STATISTICAL VIEWPOINTS ON NETWORK MODEL, PDE IDENTIFICATION, LOW-RANK MATRIX ESTIMATION AND DEEP LEARNING

A Dissertation
Presented to
The Academic Faculty

By

Namjoon Suh

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial & Systems Engineering

Georgia Institute of Technology

December 2022

STATISTICAL VIEWPOINTS ON NETWORK MODEL, PDE IDENTIFICATION, LOW-RANK MATRIX ESTIMATION AND DEEP LEARNING

Thesis committee:

Dr. Xiaoming Huo, Advisor
The H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei, Co-Advisor
The H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Sung Ha Kang
School of Mathematics
Georgia Institute of Technology

Dr. Mayya Zhilova
School of Mathematics
Georgia Institute of Technology

Dr. Ding-Xuan Zhou
School of Mathematics and Statistics
The University of Sydney

Date approved: November 18, 2022

For my Parents, Grandparents, Sister, and SJ.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude towards my advisor, Professor. Huo, Xiaoming, for introducing me to the areas of high-dimensional statistics and deep learning theory. His passion for the mathematical research on various topics in data science naturally leads me to explore many topics in the field. Furthermore, despite many setbacks I have, his firm beliefs in my research direction helped me maintain enthusiasm towards my research throughout the long journey of Ph.D. training.

I am also hugely indebted to my co-advisor, Professor. Mei, Yajun, for introducing me what it is like conducting the academic research in the field of statistics. I will never forget the joyous moment when I succeeded in reproducing the numerical experiments on Ising model after the three months of trying. It was the first research-related task I was given by him and learned the persistent and consistent thinking on the research problem is the most important morales the researcher should possess.

I would also like to thank Professors Kang, Sung Ha, Zhilova, Mayya, and Zhou, Ding-Xuan for serving as committee members in my thesis defense. During the collaboration with Professor Kang, I learned a lot about PDE identification problems via ℓ_1-penalized least-square problem. I am also thankful to Professor Mayya for giving interesting lectures on mathematical statistics and statistical hypothesis testing. Last but not least, my deep appreciations go to Professor Zhou. His expertise in learning theory and approximation theory of neural networks help me deepen my knowledge on the corresponding topics and also ignite my passion to keep working on interesting topics in these areas in the near future.

I also feel very fortunate to have an amazing group of friends who made my Ph.D. an enjoyable experience, filled with laughter and fun. First, I would like to present my appreciation to who has always been my best friend and now my wife Sujin Lee. Moreover, I treasure the incredible relations with my friends who were; my roommates, Arvind

Krishna, Jaekang Kim, Kihyun Park, Jungho Ahn, my closest friends in the department; Li-Hsiang Lin, Ana Maria, Chungjae Lee, Young-In Kim, Hoon Na, Heejune Sheen, Taesoon Kim, Mingu Kwak, Sungwoo Kim, Soonhong Kwon, Jinyong Lim; my lab mates; Hyunouk Ko, Tian-Yi Zhou, Yilling Xie, Yilling Luo, Yijin Ni, Eugene Ndiaye; friends in Korea Univ. association; Jaewoo Jeong, Jeongwon Kim, Cheonjae Yu, Jin Yu, Youngjo Kim.

Finally, I want to express my deepest gratitude to my parents, grandparents and my sister for their constant and unconditional love and support. Without them, none of my accomplishments would be possible. To them, I dedicate this thesis.

TABLE OF CONTENTS

Acknowledgments iv

List of Tables xi

List of Figures xiii

Summary xvi

Chapter 1: A Network Model that Combines Latent Factors and Sparse Graphs 1

1.1 Introduction 1

1.2 Related Work 4

1.2.1 Fused Latent and Graphical (FLaG) Model 4

1.2.2 An Extension of the Factor-based Model 5

1.2.3 Matrix Decomposition 6

1.3 Model Formulation 6

1.4 Estimation 8

1.5 Non-asymptotic error bound of the estimator 11

1.6 Numerical experiments with synthetic data 14

1.6.1 Synthetic Setting 15

1.6.2 Choosing the tuning parameters and evaluation criteria 16

1.6.3 Several Observations 19

1.6.4 Estimation Error 22

1.6.5 Computation Time 23

1.7 Applications in Real Dataset 24

1.7.1 Zachary's karate club network 24

1.7.2 U.S. political book network 25

1.7.3 Social Network of U.S. Political Blog 27

1.7.4 Citation network for statisticians 29

1.8 Discussion 32

Chapter 2: Asymptotic Theory of ℓ_1-Regularized PDE Identification from a Single Noisy Trajectory 37

2.1 Introduction 37

2.2 Related Works 41

2.3 PDE Identification via ℓ_1-PsLS 43

2.3.1 Problem Setting and Notations 43

2.3.2 Local-Polynomial Regression Estimators for Derivatives 44

2.3.3 ℓ_1-regularized Pseudo Least Squares Model 46

2.4 Recovery Theory for ℓ_1-PsLS based PDE Identification 48

2.4.1 Signed-Support Recovery 48

2.4.2 Assumptions 49

2.4.3 Statement of Main Result 50

2.4.4 Proof Strategy of Theorem 2.3.1 52

2.5 Analysis Under Sample Incoherence Matrix Assumptions 52

2.5.1 Statement of Proposition 53

2.5.2 Proof Overview of Proposition 2.4.1 54

2.5.3 Technical Contribution 55

2.6 Uniform Convergence of Sample Incoherence Matrix 57

2.7 Numerical Experiments 62

2.7.1 Experimental Setting 62

2.7.2 Numerical Verifications of Main Statements 65

2.7.3 Impact of ${\beta^*}_{\text{min}}$ in Signed-Support Recovery of ℓ_1-PsLS 66

2.8 Discussion 68

Chapter 3: High-Dimensional Multivariate Linear Regression with Weighted Nuclear Norm Regularization 72

3.1 Introduction 72

3.2 WMVR-ADMM and Convergence Analysis 75

3.3 Statistical Properties of the Estimator 78

3.3.1 Statistical Properties of $\widehat{\mathbf{\Theta}}$ under the Orthogonal Design 78

3.3.2 Estimation Error under Random Design 80

3.4 Data-driven Model Selections 83

3.4.1 Surrogate Estimator $\widehat{\mathbf{\Theta}}^{\text{SR}}$ for the GCV statistic 83

3.4.2 GCV Statistic and Weight Updates 85

3.5 Numerical Experiments 86

3.5.1 Convergence of WMVR-ADMM 86

3.5.2 Comparisons of Estimation Error with Other Methods 88

3.5.3 Application to A Real Dataset 90

3.6 Conclusion and Discussion 93

Chapter 4: A Non-parametric Regression viewpoint : Generalization of Over-parametrized Deep ReLU Network under Noisy Observations 96

4.1 Introduction 96

4.1.1 Contributions 98

4.1.2 Technical Comparisons with [142] 99

4.1.3 Additional Related works 100

4.2 Problem Formulation 101

4.2.1 Non-parametric Regression 101

4.2.2 Deep Neural Network with ReLU activation 102

4.3 Main Theory 104

4.3.1 Assumptions. 106

4.3.2 Minimax rate for recovering $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$ 107

4.3.3 Analysis of Unregularized DNN 108

4.3.4 Analysis of ℓ_2-regularized DNN 110

4.4 Conclusion 113

Chapter 5: Approximation and non-parametric estimation of functions over high-dimensional spheres via deep ReLU networks 114

5.0.1 Chapter Road map and Contributions 115

5.0.2 Related works 117

5.1 Preliminary Definitions 119

5.1.1 Definition of Deep ReLU network 119

5.1.2 Definition of Sobolev Spaces over Sphere 120

5.2 Approximation error 122

5.2.1 Comparison with [8] 124

5.2.2 Fast Approximation error in terms of d 125

5.3 Statistical risk bound 126

5.3.1 Upper-bound on excess risk 128

5.3.2 Convergence Rate of Excess Risk 129

5.4 An Open Question 130

Appendices 132

Appendix A: A Network Model that Combines Latent Factors and Sparse Graphs 133

Appendix B: Asymptotic Theory of ℓ_1-Regularized PDE Identification from a Single Noisy Trajectory 150

Appendix C: High-Dimensional Multivariate Linear Regression with Weighted Nuclear Norm Regularization 176

Appendix D: A Non-parametric Regression viewpoint : Generalization of Overparametrized Deep ReLU Network under Noisy Observations . . . 195

Appendix E: Approximation and non-parametric estimation of functions over high-dimensional spheres via deep ReLU networks 249

References 280

LIST OF TABLES

1.1 For two scenarios, our heuristic method chooses the model with $\widehat{L}$ with true rank, $\widehat{S}$ whose M_2 value is close to 1, and M_3 value is close to 0. Also note that it chooses a model whose mis-classification rate is close to 0. A number in the parentheses represents the rank of $\widehat{L}$ estimated from (γ^{HNCV}, δ^{HNCV}), (γ^{AIC}, δ^{AIC}) and (γ^{BIC}, δ^{BIC}) for each case. 20

1.2 A list of top 10 blogs that received the most links from blogs in the opposite party. 30

1.3 Top 4 edges corresponding with the pairs of papers from different communities. Authors and years of publication for the papers in each pair are also presented. In the first pair, a paper from functional analysis topic cites a paper from variable selection topic for borrowing a mathematical representation to build a theorem. But they are not related in terms of topic. 35

1.4 Grid settings for implementation of HNCV method and running times of the algorithm on each dataset. We repeat (a) and (b) of Step 2 ten times for Karate club, political book, and citation network datasets, and five times for political blog dataset. Maximum, minimum, and average computational times over the grid, G, are recorded. Interestingly, as the number of nodes grows, it turns out that running time is sensitive to the choice of tuning parameter. See Political Blog dataset. 36

2.1 Specific choices of the constants in the order of $h_N = \Theta(N^{-\frac{1}{7}})$ and $w_M = \Theta(M^{-\frac{1}{7}})$ for the experiments on Viscous Burgers equation and KdV equation are presented. 65

5.1 Notation $\tilde{\mathcal{O}}(\cdot)$ hide the logarithmic factor in n. Note that the upper-bounds for $\mathcal{N}$ in Theorem 5.3.3 (i.e., $\mathcal{N} = \mathcal{O}(Md)$) are from Theorem 5.2.1 with choices $M = \lceil n^{\frac{3d}{3d+4r}} \rceil$ and $n \ll d$. 115

A.1 Among the 151 ad-hoc citational relationships, 118 of them are formed by pairs of papers from “Mixed Community”. Remaining 31 ad-hoc edges (except for the 2 pairs of papers from “FuncAn-FunAc”) are formed by pairs of papers with different communities. 149

LIST OF FIGURES

1.1 Graphical illustrations of six synthetic networks. Nodes that share the common factors are clustered. The cross cluster links are ad-hoc Citations. All the graphs are drawn based on the algorithm in [48]. 16

1.2 Scree plots for six synthetic Networks. $\widehat{K}^{\text{Scree}}$ recovers the number of topics in the network correctly for all six cases. 20

1.3 Case 1 to 6: Plots of rows based on the two leading eigenvectors of $\widehat{L}^{\text{HNCV}}$. 22

1.4 We can observe linear growth of $\|\widehat{\Delta}^L_N\|_F^2$ and $\|\widehat{\Delta}^S\|_F^2$ on the rank parameter k (panel A) and sparsity s (panel B) for the two respective scenarios. Also note that there are no differences in quantities of $\|\widehat{\Delta}^L_N\|_F^2$ and $\|\widehat{\Delta}^S\|_F^2$ when they are obtained from Estimator 1 and Estimator 2. 23

1.5 Plot of the computational runtime (in seconds) of ADMM algorithm using the HNCV approach. 23

1.6 Plots of the first column versus the second column of $\widehat{E}_2$ on Karate (Left), Political Book (Middle), and Political Blog Network datasets (Right). The number of clusters is set as $K = 2$ and K-means clustering algorithm is applied on $\widehat{E}_2$ for the three datasets. Detected memberships of each node via K-means algorithm are colored in blue or red. 25

1.7 From left to right : Topologies of Karate club network data and U.S. political book network data. In Karate club data, node 1 represents “Mr. Hi” and node 34 represents “John A”. For Karate club data, nodes that are classified as “Hi” group are colored in orange and those classified as “John A” group are colored in lightblue. For political book dataset, nodes classified as “conservative” party are colored in orange and those classified as “liberal” party are colored in light blue. In both dataset, color coding of ad-hoc edges corresponds to any $\widehat{S}_{ij} > 0$ in respective fitted model. 26

1.8 Parametric bootstrap to check the goodness of fit of **CFSG** model (left panel) and Inner Product model (right panel). Red dotted line is an observed likelihood of **CFSG** model, whereas the observed likelihood of Inner Product model is not appeared on the histogram. 28

1.9 From left to right : Scree plots of the adjacency matrix X^{orig} and X^{sub}. . . 30

2.1 Probability of signed-support recovery $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)]$ versus the grid size of temporal dimension N, and $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty$ versus N are recorded on the same plot for Viscous Burger's equation in panel (a) and for KdV equation in panel (b), respectively. 67

2.2 Boxplots of $\|\widehat{\mathcal{Q}}_N\|_\infty$ versus N are displayed for Viscous Burgers' equation in panel (a) and KdV equation in panel (b), respectively. 67

2.3 Left panel (a) displays the curves representing $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^*)]$ versus N, when $\nu = 0.03, 0.02, 0.01, 0.005$. Right panel (b) exhibits the range of λ_N for which ℓ_1-PsLS gives the solution $\widehat{\boldsymbol{\beta}}^{\lambda}$ such that $\mathcal{S}(\widehat{\boldsymbol{\beta}}^{\lambda}) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$ with respect to N, when ν is set as 0.005. 68

3.1 Three panels display the plots of estimated sigular values versus ground truth singular values $\sigma_j^\star$. The first two panels (A) and (B) are results from WMVR-ADMM algorithm with one weight update iteration under $n = 250$. The panel (C) exhibits the result when the estimator is obtained from SNN penalized least squares under $n = 1000$. 74

3.2 Under non-orthogonal $\boldsymbol{X}$, panel (A) displays the plot of the first 50 singular values of $\widehat{\boldsymbol{\Theta}}$ versus $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$. Panel (B) exhibits the plot of the remaining 200 singular values of $\widehat{\boldsymbol{\Theta}}^{\text{LS}}$ versus $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$. 84

3.3 Convergences of $R^{(k)}$ (panel (A)) and $O^{(k)}$ (panel (B)) over the algorithm iteration index k. Regardless of random initializations, $R^{(k)}$ and $O^{(k)}$ converge to 0 and to a same objective function value, respectively. 87

3.4 The plots demonstrates the comparisons of estimation errors in terms of RMSEs from the proposed method with ANN and SNN methods under different simulation settings. The three figures in the first row (A) $\sim$ (C) are the comparison results from sample size 30, 300, and 3000, respectively, under the true rank $r^\star = 2$. Analogously, Figures (D) $\sim$ (F) (second row) are the results from $r^\star = 5$, Figures (G) $\sim$ (I) (third row) are the results from $r^\star = 8$, and Figures (J) $\sim$ (L) (fourth row) are the results from $r^\star = 11$. . . 89

3.5 Two sequences of weights in (3.19) used for the estimation (panel (A)) and the resulting RMSEs (panel (B)). Low RMSEs of WNN weights arise from the high penalization on the remaining 200 singular values, when they are compared with RMSEs of ANN weights. 90

3.6 Demonstration of the mixture components of the PAHs ($\boldsymbol{Y}$) and the electronic absorption spectrum of the 25 samples ($\boldsymbol{X}$) 91

3.7 (A) GCV Score Versus Tuning Parameters $\boldsymbol{\lambda}$, (B) Solution Path, (C) Estimated Coefficient Matrix. 92

3.8 Panel (A) exhibits the intersected region of $\mathcal{W} \leq 3$ and $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$. Panel (B) magnifies the intersected region on grid $(\omega_1, \omega_p) \in [1, 1.5] \times [1, 1.5]$. 94

A.1 Distribution of $\|e_i\|_2$ where e_i is the i^{th} row of $\widehat{E}_5^{\text{sub}} \in \mathbb{R}^{162 \times 5}$. Note that nodes classified as "Mixed Topics" are highly clustered around origin in a L_2 sense. Recall that the definition of $\widehat{E}_K$ is in Subsection 7.3. 149

D.1 Results on synthetic data. 195

SUMMARY

The phenomenal advancements in modern computational infrastructure enable the massive amounts of data acquisition in high-dimensional feature space possible. To put it more specific, the largest datasets available in the industry which often involve up to billions of samples and millions of features. The nature of datasets arising in modern science and engineering are sometimes even larger, often with the dimension of the same order as, or possibly even larger than, the sample size.

The cornerstone of modern statistics and machine learning has been a precise characterization of how well we can estimate the objects of interests under these huge high-dimensional datasets. While it remains impossible to consistently estimate in such a high-dimensional regime in general, a large body of research has investigated various structural assumptions under which statistical recovery is possible even in these seemingly ill-posed scenarios. Examples include a large line of works on sparsity [1, 2, 3, 4], low-rank assumptions [5], and more abstract generalizations of these [6, 7]. These structural assumptions on signals are often realized through specially designed norms; i.e., for inducing sparsity of either vector or matrix, entry-wise ℓ_1-norm is used; for inducing low-rank matrix, nuclear norm is used. Not only in parametric, but in non-parametric models, high-dimensional dataset is common in real world applications. A deep neural network, one of the most successful models in modern machine learning in various tasks, is a primary example of non-parametric model for function estimations. Tasks such as image classification or speech recognition often require a dataset in high-dimensional space. For the accurate function estimation avoiding the commonly known *curse of dimensionality* phenomena, some special structural assumptions on regression functions are imposed i.e., [8, 9].

Under some specific structural assumptions imposed on problems, the main emphasis in this thesis proposal is on exploring how various regularizing penalties can be utilized for estimating parameters and functions in parametric and non-parametric statistical problems.

Specifically, our main focus will be the problems in network science, PDE identification, and neural network. In the following, we describe motivations and relevant literature of each problems in detail.

Chapter 1: Network data modeling via Low rank + Sparse matrices : My first publication is about the network data modeling. While many existing community detection algorithms [10, 11, 12, 13] have focused on clustering the nodes in the network within the same communities, our model enables simultaneous inferences on the node label in the network and the ad-hoc edges that connect nodes between clusters. For instance, suppose there is a statistician who wants to find some papers that study theoretical relationships between kernel regression and neural network. Given a citation network with most recently published papers in statistical journals, models proposed in the above literature can cluster the papers on kernel regression and neural network respectively, but cannot give the statistician a satisfactory answer. The model can serve the right purpose for this statistician. We design the statistical model that can accommodate two types of dependencies encoded in the adjacency matrix of the given network: (1) Majority of the edges are generated due to the latent factors that are commonly shared among the nodes in the network; (2) Additionally, there still exist ad-hoc edges in the network that cannot be captured by the commonly shared factors, and they are relatively less common. Through a proper model parameterizations, **the latent factors** in the first component can be characterized through **a low rank matrix**, and **the ad-hoc edges** in the second component can be represented as the **a sparse matrix**. Along with this idea, the problem can be translated into decomposing model parameter into the two types of matrices; that is, **a low rank matrix** and **a sparse matrix**. Our model proves its effectiveness on various real network datasets including karate network data, political book data, and statisticians' citation network data. This work was published in Statistical Analysis and Data Mining, 2020.

Chapter 2: PDE Identification via ℓ_1-regularization : Many natural phenomena and engineering problems can be described through various Partial Differential Equation (PDE) models. (i.e., Navier-Stokes Eq, Schrödinger Eq, etc.) With the advancements in modern computing power, the acquisition of noisy dataset from the solution of these PDE models becomes accessible. My second work focuses on variable selection problem for identifying the correct PDE models that govern the data-generating process. Given the noisy observations from a certain dynamic function defined over a spatial-temporal domain, it is assumed that the dynamic function is the solution of the certain PDE model. The main goal of our work is to correctly identify the ground-truth PDE model under the noisy observations out of many possible candidate models. We prove that a minimizer from ℓ_1-penalized least-square problem can have model selection consistency guarantees for the ground-truth PDE model under three sufficient conditions on the design matrix and true signal $\beta^\star$; that is, **(1) Incoherence condition**, **(2) Minimum eigen-value condition** and **(3) $\beta^\star_{\textbf{min}}$-condition**. These three conditions also can be found in statistical literature [14, 4] where they study the model selection consistency of LASSO estimator for sparse linear regression model. This is the **first** work which gives the theoretical understandings on why ℓ_1-regularization works for PDE identification from the statistical viewpoint. This work is to appear in SIAM/ASA Journal on Uncertainty Quantification, 2022.

Chapter 3: Low-rank matrix estimation via weighted nuclear norm : Many estimators from the penalized methods suffer from biases induced from the regularizers. Researchers have put enormous efforts for reducing such biases under various statistical settings. For instance, see [15]. In my fourth work, I consider the estimation problem of low-rank matrix under the multivariate linear regression setting; that is, we consider the problem of recovering an unknown coefficient matrix $\mathbf{\Theta}^\star \in \boldsymbol{R}^{d_1 \times d_2}$ from n observations of the response vector $y_i \in \mathbb{R}^{d_2}$, $1 \leq i \leq n$, and predictor $x_i \in \mathbb{R}^{d_1}$, where the ground truth model is as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta}^\star + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{Y} = (y_1, \dots, y_n)^\top$ is an $n \times d_2$ matrix, $\boldsymbol{X} = (x_1, \dots, x_n)^\top$ is an $n \times d_1$ matrix, and $\boldsymbol{E} = (e_1, \dots, e_n)^\top$ is an $n \times d_2$ regression noise matrix. In my work, weighted nuclear norm (WNN) is used for reducing the biases in singular values of the estimated matrix. The main idea for employing WNN is to put large enough weights on the small singular values, and to put small enough weights on the large singular values of the matrix; that is, $0 \le \omega_1 \le \omega_2 \le \cdots \le \omega_p$ for $\sigma_1^\star(\boldsymbol{\Theta}^\star) \ge \sigma_2^\star(\boldsymbol{\Theta}^\star) \ge \cdots \ge \sigma_p^\star(\boldsymbol{\Theta}^\star)$, where ω_j is the j^{th} weight, which corresponds to the j^{th} singular value of the ground-truth matrix $\boldsymbol{\Theta}^\star$ denoted as $\sigma_j^\star(\boldsymbol{\Theta}^\star)$ for $1 \le j \le p = \min(d_1, d_2)$. Then, WNN is defined as $\sum_{j=1}^{p} \omega_j \sigma_j^\star(\boldsymbol{\Theta}^\star)$. However, solving the following WNN penalized least square problem (3.2) is difficult, since WNN is a non-convex function in matrix parameter space when weights are non-decreasing order. See [16].

$$\widehat{\boldsymbol{\Theta}} := \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta} \right\|_{\mathrm{F}}^2 + \boldsymbol{\lambda}_n \left\| \boldsymbol{\Theta} \right\|_{\boldsymbol{\omega},\star} \right\}. \tag{2}$$

I develop an efficient ADMM-type algorithm solving (3.2) despite its non-convexity (named as WMVR-ADMM), and study the statistical properties of $\widehat{\boldsymbol{\Theta}}$ in (3.2) under the orthogonal and random Gaussian design matrices $\mathbf{X}$, respectively. Panels in **Figure** 1**.** show the results of singular values of $\widehat{\boldsymbol{\Theta}}$ (i.e., $\widehat{\sigma}_j$) versus those of $\boldsymbol{\Theta}^\star$ (i.e, $\sigma_j^\star$) under Gaussian design $\mathbf{X}$. The first two panels (A) and (B) are results from WMVR-ADMM algorithm with one weight update iteration under $n = 250$. The panel (C) exhibits the result when the estimator is obtained from standard nuclear norm (SNN) (i.e., $\omega_1 = \omega_2 = \cdots = \omega_p = 1$) penalized least squares under $n = 1000$. The result shows that our method achieves a satisfactory de-biasing result within two iterations of loop with only sample size $n = 250$ (Panels (A) and (B)), whereas there is still a slight bias on the estimated singular value from SNN with $n = 1000$ (Panel (C)).

Chapter 4: Minimax Rate of deep ReLU net through ℓ_2-regularization : Theoretical studies on neural network models are notoriously difficult, because of its non-convex landscape of loss function in the parameter spaces. However, networks learned through noisy dataset show the good generalization abilities for the unseen data. In my third work, I studied the generalization properties of the overparameterized deep neural network (DNN) with Rectified Linear Unit (ReLU) activations. Under the non-parametric regression framework, it is assumed that the ground-truth function is from a reproducing kernel Hilbert space (RKHS) induced by a neural tangent kernel (NTK) of ReLU DNN, and a dataset is given with the noises. Without a delicate adoption of early stopping, we prove that the overparametrized DNN trained by vanilla gradient descent does not recover the ground-truth function. It turns out that the estimated DNN's L_2 prediction error is bounded away from 0. As a complement of the above result, we show that the ℓ_2-regularized gradient descent enables the overparametrized DNN to achieve the minimax optimal convergence rate of the L_2 prediction error, without early stopping. Notably, the rate we obtained is faster than $\mathcal{O}(n^{-1/2})$ known in the literature. This work is recently published in International Conference on Learning Representation (ICLR), 2022.

Chapter 5: Approximation and non-parametric estimation of functions in high dimensional spheres via deep ReLU network : In this chapter, We develop a new approximation and statistical estimation analysis of deep feed-forward neural networks (FNNs) with the Rectified Linear Unit (ReLU) activation. The functions of interests for the approximation and estimation are assumed to be from Sobolev spaces defined over the d-dimensional unit sphere with smoothness index $r > 0$. In the regime where r is in the constant order (i.e., $r = \mathcal{O}(1)$), it is shown that at most d^d active parameters are required for getting d^{-C} approximation rate for some constant $C > 0$. In the regime where the index r grows in the order of d (i.e., $r = \mathcal{O}(d)$) asymptotically, we prove the approxima-

tion error decays in the rate $d^{-d^{\beta}}$ with $0 < \beta < 1$ up to some constant factor independent of d. The required number of active parameters in the networks for the approximation increases polynomially in d as $d \to \infty$. It is also shown that bound on the excess risk has a d^d factor, when $r = \mathcal{O}(1)$, whereas it has $d^{\mathcal{O}(1)}$ factor, when $r = \mathcal{O}(d)$. We emphasize our findings by making comparisons to the results on approximation and estimation errors of deep ReLU FNN when functions are from Sobolev spaces defined over d-dimensional cube. Here, we show that with the current state-of-the-art result, d^d factor remain both in the approximation and estimation errors, regardless of the order of r.

CHAPTER 1
A NETWORK MODEL THAT COMBINES LATENT FACTORS AND SPARSE GRAPHS

1.1 Introduction

Many state-of-the-art community detection algorithms only focus on clustering nodes that share some common characteristics in a given network. The community labels in the network are unknown and the main interest is to estimate them [10, 11, 17, 18, 19, 12, 13, 20]. However, there are some cases where we need more than the label information of each node in the network. Suppose there is a statistician who wants to find some papers that study theoretical relationships between kernel regression and neural network. Given a citation network with most recently published papers in statistical journals, algorithms proposed in above literature can cluster the papers on kernel regression and papers on neural network with good statistical accuracy and reasonable computational complexity. But none of them can give a satisfactory answer to the statistician. Motivated from this idea, the present paper develops a new model that enables simultaneous inferences on node labels in the network and ad-hoc edges that connect nodes between clusters.

We review the relevant literature here. Several attempts have been made to capture interesting characteristics of networks such as degree heterogeneity, transitivity, homophily, and so on [18, 21, 22, 23]. See Wasserman and Faust [24] for more on common structures for network data. An interesting line of work in network modeling is to adopt a latent space, seeing a seminal work by Hoff et al. [25]. The key idea of the latent space model is that each node i in the network can be represented as a vector f_i in a low-dimensional latent space, which sometimes is referred as the social space in the literature [26, 25]. Nodes that are "close" in the social space are highly likely to have links among them. Many other papers generalized this approach, such as accommodating clustering effect or node homogeneity in the networks, treating the latent variables f_i's as random effects [27, 26], and many more.

To extend the latent space model, we introduce a statistical model that can accommodate pairs of connected nodes in the network, even though they are not close in latent space. More precisely, suppose that we observe a large undirected network represented by a symmetric adjacency matrix X on n nodes with $X_{ij} = X_{ji} = 1$ if nodes i and j are connected and zero otherwise. (We do not allow the self-loop, so we set the diagonal elements of the matrix X to be zero, i.e., $X_{ii} = 0$.) We design a model that can accommodate two types of dependencies encoded in the aforementioned binary matrix: (1) Majority of the edges are generated due to the latent factors that are commonly shared among the nodes in the network; (2) Additionally, there still exist ad-hoc edges in the network that cannot be captured by the commonly shared factors, and they are relatively less common.

We give more details on our statistical model. Let the observed n-by-n adjacency matrix be $X \in \mathbb{R}^{n\times n}$. The model that we are considering can be written in the following form: for any $0 < i < j \leq n$,

$$X_{ij} = X_{ji} \sim \text{Bernoulli}(P_{ij}), \text{ with } \text{logit}(P_{ij}) = \alpha + f_i^T D f_j + S_{ij}, \tag{1.1}$$

where P_{ij} is the parameter in the Bernoulli distribution, logit$(x) = \log[x/(1-x)]$, for any $x \in (0,1)$, $\alpha, S_{ij} \in \mathbb{R}$ are scalars, vectors $f_i, f_j \in \mathbb{R}^k$ are k dimensional (k is a positive integer), and diagonal matrix $D \in \mathbb{R}^{k\times k}$ has nonnegative entries. Entries of X are Bernoulli random variables that are assumed to be independent. The positively weighted inner product of latent factors f_i and f_j (i.e., $f_i^T D f_j$) corresponds to the factor model. The sparse graphical component is reflected by the presence of S_{ij}. Note that we will require that a very few of S_{ij}'s to have non-zero values (that is the matrix $S = (S_{ij})_{i,j=1,\ldots,n}$ is sparse) so that they can capture the ad-hoc dependencies of nodes in the network. Our model is named **Combined latent Factors and Sparse Graphs** model, which can be abbreviated as **CFSG**.

For the proposed model, we provide a general oracle-type inequality for the estimation error. The result is non-asymptotic; the error upper bound is a function of the number of nodes in the network. Interestingly, the result can be applied to cases when the ground-truth latent matrix in the network is either exactly or approximately low rank. Another interesting point is that we do not impose any incoherence condition [28] on the singular vectors of the factor matrix, instead we assume a milder condition called "Spikiness" on the matrix that is associated with the factor variables [29, 30]. The effectiveness of our model is demonstrated in some real-data examples.

The rest of the paper is organized as follows. Some related works are reviewed and discussed in Section 1.2. In Section 1.3, we present our model, which can encode both the latent dependent structure due to the common factors and the remaining sparse ad-hoc dependent structure. In Section 1.4, we will discuss the assumptions imposed in our model and the penalization on the likelihood function. In Section 1.5, we provide a non-asymptotic error bound of the estimator. In Section 1.6, we will present numerical experiments with synthetic data to verify the effectiveness of our method and to validate the theoretical results that are presented in Section 1.5. In Section 1.7, our model is applied to four real network datasets. We finally conclude this work in Section 1.8; some possible directions of future research are discussed as well.

Remark 1.1.1 *To make the body of this Chapter concise and to the point, we put several parts to Appendix A. In Appendix A, we provide detailed procedures on (1) how the ADMM algorithm can be employed to solve the optimization problem suggested in Section 1.4, (2) how to generate synthetic network data introduced in Subsection 1.6.1, (3) proof of the Theorem 1.5.4 presented in Section 1.5, and (4) detailed discussions on "Mixed Topics" cluster in Subsection 1.7.4.*

Remark 1.1.2 *We create a GitHub repository (https://github.com/namjoonsuh/Network-*

CFSG) to distribute R codes with documentation. The results in Section 1.6 and Section 1.7 can be reproduced by running this software.

1.2 Related Work

We describe three related works in the following, including

1. another fused latent and graphical model (Section 1.2.1), however it has very different mathematical details in formulation and analysis compared to the proposed **CFSG** model;

2. an exemplary extension of the factor-based model (Section 1.2.2), which incorporates additional external variates, however there is no sparse graph component, and

3. the large literature on matrix decomposition in general (Section 1.2.3) while our innovative contribution is to adapt the matrix decomposition strategy into the statistical modeling of network data.

1.2.1 Fused Latent and Graphical (FLaG) Model

Our work is relevant with a recent work named *Fused Latent and Graphical (FLaG) model* (Chen et al, 2016, [31]). FLaG is built to analyze the Eyesenck's Personality Questionnaire [32] that consists of questions designed to measure three aspects of human personality: Psychoticism (P), Extraversion (E) and Neuroticism (N). Questions in the questionnaire are considered as random variables, and (P), (E) and (N) are assumed to be latent attributes that are shared among the questions. Additionally, sparse graphical structure complements dependence among the questions that are not attributable to these latent factors. Thus, the resulting model contains a low-dimensional latent vector and a sparse conditional graph. Though our model (recall it is named CFSG) may seem similar to FLaG in terms of the matrix decomposition into a low-rank matrix and a sparse matrix, we work on a different model formulation in several aspects. We summarize the differences as follows.

1. In FLaG model, a collection of binary responses for each question in the questionnaire follows a joint distribution, which is a combination of the Item Response Theory (IRT) model [33] and the Ising model [34]. In CFSG, the edges in the network are modeled as random variables, whose dependent structure is characterized by the combination of the Latent Factor Analysis model and the Sparse Graphical model.

2. In FLaG, there are p questions that need to be answered, and if there are n respondents to questions, they have n independent data generated from the same distribution. In CFSG, the observed citation network can be thought of as one realization of a random graph.

3. FLaG approximates the original likelihood through constructing pseudo-likelihood function by taking advantage of conditional independence among the nodes. In CFSG, likelihood function is directly accessible due to the conditional independence among the edges.

1.2.2 An Extension of the Factor-based Model

Based upon the latent space modeling framework, Ma et al.[22] suggest a model that can incorporate additional features Z other than the edge information in the network. The additional features Z_{ij} can indicate whether the i^{th} entity and j^{th} entity of the network share the common attributes, for instance, gender. This also can be easily applied in our model by adding βZ_{ij} on the right-hand side of logistic regression model in (1.1). Here, a positive regression coefficient β reflects that if the two entities share the common attributes, the two nodes are likely to be connected. In the present paper, we choose not to pursue in this direction. It could be an interesting future work.

1.2.3 Matrix Decomposition

The proposed modeling framework is also related with the analysis of decomposing a matrix into low-rank and sparse components ([29, 28, 35, 36]). Specifically, paper [35] studies statistical inference of a multivariate Gaussian model whose precision matrix admits the form of a low-rank matrix plus a sparse matrix. The inference and optimization of the current model are different from the aforementioned cases. We will construct a regularized-likelihood function, based on what estimator will be proposed for simultaneous model selection and parameter estimation. The objective function in the optimization problem for the regularized estimator is convex, for which we will develop an efficient algorithm through the alternating direction method of multiplier (ADMM, [37, 38, 39]).

