{-\frac{2\beta}{2\beta+K}}$, which completes the proof.

2.6.1.5 Final Step

Note that by (I) and (III), it can be proven that, with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\left|\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\varphi_i^\dagger(\rho^\dagger(X_t)) - X_{it}^\star\right)U_{it}\right| \\
\lesssim& T^{1/2}K_1^{1/2}\log T\left(\left(\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\varphi_i^\dagger(\rho^\dagger(X_t)) - X_{it}^\star\right)^2\right)^{1/2} + N^{1/2}\right) + N \\
\lesssim& T^{1/2}N^{1/2}K_1^{1/2}\log T\left(T^{1/2}\left(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F_t^\star\|^2\right)^{1/2} + 1\right) + N \\
\lesssim& NT\left(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F_t^\star\|^2\right)\log^2(T).
\end{aligned}$$

Together with (II), (III) and Eq. (2.11) with $(\varphi_1, \ldots, \varphi_N) \circ \rho$ on the right hand side replaced by $(\varphi_1^\dagger, \ldots, \varphi_N^\dagger) \circ \rho^\dagger$, we conclude that with probability at least $1 - C\exp(-cT)$, $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\varphi_i^\star(F_t^\star) - \widehat{X}_{it}\right)^2$ is approximately less than

$$\begin{aligned}
&(T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1)^{1/2}\log^2(T)\left(\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it} - \varphi_i^\star(F_t^\star)\right)^2\right)^{1/2} \\
&+ \left(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F_t^\star\|^2\right)\log^2(T).
\end{aligned}$$

Note that for $a, b > 0$, the following derivation holds: $x \le a\sqrt{x} + b \Rightarrow x \le a^2 + 4b$. Therefore, the proof is completed by taking $x = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\varphi_i^\star(F_t^\star) - \widehat{X}_{it}\right)^2$, $a = (T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1)^{1/2}\log^2(T)$, and

$$b = \left(N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(X_t) - F_t^\star\|^2\right)\log^2(T).$$

□

2.6.2 Proof of Theorem 7

Proof. Regarding the in-sample performance, it is sufficient to prove that with probability at least $1 - C\exp(-cL) - C\exp(-cT)$, there exists a $\rho^\dagger \in \mathcal{F}_N^K$ such that $T^{-1}\sum_{t=1}^T \|\rho^\dagger(X_t) - F_t^\star\|^2 \lesssim L^{-1} + \epsilon + T^{-\frac{2\beta}{2\beta+K}}$. Notably, since $\|\varphi^\star\|_\infty \leq B$ and $\|W^\star\|_\infty \lesssim L^{-1}$, $\|W^\star\|_0 \asymp L$ by Assumption **??**, we observe that the domain, $W^\star\varphi^\star([-B,B]^K)$, a compact set, is contained within a bounded set $[-\tilde{C},\tilde{C}]^K$. Therefore, by the Whitney extension theorem, $\rho^\star$ can be extended to $\mathcal{H}^\beta([-\tilde{C}-1,\tilde{C}+1]^K, C)$ for some fixed constant C.

According to Proposition 4, there exist $\tilde{\rho}^\dagger \in \mathcal{F}_K^K(d_1-1, w_1, T^{5\beta+5}, B)$ such that:

$$\|\tilde{\rho}^\dagger(x+1+\tilde{C}) - \rho^\star(x)\|_\infty \lesssim T^{-\frac{\beta}{2\beta+K}}, \quad \forall x \in [-\tilde{C}-1, \tilde{C}+1]^K. \tag{2.20}$$

Let $v^\dagger \in \mathbb{R}^K$ be a vector with all elements equal to $-1-\tilde{C}$ and define $\rho^\dagger(x) := \tilde{\rho}^\dagger(\sigma_{v^\dagger}W^\star x) \in \mathcal{F}_K^K(d_1, w_1, T^{5\beta+5}, B)$. By Lemma 29, with probability at least $1-\exp(-cL)$, $\|W^\star X_t\|_\infty \leq 1+\tilde{C}$. As a consequence, by Lemma 30, Eq. (2.20) and the fact that $\rho^\star$ is Lipschitz continuous, with probability at least $1 - C\exp(-cT) - C\exp(-cL)$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \|F_t^\star - \rho^\dagger(X_t)\|^2 \lesssim \frac{1}{T}\sum_{t=1}^T \big(&\|F_t^\star - \rho^\star(W^\star X_t^\star)\|^2 + \|\rho^\star(W^\star X_t^\star) - \rho^\star(W^\star X_t)\|^2 \\
&+ \|\rho^\star(W^\star X_t) - \tilde{\rho}^\dagger(\sigma_{v^\dagger}W^\star X_t)\|^2\big) \\
\lesssim \epsilon + \frac{1}{T}\sum_{t=1}^T \Big(&\|W^\star U_t\|^2 + \|\rho^\star(W^\star X_t) - \tilde{\rho}^\dagger(W^\star X_t + 1 + \tilde{C})\|^2\Big) \lesssim L^{-1} + \epsilon + T^{-\frac{2\beta}{2\beta+K}},
\end{aligned}$$

which completes the proof for the in-sample performance.

Regarding the out-of-sample performance, by the in-sample bound and the facts that $\frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N \big(\varphi_i^\star(F_t^\star) - \widehat{\varphi}_i(\widehat{\rho}(X_t))\big)^2 \leq 4B^2$, $\exp(-cT) = o(T^{-1})$, $\exp(-cL) = o(N^{-1})$

as is assumed in Theorem 7, we have

$$\begin{aligned}
&\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\big(\varphi_i^\star(F_t^\star)-\widehat{\varphi}_i(\widehat{\rho}(X_t))\big)^2\\
\lesssim&(N^{-1}K_1+T^{-\frac{2\beta}{2\beta+K}}+L^{-1}+\epsilon)\log^4(T)+4B^2\cdot(\exp(-cT)+\exp(-cL))\\
\lesssim&(N^{-1}K_1+T^{-\frac{2\beta}{2\beta+K}}+L^{-1}+\epsilon)\log^4(T)+4B^2\cdot o(T^{-1}+N^{-1})\\
\lesssim&(N^{-1}K_1+T^{-\frac{2\beta}{2\beta+K}}+L^{-1}+\epsilon)\log^4(T). \qquad (2.21)
\end{aligned}$$

Assuming that $\{(\tilde{F}_t^\star,\tilde{U}_t,\tilde{X}_t)\}_{t=1}^T$ is an independent copy of $\{(F_t^\star,U_t,X_t)\}_{t=1}^T$, we have

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_{T+1}))-\varphi_i^\star(F_{T+1}^\star)\big)^2-\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t))-\varphi_i^\star(F_t^\star)\big)^2\right|\\
=&\left|\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(\tilde{X}_t))-\varphi_i^\star(\tilde{F}_t^\star)\big)^2-\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t))-\varphi_i^\star(F_t^\star)\big)^2\right|\\
\le&\frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(\tilde{X}_t))-\varphi_i^\star(\tilde{F}_t^\star)\big)^2-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t))-\varphi_i^\star(F_t^\star)\big)^2\right|. \qquad (2.22)
\end{aligned}$$

Lemma 33 provides an upper bound for the right-hand-side of the above inequality, which implies that

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{X}_{T+1}-\varphi_i^\star(F_{T+1}^\star)\big)^2-\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t))-\varphi_i^\star(F_t^\star)\big)^2\right|\\
\le&C(T^{-\frac{\beta}{2\beta+K}}+T^{-1/2}L^{1/2})\log^2(NT)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{X}_{T+1,i}-\varphi_i^\star(F_{T+1}^\star)\big)^2\right)^{1/2}\\
&+C(T^{-\frac{2\beta}{2\beta+K}}+T^{-1}L)\log^4(NT) \qquad (2.23)
\end{aligned}$$

for some fixed constant C. Let $a,b,c,d>0$ satisfy $|a-b|\le 2\sqrt{a}c+d$. Simple algebra gives

$a \le 2(b+d) + 4c^2$. The proof completes by applying this fact to Eq. (2.23) with

$$a = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\big(\widehat{X}_{T+1} - \varphi_i^\star(F_{T+1}^\star)\big)^2, \quad b = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{E}\big(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\big)^2,$$
$$c = \frac{C}{2}(T^{-\frac{\beta}{2\beta+K}} + T^{-1/2}L^{1/2})\log^2(NT), \quad d = C(T^{-\frac{2\beta}{2\beta+K}} + T^{-1}L)\log^4(NT),$$

and the fact that $b \lesssim (N^{-1}K_1 + T^{-\frac{2\beta}{2\beta+K}} + L^{-1} + \epsilon)\log^4(T)$ by Eq. (2.21). □

2.6.3 Proof of Theorem 8

Proof. By Assumption **??**, Eq. (2.19), $\|W^\star\|^2 \le \mathrm{Tr}((W^\star)^\top W^\star) \lesssim L^{-1}$, and Theorem 7, we have with probability at least $1 - C\exp(-cT) - C\exp(-cL)$, $\frac{1}{T}\sum_{t=1}^{T}\|\rho^\star(W^\star\widehat{X}_t) - F_t^\star\|^2$ is approximately less than

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\|\rho^\star(W^\star\widehat{X}_t) - \rho^\star(W^\star\varphi^\star(F_t^\star))\|^2 + \frac{1}{T}\sum_{t=1}^{T}\|\rho^\star(W^\star\varphi^\star(F_t^\star)) - F_t^\star\|^2 \\
\lesssim &\frac{1}{T}\sum_{t=1}^{T}\|W^\star\widehat{X}_t - W^\star\varphi^\star(F_t^\star)\|^2 + \epsilon \lesssim \frac{1}{T}\sum_{t=1}^{T}\|W^\star\|^2\|\widehat{X}_t - \varphi^\star(F_t^\star)\|^2 + \epsilon \\
\lesssim &\frac{1}{TL}\sum_{t=1}^{T}\|\widehat{X}_t - \varphi^\star(F_t^\star)\|^2 + \epsilon \lesssim (L^{-1}K_1 + NL^{-1}T^{-\frac{2\beta}{2\beta+K}} + NL^{-2} + NL^{-1}\epsilon)\log^4(T).
\end{aligned}$$

Together with the fact that $\widehat{X}_t$ is a function of $\widehat{F}_t$, i.e., $\widehat{X}_t = \widehat{\varphi}(\widehat{F}_t)$, we identify a function of $\widehat{F}_t$, given by $\rho^\star \circ W \circ \widehat{\varphi}$, that satisfies the specified condition. □

2.6.4 Proof of Corollary 3

Define

$$\tilde{\varphi}_i^\star(x) = \pi_i\varphi_i^\star(x), \quad \tilde{U}_{it} = \begin{cases} -\pi_i\varphi_i^\star(F_t^\star), & \text{with probability } 1-\pi_i, \\ (1-\pi_i)\varphi_i^\star(F_t^\star) + U_{it}, & \text{with probability } \pi_i. \end{cases}$$

By doing so, we observe that $\tilde{X}_{it} = \tilde{\varphi}_i^\star(F_t^\star) + \tilde{U}_{it}$. For this DGP, we note that Assumption 7 follows from the fact that $\pi_i \varphi_i^\star \in \mathcal{H}^\beta(\Omega, B)$. Since U_{it} is sub-Gaussian with uniformly bounded sub-Gaussian norm and $\varphi_i^\star(\cdot)$ is bounded, it holds that $\tilde{U}_{it}$ is a mean zero sub-Gaussian variables with uniformly bounded sub-Gaussian norm. In addition, since we assume U_{it} is i.i.d., it holds that given $F_t^\star$, $\tilde{U}_{it}$ is independent to each other, which verifies Assumption 6. Therefore, by Theorem 6, with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned}
&\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it} - \pi_i X_{it}^\star\right)^2 \\
\lesssim &\left(T^{-\frac{2\beta}{2\beta+K}} + N^{-1}K_1 + T^{-1}\inf_{\rho\in\mathcal{F}_N^K}\sum_{t=1}^{T}\|\rho(\tilde{X}_t) - F_t^\star\|^2\right)\log^4(T). \qquad (2.24)
\end{aligned}$$

In addition, since X_{it} is missing independently across t, it is not hard to check by Bernstein inequality that $\mathrm{P}(|\widehat{\pi}_i - \pi_i| > T^{-\frac{\beta}{2\beta+K}}\log(T)) \lesssim C\exp(-cT^{-\frac{K}{2\beta+K}}\log^2(T))$. With union bound inequality, we conclude that with probability at least $1 - CT\exp(-cT^{-\frac{K}{2\beta+K}}\log^2(T)) \gtrsim 1 - C\exp(-cT^{-\frac{K}{2\beta+K}})$, $\max_{1\le i\le N}|\widehat{\pi}_i - \pi_i| \le T^{-\frac{\beta}{2\beta+K}}\log(T)$. Together with the facts that $\widehat{\pi}_i, \pi_i \ge \pi_{\min}$ and $|\widehat{X}_{it}| \le B$, with probability at least $1 - CT\exp(-c\log^2 T)$,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{N}\left(\widehat{X}_{it}/\widehat{\pi}_i - X_{it}^\star\right)^2 \lesssim \sum_{t=1}^{T}\sum_{i=1}^{N}\left[\left(\widehat{X}_{it}/\pi_i - X_{it}^\star\right)^2 + \left(\widehat{X}_{it}/\pi_i - \widehat{X}_{it}/\widehat{\pi}_i\right)^2\right] \\
\lesssim &\sum_{t=1}^{T}\sum_{i=1}^{N}\left[\left(\widehat{X}_{it} - \pi_i X_{it}^\star\right)^2 + \left(1/\pi_i - 1/\widehat{\pi}_i\right)^2\right] \\
\lesssim &\sum_{t=1}^{T}\sum_{i=1}^{N}\left[\left(\widehat{X}_{it} - \pi_i X_{it}^\star\right)^2 + T^{-\frac{2\beta}{2\beta+K}}\log^2(T)\right].
\end{aligned}$$

Combining the above inequality and Eq. (2.24), the corollary holds.

2.7 Empirical Applications in Economics

In this section, we discuss three separate applications that illustrate the use of AEs and SAEs with economic datasets.

2.7.1 Macroeconomic Forecasting

Forecasting economic developments is essential for effective decision-making in policy formulation, consumption and investment planning, and financial forecasting. Accurate predictions help stakeholders anticipate future trends and adapt their strategies accordingly.

A variety of methods has been proposed to aggregate macroeconomic indicators for predicting individual macroeconomic variables, such as unemployment rates, inflation, and short-term interest rates. Prominent approaches include factor-based models developed by Stock and Watson [1999] and Stock and Watson [2002], Bayesian model averaging (Koop and Potter [2004] and Wright [2009]), and ensemble learning methods such as bagging and boosting (Inoue and Kilian [2008] and Bai and Ng [2009]).

Recently, there has been growing interest in methods designed to forecast multiple variables simultaneously. Examples include Bayesian reduced-rank multivariate models (Carriero et al. [2011]) and Bayesian vector autoregressions (Bańbura et al. [2010]). At the same time, forecasting techniques originally developed for single variables can often be adapted to panels by applying algorithms separately to each target variable. Stock and Watson [2012] provide a comprehensive comparison of panel forecasting methods and emphasize the strong performance of factor-based approaches in such settings.

This study focuses on forecasting an entire panel of economic indicators, employing the FRED-MD dataset compiled by McCracken and Ng [2016]. This dataset features a comprehensive range of economic indicators across categories such as output and income, labor markets, consumption, orders and inventories, money and credit, interest rates and exchange rates, prices, and the stock market. It includes 119 predictors spanning January 1960 to De-

cember 2019 at a monthly frequency.[9]

Formally, let $X_t \in \mathbb{R}^{119}$ represent the predictor variables, and let Y_{t+1} denote the target variables for the subsequent month. SAEs are employed to predict Y_{t+1}. Since the focus is on forecasting the entire panel, we set $Y_{t+1} = X_{t+1}$. On the basis of Stock and Watson [2012], we center our analysis on the residuals of a prior AR(6) fit for Y_{t+1}.[10] This approach enables us to evaluate the models' ability to predict beyond the autoregressive component. To benchmark the performance of SAEs, we compare their predictions against those of the PCR approach, which is widely regarded as a robust baseline, as highlighted by Stock and Watson [2012].

Training the SAE involves an expanding window approach. The initial 20 years of data (1960–1979) serve as the training set, while the subsequent 10 years (1980–1989) form a validation set used for tuning parameters such as the learning rate and pruning ratio. Rather than optimizing the number of factors, results are reported for each specified factor count. Performance is evaluated on the following year (1990). This process is repeated 30 times, with the training set expanding by one year and the validation set shifting accordingly at each iteration. For PCR, performance is also reported for each specified number of factors. Since it does not involve tuning parameters any more, the training and validation sets are combined to produce direct estimates.

The results in Figure 2.8 highlight a significant advantage of SAEs over PCR. Among the PCR models, the single-factor specification consistently outperforms other configurations, reinforcing the effectiveness of simplicity in this context. In contrast, SAEs, particularly SAE1 through SAE3, exhibit superior performance. The box plots for these models demonstrate

9. We adhere to the preprocessing steps outlined by McCracken and Ng [2016], including necessary transformations and the exclusion of variables with missing values in earlier years or those with long lags. These measures ensure the robustness of our analysis and facilitate comparability with other studies in the literature.

10. To avoid forward-looking bias, the AR(6) model is refitted using an expanding window of the training sample at each step.

narrower spreads and consistently positive median R^2 values across all factor counts, underscoring their robust predictive capabilities. Notably, SAE2 with two factors appears to dominate other scenarios. SAE4, however, performs poorly, showing clear signs of overfitting due to its more complex structure. In comparison, SAE5—while fully connected—has a much smaller number of parameters than SAE4. As a result, its performance is closely aligned with that of SAE1 through SAE3, albeit slightly lower. To further test the statistical significance of the improvements over PCR, we apply the Wilcoxon signed-rank test to the R^2 values. This test evaluates the null hypothesis that the median R^2 is zero against the alternative that the median is positive. The corresponding p-values are presented in Table 2.2. SAE2 dominates PCR with at least two factors, while SAE1 achieves dominance with three or more factors. Although SAE3 performs weaker, it still surpasses PCR with four factors and with three or five factors at the 10% significance level. Overall, our results underscore the advantage of the disjoint output decoder structure in SAEs. By effectively extracting nonlinear components, this architecture contributes to stable and improved macroeconomic forecasting.

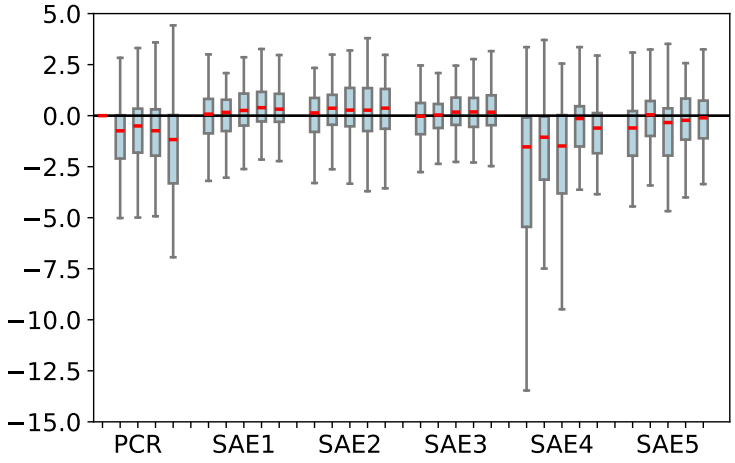


Figure 2.8: Out-of-sample R^2s in Macro Panel Forecasting

Note: This figure compares the distributions of out-of-sample R^2 values (in percentages) for SAEs and PCR across different numbers of factors (ranging from 1 to 5), benchmarked against PCR with one factor, in forecasting the cross-section of macroeconomic indicators one month ahead at a monthly frequency.

K	1	2	3	4	5
PCR	–	1.0	1.0	1.0	1.0
SAE1	0.549	0.441	0.04	0.003	0.002
SAE2	0.537	0.014	0.035	0.073	0.033
SAE3	0.836	0.647	0.073	0.047	0.085
SAE4	1.0	1.0	1.0	0.996	1.0
SAE5	1.0	0.682	1.0	0.952	0.968

Table 2.2: P-values of Wilcoxon Signed-rank Test

Note: This table reports the p-values from the Wilcoxon signed-rank test applied to out-of-sample R^2 values, benchmarked against PCR with one factor. K_1 denotes the number of factors in PCA and the number of neurons in the bottleneck layers of the SAEs.

2.7.2 Predicting the Cross-Section of Factor Returns

Our second exercise examines an asset pricing application centered on factor investing. Over decades, research has identified systematic factors that explain cross-sectional variations in expected returns, driving the popularity of factor-based investing. Like aggregate market returns, individual factor returns may exhibit predictability, presenting investment opportunities for dynamically adjusting positions on different factors based on their expected performance.

Extensive research has explored the predictability of individual factors, such as value (Cohen et al. [2003]) and momentum (Cooper et al. [2004]; Daniel and Moskowitz [2016]), as well as the joint predictability of multiple factors (Stambaugh et al. [2012]; Akbas et al. [2015]). For instance, Arnott et al. [2023] investigate factor momentum and demonstrate that past factor returns can predict future returns in the cross-section. Expanding on this, Haddad et al. [2020] use factors' book-to-market (BM) ratios—calculated by weighting individual equities' BM ratios—to predict the five principal components of factor returns, which are subsequently translated into predictions of the returns of individual factors.

Our approach differs by directly predicting the cross-section of factor returns using each factor's characteristics, such as BM ratios and momentum measures derived from factor returns over various horizons. Utilizing the SAE architecture examined in simulations, we

extract nonlinear, low-dimensional components from these factor characteristics ($X_{i,t}$), while simultaneously predicting the cross-section of factor returns ($Y_{i,t+1}$) for the next period. Unlike our focus on factor-level predictions, an important application of AEs in asset pricing is demonstrated in the work of Gu et al. [2021b], which proposes Conditional AEs to model individual stock returns by incorporating firm characteristics as covariates alongside the input returns.

Our analysis uses the extended monthly dataset from Haddad et al. [2020], accessed via Serhiy Kozak's website, spanning January 1974 to December 2019. This dataset includes 50 long-short, characteristic-sorted decile portfolios, along with their BM ratios. In addition to BM ratios, we calculate trailing 1-month (MOM1), 6-month (MOM6), and 12-month (MOM12) returns as alternative predictors. The inclusion of MOM12 requires the sample period to begin in 1975.

To train the SAE, we adopt an expanding window approach. Specifically, the first 15 years of data serve as the training set, while the following 5 years form the validation set. Performance is then evaluated in the subsequent year. This procedure is repeated 25 times, with the training set expanding by one year and the validation set shifting accordingly at each iteration. For comparison, we implement PCA as in Haddad et al. [2020], using combined training and validation sets for estimation without tuning the number of factors.

Table 2.3 reports the Sharpe ratio of a long-short portfolio constructed based on predicted factor returns. The portfolio takes long positions in the top 10 (20%) factors and short positions in the bottom 10, ranked by their predicted values. The Sharpe ratio provides an economic assessment of the prediction power.

For BM, the PCA-based approach achieves a maximum Sharpe ratio of 0.452 with 4 principal components (PCs). Using MOM1 as input yields a slightly higher Sharpe ratio of 0.654. However, PCA's performance declines with MOM6 and MOM12 inputs, with Sharpe ratios below 0.226 and 0.357, respectively, regardless of the number of factors.

	BM						MOM1				
K_1	1	2	3	4	5	K_1	1	2	3	4	5
PCA	0.158	0.351	0.33	0.452	0.425	PCA	0.133	0.529	0.475	0.654	0.597
SAE1	0.633	0.393	0.507	0.52	0.484	SAE1	0.512	0.652	0.704	0.748	0.802
SAE2	0.742	0.82	0.728	0.514	0.714	SAE2	0.904	0.867	0.802	0.727	0.738
SAE3	0.558	0.586	0.46	0.675	0.564	SAE3	0.744	0.662	0.809	1.007	0.779
SAE4	0.304	0.316	0.603	0.361	0.388	SAE4	0.638	0.831	0.946	0.962	0.999
SAE5	0.645	0.322	0.288	0.293	0.526	SAE5	0.998	0.897	0.818	0.889	0.865
	MOM6						MOM12				
PCA	-0.062	0.078	0.094	0.226	0.211	PCA	0.155	0.12	0.065	0.357	0.298
SAE1	0.612	0.515	0.525	0.507	0.706	SAE1	0.645	0.76	0.831	0.396	0.534
SAE2	0.723	0.736	0.701	0.737	0.718	SAE2	0.661	0.613	0.487	0.67	0.59
SAE3	0.41	0.529	0.611	0.673	0.519	SAE3	0.834	0.802	0.751	0.658	0.373
SAE4	0.45	0.38	0.522	0.403	0.573	SAE4	0.446	0.606	0.531	0.478	0.638
SAE5	0.568	0.701	0.743	0.592	0.532	SAE5	0.763	0.768	0.859	0.835	0.704

Table 2.3: Out-of-sample Sharpe Ratios

Note: The table summarizes the empirical performance of predicting the cross-section of expected factor returns, measured by the Sharpe ratios of long-short portfolios. These portfolios are constructed from next-month predictions by SAEs and PCA across various numbers of factors, using BM, MOM1, MOM6, and MOM12 as predictors. Portfolios are rebalanced monthly.

SAEs consistently outperform PCA in nearly all cases, often by a substantial margin. For example, with MOM1 as input, the SAE3 model achieves an impressive Sharpe ratio of 1.0 with 4 factors. Similarly, for BM, the SAE2 model with two factors achieves a Sharpe ratio of 0.82. Using MOM6 and MOM12 as inputs, the same SAE1 model achieves Sharpe ratios of 0.736 and 0.652, respectively. The strong positive predictability associated with MOM1 supports evidence of factor momentum, particularly evident in returns sorted by short-term factor performance, aligning with the findings of Arnott et al. [2023]. Notably, SAE4 and SAE5 also deliver competitive results, particularly for MOM across various horizons.

Overall, SAEs demonstrate significantly stronger performance than PCA, underscoring their superior ability to capture nonlinear predictive signals effectively.

2.7.3 Causal Analysis with Corrupted Data

In the final exercise, we revisit the study by Agarwal and Singh [2024] on causal analysis with corrupted data, focusing on the well-known Gaussian mechanism introduced for differential privacy (Dwork et al. [2006]; Dwork and Roth [2014]). This mechanism deliberately adds Gaussian noise to protect respondent privacy while preserving data utility.[11]

The economic context involves recovering the effect of import competition from China on U.S. labor markets, as studied in the influential work by Autor et al. [2013]. Their panel dataset is organized at the commuting zone (CZ) level, encompassing 722 CZs across two time periods: the 1990s and 2000s. Each CZ is represented by a vector of 30 covariates, including variables from the authors' preferred specification as well as auxiliary variables detailed in their appendix.[12]

The dataset captures multiple dimensions of the same underlying economic indicators, thereby exhibiting a pronounced factor structure. Table 2.4 highlights this low-dimensional structure by comparing the variance explained by AEs and PCA. AEs consistently recover more variance than PCA for a given number of factors. Notably, while PCA requires over 9 factors to achieve 80% explanatory power, AEs need only 3–5 components.

Next, we introduce synthetic Gaussian noise into the original 1990s dataset ($X^\star \in \mathbb{R}^{30\times 722}$), referred to as the clean data, to obtain the corrupted data X with SNRs of 4, 2, and 1. PCA and AEs are applied to X to denoise the inputs , and the MSEs between the original data $X_i^\star$ and the reconstructed output $\widehat{X}_i$ are computed for each dimension i.

11. Another common approach is to add Laplace-distributed noise. The empirical results using the Laplace mechanism are nearly identical and are therefore omitted.

12. These variables are drawn from Column 6 in Table 3 and Appendix Table 2 of Autor et al. [2013]. They include percentages of employment in manufacturing, college-educated population, and foreign-born population; percentages of employment among women and in routine occupations; average offshorability index of occupations; Census division dummies; and percentages of the working-age population: employed in manufacturing, employed in non-manufacturing, unemployed, not in the labor force, and receiving disability benefits. Additional variables include average log weekly wages (manufacturing and non-manufacturing), average benefits per capita (individual transfers, retirement, disability, medical, federal income assistance, unemployment, and TAA), and average household income per working-age adult (total and wage/salary).

K_1	1	3	5	7	9
PCA	24.1	50.9	64.9	74.1	81.7
AE1	44.5	82.0	91.1	93.8	95.8
AE2	40.8	76.9	86.2	91.6	95.1
AE3	51.2	86.2	92.4	95.3	96.7
AE4	49.8	91.2	95.1	98.0	98.8
AE5	43.5	80.0	89.2	93.3	94.6

Table 2.4: Cumulative Percentage of Variance Explained by Extracted Factors

Note: This table reports the cumulative percentage of variance explained by an incremental number of factors for PCA and AEs, respectively.

The denoised outputs are subsequently used in a causal analysis framework to assess the effectiveness of these data-cleaning methods.

Figure 2.9 presents the average MSEs across all variables. Consistent with our simulation studies, AEs demonstrate superior performance in recovering the underlying data, achieving significantly smaller MSEs compared to PCA. As the SNR decreases, performance deteriorates, resulting in larger MSEs. Among the AEs, AE4 performs the worst, as expected, due to overfitting caused by its excessive number of parameters. We next explore how these reconstruction errors affect the final causal analysis.

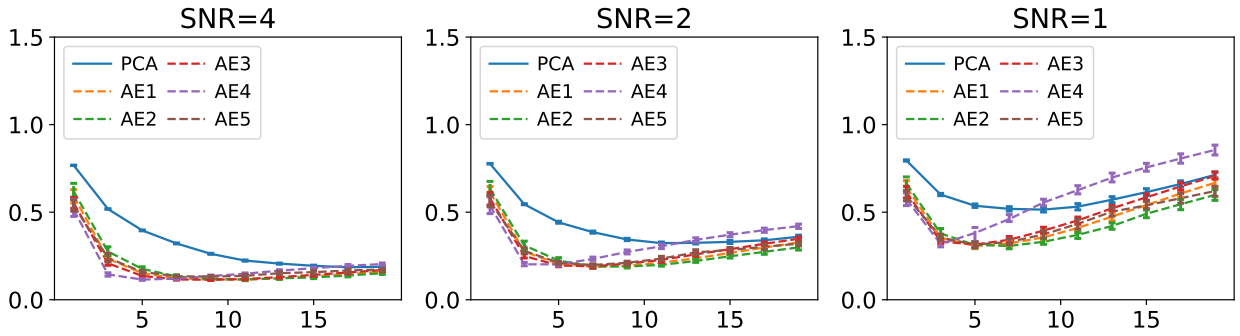


Figure 2.9: Effectiveness in Reducing Synthetic Measurement Error

Note: This figure compares the MSEs for AEs and PCA across different numbers of factors in eliminating synthetic errors from the covariates.

For consistency with Autor et al. [2013], we use their specified variables for the two-stage least squares (TSLS) estimation, while the data-cleaning step benefits from the inclusion of

auxiliary variables to enhance the recovery of the factor structure. We repeat the denoising process 100 times and present the distribution of TSLS estimates based on AEs and PCA, with the number of factors fixed at $K_1 = 5$ in Figure 2.10. For comparison, we also compute the causal effects directly using the noisy data without any cleaning. Notably, while larger noise (SNR = 1) reduces variance of TSLS estimates, it results in a greater bias. When applying TSLS to the outputs of AEs, the median of estimated causal effects from 100 experiments are close to the value reported in the original paper ($-0.596 \pm 1.96 \cdot 0.099$). Specifically, for SNR=2, the median estimates are -0.639 for AE1, -0.659 for AE2, and -0.652 for AE3. PCA, by comparison, yields a corresponding estimate of -0.674, with smaller variance than AEs but a larger bias, even exceeding that of using the noisy data directly.

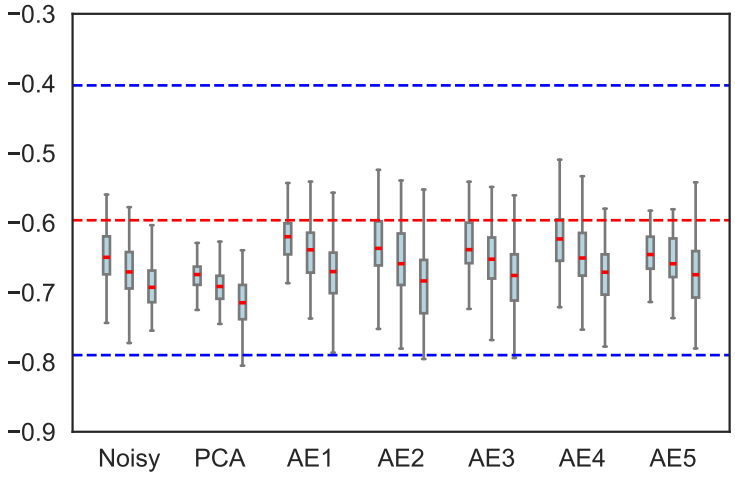


Figure 2.10: Effect of Noise Reduction on Causal Analysis Outcomes

Note: This figure presents boxplots of the TSLS estimates based on 100 repeated experiments with synthetic noise added to the covariates. The "Noisy" case represents the estimates obtained directly from the corrupted data, while the remaining boxplots correspond to estimates after applying respective data-cleaning methods using PCA and AEs. Each of the three boxes from left to right correspond to three SNRs from 4 to 1. The red dashed line denotes the causal effect derived from clean data, and the blue dashed lines mark its confidence interval.