1.3 Model Formulation

Associated with a pair of nodes i and j in the network, we denote a binary random variable X_{ij}, where $1 \leq i, j \leq n$ and n is the total number of nodes. We have $X_{ij} = 1$ if and only if node i has a link with node j; otherwise $X_{ij} = 0$. For each node i, we assume that there is an associated binary vector $f_i \in \mathbb{R}^K$, such that the kth entry of f_i, $f_{ik} = 1$, if and only if node i is related to factor $k, 1 \leq k \leq K$. Here K is the total number of underlying factors. We assume a logistic model for X_{ij}'s: for $1 \leq i, j \leq n$, the model is formally defined as:

$$\mathbb{P}(X_{ij} \mid \alpha, f_i, f_j, D) := \frac{e^{X_{ij}(\alpha + f_i^T D f_j)}}{1 + e^{\alpha + f_i^T D f_j}}, \tag{1.2}$$

where $\alpha \in \mathbb{R}$ is a parameter and $D \in \mathbb{R}^{K \times K}$ is a diagonal matrix: $D = \text{diag}\{d_1, d_2, \ldots, d_K\}$. We assume $d_i > 0$ for $1 \leq i \leq K$. Another way to put (1.2) is

$$\mathbb{P}(X_{ij} = 1 \mid \alpha, f_i, f_j, D) = \frac{\exp\left(\alpha + \sum_{k=1}^{K} f_{ik} f_{jk} d_k\right)}{1 + \exp\left(\alpha + \sum_{k=1}^{K} f_{ik} f_{jk} d_k\right)}. \tag{1.3}$$

A justification of the above model is that when both node i and node j belong to factor k, they have a higher chance to have a link one way or the other. We have assumed a common strength coefficient d_k ($1 \leq k \leq K$) for factor k, despite different nodes. We denote a matrix $F = \{f_1, f_2, \ldots, f_n\} \in \mathbb{R}^{K \times n}$. Each column i in matrix F contains the factor loadings associated with the node i ($1 \leq i \leq n$). Given the diagonal matrix D and the factor loading matrix F, we assume that X_{ij}'s are independent; therefore we have the total conditional probability function as follows:

$$\mathbb{P}(\{X_{ij}, 1 \leq i, j \leq n\} \mid \alpha, F, D) = \prod_{1 \leq i < j \leq n} \mathbb{P}(X_{ij} \mid \alpha, f_i, f_j, D) = \prod_{1 \leq i < j \leq n} \frac{e^{X_{ij}(\alpha + f_i^T D f_j)}}{1 + e^{\alpha + f_i^T D f_j}}, \tag{1.4}$$

where $\mathbb{P}(X_{ij})$ is given in (1.3). The last equation holds because X_{ij} only takes binary (i.e., 0 or 1) values. Recall that the dot product of two matrices with same dimensionality, $A, B \in \mathbb{R}^{a \times b}$, is defined as $A \bullet B = \text{trace}(A^T B) = \sum_{i=1}^{a} \sum_{j=1}^{b} a_{ij} b_{ij}$. The above (1.4) can be further rewritten as

$$\mathbb{P}(\{X_{ij}, 1 \leq i < j \leq n\} \mid \alpha, F, D) = \frac{\exp\left(\alpha \sum_{1 \leq i < j \leq n} X_{ij} + \frac{1}{2} X \bullet (F^T D F)\right)}{\prod_{1 \leq i < j \leq n} 1 + e^{\alpha + f_i^T D f_j}}, \tag{1.5}$$

where we assume $X_{ii} = 0$ for all i ($1 \leq i \leq n$) and $X_{ij} = X_{ji}$ for all i and j ($1 \leq i, j \leq n$), i.e., the matrix X is symmetric. The above delivers a factor analysis model. Various linear and nonlinear latent variable models have been studied extensively in the literature (e.g., [40, 41, 42, 43, 44, 45]).

The above specifies a latent model (or equivalently a factor model). We now describe a graphical model as follows. The graphical model will complement the latent model by characterizing links that are not interpretable via common factors. For the aforementioned binary random variable X_{ij}, $1 \leq i, j \leq n$, similarly with (1.2), we define

$$\mathbb{P}(X_{ij} \mid \alpha', S_{ij}) := \frac{e^{X_{ij}(\alpha' + S_{ij})}}{1 + e^{\alpha' + S_{ij}}}, \tag{1.6}$$

where $S_{ij} \in \mathbb{R}$, for $1 \leq i, j \leq n$, denotes the relation between nodes i and j. Note that the matrix S is introduced to capture the ad-hoc links in the graph. If we have $S_{ij} \leq 0$, then it is less likely to have a citational relationship between nodes i and j. On the other hand, if $S_{ij} > 0$, then it is more likely to have a citation link between nodes i and j. Here parameter $\alpha' \in \mathbb{R}$ plays the same role as parameter α does in model (1.2). Denote the matrix $S = \{S_{ij}, 1 \leq i, j \leq n\} \in \mathbb{R}^{n \times n}$. Assume that given the matrix S, the binary random variables X_{ij}'s are independent; consequently, we have the total conditional probability function as follows:

$$\begin{aligned} \mathbb{P}(\{X_{ij}, 1 \leq i, j \leq n\} \mid \alpha', S) &= \prod_{1 \leq i < j \leq n} \mathbb{P}(X_{ij} \mid \alpha', S_{ij}) = \prod_{1 \leq i < j \leq n} \frac{e^{X_{ij}(\alpha' + S_{ij})}}{1 + e^{\alpha' + S_{ij}}} \\ &= \frac{\exp\left(\alpha' \sum_{1 \leq i < j \leq n} X_{ij} + \frac{1}{2} X \bullet S\right)}{\prod_{1 \leq i < j \leq n} 1 + e^{\alpha' + S_{ij}}}. \end{aligned} \tag{1.7}$$

Recall that we have assumed that $X_{ii} = 0$ for all i ($1 \leq i \leq n$) and $X_{ij} = X_{ji}$ for all i and j ($1 \leq i, j \leq n$), i.e., the matrix X is symmetric. In the combined model, we integrate (1.5) and (1.7) to render the joint conditional probability function as follows:

$$\begin{aligned} \mathbb{P}(X \mid \alpha, F, D, S) &= \prod_{1 \leq i < j \leq n} \frac{e^{X_{ij}(\alpha + S_{ij} + f_i^T D f_j)}}{1 + e^{\alpha + S_{ij} + f_i^T D f_j}} \\ &= \frac{\exp\left(\alpha \sum_{1 \leq i < j \leq n} X_{ij} + \frac{1}{2} X \bullet (F^T D F) + \frac{1}{2} X \bullet S\right)}{\prod_{1 \leq i < j \leq n} \left(1 + e^{\alpha + f_i^T D f_j + S_{ij}}\right)}. \end{aligned} \tag{1.8}$$

1.4 Estimation

Note that in the model (1.8), the log-likelihood function has the form as follows:

$$\begin{aligned} \mathbb{L}(\alpha, F, D, S; X) &= \alpha \sum_{1 \leq i < j \leq n} X_{ij} + \frac{1}{2} X \bullet (F^T D F) + \frac{1}{2} X \bullet S \\ &\quad - \sum_{1 \leq i < j \leq n} \log\left(1 + e^{\alpha + f_i^T D f_j + S_{ij}}\right). \end{aligned} \tag{1.9}$$

If we consider maximizing the above log-likelihood function, we will encounter several technical issues that are listed below.

1. We would like the matrix $S \in \mathbb{R}^{n \times n}$ to have as many zero entries as possible; i.e., matrix S is *sparse.*

2. There is an identifiability issue with the formation $F^T DF$. More specifically, let $P \in \mathbb{R}^{K \times K}$ be a signed permutation matrix, then we have $P^T P = I_n$, where $I_n \in \mathbb{R}^{K \times K}$ is the identity matrix. Notice that matrix $F' = PF$ is also a factor loading matrix, and matrix $D' = PDP^T$ is still a diagonal matrix; we have

   $$F^T DF = F^T P^T PDP^T PF = (F')^T D' F',$$

   i.e., the choice of F and D is not unique.

3. We would like the number of nonzeros in each column of F to be small, reflecting that each node is associated with a small number of underlying topics.

4. Overall, the rank of matrix $F^T DF$ cannot be larger than $\min\{n, K\}$. With the application that we have in mind, in this paper, we assume that K is much smaller than n.

5. Following the approaches that were mentioned in the Introduction, we propose to relax $F^T DF$ to L, where L is a low rank matrix. Furthermore, to ensure the separation of matrices $\alpha \mathbb{1}\mathbb{1}^T$ and an arbitrary matrix L, we assume that the eigen-vector of L is centered, that is,

   $$JLJ = L \quad \text{where} \quad J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T, \tag{1.10}$$

   where $\mathbb{1}$ denotes a n-dimensional vector whose entries are all 1's. Since we have $L = F^T DF$, this condition uniquely identifies F up to a common orthogonal transformation of its columns.

Directly maximizing the objective function in (1.9) is not going to be an easy task. Consequently, the log-likelihood function in (1.9) can be rewritten as

$$\begin{aligned}\mathbb{L}_n(\alpha, L, S; X) &= \alpha \sum_{1 \le i < j \le n} X_{ij} + \frac{1}{2} X \bullet L + \frac{1}{2} X \bullet S \\ &\quad - \sum_{1 \le i < j \le n} \log\left(1 + e^{\alpha + L_{ij} + S_{ij}}\right).\end{aligned} \tag{1.11}$$

We propose a penalized likelihood estimation approach as follows:

$$(\hat{\alpha}, \widehat{L}, \widehat{S}) = \text{arg min}_{\alpha, L, S} \left\{ -\frac{1}{n} \mathbb{L}_n(\alpha, L, S; X) + \gamma \|S\|_1 + \delta \|L\|_* \right\}, \tag{1.12}$$

where $\gamma > 0$ and $\delta > 0$ are algorithmic parameters whose values will be discussed later (Section 1.6.2), the L_1 norm of matrix S is defined as $\|S\|_1 = \sum_{i \neq j} |S_{ij}|$ (Note that we do not penalize the diagonal entries of S), and nuclear norm of matrix L is defined as $\|L\|_* = \text{trace}\sqrt{(L^T L)}$. Recall that both S and L are symmetric matrices. The entries of matrix S can either be positive or negative. Note that we have imposed the diagonal entries of the matrix X to be zeros. Given that $L = F^T D F$ where matrix D is diagonal with nonnegative diagonal entries, it is easy to see that matrix L is positive semidefinite; which consequently leads to $\|L\|_* = \text{trace}(L)$, which is a linear functional to the matrix L. The nuclear norm of L mimicks the number of nonzero eigenvalues of L, which is the same as the rank of L. The regularization based on the nuclear norm was proposed in [46] and its statistical properties are studied in [47].

After we have obtained $\widehat{S}$ in (1.12), we can uncover the graphical model by investigating non-zero entries in $\widehat{S}$. On the other hand, when we have calculated $\widehat{L}$, we may not be able to find binary matrix F and nonnegative diagonal matrix D such that $\widehat{L} = F^T D F$. This is the price we have to pay for an amenable computational approach. The rank of estimated $\widehat{L}$ will be our estimate of the number of factors (i.e., the number of underlying common factors). We will discuss the issue on assigning the topic membership of each node i later in Subsection 1.6.3. Detailed description on the ADMM algorithm that we employed for optimizing (1.12) can be found in the Appendix $A.1$.

1.5 Non-asymptotic error bound of the estimator

In this section, we focus on investigating the behaviour of non-asymptotic error bound of our estimator in the context where the number of nodes in a network is explicitly tracked. Recall that we are interested in solving the following optimization problem:

$$\min_{\substack{\alpha\in R, S=S^T\\ L\succcurlyeq 0}} -\frac{1}{n}\log\prod_{1\leq i,j\leq n}\frac{\exp\left(X_{ij}\left(\alpha+L_{ij}+S_{ij}\right)\right)}{1+\exp\left(\alpha+L_{ij}+S_{ij}\right)}+\delta\|L\|_*+\gamma\|S\|_1. \tag{1.13}$$

For the convenience of theoretical investigation, we slightly modify the first term in the objective function summing over all (i,j) pairs. After scaling, due to symmetry of X, L, and S, the only difference between (1.12) and (1.13) is in the inclusion of terms in diagonal pairs $(i,i), \forall i=1,\ldots,n$. We borrow the idea of this modification from the work of [22], where they also consider the latent factor model in analyzing the embedded topics in the network but without the sparse component. Not only do this modification simplify the theoretical investigation of the estimator, but also it makes no differences in quality of the estimator, as will be demonstrated in Subsection 1.6.4.

Let $(\widehat{\alpha},\widehat{L},\widehat{S})$ be the solution to (1.13), and (α^*,L^*,S^*) be the ground truth, which governs the data generating process. Let $\widehat{\Theta}$ and Θ^* be defined respectively as $\widehat{\Theta}=\widehat{\alpha}\mathbb{1}\mathbb{1}^T+\widehat{L}+\widehat{S}$ and $\Theta^*=\alpha^*\mathbb{1}\mathbb{1}^T+L^*+S^*$. And denote the error term for each parameter as $\widehat{\Delta}^\Theta=\widehat{\Theta}-\Theta^*, \widehat{\Delta}^\alpha=\widehat{\alpha}-\alpha^*, \widehat{\Delta}^L=\widehat{L}-L^*, \widehat{\Delta}^S=\widehat{S}-S^*$. Throughout the discussion, let $P^*=\left\{\frac{\exp(\Theta^*_{ij})}{1+\exp(\Theta^*_{ij})}\right\}_{1\leq i,j\leq n}\in\mathbb{R}^{n\times n}$. We describe several assumptions before establishing the theoretical guarantees of our estimator.

Assumption 1.5.1 *(**Strong convexity**) For any $\Theta\in\mathbb{R}^{n\times n}$, define the log-likelihood in (1.13):*

$$h(\Theta)=-\frac{1}{n}\sum_{i,j}\left\{X_{ij}\Theta_{ij}-\log(1+\exp(\Theta_{ij}))\right\}.$$

We assume that $h(\Theta)$ is τ-strongly convex in a sense that lowest eigenvalue of Hessian matrix of the log-likelihood function is bounded away from zero ($\tau > 0$):

$$\nabla^2 h(\Theta) = diag\Big(vec\Big(\frac{1}{n}\frac{\exp(\Theta)}{(1+\exp(\Theta))^2}\Big)\Big) \succcurlyeq \tau I_{n^2\times n^2}.$$

For any vector a, diag(a) is the diagonal matrix with elements of a on its diagonal. For any matrix $B = [b_1, \dots, b_n] \in \mathbb{R}^{n\times n}$, vec$(B) \in \mathbb{R}^{n^2}$ is obtained by stacking $b_1, \dots, b_n$ in order. For any square matrix A and B, we have $A \succcurlyeq B$ if and only if matrix $A - B$ is positive semi-definite.

Assumption 1.5.2 *(**Identifiability of** $\alpha\mathbb{1}\mathbb{1}^T$ **and** L) To ensure the separation between $\alpha\mathbb{1}\mathbb{1}^T$ and L, we assume that the latent variables are centered, that is $JL = L$, where $J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$, where $\mathbb{1}$ denotes an all one vector in $\mathbb{R}^n$.*

Assumption 1.5.3 *(**Spikiness of** L **and Constraint on** α) We impose a spikiness condition $\|L\|_\infty \le \frac{\kappa}{\sqrt{n\times n}}$ on L, to ensure the separation of L and matrix S [29]. We would also like to note that the constraint $|\alpha| \le C\kappa$, for an absolute constant C, is included partially for obtaining theoretical guarantees.*

Under these assumptions, we present the behavior of non-asymptotic error bound of our estimator through the following theorem. In our result, we measure error using squared Frobenius norm summed across three matrices:

$$e^2\big(\widehat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}\big) := \big\|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\big\|_F^2 + \big\|\widehat{\Delta}^L\big\|_F^2 + \big\|\widehat{\Delta}^S\big\|_F^2. \tag{1.14}$$

Theorem 1.5.4 *Under the Assumptions 1.5.1, 1.5.2 and 1.5.3, if we solve the convex problem (1.13) with a pair of regularization parameter (δ, γ) satisfying*

$$\delta \ge 2\left\|\frac{1}{n}(X - P^*)\right\|_{op} \quad and \quad \gamma \ge 2\left\|\frac{1}{n}(X - P^*)\right\|_\infty + 4\kappa\tau\Big(\frac{Cn+1}{n}\Big), \tag{1.15}$$

where $\| * \|_{op}$ stands for the matrix operator norm (i.e., the largest singular value of the matrix), then there exist universal constants c_j, $j = 1, 2, 3$, for all integers $k = 1, 2, ..., n$, and $s = 1, 2, ..., n^2$, and we have the following upper bound of $e^2\big(\widehat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}\big)$:

$$e^2\big(\widehat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}\big) \le \underbrace{c_1\frac{\delta^2}{\tau^2}}_{\mathcal{K}_{\alpha^*}} + \underbrace{c_2\frac{\delta^2}{\tau^2}\left\{k + \frac{\tau}{\delta}\sum_{j=k+1}^{n}\sigma_j(L^*)\right\}}_{\mathcal{K}_{L^*}} + \underbrace{c_3\frac{\gamma^2}{\tau^2}\left\{s + \frac{\tau}{\gamma}\sum_{(i,j)\notin M}|S^*_{ij}|\right\}}_{\mathcal{K}_{S^*}}, \tag{1.16}$$

where M is an arbitrary subset of matrix indices of cardinality at most s.

We would first like to note that the result presented in Theorem 1.5.4 can be thought of as an extension of Theorem 1 presented in paper [29] to a generalized linear model. Specifically, our work considers a logistic loss function whose parameter is characterized by a sparse matrix plus a low rank matrix, whereas Agarwal, et al. [29] work on a general linear observation model whose parameter is also characterized by a sum of a low rank matrix and a sparse matrix.

Astute readers might have noticed that the upper bound in (1.16) consists of three different terms, where we denote them as $\mathcal{K}_{\alpha^*}$, $\mathcal{K}_{L^*}$ and $\mathcal{K}_{S^*}$. Each respective term is involved with estimating three model parameters: α, L and S. To be more specific, both $\mathcal{K}_{L^*}$ and $\mathcal{K}_{S^*}$ have two types of error: 1) The first one is called as an "estimation error." This error represents the statistical cost of estimating parameters that belong to the model subspace. 2) Another quantity is referred as "approximation error." This error occurs when we only focus on estimating parameters within the model subspace, and it shrinks as the model subspace becomes large.

The result of the Theorem 1.5.4 provides a family of upper-bounds, one for each indexed by a specific choice of model subspace M, and rank parameter k. In other words, this means that the subset M and the target rank k can be adaptively chosen so as to obtain the tightest upper bound. In an ideal case where L^* is an exact low rank matrix with rank

k $\big($i.e., rank$(L^*) = k\big)$ and S^* is a sparse matrix, whose support lies within the model subspace M $\big($i.e., supp$(S^*) \subset M\big)$, we can easily see "*approximation error*" terms in $\mathcal{K}_{L^*}$ $\big($i.e., $\delta \sum_{j=k+1}^{n} \sigma_j(L^*)$ $\big)$ and in $\mathcal{K}_{S^*}$ $\big($i.e., $\gamma \sum_{(i,j)\notin M} |S^*_{ij}|\big)$ disappear, giving us Frobenius error bound as follows:

$$e^2(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}) \lesssim \delta^2(k+1) + \gamma^2 s.$$

Here we use the notation $X \lesssim Y$ to denote the fact that there exists a universal absolute constant C such that $X \leq CY$.

However, in a realistic setting, we can rarely observe network data with both the exact low rank matrix L^* and the exact sparse matrix S^*. The beauty of Theorem 1.5.4 lies on the characterization of the estimation error bound in this situation as well. Suppose a tuple of network parameters, $\big(\hat{\alpha}, \widehat{L}, \widehat{S}\big)$, is obtained from a specific choice of tuning parameter pair (γ, δ). Provided that the rank of $\widehat{L}$ is $k(< n)$, then $\widehat{L}$ is a rank k approximation of L^* matrix. In $\mathcal{K}_{L^*}$, the term $\delta \sum_{j=k+1}^{n} \sigma_j(L^*)$ corresponds to the approximation error associated with the representing L^* matrix, which possibly is a full rank matrix. A similar interpretation can be applied to the terms in $\mathcal{K}_{S^*}$ as well.

1.6 Numerical experiments with synthetic data

We present our numerical experiments here. First, in Subsection 1.6.1, we introduce two synthetic scenarios that we want to explore. In Subsection 1.6.2, we describe three model selection criteria and four evaluation metrics for the selected model. Subsequently, we elaborate experimental results from the synthetic networks and several interesting findings from those results in Subsection 1.6.3. In Subection 1.6.4, a numerical experiment is presented to validate the theoretical properties that are introduced in Section 1.5. Lastly, a simple numerical experiment on running time of our algorithm is presented in Subsection 1.6.5. All numerical experiments presented in this paper are performed via statistical software R in an Intel(R) Core(TM) i7-8700 3.20GHz computer with 16 GB Ram.

1.6.1 Synthetic Setting

In this Subsection, we describe two synthetic scenarios that will be adopted in Subsections 1.6.2 and 1.6.3. Readers can refer to the Appendix $A.2$. for the detailed steps on the network data generation, together with the setting for the ground truth parameters α^*, F^*, D^* and S^*. We put astroid in the superscripts of parameters to indicate that they are the ground truth.

In each scenario, we generate three synthetic networks as follows.

1. In the first scenario, we consider three networks, in which each of them consists of nodes with only one topic. Specifically, we consider three networks

$$\{(n^{(i)}, n_1^{(i)}, K^{(i)}, |S^*|^{(i)})\}_{i=1}^3 = \{(30, 30, 3, 9), (80, 80, 4, 18), (120, 120, 5, 30)\}.$$

   For example, the notation $(n^{(1)}, n_1^{(1)}, K^{(1)}, |S^*|^{(1)}) = (30, 30, 3, 9)$ means that we generate a network with 30 nodes. All the 30 nodes correspond to a single hidden factor (correspondingly, $n_1^{(1)} = 30$). There are 3 factors embedded in the network, and 9 random ad-hoc links connect the 3 clusters of papers.

2. In the second scenario, we consider three networks, in which each of them has some nodes that can be associated with more than one factor. Particularly, we consider

$$\begin{aligned}&\{(n^{(i)}, n_2^{(i)}, n_3^{(i)}, K^{(i)}, |S^*|^{(i)})\}_{i=4}^6 \\ &\qquad = \{(120, 0, 10, 3, 18), (210, 50, 0, 3, 18), (210, 10, 10, 3, 18)\}.\end{aligned}$$

   For example in the third case, we have a network with 210 nodes in total. There are 3 factors commonly shared across the network. Among 210 nodes, 10 nodes randomly share 2 factors out of 3, other 10 nodes have 3 mixed factors, whereas the remaining

190 nodes only relate to 1 factor. The 3 clusters from these 190 nodes are connected through 18 random ad-hoc links.

All six networks that are elaborated in scenario 1 and 2 are visualized in Figure 1.1. In the first three cases, the nodes that share the common factors are clustered, and the cross clustered links (green ones) are the ad-hoc edges. In the case 4 to 6, due to the presence of nodes with multiplicity, the clustering pattern become less clear. The red edges are not the ad-hoc ones; they suppose to be within cluster edges. However due to the multiplicity and the change of the general pattern, they appear like inter-cluster edges.

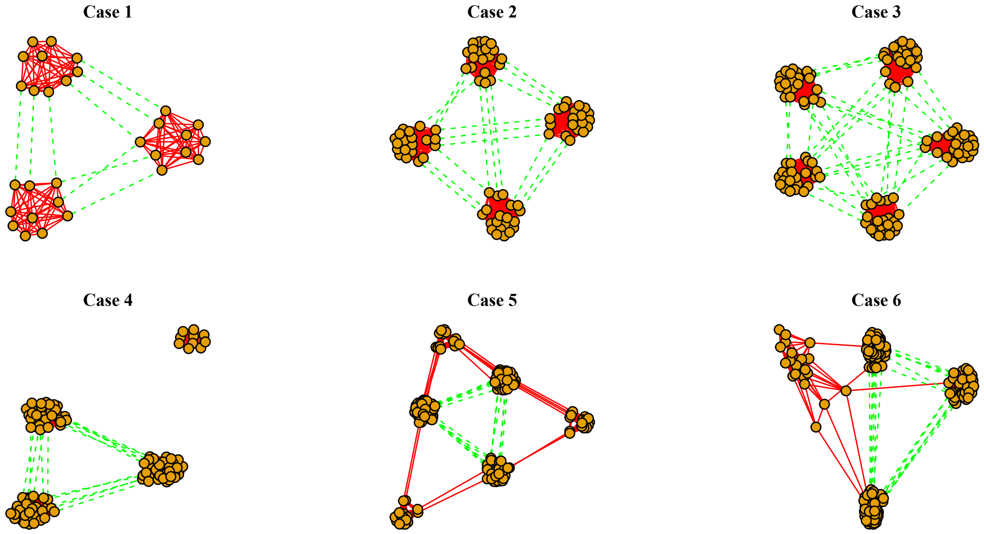


Figure 1.1: Graphical illustrations of six synthetic networks. Nodes that share the common factors are clustered. The cross cluster links are ad-hoc Citations. All the graphs are drawn based on the algorithm in [48].

1.6.2 Choosing the tuning parameters and evaluation criteria

Heuristic Network Cross-Validation. We present a two-step procedure for choosing a good pair of tuning parameters (γ, δ), which are critical in implementing the method in (1.12). The first step is to get a proper range of a grid to search over using the scree-plot analysis; the second step is to select a pair of tuning parameters using the network cross-validation. Specific procedure is elaborated as follows:

1. **First Step.** Following the scree-plot approach in Ji and Jin [49], we plot the largest 15 eigenvalues of the adjacency matrix X, and find an "elbow" point where the eigenvalues seem to level off. An index of the point, which is to the left of this elbow point, is considered as the number of the communities embedded in the network. (We will denote this number as $\widehat{K}^{\text{Scree}}$.) We want to note that the scree-plot analysis serves as a good approach for determining the range of grids to search over. With the estimate of the number of communities in the network in mind, we record the rank$(\widehat{L}^{\gamma,\delta})$ for each tuning parameter pair on a given grid. We need to go through several iterations of this recording procedure to find a proper range of grid, in which we can get rank$(\widehat{L}^{\gamma,\delta}) = \widehat{K}^{\text{Scree}}$. Here, we denote G as the grid obtained in this step. We set the grid size $|G|$ as 10×10 for numerical experiments presented in Subsections 1.6.3 and 1.6.4. For more detailed information on the grid range for each of network dataset, readers can refer the provided code.

2. **Second Step.** Given G, following the idea presented in [50, 22, 51], we suggest to use network cross-validation for choosing a proper pair of tuning parameters (γ, δ). For each pair of tuning parameters (γ, δ) on the grid, G, we do following three steps sequentially:

   (a) Randomly partition n nodes in the network into I_1 and I_2 with $|I_1| = \lfloor \frac{n}{2} \rfloor$ and $|I_2| = n - \lfloor \frac{n}{2} \rfloor$, where $| \cdot |$ denotes the cardinality of a set.

   (b) Optimization problem (1.12) is solved with the $n \times n$ graph $\tilde{X} = \{\tilde{X}_{ij}\}_{i,j=1}^{n}$ in place of X, where

   $$\tilde{X}_{ij} = \begin{cases} 0 & \text{if } i, j \in I_2 \\ X_{ij} & \text{otherwise.} \end{cases}$$

   Then test the fitted models with edges in $\{(i, j) : i \in I_2 \text{ and } j \in I_2\}$ by calculating mis-classification rate.

   (c) Repeat (a) and (b) 10 times and take the average of 10 mis-classification rates.

We choose a pair of parameters whose averaged mis-classification rate is minimal over a set of pairs $\{(\gamma, \delta) \in G : \text{rank}(\widehat{L}^{\gamma,\delta}) = \widehat{K}^{\text{Scree}}\}$. The chosen pair of tuning parameters is denoted as $(\gamma^{HNCV}, \delta^{HNCV})$, where HNCV stands for Heuristic Network Cross-Validation.

BIC and AIC. One might wonder how the traditional model selection methods, such as the Bayes Information Criterion (BIC [52]) and the Akaike information criterion (AIC), work. Recall that BIC and AIC are defined as follows:

$$\text{BIC}(M) = -2\mathbb{L}_n(\hat{\beta}(M)) + |M| \log\left(\frac{n(n-1)}{2}\right),$$

and

$$\text{AIC}(M) = -2\mathbb{L}_n(\hat{\beta}(M)) + 2|M|.$$

Here M indicates the current model, which is implicitly understood that the model is obtained from certain tuning parameter pair (γ, δ). We use $\mathbb{L}_n(\hat{\beta}(M))$ to denote the maximal log-likelihood for a given model M, and $|M|$ is the number of free parameters in M, which is determined by the number of non-zeros in $\widehat{S}^{\gamma,\delta}$ and the low-rank matrix $\widehat{L}^{\gamma,\delta}$. In detail, if we have $\text{rank}(\widehat{L}^{\gamma,\delta}) = K$, we can establish the following

$$|M| = \sum_{i<j} 1_{\{S_{ij} \neq 0\}} + nK - \frac{K(K-1)}{2} + 1;$$

since the number of free parameters in $\widehat{L}^{\gamma,\delta}$ is K plus $nK - K(K+1)/2$, which is the number of free parameters in determining K orth-normal vectors. Additional 1 in the last term is due to $\hat{\alpha}$. We want to find a pair (γ, δ), which minimizes BIC(M) or AIC(M) as a function of (γ, δ), respectively, where we denote them as follows:

$$(\gamma^{BIC}, \delta^{BIC}) := \text{arg min}_{\gamma,\delta}\text{BIC}(M), \quad (\gamma^{AIC}, \delta^{BIC}) := \text{arg min}_{\gamma,\delta}\text{AIC}(M).$$

Evaluation. We evaluate the models that are selected via our heuristic approach, BIC, and AIC by using the following four evaluation metrics:

$$\begin{aligned}
M_1 &= \mathbb{1}\big\{\text{rank}(\widehat{L}) = \text{rank}(L^*)\big\}, \\
M_2 &= \frac{\Big|\big\{(i,j) : i < j : S^*_{i,j} \neq 0 \quad \& \quad \widehat{S}_{i,j} \neq 0\big\}\Big|}{\big|\big\{(i,j) : i < j : S^*_{i,j} \neq 0\big\}\big|}, \\
M_3 &= \Big|\big\{(i,j) : i < j : S^*_{i,j} = 0 \quad \& \quad \widehat{S}_{i,j} \neq 0\big\}\Big|, \\
M_4 &= \frac{\big|\big\{\text{Mis-classified Nodes}\big\}\big|}{n},
\end{aligned}$$

where M_1 is a metric on whether the selected model recovers the true low rank structure of network, M_2 evaluates the positive selection rate of the sparse ad-hoc structure in network, M_3 evaluates the false discoveries of ad-hoc edges, and M_4 calculates the proportion of mis-classified nodes to the entire nodes in the network. With properly selected tuning parameter, M_1 will be 1, M_2 will be close to 1, and M_3 and M_4 will get close to 0. We present the evaluation results on the six networks created in Subsection 1.6.1 via the four criteria, M_1, M_2, M_3 and M_4 in Table. 1.1.

1.6.3 Several Observations

We set the Truncated-SVD parameter as $r = 15$ for all experiments performed in this Subsections 1.6.2 and 1.6.3.

1. **Node Membership.** After fitting the model with a proper pair of tuning parameters, (γ, δ), we need to determine whether the ith node relates to the kth factor or not. We denote rank of matrix $\widehat{L}$ as K. Let $\widehat{L} = UDU^T$ denote its eigenvalue-decomposition (EVD), where the columns of unitary matrix $U = [u_1, \ldots, u_K] \in \mathbb{R}^{n \times K}$ contain the eigen-vectors, and diagonal matrix $D = \text{diag}(\lambda_1, \ldots, \lambda_K) \in \mathbb{R}^{K \times K}$ stores the eigenvalues arranged in decreasing order. We apply simple K-means clustering algorithm on $\widehat{E}_K := U\sqrt{D} \in \mathbb{R}^{n \times K}$ treating each row of the matrix $\widehat{E}_K$ as a new data point.

Table 1.1: For two scenarios, our heuristic method chooses the model with $\widehat{L}$ with true rank, $\widehat{S}$ whose M_2 value is close to 1, and M_3 value is close to 0. Also note that it chooses a model whose mis-classification rate is close to 0. A number in the parentheses represents the rank of $\widehat{L}$ estimated from (γ^{HNCV}, δ^{HNCV}), (γ^{AIC}, δ^{AIC}) and (γ^{BIC}, δ^{BIC}) for each case.

	Scenario 1								
	Case 1			Case 2			Case 3		
	HNCV	AIC	BIC	HNCV	AIC	BIC	HNCV	AIC	BIC
M_1	1 (3)	0 (2)	0 (2)	1 (4)	0 (3)	0 (3)	1 (5)	0 (4)	0 (4)
M_2	9/9	0/9	0/9	17/18	0/18	0/18	30/30	30/30	30/30
M_3	4	0	0	0	0	0	0	0	0
M_4	0	0	0	0	20/80	20/80	0	24/120	24/120

	Scenario 2								
	Case 4			Case 5			Case 6		
	HNCV	AIC	BIC	HNCV	AIC	BIC	HNCV	AIC	BIC
M_1	1 (3)	0 (2)	0 (2)	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)
M_2	17/18	0/18	0/18	18/18	0/18	0/18	18/18	0/18	0/18
M_3	0	0	0	0	0	0	6	0	0
M_4	0	47/120	47/120	0	0	0	1/210	1/210	1/210

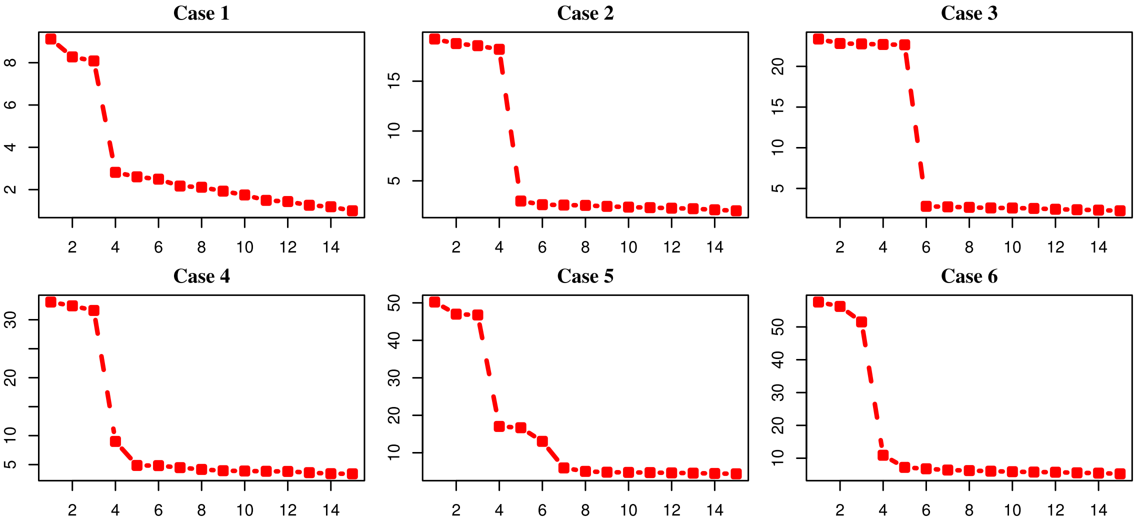


Figure 1.2: Scree plots for six synthetic Networks. $\widehat{K}^{\text{Scree}}$ recovers the number of topics in the network correctly for all six cases.

2. **Model Selection.** Choosing a good pair of tuning parameters is critical when it comes to making a good statistical inference on data. As presented in Table. 1.1, both BIC and AIC, which are well known for their model selection consistency in asymptotic setting, appear to under-estimate both the number of communities and the number of ad-hoc links in the networks in our synthetic settings. This may be caused by the fact that these traditional methods take the sample size into account, and therefore penalizes the model complexity too harshly. On the other hand, HNCV approach performs relatively better than AIC and BIC in terms of evaluation metrics M_1 and M_2. In the HNCV, scree-plot plays an important role when it comes to recovering the number of communities, and this strategy leads to good model selection results for 6 cases considered in two scenarios. See Figure 1.2.

3. **Clustering patterns on latent space.** It is interesting to observe that clustering patterns of nodes are clearly displayed on the plot of the first two leading eigenvector of $\widehat{L}^{\text{HNCV}}$. See Figure 1.3. Additionally, for cases 4, 5, and 6, it is noticed that the nodes with mixed memberships can be identified distinctly as separate clusters on the plane. In Figure 1.3, we distinguish clusters of nodes assigned by K-means algorithm with different colors. For cases 1 to 3, when applying K-means algorithm, we set the number of clusters to be detected as $K^{\text{HNCV}} = \text{rank}(\widehat{L}^{\text{HNCV}})$, $K^{\text{AIC}} = \text{rank}(\widehat{L}^{\text{AIC}})$, and $K^{\text{BIC}} = \text{rank}(\widehat{L}^{\text{HNCV}})$, respectively. For cases 4 to 6, we count the number of distinct clusters plotted on the plane, and set it as the number of clusters to be identified. Subsequently, we run the K'-means algorithm on $\widehat{E}_K$. For instance, in case 4, K' is set as 4, and K'-means algorithm is performed on $\widehat{E}_3$, where we use HNCV approach for model selection. Among three approaches suggested in Subsection 1.6.2, HNCV approach gives the most satisfactory results for the six cases in terms of the M_4 metric, seeing Table. 1.1.

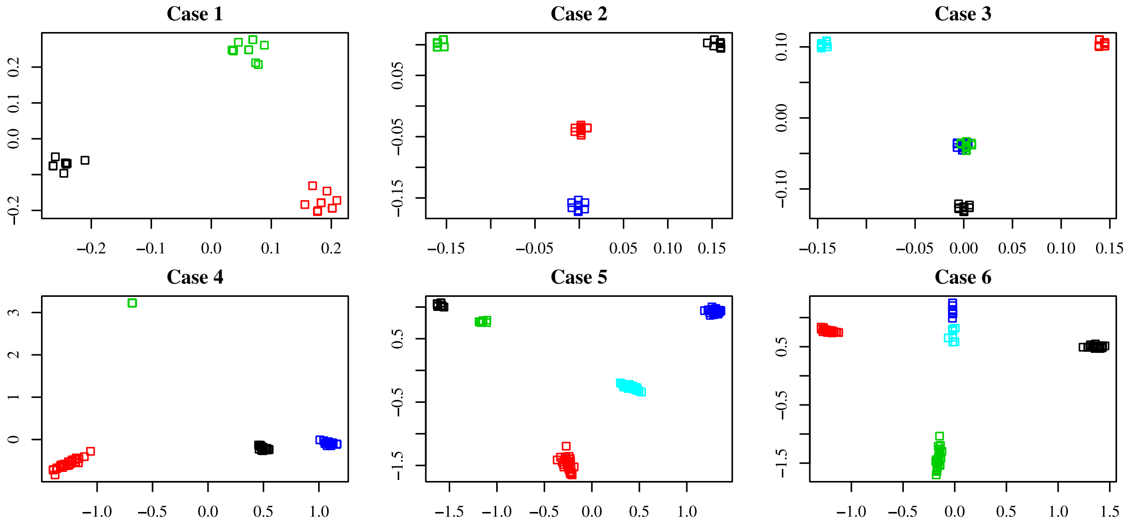


Figure 1.3: Case 1 to 6: Plots of rows based on the two leading eigenvectors of $\widehat{L}^{\text{HNCV}}$.

1.6.4 Estimation Error

In Section 1.5, we see that the estimation error — $e^2(\widehat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S})$ in (1.14) — increases linearly with the rank of L^* matrix and the cardinality of S^* matrix, where we denote them as k and s respectively (i.e., $\text{rank}(L^*) = k$ and $|S^*| = s$). In order to demonstrate the agreement between these theoretical predictions and the behaviour of the estimator in practice, two sets of experiments are designed as follows:

1. With the number of nodes and the sparsity of network being fixed as $n = 100$ and $|S^*| = 50$, we generate four networks with rank parameter $k = 4, 5, 6, 7$.

2. With the number of nodes and the rank of L^* being fixed as $n = 100$ and $\text{rank}(L^*) = 4$, we generate four networks with sparsity $|S^*| = 18, 30, 42, 54$.

All the eight networks described above are independently generated through a 5-step procedure described in Section 1.6.1. Note that we don't allow the mixed membership of each node in the networks. We normalize columns of $\widehat{L}$ and L^* to make them satisfy the spikiness condition imposed in the theoretical analysis. We define $\widehat{\Delta}_N^L = \widehat{L}_N - L_N^*$ and calculate $\|\widehat{\Delta}_N^L\|_F^2$. Also, recall $\widehat{\Delta}^S = \widehat{S} - S^*$. All eight models are selected through the HNCV approach. In Figure 1.4, we can observe linear growth of $\|\widehat{\Delta}_N^L\|_F^2$ and $\|\widehat{\Delta}^S\|_F^2$ on the rank parameter k and sparsity s in the two scenarios. Additionally, we refer to the algorithms for obtaining minimizers of optimization problems (1.12) and (1.13) as Estimator 1 and Estimator 2, respectively. Estimator 2 can be easily obtained by optimizing the objective function $(A.1)$ in Appendix $A.1$ over the whole indices $1 \leq i, j \leq n$. We can see no differences in quantities of $\|\widehat{\Delta}_N^L\|_F^2$ and $\|\widehat{\Delta}^S\|_F^2$ when the they are obtained from these two estimators.

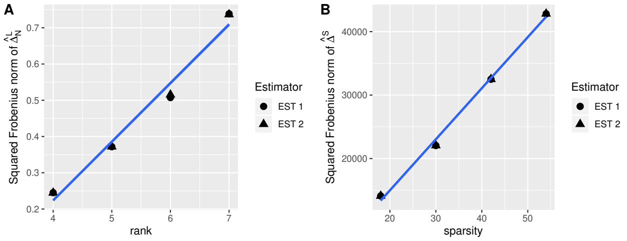


Figure 1.4: We can observe linear growth of $\|\widehat{\Delta}_N^L\|_F^2$ and $\|\widehat{\Delta}^S\|_F^2$ on the rank parameter k (panel A) and sparsity s (panel B) for the two respective scenarios. Also note that there are no differences in quantities of $\|\widehat{\Delta}_N^L\|_F^2$ and $\|\widehat{\Delta}^S\|_F^2$ when they are obtained from Estimator 1 and Estimator 2.

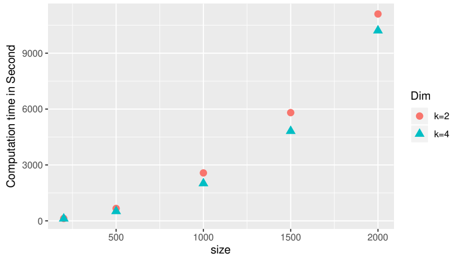


Figure 1.5: Plot of the computational runtime (in seconds) of ADMM algorithm using the HNCV approach.

1.6.5 Computation Time

To test the scalability of the proposed ADMM algorithm, we record the runtimes of simulations for different sizes of network and different dimensions of the latent vectors. We record the computation time of Step 2 of HNCV from start through finish for one pair (γ, δ). In this experiment, we simply put $(\gamma, \delta) = (0, 0)$, and stopping criteria of the algorithm as 10^{-4}. (See Section 1 in Appendix $A.1$ for the definition of stopping criteria.) For computational convenience, we repeat (a) and (b) of Step 2 five times, instead of ten times. Two sets of networks are generated where the latent dimensions are $k = 2$ and $k = 4$ with varying network size $n = 200, 500, 1000, 1500, 2000$, respectively, and set $|S^*| = 50$ for all created networks. As Figure 1.5 indicates, the runtime of algorithm do not seem to be sensitive on the latent dimensions of data. We can also observe that the algorithm can deal with networks that have around 500 to 1000 nodes within a reasonable amount of time.

1.7 Applications in Real Dataset

In this section, **CFSG** is applied to four network datasets : Zachary's karate club data, U.S. political book network data, U.S. political blog data, and Citation network of statisticians.

Remark 1.7.1 *We set the Truncated-SVD parameter as $r = 4$ for all experiments that are performed in this Section. HNCV method is used for model selection. Table 1.4 and its caption include detailed grid settings for implementation of HNCV method and running times of the algorithm on each dataset.*

Remark 1.7.2 *Note that there's a discrepancy in running time per a pair of tuning parameter for the network with 1000 nodes presented in Subsection.$1.6.5$ (around 50 mins) and that for the political blog dataset (around 2 hours). This is because of the difference in setting on stopping criteria of the algorithm, where we denote it as ε. We set $\varepsilon = 10^{-4}$ for the numerical experiment in Subsection $1.6.5$ and set $\varepsilon = 6 \times 10^{-5}$ for remaining numerical experiments including real data analysis in Section 6 and Section 7.*

1.7.1 Zachary's karate club network

In this subsection, we apply our model **CFSG** to the well-known Zachary's karate club dataset that was originally introduced in paper [53]. This dataset contains 34 nodes where each of them represents a member of university karate club. It is reported that at some point, those 34 members split into two communities, one led by "Mr. Hi" and the other led by "John A". We set the number of communities as $\mathbf{K} = 2$ and select a proper model through HNCV method proposed in Subsection 1.6.2. The model selected at a tuning parameter pair, $(\gamma^{\text{HNCV}}, \delta^{\text{HNCV}}) = (0.0126, 0.048)$, is visualized in the left panel of Figure 1.7. Nodes that are classified as "H" group are colored in orange and those classified as "A" group are colored in lightblue. Additionally, edges colored in green represent the ad-hoc edges estimated through **CFSG**. Since true labels of nodes in the network are recorded, qualities of selected model can be measured through evaluation metrics in Subsection 1.6.2. It is interesting to see that $M_1 = 1, M_2 = \frac{10}{10}, M_3 = 0, M_4 = 0$. For membership assignments of nodes in the network, see left panel of Figure 1.6 on how K-means algorithm works on $\widehat{E}_2$ for detecting communities.

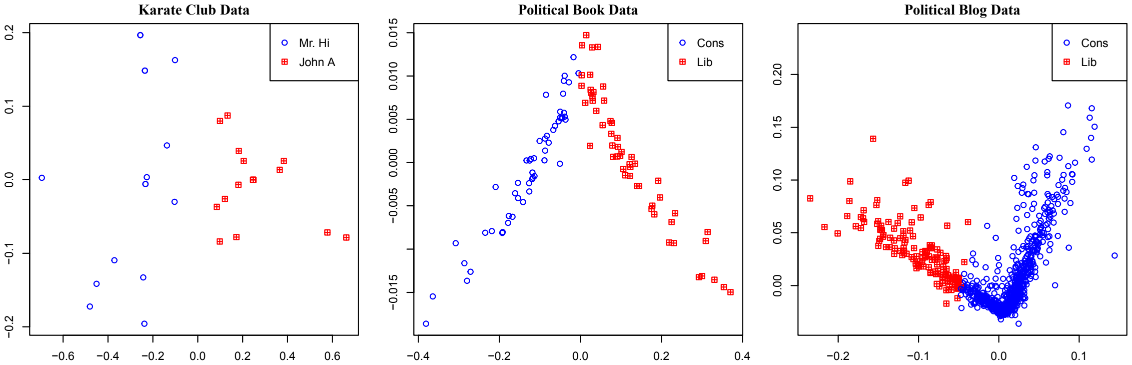


Figure 1.6: Plots of the first column versus the second column of $\widehat{E}_2$ on Karate (Left), Political Book (Middle), and Political Blog Network datasets (Right). The number of clusters is set as $K = 2$ and K-means clustering algorithm is applied on $\widehat{E}_2$ for the three datasets. Detected memberships of each node via K-means algorithm are colored in blue or red.

1.7.2 U.S. political book network

U.S. political book dataset consists of 105 political books sold by an online bookseller in the year of 2004. Nodes in the network represent books and they are connected through edges if they are frequently co-purchased by the same buyers. Original network was compiled by V. Krebs (http://www.orgnet.com) and labels of books were manually assigned by M. E. J. Newman. It is known that Newman labeled the books into three categories (liberal, conservative, and neutral) based on the reviews and descriptions of the books [54]. However, it is pointed out by several researchers that label assignments for some nodes in the network may not be accurate (seeing [54, 55, 56]). In our work, we simply view the network as having $\mathbf{K} = 2$ communities (liberal and conservative). A selected model from the HNCV method with a tuning parameter pair $(\gamma^{\text{HNCV}}, \delta^{\text{HNCV}}) = (0.004388, 0.035)$ is illustrated in the right panel of Figure 1.7. Nodes classified as “conservative” party are colored in orange and those classified as "liberal" party are colored in light blue. Estimated ad-hoc edges in $\widehat{S}$ are colored in green. It is interesting to observe that **CFSG** recovers all the 23 edges between estimated communities but for one edge (between 52^{th} and 70^{th} nodes) through the estimated $\widehat{S}$ matrix. To measure the quality of our selected model, we set the ground-truth labels of nodes in the network as suggested in the paper [55]. Ignoring the books they assigned as "Neurtral" (i.e. Node number : $1, 5, 7, 8, 49, 52, 70$) and "Not a political book" (i.e. Node number : 19), we found membership assignments of books estimated by **CFSG** are perfectly consistent with those estimated by [55]. Our model mis-classified one book, "The Bushes" (Node number: 50), as one belonging to liberal party. See middle panel of Figure 1.6 on how K-means algorithm works on $\widehat{E}_2$ for detecting communities.

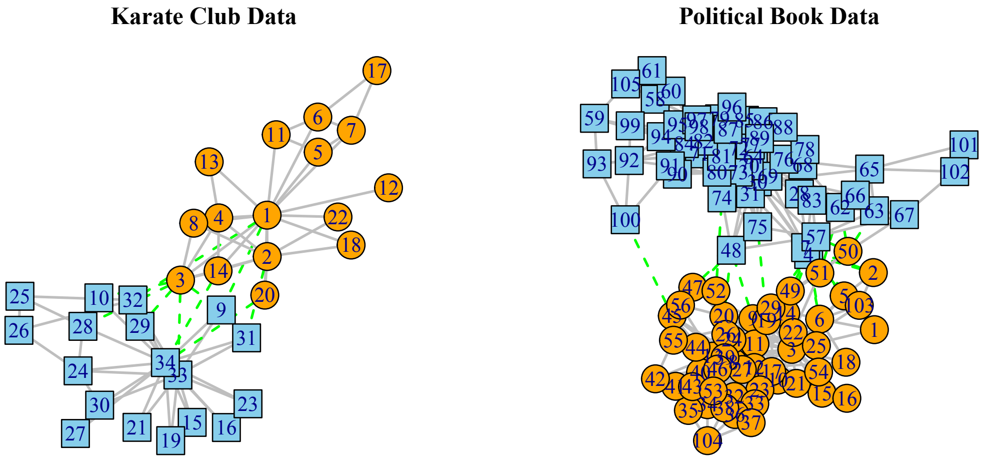


Figure 1.7: From left to right : Topologies of Karate club network data and U.S. political book network data. In Karate club data, node 1 represents "Mr. Hi" and node 34 represents "John A". For Karate club data, nodes that are classified as "Hi" group are colored in orange and those classified as "John A" group are colored in lightblue. For political book dataset, nodes classified as "conservative" party are colored in orange and those classified as "liberal" party are colored in light blue. In both dataset, color coding of ad-hoc edges corresponds to any $\widehat{S}_{ij} > 0$ in respective fitted model.

1.7.3 Social Network of U.S. Political Blog

We apply our model to the well-known political network dataset, collected by [57] shortly before the 2004 U.S. presidential election. Each node of the network is a web blog about U.S. politics and each edge indicates a hyperlink between them. (we neglect the direction of edges so that the graph is undirected.) It is believed that the blogs in the network share some common political leanings: liberal and conservative. The political leanings between two communities are significantly different, whereas are not significantly different among the nodes in the same topic. Out of 1494 blogs in the original network, we focus on the largest connected component, (i.e., a collection of nodes with non-zero degrees), which contains 1222 blogs and 16714 edges. Node numbers from 1 to 586 are labeled as "liberal" and remaining nodes are labeled as "conservative". We denote X_{Pol} as the adjacency matrix of the political blog network with 1222 nodes. We use the HNCV approach for model selection. A pair of tuning parameters, $(\gamma^{\text{HNCV}}, \delta^{\text{HNCV}}) = (0.00041, 0.012)$, gives us $\widehat{L}$ with $\text{rank}(\widehat{L}) = 2$ and $\widehat{S}$ with $|\widehat{S}| = 1755$. The resultant mis-classification rate of nodes for the blog dataset is $\frac{441}{1222}$ (i.e., $M_4 = \frac{441}{1222}$). See the right panel of Figure 1.6 on how K-means algorithm performs on $\widehat{E}_2$ of blog dataset. We make a short comment on this seemingly unsatisfactory performance of node classification on Section 1.8, which can be extended as another line of interesting future research direction.

Goodness of fit. Following the idea presented in [31], we investigate the goodness of fit of the model to the dataset via the parametric bootstrap. We denote $(\hat{\alpha}^{\text{HNCV}}, \widehat{L}^{\text{HNCV}}, \widehat{S}^{\text{HNCV}})$ as the minimizer of (1.12) evaluated at $(\gamma^{\text{HNCV}}, \delta^{\text{HNCV}})$, and generate 1000 independent adjacency matrix $\{X_b\}_{b=1}^{1000}$ from the minimizer, $(\hat{\alpha}^{\text{HNCV}}, \widehat{L}^{\text{HNCV}}, \widehat{S}^{\text{HNCV}})$. Here, we use X_b to denote the bth bootstrapped network's adjacency matrix. For each bootstrap sample, X_b, we evaluate the log-likelihood function defined in (1.11) under the parameters $(\hat{\alpha}^{\text{HNCV}}, \widehat{L}^{\text{HNCV}}, \widehat{S}^{\text{HNCV}})$, and denote it as $\ell_b^{\text{HNCV}} = \mathbb{L}_n(\hat{\alpha}^{\text{HNCV}}, \widehat{L}^{\text{HNCV}}, \widehat{S}^{\text{HNCV}}; X_b)$. The empirical distribution of $(\ell_1^{\text{HNCV}}, \ell_2^{\text{HNCV}}, \ldots, \ell_{1000}^{\text{HNCV}})$ is compared with the observed one:

$\ell_{\text{Pol}}^{\text{CFSG}} = \mathbb{L}_n(\hat{\alpha}^{\text{HNCV}}, \widehat{L}^{\text{HNCV}}, \widehat{S}^{\text{HNCV}}; X_{\text{Pol}})$. The histogram of $(\ell_1^{\text{HNCV}}, \ell_2^{\text{HNCV}}, \ldots, \ell_{1000}^{\text{HNCV}})$ is shown in the left panel of 1.8, and the observed log-likelihood $\ell_{\text{Pol}}^{\text{CFSG}} = -80012.57$ is marked by red dotted line with the p-value $= 49.85\%$ suggesting that the model fits the data reasonably well.

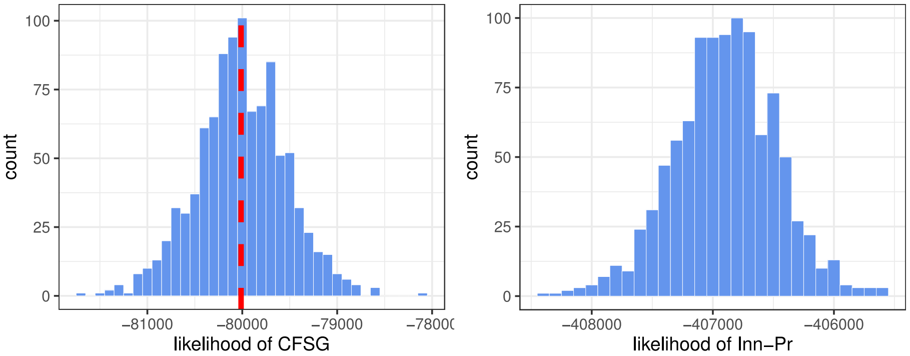


Figure 1.8: Parametric bootstrap to check the goodness of fit of **CFSG** model (left panel) and Inner Product model (right panel). Red dotted line is an observed likelihood of **CFSG** model, whereas the observed likelihood of Inner Product model is not appeared on the histogram.

For comparison purpose, we adopt the Inner-product model proposed by [22]. The paper also studies the largest connected component of political blog dataset, and fit the dataset to a following model:

$$X_{ij} = X_{ji} \sim \text{Bernoulli}(P_{ij}), \text{ with } \text{logit}(P_{ij}) = \alpha_i + \alpha_j + f_i^T f_j + \beta Z_{ij}. \tag{1.17}$$

Here, $\alpha_i, \forall i \leq 1222$, are parameters modeling degree heterogeneity, $f_i^T f_j$ is an inner-product between latent vectors, where f_i denotes ith row of matrix $F \in \mathbb{R}^{1222 \times 2}$. Lastly, as mentioned in Section 1.2.2, β is the coefficient for the observed information. We set $\beta = 0$ and obtain the local-minimizer $(\hat{\alpha}, \widehat{F})$ of log-likelihood function of $\{X_{ij}\}_{i,j=1}^{1222}$ through projected gradient descent algorithm. For more detailed inference procedure, readers can refer to [22]. After fitting the model, we check the goodness of fit of the Inner-product model via the same parametric procedure, based on 1000 bootstrap sample. The observed log-likelihood is $\ell_{\text{Pol}}^{\text{Inn-Pr}} = -214969.3$, and corresponding bootstrap distribution is shown in the right panel of 1.8. We obtain p-value $= 0$, and this suggests that the Inner-product model may not fit the data well.

Ad-hoc Edges in $\widehat{S}$. The selected model has 1755 ad-hoc edges. All non-zero entries of $\widehat{S}$ are positive. Among 1755 edges, True Positive (TP) of $\widehat{S}$ is 864. (i.e., $|\{(i,j) : 1 \leq i \leq 586,\ \ 587 \leq j \leq 1222, X_{\text{Pol},ij} = 1\ \ \&\ \ \widehat{S}_{ij} \neq 0\}| = 864$), where the number of true ad-hoc edges is 1575. (i.e., $|\{(i,j) : 1 \leq i \leq 586,\ \ 587 \leq j \leq 1222, X_{\text{Pol},ij} = 1\}| = 1575$.) Also, False Positive (FP) is 0, which means that the selected model does not emphasize the wrong connections between blogs. (i.e., $|\{(i,j) : 1 \leq i \leq 586,\ \ 587 \leq j \leq 1222, X_{\text{Pol},ij} = 0\ \ \&\ \ \widehat{S}_{ij} \neq 0\}| = 0$.) Among the blogs which form the 864 edges, we list top 10 which have the most links from the opposite party in Table 1.2. Aside from blogs which act principally as message boards or ranking sites (i.e., truthlaidbear.com, democraticunderground.com, nationalreview.com), it is noted that individual bloggers such as “andrewsullivan” and “wonkette” had the across board appeal at the period of 2004 U.S. presidential election. Our result agrees well with the observations in the original paper of the dataset [57]. It is not surprising that “Daily Kos”, a popular liberal blog, and “Instapundit”, a popular conservative blog, received a lot of attention from the opposite parties. An even closer look at the blogs in the list tells us what specific topics drew the attentions from both parties. One of the listed conservative blogs, “powerline.com”, broke the story on the CBS news’ credibility over the memos of unsubstantiated allegations about the former U.S. president, George W. Bush’s service on Texas Air National Guard in years 1972-3. Reports from powerline blog launched a flurry of discussions across the political blogs to both conservative and liberal leanings and beyond.

1.7.4 Citation network for statisticians

Recently, Ji and Jin [49] published an interesting dataset on citation network of papers from statistics journals. Specifically, this dataset is based upon all papers published from 2003 to

Table 1.2: A list of top 10 blogs that received the most links from blogs in the opposite party.

Blog URL	Political Party	Number of Links
truthlaidbear.com	“Cons”	40
dailykos.com	“Lib”	40
andrewsullivan.com	“Cons”	33
talkingpointsmemo.com	“Lib”	33
democraticunderground.com	“Lib”	31
instapundit.com	“Cons”	30
drudgereport.com	“Cons”	29
nationalreview.com/thecorner	“Cons”	20
powerlineblog.com	“Cons”	18
wonkette.com	“Lib”	18

the first half of 2012, from the four top statistical journals: *Annals of Statistics, Biometrika, Journal of American Statistical Association* (JASA), and *Journal of Royal Statistical Society (Series B)* (JRSS-B). Citational relationships of 3248 papers are given in the form of adjacency matrix. In our analysis, for computational convenience, we focus our attentions on the papers that have greater than or equal to 10 citational edges in the network of Ji and Jin [49]. After collecting papers with greater than or equal to 10 citational edges and eliminating those that have no connecting edges from the rest, we have 232 papers in total.

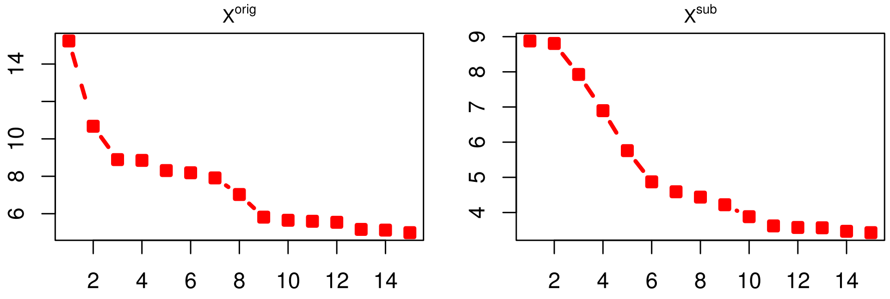


Figure 1.9: From left to right : Scree plots of the adjacency matrix X^{orig} and X^{sub}.

We denote the adjacency matrix of these 232 papers as X^{orig}. Elbow points of the scree plot from X^{orig} may be at the 3rd, 5th, or 9th largest eigenvalue, suggesting that there are from 2 to 8 embedded topics in the network (Figure 1.9). In light of this, we conduct the analysis in the following two steps:

1. First, following the approach suggested in [49], we assume that the network X^{orig} has 2 distinct topics and one giant mixed-component, which has a sub-network structure. Under this assumption, we set $\widehat{K}^{\text{scree}}$ as 3, and select a proper model via HNCV. Then, we perform K-means algorithm on matrix $\widehat{E}_3$ treating each row of the matrix as one data point.

2. Next, we restrict the network to the giant component ignoring all the edges to/from outside and obtain a subnetwork. We denote the adjacency matrix of this subnetwork as X^{sub}. We set $\widehat{K}^{\text{scree}}$ as 5, and also select a proper model through HNCV. Here, we run K-means algorithm on $\widehat{E}_5$ setting the number of clusters as 5.

In the first step, a pair of parameters, $(\gamma^{\text{HNCV}}, \delta^{\text{HNCV}}) = (0.00208, 0.0188)$, gives us $\widehat{L}$ with rank 3, and $\widehat{S}$ with $|\widehat{S}| = 23$. First topic studies on variable selection with high-dimensional data **(VarSel)**. Second topic discusses controlling false discovery rate in various statistical settings **(MulT)**. Third group, which consists of 162 papers, is hard to interpret and appears to have sub-network structures. For further investigation, we set this group as a giant component in the network, and denote corresponding component's adjacency matrix as X^{sub}. We perform a model selection as described in aforementione Step 2. A pair of tuning parameters, $(\gamma^{\text{HNCV}}, \delta^{\text{HNCV}}) = (0.00296, 0.0160)$, gives us the model with $\widehat{L}$ with rank 5, and $\widehat{S}$ with $|\widehat{S}| = 151$, and we can obtain five sub-communities as follows:

From the sub-network X^{sub}, we got four meaningful topics: Bayesian Statistics **(Bayes)**, Functional/Longitudinal Data Analysis **(FuncAn)**, Dimension Reduction **(DimRed)**, and High-dimensional Covariance Estimation **(CovEst)**. Due to the small volume of each topic, we could manually check that the false discovery for each topic is all zero. A full list of papers for each topic is provided in webpage. [1] Lastly, the sub-network X^{sub} has a big collection of papers that we refer it as "Mixed Topics" cluster **(Mixed)**. We provide a further inspection on this cluster with interesting observations in Appendix $A.4$.

Ad-hoc Edges in $\widehat{S}$. Non-zero components of $\widehat{S}$ capture the citational relationships among

[1] https://sites.google.com/site/namjoonsuh/publications

papers that are not attributable to the common topics. The selected model at Step 1 has 23 sparse edges. Among them, we provide 4 pairs of papers which have the largest estimated $\widehat{S}_{ij}^{\text{orig}}$ in Table A.1. All the 4 edges come from the pairs of papers with different topics. For instance, first pair of papers comes from Functional analysis topic and Variable selection topic. The paper from Functional analysis topic cites the paper from Variable selection for borrowing a mathematical representation to build a theorem. Though it is an essential step for building a theorem in their paper, we cannot say that two papers are closely related in terms of topic. The second pair of papers comes from Bayesian statistics topic and Variable selection topic, respectively. Specifically, authors in the paper from Bayesian statistics topic study variable selection problem under Non-parametric Bayesian framework, and compare their method with the "Adaptive Lasso". Interested readers can find a full list of papers which form the 23 ad-hoc edges in the webpage. [1]

1.8 Discussion

In this Section, we mention several directions to be explored based upon the model we propose in this work.

1. We can incorporate important characteristics such as degree heterogeneity or homophily of the network data in our model. Many researchers have been working on building interesting models to capture these characteristics [22, 23]. Following their approaches, we can easily incorporate these characteristics in our model as follows:

   $$X_{ij} = X_{ji} \sim \text{Bernoulli}(P_{ij}), \ \text{ with } \ \text{logit}(P_{ij}) = \alpha_i + \alpha_j + L_{ij} + S_{ij} + \beta Z_{ij}, \quad (1.18)$$

   where $\alpha_i \ \forall i \in \{1, 2, \ldots, n\}$, allows us to model every node in the network has different node degree and positive β denotes the regression coefficient, reflecting the fact that if the entities of the network have identical attributes in common, they are more likely to be connected. For instance, in citation network, if ith paper cites jth paper 5 times, or vice versa, we can set $Z_{ij} = 5$. It would be interesting to measure whether the goodness of fit of the data increase if we incorporate these terms.

2. To assign membership of each node, we adopt the K-means clustering algorithm on the weighted latent vectors of nodes in the network. The reason why we initially adopt K-means clustering algorithm is because of its simplicity. Furthermore, simple K-means clustering is widely used for clustering nodes in the community detection problem, which shows great successes in many applications, [19, 22, 58]. On the other hand, we found that the conventional K-means algorithm doesn't cluster nodes well for the blog dataset. (i.e., Subsection 1.7.3.) Similar to the discussion in Subsection $1.6.$ of paper [19], we conjecture that this is due to the presence of serious degree heterogeneity in the network, the weighted latent vectors are highly centered around the origin. Nonetheless, with ground-truth label, we found the weighted latent vectors of the blog dataset exhibit an interesting "EigenSpokes" pattern [59], wherein they have a clear, separate line that neatly aligns along the vertical line centered in the origin. We can observe this phenomenon for Karate club and political book dataset. (See Figure 1.6.) Paper [59] observed this "EigenSpokes" phenomenon on eigenvector versus eigenvector plot of the adjacency matrix. They employed the so-called "Chipping off" technique for detecting communities instead of using the K-means algorithm. Based upon these observations, we can explore the questions on "when" and "why" we observe the EigenSpokes phenonmenon on the weighted latent space in **CFSG**.

3. Taking into account the directions of edges in the graph is an important issue, and this may result in a completely different modeling approach, algorithm, and inference result from this in the present paper. For instance, based upon the idea of latent variable model, we can extend **CFSG** as follows :

$$X_{ij} \sim \text{Bernoulli}(P_{ij}), \text{ with } \text{logit}(P_{ij}) = \alpha_i + \alpha_j + u_i^T D v_j + S_{ij},$$

where α_i, α_j $\forall i, j \in \{1, 2, \ldots, n\}$, allow us to model every node in the network has different node degree, u_i and v_j denote two different latent factors, and sparse matrix S needs not to be symmetric. Given that we know the number of embedded communities K in the network, this formulation results in non-convex optimization problem with ℓ_1 penalization on the S matrix. We conjecture that this problem can be solved via the projected-subgradient descent with a good initialization. A similar formulation has been studied by [22], where they also solve a non-convex optimization problem based upon the latent variable model. But they work with undirected graph considering the same latent factors u_i. Moreover, instead of a sparse component S, they incorporate the term βZ that characterizes the homophily of a network. This direction can lead to a promising future research.

4. In Theorem 1.5.4, we assume strong convexity assumption on the likelihood function, given that there exists $\tau > 0$. (See Assumption 1.5.1 in Section 1.5.) However, in many high-dimensional statistical models, this assumption is violated. See an example in [7]. In our model, we also observe this assumption is violated especially when the size of network, n, grows large enough, and the entries of the ground truth of the parameter, Θ^*, are large enough. This results in very small τ, and leads to a loose bound in (1.16). (Note that the bound in (1.16) is involved with $\frac{1}{\tau^2}$.) In order to remedy this, a notion of restricted strong convexity (RSC) can be imposed, assuming that the likelihood function is strongly convex over a certain subset of parameter space. In order to establish the RSC condition on a generalized linear model, truncation argument needs to be employed, [6, 7]. This topic is left as a future research task.

Table 1.3: Top 4 edges corresponding with the pairs of papers from different communities. Authors and years of publication for the papers in each pair are also presented. In the first pair, a paper from functional analysis topic cites a paper from variable selection topic for borrowing a mathematical representation to build a theorem. But they are not related in terms of topic.

Pair	topic	Title
1	FuncAn	Properties of principal component methods for functional and longitudinal data analysis
	VarSel	Nonconcave penalized likelihood with a diverging number of parameters
2	VarSel	The adaptive lasso and its oracle properties
	Bayes	Nonparametric Bayes conditional distribution modeling with variable selection
3	DimRed	Contour projected dimension reduction
	VarSel	Factor profiled sure independence screening
4	VarSel	Factor profiled sure independence screening
	DimRed	Sliced regression for dimension reduction

Pair 1	Pair 2	Pair 3	Pair 4
P. Hall, et al. 2006	H. Zou. 2006	R. Luo, et al. 2009	H. Wang. 2012
J. Fan, et al. 2004	Y. Chung, et al. 2009	H. Wang. 2012	H. Wang, et al. 2012

Table 1.4: Grid settings for implementation of HNCV method and running times of the algorithm on each dataset. We repeat (a) and (b) of Step 2 ten times for Karate club, political book, and citation network datasets, and five times for political blog dataset. Maximum, minimum, and average computational times over the grid, G, are recorded. Interestingly, as the number of nodes grows, it turns out that running time is sensitive to the choice of tuning parameter. See Political Blog dataset.

	Karate Club	**Political Book**	**Political Blog**	**Citation Network**	
				$X^{\textbf{orig}}$	$X^{\textbf{Sub}}$
Node Number	34	105	1222	232	162
Grid Size	30	30	9	30	30
Grid Range	$\gamma \in [0.012, 0.0128]$ $\delta \in [0.04, 0.05]$	$[0.004386, 0.00439]$ $[0.035, 0.04]$	$[0.00040, 0.00042]$ $[0.011, 0.013]$	$[0.00205, 0.0021]$ $[0.0186, 0.019]$	$[0.00296, 0.003]$ $[0.016, 0.0165]$
Selected Parameter	$\gamma^{\text{HNCV}} = 0.0126$ $\delta^{\text{HNCV}} = 0.048$	0.004388 0.035	0.000411 0.012	0.00208 0.0188	0.00296 0.0160
Max (sec)	67.74	136.65	15322.15	222.80	131.37
Min (sec)	21.61	72.82	5533.924	203.03	123.29
Avg (sec)	45.21	106.84	10072.12	210.52	125.93

CHAPTER 2
ASYMPTOTIC THEORY OF ℓ_1-REGULARIZED PDE IDENTIFICATION FROM A SINGLE NOISY TRAJECTORY

2.1 Introduction

Differential equations are widely used to describe many interesting phenomena arising in scientific fields, including physics [60], social sciences [61], biomedical sciences [62], and economics [63], just to name a few. The forward problem of solving equations or simulating state variables for differential models has been extensively studied either theoretically or numerically in literature. We consider an inverse problem of learning a Partial Differential Equations (PDE) model.

More specifically, we assume that the governing PDE is a multi-variate polynomial of a subset of a prescribed dictionary containing different differential terms. Let $u(x,t)$: $\mathbb{R} \times [0,+\infty) \to \mathbb{R}$ be a real-valued function, where x be the spatial and t be the temporal variables. Suppose that within a bounded region of $\mathbb{R} \times [0,+\infty)$, $u(x,t)$ satisfies an evolutionary PDE:

$$\partial_t u = \mathcal{F}(u, \partial_x u, \partial_x^2 u, \ldots,), \quad \forall (x,t) \in \Omega \subseteq \mathbb{R} \times [0,+\infty). \tag{2.1}$$

Here, $\partial_t u$ (or u_t) denotes the partial derivative of u with respect to temporal variable, t; for $p = 0,1,2,\ldots, \partial_x^p u$ denotes the p-th order partial derivative of u with respect to spatial variable, x; $\mathcal{F}$ is an unknown polynomial mapping, and Ω is a bounded open subset of space-time domain. This format encloses various important classes of PDEs, e.g., advection-diffusion-decay equation characterizing pollutant distribution in fluid, Burgers' equation modeling the traffic flow [64], Kolmogorov-Petrovsky-Piskunov equation describing phase transitions [65], and Korteweg-de-Vries equation simulating the shallow water dynamics [66].

In our work, $\mathcal{F}$ is assumed to be a linear map, parameterized by a sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^K$: that is, u_t is represented as a linear combination of the degree 2 polynomials of arguments in $\mathcal{F}$, and only a few from a large set of potential functions are assumed to be relevant with u_t. Our goal is to estimate the correct non-zero indices of $\boldsymbol{\beta}^*$, given a single noisy trajectory of the function $u(x,t)$. Readers can refer to Subsection 2.3.1 for more detailed descriptions on the structural assumptions on $\mathcal{F}$, $\boldsymbol{\beta}^*$, and noisy trajectory. This problem setting naturally leads us to develop a two-stage method for the PDE identification based on Local-Polynomial smoothing and the ℓ_1-regularized Pseudo Least Squares (ℓ_1-PsLS) method. In the first stage, from a given noisy observation, we propose to estimate the underlying bi-variate function $u(x,t)$ and its partial derivatives with respect to its spatial and temporal dimensions via the Local-Polynomial fitting [67, 68]. In the second stage, with the constructed functions through Local-Polynomial regression, we propose to identify the correct differential terms and estimate model parameters via an ℓ_1-regularized Pseudo Least-Squares method.

We note that the two-stage method with Local-Polynomial regression has been applied in the Ordinary Differential Equations (ODE) setting. Specifically, the paper [69] established the consistency and asymptotic normality of the pseudo least squares estimator in the ODE setting, where they used Local-Polynomial regression to estimate the state variables from the noisy data. Similarly, [70, 71] studied the parameter estimation of ODE models with varying coefficients. However, these literature focused on estimating model parameters, rather than on selecting correct differential models. In the context of PDE, [72] studied PDE identification problems, using two-stage method. Authors of the paper modeled unknown PDEs using multivariate polynomials of sufficiently high order, and the best fit was chosen by minimizing the least squares error of the polynomial approximation. Nonetheless, ℓ_1 penalization for model selection was not used, and theoretical justification for their method remains underdeveloped.