2.8 Supplemental Mathematical Proofs

2.8.1 Approximation Error of Neural Networks

Proposition 4. *Assume $f \in \mathcal{H}^p([-a,a]^r, C)$ for some $p = q + s$, $q \in \mathbb{N}_0$ and $s \in (0,1]$, and $C > 0$. Then for $a \geq 1$ and ζ sufficiently large (only depends on fixed constants including p, a, C and $\|f\|_{C^q}$), when $d \asymp \lceil \log_4 \left(\zeta^{2p}\right) \rceil \cdot (\lceil \log_2(\max\{q, r\} + 1\}) \rceil + 1)$ and $w \asymp 2^r \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r$, there exists a neural network $\widehat{f}_{wide} \in \mathcal{F}_r^1(d, w, \zeta^{5p+5}, B)$ such that $\left\| f - \widehat{f}_{wide} \right\|_\infty \leq C\zeta^{-2p}$, for some constant C only depends on fixed parameters.*

Proof. The proof primarily follows the same argument as in Kohler and Langer [2021], with the main difference being the necessity to construct the neural networks used in their paper with bounded weights. First, we introduce the notations required for the proof.

Notations: Given an $r-$dimensional cube C, we denote the "bottom left" corner of C as C_{left}. Thus, each half-open cube C with side length s can be expressed as a polytope defined by $-x^{(j)} + C_{\text{left}}^{(j)} \leq 0$ and $x^{(j)} - C_{\text{left}}^{(j)} - s < 0 \quad (j \in \{1, \ldots, r\})$. Moreover, we denote by $C_\delta^0 \subset C$ the cube containing all $x \in C$ that have a distance of at least δ from the boundaries of C, i.e., a polytope defined by $-x^{(j)} + C_{\text{left}}^{(j)} \leq -\delta$ and $x^{(j)} - C_{\text{left}}^{(j)} - s < -\delta \quad (j \in \{1, \ldots, r\})$. If $\mathcal{P}$ is a partition of cubes of $[-a,a)^r$ and $x \in [-a,a)^r$, then we denote the cube $C \in \mathcal{P}$, which satisfies $x \in C$, by $C_{\mathcal{P}}(x)$. We partition $[-a,a)^r$ into ζ^r and ζ^{2r} half-open equivolume cubes of the form $[\alpha, \beta) = \left[\alpha^{(1)}, \beta^{(1)}\right) \times \cdots \times \left[\alpha^{(r)}, \beta^{(r)}\right)$, for $\alpha, \beta \in \mathbb{R}^r$, respectively. Let

$$\mathcal{P}_1 = \left\{C_{k,1}\right\}_{k \in \{1, \ldots, \zeta^r\}} \text{ and } \mathcal{P}_2 = \left\{C_{j,2}\right\}_{j \in \{1, \ldots, \zeta^{2r}\}} \tag{2.25}$$

be the corresponding partitions. We denote for each $i \in \{1, \ldots, \zeta^r\}$ those cubes of $\mathcal{P}_2$ that are contained in $C_{i,1}$ by $\widehat{C}_{1,i}, \ldots, \widehat{C}_{\zeta^r,i}$. Here we order the cubes in such a way that $\left(\widehat{C}_{k,i}\right)_{\text{left}} = \left(C_{i,1}\right)_{\text{left}} + v_k$ holds for all $k \in \{1, \ldots, \zeta^r\}, i \in \{1, \ldots, \zeta^r\}$ and for some vector v_k with entries in $\left\{0, 2a/\zeta^2, \ldots, (\zeta - 1) \cdot 2a/\zeta^2\right\}$. The vector v_k describes the position of

$\left(\widehat{C}_{k,i}\right)_{\text{left}}$ relative to $\left(C_{i,1}\right)_{\text{left}}$, and we order the above cubes such that this position is independent of i. Finally, define the identity function achieved by ReLU neural network,

$$\widehat{f}_{id}(z) = \sigma(z) - \sigma(-z) = z, \quad z \in \mathbb{R}, \text{ and}$$
$$\widehat{f}_{id}(x) = \left(\widehat{f}_{id}\left(x^{(1)}\right), \ldots, \widehat{f}_{id}\left(x^{(r)}\right)\right) = \left(x^{(1)}, \ldots, x^{(r)}\right), \quad x \in \mathbb{R}^r.$$

We further let $\widehat{f}_{id}^{t+1}(x) = \widehat{f}_{id}\left(\widehat{f}_{id}^{t}(x)\right) = x$, $t \in \mathbb{N}_+, x \in \mathbb{R}^r$. Note that the weights of these neural network are all bounded by ζ^{5p+5}.

With these notations established, the proof proceeds by constructing neural networks with bounded weights in a manner similar to the approach outlined in Kohler and Langer [2021]. We demonstrate that for all the neural networks utilized in their proof, there exists an alternative with a number of layers twenty times theirs and with weights bounded by ζ^{5p+5}. We partition the proof into three steps, which correspond to key steps $2-4$ in the proof of Theorem 2(a) of Kohler and Langer [2021].

2.8.1.1 Step I: An Initial Approximation

We construct a neural network $\widehat{f}_{\mathcal{P}_2}(x)$ that approximates the target function for all $x \in \bigcup_{j\in\{1,\ldots,\zeta^{2r}\}} \left(C_{j,2}\right)_{1/\zeta^{2p+2}}^{0}$. To prove this, we prove in Lemmas 34 and 35 that there exist $\widehat{f}_{mult,r}$ and $\widehat{f}_p$, which are weight-bounded versions of $\widehat{f}_{mult,r}$ and $\widehat{f}_p$ defined in Lemmas 8 and 5 of Kohler and Langer [2021] and are used to approximate monomials, that achieve the same approximation error as theirs but with the number of layers twenty times theirs. Also,

we adopt the definition of $\widehat{f}_{\text{ind},[a,b)}$ and $\widehat{f}_{\text{test}}\ (x,a,b,s)$ in Lemma 6 of their paper:

$$\widehat{f}_{\text{ind},[a,b)}(x) = \sigma\left(1 - \zeta^{2p+2}\cdot\sum_{i=1}^{r}\left(\sigma\left(a^{(i)} + \zeta^{-2p-2} - x^{(i)}\right) + \sigma\left(x^{(i)} - b^{(i)} + \zeta^{-2p-2}\right)\right)\right)$$

$$\begin{aligned}&\widehat{f}_{\text{test}}\ (x,a,b,s) \qquad (2.26)\\ &= \sigma\left(\widehat{f}_{id}(s) - \zeta^{4p+4}\cdot\sum_{i=1}^{r}\left(\sigma\left(a^{(i)} + \zeta^{-2p-2} - x^{(i)}\right) + \sigma\left(x^{(i)} - b^{(i)} + \zeta^{-2p-2}\right)\right)\right)\\ &\quad - \sigma\left(-\widehat{f}_{id}(s) - \zeta^{4p+4}\cdot\sum_{i=1}^{r}\left(\sigma\left(a^{(i)} + \zeta^{-2p-2} - x^{(i)}\right) + \sigma\left(x^{(i)} - b^{(i)} + \zeta^{-2p-2}\right)\right)\right).\end{aligned}$$

Note that $\widehat{f}_{\text{ind},[a,b)}(x) \in \mathcal{F}_r^1(2, 2d, \zeta^{5p+5})$ satisfies

$$\widehat{f}_{\text{ind},[a,b)}(x) = \mathbf{1}_{[a,b)}(x)$$

for $x \in K_\zeta := \{x \in \mathbb{R}^r : x^{(i)} \notin \left[a^{(i)}, a^{(i)} + \zeta^{-2p-2}\right) \cup \left(b^{(i)} - \zeta^{-2p-2}, b^{(i)}\right), \forall i \in \{1,\ldots,r\}\}$ and $|\widehat{f}_{\text{ind},[a,b)}(x) - \mathbf{1}_{[a,b)}(x)| \le 1$ for $x \in \mathbb{R}^r$. The function $\widehat{f}_{\text{test}}\ (x,a,b,s) \in \mathcal{F}_r^1(2, 2d, \zeta^{5p+5})$ satisfies $\widehat{f}_{\text{test}}\ (x,a,b,s) = s\mathbf{1}_{[a,b)}(x)$ for $x \in K_{1/R}$ and $|\widehat{f}_{\text{test}}\ (x,a,b,s) - s\mathbf{1}_{[a,b)}(x)| \le |s|$ for

$x \in \mathbb{R}^r$. Additionally, we adopt the definitions of functions in their proof of Lemma 3:

$$
\begin{aligned}
\widehat{\phi}_{1,1} &= \left(\widehat{\phi}_{1,1}^{(1)}, \ldots, \widehat{\phi}_{1,1}^{(r)}\right) = \widehat{f}_{id}^{2}(x), \\
\widehat{\phi}_{2,1} &= \left(\widehat{\phi}_{2,1}^{(1)}, \ldots, \widehat{\phi}_{2,1}^{(r)}\right) = \sum_{i \in \{1,\ldots,\zeta^r\}} \left(C_{i,1}\right)_{left} \cdot \widehat{f}_{ind,C_{i,1}}(x), \\
\widehat{\phi}_{3,1}^{(l,j)} &= \sum_{i \in \{1,\ldots,\zeta^r\}} \left(D^l f\right)\left(\left(\tilde{C}_{j,i}\right)_{left}\right) \cdot \widehat{f}_{ind,C_{i,1}}(x), \\
\widehat{\phi}_{1,2} &= \left(\widehat{\phi}_{1,2}^{(1)}, \ldots, \widehat{\phi}_{1,2}^{(r)}\right) = \widehat{f}_{id}^{2}\left(\widehat{\phi}_{1,1}\right), \\
\widehat{\phi}_{2,2}^{(i)} &= \sum_{j=1}^{\zeta^r} \widehat{f}_{\text{test}}\left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1} + v_j, \widehat{\phi}_{2,1} + v_j + \frac{2a}{\zeta^2} \cdot 1, \widehat{\phi}_{2,1}^{(i)} + v_j^{(i)}\right), \\
\widehat{\phi}_{3,2}^{(l)} &= \sum_{j=1}^{\zeta^r} \widehat{f}_{\text{test}}\left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1} + v_j, \widehat{\phi}_{2,1} + v_j + \frac{2a}{\zeta^2} \cdot 1, \widehat{\phi}_{3,1}^{(1,j)}\right),
\end{aligned}
\tag{2.27}
$$

for $j \in \{1, \ldots, \zeta^r\}$, $l \in \mathbb{N}_0^d$ with $\|l\|_1 \leq q$ and $i \in \{1, \ldots, r\}$. These functions are fundamental to the desired neural network that approximate the target function. It can be readily observed that all parameters of the aforementioned neural networks are bounded by ζ^{5p+5} for sufficiently large values of ζ given that $\|f\|_{C^q}$ is finite. Select $l_1, \ldots, l_{\binom{r+q}{r}}$ to satisfy $\left\{l_1, \ldots, l_{\binom{r+q}{r}}\right\} = \left\{(s_1, \ldots, s_d) \in \mathbb{N}_0^r : s_1 + \cdots + s_d \leq q\right\}$ and define $r_i = \frac{1}{l_i!}$. Additionally, we define monomial $m_i(x) = x^{l_i}$ and function $p\left(x, y_1, \ldots, y_{\binom{r+q}{r}}\right) = \sum_{i=1}^{\binom{r+q}{r}} r_i \cdot y_i \cdot m_i(x)$. Define $B_{\zeta,p} := \left\lceil \log_4\left(\zeta^{2p}\right)\right\rceil$. By Lemma 35, $\max_i |r_i| = 1$, and the fact that $(4 \max\{2a, \|f\|_{C^q}\})^{2q+2} < \zeta^{5p+5}$ as long as ζ is big enough, a function

$$
\widehat{f}_p \in \mathcal{F}^1_{\binom{r+q}{r}+r}\left(20 B_{\zeta,p} \cdot \lceil \log_2(\max\{q+1, 2\})\rceil, 18 \cdot (q+1) \cdot \binom{r+q}{r}, \zeta^{5p+5}\right)
$$

exists such that for all $|z^{(1)}|, \ldots, |z^{(r)}|, |y_1|, \ldots, |y_{\binom{r+q}{r}}| \leq \max\{2a, \|f\|_{C^q}\}$, it holds that

$$
\left|\widehat{f}_p\left(z, y_1, \ldots, y_{\binom{r+q}{r}}\right) - p\left(z, y_1, \ldots, y_{\binom{r+q}{r}}\right)\right| \leq C \cdot \left(\max\{2a, \|f\|_{C^q}\}\right)^{4(q+1)} \cdot 4^{-B_{\zeta,p}}
$$

for some fixed constant C. Finally, we define

$$\widehat{f}_{\mathcal{P}_2}(x) := \widehat{f}_p\left(\widehat{\phi}_{1,2}(x) - \widehat{\phi}_{2,2}(x), y_1, \ldots, y_{\binom{r+q}{r}}\right) \tag{2.28}$$

where $z := \widehat{\phi}_{1,2}(x) - \widehat{\phi}_{2,2}(x)$ and $y_v := \widehat{\phi}_{3,2}^{(l_v)}(x)$ for $v \in \left\{1, \ldots, \binom{r+q}{r}\right\}$. Lemma 36 states that $\left|\widehat{f}_{\mathcal{P}_2}(x) - f(x)\right| \leq C \cdot (\max\{2a, \|f\|_{C^q}\})^{4(q+1)} \cdot \zeta^{-2p}$ holds for all $x \in \bigcup_{j \in \{1,\ldots,\zeta^{2r}\}} \left(C_{j,2}\right)^0_{1/\zeta^{2p+2}}$, i.e., $\widehat{f}_{\mathcal{P}_2}$ approximates the target function well for certain area.

2.8.1.2 Step II: Approximating $w_{\mathcal{P}_2}(x) \cdot f(x)$

Define

$$w_{\mathcal{P}_2}(x) = \prod_{j=1}^{r} \left(1 - \frac{\zeta^2}{a} \cdot \left|\left(C_{\mathcal{P}_2}(x)\right)^{(j)}_{\text{left}} + \frac{a}{\zeta^2} - x^{(j)}\right|\right)_+. \tag{2.29}$$

This function is a linear tensorproduct B-spline which takes its maximum value at the center of $C_{\mathcal{P}_2}(x)$. In this section, we construct a neural network that effectively approximates $w_{\mathcal{P}_2}(x) \cdot f(x)$. To achieve this, we establish $\widehat{f}_{w_{\mathcal{P}_2}}$ and $\widehat{f}_{check,\mathcal{P}_2}$ in Lemmas 37 and 38, which are the weight-bounded versions of functions $\widehat{f}_{w_{\mathcal{P}_2}}$ and $\widehat{f}_{check,\mathcal{P}_2}$ defined in Lemmas 9 and 10 of Kohler and Langer [2021]. By Lemmas 36 and 37, the functions $\widehat{f}_{\mathcal{P}_2}$ and $\widehat{f}_{w_{\mathcal{P}_2}}$ can effectively approximate f and $w_{\mathcal{P}_2}$ except for a certain region, while $\widehat{f}_{check,\mathcal{P}_2}$ is used to check if the input x lies in that region. Based on this, define

$$\widehat{f}_{\mathcal{P}_2,\text{true}}(x) = \sigma\left(\widehat{f}_{\mathcal{P}_2}(x) - B_{\text{true}} \cdot \widehat{f}_{\text{check},\mathcal{P}_2}(x)\right) - \sigma\left(-\widehat{f}_{\mathcal{P}_2}(x) - B_{\text{true}} \cdot \widehat{f}_{\text{check},\mathcal{P}_2}(x)\right),$$

where $B_{\text{true}} = 2 \cdot e^{2ar} \cdot \max\{\|f\|_{C^q}, 1\}$. By Lemma 36, $|\widehat{f}_{\mathcal{P}_2}(x)| \leq B_{\text{true}}$ for all $x \in [-a, a)^r$. Intuitively, $\widehat{f}_{\mathcal{P}_2,\text{true}}$ becomes zero when x lies in the region that $\widehat{f}_{\mathcal{P}_2}$ fails to approximate the target function, while it equals $\widehat{f}_{\mathcal{P}_2}$ when x is outside that region. Since B_{true} and the weights of $\widehat{f}_{\mathcal{P}_2}$ and $\widehat{f}_{\text{check},\mathcal{P}_2}(x)$ are all bounded by ζ^{5p+5} when ζ becomes

sufficiently large, we see that the weights of $\widehat{f}_{\mathcal{P}_2,\text{true}}$ are also bounded by ζ^{5p+5}. Let $\widehat{f}_{\text{mult}} \in \mathcal{F}_2^1(20\left\lceil \log_4\left(\zeta^{2p}\right)\right\rceil, 18, \zeta^{5p+5})$ be the function defined in Eq. (2.39) with a there being replaced by $2 \cdot \max\left\{\|f\|_{\infty,[-a,a]^r}, 1\right\}$. This function is used to approximate the product of the inputs. Finally, we let

$$\widehat{f}(x) = \widehat{f}_{\text{mult}}\ \left(\widehat{f}_{w_{\mathcal{P}_2}}(x), \widehat{f}_{\mathcal{P}_2,\text{ true}}\,(x)\right). \tag{2.30}$$

Lemma 39 demonstrates that $\left|\widehat{f}(x) - w_{\mathcal{P}_2}(x) \cdot f(x)\right| \leq C \cdot (\max\{2a, \|f\|_{C^q}\})^{4(q+1)} \cdot \zeta^{-2p}$ holds for $x \in [-a, a)^r$.

2.8.1.3 Step III: Applying $\widehat{f}$ to slightly shifted partitions

Based on the previous section, we are able to approximate $w_{\mathcal{P}_2}(x) \cdot f(x)$ well by neural networks. The function $w_{\mathcal{P}_2}(x)$ plays a role of weight function that achieves its maximum at the center of $C_{\mathcal{P}_2}(x)$. Intuitively, by further defining a series of weight functions such that their sum equals one, the sum of corresponding neural network as $\widehat{f}$ can effectively approximate the target function f. The proof for this section uses this idea.

Recall the definition of $\mathcal{P}_1$ and $\mathcal{P}_2$ in Eq. (2.25). Let $\mathcal{P}_{1,1} = \mathcal{P}_1$ and $\mathcal{P}_{2,1} = \mathcal{P}_2$ and define for each $v \in \{2, 3, \ldots, 2^r\}$ partitions $\mathcal{P}_{1,v}$ and $\mathcal{P}_{2,v}$, which are modifications of $\mathcal{P}_{1,1}$ and $\mathcal{P}_{2,1}$ where at least one of the components it shifted by $\zeta^{-2}a$. That is, $\mathcal{P}_{1,v}$ is of the form $\mathcal{P}_1 + \sum_{k\in S}\zeta^{-2}ae_k$ for S being a subset of $\{1, 2, \ldots, r\}$, where e_k represents the k-th basis vector of $\mathbb{R}^r$. Accordingly, we define the function $w_{\mathcal{P}_{2,v}}$ and the neural network $\widehat{f}_v$ in the same way as Eq. (2.29) and Eq. (2.30) corresponding to the partitions $\mathcal{P}_{1,v}$ and $\mathcal{P}_{2,v}$ for $v \in \{1, \ldots, 2^r\}$, respectively. Note that $w_{\mathcal{P}_{2,1}} + \cdots + w_{\mathcal{P}_{2,2^r}} = 1$ for $x \in [-a/2, a/2]^d$, which is proven Page 32 of the Supplementary A of Kohler and Langer [2021]. Finally we let $\widehat{f}_{\text{wide}}\,(x) = \sum_{v=1}^{2^r}\widehat{f}_v(x) \in \mathcal{F}_r^1(d, w, \zeta^{5p+5})$ with $d = 100 + 20\left\lceil \log_4\left(\zeta^{2p}\right)\right\rceil \cdot (\lceil \log_2(\max\{q, r\} + 1)\rceil + 1)$

and $w = 2^r \cdot 64 \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r$. By Lemma 39,

$$|\widehat{f}_{\text{wide}}(x) - f(x)| = \left| \sum_{v=1}^{2^r} \widehat{f}_v - \sum_{v=1}^{2^r} w_{\mathcal{P}_{2,v}} \cdot f(x) \right| \leq \sum_{v=1}^{2^r} \left| \widehat{f}_v - w_{\mathcal{P}_{2,v}} \cdot f(x) \right| \lesssim \zeta^{-2p}, \quad (2.31)$$

for $x \in [-a/2, a/2]^r$. Although the result holds for $x \in [-a/2, a/2]^r$, we can always increase a if necessary. Note that $\widehat{f}_{\text{wide}} \in \mathcal{F}_r^1(d, w, \zeta^{5p+5})$ and it may not be bounded by B. However, we can add two more layers at the end in the form of $\sigma(2B - \sigma(B - \widehat{f}_{\text{wide}}\,)) - B$ to make it belong to $\mathcal{F}_r^1(d, w, \zeta^{5p+5}, 2B)$. This will not change the value of the output if $|\widehat{f}_{\text{wide}}(x)| \leq B$, so that Eq. (2.31) still holds given that $|f(x)| \leq B$. □

2.8.2 Proofs of Technical Lemmas

Lemma 27. *Under Assumption 6, given $\{F_t^\star\}_{t=1}^T$, for any constants $\{C_{it}\}_{i=1,\ldots,N,t=1,\ldots,N}$, $(\sum_{t=1}^T \sum_{i=1}^N C_{it}U_{it})/(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2)^{1/2}$ is subGaussian with its subGaussian norm bounded by some fixed constant σ_u^2.*

Proof. Recall that the subGaussian norm of Z_{it} is bounded by σ_z^2. Let $C = \{C_{it}\}_{i=1\ldots,N,t=1,\ldots,T}$, we have

$$\frac{\sum_{t=1}^T \sum_{i=1}^N C_{it}U_{it}}{\left(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2\right)^{1/2}} = \frac{\operatorname{vec}(C)^\top \operatorname{vec}(U)}{\left(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2\right)^{1/2}} = \frac{\operatorname{vec}(C)^\top \Sigma^{1/2} \operatorname{vec}(Z)}{\left(\sum_{t=1}^T \sum_{i=1}^N C_{it}^2\right)^{1/2}}.$$

Since $\operatorname{vec}(Z)$ composed of independent subGaussian random variables with bounded subGaussian norm, the subGaussian norm of the right hand side is bounded by

$$\frac{\sigma_z^2 \operatorname{vec}(C)^\top \Sigma \operatorname{vec}(C)}{\sum_{t=1}^T \sum_{i=1}^N C_{it}^2} \leq \frac{\sigma_z^2 \|\Sigma\| \|\operatorname{vec}(C)\|^2}{\sum_{t=1}^T \sum_{i=1}^N C_{it}^2} = \sigma_z^2 \|\Sigma\|. \quad \square$$

Lemma 28. *Under Assumption 6, with probability at least $1 - C\exp(-cT)$, $\sum_{t=1}^T \|U_t\|_1 \lesssim NT$.*

Proof. Recall that the subGaussian norm of Z_{it} is bounded by σ_z^2. Observe that

$$\begin{aligned}\sum_{t=1}^{T}\|U_t\|_1 \leq & \sum_{t=1}^{T} N^{1/2}\|U_t\| \leq N^{1/2}T^{1/2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}U_{it}^2\right)^{1/2} \\ = & N^{1/2}T^{1/2}\left(\operatorname{vec}(Z)^\top \Sigma \operatorname{vec}(Z)\right)^{1/2} \lesssim N^{1/2}T^{1/2}\left(\operatorname{vec}(Z)^\top \operatorname{vec}(Z)\right)^{1/2}.\end{aligned}$$

Therefore, we only need to prove $N^{-1}T^{-1}\operatorname{vec}(Z)^\top \operatorname{vec}(Z)$ is bounded by some constant. By Hansen-Wright inequality,

$$\mathrm{P}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Z_{it}^2 - \mathbb{E}Z_{it}^2) \geq \sigma_z^2\right) \leq 2\exp(-cNT).$$

By the property of sub-Gaussian variable, we have $\mathbb{E}Z_{it}^2 \leq 2\sigma_z^2$. As a consequence, it holds that $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}^2 \leq 3\sigma_z^2$ with probablity at least $1 - C\exp(-cT)$. □

Lemma 29. *Under Assumptions 6 and* **??**, *if* $\log T = o(N)$, *with probability at least* $1 - C\exp(-cL)$, $\|W^\star U_t\|_\infty \leq 1$ *for all* $1 \leq t \leq T$.

Proof. Write $\Sigma_t \in \mathbb{R}^{T\times NT}$ as the submatrix of $\Sigma^{1/2}$ formed by its $(t-1)N+1$ to tN rows. As a consequence, $U_t = \Sigma_t \operatorname{vec}(Z)$. For $i = 1, \ldots, K$, write $W_i^\star$ as the i-th row of $W^\star$. Then we have $\|W^\star U_t\|_\infty = \max_{i=1,\ldots,K}\|W_i^\star \Sigma_t \operatorname{vec}(Z)\|$. Observe that $W_i^\star \Sigma_t \operatorname{vec}(Z)$ is a subGaussian random variable with its subGaussian norm bounded by $\sigma_z^2 W_i^\star \Sigma_t \Sigma_t^\top (W_i^\star)^\top \lesssim \|W_i^\star\|^2 \|\Sigma_t\Sigma_t^\top\| \lesssim L^{-1}\|\Sigma_t\Sigma_t^\top\|$. Since $\Sigma_t\Sigma_t^\top$ is a submatrix of Σ and $\|\Sigma\| \lesssim 1$, we have $\|\Sigma_t\Sigma_t^\top\| \lesssim 1$. Therefore, the subGaussian norm of $W_i^\star\Sigma_t \operatorname{vec}(Z)$ is approximately less than L^{-1}. By the property of subGaussian variable, $\mathrm{P}(|W_i^\star\Sigma_t \operatorname{vec}(Z)| \geq 1) \leq C\exp(-cL)$. Together with union bound inequality and the fact that $\log(T) = o(L)$,

$$\mathrm{P}(\max_{1\leq t\leq T}\max_{1\leq i\leq K}|W_i^\star\Sigma_t \operatorname{vec}(Z)| \geq 1) \leq CTK\exp(-cL) \lesssim C\exp(-cL/2).$$ □

Lemma 30. *With probability at least* $1 - C\exp(-cT)$, $T^{-1}\sum_{t=1}^{T}\|W^\star U_t\|^2 \lesssim L^{-1}$.

Proof. Let $Q := \mathbb{I}_T \otimes (W^\star)^\top W^\star$. Observe that

$$T^{-1}\sum_{t=1}^{T}\|W^\star U_t\|^2 = T^{-1}\operatorname{vec}(U)^\top Q\operatorname{vec}(U) = T^{-1}\operatorname{vec}(Z)^\top\Sigma^{1/2}Q\Sigma^{1/2}\operatorname{vec}(Z).$$

Note that $\|Q\| \le \|(W^\star)^\top W^\star\| \le \operatorname{Tr}((W^\star)^\top W^\star) \lesssim L^{-1}$. Therefore, we have $\|\Sigma^{1/2}Q\Sigma^{1/2}\| \lesssim L^{-1}$. Additionally, observe that $\operatorname{rank}(\Sigma^{1/2}Q\Sigma^{1/2}) \le \operatorname{rank}(Q) \le T\operatorname{rank}(W^\star) = TK$. As a consequence, we have

$$\begin{aligned}
&T^{-1}\operatorname{Tr}(\Sigma^{1/2}Q\Sigma^{1/2}) \le \|T^{-1}\Sigma^{1/2}Q\Sigma^{1/2}\|\operatorname{rank}(\Sigma^{1/2}Q\Sigma^{1/2}) \lesssim L^{-1},\\
&\|T^{-1}\Sigma^{1/2}Q\Sigma^{1/2}\|_F^2 \lesssim \|T^{-1}\Sigma^{1/2}Q\Sigma^{1/2}\|^2\operatorname{rank}(\Sigma^{1/2}Q\Sigma^{1/2}) \lesssim T^{-1}L^{-2}.
\end{aligned}$$

By applying Hanson-Wright inequality, we conclude that

$$\mathrm{P}\left(|(1-\mathbb{E})T^{-1}\operatorname{vec}(Z)^\top\Sigma^{1/2}Q\Sigma^{1/2}\operatorname{vec}(Z)| \ge L^{-1}\right) \le C\exp(-cT).$$

Together with the fact that $\mathbb{E}T^{-1}\operatorname{vec}(Z)^\top\Sigma^{1/2}Q\Sigma^{1/2}\operatorname{vec}(Z) \asymp T^{-1}\operatorname{Tr}(\Sigma^{1/2}Q\Sigma^{1/2}) \lesssim L^{-1}$, we completes the proof. □

Lemma 31. *For any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$, it holds that*

$$|\varphi(x) - \varphi(y)| \le n_0 T^{(5\beta+5)(d+1)} w^d \|x-y\|_\infty.$$

Proof. Write $\varphi(x) = W_d\sigma_{v_d}W_{d-1}\sigma_{v_{d-1}}\dots W_1\sigma_{v_1}W_0x$ where $W_i \in \mathbb{R}^{n_{i+1}\times n_i}$. It is not hard to verify that $\|W_ix - W_iy\|_\infty \le n_iT^{5\beta+5}\|x-y\|_\infty$ and $\|\sigma_{v_i}x - \sigma_{v_i}y\|_\infty \le \|x-y\|_\infty$. Note that for two Lipschitz functions f_1 and f_2 satisfying $\|f_1(x) - f_1(y)\|_\infty \le L_1\|x-y\|_\infty$ and $\|f_2(x) - f_2(y)\|_\infty \le L_2\|x-y\|_\infty$, it holds that $\|f_2\circ f_1(x) - f_2\circ f_1(y)\|_\infty \le L_1L_2\|x-y\|_\infty$.

By repeatedly applying this fact to $W_d\sigma_{v_d}W_{d-1}\sigma_{v_{d-1}}\dots W_1\sigma_{v_1}W_0x$, we have

$$\|\varphi(x)-\varphi(y)\| \le T^{(5\beta+5)(d+1)}\Big(\prod_{i=0}^{d} n_i\Big)\|x-y\|_\infty \le n_0 T^{(5\beta+5)(d+1)} w^d \|x-y\|_\infty. \qquad \square$$

Lemma 32. *For the function class $\mathcal{F}_{n_0}^{n_{d+1}}(d,w,T^{5\beta+5},B)$ with width vector $n=(n_0,\dots,n_{d+1})$, assume the total number of non-zero weights are bounded by S. There exists a function class $\mathcal{F}_{n_0,\delta}^{n_{d+1}} \subset \mathcal{F}_{n_0}^{n_{d+1}}(d,w,T^{5\beta+5})$ such that*
(a) $|\mathcal{F}_{n_0,\delta}^{n_{d+1}}| \le \Big(8\delta^{-1}CT^{(5\beta+5)(d+1)}(1+w)^d n_0 n_{d+1} d\Big)^{2S}$
(b) For any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d,w,T^{5\beta+5},B)$, there exists $\overline{\varphi} \in \mathcal{F}_{n_0,\delta}^{n_{d+1}}$ satisfying $\|\varphi(x)-\overline{\varphi}(x)\|_\infty \le \delta$ whenever $\|x\|_\infty \le C$.