From the theoretical point of view, our paper is the first work to propose the method, ℓ_1-PsLS, with a provable guarantee in the PDE recovery problem. Our main theoretical contribution is to establish sufficient conditions for ***signed-support recovery*** of the proposed ℓ_1-PsLS in PDE identification problems. It is worth noting that the signed-support recovery is a slightly stronger criterion than the support recovery, where its primary goal is not limited to finding the non-zero indices of $\boldsymbol{\beta}^*$, but also aims at recovering the correct signs of the selected coefficients. Ensuring the correct signed-support recovery of governing dynamical system has an important practical implication since many PDEs are sensitive to the signs of coefficients. For example, changing the sign of the advection term in the transport equation reverses the moving direction, and in —ing the sign of the Laplacian term of heat equation leads to instability of the system of interest.

Our theorem states that following three main conditions are sufficient for the signed-support recovery of ℓ_1-PsLS: (i) ***minimum eigenvalue condition*** among the arguments of the map $\mathcal{F}$ supported on non-zero indices of $\boldsymbol{\beta}^*$, and (ii) ***mutual incoherence condition*** among the arguments of the map $\mathcal{F}$, and (iii) $\boldsymbol{\beta}^*_{min}$***-condition*** on $\boldsymbol{\beta}^*$. The first condition indicates that relevant feature functions should be linearly independent. The second condition states that a large number of irrelevant predictors cannot exhibit an overly strong influence on the subset of relevant predictors. The third condition says that the minimum absolute value of non-zero entries of $\boldsymbol{\beta}^*$ should be greater than a certain threshold. These conditions appear in the statistical literature on the signed-support/support recovery of LASSO [3, 4, 73, 74] in linear regression problems, and our work rigorously shows that these are also essential for the signed-support recovery of PDE identification problems.

We employ Primal-Dual Witness (PDW) construction [4] as the main proof technique for the theorem. PDW construction is a popular mathematical technique for certifying variable-selection consistency of ℓ_1-penalized M-estimation problems including LASSO. See [75, 76, 77, 78, 79, 80]. For reader's convenience, we provide a brief introduction of the technique in the Appendix B.1. However, we want to emphasize that our Theorem is not a direct result of the trivial application of the PDW construction. Our problem settings are different from those of the work [4] in two aspects, which add some delicacies to our proof:

- As will be detailed in Subsection 2.3.3, the distribution of residual vector $\boldsymbol{\tau}$ is unknown, and neither mean 0 nor independent in our setting. On the contrary, in the work of [4], each entry of the residual vector is assumed to follow centered Gaussian with $\sigma^2 > 0$ variance and independent with the others.

- In the ℓ_1-PsLS method, the feature matrix obtained via Local-Polynomial fitting from noisy data is always random and has dependent rows uniquely determined through the underlying PDE. On the other hand, [4] divided their analysis into two cases, where the feature matrix $\mathbf{F}$ is either deterministic or random. When $\mathbf{F}$ is random, it is assumed to be a Gaussian ensemble with independent rows, whose covariance matrix satisfies mutual incoherence condition.

Organization. The remainder of the paper is organized as follows. Some related literature with our work are reviewed in Section 2.2. In Section 2.3, we formally define our problem by imposing some specific structural assumptions on $\mathcal{F}$ and propose a ℓ_1-PsLS method for PDE identification. In Section 2.4, the main theorem of our work is given on the signed-support recovery of ℓ_1-PsLS with the mutual incoherence assumption on the feature matrix $\mathbf{F}$, and we provide a high-level outline of the proof. Section 2.5 is devoted to provide a similar result with that of the one in the main theorem in Section 2.4 under milder assumption: that is, mutual incoherence assumption is imposed on the estimated feature matrix $\widehat{\mathbf{F}}$; an overview of proof is furnished. Related technical difficulties for the proof and main technical contribution of the paper are also given. Section 2.6 provides two lemmas for completing the proof of the main theorem by linking the mutual incoherence assumption with the ground-truth $\mathbf{F}$ to its sampled version. In Section 2.7, we show various numerical examples to validate and demonstrate different aspects of our method. We conclude this paper in Section 2.8 with some discussion.

Notation. For sufficiently large n, we write $f(n) = \mathcal{O}(g(n))$, if there exists a constant $K > 0$ such that $f(n) \leq Kg(n)$, and $f(n) = \Omega(g(n))$ if $f(n) \geq K'g(n)$ for some constant $K' > 0$. The notation $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$. We adopt bold lower-case letters for vectors and bold upper-case letters for matrices. For a vector $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\|_1 := \sum_{i=1}^{n}|\mathbf{v}_i|$, $\|\mathbf{v}\|_2 := \sqrt{\sum_{i=1}^{n}\mathbf{v}_i^2}$, and $\|\mathbf{v}\|_\infty := \max\limits_{1\leq i\leq n}|\mathbf{v}_i|$. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{A}^T$ denotes its transpose, $\|\mathbf{A}\|_2 := \max_{\forall\|\mathbf{x}\|_2=1}\|\mathbf{Ax}\|_2$, $\|\mathbf{A}\|_\infty := \max\limits_{1\leq i\leq n}\sum_{j=1}^{m}|\mathbf{A}_{i,j}|$, $\|\mathbf{A}\|_{\infty,\infty} := \max\limits_{1\leq i\leq n, 1\leq j\leq m}|\mathbf{A}_{i,j}|$, and $\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{A}_{i,j}^2}$.

2.2 Related Works

Our work is relevant to various topics in applied mathematics and statistics. Among them, we provide two most closely related topics: (i) Regression-based framework for PDE identification, and (ii) Some theoretical results of support-recovery of LASSO [3] in linear regression setting. In this Section, we denote K as the problem dimension, s as the number of non-zero entries of model parameter, and n as the number of observations.

Regression-based Methods. Recently, various regression-based frameworks have been developed and applied for model selection and parameter estimation of dynamic data. A sparsity-promoting method was proposed in [81] for extracting the governing dynamical system, by comparing the computed velocity to a large set of potential trial functions. Under the over-determined systems of linear equations (i.e., $n \gg k$), the authors developed a sequential-thresholded least-squares method to select the correct nonlinear functions. In the follow-up study, [82] devised a weighted-ℓ_1-regularized least squares solver for improving the accuracy and robustness of the approach introduced by [81] in the presence of state-measurement noise. Several papers [83, 84, 85] also suggested sparse regression frameworks for PDE identification problems over spatial-temporal data. Specifically, [84] studied the model selection problem via LASSO under the PDE context. The author empirically showed that the method works well in various important equations such as Burgers' equation, Navier-Stokes equation, Swift-Hohenberg equation. Recently, [83] considered PDE identification problem using numerical time evolution. The authors utilized LASSO to select candidate monomials, then proposed the time evolution error to select the underlying true model. Unlike the previously mentioned literature, which was mostly empirical, [86] provided a provable guarantee on the usage of ℓ_1-norm for PDE identification problems, based upon the theoretical results from compressive sensing. Interestingly, this work imposed the *incoherence property* on the feature matrix and employed the Legendre-transform on the columns of the matrix to ensure that the property holds for every PDE recovery problem of interest. Our work imposes mutual incoherence assumption on the feature matrix, which is an analogous notion of the incoherence property. However, the important difference between our paper and [86] is that our work only allows a single trajectory, whereas [86]'s theorem requires $\Omega(s \log K)$ bursts of noisy trajectories for the exact recovery of the underlying PDE.

Support Recovery in Statistics. Support recovery or variable selection problems of LASSO have a long history in the statistical literature. In the noiseless setting, many researchers [87, 1, 2, 88, 89, 90] established sufficient conditions for either the deterministic or random predictors for the support recovery problems of linear systems via the ℓ_1-norm.

Since our work falls into the category of noisy setting, we focus more on reviewing the body of work in the noisy setting. In [91], authors studied the asymptotic behavior of the LASSO-type estimator with fixed dimension K under the general centered i.i.d. noises with variance $\sigma^2 > 0$. Both [92] and [2] independently developed sufficient conditions for the support of LASSO estimator to be contained within true support of the sparse model. Under a more general setting, when the exterior noise is i.i.d. with finite moments, [93] showed that the Irrepresentable Condition [94] is almost necessary and sufficient for LASSO's signed-support recovery for fixed K and s. Furthermore, under the Gaussian noise assumption, they showed that LASSO can still achieve signed-support recovery when K is allowed to grow exponentially faster than n. In a non-asymptotic setting, [4] established the sharp relationship of n, K, and s, required for the exact sign consistency of LASSO, where K and s are allowed to grow as n increases under mutual incoherence condition. Using a similar technique in [4], the paper [73] studied LASSO under Poisson-like model with heteroscedastic noise and show that irrepresentable condition can serve as a necessary and sufficient condition for signed-support recovery in their setting. In the context of graphical model, [95, 14] analyzed the model selection consistency of Gaussian graphical models, and [75] showed the signed-support recovery of Ising models. See [96] for a more comprehensive overview on this topic.

Remark 2.2.1 *Our work is of asymptotic nature with fixed K and s, while the number of grid points of the observed trajectory tends to infinity in both space and time.*

2.3 PDE Identification via ℓ_1-PsLS

In Subsection 2.3.1, we provide concrete problem settings on the governing PDE of (2.1) and the observed trajectory. Then, specific settings of the Local-Polynomial regression for the estimations of state variables in our paper are provided in Subsection 2.3.2. Lastly, we propose a two-stage ℓ_1-regularized Pseudo Least Squares method for PDE identification in Subsection 2.3.3.

2.3.1 Problem Setting and Notations

Based on the general form (2.1), we take $(x, t) \in [0, X_{\max}) \times [0, T_{\max})$ for some finite constants $0 < X_{\max}, T_{\max} < \infty$. It is assumed that the underlying mapping $\mathcal{F}$ is a degree 2 polynomial in terms of u and its partial derivatives $\partial_x^p u$ for $0 \leq p \leq P_{\max}$, [1] parameterized

[1] It should be noted that our setting can be generalized to higher-degrees of polynomials and functions with multiple spatial dimensions.

by a coefficient vector $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{p,q}^*, \ldots)$ with real entries; that is,

$$u_t(x,t) = \beta_0^* + \beta_1^* u + \beta_2^* \partial_x u + \beta_3^* \partial_x^2 u + \cdots + \beta_{p,q}^* \partial_x^p u \partial_x^q u + \ldots . \tag{2.2}$$

We call the monomials in the right-hand side of (2.2) as *feature variables*. We set a finite integer upper-bound, $P_{\max} > 0$, for the possible orders of the partial derivatives of u with respect to x in (2.2). Hence, we assume that $\boldsymbol{\beta}^* \in \mathbb{R}^K$, with $K = 1+2(P_{\max}+1)+\binom{P_{\max}+1}{2}$; consequently, constant and any term of the form $\partial_x^p u$ or $\partial_x^p u \partial_x^q u$, for $0 \le p, q \le P_{\max}$, are contained in (2.2). Notice that many entries of $\boldsymbol{\beta}^*$ can be zero. We denote $\mathcal{S}(\boldsymbol{\beta}^*) := \{0 \le j \le K \mid \beta_j^* \neq 0\}$, or simply $\mathcal{S}$, as the support of the coefficient vector $\boldsymbol{\beta}^*$, i.e., the set of indices of the non-zero entries. Additionally, we denote s as the cardinality of the set $\mathcal{S}$, i.e., $s := |\mathcal{S}|$.

The given data $\mathcal{D} = \{(X_i, t_n, U_i^n) \mid i = 0, \ldots, M-1; n = 0, \ldots, N-1\} \subseteq \Omega \times \mathbb{R}$ consists of $M \times N$ data, where $M, N \in \mathbb{R}$, $M, N \ge 1$. Each $(X_i, t_n) \in \Omega$ represents a space-time point, and U_i^n is a representation of $u(X_i, t_n)$ contaminated by additive Gaussian noise:

$$U_i^n = u(X_i, t_n) + \nu_i^n, \quad \nu_i^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) ,$$

whose second moment is uniformly bounded as follows: $\sup_{N,M \in \mathbb{R}} \max_{n,i} E\,|U_i^n|^2 := \eta^2 < \infty$. Here $\mathcal{N}(0, \sigma^2)$ denotes the centered normal distribution with variance $\sigma^2 > 0$.

2.3.2 Local-Polynomial Regression Estimators for Derivatives

Given data $\{(X_i, t_n, U_i^n)\}$ with $i = 0, 1, \ldots, M-1$ and $n = 0, 1, \ldots, N-1$, we employ a local quadratic regression to estimate $u_t(X_i, \cdot)$ for each fixed space point X_i and use a Local-Polynomial with degree $p+1$ to estimate $\partial_x^p u(\cdot, t_n)$ at each temporal point t_n, for each degree $p = 0, 1, \ldots, P_{\max}$. More specifically, we solve the following optimization problems:

$$\left\{\widehat{b}_j(X_i,t)\right\}_{j=0,1,2} = \underset{b_j(t)\in\mathbb{R},0\leq j\leq 2}{\text{argmin}} \sum_{n=0}^{N-1}\left(U_i^n - \sum_{j=0}^{2} b_j(t)(t_n-t)^j\right)^2 \mathcal{K}_{h_N}\left(t_n - t\right),$$

$$\text{for } i = 0,1,\ldots,M-1\ ; \tag{2.3}$$

$$\left\{\widehat{c}_j^p(x,t_n)\right\}_{j=0,1,\ldots,p+1} = \underset{c_j(t)\in\mathbb{R},0\leq j\leq p+1}{\text{argmin}} \sum_{i=0}^{M-1}\left(U_i^n - \sum_{j=0}^{p+1} c_j^p(t)(X_i-x)^j\right)^2 \mathcal{K}_{w_M}\left(X_i - x\right),$$

$$\text{for } n = 0,1,\ldots,N-1 \text{ and } p = 0,1,\ldots,P_{\max}. \tag{2.4}$$

and set $\widehat{u}_t(X_i,t) = \widehat{b}_1(X_i,t)$ and $\widehat{\partial_x^p u}(x,t_n) = p!\widehat{c}_p^p(x,t_n)$. Here h_N and $w_{p,M}$ denote the bandwidth parameters, and $\mathcal{K}_w(z) := \mathcal{K}(z/w)/w$ for some kernel function $\mathcal{K}$ with bandwidth $w > 0$. Specific choices of the order of polynomial fit for the functions $\widehat{u}_t$ and $\widehat{\partial_x^p u}$ are to strike the balance between modeling bias and variance. See Subsections 3.1 and 3.3 of Fan and Gijbels [67] for more rigorous treatments on this topic. Also the kernel $\mathcal{K}$ is assumed to be uniformly continuous and absolutely integrable with respect to Lebesgue measure on the real-line; $\mathcal{K}(z) \to 0$ as $|z| \to +\infty$; and $\int |z \ln |z||^{1/2} |d\mathcal{K}(z)| < +\infty$.

Optimization problems (2.3) and (2.4) have closed-form solutions in the form of weighted least squares estimator. See Appendix B.2. However, for theoretical investigation, we employ the notion of *equivalent kernel* [67, 68] to write the solutions as follows: for any fixed spatial point X_i, $i = 0,1,\ldots,M-1$, $\widehat{u}_t(X_i,t)$ can be written as:

$$\widehat{u}_t(X_i,t) = \frac{1}{Nh_N^2}\sum_{n=0}^{N-1}\mathcal{K}_2^*\left(\frac{t_n - t}{h_N}\right)U_i^n\big\{1 + o_{\mathbb{P}}(1)\big\}. \tag{2.5}$$

Similarly, for any fixed temporal point t_n, $n = 0,1,\ldots,N-1$, the estimation for the p-th order partial derivative takes the form:

$$\widehat{\partial_x^p u}(x,t_n) = \frac{p!}{Mw_M^{p+1}}\sum_{i=1}^{M}\mathcal{K}_p^*\left(\frac{X_i - x}{w_M}\right)U_i^n\big\{1 + o_{\mathbb{P}}(1)\big\}. \tag{2.6}$$

Here, $\mathcal{K}_j^*(z) = e_j^\top S^{-1}(1, z, \dots, z^p)^\top K(z)$ is called an equivalent kernel, where e_j denotes a unit vector with 1 on the j^{th} position; $S = (\int z^{l+s}\mathcal{K}(z)dz)_{0\le l,s\le p}$ is the moment matrix associated with kernel $\mathcal{K}$; and $o_{\mathbb{P}}(1)$ denotes a random quantity tending to zero as either N or M tends to infinity. From here, we will omit the dependency on j for the simplicity of notation when using the equivalent kernel.

Remark 2.3.1 *The most important reason for using Local-Polynomial fitting for the estimation of state variables and their derivatives is due to its rich literature on asymptotic properties and uniform convergence of the estimator [67, 97, 98, 68]. Specifically, these results allow us to explore the behavior of tail-probability of the measurement error* $\boldsymbol{\tau}$*, which is essential for the analysis of the* ℓ_1*-PsLS estimator. See Subsection 2.5.2 for more information.*

2.3.3 ℓ_1-regularized Pseudo Least Squares Model

First, we introduce matrix-vector notations for compact expressions of the problem. We let $\mathbf{u}_t \in \mathbb{R}^{NM}$ denote the vectorization of $\{u_t(X_i, t_n)\}_{i=0,\dots,M-1}^{n=0,\dots,N-1}$ in a dictionary order prioritizing the spatial dimension; that is, $\mathbf{u}_t^T = \begin{bmatrix} u_t(X_0,t_0) & u_t(X_1,t_0) & \cdots \end{bmatrix}$. Define the *feature matrix*, $\mathbf{F} \in \mathbb{R}^{NM\times K}$, as the collection of values of feature variables organized as follows:

$$\mathbf{F} := \begin{bmatrix} 1 & u(X_0,t_0) & \partial_x u(X_0,t_0) & \cdots & \partial_x^p u(X_0,t_0)\partial_x^q u(X_0,t_0) & \cdots \\ 1 & u(X_1,t_0) & \partial_x u(X_1,t_0) & \cdots & \partial_x^p u(X_1,t_0)\partial_x^q u(X_1,t_0) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots \\ 1 & u(X_{M-1},t_0) & \partial_x u(X_{M-1},t_0) & \cdots & \partial_x^p u(X_{M-1},t_0)\partial_x^q u(X_{M-1},t_0) & \cdots \\ 1 & u(X_0,t_1) & \partial_x u(X_0,t_1) & \cdots & \partial_x^p u(X_0,t_1)\partial_x^q u(X_0,t_1) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots \\ 1 & u(X_{M-1},t_{N-1}) & \partial_x u(X_{M-1},t_{N-1}) & \cdots & \partial_x^p u(X_{M-1},t_{N-1})\partial_x^q u(X_{M-1},t_{N-1}) & \cdots \end{bmatrix}.$$

With these notations, ground-truth PDE models (2.2) evaluated at the data points on the grid can be succinctly written as $\mathbf{u}_t = \mathbf{F}\boldsymbol{\beta}^*$. Note that before estimating the correct signed-support of $\boldsymbol{\beta}^*$, $\mathbf{u}_t$ and $\mathbf{F}$ need to be estimated. Conventional regression techniques such as Local-Polynomial regression, smoothing spline, among others, can be used to estimate $\mathbf{u}_t$ and columns of $\mathbf{F}$. As previously mentioned, we employ the Local-Polynomial approach. We denote $\widehat{\mathbf{u}}_t \in \mathbb{R}^{NM}$ and $\widehat{\mathbf{F}} \in \mathbb{R}^{NM\times K}$ by replacing the entries of $\mathbf{u}_t$ and $\mathbf{F}$ respectively with those of corresponding estimators. (i.e., $\widehat{(u_t)_i^n}$, $\widehat{(\partial_x^p u)_i^n}$, and $\widehat{(\partial_x^p u)_i^n (\partial_x^q u)_i^n}$.)

Let $\Delta\mathbf{u}_t = \widehat{\mathbf{u}}_t - \mathbf{u}_t$, $\Delta\mathbf{F} = \widehat{\mathbf{F}} - \mathbf{F}$ denote the difference between the obtained estimators $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ via Local-Polynomial regression and their ground-truth counterparts. With these notations, we formally obtain a regression model

$$\widehat{\mathbf{u}}_t = \widehat{\mathbf{F}}\boldsymbol{\beta}^* + \boldsymbol{\tau}\ , \quad \text{where } \boldsymbol{\tau} = \Delta\mathbf{F}\boldsymbol{\beta}^* - \Delta\mathbf{u}_t\ . \tag{2.7}$$

The natural extension for inducing sparsity of the parameter of interest is to add positively weighted ℓ_1-penalty term $\|\boldsymbol{\beta}\|_1$ to the squared loss $\|\widehat{u}_t - \widehat{F}\boldsymbol{\beta}\|_2^2$, leading to an estimator:

$$\widehat{\boldsymbol{\beta}}^{\lambda} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^K} \left\{ \frac{1}{2NM} \left\| \widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\boldsymbol{\beta} \right\|_2^2 + \lambda_N \left\| \boldsymbol{\beta} \right\|_1 \right\}, \tag{2.8}$$

where $\lambda_N > 0$ is a regularization hyper-parameter. Note that we normalize the columns of $\widehat{\mathbf{F}}$ such that $\frac{1}{\sqrt{NM}} \max_{j=1,\dots,K} \|\widehat{\mathbf{F}}_j\|_2 \leq 1$ while solving (2.8).

Observe that (2.8) is formally identical to LASSO [3] for high-dimensional sparsity recovery. Meanwhile, we should also emphasize that $\widehat{\boldsymbol{\beta}}^{\lambda}$ is not a true ℓ_1-least squares estimator, but a minimizer of the ℓ_1-least squares fit with the estimated $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$, instead of the ground-truth $\mathbf{u}_t$ and $\mathbf{F}$. Hence, we use the word ***"pseudo"*** as in [69] to emphasize the approximations of the solutions and derivatives of measurements, and call our method ℓ_1**-Pseudo Least Squares** method.

Additionally, the residual vector $\boldsymbol{\tau}$ violates conventional assumptions on residuals in linear regression, where the entries of residuals are commonly assumed to be i.i.d. centered random variables with finite variance. See [93, 4, 91]. Note that, since Local-Polynomial estimator is biased, $\boldsymbol{\tau}$ is not a mean zero random vector. Furthermore, there is no guarantee that entries of $\boldsymbol{\tau}$ are independent with each other. This arises from the fact that rows of $\mathbf{F}$ and entries of $\mathbf{u}_t$ are uniquely characterized by the underlying PDE model. Lastly, the unknown signal $\boldsymbol{\beta}^*$ makes the distribution of $\boldsymbol{\tau}$ completely inaccessible. These complexities make the study of the proposed estimator $\widehat{\boldsymbol{\beta}}^\lambda$ challenging.

2.4 Recovery Theory for ℓ_1-PsLS based PDE Identification

In subsection 2.4.1, we formally describe a signed-support recovery problem. In subsection 2.4.2, two regularity assumptions on feature matrix $\mathbf{F}$ are given for the proof of the main theorem. Then, the main theorem of this work is presented with some important remarks in subsection 2.4.3. Lastly, we provide a proof sketch of the main theorem in subsection 2.4.4.

2.4.1 Signed-Support Recovery

The main goal of this paper is to provide provable guarantees that the proposed ℓ_1-PsLS method gives asymptotically consistent estimator of $\boldsymbol{\beta}^*$ in the sense of signed-support recovery. We can formally state this problem with the adoption of $\mathbb{S}_\pm(\boldsymbol{\beta})$ notation, that is: for any vector $\boldsymbol{\beta} \in \mathbb{R}^K$, we define its extended sign vector, whose each entry is written as:

$$\mathbb{S}_\pm(\beta_i) := \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{if } \beta_i = 0, \end{cases}$$

for $i \in \{1, \ldots, K\}$. This notation encodes the *signed-support* of the vector $\boldsymbol{\beta}$. Denote $\widehat{\boldsymbol{\beta}}^\lambda$ as the unique solution of ℓ_1-PsLS. Under some regularity conditions on $\mathbf{F}$, we will show,

$$\mathbb{P}\big[\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}^\lambda) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)\big] \to 1 \quad \text{as } N, M \to +\infty,$$

where N and M denote the grid size of temporal and spatial dimensions, respectively.

2.4.2 Assumptions

We introduce two sufficient conditions frequently assumed in ℓ_1- regularized regression models for the signed-support recovery of the true signal $\boldsymbol{\beta}^*$.

1. **Minimal eigenvalue condition.** There exists some constant $C_{\min} > 0$ such that:

$$\Lambda_{\min}\left(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right) \geq C_{min}. \tag{A1}$$

   Here $\Lambda_{\min}(\mathbf{A})$ denotes the minimal eigenvalue of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $\mathbf{F}_{\mathcal{S}}$ is made of columns of $\mathbf{F}$ when the column index is in the support set $\mathcal{S}$. Note that if this condition is violated, the columns of $\mathbf{F}_{\mathcal{S}}$ would be linearly dependent, and it would be impossible to estimate the true signal β^* even in the "oracle case" when the support set $\mathcal{S}$ is known a priori.

2. **Mutual incoherence condition.** For some *incoherence parameter* $\mu \in (0, 1]$:

$$\left\|\left(\mathbf{F}_{\mathcal{S}^c}^{\top}\mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_{\infty} \leq 1 - \mu. \tag{A2}$$

   This condition states that the irrelevant predictors cannot exhibit an overly strong influence on the relevant predictors. More specifically, for each index $j \in \mathcal{S}^c$, the vector $(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}})^{-1}\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_j$ is the regression coefficient of $\mathbf{F}_j$ on $\mathbf{F}_{\mathcal{S}}$, thus, it is a measure of how well the column $\mathbf{F}_j$ aligns with the columns of $\mathbf{F}_{\mathcal{S}}$. A large μ close to 1 indicates that the columns $\{\mathbf{F}_j, j \in \mathcal{S}^c\}$ are nearly orthogonal to the columns of $\mathbf{F}_{\mathcal{S}}$, which is desirable for support recovery.

For future reference, we define $\mathcal{Q}^* := \left(\mathbf{F}_{\mathcal{S}^c}^{\top}\mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right)^{-1}$, and name it as *population incoherence matrix*. Also, define its estimated counterpart as $\widehat{\mathcal{Q}}_N := \left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)\left(\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}$, and call it *sample incoherence matrix*. Note that the dependence of the support set $\mathcal{S}$ on quantities $\mathcal{Q}^*$ and $\widehat{\mathcal{Q}}_N$ is suppressed for notational simplicity.

2.4.3 Statement of Main Result

Theorem 2.4.1 *Given the observed data set $\mathcal{D}$ whose spatial resolution is related to the temporal resolution via $M = \Theta(N^{\frac{2P_{\max}+5}{7}})$, we take the bandwidths of the kernels in* (2.3) *and* (2.4) *as $h_N = \Theta(N^{-\frac{1}{7}})$, $w_M = \Theta(M^{-\frac{1}{7}})$, respectively. Under the assumptions* (A1) *and* (A2) *imposed on the ground-truth feature matrix* $\mathbf{F}$, *suppose that the sequence of regularization hyper-parameters $\{\lambda_N\}$ satisfies $\lambda_N = \Omega\left(\frac{\sqrt{K}\ln N}{\mu N^{2/7-c}}\right)$ for some constant $0 < c < \frac{2}{7}$ independent of N. Then, the following properties hold with probability greater than $1 - \mathcal{O}\left(N^{\frac{2P_{\max}+5}{7}} \exp\left(-\frac{1}{6}N^c\right)\right) \to 1$ as $N \to \infty$:*

1. *The ℓ_1-PsLS method* (2.8) *has a unique minimizer $\widehat{\boldsymbol{\beta}}^\lambda \in \mathbb{R}^K$ with its support contained within the true support, that is $\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$, and the estimator satisfies the ℓ_∞ bound:*

$$\left\|\widehat{\boldsymbol{\beta}}^\lambda_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\right\|_\infty \le K^{3/2} C_{\min}\left(o_N(1) + \lambda_N\right) \ . \tag{2.9}$$

2. *Additionally, if the minimum value of the model parameters supported on $\mathcal{S}$ is greater than the upper-bound of* (2.9), *that is $\min_{1\le i\le s} |(\boldsymbol{\beta}^*_{\mathcal{S}})_i| > K^{3/2}C_{\min}\left(o_N(1) + \lambda_N\right)$, then $\widehat{\boldsymbol{\beta}}^\lambda$ has a correct signed-support. i.e., $\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}^\lambda) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)$.*

The overall proof sketch of Theorem 2.4.1 is described in the Subsection 2.4.4, and relevant technical propositions and Lemmas are further provided in Sections 2.4 and 2.5. Here, we give some important remarks about Theorem 2.4.1.

1. The uniqueness claim of $\widehat{\beta}^\lambda$ in (i) seems trivial since the objective function in (2.8) is strictly convex in the regime of K being fixed and $NM \to \infty$. However, we need to ensure that the minimal eigenvalue condition hold over the estimated feature

matrix $\widehat{\mathbf{F}}$, given the assumption (A1) for some $C_{\min} > 0$. We defer this statement as Lemma 2.6.2 in Section 2.6 with the detailed proof.

2. The item (i) claims that ℓ_1-PsLS does not select the arguments that are not in the support of $\boldsymbol{\beta}^*$. The item (ii) is a consequence of the sup-norm bound from (2.9): as long as $|\boldsymbol{\beta}_i^*|$ over indices $i \in \mathcal{S}$ is not small, ℓ_1-PsLS is signed-support recovery consistent.

3. The asymptotic orders of M, h_N, and w_M are specifically chosen for simplicity. Although there is certain flexibility, the spatial resolution M and the temporal resolution N (as well as h_N and w_M) need to be coordinated well to guarantee the support recovery property. This was expected in practice since we need sufficient sampling frequencies both in temporal and space to estimate the underlying dynamics. Here, the Theorem 2.4.1 present a rigorous justification for a combination of these resolutions which is sufficient for the support recovery.

4. The quantity c is derived from the Tusnády's strong approximation [98] where the error of an empirical distribution is compared with a Brownian bridge in tail probability. See Appendix B.3.1. With a larger value of c, the regularization hyper-parameter λ_N needs to remain relatively large, but the convergence is faster. Whereas for a smaller value of c, we can relax the regularization in the cost of a slower probability convergence rate.

5. The threshold of λ_N in the statement of the Theorem shows that when the number of data increases, there is more flexibility in tuning this parameter. If the incoherence parameter μ is small, or equivalently, the group of correct feature variables and the group of the others are similar, to guarantee that the support of the estimated coefficient vector is contained in the correct one, it suffices to use a large value of λ_N. Such behavior of the threshold is consistent with that described in Theorem 1 of [4].

6. The upper-bound for the ℓ_∞-norm of the coefficient error in (2.9) consists of two components. The first term $o_N(1)$ denotes a deterministic sequence converging to 0 as N increases to ∞. We want to note that this term is involved with the underlying function u as well as the choice of regression kernels and independent with the choice of feature variables selected by ℓ_1-PsLS. The second component is simple: $K^{3/2}C_{\min}\lambda_N$. When N increases, this part does not vary. This indicates that asymptotically, ℓ_1-PsLS recovers signed-support of governing PDE, as long as $\min_{1\le i\le s}|(\boldsymbol{\beta}^*_{\mathcal{S}})_i| > K^{3/2}C_{\min}\lambda_N$.

2.4.4 Proof Strategy of Theorem 2.3.1

The analysis for the proof of Theorem 2.4.1 is naturally divided into two steps as follows: In the first step, we prove a result analogous to that of the Theorem 2.4.1 by imposing incoherence assumption on the estimated feature matrix $\widehat{\mathbf{F}}$. Specifically, since $\widehat{\mathbf{F}}$ is a random matrix, we assume that for some $\mu \in (0,1]$, the event, $\{\left\|\widehat{\mathcal{Q}}_N\right\|_\infty \le 1-\mu\}$, holds with some probability at least P_μ, for some $P_\mu \in (0,1]$. Under this assumption, we prove that the success probability of signed-support recovery of ℓ_1-PsLS converges to P_μ with an exponential decay rate. This is formally stated as Proposition 2.5.1 in Subsection 2.5.1.

In the second step, we show that the success probability P_μ goes to 1, given that the ground-truth matrix $\mathbf{F}$ satisfies assumptions (A1) and (A2). This is equivalent to proving that, given the assumptions (A1) and (A2) for $\mathbf{F}$ for some $C_{\min} > 0$ and $\mu \in (0,1]$, the same assumptions hold for the estimated $\widehat{\mathbf{F}}$ in probability. We state these results formally in Lemmas 2.6.2 and 2.6.3 in Section 2.6.

2.5 Analysis Under Sample Incoherence Matrix Assumptions

In this section, we provide a proof overview of Proposition 2.5.1 and the key technical contribution of our paper. All the detailed statements and proofs of the Proposition 2.5.1 and its relevant Lemmas are relegated to the Appendix for the conciseness.

2.5.1 Statement of Proposition

We establish the signed-support consistency of ℓ_1-PsLS estimator when the assumptions are directly imposed on the estimated feature matrix $\widehat{\mathbf{F}}$, instead on the ground-truth feature matrix $\mathbf{F}$. More specifically, we assume that there exist some constants $\mu \in (0, 1]$ and $C_{\min} > 0$, such that the followings hold:

$$\mathbb{P}\left[\left\|\widehat{\mathcal{Q}}_N\right\|_{\infty} \leq 1 - \mu\right] \geq P_\mu \text{ and } \Lambda_{\min}\Big(\frac{1}{NM}\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\Big) \geq C_{\min} \quad \text{almost surely} . \tag{A3}$$

Here, $P_\mu \in [0, 1]$ denotes some probability that $\widehat{\mathcal{Q}}_N$ satisfies the incoherence assumption. Equipped with this assumption, we have the following proposition:

Proposition 2.5.1 *Given the observed data set $\mathcal{D}$, where the spatial resolution is related to the temporal resolution via $M = \Theta(N^{\frac{2P_{\max}+5}{7}})$, we take the bandwidths of the kernels in* (2.3) *and* (2.4) *as $h_N = \Theta(N^{-\frac{1}{7}})$, $w_M = \Theta(M^{-\frac{1}{7}})$, respectively. Under the assumptions in* (A3) *imposed on the estimated feature matrix $\widehat{\mathbf{F}}$, suppose that the sequence of regularization hyper-parameters $\{\lambda_N\}$ satisfies $\lambda_N = \Omega\Bigg(\frac{\sqrt{K}\ln N}{\mu N^{2/7-c}}\Bigg)$ for some constant $0 < c < \frac{2}{7}$ independent of N. Then, the following properties hold :*

1. *With probability greater than $P_\mu - \mathcal{O}\left(N^{\frac{2P_{\max}+5}{7}} \exp\big(-\frac{1}{6}N^c\big)\right) \to P_\mu$ as $N \to \infty$, the ℓ_1-PsLS method* (2.8) *has a unique minimizer $\widehat{\boldsymbol{\beta}}^\lambda \in \mathbb{R}^K$ with its support contained within the true support, that is $\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$.*

2. *With probability greater than $1 - \mathcal{O}\left(N^{\frac{2P_{\max}+5}{7}} \exp\big(-\frac{1}{6}N^c\big)\right) \to 1$ as $N \to \infty$, $\widehat{\boldsymbol{\beta}}^\lambda$ satisfies the ℓ_∞ bound:*

$$\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^\lambda - \boldsymbol{\beta}_{\mathcal{S}}^*\right\|_\infty \leq K^{3/2}C_{\min}\left(o_N(1) + \lambda_N\right) . \tag{2.10}$$

3. *Additionally, if the minimum value of model parameter supported on $\mathcal{S}$ is greater than the upper-bound of* (2.10)*, that is $\min_{1\leq i\leq s} |(\boldsymbol{\beta}_{\mathcal{S}}^*)_i| > K^{3/2}C_{\min}\left(o_N(1) + \lambda_N\right)$,*

then $\widehat{\boldsymbol{\beta}}^{\lambda}$ has a correct signed-support. (i.e., $\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}^{\lambda}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^{*})$)

We remark that the first item (i) in Proposition 2.5.1 holds with probability $P_{\mu} \leq 1$ asymptotically, while the second item (ii) holds with probability 1 asymptotically. They are not contradictory, since (i) describes the support recovery of the coefficient vector over all indices, whereas (ii) focuses on the estimation errors on entries within the true support $\mathcal{S}$. Technically speaking, proof of (i) is involved with mutual incoherence condition in (A3), whereas (ii) is involved with minimum-eigen value condition on $\widehat{\mathbf{F}}$ in (A3).

2.5.2 Proof Overview of Proposition 2.4.1

Readers can find the proof of (2.10) in the Appendix B.3.6. Here, we focus on providing the high-level idea on the proof of (i) of Propostion 2.5.1. The most important ingredient for the success of PDW construction is to establish the *strict dual feasibility* of the dual vector $\widehat{\mathbf{z}}$, when $\widehat{\mathbf{z}} \in \partial\|\widehat{\boldsymbol{\beta}}^{\lambda}\|_1$, where $\partial\|\widehat{\boldsymbol{\beta}}^{\lambda}\|_1$ is a sub-differential set of $\|\cdot\|_1$ evaluated at $\widehat{\boldsymbol{\beta}}^{\lambda}$. In other words, we need to ensure that $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_{\infty} < 1$ with high probability. (See Appendix B.1.) Through Karush–Kuhn–Tucker (KKT) condition of the optimal pair $(\widehat{\boldsymbol{\beta}}^{\lambda}, \widehat{\mathbf{z}})$ of (2.8) and settings of PDW construction, we can explicitly derive the expression of the dual vector $\widehat{\mathbf{z}}$ supported on the complement of the support set $\mathcal{S}$ as follows:

$$\widehat{\mathbf{z}}_{\mathcal{S}^c} = \widehat{\mathbf{F}}_{\mathcal{S}^c}^T \widehat{\mathbf{F}}_{\mathcal{S}} (\widehat{\mathbf{F}}_{\mathcal{S}}^T \widehat{\mathbf{F}}_{\mathcal{S}})^{-1} \widehat{\mathbf{z}}_{\mathcal{S}} + \underbrace{\frac{1}{\lambda_N MN} \widehat{\mathbf{F}}_{\mathcal{S}^c}^T \boldsymbol{\Pi}_{\mathcal{S}^{\perp}} (\Delta\mathbf{u}_t - \Delta\mathbf{F}_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}}^{*})}_{:=\tilde{Z}_{\mathcal{S}^c}} , \tag{2.11}$$

where $\boldsymbol{\Pi}_{\mathcal{S}^{\perp}}$ is an orthogonal projection operator on the column space of $\widehat{\mathbf{F}}_{\mathcal{S}}$. By the mutual incoherence condition in (A3), the first term of the right-hand side in (2.11) is upper-bounded by $1-\mu$ for some $\mu \in (0,1]$, with some probability $P_{\mu} \in [0,1]$. The remaining task is to control the tail probability of $\tilde{Z}_j$ for $j \in \mathcal{S}^c$: that is to ensure $\mathbb{P}\big[\max_{j\in\mathcal{S}^c} |\tilde{Z}_j| \geq \mu\big] \to 0$ with some exponential decay rate. With the help of Lemma B.3.1 in the Appendix, controlling the probability $\mathbb{P}\big[\|\tilde{Z}_{\mathcal{S}^c}\|_{\infty} \geq \mu\big]$ reduces to controlling $\mathbb{P}\big[\|\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^{*} - \Delta\mathbf{u}_t\|_{\infty} \geq$

$\mu\frac{\lambda_N}{\sqrt{K}}\big]$. Controlling the bound on $\mathbb{P}\big[\|\boldsymbol{\tau}\|_\infty \geq \varepsilon\big]$ for some $\varepsilon > 0$ is challenging, since the exact form of the residual distribution $\boldsymbol{\tau}$ is unknown. (Note that since $\boldsymbol{\beta}^*_{\mathcal{S}^c} = \mathbf{0} \in \mathbb{R}^{K-s}$, $\boldsymbol{\tau} = \Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}} - \Delta\mathbf{u}_t$.)

We circumvent this difficulty by using the following inequality: for some thresholds $\varepsilon_N > 0$ and $\varepsilon_M > 0$, both of which go to 0 as N and M tend to ∞, we have,

$$
\begin{aligned}
&\mathbb{P}\big[\,\|\boldsymbol{\tau}\|_\infty \geq \varepsilon_N + \varepsilon_M\big] \\
&\quad\leq \mathbb{P}\Bigg[\max_{0\leq i\leq M-1}\sup_{t\in[0,T_{\max})}|\Delta u_t(X_i,t)| \geq \varepsilon_N\Bigg] + \mathbb{P}\Bigg[\max_{\substack{1\leq k\leq s\\ 0\leq n\leq N-1}}\sup_{x\in[0,X_{\max})}|\Delta F_k(x,t_n)| \geq \frac{\varepsilon_M}{s\|\boldsymbol{\beta}^*\|_\infty}\Bigg] \\
&\quad\leq M\cdot\mathbb{P}\Bigg[\sup_{t\in[0,T_{\max})}|\Delta u_t(X_i,t)| \geq \varepsilon_N\Bigg] + sN\cdot\mathbb{P}\Bigg[\sup_{x\in[0,X_{\max})}|\Delta F_k(x,t_n)| \geq \frac{\varepsilon_M}{s\|\boldsymbol{\beta}^*\|_\infty}\Bigg].
\end{aligned}
$$

The above inequality leads us to study the uniform convergence of Local-Polynomial estimator to its ground-truth function of interest. Say, for sufficiently large enough grid size of temporal dimension N, for some $\varepsilon_N \geq 0$ that is h_N-dependent threshold and $X_i \in [0, X_{\max})$, we will achieve

$$
\mathbb{P}\Bigg[\sup_{t\in[0,T_{\max})}|\widehat{\mathbf{u}}_t(X_i,t) - \mathbf{u}_t(X_i,t)| > \varepsilon_N\Bigg] \to 0, \tag{2.12}
$$

with an exponential decay rate. As for obtaining the exponential decay rate in (2.12), we defer the detailed explanation with some intuitions in the following Subsection. It turns out that thresholds ε_N and ε_M are functions of bandwidth parameters h_N and w_M in (2.5) and (2.6). We choose correct orders of h_N and w_M so that we can ensure that the thresholds ε_N and ε_M go to zero. Then, with the proper choice on the order of λ_N together with $\mathbb{P}\big[\|\boldsymbol{\tau}\|_\infty \geq \mu\frac{\lambda_N}{\sqrt{K}}\big] \to 0$ as $N \to \infty$, we conclude the proof.

2.5.3 Technical Contribution

Several researchers have tried to achieve uniform convergence of Local-Polynomial or kernel smoothing estimators in almost sure sense. See the works [99] and [100]. However, to the best of the authors' knowledge, uniform convergence of Local-Polynomial estimator with an explicit decaying probability rate has not been studied in the literature. We provide it as a technical contribution of the present paper. Readers can find the exact statements of these results for the estimators $\widehat{\mathbf{u}}_t$ and $\widehat{\partial_x^p u}$ for $0 \leq p \leq P_{\text{max}}$ in the Appendix stated as Lemma B.3.2 and Lemma B.3.3, respectively.

Here, we provide a high-level idea of the proof of Lemma B.3.2. First, we observe that the higher-order Local-Polynomial smoothing is asymptotically equivalent to higher-order kernel smoothing through equivalent kernel theory [67]. See (2.5) and (2.6) for their equivalences in mathematical form with kernel smoothing estimators. Second, we employ the truncation idea in [97] on the Local-Polynomial estimator and decompose $\widehat{\mathbf{u}}_t(X_i, t) - \mathbf{u}_t(X_i, t)$ into three parts as follows:

$$\widehat{\mathbf{u}}_t - \mathbf{u}_t = \underbrace{\left(\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N} - \mathbb{E}\big(\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N}\big)\right)}_{\text{Asymptotic deviation of truncation error}} + \underbrace{\left(\widehat{\mathbf{u}}_t^{B'_N} - \mathbb{E}\widehat{\mathbf{u}}_t^{B'_N}\right)}_{\substack{\text{Asymptotic deviation of}\\ \text{truncated estimator}}} + \underbrace{\left(\mathbb{E}\widehat{\mathbf{u}}_t - \mathbf{u}_t\right)}_{\substack{\text{Asymptotic bias of}\\ \text{Local-Polynomial estimator}}} ,$$

where B'_N is some increasing sequence in N, and $\widehat{\mathbf{u}}_t^{B'_N}$ denotes the truncated Local-Polynomial estimator of $\mathbf{u}_t$. We control the *sup* over $t \in [0, T_{\text{max}})$ on each of the three components. The last component, *Asymptotic bias* of $\widehat{\mathbf{u}}_t$ can be obtained through the classical result from [67, 68]. The exponential decay rate comes from the first two components as follows:

1. *Asymptotic deviation of truncation error* can be decomposed into two parts. The first part, which is $\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N}$, can be easily controlled via Chernoff bound of Gaussian random variable. by using the definition of truncated estimator $\widehat{\mathbf{u}}_t^{B'_N}$. The second part, which is the expected difference $\mathbb{E}\big(\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N}\big)$, can be bounded by some deterministic function of B'_N and h_N using the similar arguments in Proposition 1 of [97].

2. *Asymptotic deviation of truncated estimator* is decomposed into two components as

well: (i) Brownian bridge and (ii) difference between some two-dimensional empirical process and the Brownian bridge. (i) can be controlled via uniform convergence of Gaussian Process using the arguments similar to [101], together with simple Markov inequality. (ii) can be controlled via Tusnády's strong uniform approximation theory [97, 98], stating that the two-dimensional empirical process can be well approximated by a certain solution path of two-dimensional Brownian bridge.

Same ideas can be employed for the uniform convergence of $\widehat{(\partial_x^p u)_i^n}$ to $(\partial_x^p u)_i^n$ and of $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}$ to $(\partial_x^p u)_i^n(\partial_x^q u)_i^n$ for $0 \le p, q \le P_{\max}$.

2.6 Uniform Convergence of Sample Incoherence Matrix

In this section, we provide two Lemmas 2.6.2 and 2.6.3 that can complete the proof of Theorem 2.4.1. Here, the minimum-eigenvalue and incoherence assumptions are imposed on the ground-truth feature matrix $\mathbf{F}$, instead on the estimated feature matrix $\widehat{\mathbf{F}}$. See (A1) and (A2). That is, there exist $C_{\min} > 0$ and $\mu \in (0, 1]$ such that the followings hold for the unknown support set $\mathcal{S}$:

$$\Lambda_{\min}\Big(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\Big) \ge C_{\min} \quad \text{and} \quad \|\mathcal{Q}^*\|_\infty \le 1 - \mu.$$

Equipped with the above assumptions, we can formally show that success probability of the sample incoherence condition P_μ in (A3) tends to 1 as $N \to \infty$.

The key step of the proofs in the following Lemmas is to control the tail probability of difference between inner-product of two arbitrary columns of $\widehat{\mathbf{F}}$ and inner-product of the two corresponding columns of ground-truth $\mathbf{F}$. This problem is challenging even if the exact distribution of any entries of $\widehat{\mathbf{F}}$ is known, since the distribution of $\sum_{k=1}^{NM}\widehat{F}_{ki}\widehat{F}_{kj}$ needs to be derived. In order to circumvent this problem, we take the advantage of the uniform convergence result of $\widehat{(\partial_x^p u)_i^n}$ for any $0 \le p \le P_{\max}$ proved in Lemma B.3.3. Additionally, we need the uniform convergence results of $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}$, $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}\widehat{(\partial_x^k u)_i^n}$, and $\widehat{(\partial_x^p u)_i^n (\partial_x^q u)_i^n (\partial_x^k u)_i^n (\partial_x^l u)_i^n}$ for $0 \leq p, q, k, l \leq P_{\text{max}}$. These convergence results are explicitly stated as Corollaries B.3.7, B.4.1, and B.4.2, with proofs in the Appendix.

Equipped with the uniform convergence results, we introduce a following Lemma stating that the distance between the matrices $\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}}$ and $\mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}$ are close enough under operator norm for large enough grid sizes.

Lemma 2.6.1 *Let $\varepsilon_M^*, \varepsilon_M^{**}, \varepsilon_M^{***}, \varepsilon_M^{****}$ be the thresholds defined in B.3.3, B.3.7, B.4.1, and B.4.2. Then for any $\varepsilon_M^{max'}$ such that*

$$\varepsilon_M^{max'} > \sqrt{s(K-s)} \max\left\{\varepsilon_M^*, \varepsilon_M^{**}, \varepsilon_M^{***}, \varepsilon_M^{****}\right\},$$

then, for $0 < c < \frac{2}{7}$, and for sufficiently large enough N, we have

$$\mathbb{P}\left[\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right\|_2 > \varepsilon_M^{max'}\right] \leq \mathcal{O}\Big(N \exp\big(-\frac{1}{6}N^c\big)\Big).$$

Now, we are ready to prove our main claims, Lemmas 2.6.2 and 2.6.3. We first state and prove the Lemma 2.6.2 asserting that if there exists $C_{\text{min}} > 0$ such that the minimum eigen-value condition holds for $\mathbf{F}_{\mathcal{S}}$, then the sample minimum eigen-value condition holds with probability converging to 1 with an exponential decay rate.

Lemma 2.6.2 *Suppose that the assumption (A1) holds with some constant $C_{min} > 0$ and $0 < c < \frac{2}{7}$, then with probability at least $1 - \mathcal{O}(N\exp(-\frac{1}{6}N^c)) \to 1$ as $N \to \infty$, we have,*

$$\Lambda_{\text{min}}\Big(\frac{1}{NM}\widehat{\mathbf{F}}_{\mathcal{S}}^T \widehat{\mathbf{F}}_{\mathcal{S}}\Big) \geq C_{\text{min}}\ .$$

Proof. Observe that we can write:

$$\begin{aligned}\Lambda_{\text{min}}\Big(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\Big) &= \frac{1}{NM}\min_{\|x\|_2=1}\left\{x^\top\Big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\Big)x + x^\top\Big(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\Big)x\right\}\\ &\leq \frac{1}{NM}\left\{y^\top\Big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\Big)y + y^\top\Big(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\Big)y\right\}\end{aligned}$$

where $y \in \mathbb{R}^K$ is a unit-norm minimal eigen-vector of $\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}$. Therefore, we can write,

$$\begin{aligned}\Lambda_{\text{min}}\bigg(\frac{1}{NM}\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\bigg) &\geq \Lambda_{\text{min}}\bigg(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\bigg) - \frac{1}{NM}\left\|\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2 \\ &\geq C_{\text{min}} - \frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right\|_2 .\end{aligned}$$

By using a similar argument used in Lemma 2.6.1, we can prove $\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right\|_2 \to 0$ with high-probability as $N \to \infty$. For any $\varepsilon_M^{\text{max}}$ such that,

$$\varepsilon_M^{\text{max}} > s \max\left\{\varepsilon_M^{*}, \varepsilon_M^{**}, \varepsilon_M^{***}, \varepsilon_M^{****}\right\}, \tag{2.13}$$

Then, we can bound the probability as follows:

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right\|_2 > \varepsilon_M^{\text{max}}\right] \\ &\quad\leq \mathbb{P}\left[\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right\|_{\text{F}} > NM\varepsilon_M^{\text{max}}\right] \leq \mathbb{P}\left[\big(NMs\big)\cdot\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right\|_{\infty,\infty} > NM\varepsilon_M^{\text{max}}\right] \\ &\quad\leq \mathbb{P}\left[\max_{n=0,\ldots,N-1}\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}}_i(x,t_n)\widehat{\mathbf{F}_j}(x,t_n) - \mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right| > \frac{\varepsilon_M^{\text{max}}}{s}\right] \\ &\quad\leq \sum_{n=0}^{N-1}\mathbb{P}\left[\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}}_i(x,t_n)\widehat{\mathbf{F}_j}(x,t_n) - \mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right| > \frac{\varepsilon_M^{\text{max}}}{s}\right] \\ &\quad\leq \mathcal{O}\bigg(N\exp\big(-\frac{1}{6}N^c\big)\bigg) .\end{aligned}$$

□

With the help of Lemma 2.6.2, we can show that the sample incoherence condition holds with high probability, given that there exists $\mu \in (0,1]$ for the ground-truth version of (A2).

Lemma 2.6.3 *Suppose that the assumption (A2) holds with some constant* $\mu \in (0,1]$ *and* $0 < c < \frac{2}{7}$*, then with probability at least* $1 - \mathcal{O}(N\exp(-\frac{1}{6}N^c)) \to 1$ *as* $N \to \infty$*, we have,*

$$\left\|\widehat{\mathcal{Q}}_N\right\|_\infty \leq 1 - \mu .$$

Proof. Motviated from [75], we begin the proof by decomposing the matrix $\big(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}$ into four parts:

$$
\begin{aligned}
\big(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} = {} & \underbrace{\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\Big(\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} - \big(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\big)^{-1}\Big)}_{:=\mathbf{T_1}} + \underbrace{\Big(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\Big)\big(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\big)^{-1}}_{:=\mathbf{T_2}} \\
& + \underbrace{\Big(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\Big)\Big(\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} - \big(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\big)^{-1}\Big)}_{:=\mathbf{T_3}} \\
& + \underbrace{\big(\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\big)\big(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\big)^{-1}}_{:=\mathbf{T_4}}.
\end{aligned}
$$

Since we know $\|\mathbf{T_4}\|_\infty \leq 1-\mu$ for some $\mu \in (0,1]$, the decomposition reduces the proof showing $\|\mathbf{T_i}\|_\infty \to \mathbf{0}$ with probability $1-\mathcal{O}(N\exp(-\frac{1}{6}N^c))$ for $i=1,2,3$.

1. Control of $\mathbf{T_1}$: Observe that we can re-factorize $\mathbf{T_1}$ as follows:

$$
\mathbf{T_1} = \big(\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\big)\big(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\big)^{\mathbf{-1}}\big[\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big]\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{\mathbf{-1}}.
$$

Then, by taking the advantage of sub-multiplicative property $\|AB\|_\infty \leq \|A\|_\infty\|B\|_\infty$ and the fact $\|\mathbf{T_4}\|_\infty \leq 1-\mu$ and $\|C\|_\infty \leq \sqrt{N}\|C\|_2$ for $C \in \mathbb{R}^{M\times N}$, we can bound $\|\mathbf{T_1}\|_\infty$ as follows:

$$
\begin{aligned}
\|\mathbf{T_1}\|_\infty &\leq \Big\|\big(\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\big)\big(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\big)^{-1}\Big\|_\infty \Big\|\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\Big\|_\infty \Big\|\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\Big\|_\infty \\
&\leq s(1-\mu)\bigg(\frac{1}{NM}\Big\|\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\Big\|_2\bigg)\bigg(NM\Big\|\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\Big\|_2\bigg) \\
&\leq \frac{s(1-\mu)}{C_{\min}}\bigg(\frac{1}{NM}\Big\|\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\Big\|_2\bigg).
\end{aligned}
$$

Note that we use $\|\big(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\|_2 \leq \frac{1}{NMC_{min}}$ with probability $1-\mathcal{O}(N\exp(-\frac{1}{6}N^c))$ in the last inequality from Lemma 6.1.

2. Control of $\mathbf{T_2}$: With similar techniques employed for controlling $\|\mathbf{T_1}\|_\infty$, we can bound $\|\mathbf{T_2}\|_\infty$ as follows:

$$\begin{aligned}
\|\mathbf{T_2}\|_\infty &\leq \left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right\|_\infty \left\|\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_\infty \\
&\leq s \left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right\|_2 \left\|\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_2 \\
&= s\left(\frac{1}{NM} \left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right\|_2\right)\left(NM \left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right\|_2\right) \\
&\leq \frac{s}{C_{\text{min}}}\left(\frac{1}{NM} \left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right\|_2\right).
\end{aligned}$$

3. Control of $\mathbf{T_3}$: To bound $\|\mathbf{T_3}\|_\infty$, we re-factorize the second argument of product in $\mathbf{T_3}$:

$$\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} - \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1} = \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\left[\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right) - \left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}$$

With the factorization, we bound $\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} - \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\|_\infty$ by using sub-multiplicative property and the fact $\|C\|_\infty \leq \sqrt{N}\|C\|_2$ for any $C \in \mathbb{R}^{M\times N}$ again:

$$\begin{aligned}
\left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} - \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_\infty &= \left\|\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\left[\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right) - \left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right\|_\infty \\
&\leq \sqrt{s}\left\|\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\left[\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right) - \left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right\|_2 \\
&\leq \sqrt{s}\left\|\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_2 \left\|\left[\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right) - \left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\right\|_2 \left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right\|_2 \\
&\leq \frac{\sqrt{s}}{NMC_{\text{min}}^2}\left(\frac{1}{NM}\left\|\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2\right). \qquad (2.14)
\end{aligned}$$

In the last inequality, we use the result of Lemma 6.1. Now we can bound $\|\mathbf{T_3}\|_\infty$ as follows:

$$\begin{aligned}
\|\mathbf{T_3}\|_\infty &= \left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right)\left(\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} - \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right)\right\|_\infty \\
&\leq \left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right\|_\infty \left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} - \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_\infty \\
&\leq \frac{s}{C_{\text{min}}}\left(\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right\|_2\right)\left(\frac{1}{NM}\left\|\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2\right),
\end{aligned}$$

where in the last inequality, we use (2.14) and $\|C\|_\infty \leq \sqrt{N}\|C\|_2$ for any $C \in \mathbb{R}^{M\times N}$. Take $\varepsilon_M^{\max''}$ such that,

$$\varepsilon_M^{\max''} > \max\left\{ \frac{C_{\min}}{s(1-\mu)}\varepsilon_M^{\max}, \frac{C_{\min}}{s}\varepsilon_M^{\max'} \right\},$$

with $\varepsilon_M^{\max}$ in (2.13) and with $\varepsilon_M^{\max'}$ in Lemma 2.6.1, respectively. Then, we have

$$\mathbb{P}\left[\forall i = 1,2,3 : \|\mathbf{T_i}\|_\infty > \varepsilon_M^{\max''} \right] \leq \mathcal{O}\Big(N \exp\big(-\frac{1}{6}N^c \big) \Big).$$

□

Verification of Lemma 2.6.3 automatically leads to the complete proof of Theorem 2.4.1, together with Proposition 2.5.1. Therefore, as long as the two assumptions (A1) and (A2) hold for $\mathbf{F}$, with sufficiently fine-grained grid points over the function $u(X,t)$, ℓ_1-PsLS can always find the correct signed-support of the given PDE model, with the minimum absolute value of $\boldsymbol{\beta}_S^*$ not too close to zero.

2.7 Numerical Experiments

In the first subsection, two PDE models and data-generating processes of respective models are introduced. In the next subsection, we verify the main statements of the Theorem 2.4.1 through numerical experiments over the PDE models described in Subsection 2.7.1. The impact of $\boldsymbol{\beta}_{\min}^*$-condition in the signed-support recovery of ℓ_1-PsLS is numerically explored in subsection 2.7.3.

2.7.1 Experimental Setting

In this subsection, we provide detailed descriptions on (i) two popular PDE models that we are going to work on throughout the Section 2.7, and on (ii) how to generate the data from respective models, and (iii) how to design the regression problem for the experiments to be presented.

Model Specification and Data Generation

Viscous Burgers' equation is a fundamental second-order semilinear PDE which is frequently employed to model physical phenomena in fluid dynamics [102] and nonlinear acoustic in dissipative media [103]. Its general form is

$$u_t = -u\partial_x u + \nu \partial_x^2 u$$

where $\nu > 0$ is the diffusion coefficient which characterizes physical quantities such as viscosity of fluid. Specifically, when $\nu = 0$, it becomes an inviscid Burgers' equation, which is a conservative system that can form shock waves. Here we consider the following viscous Burgers' equation:

$$\begin{aligned} & u_t = -u\partial_x u + \nu \partial_x^2 u\ ,\ \ 0 < x < 1, 0 < t < 0.1 \qquad (2.15) \\ & u(x,0) = \sin^2(2\pi x) + \cos^3(3\pi x)\ ,\ \ 0 \le x \le 1\ ,\ \ u(0,t) = u(1,t)\ ,\ \ 0 \le t \le 0.1. \end{aligned}$$

Korteweg–de Vries equation is well known for its solution that demonstrates the phenomenon of superposition of nonlinear waves [104], and for modeling fluid dynamics of shallow water surfaces in long and narrow channels [105]. Its dimensionless form is given

$$u_t + \partial_x^3 u + 6u\partial_x u = 0\ . \qquad (2.16)$$

In this Section, we consider the form of (2.16), whose initial solution is as follows:

$$\begin{aligned} & u(x,0) = 3.5\sin^3(4\pi x) + 1.5\exp\big(-\sin(2\pi x)(1-x)\big)\ , \\ & 0 \le x \le 1\ ,\ \ u(0,t) = u(1,t)\ ,\ \ 0 \le t \le 0.1. \end{aligned}$$

Data Generation For N-size sampling in the temporal dimension, by Theorem 2.4.1, we take $M = \lfloor N^{(2\times P_{\max}+5)/7} \rfloor$ sample size in the space dimension. We numerically solve Viscous Burgers' equation (2.15) by the Lax-Wendroff scheme on a grid with interval width $\delta t = 0.1/(100N)$ in temporal and $\delta x = 1/M$ in space, then we downsampled the data in the temporal dimension by a factor of 100; thus the resulted clean data is distributed over a grid with N nodes in time and M nodes in space. Lastly, we added i.i.d. Gaussian noise with standard deviation $\sigma = 0.25$ to the data. i.e., $\nu_i^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.25^2)$. As for solving the KdV equation (2.16), the same approaches with Viscous Burger's equation are applied, with i.i.d. Gaussian noises with standard deviation $\sigma = 0.025$.

Constructions of Regression Problems

We employ the Local-Polynomial smoothing for estimating $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ as described in Subsection 2.3.2. Regarding a choice of kernel for constructing $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$, we use the Epanechnikov kernel defined by:

$$\mathcal{K}(z) = \frac{3}{4}(1 - z^2)_+ \ , \ z \in \mathbb{R} \ ,$$

where $(\cdot)_+ := \max(0, \cdot)$. Bandwidth parameters h_N and w_M in (2.3) and (2.4) are chosen in the order of $h_N = \Theta(N^{-\frac{1}{7}})$ and $w_M = \Theta(M^{-\frac{1}{7}})$, respectively. As displayed in Table 2.1, for the experiments presented in this Section, we choose specific constant factors in the order expressions of h_N and w_M for Viscous Burgers equation and KdV equation. Regarding more detailed issues on the choices of these constants, readers can refer to Section 2.8. It is also worth noting that we do not use (2.5) and (2.6) as solutions of the optimization problems (2.3) and (2.4) for the experiments, since the expressions in (2.5) and (2.6) are derived in asymptotic settings. For the reader's convenience, We provide the closed form solutions of (2.5) and (2.6) in Appendix B.2.

For Viscous Burgers' equation with noisy data, Local-Polynomial fitting with $P_{\max} = 2$

Table 2.1: Specific choices of the constants in the order of $h_N = \Theta(N^{-\frac{1}{7}})$ and $w_M = \Theta(M^{-\frac{1}{7}})$ for the experiments on Viscous Burgers equation and KdV equation are presented.

	w_M	h_N
Viscous Burgers	$0.75M^{-\frac{1}{7}}$	$0.25N^{-\frac{1}{7}}$
KdV	$0.1M^{-\frac{1}{7}}$	$0.01N^{-\frac{1}{7}}$

is applied to construct $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$. Our goal is to identify the fifth and the sixth coefficients, $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$, of a following linear measurement via the proposed ℓ_1-PsLS model (2.8): Here, we denote $\widehat{\mathbf{u}}_x := \widehat{\partial_x \mathbf{u}}$ and $\widehat{\mathbf{u}}_{xx} := \widehat{\partial_x^2 \mathbf{u}}$.

$$\widehat{\mathbf{u}}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\widehat{\mathbf{u}} + \boldsymbol{\beta}_2\widehat{\mathbf{u}}^2 + \boldsymbol{\beta}_3\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_4\widehat{\mathbf{u}}_x^2 + \boldsymbol{\beta}_5\widehat{\mathbf{u}}\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_6\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_7\widehat{\mathbf{u}}_{xx}^2 + \boldsymbol{\beta}_8\widehat{\mathbf{u}}_x\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_9\widehat{\mathbf{u}}\widehat{\mathbf{u}}_{xx}.$$

For KdV equation, after generating the data-points, $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ are fitted through Local-Polynomial with $P_{max} = 3$. We want ℓ_1-PsLS to select $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_{10}$ as non-zero coefficients in a following linear measurement: Here, denote $\widehat{\mathbf{u}}_{xxx} := \widehat{\partial_x^3 \mathbf{u}}$.

$$\begin{aligned}\widehat{\mathbf{u}}_t = {} & \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\widehat{\mathbf{u}} + \boldsymbol{\beta}_2\widehat{\mathbf{u}}^2 + \boldsymbol{\beta}_3\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_4\widehat{\mathbf{u}}_x^2 + \boldsymbol{\beta}_5\widehat{\mathbf{u}}\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_6\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_7\widehat{\mathbf{u}}_{xx}^2 + \boldsymbol{\beta}_8\widehat{\mathbf{u}}_x\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_9\widehat{\mathbf{u}}\widehat{\mathbf{u}}_{xx} \\ & + \boldsymbol{\beta}_{10}\widehat{\mathbf{u}}_{xxx} + \boldsymbol{\beta}_{11}\widehat{\mathbf{u}}_{xxx}^2 + \boldsymbol{\beta}_{12}\widehat{\mathbf{u}}_x\widehat{\mathbf{u}}_{xxx} + \boldsymbol{\beta}_{13}\widehat{\mathbf{u}}_{xx}\widehat{\mathbf{u}}_{xxx} + \boldsymbol{\beta}_{14}\widehat{\mathbf{u}}\widehat{\mathbf{u}}_{xxx}.\end{aligned}$$

2.7.2 Numerical Verifications of Main Statements

In this subsection, we design an experiment to numerically verify following two main statements of this paper. [2]

1. *Under assumptions (A1) and (A2), and with large enough data points, there exist some $\lambda_N \geq 0$ such that ℓ_1-PsLS model* (2.8) *recovers a signed-support* $\big(\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}^{\lambda}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^*)\big)$ *of an unique PDE that admits the underlying function as a solution in probability.*

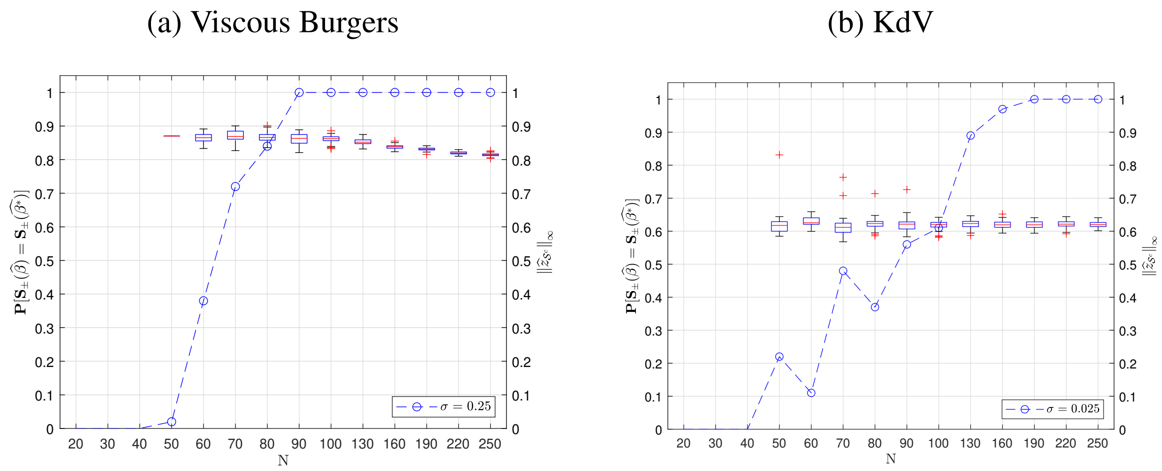
[2] Results provided in Subsections 2.7.2 and 2.7.3 can be reproduced via MATLAB codes in *https://github.com/namjoonsuh/PDE-identification*.

2. *Given assumption (A2) for some $\mu \in (0, 1]$, sampled incoherence parameter μ' converges to ground-truth μ in probability with large enough data points.*

The experiment is conducted over two PDE models, **Viscous Burgers' equation** and **KdV equation** introduced in Subsection 2.7.1. We generate the data by setting $\nu = 0.03$ in (2.15). In Figure 2.1, the probability of signed-support recovery $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^{*})]$ versus the grid size of temporal dimension N, and $\|\widehat{\mathbf{z}}_{\mathcal{S}^{\mathsf{c}}}\|_{\infty}$ versus N are recorded on the same plot for respective models. Each point on each curve, which represents $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^{*})]$, in (a) and (b) corresponds to the average over 100 trials. For each iteration, the hyper-parameter λ_N is chosen in an "optimal" way: we used the value yielding the correct number of nonzero coefficient. With the chosen λ_N, $\widehat{\mathbf{z}}_{\mathcal{S}^{\mathsf{c}}}$ is calculated as given in (2.11). Note that (2.11) can be calculated only when the ℓ_1-PsLS finds λ_N that gives the minimizer of (2.8) $\widehat{\boldsymbol{\beta}}^{\lambda}$ such that $\widehat{\boldsymbol{\beta}}^{\lambda}_{\mathcal{S}^{c}} = 0$ and $\mathcal{S}(\widehat{\boldsymbol{\beta}}^{\lambda}) \subseteq \mathcal{S}(\boldsymbol{\beta}^{*})$. For this reason, boxplots of $\|\widehat{\mathbf{z}}_{\mathcal{S}^{\mathsf{c}}}\|_{\infty}$ in (a) and (b) are drawn from the point when ℓ_1-PsLS starts to find such λ_N. For both models, $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^{*})]$ goes to 1, as we observe more data points on finer grid. Furthermore, it is worth noting that the *strict dual feasibility* condition (i.e., $\|\widehat{\mathbf{z}}_{\mathcal{S}^{\mathsf{c}}}\|_{\infty} < 1$) holds for both cases. In Figure 2.2, boxplots of $\|\widehat{\mathcal{Q}}_N\|_{\infty}$ versus N are displayed for Viscous Burgers' equation and kDV equation respectively. A dotted horizontal line in each panel represents $1 - \mu$ calculated from the ground-truth feature matrix $\mathbf{F}$. Notice that as the number of observed data gets larger, the sampled incoherence parameter goes below the dotted lines for both models.

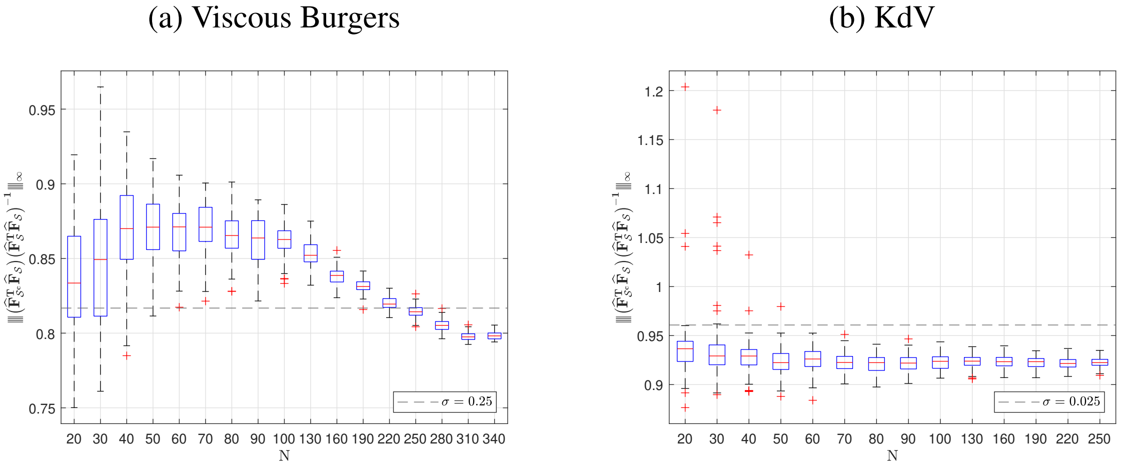
2.7.3 Impact of $\beta^{*}{}_{\text{min}}$ in Signed-Support Recovery of ℓ_1-PsLS

Theorem 2.4.1 states that as long as $\beta^{*}_{\text{min}} := \min_{i \in \mathcal{S}} |\beta^{*}_{i}|$ is beyond certain threshold, ℓ_1-PsLS is signed-support recovery consistent. In this subsection, we design an experiment to numerically confirm this claim. The experiment is performed over Viscous Burgers' equation by varying the coefficient ν in (2.15) : we set $\nu = 0.03, 0.02, 0.01, 0.005$ The


(a) Viscous Burgers

(b) KdV

Figure 2.1: Probability of signed-support recovery $\mathbb{P}[\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)]$ versus the grid size of temporal dimension N, and $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty$ versus N are recorded on the same plot for Viscous Burger's equation in panel (a) and for KdV equation in panel (b), respectively.


(a) Viscous Burgers

(b) KdV

Figure 2.2: Boxplots of $\|\widehat{\mathcal{Q}}_N\|_\infty$ versus N are displayed for Viscous Burgers' equation in panel (a) and KdV equation in panel (b), respectively.