Proof. The proof follows the same idea as Lemma 5 of Schmidt-Hieber [2020]. For any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d,w,T^{5\beta+5},B)$, write $\varphi(x) = W_d\sigma_{v_d}W_{d-1}\sigma_{v_{d-1}}\cdots W_1\sigma_{v_1}W_0x$. Dfine $A_k^+\varphi : [-C,C]^r \to \mathbb{R}^{n_k}$ and $A_k^-\varphi : \mathbb{R}^{n_{k-1}} \to \mathbb{R}$ as

$$\begin{aligned} A_k^+\varphi(x) &= \sigma_{v_k}W_{k-1}\sigma_{v_{k-1}}\cdots W_1\sigma_{v_1}W_0x, \quad \text{for } k=1,\dots,d,\\ A_k^-\varphi(x) &= W_d\sigma_{v_d}W_{d-1}\cdots W_k\sigma_{v_k}W_{k-1}x, \quad \text{for } k=1,\dots,d+1. \end{aligned}$$

By convention, we set $A_0^+\varphi(x) = A_{d+2}^+\varphi(x) = x$. Given that all the parameters of φ are bounded by $T^{5\beta+5}$ and that $\|x\|_\infty \le C$, it can by proven by mathematical induction that

$$\|A_k^+(x)\|_\infty \le T^{(5\beta+5)(k+1)} C \prod_{\ell=0}^{k-1}(n_\ell+1). \tag{2.32}$$

Additionally, by using Lemma 31, we have

$$|A_k^-\varphi(x) - A_k^-\varphi(y)| \le T^{(5\beta+5)(d-k+2)} \left(\prod_{\ell=k-1}^{d} n_\ell\right)\|x-y\|_\infty. \tag{2.33}$$

Let us fix some $\varepsilon > 0$. Suppose that $\varphi, \overline{\varphi} \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ be two neural networks such that all parameters are at most ε away from each other. Denote the parameters of φ by $(v_k, W_k)_k$ and the parameters of $\overline{\varphi}$ by $(\overline{v}_k, \overline{W}_k)_k$. Note that

$$
\begin{aligned}
\|\varphi(x) - \overline{\varphi}(x)\|_\infty &\leq \sum_{k=1}^{d+1} \left\| A_{k+1}^- \varphi \sigma_{v_k} W_{k-1} A_{k-1}^+ \overline{\varphi}(x) - A_{k+1}^- \varphi \sigma_{\overline{v}_k} \overline{W}_{k-1} A_{k-1}^+ \overline{\varphi}(x) \right\|_\infty \\
&\leq \sum_{k=1}^{d+1} T^{(5\beta+5)(d-k+1)} \left(\prod_{\ell=k}^{d} n_\ell \right) \left\| \sigma_{v_k} W_{k-1} A_{k-1}^+ \overline{\varphi}(x) - \sigma_{\overline{v}_k} \overline{W}_{k-1} A_{k-1}^+ \overline{\varphi}(x) \right\|_\infty \\
&\leq \sum_{k=1}^{d+1} T^{(5\beta+5)(d-k+1)} \left(\prod_{\ell=k}^{d} n_\ell \right) \left(\left\| \left(W_{k-1} - \overline{W}_{k-1} \right) A_{k-1}^+ \overline{\varphi}(x) \right\|_\infty + \|v_k - \overline{v}_k\|_\infty \right) \\
&\leq \varepsilon \sum_{k=1}^{d+1} T^{(5\beta+5)(d-k+1)} \left(\prod_{\ell=k}^{d} n_\ell \right) \left(n_{k-1} \left\| A_{k-1}^+ \overline{\varphi}(x) \right\|_\infty + 1 \right) \\
&\leq 4\varepsilon T^{(5\beta+5)(d+1)} C (1+w)^d n_0 d,
\end{aligned}
$$

where we apply triangle inequality in the first inequality, use Eq. (2.33) in the second inequality, the facts that $\|\sigma_v(x) - \sigma_v(y)\|_\infty \leq \|x - y\|_\infty$ and $\|\sigma_v(x) - \sigma_{\overline{v}}(x)\|_\infty \leq \|v - \overline{v}\|_\infty$ in the third inequality, and Eq. (2.32) in the last inequality. Upon selecting ε as $\frac{\delta}{4T^{(5\beta+5)(d+1)} C (1+w)^d n_0 d}$, it becomes evident that $\|\varphi(x) - \overline{\varphi}(x)\|_\infty \leq \delta$. Consequently, by discretizing the parameters with grid size ε for neural networks φ, we form a function class $\mathcal{F}_{n_0,\delta}^{n_{d+1}}$ satisfying (b). To verify (a), note that the total number of parameters is bounded by $\sum_{\ell=0}^{d} (n_\ell + 1) n_{\ell+1} \leq (d+1) 2^{-d} \prod_{\ell=0}^{d+1} (n_\ell + 1) \leq 4 n_0 n_{d+1} (1+w)^d$ and there are at most $\binom{4 n_0 n_{d+1} (1+w)^d}{S} \leq (4 n_0 n_{d+1} (1+w)^d)^S$ combinations to pick S non-zero parameters. Since all the parameters are bounded in absolute value by $T^{5\beta+5}$, we can discretize the non-zero

parameters with grid size ε and obtain for the covering number

$$\begin{aligned}|\mathcal{F}_{n_0,\delta}^{n_{d+1}}| &\le \sum_{S^\star \le S} (8\delta^{-1}CT^{(5\beta+5)(d+1)}(1+w)^d n_0 d)^{S^\star} \\ &\le (8\delta^{-1}CT^{(5\beta+5)(d+1)}(1+w)^d n_0 n_{d+1} d)^{2S}. \end{aligned}$$

□

Lemma 33. *The right-hand-side of (2.22) is approximately less than*

$$\begin{aligned}&(T^{-\frac{\beta}{2\beta+K}} + T^{-1/2}L^{1/2})\log^2(NT)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big(\widehat{X}_{T+1,i} - \varphi_i^\star(F_{T+1}^\star)\big)^2\right)^{1/2} \\ &+ (T^{-\frac{2\beta}{2\beta+K}} + T^{-1}L)\log^4(NT).\end{aligned}$$

Proof. Observe that $\varphi_i \circ \rho \in \mathcal{F}_N^1(d_1+d_2, \max(w_1,w_2), T^{5\beta+5}, B)$ and its total number of non-zero weights are approximately less than $L + T^{\frac{K}{2\beta+K}}\log T$. Therefore, by Lemma 32, there exists $\mathcal{F}_{N,\delta}^1(d_1+d_2, \max(w_1,w_2), T^{5\beta+5}, B)$ with $\delta = T^{-1}$ satifying $\log|\mathcal{F}_{N,\delta}^1| \lesssim (L+T^{\frac{K}{2\beta+K}})\log^4 NT$. In addition, for any $\varphi_i \circ \rho$ satisfying the conditions in Theorem 7, there exists a function $f(\cdot) \in \mathcal{F}_{N,\delta}^1$ such that $\|\varphi_i \circ \rho(x) - f(x)\| \le T^{-1}$ whenever $\|x\|_\infty \le 2NT$. Without loss of generality, we can assume all the functions in $\mathcal{F}_{N,\delta}^1$ are bounded by B. Write $\mathcal{F}_{N,\delta}^1 = \{f_j, j \in [|\mathcal{F}_{N,\delta}^1|]\}$ and assume $f_{j^\star} \in \mathcal{F}_{N,\delta}^1$ such that

$$\|\widehat{\varphi}_i \circ \widehat{\rho}(x) - f_{j^\star}(x)\|_\infty \le \delta = T^{-1} \tag{2.34}$$

for $\|x\|_\infty \le 2NT$. Note that by Lemma 28 and the fact that $\|X_t^\star\|_\infty, \|\tilde{X}_t^\star\|_\infty \lesssim 1$, it is not hard to check that with probability at least $1 - C\exp(-cT)$, $\max_{1\le t\le T+1}\|X_t\|_\infty, \|\tilde{X}_t\|_\infty \le 2NT$. Therefore, under this event, Eq. (2.34) holds for $X_1, \tilde{X}_1, \ldots, X_{T+1}, \tilde{X}_{T+1}$. Let us

define

$$r_{ij} := \max\left(\sqrt{T^{-1}\log|\mathcal{F}^1_{N,\delta}|}, \mathbb{E}^{1/2}\left[\left(f_j(X_{T+1}) - \varphi_i^\star\left(F^\star_{T+1}\right)\right)^2\right]\right) \text{ and}$$
$$V_i := \max_{j\in[|\mathcal{F}^1_{N,\delta}|]}\left|\frac{1}{r_{ij}B}\sum_{t=1}^{T}\left[\left(f_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \left(f_j(X_t) - \varphi_i^\star(F_t^\star)\right)^2\right]\right|.$$

Additionally, we define $r_i^\star$ as r_{ij} for $j = j^\star$, which is the same as

$$\begin{aligned} r_i^\star &= \max\left(\sqrt{T^{-1}\log|\mathcal{F}^1_{N,\delta}|}, \mathbb{E}^{1/2}\left[\left(f_{j^\star}(X_{T+1}) - \varphi_i^\star\left(F^\star_{T+1}\right)\right)^2 \middle| \{X_t\}_{t=1}^T\right]\right) \\ &\le \sqrt{T^{-1}\log|\mathcal{F}^1_{N,\delta}|} + \mathbb{E}^{1/2}\left[\left(\widehat{\varphi}_i(\widehat{\rho}(X_{T+1}) - \varphi_i^\star\left(F^\star_{T+1}\right)\right)^2 \middle| \{X_t\}_{t=1}^T\right] + T^{-1}, \end{aligned} \tag{2.35}$$

where the last part follows from triangle inequality and Eq. (2.34). Moreover, by Eq. (2.34), with probability at least $1 - C\exp(-cT)$,

$$\begin{aligned} &\left|\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\varphi}_i(\widehat{\rho}(\tilde{X}_t)) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\right)^2\right| \\ \le &\left|\frac{1}{T}\sum_{t=1}^{T}\left(f_{j^\star}(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\left(f_{j^\star}(X_t) - \varphi_i^\star(F_t^\star)\right)^2\right| + 9T^{-1}B. \end{aligned}$$

With the above inequality and the fact that $\exp(-cT) = o(T^{-1})$, we conclude

$$\begin{aligned} &\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\widehat{\varphi}_i(\widehat{\rho}(\tilde{X}_t)) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\widehat{\varphi}_i(\widehat{\rho}(X_t)) - \varphi_i^\star(F_t^\star)\right)^2\right| \\ \le &\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(f_{j^\star}(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(f_{j^\star}(X_t) - \varphi_i^\star(F_t^\star)\right)^2\right| + 10T^{-1}B \\ \le &\frac{B}{T}\mathbb{E}(V_i r_i^\star) + 10T^{-1}B \end{aligned} \tag{2.36}$$

By using Cauchy-Schwarz inequality and Eq. (2.35), we have

$$\begin{aligned}\mathbb{E}(V_i r_i^\star) &\le \left(\mathbb{E}\left(\mathbb{E}\left[\left(\widehat{\varphi}_i \circ \widehat{\rho}(X_{T+1}) - \varphi_i^\star(F_{T+1}^\star)\right)^2 \middle| \{X_t\}_{t=1}^T\right]\right)\right)^{1/2} (\mathbb{E}V_i^2)^{1/2} \\ &\quad + \sqrt{T^{-1}\log|\mathcal{F}_{N,\delta}^1|}\mathbb{E}V_i + T^{-1}\mathbb{E}V_i \qquad (2.37) \\ &= \left(\mathbb{E}\left(\widehat{\varphi}_i \circ \widehat{\rho}(X_{T+1}) - \varphi_i^\star(F_{T+1}^\star)\right)^2\right)^{1/2} (\mathbb{E}V_i^2)^{1/2} + \sqrt{T^{-1}\log|\mathcal{F}_{N,\delta}^1|}\mathbb{E}V_i + T^{-1}\mathbb{E}V_i\end{aligned}$$

Next we analyze the term V_i. For any fixed $j \in [|\mathcal{F}_{N,\delta}^1|]$, define $Y_{tj} := \left(f_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2 - \left(f_j(X_t) - \varphi_i^\star(F_t^\star)\right)^2$. Then it is straightforward to see that $\{Y_{tj}\}_{t=1}^T$ is a series of i.i.d. random variables satisfying $\mathbb{E}Y_{tj} = 0$, $|Y_{tj}| \le 4B^2$, and

$$\mathrm{Var}(Y_{tj}) = 2\,\mathrm{Var}\left(\left(f_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^2\right) \le 2\mathbb{E}\left[\left(f_j(\tilde{X}_t) - \varphi_i^\star(\tilde{F}_t^\star)\right)^4\right] \le 8B^2 r_{ij}^2.$$

With the above results, we apply Bernstein's inequality and union bound inequality to obtain

$$\mathrm{P}(V_i \ge x) = \mathrm{P}\left(\max_{j\in[|\mathcal{F}_{N,\delta}^1|]} \left|\frac{1}{r_{ij}B}\sum_{t=1}^T Y_{tj}\right| \ge x\right) \le 2|\mathcal{F}_{N,\delta}^1| \max_{j\in[|\mathcal{F}_{N,\delta}^1|]} \exp\left(-\frac{x^2}{8x/(3r_{ij}) + 16T}\right).$$

Since $r_{ij} \ge \sqrt{T^{-1}\log|\mathcal{F}_{N,\delta}^1|}$, it holds that $\mathrm{P}(V_i \ge x) \le 2\exp(-3x\sqrt{\log|\mathcal{F}_{N,\delta}^1|}/16\sqrt{T})$ whenever $x \ge 6\sqrt{T\log|\mathcal{F}_{N,\delta}^1|}$ with simple algebra. Therefore, when T is large enough,

$$\begin{aligned}\mathbb{E}V_i &= \int_0^\infty \mathrm{P}(V_i \ge x)dx \le 6\sqrt{T\log|\mathcal{F}_{N,\delta}^1|} \\ &\quad + \int_{6\sqrt{T\log|\mathcal{F}_{N,\delta}^1|}}^\infty 2|\mathcal{F}_{N,\delta}^1| \exp\left(\frac{-3x\sqrt{\log|\mathcal{F}_{N,\delta}^1|}}{16\sqrt{T}}\right) dx \\ &\le 6\sqrt{T\log|\mathcal{F}_{N,\delta}^1|} + \frac{32}{3}\sqrt{\frac{T}{\log|\mathcal{F}_{N,\delta}^1|}} \le 7\sqrt{T\log|\mathcal{F}_{N,\delta}^1|}. \qquad (2.38)\end{aligned}$$

In a similar way, it can be proven $\mathbb{E}V_i^2 \le 40T\log|\mathcal{F}_{N,\delta}^1|$. Combining Eq. (2.37) and bounds

for the moment of V_i, $\frac{B}{T}\mathbb{E}(V_i r_{ij^*})$ is asymptotically bounded by

$$\frac{B}{T}\left(\mathbb{E}^{1/2}\left[\left(\widehat{X}_{T+1,i}-\varphi_i^\star(F_{T+1}^\star)\right)^2\right]\sqrt{T\log|\mathcal{F}_{N,\delta}^1|}+\sqrt{T^{-1}\log|\mathcal{F}_{N,\delta}^1|}+\log|\mathcal{F}_{N,\delta}^1|\right).$$

Recall we have $\log|\mathcal{F}_{N,\delta}^1|\lesssim(T^{\frac{K}{2\beta+K}}+L)\log^4(NT)$, which implies that $\frac{B}{T}\mathbb{E}(V_i r_{ij^*})$ can be asymptotically bounded by

$$(T^{-\frac{2\beta}{2\beta+K}}+T^{-1}L)\log^4(NT)+(T^{-\frac{\beta}{2\beta+K}}+T^{-1/2}L^{1/2})\log^2(NT)\mathbb{E}^{1/2}\left[\left(\widehat{X}_{T+1,i}-\varphi_i^\star(F_{T+1}^\star)\right)^2\right].$$

Together with Eq. (2.36), we complete the proof. □

Lemma 34. *There exists* $\widehat{f}_{mult,r}\in\mathcal{F}_r^1(20R\lceil\log_2(r)\rceil,18r,4^{2r}a^{2r})$ *such that*

$$|\widehat{f}_{mult,r}(x)-\prod_{i=1}^r x^{(i)}|\le 4^{4r+1}\cdot a^{4r}\cdot r\cdot 4^{-R},\quad\forall x\in[-a,a]^r,$$

for any $a\ge 1$ *and any* $R\ge\log_4(2\cdot 4^{2r}\cdot a^{2r})$.

Proof. We initially demonstrate the existence of a neural network $\widehat{f}_{mult}\in\mathcal{F}_2^1(10R,18,4a^2)$ such that

$$|\widehat{f}_{mult}(x,y)-xy|\le 2a^2 4^{-R},\quad\text{for all } x,y\in[-a,a].\tag{2.39}$$

According to Lemma A.2 of Schmidt-Hieber [2020], a neural network $\widehat{\phi}\in\mathcal{F}_2^1(2R+5,6,1)$ exists, satisfying $\left|\widehat{\phi}(x,y)-xy\right|\le 2^{-2R-1}$, for all $x,y\in[0,1]$. By defining $\widehat{f}_{mult}(x,y)=4a^2\widehat{\phi}\left(\frac{x+a}{2a},\frac{y+a}{2a}\right)-a(x+y)-a^2$, we have $\widehat{f}_{mult}\in\mathcal{F}_2^1(10R,18,4a^2)$. Observe that for all

$x, y \in [-a, a]$, we have $(x+a)/2a, (y+a)/2a \in [0,1]$. Therefore,

$$\begin{aligned}|\widehat{f}_{mult}(x,y) - xy| &= \left|4a^2\widehat{\phi}\left(\frac{x+a}{2a}, \frac{y+a}{2a}\right) - a(x+y) - a^2 - xy\right| \\ &= 4a^2\left|\widehat{\phi}\left(\frac{x+a}{2a}, \frac{y+a}{2a}\right) - \frac{(x+a)(y+a)}{4a^2}\right| \leq 4a^2 2^{-2R-1},\end{aligned}$$

which implies Eq. (2.39). To construct the desired neural network, we use $\widehat{f}_{mult}$ defined in Eq. (2.39) with $a = 4^r a^r$. Therefore, $\widehat{f}_{mult}$ belongs to $\mathcal{F}_2^1(20R, 18, 4^{2r}a^{2r})$ and satisfies

$$|\widehat{f}_{mult}(x,y) - xy| \leq 2 \cdot 4^{2r} \cdot a^{2r} \cdot 4^{-R}, \tag{2.40}$$

for $x_1, x_2 \in [-4^r a^r, 4^r a^r]$ For $r > 2$, let $q := \lceil \log_2(r) \rceil$ and set

$$(z_1, \ldots, z_{2^q}) = (x^{(1)}, x^{(2)}, \ldots, x^{(r)}, 1, \ldots, 1).$$

We are now prepared to construct $\widehat{f}_{mult,r}$. In the first $20R$ layers we compute

$$(\widehat{f}_{mult}(z_1, z_2), \widehat{f}_{mult}(z_3, z_4), \ldots, \widehat{f}_{mult}(z_{2^q-1}, z_{2^q})) \in \mathbb{R}^{2^{q-1}}.$$

Now we pair neighboring entries and apply $\widehat{f}_{mult}$. The procedure continues until only one entry remains. The resulting network $\widehat{f}_{mult,r}$ belongs to $\mathcal{F}_r^1(20qR, 18r, 4^{2r}a^{2r})$. By Eq. (2.40) and $R \geq \log_4(2 \cdot 4^{2r} \cdot a^{2r})$, for any $1 \leq l \leq r$ and $z_1, z_2 \in [-(4^l - 1)a^l, (4^l - 1)a^l]$, we have

$$\left|\widehat{f}_{\text{mult}}\ (z_1, z_2)\right| \leq |z_1 \cdot z_2| + \left|\widehat{f}_{\text{mult}}\ (z_1, z_2) - z_1 \cdot z_2\right| \leq \left(4^l - 1\right)^2 a^{2l} + 1 \leq \left(4^{2l} - 1\right) \cdot a^{2l}.$$

From this we get successively that all outputs of the procedure $l \in \{1, \ldots, q-1\}$ lie in the interval $\left[-\left(4^{2^l} - 1\right) \cdot a^{2^l}, \left(4^{2^l} - 1\right) \cdot a^{2^l}\right]$, hence they are contained in the interval $[-4^r a^r, 4^r a^r]$, where Eq. (2.40) holds. Write $\widehat{z_1 z_2} = \widehat{f}_{mult}(z_1, z_2)$, $\widehat{z_3 z_4} = \widehat{f}_{mult}(z_3, z_4)$,

and $\widehat{z_1\widehat{z_2z_3}z_4} = \widehat{f}_{mult}(\widehat{z_1z_2}, \widehat{z_3z_4})$. Observe that $|\widehat{z_1\widehat{z_2z_3}z_4} - z_1z_2z_3z_4|$ is upper bounded by

$$\begin{aligned}
&|z_1z_2\widehat{z_3z_4} - z_1z_2z_3z_4| + |\widehat{z_1z_2}\widehat{z_3z_4} - z_1z_2\widehat{z_3z_4}| + |\widehat{z_1\widehat{z_2z_3}z_4} - \widehat{z_1z_2}\widehat{z_3z_4}| \\
&\leq 2 \cdot 4^{2r} \cdot a^{2r} \cdot 4^{-R} + 2 \cdot 4^2 \cdot a^2 \cdot (2 \cdot 4^{2r} \cdot a^{2r} \cdot 4^{-R}) = 2 \cdot 4^{2r} \cdot a^{2r} \cdot 4^{-R} \cdot (1 + 2 \cdot 4^2 \cdot a^2)
\end{aligned}$$

where in the second inequality we use the fact that all outputs of the first procedure lie in the interval $[-4^2 \cdot a^2, 4^2 \cdot a^2]$. By recursively apply this inequality, we obtain

$$\begin{aligned}
|\widehat{f}_{mult,r}(x) - \prod_{i=1}^{d} x_i| &\leq 2 \cdot 4^{2r} \cdot a^{2r} \cdot 4^{-R} \cdot 4^{1+2+\cdots+2^{q-1}} \cdot a^{1+2+\cdots+2^{q-1}} \cdot \left(1 + 2 + \cdots + 2^{q-1}\right) \\
&\leq 4^{2r} \cdot a^{2r} \cdot 4^{-R} \cdot 4^{2r+1} \cdot a^{2r} \cdot d = 4 \cdot 4^{4r} \cdot a^{4r} \cdot r \cdot 4^{-R}. \qquad \square
\end{aligned}$$

Lemma 35. *Let $m_1, \ldots, m_{\binom{r+n}{r}}$ denote all monomials in $\mathcal{P}_n^r$ for some $n \in \mathbb{N}_+$. Let $r_1, \ldots, r_{\binom{r+n}{r}} \in \mathbb{R}$, define*

$$p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right) = \sum_{i=1}^{\binom{r+n}{r}} r_i \cdot y_i \cdot m_i(x), \quad x \in [-a, a]^r, y_1, \ldots, y_{\binom{r+n}{r}} \in [-a, a]$$

and set $\overline{r}(p) = \max_{i \in \left\{1, \ldots, \binom{r+n}{r}\right\}} |r_i|$. Then for any $a \geq 1$ and $R \geq \log_4\left(2 \cdot 4^{2\cdot(n+1)} \cdot a^{2\cdot(n+1)}\right)$, a neural network $\widehat{f}_p \in \mathcal{F}^1_{\binom{r+n}{r}+r}(d, w, 4^{2n+2}a^{2n+2} \vee \overline{r}_p)$ with $d = 20R \cdot \lceil \log_2(n+1) \rceil$ and $w = 18 \cdot (n+1) \cdot \binom{r+n}{r}$ exists, such that

$$\left|\widehat{f}_p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right) - p\left(x, y_1, \ldots, y_{\binom{r+n}{r}}\right)\right| \leq C \cdot \overline{r}(p) \cdot a^{4(n+1)} \cdot 4^{-R}$$

for $x \in [-a, a]^r, y_1, \ldots, y_{\binom{r+n}{r}} \in [-a, a]$, where C is a constant only depends on d and n.

Proof. By Lemma 34, we know there exists a neural network $\widehat{f}_{m_i} : \mathbb{R}^{r+1} \to \mathbb{R}$ which belongs

to $\mathcal{F}^1_{r+1}\left(20R\cdot\lceil\log_2(n+1)\rceil, 18\cdot(n+1), 4^{2n+2}a^{2n+2}\right)$ such that

$$|\widehat{f}_{m_i}(x,y_i)-y_i m_i(x)|\leq 4\cdot 4^{4(n+1)}\cdot a^{4(n+1)}\cdot(n+1)\cdot 4^{-R}.$$

Define $\widehat{f}_p:=\sum_{i=1}^{\binom{r+n}{r}} r_i\cdot\widehat{f}_{m_i}(x,y_i)$, it is not hard to verify that it belongs to

$$\mathcal{F}^1_{\binom{r+n}{r}+r}(d,w,4^{2n+2}a^{2n+2}\vee\overline{r}_p)$$

and satisfies

$$\begin{aligned}\left|\widehat{f}_p\left(x,y_1,\ldots,y_{\binom{r+n}{r}}\right)-p\left(x,y_1,\ldots,y_{\binom{r+n}{r}}\right)\right| &\leq \sum_{i=1}^{\binom{r+n}{r}}|r_i|\cdot\left|y_i\cdot m_i(x)-\widehat{f}_{m_i}(x,y_i)\right| \\ &\leq \binom{r+n}{r}\cdot\overline{r}(p)\cdot 4\cdot 4^{4\cdot(n+1)}\cdot a^{4\cdot(n+1)}\cdot(n+1)\cdot 4^{-R}.\end{aligned}$$

□

Lemma 36. *The neural network $\widehat{f}_{\mathcal{P}_2}$ defined in Eq. (2.28) belongs to $\mathcal{F}^1_r(d,w,\zeta^{5p+5})$ with*

$$d=80+20B_{\zeta,p}\cdot\lceil\log_2(\max\{q+1,2\})\rceil$$

and

$$w=\max\left(\binom{r+q}{r}\cdot\zeta^r\cdot 2\cdot(2+2r)+2r, 18\cdot(q+1)\cdot\binom{r+q}{r}\right).$$

Additionally, as long as ζ is large enough (only depends on fixed constants), it holds that

$$\left|\widehat{f}_{\mathcal{P}_2}(x)-f(x)\right|\leq C\cdot\left(\max\left\{2a,\|f\|_{C^q}\right\}\right)^{4(q+1)}\cdot\zeta^{-2p}$$

for all $x\in\bigcup_{j\in\{1,\ldots,\zeta^{2r}\}}\left(C_{j,2}\right)^0_{1/\zeta^{2p+2}}$. The neural network value is bounded by $\left|\widehat{f}_{\mathcal{P}_2}(x)\right|\leq 2\cdot e^{2ad}\cdot\max\left\{\|f\|_{C^q},1\right\}$ for all $x\in[-a,a)^r$.

Proof. The proof for this lemma follows the same argument as Lemma 3 in Kohler and Langer [2021] and is thus omitted. The neural networks we use to construct $\widehat{f}_{\mathcal{P}_2}$ have the

same approximation error as theirs. The only distinction between our neural network and theirs is twofold: our neural network has a layer count at most twenty times theirs, and the weights are bounded by ζ^{5p+5}. □

Lemma 37. *Let $1 \leq a < \infty$ and $\zeta \geq 4^{4r+1}r$. For $w_{\mathcal{P}_2}$ defined in Eq. (2.29), there exists a neural network $\widehat{f}_{w_{\mathcal{P}_2}} \in \mathcal{F}_r^1\left(d, w, \zeta^{5p+5}\right)$, with $d = 100 + 20\left\lceil \log_4\left(\zeta^{2p}\right)\right\rceil \cdot \lceil \log_2(r)\rceil$ and $w = \max\{18r, 2r + r \cdot \zeta^r \cdot 2 \cdot (2+2r)\}$, such that $\left|\widehat{f}_{w_{\mathcal{P}_2}}(x) - w_{\mathcal{P}_2}(x)\right| \leq 4^{4r+1} \cdot r \cdot \zeta^{-2p}$ for $x \in \bigcup_{i\in\{1,\ldots,\zeta^{2r}\}} \left(C_{i,2}\right)^0_{1/\zeta^{2p+2}}$ and $|\widehat{f}_{w_{\mathcal{P}_2}}(x)| \leq 2$ for $x \in [-a,a)^r$.*

Proof. We apply $\widehat{\phi}_{1,2}$ and $\widehat{\phi}_{2,2}$ defined in Eq. (2.27). Let

$$\begin{aligned}\widehat{f}_{w_{\mathcal{P}_2},j}(x) =& \sigma\left(\frac{\zeta^2}{a} \cdot \left(\widehat{\phi}^{(j)}_{1,2} - \widehat{\phi}^{(j)}_{2,2}\right)\right) - 2 \cdot \sigma\left(\frac{\zeta^2}{a} \cdot \left(\widehat{\phi}^{(j)}_{1,2} - \widehat{\phi}^{(j)}_{2,2} - \frac{a}{\zeta^2}\right)\right) \\ &+ \sigma\left(\frac{\zeta^2}{a} \cdot \left(\widehat{\phi}^{(j)}_{1,2} - \widehat{\phi}^{(j)}_{2,2} - \frac{2\cdot a}{\zeta^2}\right)\right).\end{aligned}$$

Given that the weights of $\widehat{\phi}_{1,2}$ and $\widehat{\phi}_{2,2}$ are bounded by ζ^{5p+5}, it is straightforward to see that the weights of $\widehat{f}_{w_{\mathcal{P}_2},j}(x)$ are also bounded by ζ^{5p+5}. We then construct the required function $\widehat{f}_{w_{\mathcal{P}_2}}$ by $\widehat{f}_{w_{\mathcal{P}_2}}(x) = \widehat{f}_{mult,d}\left(\widehat{f}_{w_{\mathcal{P}_2},1}(x), \ldots, \widehat{f}_{w_{\mathcal{P}_2},d}(x)\right)$ where we set $a = 1$ and $R = 20\lceil \log_4(\zeta^{2p})\rceil$ in Lemma 34. Given that $\widehat{f}_{w_{\mathcal{P}_2},j}$ used here is totally the same as the one in Lemma 9 of Kohler and Langer [2021], and that $\widehat{f}_{mult,d}$ we use achieve the same approximation error as theirs. The remaining proof follows by totally the same argument as Lemma 9 of Kohler and Langer [2021] and is thus omitted. □

Lemma 38. *Let $1 \leq a < \infty$. There exists a neural network*

$$\widehat{f}_{check,\mathcal{P}_2} \in \mathcal{F}_r^1\left(100, 2r + \left(4r^2 + 4r\right) \cdot \zeta^r, \zeta^{5p+5}\right)$$

satisfying

$$\widehat{f}_{check,\mathcal{P}_2}(x) = 1_{\bigcup_{i\in\{1,\ldots,\zeta^{2r}\}} C_{i,2}\setminus\left(C_{i,2}\right)^0_{1/\zeta^{2p+2}}}(x)$$

for $x \notin \bigcup_{i\in\{1,\ldots,\zeta^{2r}\}} \left(C_{i,2}\right)^0_{1/\zeta^{2p+2}} \setminus \left(C_{i,2}\right)^0_{2/\zeta^{2p+2}}$ *and* $\widehat{f}_{check,\mathcal{P}_2}(x) \in [0,1]$ *for* $x \in [-a,a)^r$.

Proof. We adopt the definition of $\widehat{f}_{\text{check},\mathcal{P}_2}$ in Lemma 10 of Kohler and Langer [2021]:

$$\begin{aligned}
&\widehat{f}_{\text{check},\mathcal{P}_2}(x) = 1 - \sigma\left(1 - \widehat{f}_2(x) - \widehat{f}^2_{id}\left(\widehat{f}_1(x)\right)\right), \text{ where}\\
&\widehat{f}_1(x) := 1 - \sum_{k=1}^{\zeta^r} \widehat{f}_{ind,\left(C_{k,1}\right)^0_{1/\zeta^{2p+2}}}(x) \text{ and} \qquad (2.41)\\
&\widehat{f}_2(x) := 1 - \sum_{j=1}^{\zeta^r} \widehat{f}_{\text{test}}\left(\widehat{f}^2_{id}(x), \widehat{\phi}_{2,1} + v_j + \frac{1}{\zeta^{2p+2}}\cdot 1, \widehat{\phi}_{2,1} + v_j + \frac{2a}{\zeta^2}\cdot 1 - \frac{1}{\zeta^{2p+2}}\cdot 1, 1\right).
\end{aligned}$$

Here $\widehat{f}_{ind,[a,b)}$ and $\widehat{f}_{test}$ are defined in Eq. (2.26) and the function $\widehat{\phi}_{2,1}$ is defined in Eq. (2.27). Since these functions are also totally the same as theirs, to prove this lemma, we only need to verify that the weights of $\widehat{f}_{\text{check}}$ are contained in the interval $[-\zeta^{5p+5}, \zeta^{5p+5}]$. Given that the weights of $\widehat{f}_{ind,[a,b)}$, $\widehat{f}_{\text{test}}$, and $\widehat{\phi}_{2,1}$ are all contained in the interval $[-\zeta^{5p+5}, \zeta^{5p+5}]$, based on the definition, it is evident that the weights of $\widehat{f}_1$ and $\widehat{f}_2$ are also contained in the interval $[-\zeta^{5p+5}, \zeta^{5p+5}]$, so is $\widehat{f}_{\text{check},\mathcal{P}_2}$. □

Lemma 39. *The function* $\widehat{f}$ *defined in Eq. (2.30) belongs to* $\mathcal{F}^1_r(d, w, \zeta^{5p+5})$ *with* $d = 100 + 20\left\lceil \log_4\left(\zeta^{2p}\right)\right\rceil \cdot \left(\left\lceil \log_2\left(\max\{q,r\}+1\right)\right\rceil + 1\right)$, *and* $w = 64 \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r$. *Additionally, as long as* ζ *is large enough (only depends on fixed constants), it holds that*

$$\left|\widehat{f}(x) - w_{\mathcal{P}_2}(x)\cdot f(x)\right| \leq C \cdot \left(\max\left\{2a, \|f\|_{C^q}\right\}\right)^{4(q+1)} \cdot \zeta^{-2p}$$

for some fixed constant C *and* $x \in [-a,a)^r$.

Proof. The proof for this lemma follows the same argument as Lemma 7 in Kohler and Langer [2021] and is thus omitted. The neural networks we use to construct $\widehat{f}$ have the same approximation error as theirs but with slightly more layers and bounded weights. □

CHAPTER 3

RECURRENT NEURAL NETWORKS MEET TIME SERIES: A THEORETICAL PERSPECTIVE

3.1 Introduction

Recurrent Neural Networks (RNNs), whose fundamental idea can be traced back to Hopfield [1982], with the theoretical foundation established in Rumelhart et al. [1986] and the popularized simple RNNs developed in Elman [1990], represent a class of artificial neural networks specifically designed for processing sequential data. The core concept underlying RNNs is the ability to retain and utilize information from previous inputs. This is realized through the use of a recurrent unit, which updates a hidden layer at each time step based on the current input and the preceding hidden state.

This architecture allows RNNs to capture dependencies across different points in a sequence, making them highly effective in modeling time series data. Time series models are widely employed in disciplines such as statistics, economics, finance, and meteorology, as discussed by Shumway et al. [2000], Tsay [2005], Fan and Yao [2008], and Box et al. [2015]. Although RNNs have gained considerable popularity in various machine learning domains, the statistical theoretical foundations supporting their application to time series data remain largely unexplored.

In this paper, we investigate the statistical properties of RNNs within the framework of time series models. Specifically, we address several critical questions aimed at enhancing the theoretical understanding of RNNs. These questions include whether RNNs can accurately forecast future values of a time series and, if so, the corresponding statistical error bounds. Furthermore, we examine the influence of model parameters, such as the number of neurons in the recurrent layer and the width of other layers, on the determination of the error bounds. We also discuss how to select these parameters to achieve an optimal convergence

rate under various model configurations. Addressing these questions from a theoretical perspective provides new insights and opens avenues for novel applications of RNNs across diverse domains, as previously mentioned.