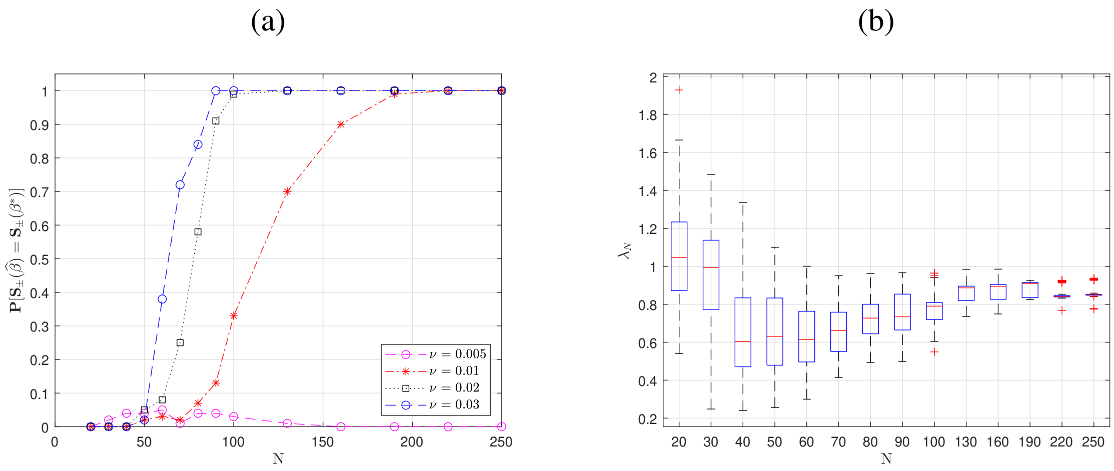


Figure 2.3: Left panel (a) displays the curves representing $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^*)]$ versus N, when $\nu = 0.03, 0.02, 0.01, 0.005$. Right panel (b) exhibits the range of λ_N for which ℓ_1-PsLS gives the solution $\widehat{\boldsymbol{\beta}}^{\lambda}$ such that $\mathcal{S}(\widehat{\boldsymbol{\beta}}^{\lambda}) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$ with respect to N, when ν is set as 0.005.

Figure 2.3 (a) displays the curves representing $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^*)]$ versus N for each of the four cases. Each point on each curve represents the average over 100 trials. The Figure 2.3 (b) exhibits the range of λ_N for which ℓ_1-PsLS finds the support of $\widehat{\boldsymbol{\beta}}^{\lambda}$ that is contained within the true support, when ν is set as 0.005. More specifically, boxplots in (b) record the range of λ_N that picks $\widehat{\mathbf{u}}_{xx}$ as the selected argument. In (a), we can check that, as the magnitude of $\min_{i \in \mathcal{S}} |\boldsymbol{\beta}_i^*|$ decreases from 0.03 to 0.01, ℓ_1-PsLS requires more data-points for the signed-support recovery, and when $\min_{i \in \mathcal{S}} |\boldsymbol{\beta}_i^*|$ drops to 0.005, ℓ_1-PsLS fails to recover the governing PDE. On the other hand, (b) says that there exists a range of λ_N for which ℓ_1-PsLS can still recover a subset of $\boldsymbol{\beta}^*$, while the perfect signed-support recovery is difficult.

2.8 Discussion

We present future directions that can be further explored based on our ℓ_1-PsLS method.

1. Recall that our theory utilizes the equivalent kernel theory for Local-Polynomial regression [67], stating that the higher-order Local-Polynomial smoothing is asymp-

totically equivalent to higher-order kernel smoothing. Due to this construction, our theory cannot characterize the convergence behavior of signed-support recovery of ℓ_1-PsLS, when the number of observations is small. We conjecture that the uniform convergence rate of the Local-Polynomial estimator with exponential decay can be obtained in a non-asymptotic sense, by using a similar technique employed in [106]. They impose an assumption that the regression function belongs to the Hölder class. They manipulate the closed-form solution of the Local-Polynomial estimator so that the difference of the estimator and the regression function has a special form that can be controlled by the Bernstein's inequality. It would be an interesting research direction to see whether this technique can be employed in our setting.

2. The choice of the bandwidth parameter is essential in Local-Polynomial fitting, thereby having a significant impact on support recovery of PDE problem via ℓ_1-PsLS. It is worth noting that [69] employed the substitution method in [107] based on the asymptotic Mean Integrated Squared Error for the specific choices of the constant factors of the bandwidth parameter. However, the method is only limited to the local-quadratic estimator and is not applicable to our setting, which requires a higher-order smoothing estimator. In our numerical experiments, we choose the constant factors of bandwidth parameters h_N and w_M manually. It only provides an ad-hoc guidance of bandwidth selection. Developing a data-driven bandwidth selection procedure for ℓ_1-PsLS is a worthy topic for future research.

3. In practice, we need to set P_{max} large so as to avoid the model misspecification. Specifically, when P_{max} is set to be very large, the dimension of columns of $\widehat{\mathbf{F}}$ can be approximated as $K \approx \big(P_{\max} + 1\big)^2$ in our problem setting. (Recall that we set $K = 1 + 2(P_{\max} + 1) + \binom{P_{\max}+1}{2}$.) Under finite grid size NM, it is a possible scenario in which we have $K \gg NM$. Can we reduce the computational burdens in this case? As one possible direction, we can think of using the Sure Indepen-

dence Screening (SIS) process [108] before solving ℓ_1-PsLS in (3.7). SIS is a dimension reduction technique before implementing variable selection algorithms, such as LASSO, SCAD, LARS, etc. In our case, for implementing SIS, we need to compute the marginal correlation between the response vector $\mathbf{u}_t$ and columns in $\widehat{\mathbf{F}}$, denoted as $\omega = \widehat{\mathbf{F}}^\top \widehat{\mathbf{u}}_t \in \mathbb{R}^K$. The paper [108] proved that with a certain choice of d, it is guaranteed that all the relevant predictors in $\widehat{\mathbf{F}}$ with $\widehat{\mathbf{u}}_t$ are included under regularity conditions on $\widehat{\mathbf{F}}$. Then, we may choose the largest d entries of the vector $|\omega|$, such that $K \gg NM \gg d$. The computational complexities of solving (3.7) via the well-known LARS algorithm [109] is known to be in the order of $\mathcal{O}\big(NMp{\cdot}\min(NM, p)\big)$, where p is set to be K before implementing the SIS and d after implementing the SIS. However, we need further studies on whether SIS will work well in the PDE identification problem, with theoretical guarantees. We leave this as a future work.

4. As one of the referees mentioned, the Theorem 2.4.1 cannot provide a guideline in practice, whether the selected model excludes crucial terms or even includes the irrelevant terms. To the best of the our knowledge, this is largely an open problem in the PDE identification context. From statisticians' viewpoint, we can suggest constructing a hypothesis testing, for $j \in \{1, 2, \ldots, K\}$, $H_0 : \widehat{\beta}_j^\lambda = 0$ v.s. $H_1 : \widehat{\beta}_j^\lambda \neq 0$. However, this is a challenging problem since we need to derive the distribution of the estimated coefficient $\widehat{\beta}_j^\lambda$ for each $j \in \{1, \ldots, K\}$. We are aware of the work [110] on constructing the confidence intervals of the LASSO estimator $\widehat{\beta}_j^\lambda$ under the classical $i.i.d.$ centered normal error distribution. Nonetheless, this assumption is not applicable in our problem setting, and requires further investigations. We leave this problem as a future work.

5. It is worth noting that our paper is about model selection consistency of PDEs under noisy data and we consider the study on the estimation accuracy of the selected model is beyond the scope of our work. Nevertheless, it is still of importance to in-

vestigate whether the regression-based PDEs give a solution that closely resembles the original one. In practice, we suggest using the least-squares estimate with the the selected features through ℓ_1-PsLS; that is, given that the ℓ_1-PsLS selects the true support set $\mathcal{S}$, then the least-squares estimate has a form: $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}} := \left(\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{u}}_t$. Note that $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ can avoid the bias introduced from λ_N and gives the consistent estimate than ℓ_1-PsLS. Although not reported in the paper, we verify that the least-squares estimate $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ works pretty well for the cases of KdV and viscous Burger's equations in Section 2.7 in terms of estimation. We leave the study on the theoretical properties of this estimator as the future work. For more specific application with smaller data, there are related works with more refined model selection procedure, including [83] and [111]. We refer the readers these works and references therein.

CHAPTER 3
HIGH-DIMENSIONAL MULTIVARIATE LINEAR REGRESSION WITH WEIGHTED NUCLEAR NORM REGULARIZATION

3.1 Introduction

We consider the problem of recovering an unknown coefficient matrix $\boldsymbol{\Theta}^\star \in \boldsymbol{R}^{d_1 \times d_2}$ from n observations of the response vector $y_i \in \mathbb{R}^{d_2}$, $1 \leq i \leq n$, and predictor $x_i \in \mathbb{R}^{d_1}$, where the ground truth model is as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta}^\star + \boldsymbol{E}, \tag{3.1}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ is an $n \times d_2$ matrix, $\boldsymbol{X} = (x_1, \ldots, x_n)^\top$ is an $n \times d_1$ matrix, and $\boldsymbol{E} = (e_1, \ldots, e_n)^\top$ is an $n \times d_2$ regression noise matrix. The vectors $\{e_j\}_{j=1}^n$ are independently sampled from $\mathcal{N}(0, \sigma^2 \cdot \mathcal{I}_{d_2 \times d_2})$ with variance parameter $\sigma^2 > 0$. Throughout the paper, we write $p :=$ min(d_1, d_2), $r^\star :=$ rank$(\boldsymbol{\Theta}^\star)$ and $\mathcal{I}_{m \times m}$ as an $m \times m$ identity matrix. The observational model (3.1) is referred to as a multivariate linear regression model in the statistics literature. This model is particularly attractive when there exists a dependence structure in the multivariate response, where the response matrix $\boldsymbol{Y}$ can be represented with a linear combination of only a small number of linearly transformed predictors. The situation is induced from the assumption that the coefficient matrix $\boldsymbol{\Theta}^\star$ has a low rank, that is $r^\star \ll p$.

To estimate $\boldsymbol{\Theta}^\star$ with a low rank structure, given the noisy measurement pair $(\boldsymbol{X}, \boldsymbol{Y})$, previous research has incorporated weighted nuclear norm (WNN) penalization with least square method for various applications, such as in computer vision ([112, 113, 114, 115]), biogenesis ([116]), and wireless system ([117]), but we are not aware of papers that develop estimation properties for penalized least square estimator with WNN penalty, which can be expressed as

$$\widehat{\boldsymbol{\Theta}} := \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta} \right\|_{\mathrm{F}}^2 + \boldsymbol{\lambda}_n \left\| \boldsymbol{\Theta} \right\|_{\boldsymbol{\omega},\star} \right\} \tag{3.2}$$

with the weighted nuclear norm

$$\left\| \boldsymbol{\Theta} \right\|_{\boldsymbol{\omega},\star} = \sum_{j=1}^{p} \omega_j \sigma_j(\boldsymbol{\Theta}), \tag{3.3}$$

where $\sigma_j(\boldsymbol{\Theta})$ means the j^{th} largest singular value of a matrix $\boldsymbol{\Theta} \in \boldsymbol{R}^{d_1 \times d_2}$, $\boldsymbol{\omega} = (\omega_1, ..., \omega_p)$, ω_j is a non-negative weight assigned to $\sigma_j(\boldsymbol{\Theta})$, $\boldsymbol{\lambda}_n \geq 0$ is a hyper-parameter, and $\| \cdot \|_{\mathrm{F}} := \sqrt{\sum_{j=1}^{p} \sigma_j(\cdot)^2}$ is the Frobenius norm.

On the other hand, there are a myriad of papers that studied the statistical properties of estimators related to standard nuclear norm (SNN) penalization, which is a special case of WNN with $\omega_1 = \cdots = \omega_p = 1$. Among them, [118] studied the least-squares problem with SNN penalization. Other literature studied the SNN under a more general model than multivariate linear regression (3.1) called trace regression ([47], [119], [5], [120]). Additionally, [121] investigated the SNN problem under a generalized trace regression problems for categorical responses, and [122] worked on obtaining the same minimax rate of a trace regression problem with [5] under the heavy-tail assumption on the design matrix and observational noise. The estimators from these SNN methods may still suffer higher bias than the estimator from the WNN method and generally has an estimator associated with higher rank. To support this observation, we demonstrate a simulation example that compares our WNN method with the SNN method ((3.2) with $\omega_1 = \cdots = \omega_p = 1$ in (3.3)) for estimating the singular values of $\boldsymbol{\Theta}^\star$. The simulation setting is described in details in Section 5.1. The result is summarized in Figure 1, and shows that our method achieves a satisfactory result within two iterations of loop with sample size $n = 250$ (Panel (A)), whereas there is still a slight bias on each of the estimated singular value from SNN with $n = 250$ (Panel B) and even with $n = 1000$ (Panel (C)). Such phenomenon is observed because WNN possesses flexibility to put the small weights on large singular values to

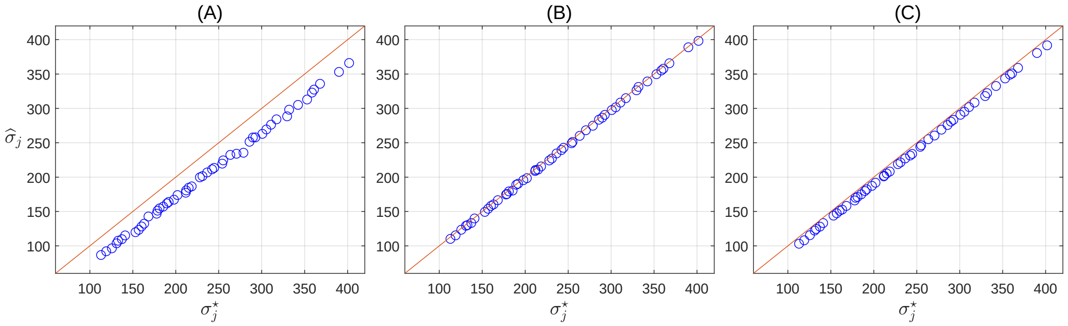


Figure 3.1: Three panels display the plots of estimated sigular values versus ground truth singular values $\sigma_j^\star$. The first two panels (A) and (B) are results from WMVR-ADMM algorithm with one weight update iteration under $n = 250$. The panel (C) exhibits the result when the estimator is obtained from SNN penalized least squares under $n = 1000$.

reduce the bias and to put the large weights on small singular values to encourage the estimated matrix to have a low rank.

Developing theoretical properties for the estimator from optimization problem (3.2) with (3.3) is non-trivial because finding the estimator is a non-convex problem under the desirable weights $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_p$. This makes the problem quite different from the adaptive lasso ([15]) in the context of sparse linear regression, which is always a convex optimization problem once the weights are fixed. [16] may be most related to our paper but they considered the WNN penalization on $\boldsymbol{X\Theta}$ instead of directly on $\boldsymbol{\Theta}$, which still does not provide the statistical properties for (3.2) with (3.3). Their theoretical analysis of [16] is focused on the behavior of prediction error, not like the estimation error developed in this paper.

Our contribution can be summarized into threefolds. First, we extend the classical alternative direction method of multipliers (ADMM) algorithm [37] to solve (3.2) with (3.3) and show that the sequence of tuples generated from the suggested algorithm converges uniquely to a stationary point of the augmented Lagrangain function. Furthermore, several estimation proprieties of the proposed estimator is derived, which provides more insights in understanding the proposed estimator from WMVR-ADMM algorithm. In this perspecitve, our paper provides a theoretical explanation on the role of weights for estimating the ground-truth coefficient matrix, and provide a non-asymptotic convergence rate of its singular values to its ground-truth counterparts. To be more specific, the result shows that the smaller weights compared to 2σ are desirable for estimating the non-zero $\sigma_j(\mathbf{\Theta}^\star)$s, whereas the larger weights than 2σ are required for estimating zero $\sigma_j(\mathbf{\Theta}^\star)$s under a proper choice of $\boldsymbol{\lambda}_n$. Under a Gaussian random design setting, we derive the minimax rate of the estimation error by adopting the technique used by [5] under high-dimensional regime (i.e., $n \ll d_1 d_2$). Finally, we develop a data-driven method for choosing the value of the tuning weights parameters $w_1, \cdots, w_p$ and the penalty parameter λ_n in (3.2) for practical use.

The rest of the paper is organized as follows. In Section 3.2, we introduce the details of WMVR-ADMM and provide a theorem on the algorithm's convergence guarantees. In Section 3.3, statistical properties of the estimator are provided. First, in the orthogonal design setting, the non-asymptotic convergence rate of the singular values from the proposed estimator, $\{\sigma_j(\widehat{\mathbf{\Theta}})\}_{j=1}^p$, is provided. Second, under a Gaussian random design, we obtain the minimax rate of the estimation error. In Section 3.4, a two-stage data-driven method for updating weights and choosing the regularization parameter is detailed. In Section 3.5, we compare the performance of our estimator with SNN and an estimator from [16] in terms of estimation error under various model parameter settings, and apply our algorithm to a real data set to demonstrate the validity of WMVR-ADMM in practice. Finally, we conclude our paper with the discussion section.

3.2 WMVR-ADMM and Convergence Analysis

To develop an algorithm for solving the non-convex optimization problem (3.2), we start with reformulating (3.2) as follows:

$$\min_{\Theta,\mathbf{\Gamma}} \left\{ f(\mathbf{\Theta}) + g(\mathbf{\Gamma}) \right\} \qquad \textbf{s.t.} \qquad \mathbf{\Theta} = \mathbf{\Gamma} \in \mathbb{R}^{d_1 \times d_2}, \tag{3.4}$$

by letting $f(\boldsymbol{\Theta}) := \boldsymbol{\lambda}_n \|\boldsymbol{\Theta}\|_{\omega,\star}$ and $g(\boldsymbol{\Gamma}) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Gamma}\|_{\mathrm{F}}^2$. This reformulation naturally leads to the construction of an augmented lagrangian function $\mathcal{L}_\rho\big(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}\big)$: For any $\rho > 0$ and dual variable $\boldsymbol{\Lambda} \in \mathbb{R}^{d_1 \times d_2}$, we define,

$$\mathcal{L}_\rho\big(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}\big) := f(\boldsymbol{\Theta}) + g(\boldsymbol{\Gamma}) + \mathbf{tr}\big(\boldsymbol{\Lambda}^\top\big(\boldsymbol{\Theta} - \boldsymbol{\Gamma}\big)\big) + \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Gamma}\|_{\mathrm{F}}^2. \tag{3.5}$$

Then, we solve the following three optimization problems repeatedly until primal and dual feasibility condition hold; that is, repeat **Steps 1-3**,

$$\begin{aligned}
&\textbf{Step 1.} \quad \boldsymbol{\Theta}^{(k+1)} = \underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \mathcal{L}_\rho\big(\boldsymbol{\Theta}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big),\\
&\textbf{Step 2.} \quad \boldsymbol{\Gamma}^{(k+1)} = \underset{\boldsymbol{\Gamma} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}^{(k)}\big),\\
&\textbf{Step 3.} \quad \boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big),
\end{aligned}$$

until $\|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\|_{\mathrm{F}} \leq 10^{-7}$ and $\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}} \leq 10^{-7}$. Here, we denote the tuple $(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})$ as the updated parameters at k^{th} iteration of the algorithm. Note that the non-convexity of the landscape of the objective function in **Step 1** arises from the WNN (i.e., $\|\cdot\|_{\omega,\star}$) over $\boldsymbol{\Theta}$ with fixed $\boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}$, whereas the objective function in **Step 2** is a simple quadratic function of $\boldsymbol{\Gamma}$ with fixed $\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Lambda}^{(k)}$.

The algorithm is conducted by initializing $\boldsymbol{\Theta}^{(0)} = \boldsymbol{\Gamma}^{(0)} = \boldsymbol{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{d_1 \times d_2}$. Next, the key of our algorithm is that a closed-form solution of **Step 1** can be obtained, even if it is a non-convex problem. We state the result in Lemma 3.2.1 whose proof is deferred in Section A of Supplemental Material.

Lemma 3.2.1 *Let $\boldsymbol{\Theta}^{(k+1)}$ be the minimizer of* ***Step 1****. Denote $\boldsymbol{B}^{(k)} := -\boldsymbol{\Lambda}^{(k)} + \rho \cdot \boldsymbol{\Gamma}^{(k)}$ and* *its SVD as $\boldsymbol{U}^{\boldsymbol{B}}\boldsymbol{D}^{\boldsymbol{B}}\big(\boldsymbol{V}^{\boldsymbol{B}}\big)^\top$. Then, for any fixed $\boldsymbol{\lambda}_n, \rho \geq 0$ and $0 \leq \omega_1 \leq \cdots \leq \omega_p$,*

$$\begin{aligned}
&\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{U}^{\boldsymbol{B}}\mathcal{S}_{\boldsymbol{\lambda}_n\omega}\big(\boldsymbol{D}^{\boldsymbol{B}}\big)\big(\boldsymbol{V}^{\boldsymbol{B}}\big)^\top, \\
&\mathcal{S}_{\boldsymbol{\lambda}_n\omega}\big(\boldsymbol{D}^{\boldsymbol{B}}\big) = \mathit{diag}\bigg\{\max\bigg\{\frac{1}{\rho}\big(\sigma_j(\boldsymbol{B}^{(k)}) - \boldsymbol{\lambda}_n w_j\big), 0\bigg\}, j = 1, \ldots, p\bigg\}.
\end{aligned}$$

Furthermore, if all the non-zero singular values of $\boldsymbol{B}^{(k)}$ are distinct, then the solution $\boldsymbol{\Theta}^{(k+1)}$ is unique.

For the optimization problem in **Step 2**, it can be rewritten and solved as follows:

$$\begin{aligned}
\boldsymbol{\Gamma}^{(k+1)} &= \operatorname*{argmin}_{\boldsymbol{\Gamma}\in\mathbb{R}^{d_1\times d_2}} \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}^{(k)}\big) \\
&= \operatorname*{argmin}_{\boldsymbol{\Gamma}\in\mathbb{R}^{d_1\times d_2}} \bigg\{\mathbf{tr}\bigg(\boldsymbol{\Gamma}^\top\bigg(\frac{1}{2n}\boldsymbol{X}^\top\boldsymbol{X} + \frac{\rho}{2}\cdot\mathcal{I}_{d_1\times d_1}\bigg)\boldsymbol{\Gamma} - \bigg(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{X} + \rho\cdot\boldsymbol{\Theta}^{(k+1)} + \boldsymbol{\Lambda}^{(k)}\bigg)^\top\boldsymbol{\Gamma}\bigg)\bigg\} \qquad (3.6) \\
&= \bigg(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \rho\cdot\mathcal{I}_{d_1\times d_1}\bigg)^{-1}\bigg(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{X} + \rho\cdot\boldsymbol{\Theta}^{(k+1)} + \boldsymbol{\Lambda}^{(k)}\bigg). \qquad (3.7)
\end{aligned}$$

Note that the quadratic equation (3.6) always has an unique minimizer (3.7) as long as $\rho > 0$. With the updated $\boldsymbol{\Theta}^{(k+1)}$ and $\boldsymbol{\Gamma}^{(k+1)}$ from **Steps 1**, **2**, we can easily update $\boldsymbol{\Lambda}^{(k)}$ to $\boldsymbol{\Lambda}^{(k+1)}$ through **Step 3**. The final output of WMVR-ADMM is a minimizer of $\mathcal{L}_\rho\big(\boldsymbol{\Theta}, \boldsymbol{\Gamma}^{(\mathcal{T}-1)}, \boldsymbol{\Lambda}^{(\mathcal{T}-1)}\big)$ in **Step 1**, where $\mathcal{T}$ denotes the last iteration index of the algorithm. Therefore, as long as all the non-zero singular values of $\boldsymbol{B}^{(\mathcal{T})}$ are distinct, then the results estimator from repeating **Steps 1-3** has an unique solution. The entire implementation is summarized in Algorithm 1 and named Weighted Multi-Variate-Regression-ADMM (WMVR-ADMM) algorithm. Note that WMVR-ADMM algorithm can be easily extended to trace regression model, which is a general model of multivariate linear regression model.[1] In order for the concise presentation of the paper, we defer the detailed descriptions of the extended algorithm to trace regression model in the Section G of Supplemental Material.

[1] Refer [5] for checking how to translate MVLR to trace regression model.

```
Input : A measurement pair $\big(\boldsymbol{X}, \boldsymbol{Y}\big)$, $\boldsymbol{\lambda}_n \geq 0$ and weights $0 \leq \omega_1 \leq \cdots \leq \omega_p$.
Initialization : $\boldsymbol{\Theta}^{(0)} = \boldsymbol{\Gamma}^{(0)} = \boldsymbol{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{d_1 \times d_2}$.
Repeat the following Steps :
    Step 1. Let $\boldsymbol{B}^{(\boldsymbol{k})} := -\boldsymbol{\Lambda}^{(k)} + \rho \cdot \boldsymbol{\Gamma}^{(k)}$.    $\boldsymbol{B}^{(\boldsymbol{k})} = \boldsymbol{U}^{\mathbf{B}} \boldsymbol{D}^{\mathbf{B}} \big(\boldsymbol{V}^{\mathbf{B}}\big)^{\top}$.
            Set $\mathcal{S}_{\boldsymbol{\lambda}_n \omega}\big(\boldsymbol{D}^{\mathbf{B}}\big) = \mathbf{diag}\Big\{ \max \Big\{ \frac{1}{\rho}\big(\sigma_j(\boldsymbol{B}^{(k)}) - \boldsymbol{\lambda}_n w_j\big), 0 \Big\}$ for
$j = 1, \ldots, p \Big\}$.
            $\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{U}^{\mathbf{B}} \mathcal{S}_{\boldsymbol{\lambda}_n \omega}\big(\boldsymbol{D}^{\mathbf{B}}\big) \big(\boldsymbol{V}^{\mathbf{B}}\big)^{\top}$
    Step 2. $\boldsymbol{\Gamma}^{(k+1)} = \big(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X} + \rho \cdot \mathcal{I}_{d_1 \times d_1}\big)^{-1} \big(\frac{1}{n}\boldsymbol{Y}^{\top}\boldsymbol{X} + \rho \cdot \boldsymbol{\Theta}^{(k+1)} + \boldsymbol{\Lambda}^{(k)}\big)$.
    Step 3. $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big)$.
Until $\|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\|_{\mathrm{F}} \leq 10^{-7}$ and $\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}} \leq 10^{-7}$.
Output : $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(k+1)}$.
```

Algorithm 1: ADMM for Weighted Multi-Variate Regression. (WMVR-ADMM)

The convergence of the proposed algorithm is shown in Theorem 3.2.2 with its proof given in Section B of Supplemental Material. The proof is motivated from [123] and [115].

Theorem 3.2.2 *Set $\rho > 2L_{\nabla g}$ with $L_{\nabla g} := \sigma_1\big(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\big)$. The sequence $\{(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})\}_{k \geq 1}$ from WMVR-ADMM converges globally to the unique stationary point of $\mathcal{L}_\rho(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda})$.*

The threshold for penalty parameter ρ can be computed from data. In the theorem, "globally" means regardless of where the initial point under non-convex landscape of (3.5). Theorem 3.2.2 is verified numerically in Section 3.5.1. The uniqueness claim is due from the fact that the augmented lagrangian function (3.5) is a Kurdyka-Lojasiewicz (KL) function. This is further elaborated in Section B of Supplemental Material with relevant references.

3.3 Statistical Properties of the Estimator

3.3.1 Statistical Properties of $\widehat{\boldsymbol{\Theta}}$ under the Orthogonal Design

We first study the convergent rate of the estimated eigenvalues under orthogonal design setting is studied, which sheds lights on understanding the role of weights on the estimation of singular values. The result is summarized in the following proposition.

Proposition 3.3.1 *Let* $\widehat{\boldsymbol{U}}^{LS}\widehat{\boldsymbol{D}}^{LS}\big(\widehat{\boldsymbol{V}}^{LS}\big)^\top$ *be SVD of the least-squares estimator* $\widehat{\boldsymbol{\Theta}}^{LS} := \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$. *Then, under the orthogonal design (i.e.,* $\boldsymbol{X}^\top\boldsymbol{X} = n\boldsymbol{I}_{d_1\times d_1}$*), SVD of the minimizer of* (3.2) *has the following closed-form solution:* $\widehat{\boldsymbol{\Theta}} := \widehat{\boldsymbol{U}}^{LS}\widehat{\boldsymbol{D}}\big(\widehat{\boldsymbol{V}}^{LS}\big)^\top$*, where the diagonal entry of* $\widehat{\boldsymbol{D}}$ *is:* $\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big) = \max\big(\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{LS}\big) - \boldsymbol{\lambda}_n\omega_j, 0\big)$ *for* $j = 1,\dots,p$. *Furthermore, suppose* $\boldsymbol{\lambda}_n = \sqrt{\frac{d_1+d_2}{n}}$. *Then, with probability at least* $1-2\exp(-(\sqrt{d_1}+\sqrt{d_2})^2/2)$*, we have,*

$$\Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big) - \sigma_j\big(\boldsymbol{\Theta}^\star\big)\Big| \leq \max\big(4\sigma, 2\omega_j\big)\cdot\sqrt{\frac{d_1+d_2}{n}}, \tag{3.8}$$

for j *such that* $\sigma_j\big(\boldsymbol{\Theta}^\star\big) > 0$. *With the same probability bound, we have,*

$$\Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big)\Big| \leq \min\big(2\sigma, \omega_j\big)\cdot\sqrt{\frac{d_1+d_2}{n}}, \tag{3.9}$$

for j *such that* $\sigma_j\big(\boldsymbol{\Theta}^\star\big) = 0$.

The proof of Proposition 3.3.1 can be found in Section C of Supplemental Material. Based on the closed-form solution of $\widehat{\boldsymbol{\Theta}}$ in Proposition 3.3.1, under the orthogonal design assumption, each estimated singular value has a form $\max\big(\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big) - \boldsymbol{\lambda}_n\omega_j, 0\big)$ for $j \in \{1,\dots,p\}$. Then, for the fixed $\boldsymbol{\lambda}_n$, it is easy to see that the large weights for small singular values of $\widehat{\boldsymbol{\Theta}}^{\text{LS}}$ can induce the sparsity among the singular values of $\widehat{\boldsymbol{\Theta}}$. Furthermore, the proposition states that with an appropriate choice of tuning parameter $\boldsymbol{\lambda}_n$, the singular values of the $\widehat{\boldsymbol{\Theta}}$ are consistently estimated. Bounds in (3.8) and (3.9) provide us with the guideline for the choices of weights. That is, for the set of indices of $\sigma_j\big(\boldsymbol{\Theta}^\star\big) > 0$, the corresponding weights ω_js need to be set lower than the twice of variance size of the measurement error σ, whereas, for the set of indices whose $\sigma_j\big(\boldsymbol{\Theta}^\star\big) = 0$, the corresponding weights can be set even higher than 2σ. This is consistent with our intuition that we need small weights for estimating non-zero singular values of $\boldsymbol{\Theta}^\star$, whereas large weights are required for the consistent estimation of zero singular values of $\boldsymbol{\Theta}^\star$.

3.3.2 Estimation Error under Random Design

We further study the estimation error under a random design assumption in the Frobenius norm (i.e., $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\|_{\mathrm{F}}^2$). Proving the main results requires extra technical assumptions: (I) A design matrix $\boldsymbol{X}$ is assumed to be random, whose rows are independently sampled from d_1-variate $\mathcal{N}(0, \mathbf{\Sigma})$ distribution for some positive definite covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d_1 \times d_1}$, and (II) The exact low rank assumption of $\mathbf{\Theta}^\star$ is relaxed to a nearly low-rank matrix by requiring that the $\{\sigma_j\big(\mathbf{\Theta}^\star\big)\}_{j=1}^p$ decays fast enough. Specifically, for a parameter $q \in [0, 1]$ and a radius $r^\star$, we assume that

$$\mathbf{\Theta}^\star \in \mathbb{B}_q(r^\star) := \bigg\{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \sum_{j=1}^{p} \big|\sigma_j\big(\mathbf{\Theta}^\star\big)\big|^q \leq r^\star\bigg\}.$$

Note that when $q = 0$, the set $\mathbb{B}_q(r^\star)$ becomes the set of matrices with rank at most $r^\star$, and (III) $\widehat{\mathbf{\Theta}}$ is a global minimizer of (3.2).

Additionally, we also need to define two extra technical terminology to understand more insights of the estimation errors: (I) restricted strong convexity of the cost function $\mathcal{L}_n(\mathbf{\Theta}) := \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Theta}\right\|_{\mathrm{F}}^2$ around $\Theta^\star$ and (II) the characterization of set where the associated error matrix $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star$ belongs. In high-dimensional setting where $n \ll d_1 d_2$, although the function $\mathcal{L}_n(\mathbf{\Theta})$ might be curved in some directions, there are $\big(d_1 d_2 - n\big)$ directions where it is flat up to second order. We hope that the associated error matrix $\widehat{\mathbf{\Delta}}$ lies in some directions $\mathcal{C} \subseteq \mathbb{R}^{d_1 \times d_2}$ where the $\mathcal{L}_n(\mathbf{\Theta})$ is curved. This notion is expressed as follows: for some positive constant $\kappa > 0$,

$$\mathcal{E}_n\big(\widehat{\mathbf{\Delta}}\big) \geq \kappa \|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2 \qquad \text{for all} \quad \widehat{\mathbf{\Delta}} \in \mathcal{C}, \tag{3.10}$$

where $\mathcal{E}_n\big(\widehat{\mathbf{\Delta}}\big)$ denotes the first order Taylor-expansion error of $\mathcal{L}_n(\cdot)$ around $\mathbf{\Theta}^\star$. In other words, we call $\mathcal{E}_n\big(\widehat{\mathbf{\Delta}}\big)$ succeeds *"restricted strong convexity"* (RSC) over the set $\mathcal{C}$ if there exists $\kappa > 0$. Fortunately, we can prove that the RSC condition indeed holds with $\kappa =$

$\frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}$ in high probability over $\mathbb{R}^{d_1 \times d_2}$ under multivariate regression model with Gaussian ensemble [2], where $\sigma_{\min}(\boldsymbol{\Sigma})$ denotes a minimum eigenvalue of $\boldsymbol{\Sigma}$.

Before we formally state the lemma that characterizes the set $\mathcal{C}$, let us introduce relevant notation. Denote $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ as the left and right singular matrices of $\boldsymbol{\Theta}^\star$. The $\mathcal{M}_r\big(\mathcal{U}, \mathcal{V}\big)$ (resp. $\overline{\mathcal{M}}_r^\perp\big(\mathcal{U}, \mathcal{V}\big)$) corresponds to subspace of matrices with non-zero left and right singular vectors associated with the first r (resp. remaining $(p-r)$) columns of $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$. That is, for any given integer $r \le p$, we have

$$\mathcal{M}_r\big(\mathcal{U}, \mathcal{V}\big) = \big\{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \textbf{colspan}(\boldsymbol{\Theta}) \subseteq \mathcal{U}, \quad \textbf{rowspan}(\boldsymbol{\Theta}) \subseteq \mathcal{V}\big\}$$
$$\overline{\mathcal{M}}_r^\perp\big(\mathcal{U}, \mathcal{V}\big) = \big\{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \textbf{colspan}(\boldsymbol{\Theta}) \subseteq \mathcal{U}^\perp, \quad \textbf{rowspan}(\boldsymbol{\Theta}) \subseteq \mathcal{V}^\perp\big\}.$$

Then, $\mathcal{U}$ and $\mathcal{V}$ are the r-dimensional subspaces of vectors from the first r columns of matrices $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$. Moreover, $\mathcal{U}^\perp$ and $\mathcal{V}^\perp$ denote the subspaces orthogonal to $\mathcal{U}$ and $\mathcal{V}$, respectively, and $\textbf{colspan}(\boldsymbol{\Theta})$ and $\textbf{rowspan}(\boldsymbol{\Theta})$ denote the column space and row space of $\boldsymbol{\Theta}$. Hereafter, we will omit $\mathcal{U}$ and $\mathcal{V}$ from the notations, if they are clear from the context. The notation can be used to characterize the set $\mathcal{C}$ as shown in the lemma below:

Lemma 3.3.2 *Suppose $\widehat{\boldsymbol{\Theta}}$ is an global minimizer of the* (3.2) *obtained from WMVR-ADMM, with the associated matrix $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star$. Set the weights $\frac{1}{2} < \omega_1 \le \cdots \le \omega_p$ and suppose regularization parameter is chosen such that $\boldsymbol{\lambda}_n \ge \frac{2}{n}\left\|\boldsymbol{X}^\top \boldsymbol{E}\right\|_{op}$. Let $\|\cdot\|_\star := \sum_{j=1}^p \sigma_j(\cdot)$. Then, for a positive integer $r \le p$, we have*

$$\mathcal{C}\big(\omega; r; \delta\big) := \left\{\widehat{\boldsymbol{\Delta}} \in \mathbb{R}^{d_1 \times d_2} : \|\widehat{\boldsymbol{\Delta}}''\|_\star \le \frac{2w_p}{w_1 - \frac{1}{2}} \sum_{j=r+1}^{p} \sigma_j\big(\boldsymbol{\Theta}^\star\big) + \frac{2w_p - w_1 + \frac{1}{2}}{w_1 - \frac{1}{2}} \cdot \|\widehat{\boldsymbol{\Delta}}'\|_\star\right\}, \tag{3.11}$$

where $\widehat{\boldsymbol{\Delta}}'' \in \Pi_{\overline{\mathcal{M}}_r^\perp}\big(\widehat{\boldsymbol{\Delta}}\big)$ and $\widehat{\boldsymbol{\Delta}}' = \widehat{\boldsymbol{\Delta}} - \widehat{\boldsymbol{\Delta}}''$. Let $\Pi_{\overline{\mathcal{M}}_r^\perp}$ denote the projection operator onto the subspace $\overline{\mathcal{M}}_r^\perp$.