Our contributions are as follows. First, to the best of our knowledge, this paper is the first to establish statistical error bounds for RNNs applied to nonlinear time series models. In particular, we are the first to thoroughly investigate the theoretical properties of deep and nonlinear RNNs, as well as the first to nonparametrically estimate the nonlinear autoregressive and moving-average model with exogenous variables (NARMAX) model. In deriving the statistical bound, we provide upper bounds for both the approximation error and the estimation error of RNNs. The approximation error is determined by the smoothness of the target function and the dimensionality of the input. Compared with existing results on the approximation error of Deep Neural Networks (DNNs), our contribution is novel in analyzing the approximation error of the nonlinear moving average components, which is captured through the recurrent layer. The estimation error, or generalization error, is governed by the architecture of RNNs, including the width and depth of the hidden layers. Most existing research is limited to single-layer RNNs, whereas we extend our analysis to RNNs with an arbitrary number of hidden layers. Second, based on the error bounds we derive, we examine several model configurations and discuss how the width of the recurrent layer and other hidden layers should be chosen to achieve the optimal convergence rate. Specifically, in cases where the moving average component is linear, the error bound can approach the minimax optimal rate for independent data, depending on the mixing properties of the model.

Towards the theoretical understanding of RNNs, various studies have explored different aspects. For example, Barabanov and Prokhorov [2002] examined the global Lyapunov stability of RNNs, Chen et al. [2018] investigated the generalization bounds for single-layer RNNs, Allen-Zhu et al. [2019] analyzed the convergence speed of RNN training, and Erichson

et al. [2020] modeled RNNs as continuous-time dynamical systems to study the evolution of hidden states. Li et al. [2022] focused on the approximation properties of linear RNNs. Schäfer and Zimmermann [2006], hoon Song et al. [2023] studied the universal approximation for discrete-time setting and Funahashi and Nakamura [1993], Maass et al. [2007], Li et al. [2022] for continuous-time setting. Additionally, Jiao et al. [2024] established a relationship between the approximation errors of RNNs and DNNs and applied it to analyze a nonlinear AR model.

Our paper also contributes to the growing body of literature on the theoretical properties of neural networks. Barron [1993] and Chen and White [1999] studied single-layer neural networks with smooth activation functions. Yarotsky [2017] explored the approximation errors of DNNs with the ReLU activation function. Schmidt-Hieber [2020] and Farrell et al. [2021] derived optimal error rates for sparse DNNs, and Kohler and Langer [2021] demonstrated that fully-connected DNNs can achieve these optimal rates. Furthermore, Nakada and Imaizumi [2020] and Jiao et al. [2023] showed the potential of DNNs to overcome the curse of dimensionality when the input data lie on a manifold.

Despite the limited theoretical understanding of RNN's properties, RNNs have become increasingly popular in a variety of fields. In computer science, they are frequently used in tasks such as natural language processing [Mikolov et al., 2010, Sarzynska-Wawer et al., 2021], speech recognition [Graves et al., 2013, Hannun et al., 2014, Bahdanau et al., 2016], machine translation [Sutskever et al., 2014, Bahdanau et al., 2014, Cho et al., 2014] and image captioning [Vinyals et al., 2015, Xu et al., 2015, Karpathy and Fei-Fei, 2014]. In economics and finance, RNNs have been adopted for foreign exchange rate forecasting [Kuan and Liu, 1995], inflation prediction [Almosova and Andresen, 2023], asset pricing [Fischer and Krauss, 2018, Chen et al., 2024], and realized volatility forecasting [Bucci, 2020]. Moreover, climate science [Lipton, 2015], biology [Tang et al., 2019], healthcare [Miotto et al., 2018], robotics [Soori et al., 2023] and generative AI [Briot et al., 2017] also benefit from RNNs. The flexibil-

ity and powerful sequence learning capabilities of RNNs make them highly applicable across these diverse domains, providing state-of-the-art solutions for complex temporal prediction problems.

In recent years, several variants of RNNs have emerged due to their wide applicability. The most notable examples are long short-term memory (LSTM) networks [Hochreiter and Schmidhuber, 1997] and gated recurrent units (GRUs) [Cho et al., 2014]. Other important variants include convolutional recurrent neural networks (CRNNs) [Choi et al., 2017, Tan and Wang, 2018], bidirectional recurrent neural networks (BRNNs) [Schuster and Paliwal, 1997], NARX neural networks [Siegelmann et al., 1997], neural Turing machines (NTMs) [Graves et al., 2014], structured state space sequence models (S4) [Gu et al., 2021a], Mamba [Gu and Dao, 2023], minimal gated units (MGUs) [Zhou et al., 2016], and GraphRNN [You et al., 2018], among others.

This chapter is organized as follows: Section 3.2 introduces the architecture of RNNs and presents the data generating process models. The main theoretical results are also given in this section. Section 3.3 conducts simulations to validate our theoretical predictions and compare RNNs with some classical models. Section 3.5 gives the proof of our main theoretical results and supporting lemmas are provided in the supplementary materials.

Notation: We denote the set of positive integers by $\mathbb{N}_+$. Furthermore, we represent the set that includes both zero and all positive integers, $0 \cup \mathbb{N}_+$, by $\mathbb{N}_0$. For $n \in \mathbb{N}_+$, we use $[n]$ to denote the set $\{1, \ldots, n\}$. for any $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer smaller than or equal to x and $\lceil x \rceil$ is the smallest integer greater than or equal to x. For any $x, y \in \mathbb{R}$, $x \vee y$ represents the maximum value between x and y. For a vector $x = (x_1, \ldots, x_d)^\top$, we define the p-norm as $\|x\|_p := (\sum_{i=1}^d |x_i|^p)^{\frac{1}{p}}$ and the infinity norm as $\|x\|_\infty = \max_i |x_i|$. For a matrix $A \in \mathbb{R}^{n \times m}$, we use $\|A\|$ and $\|A\|_\infty$ to represent its spectral norm and the maximum absolute value of its entries respectively. Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, we use $\|f\|_\infty$ to denote its maximum absolute value on its domain and $\|f\|_{\infty,[-a,a]^n}$ when the domain is $[-a, a]^n$. We say $f \in \mathcal{C}^\beta$

if it has β continuous derivatives on its domain. Furthermore, we define the norm $\|\cdot\|_{\mathcal{C}^\beta}$ of the smooth function space $\mathcal{C}^\beta$ by $\|f\|_{\mathcal{C}^\beta} := \max\left\{\|D^\alpha f\|_\infty : \|\alpha\|_1 \leq \beta, \alpha \in \mathbb{N}_0^d\right\}$. For a set $\mathcal{F}$, we use $|\mathcal{F}|$ to indicate the number of elements within the set. For two vectors $x, y \in \mathbb{R}^n$, we write $x \leq y$ when $x_i \leq y_i$ for $1 \leq i \leq n$. We use the notation $x_n \lesssim y_n$ when there exists a constant C such that $x_n \leq C y_n$ holds for sufficiently large n. If $x_n \lesssim y_n$ and $y_n \lesssim x_n$, we write $x_n \asymp y_n$ for short.

3.2 Model and Main Results

3.2.1 Model

Let $\{Y_t\}_{t=1}^T$ represent an observed sequence from a stationary time series. To account for potential nonlinearity in such processes, a common approach is to use a nonlinear autoregressive (NAR) model, expressed as:

$$Y_t = \psi^\star(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}) + \varepsilon_t, \tag{3.1}$$

where ε_t denotes the noise term, and p is a known parameter. When p is unknown, various statistical techniques can be applied to estimate it (see the review by Härdle et al. 1997). In this study, we employ a nonparametric framework, assuming that the function $\psi^\star$ is unknown.

The model in Eq. (3.1) can be viewed as a nonparametric regression problem. Numerous nonparametric techniques are available for such regression problems when the observations are independent. Hart [1996] shows that these methods can also be extended to time series with "short memory", where the dependence between future and past values weakens over time. Various mixing conditions are used to quantify this dependence. In this paper, we focus on the α-mixing condition.

Definition 5. *Let $\{Y_t\}_{t=-\infty}^{\infty}$ be a stationary random process on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. For any $i, j \in \mathbb{Z} \cup \{-\infty, \infty\}$ with $i \leq j$, let $\mathfrak{F}_i^j$ denote the σ-algebra generated by $\{Y_k, i \leq k \leq j\}$. The process $\{Y_t\}_{t=-\infty}^{\infty}$ is said to be α-mixing if*

$$\alpha(j) := \sup_{A \in \mathfrak{F}_{-\infty}^{0},\ B \in \mathfrak{F}_j^{\infty}} |\mathrm{P}(A \cap B) - \mathrm{P}(A)P(B)| \to 0 \text{ as } j \to \infty.$$

Here, $\alpha(j)$ is called the α-mixing coefficient.

Under the α-mixing condition, nonparametric regression techniques can use $(Y_{t-1}, \ldots, Y_{t-p})$ as inputs and Y_t as the target to estimate the model. Common approaches include kernel regression [Andrews, 1991], local polynomial regression [Fan, 2018], sieve estimation [Chen and Shen, 1998b], random forests [Biau, 2012]. In particular, [Men et al., 2024] applies DNNs to estimate NAR models and provide statistical error bound. While these methods effectively capture the dependence between $(Y_{t-1}, \ldots, Y_{t-p})$ and Y_t, they fail to account for potential dependence within the noise term. To address this, we propose the following NARMAX(p, q) model,

$$Y_t = \phi^{\star}(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, \varepsilon_{t-1}, \varepsilon_{t-2} \ldots, \varepsilon_{t-q}, X_{t-1}) + \varepsilon_t, \tag{3.2}$$

where $\phi^{\star}$ is an unknown function, X_{t-1} is a k-dimensional exogenous input vector, ε_t denotes the noise term, and $\{X_t\}$ is independent with $\{\varepsilon_t\}$.

To estimate $\phi^{\star}$, we decompose it into two parts, one only consists of the observable terms and the other also includes the unobservable noise term. Specifically, we define

$$\begin{aligned}
&\psi^{\star}(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, X_{t-1}) \\
&:= \mathbb{E}(\phi^{\star}(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, \varepsilon_{t-1}, \varepsilon_{t-2} \ldots, \varepsilon_{t-q}, X_{t-1}) | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, X_{t-1}),
\end{aligned}$$

and define

$$\eta^\star(Y_{t-1},\ldots,Y_{t-p},\varepsilon_{t-1},\varepsilon_{t-2}\ldots,\varepsilon_{t-q},X_{t-1})$$
$$:=\phi^\star(Y_{t-1},Y_{t-2},\ldots,Y_{t-p},\varepsilon_{t-1},\varepsilon_{t-2}\ldots,\varepsilon_{t-q},X_{t-1})-\psi^\star(Y_{t-1},Y_{t-2},\ldots,Y_{t-p},X_{t-1}).$$

As a consequence, we can write the model into the form of:

$$Y_t=\psi^\star(Y_{t-1},Y_{t-2},\ldots,Y_{t-p},X_{t-1})+\eta^\star(Y_{t-1},\ldots,Y_{t-p},\varepsilon_{t-1},\varepsilon_{t-2}\ldots,\varepsilon_{t-q},X_{t-1})+\varepsilon_t.$$

We note that $\psi^\star$ can be approximated via any nonparametric regression techniques like DNNs while $\eta^\star$ cannot. To the best of our knowledge, no statistical methods have yet been developed to consistently estimate $\eta^\star$.

For this process, we define

$$Y_t^\star:=\psi^\star(Y_{t-1},Y_{t-2},\ldots,Y_{t-p},X_{t-1})+\eta^\star(Y_{t-1},\ldots,Y_{t-p},\varepsilon_{t-1},\varepsilon_{t-2}\ldots,\varepsilon_{t-q},X_{t-1}).$$

When ε_t is independent to each other, $Y_t^\star$ represents the best predictor given historical information, i.e.,

$$Y_t^\star:=\mathbb{E}(Y_t|\mathfrak{F}_{-\infty}^{t-1}),\tag{3.3}$$

where we define $\mathfrak{F}_i^j$ as the σ-algebra generated by $\{(Y_k,\varepsilon_k,X_k),i\le k\le j\}$ here and throughout the rest of paper.

Next, we make assumptions regarding the bounds on various norms of U_t, Y_t, and X_t, as well as their mixing conditions and smoothness conditions on the functions.

Assumption 8. *The noise ε_t is i.i.d. and $\mathbb{E}(\varepsilon_t|\mathcal{F}_{-\infty}^{t-1})=0$. Furthermore, there exists a constant B such that $\mathrm{P}(\max(|\varepsilon_t|,|Y_t|,|Y_t^\star|,\|X_t\|_\infty)\le B)=1$ for all t.*

The above assumption ensures that the support of the functions $\psi^\star$ and $\eta^\star$ are compact, which is standard in nonparametrics. We note that when $\eta^\star$ is linear, the Gaussian ARMA processes clearly do not satisfy Assumption 8. However, certain non-Gaussian ARMA, bilinear, nonlinear ARX, and ARCH processes could have compact support and thus may satisfy Assumption 8.

Next we introduce the notion of smoothness.

Definition 6 (Hölder Ball)**.** *Let $\beta > 0$ and Ω be a subset of some Euclidean space, the Hölder ball of functions $\mathcal{H}^\beta(\Omega, B)$ is defined as*

$$\left\{ f : \Omega \to \mathbb{R}, \max_{\alpha, |\alpha| \leq \lfloor \beta \rfloor} \sup_{x \in \Omega} |D^\alpha f(x)| + \max_{\alpha : \|\alpha\|_1 = \lfloor \beta \rfloor} \sup_{\substack{x, x' \in \Omega \\ x \neq x'}} \frac{\left| D^\alpha f(x) - D^\alpha f\left(x'\right) \right|}{\|x - x'\|^{\beta - \lfloor \beta \rfloor}} \leq B \right\},$$

where $\lfloor \beta \rfloor$ represents the largest integer strictly smaller than β.

Assumption 9. *There exist $\beta \in \mathbb{N}_+$, such that $\psi^\star \in \mathcal{H}^\beta([-B, B]^{p+k}, B)$. Similarly, we suppose $\eta^\star \in \mathcal{H}^\beta([-B, B]^{p+q+k}, B)$. In addition, we assume that $\theta_i := \left\| \partial_{x_{p+i}} \eta^\star(x) \right\|_{\infty, [-B,B]^{p+q+k}}$, $\forall 1 \leq i \leq q$, satisfying $\theta_1 + \theta_2 + \cdots + \theta_q =: \theta < 1$.*

The condition on $\eta^\star$ in Assumption 9 provides a sufficient criterion for ensuring the model's invertibility, meaning that ε_t can be consistently recovered from Y_t as $t \to \infty$. In the linear case, the invertibility assumption is less restrictive than the one we impose here. We leave the exploration of more general conditions for nonlinear invertibility to future work.

Assumption 10. *The stationary process $\{(Y_t, \varepsilon_t, X_t)\}_{t=-\infty}^{\infty}$ is exponentially strongly mixing, with its α-mixing coefficient $\{\alpha_y(j)\}$ satisfying*

$$\alpha_y(j) \leq \bar{\alpha}_y \exp(-c_y j^\kappa) \tag{3.4}$$

for some positive constants $\bar{\alpha}_y$, c_y, and κ.

The above assumption requires that the dependencies in $(Y_t, \varepsilon_t, X_t)$ decay at an exponential rate, a condition commonly used in the literature [Mikosch et al., 2000, Horowitz, 2003]. It is worth noting that α-mixing is weaker than β-, ρ-, and ψ-mixing. Therefore, our theoretical results remain valid as long as these mixing coefficients decay at the same rate.

3.2.2 Recurrent Neural Networks

Having introduced the DGP, we introduce the architecture of RNNs in this section. We begin by introducing the rectified linear unit (ReLU) activation function defined as $\sigma(x) = \max(x, 0)$. Given $v = (v_1, \ldots, v_r) \in \mathbb{R}^r$, we define the shifted activation function $\sigma_v : \mathbb{R}^r \to \mathbb{R}^r$ as:

$$\sigma_v \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} \sigma(y_1 - v_1) \\ \vdots \\ \sigma(y_r - v_r) \end{pmatrix}.$$

The neural network architecture (d, w, n) consists of a positive integer d, representing the number of hidden layers or depth, and a width vector $n = (n_0, \ldots, n_{d+1}) \in \mathbb{N}^{d+2}$ such that $n_i \leq w$ for $i = 1, \ldots, d$. Note that n_0 and n_{d+1} represent the input and output dimensions, respectively. A neural network with the architecture (d, w, n) can be expressed as a function of the form

$$f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{d+1}}, x \to f(x) = W_d \sigma_{v_d} W_{d-1} \sigma_{v_{d-1}} \cdots W_1 \sigma_{v_1} W_0 x, \tag{3.5}$$

where W_i is an $n_{i+1} \times n_i$ weight matrix, and $v_i \in \mathbb{R}^{n_i}$ is a shift vector.

In this paper, we adopt a setup similar to Schmidt-Hieber [2020], where the parameters

of the neural network are bounded. The function space of the neural network is defined as:

$$\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C, B) := \left\{ f \text{ of the form (3.5)} : \max_{j=0,\ldots,d} \left\|W_j\right\|_\infty \vee \left|v_j\right|_\infty \leq C, \|f\|_\infty \leq B \right\}.$$

For simplicity, we omit the width vector n here. In this study, we set $C = T^{5\beta+5}$, which ensures that this constraint is naturally satisfied in most scenarios.

Unlike classical neural networks, RNNs process data across multiple time steps by adopting the fundamental building block called the recurrent unit. This unit maintains a hidden state, essentially a form of memory, which is updated at each time step based on the current input and the previous hidden state. This feedback process allows the network to learn from past inputs and incorporate this into its current processing. Denote the embedding layer (i.e., hidden state) as H_t, RNNs can be expressed as:

$$\begin{aligned} H_t &= \sigma(\rho(I_t) + W_h H_{t-1} + v_h), \\ O_t &= \varphi(H_t), \end{aligned}$$

where I_t and O_t represent the input and the output sequence, respectively. Here, $H_t \in \mathbb{R}^{w_h}$ is the embedding layer, typically the narrowest layer of the RNN. The terms W_h and v_h denote the weight matrix and bias vector for this embedding layer, while ρ and φ are DNNs. The architecture of a standard RNN is illustrated in Figure 3.1. To appropriately bound the estimation error, we restrict $W_h \in \mathbb{R}^{w_h \times w_h}$ to satisfy the following conditions:

$$W_h\text{: upper triangular matrix with } \|W_h\|_\infty \leq T^{5\beta+5} \text{ and } \|\operatorname{diag}(W_h)\|_\infty \leq 1 - \varepsilon, \quad (3.6)$$

for some fixed positive constant $\varepsilon < 1 - \theta$. Formally, we define the RNN function class $\mathcal{F}_{rnn}^{w_h}$

as

$$\mathcal{F}_{rnn}^{w_h} := \big\{(\rho, W_h, v_h, \varphi) : \rho \in \mathcal{F}_{p+k}^{w_h}(d, w, T^{5\beta+5}, B), \quad W_h \text{ of the form (3.6)},$$
$$\|v_h\|_\infty \leq T^{5\beta+5}, \varphi \in \mathcal{F}_{w_h}^{1}(d, w, T^{5\beta+5}, B)\big\}. \tag{3.7}$$

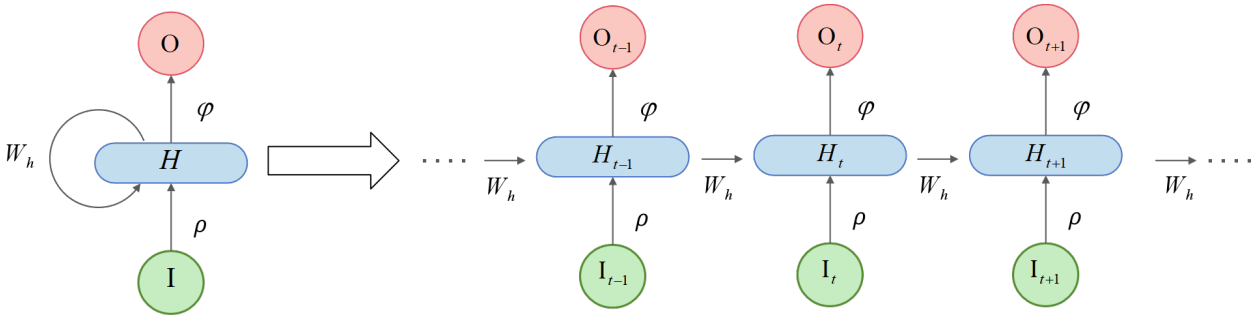


Figure 3.1: The architecture of RNNs.

3.2.3 Main Results

3.2.3.1 ARMA model as a special case

As defined by Eq. (3.11), RNNs differ from DNNs due to their capacity to retain and utilize information from previous inputs. This feature is closely connected with ARMA models. In particular, we demonstrate that linear RNNs are equivalent to ARMA model.

To fix ideas, we consider ARMA$(1,1)$ model:

$$Y_t = \theta_1 Y_{t-1} + \varepsilon_t + \theta_2 \varepsilon_{t-1}, \tag{3.8}$$

where θ_1, θ_2 are the parameters for AR term and MA term respectively, and ε_t is the noise term. To estimate the parameters in Eq. (3.8), a classical method is the least squares estimator. The details can be found in Chapter 3 of Shumway et al. [2000]. The solution is

given by

$$\hat{\theta} = \arg \min_{\tilde{\theta}=(\tilde{\theta}_1,\tilde{\theta}_2)} \sum_{t=2}^{T} (Y_t - \tilde{\theta}_1 Y_{t-1} - \tilde{\theta}_2 \tilde{\varepsilon}_{t-1})^2, \quad \tilde{\varepsilon}_{t-1} := Y_{t-1} - \tilde{\theta}_1 Y_{t-2} - \tilde{\theta}_2 \tilde{\varepsilon}_{t-2}, \tag{3.9}$$

where $Y_0 = \tilde{\varepsilon}_0 = 0$. Considering a two-layer RNN which has only one neuron in the hidden layer with a linear activation function. The model solves the following optimization problem:

$$\hat{w} = \arg \min_{\tilde{w}=(\tilde{w}_0,\tilde{w}_1,\tilde{w}_h)} \sum_{t=2}^{T} \left(Y_t - \tilde{w}_1(\tilde{w}_0 Y_{t-1} + \tilde{w}_h H_{t-1})\right)^2, \quad H_{t-1} := \tilde{w}_0 Y_{t-2} + \tilde{w}_h H_{t-2}, \tag{3.10}$$

where $Y_0 = H_0 = 0$. Proposition 5 establishes a connection between these two solutions.

Proposition 5. *For any solution $(\hat{\theta}_1, \hat{\theta}_2)$ to Eq. (3.9), $(\hat{w}_0, \hat{w}_1, \hat{w}_h)$ satisfying $\hat{w}_0\hat{w}_1 = \hat{\theta}_1 + \hat{\theta}_2$ and $\hat{w}_h = -\hat{\theta}_2$ is a solution to Eq. (3.10). In addition, their outputs $\hat{Y}_t$ are equivalent.*

3.2.3.2 General case

Recall that the best predictor is defined as $Y_t^\star := \mathbb{E}(Y_t|\mathfrak{F}_{-\infty}^{t-1})$. In this paper, we study the mean squared error between the output of the RNN and the best predictor.

Consider the RNN that minimizes the following objective:

$$\begin{aligned} &\min_{(\varphi,W_h,v_h,\rho)\in\mathcal{F}_{rnn}^{w_h}} \sum_{t=p+1}^{T} (Y_t - \varphi(H_{\rho,t}))^2, \\ &\text{subject to} \quad H_{\rho,t} = \sigma_{v_h}(W_h H_{\rho,t-1} + \rho(Y_{t-1}, \ldots, Y_{t-p}, X_{t-1})), \text{ for } t \geq p+1, \\ &\qquad\qquad \text{and } H_{\rho,t} = 0_{w_h}, \text{ for } t \leq p. \end{aligned} \tag{3.11}$$

Here, we use $H_{\rho,t}$ to emphasize its dependency on ρ.

Before introducing our main result, we define

$$\iota_\eta := \inf_{\eta \in \mathcal{F}^1_{p+q+k}(1, T^{\alpha_2}, T^{2\beta+2}, B)} \|\eta^\star - \eta\|_{\infty, [-3B, 3B]^{p+q+k}}.$$

Then we have the following theorem:

Theorem 9. *Under Assumptions 8 to 10, assume that* $w \asymp T^{\alpha_1}$, $w_h \asymp T^{\alpha_2} \log T$, $d \asymp \log T$, *for some* $0 \le \alpha_1, \alpha_2 \le 1$, *and* $\iota_\eta \to 0$. *For any fixed* $h \in \mathbb{N}^+$, *it holds that*

$$\mathbb{E}(Y^\star_{T+h} - \hat{\varphi}(H_{\hat{\rho}, T+h}))^2 \lesssim T^{-\alpha_1 \cdot \frac{4\beta}{p+k}} + \iota_\eta^2 + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)} \log^9 T,$$

where $H_{\hat{\rho}, T+h}$ *is iterated as in Eq. (3.11) and* $Y_{T+1}, \ldots, Y_{T+h-1}$ *are assumed to be observed.*

The proof of this theorem is given in Section 3.5. Here $\eta \in \mathcal{F}^1_{p+q+k}(1, T^{\alpha_2}, T^{2\beta+2}, B)$ means that the function η is a single-layer neural network with T^{α_2} neurons.

The error bound in Theorem 9 consists of three components. The first term, $T^{-\alpha_1 \cdot \frac{4\beta}{p+k}}$, represents the approximation error generated by fitting the function $\varphi^\star$. The second term, ι_η^2, is the approximation error from fitting the function $\eta^\star$ using the recurrent layer. The third term, $\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)} \log^9 T$, represents the estimation error.

We analyze two special cases. First, when the function $\eta^\star$ is linear. Under this case, by choosing $\alpha_2 = 0$, the approximation error for fitting the function $\eta^\star$ is zero, allowing us to achieve the following bound:

Corollary 5. *When* $\eta^\star$ *is linear or continuous piecewise linear with finite knots, with* $w \asymp T^{\frac{p+k}{2(2\beta+p+k)} \cdot \frac{\kappa}{1+\kappa}}$ *and* $w_h \asymp d \asymp \log T$, *under Assumptions 8 to 10, it holds that*

$$\mathbb{E}(Y^\star_{T+h} - \hat{\varphi}(H_{\hat{\rho}, T+h}))^2 \lesssim T^{-\frac{2\beta}{2\beta+p+k} \cdot \frac{\kappa}{1+\kappa}} \log^{10} T,$$

where $H_{\hat{\rho}, T+h}$ *is iterated as in Eq. (3.11), and* $Y_{T+1}, \ldots, Y_{T+h-1}$ *are treated as known.*

This rate coincides with the rate given by Men et al. [2024] for nonlinear autoregressive models with no moving average component by using DNNs.

Second, for general cases, Mao and Zhou [2023] provides a detailed upper bound for ι_η, which is related to the smoothness of $\eta^\star$ (i.e., β) and the input dimension of $\eta^\star$ (i.e., $p+q+k$) as follows[1]:

$$\iota_\eta \lesssim \begin{cases} (\log T)^{\frac{3}{2}+(p+q+k)}(T^{\alpha_2})^{-\frac{\beta}{p+q+k}\frac{(p+q+k)+2}{(p+q+k)+4}}, & \text{if } \beta \le \frac{p+q+k}{2}+2, \\ (\log T)^{\frac{1}{2}}(T^{\alpha_2})^{-\frac{1}{2}-\frac{1}{p+q+k}}, & \text{if } \beta > \frac{p+q+k}{2}+2, \end{cases}$$

where $\iota_\eta := \inf_{\eta \in \mathcal{F}^1_{p+q+k}(1,T^{\alpha_2},T^{2\beta+2},B)} \|\eta^\star - \eta\|_{\infty,[-3B,3B]^{p+q+k}}$.

Then we have the following conclusion:

Corollary 6. *With $w \asymp T^{\alpha_1}$, $w_h \asymp T^{\alpha_2}\log T$ and $d \asymp \log T$, for some $0 \le \alpha_1, \alpha_2 \le 1$, under Assumptions 8 to 10, for any fixed $h \in \mathbb{N}^+$, we have*

$$\begin{aligned} &\mathbb{E}(Y^\star_{T+h} - \hat{\varphi}(H_{\hat{\rho},T+h}))^2 \\ &\lesssim \begin{cases} T^{-\frac{\beta}{\beta+4(p+q+k+4)}\cdot\frac{\kappa}{\kappa+1}}\left((\log T)^{\frac{3}{2}+(p+q+k)} \vee \log^{10} T\right), & \textit{when } \beta \le \frac{p+q+k}{2}+2, \\ T^{-\frac{\kappa}{9(\kappa+1)}}\log^{10} T, & \textit{when } \beta > \frac{p+q+k}{2}+2, \end{cases} \end{aligned}$$

where $H_{\hat{\rho},T+h}$ is iterated as in Eq. (3.11) and $Y_{T+1},\ldots,Y_{T+h-1}$ are treated as known.

3.3 Numerical Studies

In this section, we evaluate the numerical performance of the RNNs. We consider the prediction performance of the RNNs by considering various DGPs and comparing the results with those obtained from ARMA and DNNs, two representative linear and nonlinear techniques. We demonstrate the superior generalizability and effectiveness of the RNNs.

1. The parameter condition in Mao and Zhou [2023] satisfies our restriction on the model parameters.

We consider the following six different data-generation processes. For Model 1 to Model 3, we assume ε_t follows a uniform distribution so that Assumption 8 is satisfied. In Model 4 to Model 6, Gaussian noise is applied.

- MODEL 1: We assume that Y_t follows an ARMA$(1,1)$ process:

$$Y_t = \theta_1 Y_{t-1} + \varepsilon_t + \theta_2 \varepsilon_{t-1}.$$

  Here we assume ε_t follows a uniform distribution, $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{U}[-1,1]$, and we set $\theta_1 = \theta_2 = 0.5$. In Table 3.1, we also vary their values to check the robustness of our results.

- MODEL 2: We assume that Y_t follows the following functional coefficient autoregressive (FAR) model [Chen and Tsay, 1993, Hastie and Tibshirani, 1993, Cai et al., 2000] with a linear MA noise:

$$Y_t = \theta_1 \exp(-Y_{t-1}^2 + 1) Y_{t-1} + \varepsilon_t + \theta_2 \varepsilon_{t-1},$$

  where $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{U}[-1,1]$ and we set $\theta_1 = \theta_2 = 0.5$ here.

- MODEL 3: We assume that Y_t follows the bilinear model [Rao, 1981, Rao and Gabr, 2012] as follows:

$$Y_t = \theta_1 Y_{t-1} + \varepsilon_t + \theta_2 (\varepsilon_{t-1} + Y_{t-1}\varepsilon_{t-1}),$$

  where $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{U}[-1,1]$ and we set $\theta_1 = \theta_2 = 0.5$ here.

- MODEL 4: We assume that Y_t follows a model combining a linear regression term and a polynomial moving average term, as introduced by [Robinson, 1977, Hsieh, 1991,

Campbell et al., 1998]:

$$Y_t = \theta_1 Y_{t-1} + \varepsilon_t + \theta_2 \varepsilon_{t-1}^2,$$

where $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and we set $\theta_1 = \theta_2 = 0.5$ here.

- MODEL 5: We assume that Y_t follows the logistic smooth transition autoregressive (LSTAR) model[Chan and Tong, 1986, Teräsvirta, 1994, Eitrheim and Teräsvirta, 1996] as follows:

$$Y_t = \theta_1 Y_{t-1} + \theta_2 \left(1 + \exp\left(-\gamma \left(Y_{t-1} - c\right)\right)\right)^{-1} Y_{t-1} + \varepsilon_t + \theta_3 \varepsilon_{t-1},$$

where $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and we set $\theta_1 = 0.2$, $\theta_2 = \theta_3 = 0.5$, $\gamma = 20$ and $c = 1$ here.

- MODEL 6: We assume that Y_t follows the threshold autoregressive moving average (TARMA) model [Hansen, 2000, Tong, 2012] as follows:

$$Y_t = \theta_1 |Y_{t-1}| + \varepsilon_t + \theta_2 \varepsilon_{t-1} I(\varepsilon_{t-1} \leq 0) + \theta_3 \varepsilon_{t-1} I(\varepsilon_{t-1} > 0) + \sum_{i=1}^{7} \beta_i X_{i,t-1},$$

where $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, I is the indicator function, and we set $\theta_1 = 0.2$, $\theta_2 = -0.2$, $\theta_3 = 0.8$, $\beta_1 = \beta_2 = \beta_3 = -0.2$ and $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0.2$ here.

Unlike a standard RNN, our theory imposes constraints on the weight matrix W_h, requiring it to be an upper triangular matrix with diagonal elements less than $1 - \varepsilon$ (as specified in Eq. (3.6)). To assess whether this condition impacts the performance, we compare our model (RNN1, with the restricted W_h) to the canonical RNN (RNN2, without such a restriction on W_h).

For these six DGPs, Model 1 is a linear process and Model $2 - 6$ are nonlinear process. In addition, the setup of Model 6 is similar to that in the empirical study, both having 7

exogenous variables, and the out-of-sample R^2 is also similar (around 5% when $T = 200 \sim 400$). We perform 100 independent replications and report the prediction mean squared error on the test set. We use the first 80% of the data as the training set and the remaining 20% is used as the validation set to select the tuning parameters. For ARMA, its tuning parameters are the lagged order (p, q). For RNNs and DNNs, we tune the network structure $\{(4), (4, 4, 4), (8, 4, 8)\}$, and the learning rate $\{0.001, 0.01, 0.1\}$. After training the models, we generate 100 additional observations as our test data to evaluate these models' performance.