[2] See Lemma 2 in [5] and Section E of Supplemental Material.

A detailed proof of Lemma 3.3.2 is deferred in Section D of Supplemental Material. The lemma shows that the subset $\mathcal{C}$ corresponds to the matrices $\widehat{\mathbf{\Delta}}$ for which the quantity $\|\widehat{\mathbf{\Delta}}''\|_\star$ is relatively small compared to the weighted sum of $\|\widehat{\mathbf{\Delta}}'\|_\star$ and $(p-r)$ remaining singular values of $\mathbf{\Theta}^\star$. The weights put in $\|\widehat{\mathbf{\Delta}}'\|_\star$ and $\sum_{j=r+1}^{p} \sigma_j\big(\mathbf{\Theta}^\star\big)$ are functions of a pair (ω_1, ω_p), and this pair characterizes size of the subset $\mathcal{C}$. We restrict the case $\omega_1 > \frac{1}{2}$ for a technical reason. The closer ω_1 gets to $\frac{1}{2}$ and the larger ω_p we have, the bigger the size of $\mathcal{C}$ becomes. Also, Lemma 3.3.2 shows that plugging in $\omega_1 = \cdots = \omega_p = 1$ recovers one of constraints that are used to define the set in Lemma 1 of [5]. A notable difference between the set in (3.11) and the set defined in [5] is the existence of the constraint, $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} \geq \delta$, where $\delta > 0$ is a tolerance parameter. This constraint is used to eliminate the open ball that is contained within the set $\mathcal{C}$, to ensure RSC condition holds over $\mathcal{C}$, even when $\mathcal{E}_n\big(\widehat{\mathbf{\Delta}}\big)$ fails strong convexity in a global sense. Nonetheless, as previously mentioned, since strong convexity o $\mathcal{L}_n(\mathbf{\Theta})$ holds globally in our problem setting, the constraint $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} \geq \delta$ is not required.

With the RSC condition and Lemma 3.3.2, we can further show that the estimation error converges to 0 at a minimax rate, whose proof is given in Section E of Supplemental Material.

Theorem 3.3.3 *The regularization parameter is chosen such that* $\boldsymbol{\lambda}_n = 10\sigma\|\mathbf{\Sigma}\|_{op}\sqrt{\frac{d_1+d_2}{n}}$ *and weights are set as* $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$. *Define* $\mathcal{W} := \frac{w_p\left(2w_p - w_1 + \frac{1}{2}\right)}{w_1 - \frac{1}{2}}$. *Then, there are universal constants* $\{c_i, i = 1, 2, 3\}$ *such that any minimizer* $\widehat{\mathbf{\Theta}}$ *of* (3.2) *satisfies the following bound:*

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\right\|_F^2 \leq c_1 \mathcal{W}^2 \left(\frac{\sigma^2\|\mathbf{\Sigma}\|_{op}^2}{\sigma_{min}^2(\mathbf{\Sigma})}\right)^{1-q/2} \cdot r^\star \left(\frac{d_1+d_2}{n}\right)^{1-q/2}. \tag{3.12}$$

with probability at least $1 - c_2\exp(-c_3(d_1+d_2))$.

Here, $\|\mathbf{\Sigma}\|_{\mathrm{op}}$ denotes the spectral norm of the matrix $\mathbf{\Sigma}$. Notably, when $\mathbf{\Theta}^\star \in \mathbb{B}_q(r^\star)$ is an exact rank $r^\star$ matrix (i.e., $q = 0$) and $\mathbf{\Sigma} = \mathcal{I}_{d_1 \times d_1}$, convergence rate of the estimation error becomes $\mathcal{O}\big(\mathcal{W}^2\frac{\sigma^2 r^\star(d_1+d_2)}{n}\big)$ up to a constant factor. The quantity $r^\star(d_1+d_2)$ counts the degrees of freedom in the model, and the rate is known to be minimax optimal for estimating a $d_1 \times d_2$ matrix with rank $r^\star$. See [5, 120, 124]. It is worth noting that the information on weights is solely encoded in factor $\mathcal{W}$. This factor allows a natural comparison of estimation rates between SNN and WNN, and we defer the discussion on this comparison to Section 3.6.

3.4 Data-driven Model Selections

3.4.1 Surrogate Estimator $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$ for the GCV statistic

A surrogate estimator $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$ is developed for approximating the estimator $\widehat{\boldsymbol{\Theta}}$. From Proposition 3.3.1, we know $\widehat{\boldsymbol{\Theta}} := \widehat{\boldsymbol{U}}^{\text{LS}}\widehat{\boldsymbol{D}}\big(\widehat{\boldsymbol{V}}^{\text{LS}}\big)^\top$, where the diagonal entry of $\widehat{\boldsymbol{D}}$ is: $\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big) = \max\big(\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big) - \boldsymbol{\lambda}_n\omega_j, 0\big)$ for $j \in \{1,\dots,p\}$. Hereafter, for the convenience of notation, we denote $\widehat{d}_j := \sigma_j\big(\widehat{\boldsymbol{\Theta}}\big)$, for $j \in \{1,\dots,p\}$. Then, we define the following matrix $\boldsymbol{K} \in \mathbb{R}^{d_1\times d_1}$:

$$\boldsymbol{K} := \widehat{\boldsymbol{U}}^{\text{LS}}\widehat{\boldsymbol{D}}^{\text{K}}\big(\widehat{\boldsymbol{U}}^{\text{LS}}\big)^\top := \sum_{j=1}^{\widehat{r}} \frac{\omega_j}{\widehat{d}_j}\widehat{\boldsymbol{U}}_j^{\text{LS}}\big(\widehat{\boldsymbol{U}}_j^{\text{LS}}\big)^\top, \tag{3.13}$$

where $\widehat{r}$ denotes the cardinality of a set $\{j : \widehat{d}_j > 0\}$. We provide the following proposition, whose proof is deferred in Section F of Supplemental Material.

Proposition 3.4.1 *For a fixed $\boldsymbol{K}$ that is defined in* (3.13)*, we denote $\widehat{\boldsymbol{\Theta}}^{SR}$ as the minimizer of the following surrogate optimization problem :*

$$\widehat{\boldsymbol{\Theta}}^{SR} := \operatorname*{argmin}_{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\Theta\|_F^2 + \frac{\boldsymbol{\lambda}_n}{2}\boldsymbol{tr}\big(\boldsymbol{\Theta}^\top\boldsymbol{K}\boldsymbol{\Theta}\big)\right\}. \tag{3.14}$$

Then, under orthogonal design (i.e., $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_{d_1 \times d_1}$), $\widehat{\boldsymbol{\Theta}}^{SR} = \widehat{\boldsymbol{U}}^{LS}\widehat{\boldsymbol{D}}^{SR}\left(\widehat{\boldsymbol{V}}^{LS}\right)^\top$, where

$$\widehat{\boldsymbol{D}}^{SR}_{jj} = \begin{cases} \widehat{d}_j & j = 1, 2, \ldots, \widehat{r}, \\ \sigma_j\left(\widehat{\boldsymbol{\Theta}}^{LS}\right) & j = \widehat{r}+1, \ldots, p. \end{cases}$$

Note that as long as $\widehat{\boldsymbol{\Theta}}^{\text{LS}}$ is a full-rank, $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$ is a full-rank matrix whose first $\widehat{r}$ singular values are identical to those of $\widehat{\boldsymbol{\Theta}}$, and remaining $(p - \widehat{r})$ singular values are equal to the corresponding singular values of $\widehat{\boldsymbol{\Theta}}^{\text{LS}}$. Although the result of Proposition 3.4.1 is stated under orthogonal design assumption, we also empirically demonstrate that the same results hold under non-orthogonal design in Figure 3.2. Specifically, under the same experimental setting of Figure 3.1, $\widehat{\boldsymbol{\Theta}}$ is a minimizer of (3.2) obtained via WMVR-ADMM with the weight updating scheme, which will be defined later. In this experiment, the minimum absolute off-diagonal entry of $\boldsymbol{X}^\top \boldsymbol{X}$ is 0.00157, which implies $\boldsymbol{X}$ is a non-orthogonal design. The result in Figure 3.2 is consistent with the statement in Proposition 3.4.1.

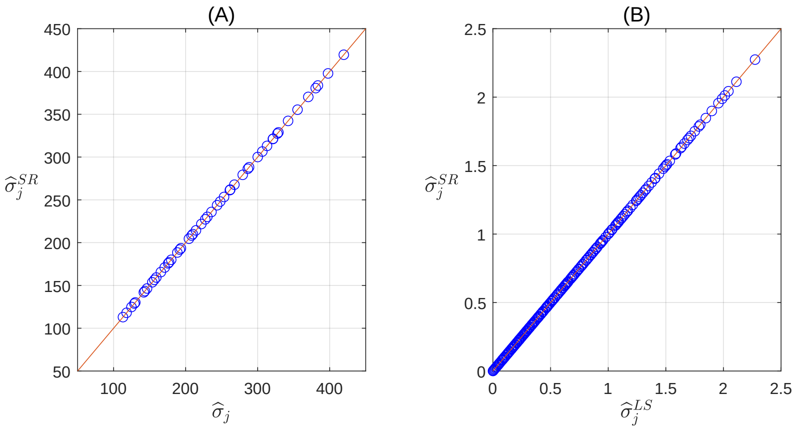


Figure 3.2: Under non-orthogonal $\boldsymbol{X}$, panel (A) displays the plot of the first 50 singular values of $\widehat{\boldsymbol{\Theta}}$ versus $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$. Panel (B) exhibits the plot of the remaining 200 singular values of $\widehat{\boldsymbol{\Theta}}^{\text{LS}}$ versus $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$.

3.4.2 GCV Statistic and Weight Updates

The closed form solution of the surrogate estimator (3.4.1) can be used to tune the hyperparameters $\boldsymbol{\lambda_n}$ and weights $\omega_1, \cdots, \omega_p$ in (3.2) with (3.3). We divide the process for tuning the parameters into two procedures. In the first procedure, we propose the following iterative algorithm that alternates between estimating $\boldsymbol{\Theta}^\star$ and updating weights.

(I) Set the iteration count ℓ to 0 and weights $\omega_1^{(0)} = \cdots = \omega_p^{(0)} = 1$.

(II) For the fixed $\boldsymbol{\lambda}_n$, solve (3.2) via WMVR-ADMM with the weights $\{\omega_j^{(\ell)}\}_{j=1}^p$, and denote the solution as $\widehat{\boldsymbol{\Theta}}^{(\ell)}$.

(III) Update weights : for each $j \in \{1, \ldots, p\}$,

$$\omega_j^{(\ell+1)} = \frac{1}{\sigma_j(\widehat{\boldsymbol{\Theta}}^{(\ell)}) + \epsilon}. \tag{3.15}$$

(IV) Terminate until convergence or when ℓ attains a pre-specified maximum number of iterations. Otherwise, increment ℓ and go to step (II).

The introduced parameter $\epsilon > 0$ in step (III) guarantees that, for any $j \in \{1, \ldots, p\}$, the $(\ell + 1)^{\text{th}}$ updated weight $\omega_j^{(\ell+1)}$ is computable, even when $\sigma_j(\widehat{\boldsymbol{\Theta}}^{(\ell)}) = 0$. The recovery process of $\boldsymbol{\Theta}^\star$ is reasonably robust to the choice of ϵ, and we set $\epsilon = 10^{-3}$ hereafter. The choice $\epsilon = 10^{-3}$ may appear a little bit arbitrary, but works well in practice. The resulting estimator from the first procedure is denoted by $\mathrm{W}(\boldsymbol{\lambda_n})$. We use a superscript W in $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\boldsymbol{\lambda_n})$ to indicate that the estimator is a converged solution from weight updating procedure introduced above, and use $\boldsymbol{\lambda_n}$ to denote the estimator is obtained from a fixed hyper-tuning parameter $\boldsymbol{\lambda_n}$.

The second procedure is designed for choosing parameter $\boldsymbol{\lambda_n}$. We develop a GCV type of statistic [125], which is more computationally efficient than the ordinary CV (Cross Validation) method, especially in large scale problems. This can be done by using the surrogate estimator $\widehat{\boldsymbol{\Theta}}^{\text{SR}}$ for approximating the degrees of freedom of $\widehat{\boldsymbol{\Theta}}^{\text{W}}(\boldsymbol{\lambda_n})$ from the first procedure. That is, given $\widehat{\boldsymbol{\Theta}}^{\text{W}}(\boldsymbol{\lambda_n})$, we can construct $\boldsymbol{K}^{\text{W}}$ from (3.13). Then, by proposition 3.4.1, we can define the projection matrix (hat matrix) for the regression problem (3.14) by $\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\lambda}_n n\boldsymbol{K}^{\text{W}}\right)^{-1}\boldsymbol{X}^\top$ and approximate the degrees of freedom of $\widehat{\boldsymbol{\Theta}}^{\text{W}}(\boldsymbol{\lambda_n})$ as

$$\text{df}(\boldsymbol{\lambda}_n) \approx d_2\mathbf{tr}\big(\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\lambda}_n n\boldsymbol{K}^{\text{W}}\right)^{-1}\boldsymbol{X}^\top\big). \tag{3.16}$$

Thus, the GCV score for $\widehat{\boldsymbol{\Theta}}^{\text{W}}(\lambda_n)$ is given by

$$\text{GCV}(\boldsymbol{\lambda}_n) := \frac{\mathbf{tr}\big(\big(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{W}}(\lambda_n)\big)\big(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{W}}(\lambda_n)\big)^\top\big)}{d_1 d_2 - \text{df}(\boldsymbol{\lambda}_n)}, \tag{3.17}$$

and the optimal $\boldsymbol{\lambda}_n^\star$ for which GCV($\boldsymbol{\lambda}_n$) is obtained by minimizing the GCV score (3.17) over the search range $\boldsymbol{\lambda}_n \in [0, \mathcal{T}]$.

3.5 Numerical Experiments

3.5.1 Convergence of WMVR-ADMM

To demonstrate the convergence of the proposed algorithm WMVR-ADMM through some simulation studies, we note that the convergences of the algorithm can be observed through the following two quantities:

1. For checking the Primal residual convergence (i.e., $\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)} \to 0$ as $k \to \infty$), and $\boldsymbol{\Gamma}^{(k)}$ convergence (i.e., $\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)} \to 0$ as $k \to \infty$), we consider

$$R^{(k)} := \|\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}\|_{\text{F}}^2 + \|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\|_{\text{F}}^2.$$

2. For checking the objective convergence, we consider

$$O^{(k)} := \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}^{(k)}\right\|_{\text{F}}^2 + \boldsymbol{\lambda}_n\left\|\boldsymbol{\Theta}^{(k)}\right\|_{\boldsymbol{\omega},\star}.$$

The coefficient matrix of the simulation setting is generated from $\boldsymbol{\Theta}^\star = \boldsymbol{A}\boldsymbol{B}^\top \in \mathbb{R}^{250\times 50}$, where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{250\times 50}$ with each entry from $\mathcal{N}(0,1)$. Each entry of $\boldsymbol{X} \in \mathbb{R}^{n\times d_1}$ is sampled from $\mathcal{N}(0,1)$, and the noise matrix $\boldsymbol{E}$ are independently chosen from another $\mathcal{N}(0, \mathcal{I}_{d_2\times d_2})$. Variance parameter σ is set as 1 with hyper-tuning parameter $\boldsymbol{\lambda}_n$ is set as $5\sqrt{\frac{d_1+d_2}{n}}$, where $d_1 = d_2 = 250$. Under the simulation setting with $n = 250$, we vary the initialized tuple matrix values $(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \boldsymbol{\Lambda}^{(0)})$ of WMVR-ADMM in Algorithm 1. Entries of three matrices are sampled from $\mathcal{N}(0, \nu^2)$, where $\nu = \{0, 0.1, 0.2, 0.5, 1, 1.5\}$. Weights $\{\omega_j\}_{j=1}^p$ are updated once, and with the updated weights, $R^{(k)}$ and $O^{(k)}$ are calculated with the same data set $(\boldsymbol{X}, \boldsymbol{Y})$ over all simulation scenarios. The resulting $R^{(k)}$ and $O^{(k)}$ values are demonstrated in Figure 3.3, and the figure shows that both $R^{(k)}$ and $O^{(k)}$ converge to 0 as k increases, regardless of the initializations of algorithm. This observation is consistent with the claims in Theorem 3.2.2, and implies that the the converged solutions from WMVR-ADMM have the same objective value on the non-convex landscape of problem (3.2).

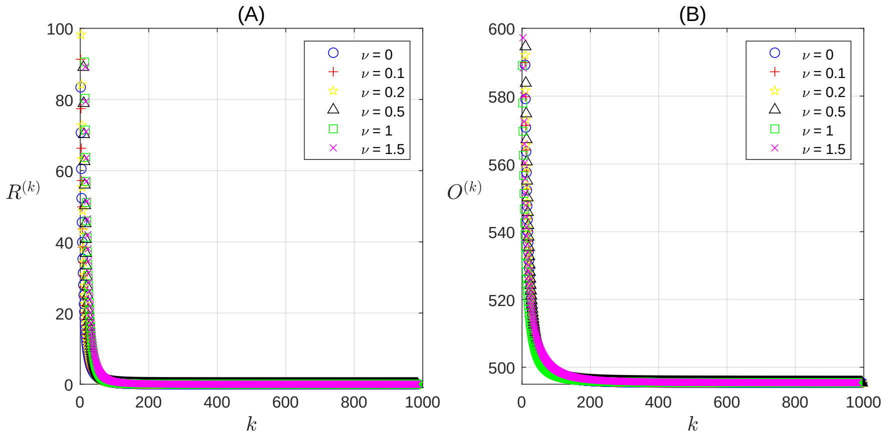


Figure 3.3: Convergences of $R^{(k)}$ (panel (A)) and $O^{(k)}$ (panel (B)) over the algorithm iteration index k. Regardless of random initializations, $R^{(k)}$ and $O^{(k)}$ converge to 0 and to a same objective function value, respectively.

3.5.2 Comparisons of Estimation Error with Other Methods

The simulation setting is following the setting in previous subsection with is $d_1 = 25$, $d_2 = 25$, and the value $r^\star$ is the rank of the ground truth matrix and is chosen to be $2, 5, 8$, and 11. The sample sizes n are set to be 30, 300, and 3000, and the simulation is repeated 100 times. The estimation errors of the proposed method are recorded in terms of the root mean squared errors (RMSE) between the estimated coefficient matrix and the ground-truth matrix for each simulation. The results are compared with those from SNN and ANN methods ([16]). Recall SNN estimator is equivalent to the model (3.1) with $\{\omega_j\}_{j=1}^{p}$. As for ANN estimator, let $\widehat{\boldsymbol{U}}^{\text{XLS}}\widehat{\boldsymbol{D}}^{\text{XLS}}\big(\widehat{\boldsymbol{V}}^{\text{XLS}}\big)^\top$ be SVD of the matrix $\boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{LS}} := \boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$. Then, the estimator from Corollary 1 in [16] has a closed-form solution as:

$$\widehat{\boldsymbol{\Theta}}^{\text{ANN}} = \widehat{\boldsymbol{\Theta}}^{\text{LS}}\widehat{\boldsymbol{V}}^{\text{XLS}}\big(\widehat{\boldsymbol{D}}^{\text{XLS}}\big)^{-1}\mathcal{S}_{\boldsymbol{\lambda}_n\omega}\big(\widehat{\boldsymbol{D}}^{\text{XLS}}\big)\big(\widehat{\boldsymbol{V}}^{\text{XLS}}\big)^\top, \tag{3.18}$$

where $\mathcal{S}_{\boldsymbol{\lambda}_n\omega}\big(\widehat{\boldsymbol{D}}^{\text{XLS}}\big) = \text{diag}\Bigg\{\max\Bigg\{\sigma_j(\widehat{\boldsymbol{D}}^{\text{XLS}}) - \boldsymbol{\lambda}_n w_j, 0\Bigg\} \text{ for } j = 1, \ldots, p\Bigg\}$. The three methods WMVR-ADMM, SNN, and ANN include parameters needed to be tuned, and we use the GCV tuning method from Section 4 for the tuning.

All results are demonstrated in Figure 3.4. The first row of Figure 3.4 shows the performance of all methods under the case whose ground-truth matrix is rank 2 ($r^\star = 2$), and we observe that the averages of RMSEs from the WMVR-ADMM method are smaller than those from other methods for all sample size cases. The second to fourth rows of Figures 3.4 presents the RMSE results from rank $r^\star = 5, 8$, and 11 cases, and the proposed methods are still better than other methods in almost all cases. Additionally, the panels in Figure 3.4 demonstrate that the RMSEs from the proposed estimator decrease to 0 as the sample size increases. This shows the consistency property of the proposed estimator empirically.

To show the effectiveness of the proposed weight updating scheme in Section 3.4, we

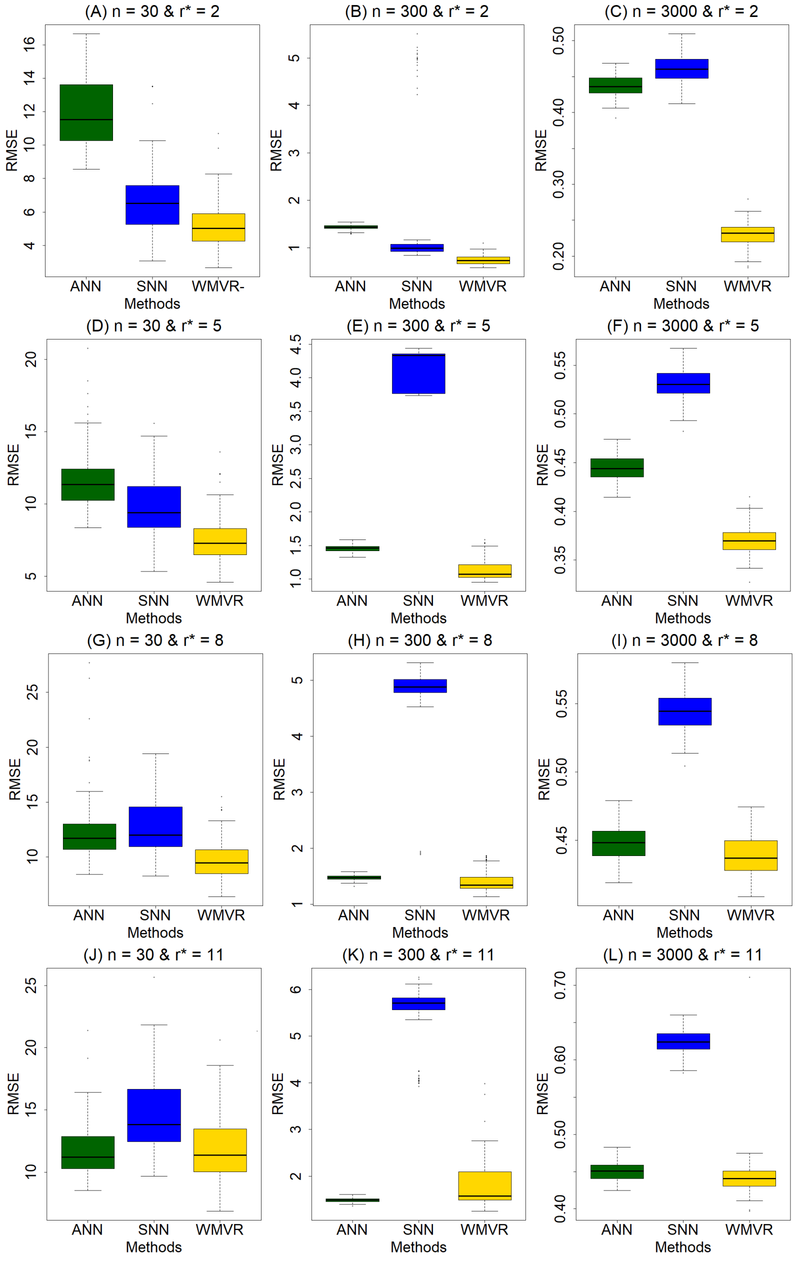


Figure 3.4: The plots demonstrates the comparisons of estimation errors in terms of RMSEs from the proposed method with ANN and SNN methods under different simulation settings. The three figures in the first row (A) $\sim$ (C) are the comparison results from sample size 30, 300, and 3000, respectively, under the true rank $r^\star = 2$. Analogously, Figures (D) $\sim$ (F) (second row) are the results from $r^\star = 5$, Figures (G) $\sim$ (I) (third row) are the results from $r^\star = 8$, and Figures (J) $\sim$ (L) (fourth row) are the results from $r^\star = 11$.

compare the weight setting suggested in [16] with our method. For comparison, we revisit the synthetic setting used in Figure 3.1. Let $\widehat{\mathbf{\Theta}}^{(1)}$ be the SNN estimator, and denote $\boldsymbol{\omega}^{\text{WNN}}$ and $\boldsymbol{\omega}^{\text{ANN}}$ be the weight settings introduced in (3.15) and [16], which means

$$\omega_j^{\text{WNN}} = \left(\sigma_j(\widehat{\mathbf{\Theta}}^{(1)}) + 10^{-3}\right)^{-1}, \qquad \omega_j^{\text{ANN}} = \sigma_j\left(\boldsymbol{X}\widehat{\mathbf{\Theta}}^{\text{LS}}\right)^{-2}, \qquad j = 1, \ldots, 250. \quad (3.19)$$

The two types of weights is implemented in Algorithm 1 to evaluate RMSEs. The results are recorded in Figure 3.5: Panel (A) shows the two sequences of averaged weights $\{\omega_j\}_{j=1}^{250}$ in (3.19) used for the estimation in logarithmic scale, and panel (B) exhibits 100 RMSEs with the respective weight scheme. While the difference of the first 50 weights between two weight schemes is negligible, the effect of WNN-weight scheme is dramatized for penalizing the remaining 200 singular values in comparison to ANN-weight scheme, and this results in lower RMSEs in panel (B).

3.5.3 Application to A Real Dataset

The proposed method is applied to an important application in this section. The application is about a study of Polycyclic Aromatic Hydrocarbons (PAHs) from Section 2.2.2 of [126].

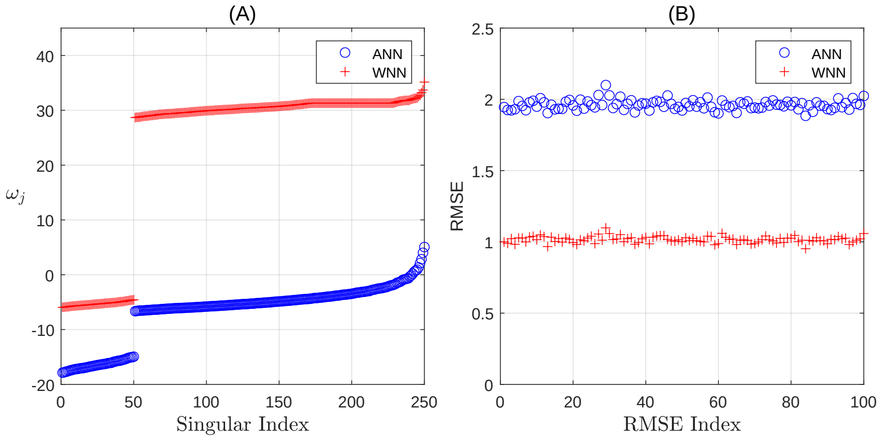


Figure 3.5: Two sequences of weights in (3.19) used for the estimation (panel (A)) and the resulting RMSEs (panel (B)). Low RMSEs of WNN weights arise from the high penalization on the remaining 200 singular values, when they are compared with RMSEs of ANN weights.

PAHs are ubiquitous environmental contaminants generated primarily during the incomplete combustion of some organic substances, such as coal, oil, rubbish, and wood. They are linked with the causes of tumors and their effects on reproduction. PAHs are widely used in industry or medicines to make dyes, plastics, and pesticides.

The dataset includes 10 PAHs, which is pyrene (Py), acenaphthene (Ace), anthracene (Anth), acenaphthylene (Acy), chrysene (Chry), benzanthracene (Benz), fluoranthene (Fluora), fluorene (Fluore), naphthalene (Nap), and phenanthracene (Phen), and 25 complex mixtures of certain concentrations (with unit milligrams per liter) of these PAHs were recorded, which indicates $n = 25$ and $d_1 = 10$ in model (3.1). The mean and range values of these mixtures of certain concentrations are plotted in Panel (A) of Figure 3.6. From each of these mixtures, an electronic absorption spectrum is computed, The spectrum are digitized at 5 nm intervals 27 wavelength channels from 220 nm to 350 nm, as shown in in Panel (B) of Figure 3.6. This means there are 27 columns for $\boldsymbol{X}_2$ in model (3.1) ($d_2 = 27$). More details about the dataset can be found in Section 5.1.2 of [127] and Section 2.2.2 of [126].

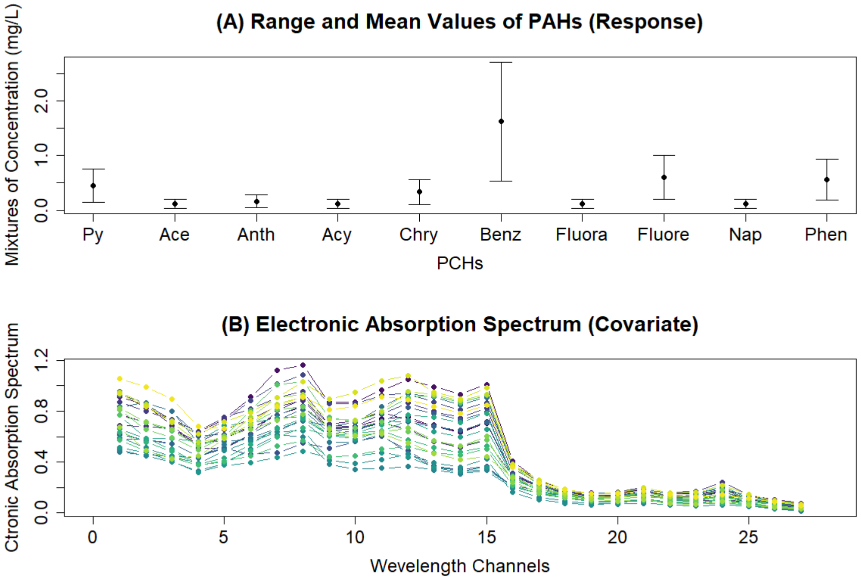


Figure 3.6: Demonstration of the mixture components of the PAHs ($\boldsymbol{Y}$) and the electronic absorption spectrum of the 25 samples ($\boldsymbol{X}$)

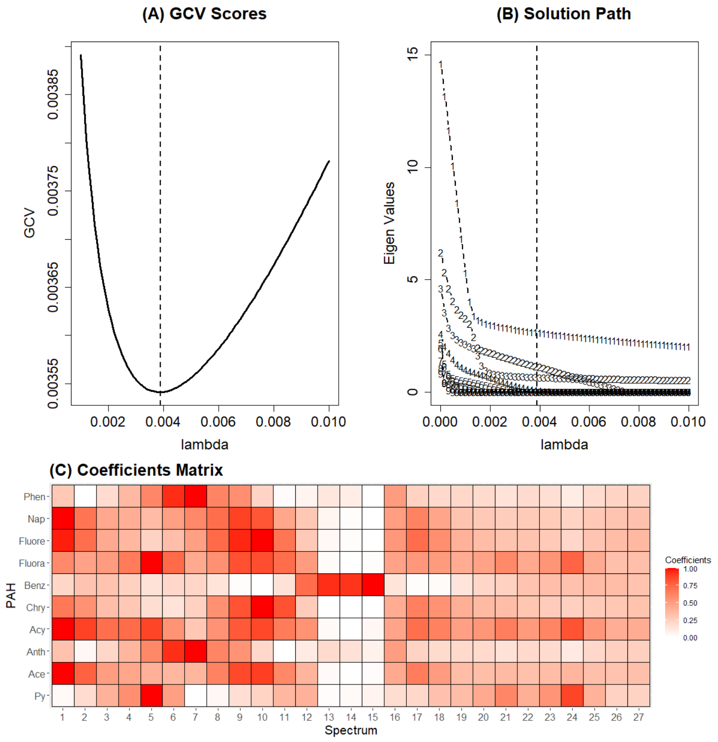


Figure 3.7: (A) GCV Score Versus Tuning Parameters $\boldsymbol{\lambda}$, (B) Solution Path, (C) Estimated Coefficient Matrix.

We are mainly interested in using WMVR-ADMM to understand the association between the concentrations from PAHs (Figure 3.6 (A)) and the electronic absorption spectrum (Figure 3.6 (B)) through model (3.1). The method is conducted by following Algorithm 1, and the optimal tuning parameter $\boldsymbol{\lambda}_n$ and weights $\boldsymbol{w}$ are selected by the proposed GCV criterion described in Section 4. The resulting GCV scores are plotted in Figure 3.7 (A) with respect to value $\boldsymbol{\lambda}_n$, showing the selected $\boldsymbol{\lambda}_n$ is around 0.039. The estimated eigenvalues with respect to $\boldsymbol{\lambda}_n$ are plotted in Figure 3.7 (B), and under the optimal $\boldsymbol{\lambda}_n$ and weights from the GCV criterion, the estimated coefficient matrix is rank 5. The estimated coefficients are demonstrated in a heatmap as shown in Figure 3.7 (C). The figure shows that for each PAH, only a few important channels can be used to determine the concentrations because only some coefficients are relatively large. Additionally, these larger coefficients are usually from smaller column numbers in the heatmap. Thus,this shows the channels with smaller wavelengths are more important than larger wavelength channels.

3.6 Conclusion and Discussion

We propose an ADMM-based method for solving the multivariate regression problem with WNN penalty. Under non-decreasing order of weights, the WNN is a non-convex function, and induce non-convexity of WNN penalized least-squares problem in (3.1) over the parameter space. The provided algorithm is shown to converge uniquely to one of stationary points of augmented Lagrangian function. The statistical properties of the estimator are investigated under orthogonal design, providing some insights on the choices of weights for the estimation. Furthermore, the minimax convergence rate of the estimation error is derived under random Gaussian design setting. In simulation studies, we demonstrate followings: (I) under random initializations, solutions of (3.2) via WMVR-ADMM algorithm converge to a certain estimator whose objective values are same (II) the WNN method outperforms SNN [118] and ANN [16] under synthetic settings, (III) the effect of our suggested weight updating scheme is verified through the comparison with the weight setting by [16]. Lastly the application to the real data set shows the effectiveness of our method. Nonetheless, there are several remaining open questions which require further investigations in the future. We summarize them as follows.

1. A question on whether the non-convex ADMM can achieve the global minimizer of (3.2) is a well-known open question. Although empirical results on the convergence of WMVR-ADMM are provided in Section 3.5, they still cannot verify the converged solution is a global minimizer of (3.2). We leave both empirical and theoretical justifications on this issue as important open problems. Under SNN setting, it is proved that there exists a primal-dual pair of (3.2) which satisfies the strong duality [128]. Therefore, the existence of saddle point on $\mathcal{L}_0$ can be ensured, so that the global minimizer of (3.2) can be proved through the classical techniques in [37].

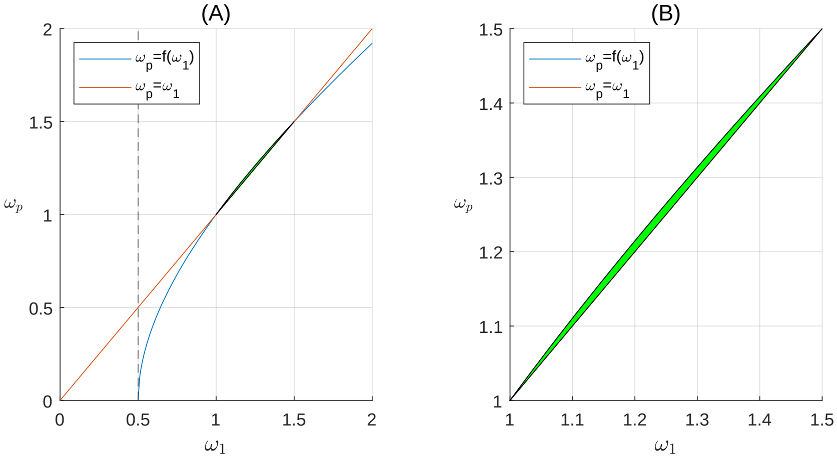


Figure 3.8: Panel (A) exhibits the intersected region of $\mathcal{W} \leq 3$ and $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$. Panel (B) magnifies the intersected region on grid $(\omega_1, \omega_p) \in [1, 1.5] \times [1, 1.5]$.