The results are shown in Figure 3.2. From Figure 3.2, we observe that the prediction mean squared error of RNNs decreases as T increases, consistent with the conclusion in Theorem 9. Moreover, RNNs outperform the other three models for nonlinear processes (Model $2-6$) and perform only slightly worse than ARMA for linear processes (Model 1), with this gap tending to zero as T increases. This highlights the advantage of RNNs, which exhibit strong fitting capabilities for both linear and nonlinear processes. The results also demonstrate the effectiveness and applicability of RNNs, underscoring their clear advantage in handling complex time series compared to linear models (ARMA) or nonlinear models that ignore the lagged noise terms (DNNs). Moreover, in Model $4-6$, we considered the Gaussian-distributed noise instead of the uniformly distributed noise, which is unbounded and falls outside the scope of our theoretical assumptions. Nonetheless, the RNN's performance remains strong, demonstrating the robustness of the model. Moreover, comparing the results of restricted W_h and unrestricted W_h, we observe that the constraints on W_h do not adversely affect the convergence of RNNs. In fact, in most scenarios, the model with the restricted W_h performs better than the model without the restriction. This suggests that the constraint does not degrade the model's performance.

To study the performance of RNNs under different parameter settings, we vary parameters (θ_1, θ_2) in Models 1-3 for further study. We consider nine scenarios where $\theta_1 \in \{0.2, 0.5, 0.8\}$ and $\theta_2 \in \{0, 0.2, 0.5, 0.8\}$. The results are shown in Table 3.1. We can observe

that as the parameter scale, especially the parameters related to the lagged noise terms, increases, the performance of RNNs gradually surpasses that of DNNs. When $\theta_2 = 0$, all these three models no longer include the lagged noise terms. In this situation, DNNs performs slightly better than RNNs, however, the advantage is not substantial. This aligns with expectations, as in the absence of lagged noise terms, the hidden layer of RNNs may lead to overfitting, which results in a degradation of out-of-sample performance. Furthermore, RNNs usually outperforms ARMA in the nonparametric cases (Models 2 and 3). And the abnormal results of Model 3 when $(\theta_1, \theta_2) = (0.8, 0.8)$ is because the model is no longer stationary under this case.

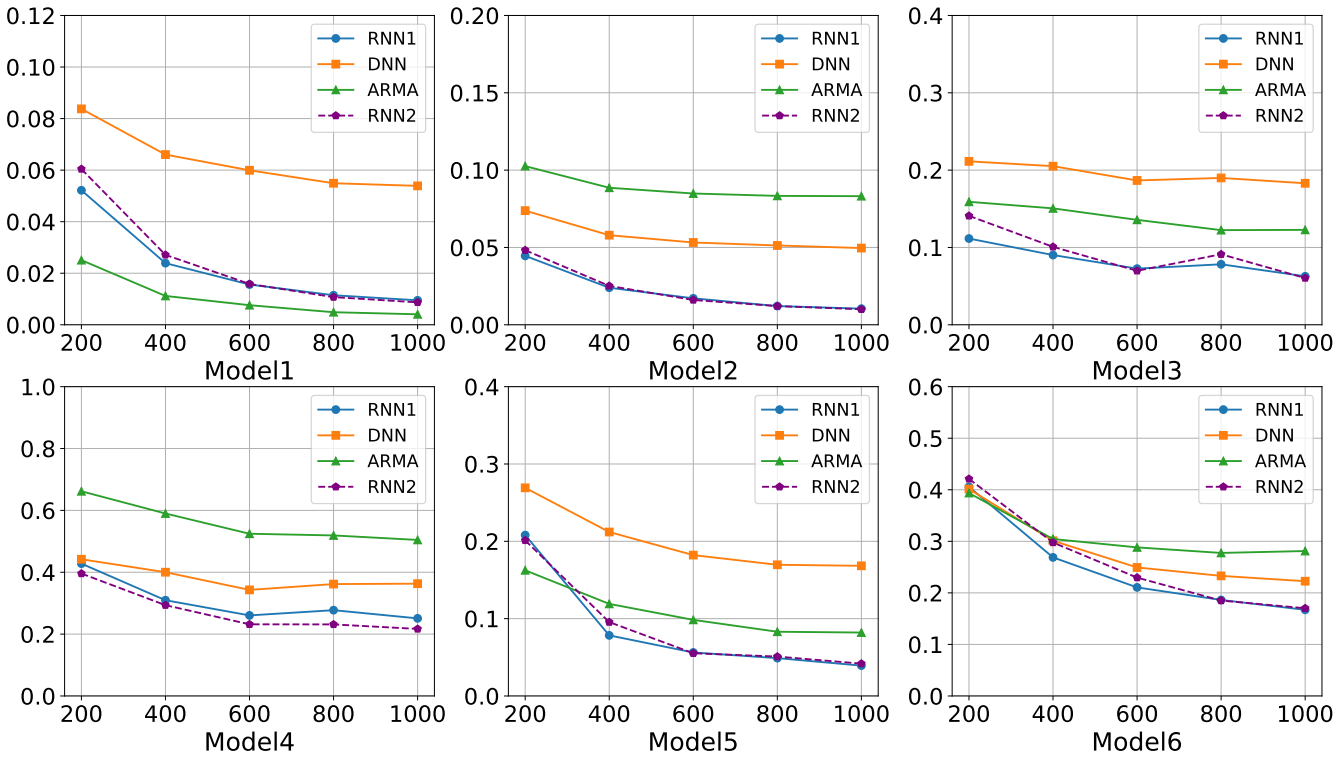


Figure 3.2: RNNs' Predictive Performance for Different Models

Table 3.1: RNNs' Predictive Performance under Different Parameters

	Model 1			Model 2			Model 3		
(θ_1, θ_2)	RNN	DNN	ARMA	RNN	DNN	ARMA	RNN	DNN	ARMA
$(0.2, 0)$	0.006	0.005	0.003	0.008	0.007	0.007	0.006	0.005	0.003
$(0.2, 0.2)$	0.007	0.006	0.003	0.010	0.009	0.009	0.006	0.005	0.007
$(0.2, 0.5)$	0.008	0.032	0.004	0.012	0.017	0.014	0.009	0.010	0.031
$(0.2, 0.8)$	0.009	0.114	0.003	0.016	0.061	0.028	0.021	0.053	0.102
$(0.5, 0)$	0.008	0.006	0.003	0.013	0.011	0.051	0.008	0.007	0.003
$(0.5, 0.2)$	0.008	0.012	0.004	0.013	0.012	0.061	0.008	0.007	0.008
$(0.5, 0.5)$	0.011	0.052	0.003	0.015	0.024	0.083	0.013	0.029	0.049
$(0.5, 0.8)$	0.011	0.155	0.004	0.020	0.072	0.126	0.134	0.238	0.233
$(0.8, 0)$	0.009	0.007	0.003	0.018	0.013	0.187	0.010	0.007	0.003
$(0.8, 0.2)$	0.012	0.019	0.003	0.018	0.018	0.190	0.015	0.023	0.016
$(0.8, 0.5)$	0.017	0.080	0.003	0.018	0.032	0.200	0.166	0.229	0.207
$(0.8, 0.8)$	0.014	0.198	0.003	0.023	0.080	0.259	3.447	3.274	1.834

3.4 Empirical Studies

In this section, we apply RNNs to monthly exchange rate data, which are notoriously difficult to predict [Engel and West, 2005, Rossi, 2013]. Specially, we define the return as $Y_{t+1} = \ln p_{t+1} - \ln p_t$, where p_t is the monthly bilateral exchange rate of the British pound (UK), Swiss Franc (SZ) or Japanese yen (JP) relative to the US dollar. In exchange rate forecasting studies, the driftless random walk model is commonly used as a benchmark [Wright, 2008, Rossi, 2013], which predicts $\ln p_{t+1}$ as $\ln p_t$ at time t. This is because many researchers have found that the simple random walk model often outperforms more complex economic models in predicting exchange rates [Meese and Rogoff, 1983, Mark, 1995]. Here, we compare the predictive performance of RNNs with ARMA, DNNs, the Kernel regression, and the Kernel Ridge regression[2].

For predicting exchange rate returns, we use the following seven exogenous variables (i.e., X_t): (i) the relative stock price (foreign vs. US), (ii) the relative annual stock price growth

2. Here, the candidates network structure of the RNN and DNN are the same as the setup in Numerical Studies. The parameters of the Kernel (bandwidth) and KernelRidge (bandwidth and regularization parameter) are also determined through tuning, ensuring that the selected optimal tunings do not spike at the boundary.

rate, (iii) the relative short-term interest rate, (iv) the relative term spread (long minus short term rates, motivated by the findings of [Bekaert and Hodrick, 1992, Clarida et al., 2003]), (v) the exchange rate return from the previous month, (vi) the sign of the exchange rate return from the previous month, and (vii) the relative inflation rate; see Wright [2008] and Chen and Liu [2023]. The sampling period spans from January 1990 to December 2023, including 34 years of data. All the variables have been transformed to stationary according to Wright [2008] and Chen and Liu [2023].

To fit these models, we use an expanding windows approach. Specifically, we select the first 180 months (15 years) as the initial training set and the next 60 months (5 years) as the validation set to tune the hyper-parameters. In each window, the model predicts exchange rate returns for the following 12 months. This procedure is repeated 14 times, with the training set expanding by one year and the validation set shifting accordingly at each iteration. Thus, we have a total of 168 monthly forecasts. For comparison, we use random walk prediction as the benchmark to calculate the out-of-sample R^2.

The results are presented in Table 3.2, where we report the R^2_{oos} values for different models across various lag orders $p = 1, 2, 3$ respectively. Across all settings, RNNs consistently outperform the random walk, demonstrating its robustness and stability in exchange rate return prediction. Notably, RNNs perform well when the lag order $p = 1$ or $p = 2$. DNNs, on the other hand, perform well with $p = 3$, likely compensating for the residual terms by incorporating a higher lag order.

3.5 Mathematical Proofs

Throughout the proof, the constants c and C are fixed and only depend on other fixed constants, which may vary from line to line.

Table 3.2: Percentage R^2_{oos} benchmark to random walk for various methods.

		RNN	ARMA	DNN	ARX	Kernel	KernelRidge
$p=1$	UK	5.368	−1.854	2.895	−2.050	4.144	2.953
	SZ	3.453	−7.988	0.589	−7.095	−4.607	−7.384
	JP	9.833	9.431	7.443	7.881	1.180	2.163
$p=2$	UK	2.554	−4.541	2.624	−6.674	2.375	2.227
	SZ	3.545	−7.517	2.711	−13.518	−1.077	−10.196
	JP	9.888	6.097	4.574	7.043	2.477	4.419
$p=3$	UK	2.804	−4.156	4.568	−8.091	3.215	1.287
	SZ	1.680	−4.060	2.684	−14.658	−1.607	−9.486
	JP	6.212	−3.771	4.847	7.172	3.095	2.405

3.5.1 Proof of Proposition 5

For any $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$, the following equation holds for the residuals from ARMA estimation

$$Y_t - \tilde{\theta}_1 Y_{t-1} - \tilde{\theta}_2 \tilde{\varepsilon}_{t-1} = Y_t - (\tilde{\theta}_1 + \tilde{\theta}_2) \sum_{j=1}^{t-1} (-\tilde{\theta}_2)^{j-1} Y_{t-j}.$$

Similarly, for any $(\tilde{w}_0, \tilde{w}_1, \tilde{w}_h)$, the following equation holds for the residuals from RNNs,

$$Y_t - \tilde{w}_1 \left(\tilde{w}_0 Y_{t-1} + \tilde{w}_h H_{t-1}\right) = Y_t - \tilde{w}_1 \tilde{w}_0 \sum_{j=1}^{t-1} (\tilde{w}_h)^{j-1} Y_{t-j}.$$

Based on the above equations, we conclude that if $(\hat{\theta}_1, \hat{\theta}_2)$ is a solution to the ARMA optimization problem, $(\hat{w}_0, \hat{w}_1, \hat{w}_h)$ satisfying $\hat{w}_0 \hat{w}_1 = \hat{\theta}_1 + \hat{\theta}_2$ and $\hat{w}_h = -\hat{\theta}_2$ is a solution to the RNN optimization problem. Additionally, we have

$$\begin{aligned}\hat{\theta}_1 Y_{t-1} + \hat{\theta}_2 \hat{\varepsilon}_{t-1} = (\hat{\theta}_1 + \hat{\theta}_2) \sum_{j=1}^{t-1} (-\hat{\theta}_2)^{j-1} Y_{t-j} &= \hat{w}_1 \hat{w}_0 \sum_{j=1}^{t-1} (\hat{w}_h)^{j-1} Y_{t-j} \\ &= \hat{w}_1 \left(\hat{w}_0 Y_{t-1} + \hat{w}_h \hat{H}_{t-1}\right),\end{aligned}$$

which completes the proof.

3.5.2 Proof of Theorem 9

For notational simplicity, we write

$$\inf_{\eta \in \mathcal{F}^1_{p+q+k}(1, T^{\alpha_2}, T^{2\beta+2}, B)} \|\eta^\star - \eta\|_{\infty, [-3B, 3B]^{p+q+k}} = \iota_\eta.$$

Note that for all $(\rho, W_h, v_h, \varphi) \in \mathcal{F}^{w_h}_{rnn}$, the following inequality holds:

$$\sum_{t=p+1}^{T} (Y_t - \hat{\varphi}(H_{\hat{\rho},t}))^2 \leq \sum_{t=p+1}^{T} (Y_t - \varphi(H_{\rho,t}))^2.$$

Based on this inequality and recognizing that $Y_t = Y_t^\star + \varepsilon_t$ and $\hat{Y}_t := \hat{\varphi}(H_{\hat{\rho},t})$, we obtain:

$$\begin{aligned}
\sum_{t=p+1}^{T} \big(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t})\big)^2 &= \sum_{t=p+1}^{T} \Big[\big(Y_t - \hat{\varphi}(H_{\hat{\rho},t})\big)^2 - \varepsilon_t^2 + 2\hat{\varphi}(H_{\hat{\rho},t})\varepsilon_t - 2Y_t^\star \varepsilon_t\Big] \\
&\leq \sum_{t=p+1}^{T} \Big[\big(Y_t - \varphi(H_{\rho,t})\big)^2 - \varepsilon_t^2 + 2\hat{Y}_t \varepsilon_t - 2Y_t^\star \varepsilon_t\Big] \qquad (3.12) \\
&= \sum_{t=p+1}^{T} \Big[\big(Y_t^\star - \varphi(H_{\rho,t})\big)^2 + 2\varepsilon_t \Big(\big(Y_t^\star - \varphi(H_{\rho,t})\big) + \big(\hat{Y}_t - Y_t^\star\big)\Big)\Big].
\end{aligned}$$

Below we first establish three inequalities, denoted as (I) through (III), with their respective proofs provided subsequently.

(I) For any fixed $(\rho, W_h, v_h, \varphi) \in \mathcal{F}^{w_h}_{rnn}$ and fixed positive constant c, with probability at least $1 - C\exp(-c\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T)$ for some constant C, the following

holds:

$$\begin{aligned}
&\left|T^{-1}\sum_{t=p+1}^{T}\varepsilon_t\big(Y_t^\star-\varphi(H_{\rho,t})\big)\right| \\
\lesssim&\left(T^{-1}\sum_{t=p+1}^{T}\mathbb{E}\big(Y_t^\star-\varphi(H_{\rho,t})\big)^2\right)^{1/2}(\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&+\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T.
\end{aligned}$$

(II): With probability at least $1-C\exp(-\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\}\log^5 T)$, the following inequality holds:

$$\begin{aligned}
&\left|T^{-1}\sum_{t=p+1}^{T}\varepsilon_t\big(Y_t^\star-\hat{\varphi}(H_{\hat{\rho},t})\big)\right| \\
\lesssim&\left(T^{-1}\sum_{t=p+1}^{T}\mathbb{E}\big(Y_t^\star-\hat{\varphi}(H_{\hat{\rho},t})\big)^2\right)^{1/2}(\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&+\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T.
\end{aligned}$$

(III): There exists a fixed $(\rho^\dagger,W_h^\dagger,v_h^\dagger,\varphi^\dagger)\in\mathcal{F}_{rnn}^{w_h}$ such that

$$\left|T^{-1}\sum_{t=p+1}^{T}\big(Y_t^\star-\varphi^\dagger(H_{\rho^\dagger,t})\big)^2\right|\lesssim T^{-\alpha_1\cdot\frac{4\beta}{p+k}}+\iota_\eta^2+T^{-1}\log T.$$

3.5.2.1 Proof of (I)

Recall that

$$H_{\rho,t}=\sigma_{v_h}\left(\rho(Y_{t-1},\ldots,Y_{t-p},X_{t-1})+W_hH_{\rho,t-1}\right).$$

For any $p+1 \le t \le T$, we define

$$
\begin{aligned}
H_{\rho,t}^{0} &= \sigma_{v_h}\left(\rho(Y_{t-1},\ldots,Y_{t-p},X_{t-1})\right), \\
H_{\rho,t}^{\ell} &= \sigma_{v_h}\left(\rho(Y_{t-1},\ldots,Y_{t-p},X_{t-1}) + W_h H_{\rho,t-1}^{\ell-1}\right), \qquad \text{for } \ell \ge 1 \text{ and } t > \ell.
\end{aligned} \tag{3.13}
$$

By convention, we set $H_{\rho,t}^{\ell} = 0_{w_h}$ for $1 \le t \le p$. Recalling the definition of $\mathfrak{F}_i^j$ in Eq. (3.3), it is straightforward to verify that $H_{\rho,t}^{\ell} \in \mathfrak{F}_{t-p-\ell}^{t-1}$ for $t \ge p+\ell+1$. On the other hand, by Lemma 44, when $t \ge p+\ell+1$, it holds that $\|H_{\rho,t}^{\ell} - H_{\rho,t}\|_\infty \le C(1-\varepsilon)^{\ell-w_h}(2\ell T^{5\beta+5})^{w_h}$ for some constant C. Note that $\varphi \in \mathcal{F}_{w_h}^1(d, w, T^{5\beta+5}, B)$, together with Lemma 41, we have

$$
|\varphi(H_{\rho,t}) - \varphi(H_{\rho,t}^{\ell})| \le C w_h T^{(5\beta+5)(d+1)} w^d (1-\varepsilon)^{\ell-w_h}(2\ell T^{5\beta+5})^{w_h}, \quad \forall t \ge p+\ell+1.
$$

Recalling that $w \asymp T^{\alpha_1}$, $w_h \asymp T^{\alpha_2}\log T$ for some $0 \le \alpha_1, \alpha_2 \le 1$ and $d \asymp \log T$, it is straightforward to verify that, by letting $\ell = C w_h \log^2 T$ for some fixed C, we have

$$
|\varphi(H_{\rho,t}) - \varphi(H_{\rho,t}^{\ell})| \lesssim T^{-1}. \tag{3.14}
$$

Given this, we first establish an upper bound for $T^{-1}\sum_{t=p+\ell+1}^{T} \varepsilon_t\big(Y_t^\star - \varphi(H_{\rho,t}^{\ell})\big)$. Define

$$
D_t := \varepsilon_t\big(Y_t^\star - \varphi(H_{\rho,t}^{\ell})\big) \in \mathfrak{F}_{t-p-\ell}^{t}. \tag{3.15}
$$

Based on this, we establish in Lemma 45 that for the stationary process $\{D_t\}_{t=p+\ell+1}^{T}$,

$$
\alpha_d(j) \le \bar{\alpha}_d \exp(-c_d j^{\kappa}),
$$

where $\bar{\alpha}_d = \max(\bar{\alpha}_y, 1)\exp(c_y)$ and $c_d = c_y(\ell+p+1)^{-\kappa}$. Additionally, since $\mathbb{E}(\varepsilon_t|\mathcal{F}_{-\infty}^{t-1}) = 0$, $|\varepsilon_t| \le B$, $|Y_t^\star| \le B$, and $|\varphi(H_{\rho,t}^{\ell})| \le B$, we have $\mathbb{E}D_t = 0$, $|D_t| \le 4B^2$, and $\mathbb{E}|D_t|^2 \le B^2\mathbb{E}\big(Y_t^\star - \varphi(H_{\rho,t}^{\ell})\big)^2$. Applying Lemma 40 to $\{D_t\}_{t=p+\ell+1}^{T}$ and noting that $c_d \asymp (w_h \log^2 T)^{-\kappa}$,

it follows that for all $\tau \in \mathbb{R}^+$ and $0 \le \zeta \le (4B^2)^{-1}$,

$$\mathrm{P}\left(\left|\frac{1}{T}\sum_{t=p+\ell+1}^{T} D_t\right| \ge \frac{\tau}{3\zeta T^{(\alpha)}} + \frac{3\zeta\mathbb{E}|D_t|^2}{2(1-4B^2\zeta)}\right) \le 2(1+4\exp(-2)\bar{\alpha}_d)\exp(-\tau),$$

where $T^{(\alpha)} \asymp (Tw_h^{-1}\log^{-2}T)^{\frac{\kappa}{\kappa+1}} = (T^{1-\alpha_2}\log^{-3}T)^{\frac{\kappa}{\kappa+1}}$. By letting

$$\zeta = \min((\mathbb{E}|D_t|^2)^{-1/2}(\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T, \log^{-1}T)$$
$$\text{and } \tau = c\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T,$$

it follows that, with probability at least $1 - C\exp(-c\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T)$,

$$\left|\frac{1}{T}\sum_{t=p+\ell+1}^{T} D_t\right| \lesssim (\mathbb{E}|D_t|^2)^{1/2}(\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T$$
$$+\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T.$$

Combining this with the fact that $\ell \asymp w_h\log^2 T$, $\left|T^{-1}\sum_{t=p+1}^{p+\ell} D_t\right| \le T^{-1}\ell\cdot 4B^2 \asymp T^{-1}w_h\log^2 T$, $w_h \asymp T^{\alpha_2}\log T$ and $\mathbb{E}|D_t|^2 \lesssim \mathbb{E}\big(Y_t^\star - \varphi(H_{\rho,t}^{\ell})\big)^2$ and using Eq. (3.14), we

obtain that, with probability at least $1 - C\exp(-c\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T)$,

$$
\begin{aligned}
&\left|\frac{1}{T}\sum_{t=p+1}^{T}\varepsilon_t\big(Y_t^\star - \varphi(H_{\rho,t})\big)\right| \lesssim \left|\frac{1}{T}\sum_{t=p+1}^{T} D_t\right| + T^{-1} \\
&\lesssim \left(\frac{1}{T}\sum_{t=p+\ell+1}^{T}\mathbb{E}\big(Y_t^\star - \varphi(H_{\rho,t}^{\ell})\big)^2\right)^{1/2} (\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&\qquad + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T \\
&\lesssim \left(\frac{1}{T}\sum_{t=p+1}^{T}\mathbb{E}\big(Y_t^\star - \varphi(H_{\rho,t}^{\ell})\big)^2\right)^{1/2} (\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&\qquad + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T \\
&\lesssim \left(\frac{1}{T}\sum_{t=p+1}^{T}\mathbb{E}\big(Y_t^\star - \varphi(H_{\rho,t})\big)^2\right)^{1/2} (\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&\qquad + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T,
\end{aligned}
$$

where $-1+\alpha_2 < 2\alpha_1+\alpha_2 - \frac{\kappa}{\kappa+1}(1-\alpha_2)$ is used in the above inequalities.

3.5.2.2 Proof of (II)

By Lemma 46, there exists a function class $\mathcal{F}_{rnn,\delta}^{w_h} \subset \mathcal{F}_{rnn}^{w_h}$, which is a $\delta-$covering of $\mathcal{F}_{rnn}^{w_h}$ with $\delta = T^{-1}$. Denote this class as $\big\{(\overline{\rho}_j, \overline{W}_{h,j}, \overline{v}_{h,j}, \overline{\varphi}_j) : 1 \le j \le |\mathcal{F}_{rnn,\delta}^{w_h}|\big)\}$. Therefore, for any $(\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn}^{w_h}$, there exists a corresponding $(\overline{\rho}_j, \overline{W}_{h,j}, \overline{v}_{h,j}, \overline{\varphi}_j) \in \mathcal{F}_{rnn,\delta}^{w_h}$ such that

$$
|\varphi(H_{\rho,t}) - \overline{\varphi}_j(H_{\overline{\rho}_j,t})| \le T^{-1}, \quad \text{for } t \gtrsim w_h \log^2 T.
$$

In particular, there exists a $(\overline{\rho}_{j^\star}, \overline{W}_{h,j^\star}, \overline{v}_{h,j^\star}, \overline{\varphi}_{j^\star})$ such that $|\hat{\varphi}(H_{\hat{\rho},t}) - \overline{\varphi}_{j^*}(H_{\overline{\rho}_{j^*},t})| \leq T^{-1}$. Additionally, by Lemma 46, we have

$$\log |\mathcal{F}_{rnn,\delta}^{w_h}| \lesssim T^{2\alpha_1+\alpha_2} \log^5 T + T^{3\alpha_2} \log^6 T.$$

Note that we can choose the constant c in (I) sufficiently large so that

$$-c \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} \log^5 T + \log |\mathcal{F}_{rnn,\delta}^{w_h}| \leq -T^{2\alpha_1+\alpha_2} \log^5 T.$$

Therefore, by applying the union bound inequality to (I), we obtain,

$$\begin{aligned}
&\mathrm{P}\Bigg(\left| T^{-1} \sum_{t=p+1}^{T} \varepsilon_t \big(Y_t^\star - \overline{\varphi}_j(H_{\overline{\rho}_j,t})\big) \right| \gtrsim \left(T^{-1} \sum_{t=p+1}^{T} \mathbb{E}\big(Y_t^\star - \varphi(H_{\rho,t}^{\ell})\big)^2 \right)^{1/2} \\
&\qquad \times (\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\})^{1/2} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2} \log^4 T \\
&\qquad + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)} \log^9 T, \forall 1 \leq j \leq |\mathcal{F}_{rnn,\delta}^{w_h}| \Bigg) \\
&\leq C |\mathcal{F}_{rnn,\delta}^{w_h}| \exp(-c \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} \log^5 T) \\
&\leq C \exp(-\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} \log^5 T). \qquad (3.16)
\end{aligned}$$

As a consequence, with probability at least $1 - C \exp(-\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} \log^5 T)$,

$$\begin{aligned}
&\left| T^{-1} \sum_{t=p+1}^{T} \varepsilon_t \big(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t})\big) \right| \lesssim \left| T^{-1} \sum_{t \asymp w_h \log^2 T}^{T} \varepsilon_t \big(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t})\big) \right| + T^{-1+\alpha_2} \log^3 T \\
&\lesssim \left(T^{-1} \sum_{t=p+1}^{T} \mathbb{E}\big(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t})\big)^2 \right)^{1/2} (\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\})^{1/2} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2} \log^4 T \\
&\qquad + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)} \log^9 T.
\end{aligned}$$

Here in the first inequality, we use the fact that $\ell \asymp T^{\alpha_2}\log^3 T$. In the second and the fourth inequality, we use the bound $|\hat{\varphi}(H_{\hat{\rho},t}) - \overline{\varphi}_{j^*}(H_{\overline{\rho}_{j^*},t})| \leq T^{-1}$. In the third inequality, we use Eq. (3.16) and in the last inequality, we use the fact that $T^{-1+\alpha_2}\log^3 T + T^{-1} = o(\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T)$.

3.5.2.3 Proof of (III)

Note that $w_h = CT^{\alpha_2}\log T$, we define $m = T^{\alpha_2}$ and $\tilde{m} = m + q - 1$. Without loss of generality, we assume C is large enough such that $\tilde{w}_h = q\tilde{m}\log_{1/\theta} T + 1 < w_h$. We verify the existence of $(\rho^\dagger, W_h^\dagger, v_h^\dagger, \varphi^\dagger) \in \mathcal{F}_{rnn}^{\tilde{w}_h}$ that satisfies (III). Based on this, we can add $w_h - \tilde{w}_h$ "useless" neurons into the recurrent layer with zero weights and biases. This approach ensures that the network's output remains unchanged, thereby satisfying (III). For ease of notation, we still use w_h to denote $\tilde{w}_h$.

By Assumption 9, $\psi^\star \in \mathcal{H}^\beta([-B,B]^{p+k}, B)$. According to Lemma 43, there exists $\psi^\dagger \in \mathcal{F}_{p+k}^1(d, w, T^{5\beta+5}, B)$ such that

$$|\psi^\dagger(x) - \psi^\star(x)| \leq T^{-\alpha_1\cdot\frac{2\beta}{p+k}}, \quad \forall x \in [-B,B]^{p+k}. \tag{3.17}$$

Assume that $|a_1|, \ldots, |a_m|, |v_1|, \ldots, |v_m|, \|b_1\|_\infty, \ldots, \|b_m\|_\infty \leq T^{2\beta+2}$ and

$$\eta^\dagger(x) = \sum_{i=1}^m a_i \sigma_{v_i}\left(b_i^\top x\right) \in \mathcal{F}_{p+q+k}^1(1, T^{\alpha_2}, T^{2\beta+2}, B)$$

satisfying

$$\left\|\eta^\star - \eta^\dagger\right\|_{\infty,[-3B,3B]^{p+q+k}} = \inf_{\eta\in\mathcal{F}_{p+q+k}^1(1,T^{\alpha_2},T^{2\beta+2},B)} \|\eta^\star - \eta\|_{\infty,[-3B,3B]^{p+q+k}} = \iota_\eta. \tag{3.18}$$

Denote $1_{\tilde{m}} \in \mathbb{R}^{\tilde{m}}$ to be a vector where all elements are equal to 1, $0_{\tilde{m}} \in \mathbb{R}^{\tilde{m}}$ to be a vector

where all elements are equal to 0, and $O_{\tilde{m}} \in \mathbb{R}^{\tilde{m}\times\tilde{m}}$ to be a matrix where all elements are zero. Let $b_i = (b_{i,1}, \ldots, b_{i,p}, b_{i,p+1}, \ldots, b_{i,p+q}, \ldots, b_{i,p+q+k}) \in \mathbb{R}^{p+q+k}$, we define

$$W = \begin{bmatrix} -b_{1,p+1}a_1 & \cdots & -b_{1,p+1}a_m & -b_{1,p+2} & \cdots & \cdots & \cdots & -b_{1,p+q} \\ -b_{2,p+1}a_1 & \cdots & -b_{2,p+1}a_m & -b_{2,p+2} & \cdots & \cdots & \cdots & -b_{2,p+q} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \vdots \\ -b_{m,p+1}a_1 & \cdots & -b_{m,p+1}a_m & -b_{m,p+2} & \cdots & \cdots & \cdots & -b_{m,p+q} \\ -a_1 & \cdots & -a_m & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{\tilde{m}\times\tilde{m}}.$$

We choose $W_h^\dagger \in \mathbb{R}^{w_h \times w_h}$, $v_h^\dagger \in \mathbb{R}^{w_h}$ and $\rho^\dagger(x) \in \mathbb{R}^{w_h}$ as follows:

$$W_h^\dagger = \begin{bmatrix} O_{\tilde{m}} & W & O_{\tilde{m}} & \cdots & \cdots & O_{\tilde{m}} & V_1 \\ O_{\tilde{m}} & O_{\tilde{m}} & W & O_{\tilde{m}} & \cdots & O_{\tilde{m}} & V_1 \\ \vdots & \ddots & \vdots & \ddots & \ddots & \vdots & V_1 \\ \vdots & \ddots & \vdots & \ddots & \ddots & \vdots & \vdots \\ O_{\tilde{m}} & \cdots & \cdots & \cdots & O_{\tilde{m}} & W & V_1 \\ O_{\tilde{m}} & \cdots & \cdots & \cdots & O_{\tilde{m}} & O_{\tilde{m}} & 0_{\tilde{m}} \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 0 \end{bmatrix}, v_h^\dagger = \begin{bmatrix} V_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ V_2 \\ -B \end{bmatrix}, \rho^\dagger(x) = \begin{bmatrix} V_3 \\ \vdots \\ \vdots \\ \vdots \\ V_3 \\ V_2 \\ \psi^\dagger(x) \end{bmatrix}, \quad (3.19)$$

where $V_1 = (-b_{1,p+1}, \ldots, -b_{m,p+1}, -1, 0_{q-2}^\top)^\top \in \mathbb{R}^{\tilde{m}}$, $V_2 = (v_1, \ldots, v_m, -B, \ldots, -B) \in \mathbb{R}^{\tilde{m}}$, and $V_3 = (b_{1,p+1}x_1 + \sum_{i=1}^{p} b_{1,i}x_i + \sum_{i=p+q+1}^{p+q+k} b_{1,i}x_i + \sum_{i=p+1}^{p+q} b_{1,i}B, \ldots, b_{m,p+1}x_1 + \sum_{i=1}^{p} b_{m,i}x_i + \sum_{i=p+q+1}^{p+q+k} b_{m,i}x_i + \sum_{i=p+1}^{p+q} b_{m,i}B, x_1 + B, -B1_{q-2}^\top)^\top \in \mathbb{R}^{\tilde{m}}$. Finally, for

$x = (x_1, \ldots, x_{w_h}) \in \mathbb{R}^{w_h}$ we define

$$\varphi^\dagger(x) = \sigma\left(2B - \sigma\left(B - \left(\sum_{i=1}^{m} a_i x_i + x_{w_h} - B\right)\right)\right) - B.$$

It is not hard to verify that

$$\varphi^\dagger(x) = \begin{cases} \sum_{i=1}^{m} a_i x_i + x_{w_h} - B, & \text{if } |\sum_{i=1}^{m} a_i x_i + x_{w_h} - B| \le B, \\ B, & \text{if } \sum_{i=1}^{m} a_i x_i + x_{w_h} - B \ge B, \\ -B, & \text{if } \sum_{i=1}^{m} a_i x_i + x_{w_h} - B \le -B. \end{cases} \tag{3.20}$$

Though $\varphi^\dagger \in \mathcal{F}^1_{w_h}(2, 1, T^{5\beta+5}, B)$, we can add several useless neurons with zero weights and bias and several hidden layers which achieve identity map by using the fact that $x = \sigma(x) - \sigma(-x)$ into $\varphi^\dagger$ to make it belong to $\mathcal{F}^1_{w_h}(d, w, T^{5\beta+5}, B)$.

Below, we demonstrate that $(\rho^\dagger, W_h^\dagger, v_h^\dagger, \varphi^\dagger)$ satisfies the required condition. Define $I_{k,i} = I(k - \lfloor |\frac{k-2}{q}| \rfloor q - i > 0)$, where $I(\cdot)$ is the indicator function, $Y_{(t-1:t-p)} = (Y_{t-1}, \ldots, Y_{t-p})$, and $\varepsilon_{(t-1:t-q)} = (\varepsilon_{t-1}, \ldots, \varepsilon_{t-q})$. By Lemma 47 with $r = \lceil \log_{1/\theta} T \rceil q$ and the fact that $\sum_{i=0}^{\infty} \theta^i \lesssim 1$, when $t \ge \max\{p, q\} + \lceil \log_{1/\theta} T \rceil q$, it holds that

$$\left| \sum_{i=1}^{m} a_i H_{i,\rho^\dagger,t} - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1}) \right| \lesssim \iota_\eta + T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + T^{-1}. \tag{3.21}$$

By Eq. (3.20) and the fact that $|Y_t^\star| \le B$, we have

$$\left(Y_t^\star - \varphi^\dagger(H_{\rho^\dagger,t})\right)^2 \le \left(Y_t^\star - \left(\sum_{i=1}^{m} a_i H_{i,\rho^\dagger,t} + H_{w_h,\rho^\dagger,t} - B\right)\right)^2.$$

Moreover, by Eq. (3.32), we have $H_{w_h,\rho^\dagger,t} = \psi^\dagger(Y_{(t-1:t-p)}, X_{t-1}) + B$. Therefore, using the facts that $w_h \asymp T^{\alpha_2} \log T$ and letting $t_1 = C \log T \ge \max\{p, q\} + \lceil \log_{1/\theta} T \rceil q$ for some

sufficiently large constant C, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=p+1}^{T}\big(Y_t^\star-\varphi^\dagger(H_{\rho^\dagger,t})\big)^2\\
&=\frac{1}{T}\sum_{t=t_1+p+1}^{T}\big(Y_t^\star-\varphi^\dagger(H_{\rho^\dagger,t})\big)^2+\frac{1}{T}\sum_{t=p+1}^{t_1+p+1}\big(Y_t^\star-\varphi^\dagger(H_{\rho^\dagger,t})\big)^2\\
&\le\frac{1}{T}\sum_{t=t_1+p+1}^{T}\Bigg(\bigg(\sum_{i=1}^{m}a_iH_{i,\rho^\dagger,t}-\eta^\star(Y_{(t-1:t-p)},\varepsilon_{(t-1:t-q)},X_{t-1})\bigg)\\
&\qquad+\Big(\psi^\dagger(Y_{(t-1:t-p)},X_{t-1})-\psi^\star(Y_{(t-1:t-p)},X_{t-1})\Big)\Bigg)^2+\frac{1}{T}\sum_{t=p+1}^{t_1+p+1}\big(Y_t^\star-\varphi^\dagger(H_{\rho^\dagger,t})\big)^2\\
&\lesssim\Big(\iota_\eta+T^{-\alpha_1\cdot\frac{2\beta}{p+k}}+T^{-1}\Big)^2+T^{-\alpha_1\cdot\frac{4\beta}{p+k}}+T^{-1}\log T\lesssim\iota_\eta^2+T^{-\alpha_1\cdot\frac{4\beta}{p+k}}+T^{-1}\log T.
\end{aligned}
$$

Therefore, the inequality (III) follows.

3.5.2.4 Combining (I)-(III)

Building on (I)-(III), and substituting $(\hat{\rho},\hat{W}_h,\hat{v}_h,\hat{\varphi})$ in Eq. (3.12) with $(\rho^\dagger,W_h^\dagger,v_h^\dagger,\varphi^\dagger)$, we have with probability at least $1-C\exp(-\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\}\log^5 T)$,

$$
\begin{aligned}
&T^{-1}\sum_{t=p+1}^{T}(Y_t^\star-\hat{Y}_t)^2\\
&\lesssim\Bigg(T^{-1}\sum_{t=p+1}^{T}\mathbb{E}\big(Y_t^\star-\hat{\varphi}(H_{\hat{\rho},t})\big)^2\Bigg)^{1/2}(\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T\\
&\qquad+\max\{T^{2\alpha_1+\alpha_2},T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T+\iota_\eta^2+T^{-\alpha_1\cdot\frac{4\beta}{p+k}}.
\end{aligned}
\tag{3.22}
$$

Given that $T^{-1}\sum_{t=p+1}^{T}(Y_t^\star - \hat{Y}_t)^2 \le 16B^2$, it follows that $T^{-1}\sum_{t=p+1}^{T}\mathbb{E}(Y_t^\star - \hat{Y}_t)^2$ is asymptotically bounded by

$$\begin{aligned}
&\left(T^{-1}\sum_{t=p+1}^{T}\mathbb{E}\big(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t})\big)^2\right)^{1/2}(\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&\quad + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T + \iota_\eta^2 + T^{-\alpha_1\cdot\frac{4\beta}{p+k}} \\
&\quad + C\exp(-\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T)\cdot 16B^2.
\end{aligned}$$

Since

$$\exp(-\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T) = o(\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T),$$

we thus obtain

$$\begin{aligned}
&T^{-1}\sum_{t=p+1}^{T}\mathbb{E}(Y_t^\star - \hat{Y}_t)^2 \\
&\lesssim \left(T^{-1}\sum_{t=p+1}^{T}\mathbb{E}\big(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t})\big)^2\right)^{1/2}(\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&\qquad + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T + \iota_\eta^2 + T^{-\alpha_1\cdot\frac{4\beta}{p+k}}.
\end{aligned}$$

Using the fact that $x \le a\sqrt{x} + b$ implies $x \le a^2 + b$, we obtain the following bound:

$$T^{-1}\sum_{t=p+1}^{T}\mathbb{E}(\hat{Y}_t - Y_t^\star)^2 \lesssim \iota_\eta^2 + T^{-\alpha_1\cdot\frac{4\beta}{p+k}} + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T. \tag{3.23}$$

Applying to Eq. (3.22), with probability at least $1-C\exp(-\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T)$,

$$T^{-1}\sum_{t=p+1}^{T}(\hat{Y}_t - Y_t^\star)^2 \lesssim \iota_\eta^2 + T^{-\alpha_1\cdot\frac{4\beta}{p+k}} + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T. \tag{3.24}$$

3.5.2.5 Generalization Performance

Based on Eq. (3.14), it holds that $|\hat{\varphi}(H_{\hat{\rho},T+h}) - \hat{\varphi}(H^\ell_{\hat{\rho},T+h})| \lesssim T^{-1}$ for some sufficiently large T. Therefore, to prove Theorem 9, it suffices to show that $\mathbb{E}(Y^\star_{T+h} - \hat{\varphi}(H^\ell_{\hat{\rho},T+h}))^2 \lesssim \iota_\eta^2 + T^{-\alpha_1\cdot\frac{4\beta}{p+k}} + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T$. For any fixed $(\overline{\rho}_j, \overline{W}_{h,j}, \overline{v}_{h,j}, \overline{\varphi}_j) \in \mathcal{F}^{w_h}_{rnn,\delta}$ with $1 \le j \le |\mathcal{F}^{w_h}_{rnn,\delta}|$, the process $(Y_t^\star - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,t}))^2$ is stationary when $t \ge p+\ell+1$. Therefore, we have:

$$\mathbb{E}(Y_t^\star - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,t}))^2 = \mathbb{E}(Y^\star_{T+h} - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,T+h}))^2, \quad \forall p+\ell+1 \le t \le T.$$

Let $D_{t,j} := (Y_t^\star - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,t}))^2 - \mathbb{E}(Y^\star_{T+h} - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,T+h}))^2$. It holds that $\mathbb{E}D_{t,j} = 0$, $|D_{t,j}| \le 4B^2$, and

$$\mathbb{E}D_{t,j}^2 \le \mathbb{E}(Y_t^\star - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,t}))^4 \le 4B^2\mathbb{E}(Y_t^\star - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,t}))^2.$$

Using the same methods as in Lemma 45, it can be shown that $\{D_{t,j}\}_{t=p+\ell+1}^T$ is a strongly mixing sequence with

$$\alpha_{d,j}(k) \le \bar{\alpha}_{d,j}\exp(-c_{d,j}k^\kappa),$$

where $\bar{\alpha}_{d,j} = \max(\bar{\alpha}_y, 1)\exp(c_y)$ and $c_{d,j} = c_y(\ell+p+1)^{-\kappa}$. Following a similar line of reasoning as in the proof of (I), with probability at least $1-C\exp(-c\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T)$,

$$\left|\frac{1}{T}\sum_{t=p+1}^{T} D_{t,j}\right| \lesssim \left(\frac{1}{T}\sum_{t=p+\ell+1}^{T}\mathbb{E}(Y_t^\star - \overline{\varphi}_j(H^\ell_{\overline{\rho}_j,t}))^2\right)^{1/2} (\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2} \times T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T. \tag{3.25}$$

By applying the union bound inequality to the result above, we can choose the constant c sufficiently large so that,

$$|\mathcal{F}^{w_h}_{rnn,\delta}|\exp(-c\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T) \le \exp(-\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T),$$

we conclude that Eq. (3.25) holds uniformly for all $1 \le j \le |\mathcal{F}^{w_h}_{rnn,\delta}|$. Moreover, there exists $j^\star$ such that when $t \ge p+\ell+1$, $|\hat{\varphi}(H_{\hat{\rho},t}) - \overline{\varphi}_{j^*}(H_{\overline{\rho}_{j^*},t})| \le T^{-1}$ and thus $|\hat{\varphi}(H^\ell_{\hat{\rho},t}) - \overline{\varphi}_{j^*}(H^\ell_{\overline{\rho}_{j^*},t})| \lesssim T^{-1}$.

Therefore, with probability at least $1 - C\exp(-c\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}\log^5 T)$,

$$\begin{aligned}
&\left|T^{-1}\sum_{t=p+1}^{T}\left[(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t}))^2 - \mathbb{E}(Y^\star_{T+h} - \hat{\varphi}(H_{\hat{\rho},T+h}))^2\right]\right| \\
\lesssim &\left|T^{-1}\sum_{t=p+\ell+1}^{T}\left[(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t}))^2 - \mathbb{E}(Y^\star_{T+h} - \hat{\varphi}(H_{\hat{\rho},T+h}))^2\right]\right| + T^{-1+\alpha_2}\log^3 T \\
\lesssim &\left|T^{-1}\sum_{t=p+\ell+1}^{T}\left[(Y_t^\star - \hat{\varphi}(H^\ell_{\hat{\rho},t}))^2 - \mathbb{E}(Y^\star_{T+h} - \hat{\varphi}(H^\ell_{\hat{\rho},T+h}))^2\right]\right| + T^{-1} + T^{-1+\alpha_2}\log^3 T \\
\lesssim &\left(T^{-1}\sum_{t=p+\ell+1}^{T}\mathbb{E}\big(Y_t^\star - \hat{\varphi}(H_{\hat{\rho},t})\big)^2\right)^{1/2} (\max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\})^{1/2}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)/2}\log^4 T \\
&+ \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2}\log T\}T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)}\log^9 T.
\end{aligned}$$

In both the first and last inequalities, we use the fact that $|\hat{\varphi}(H_{\hat{\rho},t}) - \hat{\varphi}(H^{\ell}_{\hat{\rho},t})| \lesssim T^{-1}$ when $t \geq p + \ell + 1$. Combining this with Eq. (3.23) and Eq. (3.24), we confirm that

$$\mathbb{E}(Y^{\star}_{T+h} - \hat{\varphi}(H_{\hat{\rho},T+h}))^2 \lesssim T^{-\alpha_1 \cdot \frac{4\beta}{p+k}} + \iota_{\eta}^2 + \max\{T^{2\alpha_1+\alpha_2}, T^{3\alpha_2} \log T\} T^{-\frac{\kappa}{\kappa+1}(1-\alpha_2)} \log^9 T,$$

thereby completing the proof.

3.5.3 Proof of Corollary 5

Note that linear (piecewise linear) function $\eta^{\star}$ is always Lipschitz continuous, then by Lemma 48, we have

$$\iota_{\eta} \lesssim T^{-w_h}.$$

By the results in Theorem 9 and note that $w \asymp T^{\frac{p+k}{2(2\beta+p+k)} \cdot \frac{\kappa}{1+\kappa}}$ and $w_h \asymp d \asymp \log T$, we have

$$\begin{aligned}\mathbb{E}(Y^{\star}_{T+h} - \hat{\varphi}(H_{\hat{\rho},T+h}))^2 &\lesssim T^{-\frac{p+k}{2(2\beta+p+k)} \cdot \frac{\kappa}{1+\kappa} \cdot \frac{4\beta}{p+k}} + T^{-\log T} + T^{\frac{p+k}{2\beta+p+k} \cdot \frac{\kappa}{1+\kappa}} T^{-\frac{\kappa}{\kappa+1}} \log^{10} T \\ &\sim T^{-\frac{2\beta}{2\beta+p+k} \cdot \frac{\kappa}{1+\kappa}} \log^{10} T,\end{aligned}$$

which completes the proof.

3.6 Some useful Lemmas

Lemma 40. *Let $\{Y_t\}_{t=-\infty}^{\infty}$ be a stationary α-mixing process on the probability space $(\Omega, \mathcal{F}, P)$ with the mixing coefficient satisfying $\alpha(j) \leq \bar{\alpha} \exp(-cj^{\kappa})$. Let an integer $T \geq 1$ be given. Assume f is some real-valued Borel measurable function and $|f(Y_t)| \leq C$ and that $E[f(Y_t)] = 0$. Set*

$$T^{(\alpha)} = \left\lfloor T \left\lceil \left(\frac{8T}{c} \right)^{1/(\kappa+1)} \right\rceil^{-1} \right\rfloor.$$

Then for all $T^{(\alpha)} \geq 2$, $\tau \in \mathbb{R}^+$, *and* $0 \leq \zeta \leq C^{-1}$,

$$\mathrm{P}\left(\frac{1}{T}\sum_{t=1}^{T} f(Y_t) \geq \frac{\tau}{3\zeta T^{(\alpha)}} + \frac{3\zeta \mathbb{E}(f(Y_t))^2}{2(1-\zeta C)}\right) \leq (1 + 4\exp(-2)\bar{\alpha})\exp(-\tau).$$

Proof. Please refer to Theorem 4.3 in Modha and Masry [1996]. □

Definition 7. *Denote the function space* $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C)$ *as follows:*

$$\mathcal{F}_{n_0}^{n_{d+1}}(d, w, C) := \left\{ f \text{ of the form } (3.5) : \max_{j=0,\ldots,d} \left\|W_j\right\|_\infty \vee \left|v_j\right|_\infty \leq C \right\}.$$

Lemma 41. *For any DNN* $\varphi \in \mathcal{F}_{n_o}^{n_{d+1}}(d, w, T^{5\beta+5})$, *it holds that*

$$\|\varphi(x) - \varphi(y)\|_\infty \leq n_0 T^{(5\beta+5)(d+1)} w^d \|x - y\|_\infty.$$

Proof. Write $\varphi = W_d \sigma_{v_d} W_{d-1} \sigma_{v_{d-1}} \ldots W_1 \sigma_{v_1} W_0 x$ where $W_i \in \mathbb{R}^{n_{i+1} \times n_i}$. It is straightforward to verify that $\|W_i x - W_i y\|_\infty \leq n_i T^{5\beta+5} \|x - y\|_\infty$ and $\|\sigma_{v_i}(x) - \sigma_{v_i}(y)\|_\infty \leq \|x - y\|_\infty$. Note that for two Lipschitz functions f_1 and f_2 satisfying $\|f_1(x) - f_1(y)\|_\infty \leq L_1 \|x - y\|_\infty$ and $\|f_2(x) - f_2(y)\|_\infty \leq L_2 \|x - y\|_\infty$, it holds that $\|f_2 \circ f_1(x) - f_2 \circ f_1(y)\|_\infty \leq L_1 L_2 \|x - y\|_\infty$. By repeatedly applying this property to $W_d \sigma_{v_d} W_{d-1} \sigma_{v_{d-1}} \ldots W_1 \sigma_{v_1} W_0 x$, we conclude that

$$\|\varphi(x) - \varphi(y)\|_\infty \leq T^{(5\beta+5)(d+1)} \Big(\prod_{i=0}^{d} n_i \Big) \|x - y\|_\infty \leq n_0 T^{(5\beta+5)(d+1)} w^d \|x - y\|_\infty,$$

where the last inequality holds due to $n_1, \ldots, n_d \leq w$. □

Definition 8 (Covering number and Metric Entropy)**.** *For any function class* $\mathcal{F}$ *with equipped norm* $\|\cdot\|$, *we say that* $\mathcal{F}'$ *is a* δ*-cover for* $\mathcal{F}$ *if, for all* $\varphi \in \mathcal{F}$, *there exists* $\varphi' \in \mathcal{F}'$ *such that* $\|\varphi - \varphi'\| \leq \delta$. *The* δ*-covering number of* $\mathcal{F}$, *denoted* $\mathcal{N}(\delta, \|\cdot\|, \mathcal{F})$ *is the size of the smallest* δ*-covering. The metric entropy is the log covering number* $\log \mathcal{N}(\delta, \|\cdot\|, \mathcal{F})$.

Lemma 42. *For the function class $\mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5}, B)$ with width vector $n = (n_0, \ldots, n_{d+1})$, assume the total number of non-zero weights are bounded by S. There exists a function class $\mathcal{F}_{n_0,\delta}^{n_{d+1}} \subset \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ such that*
(a) $|\mathcal{F}_{n_0,\delta}^{n_{d+1}}| \leq \left(8\delta^{-1}CT^{(5\beta+5)(d+1)}(1+w)^d n_0 d\right)^{4(w^2 d + n_0 w)}$
(b) For any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5}, 3B)$, there exists $\overline{\varphi} \in \mathcal{F}_{n_0,\delta}^{n_{d+1}}$ satisfying $\|\varphi(x) - \overline{\varphi}(x)\|_\infty \leq \delta$ whenever $\|x\|_\infty \leq C$.

Proof. The proof follows the same idea as Lemma 5 of Schmidt-Hieber [2020]. For any $\varphi \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5}, 3B)$, write

$$\varphi(x) = W_d \sigma_{v_d} W_{d-1} \sigma_{v_{d-1}} \cdots W_1 \sigma_{v_1} W_0 x.$$

Define $A_k^+\varphi : [-C, C]^{n_0} \to \mathbb{R}^{n_k}$ and $A_k^-\varphi : \mathbb{R}^{n_{k-1}} \to \mathbb{R}^{n_{d+1}}$ as

$$\begin{aligned} A_k^+\varphi(x) &= \sigma_{v_k} W_{k-1} \sigma_{v_{k-1}} \cdots W_1 \sigma_{v_1} W_0 x, \quad \text{for } k = 1, \ldots, d, \\ A_k^-\varphi(x) &= W_d \sigma_{v_d} W_{d-1} \cdots W_k \sigma_{v_k} W_{k-1} x, \quad \text{for } k = 1, \ldots, d+1. \end{aligned}$$

By convention, we set $A_0^+\varphi(x) = A_{d+2}^+\varphi(x) = x$. Considering that all the parameters of φ are bounded by $T^{5\beta+5}$ and that $\|x\|_\infty \leq C$, it can by proven by mathematical induction that

$$\|A_k^+(x)\|_\infty \leq T^{(5\beta+5)(k+1)} C \prod_{\ell=0}^{k-1} (n_\ell + 1). \tag{3.26}$$

Additionally, by using Lemma 41, we have

$$|A_k^-\varphi(x) - A_k^-\varphi(y)| \leq T^{(5\beta+5)(d-k+2)} \left(\prod_{\ell=k-1}^{d} n_\ell \right) \|x - y\|_\infty. \tag{3.27}$$

Let us fix some $\varepsilon > 0$. Suppose that $\varphi, \overline{\varphi} \in \mathcal{F}_{n_0}^{n_{d+1}}(d, w, T^{5\beta+5})$ be two neural networks such

that all parameters are at most ε away from each other. Denote the parameters of φ by $(v_k, W_k)_k$ and the parameters of $\overline{\varphi}$ by $(\overline{v}_k, \overline{W}_k)_k$. Note that

$$\begin{aligned}\|\varphi(x) - \overline{\varphi}(x)\|_\infty &\le \sum_{k=1}^{d+1} \left\| A_{k+1}^- \varphi\left(\sigma_{v_k} W_{k-1} A_{k-1}^+ \overline{\varphi}(x)\right) - A_{k+1}^- \varphi\left(\sigma_{\overline{v}_k} \overline{W}_{k-1} A_{k-1}^+ \overline{\varphi}(x)\right)\right\|_\infty \\ &\le 4\varepsilon T^{(5\beta+5)(d+1)} C(1+w)^d n_0 d,\end{aligned}$$

where we apply triangle inequality in the first inequality, use Eq. (3.27) in the second inequality, use the facts that $\|\sigma_v(x) - \sigma_v(y)\|_\infty \le \|x-y\|_\infty$ and $\|\sigma_v(x) - \sigma_{\overline{v}}(x)\|_\infty \le \|v - \overline{v}\|_\infty$ in the third inequality, and Eq. (3.26) in the last inequality.

Upon selecting ε as $\frac{\delta}{4T^{(5\beta+5)(d+1)} C(1+w)^d n_0 d}$, it becomes evident that $\|\varphi(x) - \overline{\varphi}(x)\|_\infty \le \delta$. Consequently, by discretizing the parameters with grid size ε for neural networks φ, we form a function class $\mathcal{F}_{n_0,\delta}^{n_{d+1}}$ satisfying (b). To verify (a), note that the total number of parameters is bounded by $\sum_{\ell=0}^{d} (n_\ell + 1)\, n_{\ell+1} \le 4(w^2 d + n_0 w)$. Since all the parameters are bounded in absolute value by $T^{5\beta+5}$, we can discretize the non-zero parameters with grid size ε and obtain for the covering number

$$|\mathcal{F}_{n_0,\delta}^{n_{d+1}}| \le (8\delta^{-1} C T^{(5\beta+5)(d+1)} (1+w)^d n_0 d)^{4(w^2 d + n_0 w)}. \qquad \square$$

Lemma 43. *Assume $f \in \mathcal{H}^\beta([-a,a]^r, B)$ for some $\beta = q + s$, $q \in \mathbb{N}_0$ and $s \in (0,1]$, and $B > 0$. Then, for $a \ge 1$ and ζ sufficiently large (depending only on fixed constants including β, a, B and $\|f\|_{C^q}$), when*

$$\begin{aligned} d &\asymp \left\lceil \log_4\left(\zeta^{2\beta}\right)\right\rceil \cdot (\lceil \log_2(\max\{q, r\} + 1\})\rceil + 1), \text{ and} \\ w &\asymp 2^r \cdot \binom{r+q}{r} \cdot r^2 \cdot (q+1) \cdot \zeta^r, \end{aligned}$$

there exists a neural network $\hat{f}_{wide} \in \mathcal{F}_r^1(d, w, \zeta^{5\beta+5}, B)$ *such that*

$$\left\| f - \hat{f}_{wide} \right\|_\infty \leq C\zeta^{-2\beta}, \tag{3.28}$$

for some constant C *depends only on fixed parameters.*

Proof. The proof primarily follows the same argument as in Kohler and Langer [2021], with the key difference being the necessity to construct the neural networks used in their paper with bounded weights. For a detailed proof, please refer to Shen and Xiu [2024]. □

Lemma 44. *When* $\varepsilon < B$ *and* $\ell \geq \max(\varepsilon^{-1}, 2)$*, for* $H_{\rho,t}^\ell$ *defined in Eq. (3.13), there exists a constant* C *such that*

(i) $\|H_{\rho,t}\|_\infty, \|H_{\rho,t}^\ell\|_\infty \leq C\left(\varepsilon^{-1}T^{5\beta+5}\right)^{w_h}$, $\forall t \geq p$.

(ii) $\|H_{\rho,t}^\ell - H_{\rho,t}\|_\infty \leq C(1-\varepsilon)^{\ell - w_h}(2\ell T^{5\beta+5})^{w_h}$, $\forall t \geq p + \ell + 1$.

Proof. First, we prove (i). Observe that

$$\begin{aligned}
|H_{\rho,t}| =& |\sigma_{v_h}(\rho(Y_{t-1}, \ldots, Y_{t-p}, X_{t-1}) + W_h H_{\rho,t-1})| \\
\leq& |v_h| + |\rho(Y_{t-1}, \ldots, Y_{t-p}, X_{t-1})| + |W_h H_{\rho,t-1}| \leq T^{5\beta+5} + B + |W_h H_{\rho,t-1}|.
\end{aligned}$$

Here $|\cdot|$ represents the elementwise absolute value, $\leq$ denotes elementwise smaller, and the last inequality is due to Eq. (3.7). Let $H_{k,\rho,t}$ represent the k-th element of $H_{\rho,t}$. Using Eq. (3.6), as long as $T^{5\beta+5} \geq B$, we have

$$\begin{aligned}
|H_{w_h,\rho,t}| &\leq 2T^{5\beta+5} + (1-\varepsilon)|H_{w_h,\rho,t-1}|, \\
|H_{w_h-1,\rho,t}| &\leq 2T^{5\beta+5} + (1-\varepsilon)|H_{w_h-1,\rho,t-1}| + T^{5\beta+5}|H_{w_h,\rho,t-1}|, \\
&\vdots \\
|H_{1,\rho,t}| &\leq 2T^{5\beta+5} + (1-\varepsilon)|H_{1,\rho,t-1}| + T^{5\beta+5}\sum_{i=2}^{w_h}|H_{i,\rho,t-1}|.
\end{aligned}$$

From the first inequality, given that $H_{\rho,p} = 0_{w_h}$, it is straightforward to verify that $|H_{w_h,\rho,t}| \leq 2T^{5\beta+5}\varepsilon^{-1}$ for all $t \geq p$. Consequently, from the second inequality, we obtain $|H_{w_h-1,\rho,t}| \leq 2T^{5\beta+5}\varepsilon^{-1} + 2T^{5\beta+5} \cdot T^{5\beta+5}\varepsilon^{-2}$. Using the same argument, it can be shown

$$|H_{i,\rho,t}| \leq 2T^{5\beta+5}\left(\varepsilon^{-1} + T^{5\beta+5}\varepsilon^{-2} + \cdots + (T^{5\beta+5})^{w_h-i}\varepsilon^{-(w_h-i+1)}\right).$$

Note that

$$\begin{aligned}
&2T^{5\beta+5}\left(\varepsilon^{-1} + T^{5\beta+5}\varepsilon^{-2} + \cdots + (T^{5\beta+5})^{w_h-i}\varepsilon^{-(w_h-i+1)}\right) \\
=&2T^{5\beta+5}\varepsilon^{-1}\frac{\left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h-i+1} - 1}{T^{5\beta+5}\varepsilon^{-1} - 1} \leq 4\left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h-i+1}.
\end{aligned}$$

Therefore, we conclude that $|H_{i,\rho,t}| \leq 4\left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h-i+1}$ for all $t \geq p$ and $1 \leq i \leq w_h$. Similarly, it can be shown that $|H^{\ell}_{i,\rho,t}| \leq 4\left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h-i+1}$. Then (i) is verified.

Now, we study (ii). Note that $|\sigma_{v_h}(x) - \sigma_{v_h}(y)| \leq |x-y|$ for any $v, x, y \in \mathbb{R}^{w_h}$, where all the operators are applied elementwise. Therefore, for $\ell \geq 1$,

$$|H^{\ell}_{\rho,t} - H_{\rho,t}| \leq |W_h H_{\rho,t-1} - W_h H^{\ell-1}_{\rho,t-1}|.$$

By Eq. (3.6) and the inequality above, we have

$$\begin{aligned}
|H^{\ell}_{w_h,\rho,t} - H_{w_h,\rho,t}| &\leq (1-\varepsilon)|H_{w_h,\rho,t-1} - H^{\ell-1}_{w_h,\rho,t-1}|, \\
&\vdots \\
|H^{\ell}_{1,\rho,t} - H_{1,\rho,t}| &\leq (1-\varepsilon)|H_{1,\rho,t-1} - H^{\ell-1}_{1,\rho,t-1}| + \sum_{i=2}^{w_h} T^{5\beta+5}|H_{i,\rho,t-1} - H^{\ell-1}_{i,\rho,t-1}|. \quad (3.29)
\end{aligned}$$

We then prove $|H^{\ell}_{i,\rho,t} - H_{i,\rho,t}| \leq 8(1-\varepsilon)^{\ell-w_h+i}(2\ell T^{5\beta+5})^{w_h-i+1}$ by mathematical induction.

First, for the base case $i = w_h$, from the first inequality in Eq. (3.29) and the fact that

$|H_{i,\rho,t}|, |H^{\ell}_{i,\rho,t}| \le 4\left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h-i+1}$, when $t > p+\ell$, we have

$$|H^{\ell}_{w_h,\rho,t} - H_{w_h,\rho,t}| \le (1-\varepsilon)|H_{w_h,\rho,t-1} - H^{\ell-1}_{w_h,\rho,t-1}| \le \cdots \le (1-\varepsilon)^{\ell}|H_{w_h,\rho,t-\ell} - H^{0}_{w_h,\rho,t-\ell}|$$
$$\le 8(1-\varepsilon)^{\ell}T^{5\beta+5}\varepsilon^{-1} \le 8(1-\varepsilon)^{\ell}(2\ell T^{5\beta+5})^{1},$$

where we use the fact that $\ell \ge \varepsilon^{-1}$. This means the statement holds for $i = w_h$.

Second, in the inductive step, we assume that the statement is true for $i = k+1, \ldots, w_h$, and we need to prove the statement holds for $i = k$. Observe that for $t > p+\ell$,

$$|H^{\ell}_{k,\rho,t} - H_{k,\rho,t}| \le (1-\varepsilon)|H_{k,\rho,t-1} - H^{\ell-1}_{k,\rho,t-1}| + \sum_{i=k+1}^{w_h} T^{5\beta+5}|H_{i,\rho,t-1} - H^{\ell-1}_{i,\rho,t-1}|$$
$$\le (1-\varepsilon)|H_{k,\rho,t-1} - H^{\ell-1}_{k,\rho,t-1}| + T^{5\beta+5}\sum_{i=k+1}^{w_h} 8(1-\varepsilon)^{\ell-w_h+i-1}(2\ell T^{5\beta+5})^{w_h-i+1}$$
$$\le (1-\varepsilon)^2|H_{k,\rho,t-2} - H^{\ell-2}_{k,\rho,t-2}| + (1-\varepsilon)T^{5\beta+5}\sum_{i=k+1}^{w_h} 8(1-\varepsilon)^{\ell-w_h+i-2}(2\ell T^{5\beta+5})^{w_h-i+1}$$
$$+ T^{5\beta+5}\sum_{i=k+1}^{w_h} 8(1-\varepsilon)^{\ell-w_h+i-1}(2\ell T^{5\beta+5})^{w_h-i+1}$$
$$= (1-\varepsilon)^2|H_{k,\rho,t-2} - H^{\ell-2}_{k,\rho,t-2}| + 2T^{5\beta+5}\sum_{i=k+1}^{w_h} 8(1-\varepsilon)^{\ell-w_h+i-1}(2\ell T^{5\beta+5})^{w_h-i+1}.$$

Iteratively, we obtain

$$|H^{\ell}_{k,\rho,t} - H_{k,\rho,t}|$$
$$\le (1-\varepsilon)^{\ell}|H_{k,\rho,t-\ell} - H^{0}_{k,\rho,t-\ell}| + \ell T^{5\beta+5}\sum_{i=k+1}^{w_h} 8(1-\varepsilon)^{\ell-w_h+i-1}(2\ell T^{5\beta+5})^{w_h-i+1}$$
$$\le 8(1-\varepsilon)^{\ell}\left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h-k+1} + \ell T^{5\beta+5}\sum_{i=k+1}^{w_h} 8(1-\varepsilon)^{\ell-w_h+i-1}(2\ell T^{5\beta+5})^{w_h-i+1}$$
$$=: Q_1 + Q_2.$$

Note that for Q_1, since $\ell \geq \varepsilon^{-1}$ and $w_h - k \geq 1$, it holds that

$$\begin{aligned} Q_1 &= 8(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1}\left(\frac{1-\varepsilon}{2\ell\varepsilon}\right)^{w_h-k} \cdot \frac{1}{2\ell\varepsilon} \\ &\leq 8(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1} \cdot 0.5 \cdot 0.5 \\ &= 2(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1}. \end{aligned}$$

For Q_2, we have

$$\begin{aligned} Q_2 &= 4(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1} \sum_{i=k+1}^{w_h} (1-\varepsilon)^{i-k-1}(2\ell T^{5\beta+5})^{k-i+1} \\ &= 4(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1} \sum_{i=0}^{w_h-k-1} (1-\varepsilon)^{i}(2\ell T^{5\beta+5})^{-i} \\ &\leq 4(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1} \left(1 - \frac{1-\varepsilon}{2\ell T^{5\beta+5}}\right)^{-1} \\ &\leq 4(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1} \cdot \frac{4}{3}, \end{aligned}$$

where in the last inequality we use $\frac{1-\varepsilon}{2\ell T^{5\beta+5}} \leq \frac{1}{4}$ since $\ell \geq 2$. In total, for $t > p + \ell$, we have

$$|H^{\ell}_{k,\rho,t} - H_{k,\rho,t}| \leq Q_1 + Q_2 \leq 8(1-\varepsilon)^{\ell - w_h + k}(2\ell T^{5\beta+5})^{w_h - k+1},$$

which implies (ii). □

Lemma 45. *For the the stationary process $\{D_t\}_{t=p+\ell+1}^{T}$ defined in Eq. (3.15), its α-mixing coefficient satisfies*

$$\alpha_d(j) \leq \max(\bar{\alpha}_y, 1)\exp(c_y)\exp(-c_y(\ell+p+1)^{-\kappa}j^{\kappa}) := \bar{\alpha}_d \exp(-c_d j^{\kappa}),$$

where $\bar{\alpha}_d = \max(\bar{\alpha}_y, 1)\exp(c_y)$ and $c_d = c_y(\ell+p+1)^{-\kappa}$.

Proof. Since $D_t \in \mathfrak{F}_{t-p-\ell}^{t}$, the stationary process $\{D_t\}_{t=p+\ell+1}^{T}$ forms an α-mixing sequence

satisfying

$$\alpha_d(j) \leq \alpha_y(j-p-\ell).$$

By convention, we set $\alpha_y(j) = 1$ for $j \leq 0$. We then consider two scenarios:

(i) $j-p-\ell \geq 1$: Noting that for $1 \leq b \leq a$, we have $\frac{a}{b} + b \leq a+1$. Therefore, by setting $a = j$ and $b = p+\ell+1$, we have $\frac{j}{p+\ell+1} \leq j-p-\ell$, which implies

$$\alpha_d(j) \leq \alpha_y(j-p-\ell) \leq \bar{\alpha}_y \exp(-c_y(j-p-\ell)^\kappa) \leq \bar{\alpha}_y \exp(-c_y(\ell+p+1)^{-\kappa} j^\kappa).$$

(ii) $j-p-\ell \leq 0$: In this case, we have

$$\alpha_y(j-p-\ell) \leq 1 \leq \exp(c_y) \exp(-c_y(\ell+p+1)^{-\kappa} j^\kappa).$$

Combining the results from (i) and (ii), the lemma follows. □

Lemma 46. *For $\mathcal{F}_{rnn}^{w_h}$ defined in Eq. (3.7) satisfying the conditions in Theorem 9, there exists a function class $\mathcal{F}_{rnn,\delta}^{w_h} \subset \mathcal{F}_{rnn}^{w_h}$ for $\delta = T^{-1}$ with*

$$\log |\mathcal{F}_{rnn,\delta}^{w_h}| \lesssim w^2 w_h \log^4(T) + w_h^3 \log^3(T),$$

such that for any $(\rho, W_h, v_h, \varphi) \in \mathcal{F}_{rnn}^{w_h}$, there exists a $(\overline{\rho}, \overline{W}_h, \overline{v}_h, \overline{\varphi}) \in \mathcal{F}_{rnn,\delta}^{w_h}$, satisfying for all $t \gtrsim w_h \log^2 T$, $|\varphi(H_{\rho,t}) - \overline{\varphi}(H_{\overline{\rho},t})| \leq \delta = T^{-1}$.

Proof. Recall $H_{\rho,t}^\ell$ as defined in Eq. (3.13). By Lemma 44 (ii), for $t \geq p+\ell+1$, we have

$$\|H_{\rho,t} - H_{\rho,t}^\ell\|_\infty \leq C(1-\varepsilon)^{\ell - w_h} (2\ell T^{5\beta+5})^{w_h}.$$

Now, assuming for temporarily that

$$\|W_h - \overline{W}_h\|_\infty, \|v_h - \overline{v}_h\|_\infty, \|\rho - \overline{\rho}\|_\infty \leq \xi, \tag{3.30}$$

we obtain

$$\begin{aligned}\|H_{\rho,t} - H_{\overline{\rho},t}\|_\infty &\le \left\|H^\ell_{\rho,t} - H^\ell_{\overline{\rho},t}\right\|_\infty + \left\|H_{\rho,t} - H^\ell_{\rho,t}\right\|_\infty + \left\|H_{\overline{\rho},t} - H^\ell_{\overline{\rho},t}\right\|_\infty \\ &\le \left\|H^\ell_{\rho,t} - H^\ell_{\overline{\rho},t}\right\|_\infty + 2C(1-\varepsilon)^{\ell - w_h}(2\ell T^{5\beta+5})^{w_h}, \quad \forall t \ge p+\ell+1.\end{aligned}$$

By the inequality $|\sigma_v(x) - \sigma_u(y)| \le |u-v| + |x-y|$ for $u, v, x, y \in \mathbb{R}$ and applying Lemma 44 (i), we have

$$\begin{aligned}&\left\|H^\ell_{\rho,t} - H^\ell_{\overline{\rho},t}\right\|_\infty \\ \le& 2\xi + \|W_h H^{\ell-1}_{\overline{\rho},t-1} - \overline{W}_h H^{\ell-1}_{\overline{\rho},t-1}\|_\infty + \|W_h H^{\ell-1}_{\rho,t-1} - W_h H^{\ell-1}_{\overline{\rho},t-1}\|_\infty \\ \le& 2C\xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h} + w_h T^{5\beta+5}\|H^{\ell-1}_{\rho,t-1} - H^{\ell-1}_{\overline{\rho},t-1}\|_\infty,\end{aligned}$$

where the last inequality holds by using the fact that $2\xi \le C\xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h}$, $\|W_h H^{\ell-1}_{\overline{\rho},t-1} - \overline{W}_h H^{\ell-1}_{\overline{\rho},t-1}\|_\infty \le w_h \|W_h - \overline{W}_h\|_\infty \|H^{\ell-1}_{\overline{\rho},t-1}\|_\infty \le C\xi w_h (\varepsilon^{-1}T^{5\beta+5})^{w_h}$ and $\|W_h\|_\infty \le T^{5\beta+5}$.

By repeatedly applying the above inequality, we have

$$\begin{aligned}\left\|H^\ell_{\rho,t} - H^\ell_{\overline{\rho},t}\right\|_\infty \le& 2C\xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h} \sum_{k=0}^{\ell-1} (w_h T^{5\beta+5})^k + (w_h T^{5\beta+5})^\ell \|H^0_{\rho,t-\ell} - H^0_{\overline{\rho},t-\ell}\|_\infty \\ \le& 2C\xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h} (w_h T^{5\beta+5})^\ell + (w_h T^{5\beta+5})^\ell \|H^0_{\rho,t-\ell} - H^0_{\overline{\rho},t-\ell}\|_\infty \\ =& 2C\xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h} (w_h T^{5\beta+5})^\ell + (w_h T^{5\beta+5})^\ell \\ & \cdot \|\sigma_{v_h}(\rho(Y_{t-\ell-1}, \ldots, Y_{t-\ell-p}, X_{t-\ell})) - \sigma_{\overline{v}_h}(\overline{\rho}(Y_{t-\ell-1}, \ldots, Y_{t-\ell-p}, X_{t-\ell}))\|_\infty \\ \le& 2C\xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h} (w_h T^{5\beta+5})^\ell + 2\xi (w_h T^{5\beta+5})^\ell \\ \le& 4C\xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h} (w_h T^{5\beta+5})^\ell.\end{aligned}$$

Therefore, for $t \geq p + \ell + 1$ and $\ell \geq \max(\varepsilon^{-1}, 2)$, we have

$$\begin{aligned}\|H_{\rho,t} - H_{\bar{\rho},t}\|_\infty \lesssim & \xi w_h \left(T^{5\beta+5}\varepsilon^{-1}\right)^{w_h} (w_h T^{5\beta+5})^\ell + (1-\varepsilon)^{\ell - w_h} (2\ell T^{5\beta+5})^{w_h} \\ \leq & (2\ell T^{5\beta+5})^{w_h} \left(\xi w_h (w_h T^{5\beta+5})^\ell + (1-\varepsilon)^{\ell - w_h}\right).\end{aligned}$$

Choosing $\ell = C_1 w_h \log^2 T$ and $\xi = \exp(-C_2 w_h \log^3 T)$. By the fact that $w_h \leq T$, we have

$$\begin{aligned}\xi w_h (w_h T^{5\beta+5})^\ell &= \exp(-C_2 w_h \log^3 T + C_1 w_h \log^2 T \log(w_h T^{5\beta+5}) + \log w_h) \\ &\leq \exp(-C_2 w_h \log^3 T + C_1 w_h \log^2 T \log(T^{5\beta+6}) + \log T) \\ &= \exp(-C_2 w_h \log^3 T + C_1 (5\beta+6) w_h \log^3 T + \log T),\end{aligned}$$

and

$$(1-\varepsilon)^{\ell - w_h} = \exp(C_1 w_h \log^2 T \log(1-\varepsilon) - w_h \log(1-\varepsilon)).$$

From the above two inequalities, it is not hard to check that for any constant C, there exist fixed $C_1, C_2 \geq 0$ such that

$$\xi w_h (w_h T^{5\beta+5})^\ell, (1-\varepsilon)^{\ell - w_h} \leq \exp(-C w_h \log^2 T).$$

As a consequence,

$$\|H_{\rho,t} - H_{\bar{\rho},t}\|_\infty \lesssim (2\ell T^{5\beta+5})^{w_h} \exp(-C w_h \log^2 T).$$

Together with Lemma 41, for $\varphi, \overline{\varphi} \in \mathcal{F}_{w_h}^1(d, w, T^{5\beta+5})$, we have

$$\begin{aligned}
&|\varphi(H_{\rho,t}) - \overline{\varphi}(H_{\overline{\rho},t})| \leq |\varphi(H_{\rho,t}) - \varphi(H_{\overline{\rho},t})| + |\varphi(H_{\overline{\rho},t}) - \overline{\varphi}(H_{\overline{\rho},t})| \\
\lesssim & w_h T^{(5\beta+5)(d+1)} w^d \|H_{\rho,t} - H_{\overline{\rho},t}\|_\infty + \|\varphi - \overline{\varphi}\|_\infty \\
\lesssim & w_h T^{(5\beta+5)(d+1)} w^d (2\ell T^{5\beta+5})^{w_h} \exp(-C w_h \log^2 T) + \|\varphi - \overline{\varphi}\|_\infty.
\end{aligned}$$

Observe that

$$\begin{aligned}
& w_h T^{(5\beta+5)(d+1)} w^d (2\ell T^{5\beta+5})^{w_h} \exp(-C w_h \log^2 T) \\
= & \exp\left(\log(w_h) + (5\beta+5)(d+1)\log(T) + d\log w + w_h \log(2\ell T^{5\beta+5}) - C w_h \log^2 T\right).
\end{aligned}$$

Since $d \asymp \log(T)$, $w \leq T$, $w_h \leq T$, and $\log \ell \lesssim \log T$, it is easy to verify that by properly choosing the constant $C > 0$, it holds that.

$$w_h T^{(5\beta+5)(d+1)} w^d (2\ell T^{5\beta+5})^{w_h} \exp(-C w_h \log^2 T) \lesssim (2T)^{-1}.$$

Therefore, as long as $\|\varphi - \overline{\varphi}\|_\infty \leq (2T)^{-1}$ and Eq. (3.30), we have $|\varphi(H_{\rho,t}) - \overline{\varphi}(H_{\overline{\rho},t})| \leq T^{-1}$ for $t \geq p + C_1 w_h \log^2 T + 1 \asymp w_h \log^2 T$, which meets the requirement. Hence we need at most a total of $\mathcal{N}_1\mathcal{N}_2\mathcal{N}_3\mathcal{N}_4$ functions to achieve a δ-covering of $\mathcal{F}_{rnn}^{w_h}$, where

$$\begin{aligned}
\mathcal{N}_1 &:= \mathcal{N}((2T)^{-1}, \|\cdot\|_{\infty,[-C(\varepsilon^{-1}T^{5\beta+5})^{w_h}, C(\varepsilon^{-1}T^{5\beta+5})^{w_h}]^{w_h}}, \mathcal{F}_{w_h}^1(d, w, T^{5\beta+5}, B)), \\
\mathcal{N}_2 &:= \mathcal{N}(\xi, \|\cdot\|_{\infty,[-B,B]^{p+k}}, \mathcal{F}_{p+k}^{w_h}(d, w, T^{5\beta+5}, 3B)), \\
\mathcal{N}_3 &:= \mathcal{N}(\xi, \|\cdot\|_\infty, \{v_h \in \mathbb{R}^{w_h} : \|v_h\|_\infty \leq T^{5\beta+5}\}), \\
\mathcal{N}_4 &:= \mathcal{N}(\xi, \|\cdot\|_\infty, \{W_h : W_h \text{ of Form Eq. (3.6)}\}).
\end{aligned}$$

By Lemma 42(a) and the fact that $\xi = \exp(-C_2 w_h \log^3 T)$, we have

$$\log \mathcal{N}_1 \lesssim w^2 w_h \log^3(T) \text{ and } \log \mathcal{N}_2 \lesssim w^2 w_h \log^4(T).$$

Moreover, it is straightforward to verify that

$$\log \mathcal{N}_3 \lesssim w_h^2 \log^3(T) \text{ and } \log \mathcal{N}_4 \lesssim w_h^3 \log^3(T).$$

Therefore, we conclude that $\log |\mathcal{F}^{w_h}_{rnn,\delta}| \leq \sum_{i=1}^4 \log \mathcal{N}_i \lesssim w^2 w_h \log^4(T) + w_h^3 \log^3(T)$. □

Lemma 47. *For $(\rho^\dagger, W_h^\dagger, v_h^\dagger, \varphi^\dagger)$ defined in Eq. (3.19) and $I_{r,i} = I(r - \lfloor \frac{r-2}{q} \rfloor q - i > 0)$, it holds that for $1 \leq r \leq \lceil \log_{1/\theta} T \rceil q$ and $t \geq \max\{p, q\} + r$,*

$$\begin{aligned}
&\left| \sum_{i=1}^m a_i H_{w_h - r\tilde{m} + i - 1, \rho^\dagger, t} - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1}) \right| \\
&\leq \left(\sum_{j=0}^{r-2} \theta^j \right) \iota_\eta + \left(\sum_{j=1}^{r-1} \theta^j \right) T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + 2\theta^{\lfloor \frac{r-2}{q} \rfloor + I_{r,1}} B, \\
&\left| H_{w_h - r\tilde{m} + m, \rho^\dagger, t} - \sigma_{-B}(\varepsilon_{t-1}) \right| \leq \left(\sum_{j=0}^{r-3} \theta^j \right) \iota_\eta + \left(\sum_{j=0}^{r-2} \theta^j \right) T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + 2\theta^{\lfloor \frac{r-2}{q} \rfloor + I_{r,2}} B, \\
&\quad \vdots \qquad (3.31) \\
&\left| H_{w_h - (r-1)\tilde{m} - 1, \rho^\dagger, t} - \sigma_{-B}(\varepsilon_{t-q+1}) \right| \leq \left(\sum_{j=0}^{r-q-1} \theta^j \right) \iota_\eta + \left(\sum_{j=0}^{r-q} \theta^j \right) T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + 2\theta^{\lfloor \frac{r-2}{q} \rfloor + I_{r,q}} B,
\end{aligned}$$

where $Y_{(t-1:t-p)} = (Y_{t-1}, \ldots, Y_{t-p})$ and $\varepsilon_{(t-1:t-q)} = (\varepsilon_{t-1}, \ldots, \varepsilon_{t-q})$. By convention, we set $\sum_{j=0}^n \theta^j$ as zero when $n < 0$,.

Proof. Recall that

$$H_{\rho^\dagger, t} = \sigma_{v_h^\dagger} \left(W_h^\dagger H_{\rho^\dagger, t-1} + \rho^\dagger(Y_{(t-1:t-p)}, X_{t-1}) \right).$$

By the definition of $v_h^\dagger$, $W_h^\dagger$ and $\rho^\dagger$ in Eq. (3.19) and note that $|\psi^\dagger(Y_{(t-1:t-p)}, X_{t-1})| \leq B$ which is defined in Eq. (3.17), as well as the definition of $H_{\rho,t}$ in Eq. (3.11), we have

$$\begin{aligned} &H_{w_h,\rho^\dagger,t} = \sigma_{-B}\left(\psi^\dagger(Y_{(t-1:t-p)}, X_{t-1})\right) = \psi^\dagger(Y_{(t-1:t-p)}, X_{t-1}) + B, \text{ for } t > p, \\ &H_{w_h-\tilde{m},\rho^\dagger,t} = H_{w_h-\tilde{m}+1,\rho^\dagger,t} = \cdots = H_{w_h-1,\rho^\dagger,t} = 0. \end{aligned} \tag{3.32}$$

Moreover, define $\tilde{b}_i = (b_{i,p+q+1}, \ldots, b_{i,p+q+k}) \in \mathbb{R}^k$, for any $2 \leq r \leq \lceil \log_{1/\theta} T \rceil q$ and $t > p+1$, by Eq. (3.32), we have

$$\begin{aligned} &H_{w_h-r\tilde{m}+i-1,\rho^\dagger,t} \\ =&\sigma_{v_i}\Bigg(b_{i,p+1}\Big(Y_{t-1} + B - (\psi^\dagger(Y_{(t-2:t-p-1)}, X_{t-2}) + B) - \sum_{j=1}^{m} a_j H_{w_h-(r-1)\tilde{m}+j-1,\rho^\dagger,t-1}\Big) \\ &+ b_{i,p+2}(-H_{w_h-(r-1)\tilde{m}+m,\rho^\dagger,t-1} + B) + \cdots + b_{i,p+q}(-H_{w_h-(r-2)\tilde{m}-1,\rho^\dagger,t-1} + B) \\ &+ b_{i,1}Y_{t-1} + \cdots + b_{i,p}Y_{t-p} + \tilde{b}_i^\top X_{t-1}\Bigg), \end{aligned} \tag{3.33}$$

for $1 \leq i \leq m$,

$$H_{w_h-r\tilde{m}+m,\rho^\dagger,t} = \sigma_{-B}\left(-\sum_{j=1}^{m} a_j H_{w_h-(r-1)\tilde{m}+j-1,\rho^\dagger,t-1} - H_{w_h,\rho^\dagger,t-1} + Y_{t-1} + B\right) \tag{3.34}$$

$$=\sigma_{-B}\left(Y_{t-1} + B - (\psi^\dagger(Y_{(t-2:t-p-1)}, X_{t-2}) + B) - \sum_{j=1}^{m} a_j H_{w_h-(r-1)\tilde{m}+j-1,\rho^\dagger,t-1}\right),$$

and

$$H_{w_h-r\tilde{m}+m+i,\rho^\dagger,t} = \sigma_{-B}\left(H_{w_h-(r-1)\tilde{m}+m+i-1,\rho^\dagger,t-1} - B\right), \tag{3.35}$$

for $1 \leq i \leq q-2$.

In the following, we prove Eq. (3.31). First, we prove the case $r = 1$. By Eq. (3.32), for $t \geq q + 1$,

$$\begin{aligned}&\left|\sum_{i=1}^{m} a_i H_{w_h-\tilde{m}+i-1,\rho^\dagger,t} - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\right| \\ &= \left|\eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\right| \leq 2B,\end{aligned}$$

and

$$\begin{aligned}&\left|H_{w_h-\tilde{m}+m,\rho^\dagger,t} - \sigma_{-B}(\varepsilon_{t-1})\right| = |\sigma_{-B}(\varepsilon_{t-1})| \leq 2B, \\ &\qquad\vdots \\ &\left|H_{w_h-1,\rho^\dagger,t} - \sigma_{-B}(\varepsilon_{t-q+1})\right| = \left|\sigma_{-B}(\varepsilon_{t-q+1})\right| \leq 2B.\end{aligned}$$

Therefore, when $r = 1$, Eq. (3.31) holds.

Next, we use mathematical induction to prove the conclusion. Assume Eq. (3.31) holds for some $r < \lceil \log_{1/\theta} \tilde{m} \rceil q$. We prove the case for $r + 1$.

Define $(\psi^\dagger - \psi^\star)(x) := \psi^\dagger(x) - \psi^\star(x)$ and $\widetilde{\varepsilon}_{(t-1:t-q)} := (\widetilde{\varepsilon}_{t-1}, \cdots, \widetilde{\varepsilon}_{t-q})$ where

$$\begin{aligned}\widetilde{\varepsilon}_{t-1} &= (\psi^\star - \psi^\dagger)(Y_{(t-2:t-p-1)}, X_{t-2}) + \varepsilon_{t-1} + \eta^\star(Y_{(t-2:t-p-1)}, \varepsilon_{(t-2:t-q-1)}, X_{t-2}) \\ &\quad - \sum_{j=1}^{m} a_j H_{w_h-r\tilde{m}+j-1,\rho^\dagger,t-1} \\ \widetilde{\varepsilon}_{t-2} &= -H_{w_h-r\tilde{m}+m,\rho^\dagger,t-1} + B, \ldots, \widetilde{\varepsilon}_{t-q} = -H_{w_h-(r-1)\tilde{m}-1,\rho^\dagger,t-1} + B.\end{aligned}$$

By Eq. (3.33), we have

$$
\begin{aligned}
&\left|\sum_{i=1}^{m} a_i H_{w_h-(r+1)\tilde{m}+i-1,\rho^\dagger,t} - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\right| \\
&= \Bigg|\eta^\dagger\Bigg((\psi^\star - \psi^\dagger)(Y_{(t-2:t-p-1)}, X_{t-2}) + \varepsilon_{t-1} + \eta^\star(Y_{(t-2:t-p-1)}, \varepsilon_{(t-2:t-q-1)}, X_{t-2}) \\
&\quad - \sum_{j=1}^{m} a_j H_{w_h-r\tilde{m}+j-1,\rho^\dagger,t-1}, -H_{w_h-r\tilde{m}+m,\rho^\dagger,t-1} + B, \cdots, -H_{w_h-(r-1)\tilde{m}-1,\rho^\dagger,t-1} + B, \\
&\qquad Y_{t-1}, \cdots, Y_{t-p}, X_{t-1}\Bigg) - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\Bigg| \\
&=: \left|\eta^\dagger\left(Y_{(t-1:t-p)}, \widetilde{\varepsilon}_{(t-1:t-q)}, X_{t-1}\right) - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\right|, \qquad (3.36)
\end{aligned}
$$

where $\eta^\dagger(x) = \sum_{i=1}^{m} a_i \sigma_{v_i}\left(b_i^\top x\right)$ is defined in Eq. (3.18). Define $(\eta^\dagger - \eta^\star)(x) := \eta^\dagger(x) - \eta^\star(x)$. By Assumption 8, we have $|\varepsilon_t|, \|Y_{(t-1:t-p)}\|_\infty, \|X_t\|_\infty \le B$. In addition, by Eq. (3.31) for r and Eq. (3.17), note that $\iota_\eta \to 0$ and $T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} \to 0$ when $T \to \infty$, we have

$$
|\widetilde{\varepsilon}_{t-1}| \le T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + B + \left(\sum_{j=0}^{r-2} \theta^j\right) \iota_\eta + \left(\sum_{j=1}^{r-1} \theta^j\right) T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + 2\theta^{\lfloor \frac{r-2}{q} \rfloor + I_{r,1}} B \le 3B,
$$

and

$$
\begin{aligned}
|\widetilde{\varepsilon}_{t-2}| &= \left|H_{w_h-r\tilde{m}+m,\rho^\dagger,t-1} - \sigma_{-B}(\varepsilon_{t-2}) + \varepsilon_{t-2}\right| \\
&\le \left(\sum_{j=0}^{r-3} \theta^j\right) \iota_\eta + \left(\sum_{j=0}^{r-2} \theta^j\right) T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + 2\theta^{\lfloor \frac{r-2}{q} \rfloor + I_{r,2}} B + B \le 3B, \\
&\vdots \\
|\widetilde{\varepsilon}_{t-q}| &\le \left|H_{w_h-(r-1)\tilde{m}-1,\rho^\dagger,t-1} - \sigma_{-B}(\varepsilon_{t-q}) + \varepsilon_{t-q}\right| \\
&\le \left(\sum_{j=0}^{r-q-1} \theta^j\right) \iota_\eta + \left(\sum_{j=0}^{r-q} \theta^j\right) T^{-\alpha_1 \cdot \frac{2\beta}{p+k}} + 2\theta^{\lfloor \frac{r-2}{q} \rfloor + I_{r,q}} B + B \le 3B.
\end{aligned}
$$

Then we have $\widetilde{\varepsilon}_{(t-1:t-q)}, Y_{(t-1:t-p)}, X_{t-1} \in [-3B, 3B]^{p+q+k}$. Applying Eq. (3.18), Eq. (3.31) for r, and triangular inequality to the right hand side of Eq. (3.36) leads to

$$
\begin{aligned}
&\left|\sum_{i=1}^{m} a_i H_{w_h-(r+1)\tilde{m}+i-1,\rho^\dagger,t} - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\right| \\
&\leq \left|(\eta^\dagger - \eta^\star)\left(Y_{(t-1:t-p)}, \widetilde{\varepsilon}_{(t-1:t-q)}, X_{t-1}\right)\right| \\
&\quad + \left|\eta^\star\left(Y_{(t-1:t-p)}, \widetilde{\varepsilon}_{(t-1:t-q)}, X_{t-1}\right) - \eta^\star(Y_{(t-1:t-p)}, \varepsilon_{(t-1:t-q)}, X_{t-1})\right| \\
&\leq \iota_\eta + \theta_1\Big|(\psi^\star - \psi^\dagger)(Y_{(t-2:t-p-1)}, X_{t-2}) + \eta^\star(Y_{(t-2:t-p-1)}, \varepsilon_{(t-2:t-q-1)}, X_{t-2}) \\
&\quad - \sum_{j=1}^{m} a_j H_{w_h-r\tilde{m}+j-1,\rho^\dagger,t-1}\Big| + \sum_{i=2}^{q}\theta_i\left|H_{w_h-r\tilde{m}+m+i-2,\rho^\dagger,t-1} - \varepsilon_{t-i} - B\right| \\
&\leq \iota_\eta + \theta_1\left(T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + \sum_{j=0}^{r-2}\theta^j\iota_\eta + \sum_{j=1}^{r-1}\theta^j T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + 2\theta^{\lfloor\frac{r-2}{q}\rfloor+I_{r,1}}B\right) \\
&\quad + \sum_{i=2}^{q}\theta_i\left(\sum_{j=0}^{r-i-1}\theta^j\iota_\eta + \sum_{j=0}^{r-i}\theta^j T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + 2\theta^{\lfloor\frac{r-2}{q}\rfloor+I_{r,i}}B\right) \\
&\leq \sum_{j=0}^{r-1}\theta^j\iota_\eta + \sum_{j=1}^{r}\theta^j T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + 2\theta^{\lfloor\frac{r-1}{q}\rfloor+I_{r+1,1}}B,
\end{aligned}
\tag{3.37}
\tag{3.38}
$$

where we use Assumption 9 in the second inequality and in the last inequality, we use $\theta_1 + \cdots + \theta_q = \theta < 1$ and the inequality

$$
2\theta_1\theta^{\lfloor\frac{r-2}{q}\rfloor+I_{r,1}}B + \cdots + 2\theta_q\theta^{\lfloor\frac{r-2}{q}\rfloor+I_{r,q}}B \leq 2(\theta_1 + \cdots + \theta_q)\theta^{\lfloor\frac{r-2}{q}\rfloor+I_{r,q}}B \leq 2\theta^{\lfloor\frac{r-1}{q}\rfloor+I_{r+1,1}}B,
$$

which the first inequality is due to the fact that $I_{r,q} \leq I_{r,q-1} \leq \cdots \leq I_{r,1}$ and the last

inequality is due to the fact that

$$\begin{aligned}\lfloor\frac{r-1}{q}\rfloor + I_{r+1,1} \leq & \lfloor\frac{r-1}{q}\rfloor + 1 = \lfloor\frac{r-2}{q}\rfloor + 1 + I(r = nq+1 \text{ for some } n \in \mathbb{N}_0) \\ \leq & \lfloor\frac{r-2}{q}\rfloor + 1 + I(r - \lfloor\frac{r-2}{q}\rfloor q - q > 0) = 1 + \lfloor\frac{r-2}{q}\rfloor + I_{r,q}.\end{aligned}$$

Moreover, by Eq. (3.34) and the inequality $|\sigma_{-B}(x) - \sigma_{-B}(y)| \leq |x-y|$, we have

$$\begin{aligned}&\left|H_{w_h-(r+1)\tilde{m}+m,\rho^\dagger,t} - \sigma_{-B}(\varepsilon_{t-1})\right| \\ &\leq T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + (1+\theta+\cdots+\theta^{r-2})\iota_\eta + (\theta+\theta^2+\cdots+\theta^{r-1})T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + 2\theta^{\lfloor\frac{r-2}{q}\rfloor + I_{r,1}}B \\ &= (1+\theta+\cdots+\theta^{r-2})\iota_\eta + (1+\theta+\theta^2+\cdots+\theta^{r-1})T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + 2\theta^{\lfloor\frac{r-1}{q}\rfloor + I_{r+1,2}}B,\end{aligned}$$

where in the last inequality we use the first inequality of Eq. (3.31) for r and in the last equation we use the fact that $\forall 1 \leq j \leq q$

$$\begin{aligned}\lfloor\frac{r-2}{q}\rfloor + I_{r,j} &= \lfloor\frac{r-1}{q}\rfloor - I(r = nq+1 \text{ for some } n \in \mathbb{N}_0) + I(r - \lfloor\frac{r-2}{q}\rfloor q - j > 0) \\ &= \lfloor\frac{r-1}{q}\rfloor + I(r - \lfloor\frac{r-1}{q}\rfloor q - j > 0) = \lfloor\frac{r-1}{q}\rfloor + I_{r+1,j+1}.\end{aligned}$$

Finally, by Eq. (3.35), for $1 \leq j \leq q-2$, we have,

$$\begin{aligned}&\left|H_{w_h-(r+1)\tilde{m}+m+j,\rho^\dagger,t} - \sigma_{-B}(\varepsilon_{t-j-1})\right| \\ \leq & |H_{w_h-r\tilde{m}+m+j-1,\rho^\dagger,t-1} - B - \varepsilon_{t-j-1}| = |H_{w_h-r\tilde{m}+m+j-1,\rho^\dagger,t-1} - \sigma_{-B}(\varepsilon_{t-j-1})| \\ \leq & (1+\theta+\cdots+\theta^{r-j-2})\iota_\eta + (1+\theta+\theta^2+\cdots+\theta^{r-j-1})T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + 2\theta^{\lfloor\frac{r-2}{q}\rfloor + I_{r,j+1}}B \\ = & (1+\theta+\cdots+\theta^{r-j-2})\iota_\eta + (1+\theta+\theta^2+\cdots+\theta^{r-j-1})T^{-\alpha_1\cdot\frac{2\beta}{p+k}} + 2\theta^{\lfloor\frac{r-1}{q}\rfloor + I_{r+1,j+2}}B,\end{aligned} \tag{3.39}$$

where in the last inequality, we use Eq. (3.31) for r.

Combining Eq. (3.38), Eq. (**??**) and Eq. (3.39), we complete the proof for the $r+1$ case, which completes the proof of Lemma 47. □

Lemma 48. *For any Lipschitz continuous function $\eta(\cdot)$ with Lipschitz constant bounded by θ, there exists $w_{\eta,1}, \dots w_{\eta,m}$ and a constant B such that such that $|\sum_{i=1}^{j} w_{\eta,i}| \leq \theta$, $|w_{\eta,i}| \leq 2\theta$, for $i = 1, \dots, m$, and for any $x \in [-B, B]$,*

$$\left| \eta(x) - \sum_{i=1}^{m+1} w_{\eta,i} \sigma_{v_i}(x) - b_\eta \right| \leq 4B\theta m^{-1}, \tag{3.40}$$

where $v_i := -B + \frac{2B}{m}(i-1)$.

Proof. First, we construct a neural network that is a piecewise linear approximation of $\eta(x) - b_\eta$ with

$$\sum_{i=1}^{m} w_{\eta,i} \sigma_{v_i}(v_j) = \eta(v_j) - b_\eta, \qquad \forall 1 \leq j \leq m+1.$$

Since $v_1 < v_2 < \cdots < v_m$, $\sigma_{v_i}(v_j) = 0$ for $j \leq i$, and $\sigma_{v_i}(v_j) = v_j - v_i$ for $j > i$, the above equation is equivalent to

$$\begin{aligned}
0 &= \eta(v_1) - b_\eta \\
w_{\eta,1}(v_2 - v_1) &= \eta(v_2) - b_\eta, \\
w_{\eta,1}(v_3 - v_1) + w_{\eta,2}(v_3 - v_2) &= \eta(v_3) - b_\eta, \\
&\vdots \\
\sum_{i=1}^{m} w_{\eta,i}(v_{m+1} - v_i) &= \eta(v_{m+1}) - b_\eta.
\end{aligned}$$

Solving the above equations, we get $b_\eta = \eta(-B)$ and

$$w_{\eta,1} = \frac{\eta(v_2) - \eta(v_1)}{v_2 - v_1}, \quad w_{\eta,1} + w_{\eta,2} = \frac{\eta(v_3) - \eta(v_2)}{v_3 - v_2}, \dots, \sum_{i=1}^{m} w_{\eta,i} = \frac{\eta(v_{m+1}) - \eta(v_m)}{v_{m+1} - v_m}.$$

In addition, note that $v_{i+1} - v_i = \frac{2B}{m}$, for any $1 \le j \le m$, we have

$$\left|\frac{2B}{m}w_{\eta,1} + \cdots + \frac{2B}{m}w_{\eta,j}\right| = \left|\eta(v_{j+1}) - \eta(v_j)\right| \le \theta|v_{j+1} - v_j| = \frac{2B}{m}\theta,$$

and

$$\left|\frac{2B}{m}w_{\eta,i}\right| = |\eta(v_{j+1}) - \eta(v_{j-1})| \le \theta|v_{j+1} - v_{j-1}| = \frac{4B}{m}\theta$$

which implies that $|\sum_{i=1}^{j} w_{\eta,i}| \le \theta$ and $|w_{\eta,i}| \le 2\theta$. Therefore, the choice of $w_{\eta,i}$ satisfies the constraints in Lemma 48.

To verify Eq. (3.40), we use the fact that $|\sigma(x)-\sigma(y)| \le |x-y|$ and assume $v_j \le x \le v_{j+1}$ for some $1 \le j \le m$ to obtain

$$\begin{aligned}
&\left|\eta(x) - \sum_{i=1}^{m} w_{\eta,i}\sigma_{v_i}(x) - b_\eta\right| \\
&\le \left|\eta(v_j) - \sum_{i=1}^{m} w_{\eta,i}\sigma_{v_i}(v_j) - b_\eta\right| + \left|\eta(x) - \eta(v_j)\right| + \left|\sum_{i=1}^{m} w_{\eta,i}\sigma_{v_i}(x) - \sum_{i=1}^{m} w_{\eta,i}\sigma_{v_i}(v_j)\right| \\
&\le \left|\eta(x) - \eta(v_j)\right| + \left|\sum_{i=1}^{j} w_{\eta,i}\right| |x - v_j| \le \frac{2B}{m}\theta + \theta|x - v_j| \le 4B\theta m^{-1}. \qquad \square
\end{aligned}$$

REFERENCES

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, June 2021.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Anish Agarwal and Rahul Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy, 2024.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Journal of the American Statistical Association*, 116(536):1731–1745, 2021.

Seung C. Ahn and Alex R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013a.

Seung C. Ahn and Alex R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013b.

Ferhat Akbas, Will J. Armstrong, Sorin Sorescu, and Avanidhar Subrahmanyam. Smart money, dumb money, and capital market anomalies. *Journal of Financial Economics*, 118 (2):355–382, 2015.

Sina Alemohammad. The recurrent neural tangent kernel. Master's thesis, Rice University, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *Advances in neural information processing systems*, 32, 2019.

Anna Almosova and Niek Andresen. Nonlinear inflation forecasting with recurrent neural networks. *Journal of Forecasting*, 42(2):240–259, 2023.

Yasuo Amemiya and Ilker Yalcin. Nonlinear Factor Analysis as a Statistical Method. *Statistical Science*, 16(3):275 – 294, 2001.

Stanislav Anatolyev and Anna Mikusheva. Factor models with many assets: strong factors, weak factors, and the two-pass procedure. *Journal of Econometrics*, 229(1):103–126, 2022.

P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100 – 1120, 1982.

T. W. Anderson and Herman Rubin. Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 20 (1):46 – 63, 1949.

Paolo Andreini, Cosimo Izzo, and Giovanni Ricco. Deep dynamic factor models. *arXiv preprint arXiv:2007.11887*, 2020.

Donald W. K. Andrews. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, 4(3):458–467, 1988.

Donald W. K. Andrews and Xu Cheng. Estimation and inference with weak, semi-strong, and strong identification. *Econometrica*, 80(5):2153–2211, 2012.

Donald WK Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

Isaiah Andrews and Anna Mikusheva. Conditional inference with a functional nuisance parameter. *Econometrica*, 84(4):1571–1612, 2016.

Isaiah Andrews, James H. Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753, 2019.

Eric Ghysels Andrii Babii and Jonas Striaukas. Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3): 1094–1106, 2022.

Andrew Ang, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang. The cross-section of volatility and expected returns. *The Journal of Finance*, 61(1):259–299, 2006.

Patrick Ango-Nze and Ricardo Rios. Density estimation in l_∞ norm for mixing processes. *Journal of statistical planning and inference*, 83(1):75–90, 2000.

J. D. Angrist, G. W. Imbens, and A. B. Krueger. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67, 1999.

M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Robert D Arnott, Vitali Kalesnik, and Juhani T Linnainmaa. Factor momentum. *The Review of Financial Studies*, 36(8):3034–3070, 01 2023.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116, 10 2017.

Maximilian Auffhammer and Ryan Kellogg. Clearing the air? the effects of gasoline content regulation on air quality. *American Economic Review*, 101(6):2687–2722, 2011.

David H Autor, David Dorn, and Gordon H Hanson. The china syndrome: Local labor market effects of import competition in the united states. *American economic review*, 103 (6):2121–2168, 2013.

Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

Jushan Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Boosting diffusion indices. *Journal of Applied Econometrics*, 24 (4):607–629, 2009.

Jushan Bai and Serena Ng. Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763, 2021.

Jushan Bai and Serena Ng. Approximate factor models with weaker loadings. *Journal of Econometrics*, 235(2):1893–1916, 2023a.

Jushan Bai and Serena Ng. Approximate factor models with weaker loadings. *Journal of Econometrics*, 235(2):1893–1916, 2023b. ISSN 0304-4076.

Z. Bai and J.W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer New York, 2009. ISBN 9781441906618.

Malcolm Baker and Jeffrey Wurgler. The equity share in new issues and aggregate stock returns. *The Journal of Finance*, 55(5):2219–2257, 2000.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2:53–58, 1989.

Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023.

Ravi Bansal and S. Viswanathan. No arbitrage and arbitrage pricing: A new approach. *The Journal of Finance*, 48(4):1231–1262, 1993.

Yong Bao and Aman Ullah. Expectation of quadratic forms in normal and nonnormal variables with applications. *Journal of Statistical Planning and Inference*, 140(5):1193–1205, 2010.

Nikita E Barabanov and Danil V Prokhorov. Stability analysis of discrete-time recurrent neural networks. *IEEE Transactions on Neural Networks*, 13(2):292–303, 2002.

Robert J. Barro. Economic growth in a cross section of countries. *The Quarterly Journal of Economics*, 106(2):407–443, 1991.

Robert J. Barro and Jong-Wha Lee. Sources of economic growth. *Carnegie-Rochester Conference Series on Public Policy*, 40:1–46, 1994. ISSN 0167-2231.

A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

D. J. Bartholomew. Factor analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, 42:293–321, 1980.

M. S. Bartlett. Factor analysis in psychology as a statistician sees it. In *Uppsala Symposium on Psychological Factor Analysis*, pages 23–34, Stockholm, 1953. Almqvist and Wiksell.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019a.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117:30063 – 30070, 2019b.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 2019.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

Marta Bańbura, Domenico Giannone, and Lucrezia Reichlin. Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

Geert Bekaert and Robert J Hodrick. Characterizing predictable components in excess returns on equity and foreign exchange markets. *The Journal of Finance*, 47(2):467–509, 1992.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650, 2013a.

Alexandre Belloni, Victor Chernozhukov, and Christian B. Hansen. *Inference for High-Dimensional Sparse Econometric Models*, volume 3 of *Econometric Society Monographs*, pages 245–295. Cambridge University Press, 2013b.

Alexandre Belloni, Victor Chernozhukov, and Christian B. Hansen. *Inference for High-Dimensional Sparse Econometric Models*, volume 3 of *Econometric Society Monographs*, pages 245–295. Cambridge University Press, 2013c.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 1985. ISBN 9783540960980. URL `https://books.google.com/books?id=IP1QAAAAMAAJ`.

Ben S. Bernanke and Jean Boivin. Monetary policy in a data-rich environment. *Journal of Monetary Economics*, 50(3):525–546, 2003.

Monika Bhattacharjee and Arup Bose. Supplement to "large sample behaviour of high dimensional autocovariance matrices", 2015.

Monika Bhattacharjee and Arup Bose. Large sample behaviour of high dimensional autocovariance matrices. 2016.

G. Biau and L. Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer International Publishing, 2015. ISBN 9783319253886.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106:21068 – 21073, 2009.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294, 2007.

Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J. Davison. Codeslam – learning a compact, optimisable representation for dense visual slam, 2018.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643, 2022.

Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*, 2017.

Nicola Borri, Denis Chetverikov, Yukun Liu, and Aleh Tsyvinski. One factor to bind the cross-section of returns. Working Paper 32365, National Bureau of Economic Research, April 2024.

Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. Deep learning techniques for music generation - a survey. *ArXiv*, abs/1709.01620, 2017.

Lawrence D. Brown and Eitan Greenshtein. Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 37(4):1685 – 1704, 2009. doi:10.1214/08-AOS630. URL `https://doi.org/10.1214/08-AOS630`.

Andrea Bucci. Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, 18(3):502–531, 2020.

Jian-Feng Cai, Emmanuel J. Candes, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Zongwu Cai, Jianqing Fan, and Qiwei Yao. Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956, 2000.

John Y. Campbell and Samuel B. Thompson. Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *The Review of Financial Studies*, 21 (4):1509–1531, 2007. ISSN 0893-9454.

John Y Campbell, Andrew W Lo, A Craig MacKinlay, and Robert F Whitelaw. The econometrics of financial markets. *Macroeconomic Dynamics*, 2(4):559–562, 1998.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J. Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Andrea Carriero, George Kapetanios, and Massimiliano Marcellino. Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26(5):735–761, 2011.

Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing, 2023.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51:1281–1304, 1983.

Kung Sik Chan and Howell Tong. On estimating thresholds in autoregressive models. *Journal of time series analysis*, 7(3):179–190, 1986.

Olivier Chapelle, Vladimir Naumovich Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131–159, 2002.

David A. Chapman. Approximating the asset pricing kernel. *The Journal of Finance*, 52(4): 1383–1410, 1997.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8 (1):6085, 2018.

Andrew Y. Chen and Tom Zimmermann. Open source cross-sectional asset pricing. *Critical Finance Review*, 27(2):207–264, 2022.

B.B. Chen and G.M. Pan. Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli*, 18(4):1405 – 1420, 2012. doi:10.3150/11-BEJ381. URL `https://doi.org/10.3150/11-BEJ381`.

D. L. Chen and S. Yeh. Growth under the shadow of expropriation? the economic impacts of eminent domain. Mimeo, Toulouse School of Economics, 2012.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Rong Chen and Ruey S Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993.

Xiaohong Chen and Xiaotong Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, 66(2):289–314, 1998a.

Xiaohong Chen and Xiaotong Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314, 1998b.

Xiaohong Chen and H. White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

Yi-Ting Chen and Chu-An Liu. Model averaging for asymptotically optimal combined forecasts. *Journal of Econometrics*, 235(2):592–607, 2023.

Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115:1756 – 1770, 2017.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116 (46):22931–22937, 2019.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep convolutional autoencoder-based lossy image compression. In *2018 Picture Coding Symposium (PCS)*, pages 253–257. IEEE, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

Richard H Clarida, Lucio Sarno, Mark P Taylor, and Giorgio Valente. The out-of-sample success of term structure models as exchange rate predictors: a step beyond. *Journal of International Economics*, 60(1):61–83, 2003.

Randolph B. Cohen, Christopher Polk, and Tuomo Vuolteenaho. The value spread. *The Journal of Finance*, 58(2):609–641, 2003.

Jasmine Collins, Jascha Sohl-Dickstein, and David Sussillo. Capacity and trainability in recurrent neural networks. *arXiv preprint arXiv:1611.09913*, 2016.

Gregory Connor and Robert A Korajczyk. Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of financial economics*, 15(3): 373–394, 1986.

Michael J. Cooper, Roberto C. Gutierrez Jr., and Allaudeen Hameed. Market states and momentum. *The Journal of Finance*, 59(3):1345–1365, 2004.

R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011. ISBN 9781139504966.

Hengjian Cui, Wenwen Guo, and Wei Zhong. Test for high-dimensional regression coefficients using refitted cross-validation variance estimation. *The Annals of Statistics*, 46(3):958 – 988, 2018.

Kent Daniel and Tobias J. Moskowitz. Momentum crashes. *Journal of Financial Economics*, 122(2):221–247, 2016.

Marcel de Jeu. Determinate multidimensional measures, the extended carleman theorem and quasi-analytic weights. *The annals of probability*, 31(3):1205–1227, 2003.

Serge De Valk, Daiane de Mattos, and Pedro Ferreira. Nowcasting: An r package for predicting economic variables using dynamic factor models. *The R Journal*, 11(1):230–244, 2019.

Geert Dhaene and Koen Jochmans. Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3):991–1030, 02 2015.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1 – 37, 2016.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962 – 994, 2004.

III Donohue, John J. and Steven D. Levitt. The Impact of Legalized Abortion on Crime. *The Quarterly Journal of Economics*, 116(2):379–420, 2001.

Paul Doukhan. Mixing: Properties and examples. In *New York: Springer-Verlag*, 1994.

Peter D Dueben and Peter Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10): 3999–4009, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2012. ISBN 9781139492133. URL `https://books.google.com/books?id=kfE7VkqbGawC`.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004.

Øyvind Eitrheim and Timo Teräsvirta. Testing the adequacy of smooth transition autoregressive models. *Journal of econometrics*, 74(1):59–75, 1996.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Charles Engel and Kenneth D West. Exchange rates and fundamentals. *Journal of political Economy*, 113(3):485–517, 2005.

N. Benjamin Erichson, Omri Azencot, Alejandro F. Queiruga, and Michael W. Mahoney. Lipschitz recurrent neural networks. *ArXiv*, abs/2006.12070, 2020.

Jamshid Etezadi-Amoli and Roderick P McDonald. A second generation nonlinear factor analysis. *Psychometrika*, 48(3):315–342, 1983.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Eugene F Fama and James D MacBeth. Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636, 1973.

Jianqing Fan. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Yuan Liao. Learning latent factors from diversified projections and its applications to overestimated and weak factors. *Journal of the American Statistical Association*, 117(538):909–924, 2022. doi:10.1080/01621459.2021.1906130.

Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2008.

Jianqing Fan, Yuling Yan, and Yuheng Zheng. When can weak latent factors be statistically inferred?, 2024.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Jon Faust and Jonathan H. Wright. Comparing greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business & Economic Statistics*, 27(4):468–479, 2009.

Charles Fefferman. Whitney's extension problem for. *Annals of mathematics*, pages 313–359, 2006.

Yingjie Feng. Optimal estimation of large-dimensional nonlinear factor models, 2023.

Miguel A. Ferreira and Pedro Santa-Clara. Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics*, 100(3):514–537, 2011. ISSN 0304-405X.

Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2): 654–669, 2018.

Nikolaus Fortelny and Christoph Bock. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome biology*, 21:1–36, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv: Learning*, 2018.

Hugo Freeman and Martin Weidner. Linear panel regressions with two-way unobserved heterogeneity. *Journal of Econometrics*, 237(1):105498, 2023.

Simon Freyaldenhoven. Factor models with local factors — determining the number of relevant factors. *Journal of Econometrics*, 229(1):80–102, 2022. ISSN 0304-4076.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, 6(6):801–806, 1993.

Walter Gander, Gene H. Golub, and Urs von Matt. A constrained eigenvalue problem. *Linear Algebra and its Applications*, 114-115:815–839, 1989. ISSN 0024-3795.

Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

Edward I. George and Robert E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, 2008a. ISSN 0304-3932.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of monetary economics*, 55(4):665–676, 2008b.

Domenico Giannone, Michele Lenza, and Giorgio E. Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2022.

Stefano Giglio, Dacheng Xiu, and Dake Zhang. Test assets and weak factors. Working Paper 29002, National Bureau of Economic Research, 2021.

Stefano Giglio, Dacheng Xiu, and Dake Zhang. Prediction when factors are weak. *Chicago Booth Research Paper*, (23-04), 2023a.

Stefano Giglio, Dacheng Xiu, and Dake Zhang. Prediction when factors are weak. *Chicago Booth Research Paper*, (23-04), April 2023b. Available at SSRN.

Stefano Giglio, Dacheng Xiu, and Dake Zhang. Test assets and weak factors. *The Journal of Finance*, 80(1):259–319, 2025.

Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246, 2016a.

Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pages 241–246. IEEE, 2016b.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}_n$. In *Geometric Aspects of Functional Analysis*, pages 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017.

Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26(none): 1 – 22, 2021. doi:10.1214/21-EJP606. URL `https://doi.org/10.1214/21-EJP606`.

Philippe Goulet Coulombe, Maxime Leroux, Dalibor Stevanovic, and Stéphane Surprenant. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964, 2022.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *ArXiv*, abs/1410.5401, 2014.

Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

Michael Griebel and Helmut Harbrecht. Approximation of bi-variate functions: Singular value decomposition versus sparse grids. *IMA Journal of Numerical Analysis*, 34, 01 2014.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021a.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Shihao Gu, Bryan Kellly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222:429–450, 2021b.

V. Guillemin and A. Pollack. *Differential Topology*. Mathematics Series. Prentice-Hall, 1974. ISBN 9780132126052.

Wenxuan Guo and Panos Toulis. Invariance-based inference in high-dimensional regression with finite-sample guarantees, 2023.

Amirhossein Habibian, Ties van Rozendaal, Jakub M. Tomczak, and Taco S. Cohen. Video compression with rate-distortion autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Valentin Haddad, Serhiy Kozak, and Shrihari Santosh. Factor timing. *The Review of Financial Studies*, 33:1980–2018, 05 2020.

Peter Hall and Jiashun Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686 – 1732, 2010.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Gregory Frederick Diamos, Erich Elsen, Ryan J. Prenger, Sanjeev Satheesh, Shubho Sengupta, Vinay Rao, Adam Coates, and A. Ng. Deep speech: Scaling up end-to-end speech recognition. *ArXiv*, abs/1412.5567, 2014.

Bruce E Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000.

Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

D. L. Hanson and F. T. Wright. A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables. *The Annals of Mathematical Statistics*, 42(3):1079 – 1083, 1971.

Wolfgang Härdle, Helmut Lütkepohl, and Rong Chen. A review of nonparametric time series analysis. *International statistical review*, 65(1):49–72, 1997.

Jeffrey D. Hart. Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics*, 6(2-3):115–142, 1996.

Campbell R. Harvey and Akhtar Siddique. Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3):1263–1295, 2000.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann, 1992.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2015. ISBN 9781498712170. URL `https://books.google.com/books?id=f-A_CQAAQBAJ`.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(4):757–779, 1993.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022.

K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society.

Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427, 2021.

Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Chang hoon Song, Geonho Hwang, Jun ho Lee, and Myungjoo Kang. Minimal width for universal property of deep rnn. *Journal of Machine Learning Research*, 24(121):1–41, 2023.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Joel L Horowitz. Bootstrap methods for markov processes. *Econometrica*, 71(4):1049–1082, 2003.

David A Hsieh. Chaos and nonlinear dynamics: application to financial markets. *The journal of finance*, 46(5):1839–1877, 1991.

Florian Huber and Martin Feldkircher. Adaptive shrinkage in bayesian vector autoregressive models. *Journal of Business & Economic Statistics*, 37(1):27–39, 2019.

Philippe Huber, Elvezio Ronchetti, and Maria-Pia Victoria-Feser. Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(4):893–908, 2004.

Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Yuri I. Ingster, Alexandre B. Tsybakov, and Nicolas Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4(none):1476 – 1526, 2010. doi:10.1214/10-EJS589. URL `https://doi.org/10.1214/10-EJS589`.

Atsushi Inoue and Lutz Kilian. How useful is bagging in forecasting economic time series? a case study of u.s. consumer price inflation. *Journal of the American Statistical Association*, 103(482):511–522, 2008.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1):65–91, 1993.

Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647 – 1684, 2009. doi:10.1214/08-AOS638. URL `https://doi.org/10.1214/08-AOS638`.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

Yuling Jiao, Yang Wang, and Bokai Yan. Approximation bounds for recurrent neural networks with application to regression. *arXiv preprint arXiv:2409.05577*, 2024.

Jiashun Jin and Zheng Tracy Ke. Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statistica Sinica*, 26(1):1–34, 2016.

Shuting Jin, Xiangxiang Zeng, Feng Xia, Wei Huang, and Xiangrong Liu. Application of deep learning methods in biological networks. *Briefings in bioinformatics*, 22(2):1902–1917, 2021.

Ian T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):300–303, 1982.

David Alan Jones and David Roxbee Cox. Nonlinear autoregressive processes. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 360(1700):71–95, 1978.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2014.

Michael P. Keane and Kenneth I. Wolpin. The career decisions of young men. *Journal of Political Economy*, 105(3):473–522, 1997.

Bryan Kelly and Seth Pruitt. Market expectations in the cross-section of present values. *The Journal of Finance*, 68(5):1721–1756, 2013.

Bryan T. Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *The Journal of Finance*, 2023. Forthcoming.

David A. Kenny and Charles M. Judd. Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96(1):201–210, 1984.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020.

Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. *arXiv preprint arXiv:1711.00811*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Frank Kleibergen. Generalizing weak instrument robust iv statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of Econometrics*, 139(1):181–216, 2007.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231 – 2249, 2021.

Michal Kolesár, Ulrich K. Müller, and Sebastian T. Roelsgaard. The fragility of sparsity, 2024.

Gary Koop and Simon Potter. Forecasting in dynamic factor models using bayesian model averaging. *The Econometrics Journal*, 7(2):550–565, 2004.

Alan Kraus and Robert H. Litzenberger. Skewness preference and the valuation of risk assets. *The Journal of Finance*, 31(4):1085–1100, 1976.

Chung-Ming Kuan and Tung Liu. Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of applied econometrics*, 10(4):347–364, 1995.

Brian Kulis, Mátyás Sustik, and Inderjit Dhillon. Learning low-rank kernel matrices. ICML '06, pages 505–512, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Lei Le, Andrew Patterson, and Martha White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. *Advances in neural information processing systems*, 31, 2018.

Y. LeCun. *Connexionist Learning Models*. Phd thesis, Universite Pierre et Marie Curie (Paris), 1987.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

Daniel LeJeune, Hamid Javadi, and Richard Baraniuk. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 3525–3535. PMLR, 2020.

Yue Li, Ilmun Kim, and Yuting Wei. Randomized tests for high-dimensional regression: A more efficient and powerful solution. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc., 2020. ISBN 9781713829546.

Zeng Li, Clifford Lam, Jianfeng Yao, and Qiwei Yao. Supplement to "on testing for high-dimensional white noise", 2019a.

Zeng Li, Clifford Lam, Jianfeng Yao, and Qiwei Yao. On testing for high-dimensional white noise. 2019b.

Zhong Li, Jiequn Han, E Weinan, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers. *The Annals of Statistics*, 50(3):1669 – 1695, 2022.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2 (3):18–22, 2002.

F. Liese and K.J. Miescke. *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer Series in Statistics. Springer New York, 2008.

Zachary C Lipton, David C Kale, Randall Wetzel, et al. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 56(56):253–270, 2016.

Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *ArXiv*, abs/1506.00019, 2015.

Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Hao Liu, Alex Havrilla, Rongjie Lai, and Wenjing Liao. Deep nonparametric estimation of intrinsic data structures by chart autoencoders: Generalization error and robustness. *Applied and Computational Harmonic Analysis*, 68:101602, 2024.

Sifan Liu and Edgar Dobriban. Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*, 2020.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, volume 2013, pages 436–440, 2013.

Wolfgang Maass, Prashant Joshi, and Eduardo D Sontag. Computational aspects of feedback in neural circuits. *PLoS computational biology*, 3(1):e165, 2007.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders, 2016.

Tong Mao and Ding-Xuan Zhou. Rates of approximation by relu shallow neural networks. *Journal of Complexity*, 79:101784, 2023.

Nelson C Mark. Exchange rates and fundamentals: Evidence on long-horizon predictability. *The American Economic Review*, pages 201–218, 1995.

Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pages 52–59. Springer, 2011.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80): 2287–2322, 2010.

Michael W. McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

R.P. McDonald. A general approach to nonlinear factor analysis. *Psychometrika*, 27(4): 397–415, 1962.

Richard A Meese and Kenneth Rogoff. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, 14(1-2):3–24, 1983.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 25–28 Jun 2019.

Yueyang Men, Liang Li, Ziqing Hu, and Yongli Xu. Learning rates of deep nets for geometrically strongly mixing sequence. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Carl D. Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, USA, 2000. ISBN 0898714540.

Tomas Mikolov, Martin Karafiát, Lukávs Burget, Jan vCernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, 2010.

Thomas Mikosch et al. Limit theory for the sample autocorrelations and extremes of a garch (1, 1) process. *The Annals of Statistics*, 28(5):1427–1451, 2000.

Léo Miolane and Andrea Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313 – 2335, 2021.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6): 1236–1246, 2018.

Dharmendra S Modha and Elias Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42(6):2133–2145, 1996.

Marcelo Moreira. Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, 152(2):131–140, 2009.

I. Moustaki and M. Knott. Generalized latent trait models. *Psychometrika*, 65:391–411, 2000.

S. Nagel. *Machine Learning in Asset Pricing*. Princeton Lectures in Finance. Princeton University Press, 2021. ISBN 9780691218717.

Vinod Nair and Geoffrey E. Hinton. 3d object recognition with deep belief nets. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 1339–1347, Red Hook, NY, USA, 2009. Curran Associates Inc.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21 (174):1–38, 2020.

Whitney K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167, 1991. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/2938179`.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

José Luis Montiel Olea and Carolin Pflueger. A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369, 2013.

Alexei Onatski. Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92, 02 2005.

Alexei Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

David Pollard. *Convergence of Empirical Processes*. Springer-Verlag, New York, 1984.

Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.

Jiayu Qiu, Bin Wang, and Changjun Zhou. Forecasting stock prices with long-short term memory neural network based on attention mechanism. *PloS one*, 15(1):e0227222, 2020.

T Subba Rao. On the theory of bilinear time series models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 43(2):244–255, 1981.

T Subba Rao and MM Gabr. *An introduction to bispectral analysis and bilinear time series models*, volume 24. Springer Science & Business Media, 2012.

David E. Rapach, Jack K. Strauss, and Guofu Zhou. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2):821–862, 2010.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Holger Rauhut and Rachel Ward. Interpolation via weighted $\ell 1$ minimization. *Applied and Computational Harmonic Analysis*, 40(2):321–351, 2016.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

Markus Reiß and Martin Wahl. Nonasymptotic upper bounds for the reconstruction error of PCA. *The Annals of Statistics*, 48(2):1098 – 1123, 2020.

Dan S. Rickman and Hongbo Wang. Two tales of two u.s. states: Regional fiscal austerity and economic performance. *Regional Science and Urban Economics*, 68:46–55, 2018.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 833–840, Madison, WI, USA, 2011. Omnipress.

Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

Herbert Robbins. The Empirical Bayes Approach to Statistical Decision Problems. *The Annals of Mathematical Statistics*, 35(1):1 – 20, 1964. doi:10.1214/aoms/1177703729. URL `https://doi.org/10.1214/aoms/1177703729`.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

PM Robinson. The estimation of a nonlinear moving average model. *Stochastic processes and their applications*, 5(1):81–90, 1977.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970. ISBN 9780691015866.

Veronika Rovcková and Edward I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

Barbara Rossi. Exchange rate predictability. *Journal of economic literature*, 51(4):1063–1119, 2013.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic communications in probability*, 18, 06 2013a.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013b.

David E. Rumelhart and James L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. 1987.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Eleanor Sanderson and Frank Windmeijer. A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics*, 190(2):212–221, 2016.

Terence D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2(6):459–473, 1989.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021.

Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. In *Artificial Neural Networks–ICANN 2006: 16th International Conference, Athens, Greece, September 10-14, 2006. Proceedings, Part I 16*, pages 632–640. Springer, 2006.

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

Johannes Schmidt-Hieber. The kolmogorov–arnold representation theorem revisited. *Neural Networks*, 137:119–126, 2021.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020.

Haidong Shao, Hongkai Jiang, Huiwei Zhao, and Fuan Wang. A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. *Mechanical Systems and Signal Processing*, 95:187–204, 2017.

J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. Eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In *Algorithmic Learning Theory: 13th International Conference, ALT 2002*, volume 2533 of *Lecture Notes in Computer Science*, pages 23–40. Springer-Verlag, 2002.

J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and the generalisation error of kernel pca. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005.

Zhouyu Shen and Dacheng Xiu. On the theory of autoencoders. *Working paper*, 2024.

Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 1394–1401. Ieee, 2018.

Hava T Siegelmann, Bill G Horne, and C Lee Giles. Computational capabilities of recurrent narx neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(2):208–215, 1997.

Igor Silin and Jianqing Fan. Canonical thresholding for nonsparse high-dimensional linear regression. *The Annals of Statistics*, 50(1):460 – 486, 2022.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958.

A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. CRC Press, Boca Raton, FL, 2004.

Stephan Smeekes and Etienne Wijler. Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting*, 34(3):408–430, 2018. ISSN 0169-2070.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, dec 2006.

Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 3:54–70, 2023.

Paul Speckman. Spline Smoothing and Optimal Rates of Convergence in Nonparametric Regression Models. *The Annals of Statistics*, 13(3):970 – 983, 1985.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997a.

Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997b.

Robert F. Stambaugh, Jianfeng Yu, and Yu Yuan. The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, 104(2):288–302, 2012. Special Issue on Investor Sentiment.

Ingo Steinwart and Andreas Christmann. Fast learning from non-iid observations. *Advances in neural information processing systems*, 22, 2009.

James H Stock and Mark W Watson. Forecasting inflation. *Journal of monetary economics*, 44(2):293–335, 1999.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

James H. Stock and Mark W. Watson. Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493, 2012.

James H. Stock and Jonathan H. Wright. Gmm with weak identification. *Econometrica*, 68 (5):1055–1096, 2000.

James H. Stock and Motohiro Yogo. Testing for weak instruments in linear iv regressions. In *Identification and Inference for Econometric Models*, pages 80–108. Cambridge University Press, Cambridge, 2005.

James H Stock, Jonathan H Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529, 2002.

Charles J. Stone. Optimal Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, 8(6):1348 – 1360, 1980.

Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982.

Weijie Su, Małgorzata Bogdan, and Emmanuel Candès. False discoveries occur early on the lasso path. *The Annals of Statistics*, 45(5):2133–2150, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215, 2014.

Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, volume 2018, pages 3229–3233, 2018.

Binhua Tang, Zixiang Pan, Kang Yin, and Asif Khateeb. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in genetics*, 10:214, 2019.

Pham Dinh Tao and Le Thi Hoai An. A d.c. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Timo Teräsvirta. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89(425):208–218, 1994.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. In *International Conference on Learning Representations*, 2017.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France, 2015. PMLR.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64 (8):5592–5628, 2018.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

Peter Tino, Christian Schittenkopf, and Georg Dorffner. Financial volatility trading using recurrent neural networks. *IEEE transactions on neural networks*, 12(4):865–874, 2001.

Howell Tong. *Threshold models in non-linear time series analysis*, volume 21. Springer Science & Business Media, 2012.

Howell Tong and Keng S Lim. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(3):245–268, 1980.

Ruey S. Tsay. Analysis of financial time series. *Technometrics*, 48:316 – 316, 2005.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning, 2018.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023. URL `http://jmlr.org/papers/v24/22-1398.html`.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24(1), March 2024.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Yoshimasa Uematsu and Takashi Yamagata. Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1):213–227, 2022.

Aman Ullah. *Finite sample econometrics*. OUP Oxford, 2004.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. ICML '08, pages 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery.

Pascal Vincent, H. Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Fa Wang. Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions. *Journal of Econometrics*, 229(1):180–200, 2022.

Runmin Wang and Xiaofeng Shao. Hypothesis testing for high-dimensional time series via self-normalization. *The Annals of Statistics*, 48(5):2728 – 2758, 2020. doi:10.1214/19-AOS1904. URL `https://doi.org/10.1214/19-AOS1904`.

Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791 – 2823, 2020.

Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *International conference on machine learning*, pages 6607–6617. PMLR, 2019.

Liu Wei, Huazhen Lin, Shurong Zheng, and Jin Liu. Generalized factor model for ultra-high dimensional correlated variables with mixed types. *Journal of the American Statistical Association*, 118:1–42, 10 2021.

Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. ICML '04, page 106, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385.

Ivo Welch and Amit Goyal. A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21(4):1455–1508, 2007.

Jonathan H Wright. Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics*, 146(2):329–341, 2008.

Jonathan H. Wright. Forecasting us inflation by bayesian model averaging. *Journal of Forecasting*, 28(2):131–144, 2009.

Kaspar Wüthrich and Ying Zhu. Omitted variable bias of lasso-based inference methods: A finite sample analysis. *The Review of Economics and Statistics*, 105(4):982–997, 2023.

Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pages 5708–5717. PMLR, 2018.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M. Dobaie. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273: 643–649, 2018.

Junhai Zhai, Sufang Zhang, Junfen Chen, and Qiang He. Autoencoder and its various variants. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 415–419. IEEE, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3): 107–115, February 2021.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567 – 1594, 2008.

Peng Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Guo-Bing Zhou, Jianxin Wu, Chen-Lin Zhang, and Zhi-Hua Zhou. Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing*, 13(3): 226–234, 2016.

Hong-Tu Zhu and Sik-Yum Lee. Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology*, 52(2):225–242, 1999.

Jinfeng Zhuang, Jialei Wang, Steven C. H. Hoi, and Xiangyang Lan. Unsupervised multiple kernel learning. In Chun-Nan Hsu and Wee Sun Lee, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 20 of *Proceedings of Machine Learning Research*, pages 129–144, South Garden Hotels and Resorts, Taoyuan, Taiwain, 14–15 Nov 2011. PMLR.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.