Nonetheless, we need further investigation whether these conditions can be used under our WNN setting with non-decreasing weights.

2. As previously mentioned in the remark of Theorem 3.3.3, $\mathcal{W} := w_p\big(2w_p - w_1 + \frac{1}{2}\big)/(w_1 - \frac{1}{2})$ is a sole factor that accounts for the effects of weights in the convergence rate of (3.12). This result naturally leads us to ask the question; "Under which pair of (ω_1, ω_p), does the estimator from WNN have a faster convergence rate than the one from SNN?". Under the same choices of tuning parameter $\boldsymbol{\lambda}_n$, a naive way for the comparison is to plug $\omega_1 = \omega_p = 1$ in $\mathcal{W}$. That is, we want to find a pair of (ω_1, ω_p) for which $\mathcal{W} \leq 3$ and $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$. The intersected region is illustrated in Figure 3.8. From our empirical experiences, the region of (ω_1, ω_p), for which WNN is superior than SNN in terms of estimation, is much larger than it is presented in Figure 3.8. This problem arises from the tightness of the subset $\mathcal{C}$ we derive in Lemma 3.3.2. In order to avoid this problem, we suspect that the different approach from using RSC condition of cost function is needed. A paper [129], recently appeared on arXiv, introduces a technique which takes the advantage of controlling the covering number of projection operators corresponding to the subspaces spanned by the design. They consider a problem of solving nuclear norm penal-

ized least squares problem, and their technique is independent from RSC condition. It would be an interesting open problem if their technique can be employed in our problem for obtaining a bigger intersected region than that in Figure 3.8.

CHAPTER 4
A NON-PARAMETRIC REGRESSION VIEWPOINT : GENERALIZATION OF OVERPARAMETRIZED DEEP RELU NETWORK UNDER NOISY OBSERVATIONS

4.1 Introduction

Over the past few years, Neural Tangent Kernel (NTK) [130, 131, 132, 133] has been one of the most seminal discoveries in the theory of neural network. The underpinning idea of the NTK-type theory comes from the observation that in a wide-enough neural net, model parameters updated by gradient descent (GD) stay close to their initializations during the training, so that the dynamics of the networks can be approximated by the first-order Taylor expansion with respect to its parameters at initialization. The linearization of learning dynamics on neural networks has been helpful in showing the linear convergence of the training error on both overparametrized shallow [134, 135] and deep neural networks [136, 137, 138], as well as the characterizations of generalization error on both models [139, 140]. These findings clearly lead to the equivalence between learning dynamics of neural networks and the kernel methods in reproducing kernel Hilbert spaces (RKHS) associated with NTK. [1] Specifically, [139] provided the $\mathcal{O}(n^{-1/2})$ generalization bound of shallow neural network, where n denotes the training sample size.

Recently, in the context of nonparametric regression, two papers, [141] and [142], showed that neural network can obtain the convergence rate faster than $\mathcal{O}(n^{-1/2})$ by specifying the complexities of target function and hypothesis space. Specifically, [141] showed that the shallow neural network with smoothly approximated ReLU (swish, see [143]) activa-

[1] Henceforth, we denote $\mathcal{H}_1^{\mathbf{NTK}}$ and $\mathcal{H}_L^{\mathbf{NTK}}$ as RKHSs induced from NTK of shallow $L = 1$ and deep neural networks $L \geq 2$ with ReLU activations, respecitvely.

tion trained via ℓ_2-regularized averaged stochastic gradient descent (SGD) can recover the target function from RKHSs induced from NTK with swish activation. Similarly, [142] showed that a shallow neural network with ReLU activation trained via ℓ_2-regularized GD can generalize well, when the target function (i.e., $f_\rho^\star$) is from $\mathcal{H}_1^{\textbf{NTK}}$. Notably, the rate that the papers [141] and [142] obtained is minimax optimal, meaning that no estimators perform substantially better than the ℓ_2-regularized GD or averaged SGD algorithms for recovering functions from respective function spaces. Nevertheless, these results are restricted to shallow neural networks, and cannot explain the generalization abilities of deep neural network (DNN). Similarly with [139], [140] obtained the $\mathcal{O}(n^{-1/2})$ generalization bound, showing that the SGD generalize well for $f_\rho^\star \in \mathcal{H}_L^{\textbf{NTK}}$, when $f_\rho^\star$ has a bounded RKHS norm. However, the rate they obtained is slower than the minimax rate we can actually achieve. Furthermore, their results become vacuous under the presence of additive noises on the data set. Motivated from these observations, the fundamental question in this study is as follows:

When the noisy dataset is generated from a function from $\mathcal{H}_L^{NTK}$, does the overparametrized DNN obtained via (ℓ_2-regularized) GD provably generalize well the unseen data?

We consider a neural network that has $L \geq 2$ hidden layers with width $m \gg n$. (i.e., overparametrized deep neural network.) We focus on the least-squares loss and assume that the activation function is ReLU. A positivity assumption of NTK from ReLU DNN is imposed, meaning that $\lambda_\infty > 0$, where λ_∞ denotes the minimum eigenvalue of the NTK. We give a more formal mathematical definition of ReLU DNN in the following Subsection 4.2.2. Under these settings, we provide an affirmative answer to the above question by investigating the behavior of L_2-prediction error of the obtained neural network with respect to GD iterations.

4.1.1 Contributions

Our derivations of algorithm-dependent prediction risk bound require the analysis on training dynamics of the estimated neural network through (regularized) GD algorithm. We include these results as the contributions of our paper, which can be of independent interests as well.

- In an unregulaized case, under the assumption $\lambda_{\infty} > 0$, we show that the training loss converges to 0 at a linear rate. As will be detailed in subsection 4.3.3, this is the different result from the seminal work of [136], where they also prove a linear convergence of training loss of ReLU DNN, but under different data distribution assumption.

- We show that the DNN updated via vanilla GD does not recover the ground truth function $f_{\rho}^{\star} \in \mathcal{H}_{L}^{\mathbf{NTK}}$ under noisy observations, if the DNN is trained for either too short or too long: that is, the prediction error is bounded away from 0 by some constant as n goes to infinity.

- In regularized case, we prove the mean-squared error (MSE) of DNN is upper bounded by some positive constant. Additionally, we proved the dynamics of the estimated neural network get close to the solution of kernel ridge regression associated with NTK from ReLU DNN.

- We show that the ℓ_2-regularization can be helpful in achieving the minimax optimal rate of the prediction risk for recovering $f_{\rho}^{\star} \in \mathcal{H}_{L}^{\mathbf{NTK}}$ under the noisy data. Specifically, it is shown that after some iterations of ℓ_2-regularized GD, the minimax optimal rate (which is $\mathcal{O}\big(n^{-\frac{d}{2d-1}}\big)$, where d is a feature dimension.) can be achieved.

Note that our paper is an extension of [142] to DNN model, showing that the ℓ_2-regularized DNN can achieve a minimax optimal rate of prediction error for recovering $f_{\rho}^{\star} \in \mathcal{H}_{L}^{\mathbf{NTK}}$. However, we would like to emphasize that our work is not a trivial application of their work from at least two technical aspects. These aspects are more detailed in the following subsection.

4.1.2 Technical Comparisons with [142]

Firstly, in the analysis of training loss of regularized shallow neural-net, [142] begin the proof by decomposing the difference between two individual predictions into two terms: one that is related with the gram matrix evaluated at each iteration of the algorithm and the perturbation term. Henceforth, we name this decompostion as “Gram+Pert” decomposition. This decomposition can be checked with the equality (E.2) in the supplementary PDF of [142]. The key ingredients for the decomposition are (1) the simple gradient structure of the shallow neural net, and (2) the partitioning of the nodes in the hidden-layer into two sets: a set of nodes whose activation patterns change from their initializations during training, and the complement of the set. This construction of the sets peels off the ReLU activation in the difference so that the GD algorithm can be involved in the analysis. However, because of the compositional structure of the network, the same nodes partitioning technique cannot be applied for obtaining the decomposition in the DNN setting with ReLU activation. To avoid this difficulty, we employ a specially designed diagonal matrix $\widetilde{\mathbf{\Sigma}}$ and this matrix can peel off the ReLU function for each layer of the network. (See the definition of $\widetilde{\mathbf{\Sigma}}$ in the proof of Theorem 2.4.1 in the Appendix.) Recursive applications of this diagonal matrix across the entire hidden layers enable the Gram+Pert decomposition in our setting. It should be noted that the diagnoal matrix $\widetilde{\mathbf{\Sigma}}$ had been employed in [138], which analyzed the behavior of training loss of classification problem via ReLU DNN under logistic loss. However, since their result is dependent on different data distribution assumption under the different loss function from ours, they didn’t employ the Gram+Pert decomposition. Thus their technical approaches are different from ours.

Secondly, [142] directly penalized the weight parameter $\mathbf{W}$ by adding $\|\mathbf{W}\|_F^2$ to the objective function. The ℓ_2-regularization solely on the $\mathbf{W}$ has an effect of pushing the weight towards the origin. This makes $\|\mathbf{W}^{(k)} - \mathbf{W}^{(0)}\|_2 \leq \mathcal{O}(1)$[2], allowing most activation patterns of the nodes in the hidden layer can change during the training, even in over-parametrized setting. Here, $\mathbf{W}^{(k)}$ denotes the updated weight parameter at kth itertaion of algorithm, and $\|\cdot\|_2$ denotes the spectral norm of the matrix. Nonetheless, this doesn't affect the analysis on obtaining the upper-bound of MSE in shallow neural net, since the network has only a single hidden layer. In contrast, in the DNN setting, we allow the non-convex interactions of parameters across the hidden layers. To the best of our knowledge, a technique for controlling the size of ℓ_2-norm of network gradient has not been developed under this setting, yet. We circumvent this difficulty by regularizing the distance between the updated and the initialized parameter, instead by directly regularizing the updated parameter. This ensures that the updated parameter by ℓ_2-regularized GD stays in a close neighborhood to its initialization, so that with heavy over-parametrization, the dynamics of network becomes linearized in parameter and we can ignore the non-convex interactions of parameters across the hidden layers. Specifically, under suitable model parameter setting, we prove that $\|\mathbf{W}_\ell^{(k)} - \mathbf{W}_\ell^{(0)}\|_2 \leq \widetilde{\mathcal{O}}_{\mathbb{P}}\big(\frac{1}{\sqrt{m}}\big)$ over all $\ell \in \{1, \ldots, L\}$. Here, $\widetilde{\mathcal{O}}_{\mathbb{P}}(\cdot)$ hides the dependencies on the model parameters; L, ω, and n. This result allows us to adopt the so-called "Forward Stability" argument developed by [136], and eventually leads to the control of network gradient under ℓ_2 sense.

4.1.3 Additional Related works

There has been another line of work trying to characterize the generalizabilities of DNN under noisy observation settings. Specifically, it has been shown that the neural network model can achieve minimax style optimal convergence rates of L_2-prediction risk both in regression [145, 146, 8] and classification [147] problems. Nonetheless, a limitation of the aforementioned papers is that they assume an adequate minimizer of the empirical risk can

[2] This was empirically shown to be true in paper [144]. See Figure 3 in their paper. We provide a brief mathematical explanation on why this result is hard to be shown in Appendix D.3.

be obtained. In other words, the mathematical proofs of their theorems do not correspond to implementable algortihms.

Recently, several papers, which study the generalization properties of neural network with algorithmic guarantees, appear online. Specifically, [148] showed that the data interpolants obtained through DNN by vanilla GD is inconsistent. This result is consistent with our result, but they consider the overparametrized DNN that is a linear combination of $\Omega(n^{10d^2})$ smaller neural network, and the activation function they consider is sigmoid function, which is smooth and differentiable. Along this line of research, [149] (regression) and [150] (classification) showed that when training overparametrized shallow neural network, early stopping of vanilla GD enables us to obtain consistent estimators.

Notation. We use the following notation for asymptotics: For sufficiently large n, we write $f(n) = \mathcal{O}(g(n))$, if there exists a constant $K > 0$ such that $f(n) \leq Kg(n)$, and $f(n) = \Omega(g(n))$ if $f(n) \geq K'g(n)$ for some constant $K' > 0$. The notation $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$. Let $\langle A, B\rangle_{\text{Tr}} := \text{Tr}(A^\top B)$ for the two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$. We adopt the shorthand notation denoting $[n] := \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$.

4.2 Problem Formulation

4.2.1 Non-parametric Regression

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ be the measureable feature space and output space. We denote ρ as a joint probability measure on the product space $\mathcal{X} \times \mathcal{Y}$, and let $\rho_{\mathcal{X}}$ be the marginal distribution of the feature space $\mathcal{X}$. We assume that the noisy data-set $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ are generated from the non-parametric regression model $\mathbf{y}_i = f_\rho^\star(\mathbf{x}_i) + \varepsilon_i$, where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1^2)$ for $i = 1, \dots, n$. Let $\widehat{f}_{W(k)}(\cdot)$ be the value of neural network evaluated with the parameters $\mathbf{W}$ at the k-th iterations of GD update rule. At $k = 0$, we randomly initialize the weight parameters in the model following He initialization [151] with a slight modification.

Then, the L_2 prediction risk is defined as the difference between two expected risks (i.e., excess risk) $\mathcal{R}\big(\widehat{f}_{W(k)}\big) := \mathbb{E}_{\rho\sim(\mathbf{x},\mathbf{y})}\big[\big(\mathbf{y} - \widehat{f}_{W(k)}(\mathbf{x})\big)^2\big]$ and $\mathcal{R}(f_\rho^\star) := \mathbb{E}_{\rho\sim(\mathbf{x},\mathbf{y})}\big[\big(\mathbf{y} - f_\rho^\star(\mathbf{x})\big)^2\big]$, where $f_\rho^\star(\mathbf{x}) := \mathbb{E}[\mathbf{y}|\mathbf{x}]$. Then, we can easily show the prediction risk has a following form:

$$\mathcal{R}\big(\widehat{f}_k, f_\rho^\star\big) := \mathcal{R}\big(\widehat{f}_{W(k)}\big) - \mathcal{R}\big(f^\star\big) = \mathbb{E}_{\rho_{\mathbf{x}},\varepsilon}\Big[\big(\widehat{f}_{W(k)}(\mathbf{x}) - f_\rho^\star(\mathbf{x})\big)^2\Big]. \tag{4.1}$$

Note that the expectation is taken over the marginal probability measure of feature space, $\rho_{\mathbf{x}}$, and the noise of the data, ε. However, the (4.1) is still a random quantity due to the randomness of the initialized parameters $\big(\mathbf{W}_\ell^{(0)}\big)_{\ell=1,\ldots,L}$.

4.2.2 Deep Neural Network with ReLU activation

Following the setting introduced in [136], we consider a fully-connected deep neural networks with L hidden layers and m network width. For $L \geq 2$, the output of the network $f_{\mathbf{W}}(\cdot) \in \mathbb{R}$ with input data $\mathbf{x} \in \mathcal{X}$ can be formally written as follows:

$$f_{\mathbf{W}}(\mathbf{x}) = \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}}\sigma\big(\mathbf{W}_L\sigma\big(\mathbf{W}_{L-1}\cdots\sigma\big(\mathbf{W}_1\mathbf{x}\big)\cdots\big)\big), \tag{4.2}$$

where $\mathcal{S}^{d-1}$ is a unit sphere in d-dimensional euclidean space, $\sigma(\cdot)$ is an entry-wise activation function, $\mathbf{W}_1 \in \mathbb{R}^{m\times d}$, $\mathbf{W}_2, \ldots, \mathbf{W}_L \in \mathbb{R}^{m\times m}$ denote the weight matrices for hidden layers and $\mathbf{v} \in \mathbb{R}^{m\times 1}$ denote the weight vector for the output layer. Following the existing literature, we will consider ReLU activation function $\sigma(x) = \max(x, 0)$, which is the most commonly used activation function by practitioners.

Random Initialization. Each entries of weight matrices in hidden layers are assumed to be generated from $\big(\mathbf{W}_{i,j}\big)_{\ell=1,\ldots,L} \sim \mathcal{N}(0, \frac{2}{m})$, and entries of the output layer are drawn from $\mathbf{v}_j \sim \mathcal{N}(0, \frac{\omega}{m})$. This initialization scheme helps the forward propagation neither explode nor vanish at the initialization, seeing [136, 137, 138]. Note that we initialize the parameters in the last layer with variance $\frac{\omega}{m}$, where $\omega \leq 1$ is a model parameter to be chosen later for technical convenience.

Unregularized GD update rule. We solve a following ℓ_2-loss function with the given dataset $\mathcal{D}$:

$$\mathcal{L}_{\mathbf{S}}(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{n}\big(y_i - f_{\mathbf{W}}(\mathbf{x}_i)\big)^2. \tag{4.3}$$

Let $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ be the initialized weight matrices introduced above, and we consider a following gradient descent update rule:

$$\mathbf{W}_\ell^{(k)} = \mathbf{W}_\ell^{(k-1)} - \eta\nabla_{\mathbf{W}_\ell}\big(\mathcal{L}_{\mathcal{S}}(\mathbf{W}_\ell^{(k-1)})\big), \quad \ell \in [L], \quad k \geq 1, \tag{4.4}$$

where $\nabla_{\mathbf{W}_\ell}\big(\mathcal{L}_{\mathcal{S}}(\cdot)\big)$ is a partial gradient of the loss function $\mathcal{L}_{\mathcal{S}}(\cdot)$ with respect to the ℓ-th layer parameters $\mathbf{W}_\ell$, and $\eta > 0$ is the learning rate of the gradient descent.

ℓ_2-regularized GD update rule. The estimator is obtained by minimizing a ℓ_2-regularized function;

$$\mathbf{\Phi}_D(\mathbf{W}) := \mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D\big) + \frac{\mu}{2}\sum_{\ell=1}^{L}\Big\|\mathbf{W}_{D,\ell} - \mathbf{W}_{D,\ell}^{(0)}\Big\|_F^2. \tag{4.5}$$

Naturally, we update the model parameters $\big\{\mathbf{W}_{D,\ell}\big\}_{\ell=1,\ldots,L}$ via modified GD update rule:

$$\mathbf{W}_{D,\ell}^{(k)} = \big(1 - \eta_2\mu\big)\mathbf{W}_{D,\ell}^{(k-1)} - \eta_1\nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D^{(k-1)}\big)\big] + \eta_2\mu\mathbf{W}_{D,\ell}^{(0)}, \quad \forall\ell \in [L], \quad \forall k \geq 1. \tag{4.6}$$

The notations η_1, η_2 are step sizes, and $\mu > 0$ is a tuning parameter on regularization. We adopt the different step sizes for the partial gradient and regularized term for the theoretical conveniences. Furthermore, we add the additional subscript D to the update rule (4.6) to denote the variables are under the regularized GD update rule. Recall that the $\mathbf{W}_{D,\ell}^{(0)}$ are initialized parameters same with the unregularized case. For simplicity, we fix the output layer, and train L hidden layers for both unregularized and regularized cases.

4.3 Main Theory

First, we describe the neural tangent kernel (NTK) matrix of (4.2), which is first proposed by [131] and further studied by [130, 152, 132, 153]. NTK matrix of DNN is a L-times recursively defined $n \times n$ kernel matrix, whose entries are the infinite-width limit of the gram matrix. Let $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\cdot)\big]$ be the gradient of the ReLU DNN (4.2) with respect to the weight matrix in the ℓth hidden layer at random initialization. Note that when $\ell = 1$, $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\cdot)\big] \in \mathbb{R}^{m\times d}$ and when $\ell \in \{2,\dots,L\}$, $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\cdot)\big] \in \mathbb{R}^{m\times m}$. Then, as $m \to \infty$,

$$\mathbf{H}(0) := \left(\frac{1}{m}\sum_{\ell=1}^{L}\big\langle \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_i)\big], \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_j)\big]\big\rangle_{\text{Tr}}\right)_{n\times n} \to \mathbf{H}_L^\infty, \tag{4.7}$$

where $\mathbf{H}_L^\infty := \big\{\mathbf{Ker}(\mathbf{x}_i,\mathbf{x}_j)\big\}_{i,j=1}^n$. Here, $\mathbf{Ker}(\cdot,\cdot)$ denotes a NTK function of (4.2) to be defined as follows:

Definition 4.3.1 ***(NTK function of (4.2)).*** *For any* $\mathbf{x},\mathbf{x}' \in \mathcal{X}$ *and* $\ell \in [L]$*, define*

$$\begin{aligned}
\Phi^{(0)}(\mathbf{x},\mathbf{x}') &= \langle \mathbf{x},\mathbf{x}'\rangle\,,\\
\Theta^{(\ell)}(\mathbf{x},\mathbf{x}') &= \begin{pmatrix}\Phi^{(\ell-1)}(\mathbf{x},\mathbf{x}) & \Phi^{(\ell-1)}(\mathbf{x},\mathbf{x}')\\ \Phi^{(\ell-1)}(\mathbf{x}',\mathbf{x}) & \Phi^{(\ell-1)}(\mathbf{x}',\mathbf{x}')\end{pmatrix} \in \mathbb{R}^{2\times 2},\\
\Phi^{(\ell)}(\mathbf{x},\mathbf{x}') &= 2\cdot \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\Theta^{(\ell)})}\big[\sigma(u)\cdot\sigma(v)\big], \quad \textit{and}\\
\dot{\Phi}^{(\ell)}(\mathbf{x},\mathbf{x}') &= 2\cdot \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\Theta^{(\ell)})}\big[\dot{\sigma}(u)\cdot\dot{\sigma}(v)\big],
\end{aligned}$$

where $\dot{\sigma}(u) = \mathbb{1}\big(u \geq 0\big)$*. Then, we can derive the final expression of NTK function of (4.2)* *as follows:*

$$\mathbf{Ker}(\mathbf{x}, \mathbf{x}') = \frac{\omega}{2} \cdot \sum_{\ell=1}^{L} \left(\Phi^{(\ell-1)}(\mathbf{x}, \mathbf{x}') \cdot \prod_{\ell'=\ell}^{L} \dot{\Phi}^{(\ell')}(\mathbf{x}, \mathbf{x}') \right). \tag{4.8}$$

The expression in (4.8) is adapted from [140]. As remarked in [140], a coefficient 2 in $\Phi^{(\ell)}$ and $\dot{\Phi}^{(\ell)}$ remove the exponential dependence on the network depth L in the NTK function. However, when compared with the NTK formula in [140], (4.8) is different from two aspects: (1) An additional factor ω in (4.8)) comes from the difference in initialization settings of the output layer, in which [140] considers $v_j \sim \mathcal{N}(0, \frac{1}{m})$, whereas we consider $v_j \sim \mathcal{N}(0, \frac{\omega}{m})$. (2) $\Phi^{(L)}$ is not added in the final expression of (4.8)), whereas it is added in the definition provided in [140]. This is because we only train the L hidden layers but fix the output layer, while [140] train the entire layers of the network including the output layer.

As already been pointed by several papers, [154] and [131], it can be proved that the NTK function (4.8) is a positive semi-definite kernel function. Furthermore, [154] prove that the expectations in Φ and $\dot{\Phi}$ have closed form solutions, when the covariance matrices have the form $\left(\begin{smallmatrix} 1 & t \\ t & 1 \end{smallmatrix}\right)$ with $|t| \leq 1$:

$$\begin{aligned} \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\Theta^{(\ell)})} \big[\sigma(u) \cdot \sigma(v)\big] &= \frac{1}{2\pi} \left(t \cdot (\pi - \arccos(t)) + \sqrt{1-t^2} \right), \\ \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\Theta^{(\ell)})} \big[\dot{\sigma}(u) \cdot \dot{\sigma}(v)\big] &= \frac{1}{2\pi} \big(\pi - \arccos(t)\big). \end{aligned} \tag{4.9}$$

Clearly, (4.8) is symmetric and continuous on the product space $\mathcal{X} \times \mathcal{X}$, from which it can be implied that **Ker**$(\cdot, \cdot)$ is a Mercer kernel inducing an unique RKHS. Following [155], we define the RKHS induced by (4.8) as:

Definition 4.3.2 ***(NTK induced RKHS).*** *For some integer $p \in \mathbb{N}$, set of points $\{\tilde{\mathbf{x}}_j\}_{j=1}^{p} \subset \mathcal{X}$, and weight vector $\alpha := \{\alpha_1, \ldots, \alpha_p\} \in \mathbb{R}^p$, define a complete vector space of functions,*

$f : \mathcal{X} \to \mathbb{R}$,

$$\mathcal{H}_L^{\textit{NTK}} := cl\Bigg(\bigg\{f(\cdot) = \sum_{j=1}^{p} \alpha_j \mathbf{Ker}(\cdot, \tilde{\mathbf{x}}_j)\bigg\}\Bigg), \tag{4.10}$$

where $cl(\cdot)$ denotes closure.

In the remaining of our work, we assume the regression function $f_\rho^\star(\mathbf{x}) := \mathbb{E}[\mathbf{y}|\mathbf{x}]$ belongs to $\mathcal{H}_L^{\mathbf{NTK}}$.

4.3.1 Assumptions.

In this subsection, we state the assumptions imposed on the data distribution with some remarks.

(A1) $\rho_{\mathcal{X}}$ is an uniform distribution on $\mathcal{S}^{d-1} := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$, and noisy observations are assumed to be bounded. (i.e., $\rho_{\mathbf{x}} \sim \mathbf{Unif}\big(\mathcal{S}^{d-1}\big)$, $\mathbf{y}_i = \mathcal{O}(1)$, $\forall i \in [n]$.)

(A2) Draw n i.i.d. samples $\{\mathbf{x}_i, f_\rho^\star(\mathbf{x}_i)\}_{i=1}^n$ from the joint measure ρ. Then, with probability at least $1 - \delta$, we have $\lambda_{\text{min}}\big(\mathbf{H}_L^\infty\big) = \lambda_\infty > 0$.

Remark 4.3.3

- When the feature space is restricted on the unit sphere, the NTK function in (4.8) becomes rotationally invariant zonal kernel. This setting allows to adopt the results of spectral decay of (4.8) in the basis of spherical harmonic polynomials for measuring the complexity of hypothesis space, $\mathcal{H}_L^{\mathbf{NTK}}$. See the subsection 4.3.2 and references therein.

- Assumption (A2) is commonly employed in NTK related literature for proving global convergence of training error and generalization error of both deep and shallow neural network, [135, 152, 139]. Note that the (A2) holds as long as no two $\mathbf{x}_i$ and $\mathbf{x}_j$ are parallel to each other, which is true for most of the real-world distributions. See the proof of this claim in [152].

4.3.2 Minimax rate for recovering $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$

The obtainable minimax rate of L_2-prediction error is directly related with the complexity of function space of interest. In our setting, the complexity of RKHS $\mathcal{H}_L^{\mathbf{NTK}}$ can be characterized by the eigen-decay rate of the NTK function. Since $\mathbf{Ker}(\mathbf{x}, \mathbf{x}')$ is defined on the sphere, the decomposition can be given in the basis of spherical harmonics as follows:

$$\mathbf{Ker}(\mathbf{x}, \mathbf{x}') = \sum_{k=0}^{\infty} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(\mathbf{x}) Y_{k,j}(\mathbf{x}'),$$

where $Y_{k,j}, j = 1, \ldots, N(d,k)$ are spherical harmonic polynomials of degree k and $\{\mu_k\}_{k=0}^{\infty}$ are non-negative eigenvalues. Recently, several researchers, both empirically [156] and theoretically [157, 158, 159], showed that, for large enough harmonic function frequency k, the decay rate of the eigenvalues μ_k is in the order of $\Theta\big(k^{-d}\big)$ [3]. Given this result and the fact $N(d,k) = \frac{2k+d-3}{k}\binom{k+d-3}{d-2}$ grows as k^{d-2} for large k, it can be easily shown $\lambda_j = \Theta\big(j^{-\frac{d}{d-1}}\big)$, when $\mathbf{Ker}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$, for eigen-values $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$. Furthermore, it is a well known fact that if the eigenvalues decay at the rate $\lambda_j = \Theta(j^{-2\nu})$, then the corresponding minimax rate for estimating function in RKHS is $\mathcal{O}\big(n^{-\frac{2\nu}{2\nu+1}}\big)$, [161, 162, 142]. By setting $2\nu = \frac{d}{d-1}$, we can see the minimax rate for recovering $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$ is $\mathcal{O}\big(n^{-\frac{d}{2d-1}}\big)$.

Remark 4.3.4 *We defer all the technical proofs of the Theorems in subsections* 4.3.3 *and* 4.3.4 *in the Appendix for conciseness of the paper. We also provide numerical experiments which can corroborate our theoretical findings in the Appendix D.1.*

[3] In shallow neural network with ReLU activation without bias terms, it is shown that μ_k satisfy μ_0, $\mu_1 > 0$, $\mu_k = 0$ if $k = 2j+1$ with $j \geq 1$, and otherwise $\mu_k = \Theta\big(k^{-d}\big)$. See [160]. However, in ReLU DNN, it is shown that these parity constraints can be removed even without bias terms and μ_k achieves $\Theta\big(k^{-d}\big)$ decay rate for large enough k. Readers can refer [159] for this result.

4.3.3 Analysis of Unregularized DNN

In this subsection, we provide the results on the training loss of DNN estimator obtained via minimizing unregularized ℓ_2-loss (4.3) and on the corresponding estimator's L_2-prediction risk $\mathcal{R}\big(\widehat{f}_k, f_\rho^\star\big)$.

Theorem 4.3.5 ***(Optimization)*** *For some $\delta \in [0,1]$, set the width of the network as $\frac{m}{\log^3(m)} \geq \Omega\big(\frac{\omega^7 n^8 L^{18}}{\lambda_\infty^8 \delta^2}\big)$, and set the step-size of gradient descent as $\eta = \mathcal{O}\big(\frac{\lambda_\infty}{n^2 L^2 m}\big)$. Then, with probability at least $1-\delta$ over the randomness of initialized parameters $\mathcal{W}^{(0)} := \big\{\mathbf{W}_\ell^{(0)}\big\}_{\ell=1}^{L+1}$ with $\mathbf{W}_{L+1}^{(0)} = \mathbf{v}$, we have for $k = 0, 1, 2, \dots,$*

$$\mathcal{L}_{\mathbf{S}}\big(\mathcal{W}^{(k)}\big) \leq \bigg(1 - \frac{\eta m \lambda_\infty}{2}\bigg)^k \mathcal{L}_{\mathbf{S}}\big(\mathcal{W}^{(0)}\big). \tag{4.11}$$

In other words, the training loss drops to 0 at a linear rate.

We acknowledge a series of past works [136, 152] have similar spirits with those in Theorem 4.3.5. However, it is worth noting that their results are not applicable in our problem settings and data assumptions. Specifically, the result of [152] is based on the smooth and differentiable activation function, whereas the Theorem 4.3.5 is about the training error of ReLU activation function, which is not differentiable at 0. Furthermore, the result of [136] relies on ϕ-separateness assumption stating that the every pair of feature vectors $\big\{\mathbf{x}_i, \mathbf{x}_j\big\}_{i\neq j}^n$ is apart from each other by some constant $\phi > 0$ in a Euclidean norm. In our work, the positivity assumption on the minimum eigenvalue of the NTK is imposed (i.e., $\lambda_\infty > 0$).

Remark 4.3.6 *Reducing the order of network width is definitely another line of interesting research direction. We are aware of some works in literature, but we chose not to adopt the techniques since this can make the analysis overly complicated. To the best of our knowledge, the paper that most neatly summarizes this line of literature is [163]. See the* *table in page* 3 *in their paper. The order of width they obtained is* $\Omega\big(\frac{n^8 L^{12}}{\phi^8}\big)$*, where they impose* ϕ*-separateness assumption.*

Remark 4.3.7 *There has been an attempt to make a connection between the positivity and* ϕ*-separateness assumptions. Recently, [163] proved the relation* $\lambda_\infty = \Omega\big(\phi n^{-2}\big)$ [4] *in a shallow-neural net setting. See Proposition* $3.6.$ *of their work. However, it is still an open question on whether this relation holds in DNN setting as well. The results in Theorem 4.3.5 suggest a positive conjecture on this question. Indeed, plugging the relation* $\lambda_\infty = \Omega\big(\phi n^{-2}\big)$ *in (4.11) and in the* $\eta = \mathcal{O}\big(\frac{\lambda_\infty}{n^2 L^2 m}\big)$ *yield the discount factor* $\big(1-\Omega\big(\frac{\eta m \phi}{n^2}\big)\big)^k$ *and step-size* $\eta = \mathcal{O}\big(\frac{\phi}{n^4 L^2 m}\big)$*, which are exactly the same orders as presented in [136]. See Theorem* 1 *of their ArXiv version paper for the clear comparison. We leave the proof of this conjecture as a future work.*

Theorem 4.3.8 ***(Generalization)*** *Let* $f_\rho^\star \in \mathcal{H}_L^{\boldsymbol{NTK}}$*. Fix a failure probability* $\delta \in [0,1]$*. Set the width of the network as* $\frac{m}{\log^3(m)} \geq \Omega\big(\frac{\omega^7 n^8 L^{18}}{\lambda_\infty^8 \delta^2}\big)$*, the step-size of gradient descent as* $\eta = \mathcal{O}\big(\frac{\lambda_\infty}{n^2 L^2 m}\big)$*, and the variance parameter* $\omega \leq \mathcal{O}\big(\big(\frac{\lambda_\infty \delta}{n}\big)^{2/3}\big)$*. Then, if the GD iteration* $k \geq \Omega\big(\frac{\log(n)}{\eta m \lambda_\infty}\big)$ *or* $k \leq \mathcal{O}\big(\frac{1}{\eta m \omega L}\big)$*, with probability at least* $1-\delta$ *over the randomness of initialized parameters* $\mathcal{W}^{(0)}$*, we have*

$$\mathcal{R}\big(\widehat{f}_k, f_\rho^\star\big) = \Omega(1).$$

This theorem states that if the network is trained for too long or too short, the L_2-prediction error of $\widehat{f}_{\mathbf{W}^{(k)}}$ is bounded away from 0 by some constant factor. Specifically, the former scenario indicates that the overfitting can be harmful for recovering $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$ given the noisy observations.

Remark 4.3.9 *Readers should note that the Theorem 3.3.3 does not consider if the GD algorithm can achieve low prediction risk* $\mathcal{R}\big(\widehat{f}_k, f_\rho^\star\big)$ *over the range of iterations* $(\eta m \omega L)^{-1} \lesssim$

[4] We conjecture that this is not the tightest lower bound on λ_∞. Recently, [164] proves that $\lambda_\infty \gtrsim d/n$ in shallow neural net setting. See Lemma 5.3 in their paper.

$k \lesssim (\eta m \lambda_\infty)^{-1} \log(n)$. *In the numerical experiment to be followed in Appendix D.1, we observe that for some algorithm iterations k^*, the risk indeed decreases to the same minimum as low as the ℓ_2-regularized algorithm can achieve, and increases again. This observation implies that the unregularized algorithm can achieve the minimax rate of prediction risk. However, analytically deriving a data-dependent stopping time k^* in our scenario requires further studies, since we need a sharp characterization of eigen-distribution of NTK matrix of ReLU DNN, denoted as $\mathbf{H}_L^\infty$ in this paper. Readers can refer the Theorem* 4.2. *of [142] in shallow-neural network and equation* (6) *in [161] in kernel regression context on how to compute $k^\star$ with the given eigen-values of the associated kernel matrices.*

4.3.4 Analysis of ℓ_2-regularized DNN

In this subsection, we study the training dynamics of ℓ_2-regularized DNN and the effects of the regularization for obtaining the minimax optimal convergence rate of L_2-prediction risk. In the results to be followed, we set the orders of model parameters μ, η_1, η_2 in (4.6), and a variance parameter of output layer, ω as follows:

$$\mu = \Theta\bigg(n^{\frac{d-1}{2d-1}}\bigg), \quad \eta_1 = \Theta\bigg(\frac{1}{m} n^{-\frac{3d-2}{2d-1}}\bigg), \quad \eta_2 = \Theta\bigg(\frac{1}{L} n^{-\frac{3d-2}{2d-1}}\bigg), \quad \omega = \mathcal{O}\bigg(\frac{1}{L^{3/2}} n^{-\frac{5d-2}{2d-1}}\bigg). \tag{4.12}$$

Theorem 4.3.10 ***(Optimization)*** *Suppose we minimize ℓ_2-regularized objective function (4.5) via modified GD (4.6). Set the network width $\frac{m}{\log^3(m)} \geq \Omega\big(\frac{L^{20} n^{24}}{\delta^2}\big)$ and model parameters as in (4.12). Then, with probability at least $1-\delta$, the mean-squared error follows*

$$\mathcal{L}_{\mathbf{S}}\big(\mathcal{W}_D^{(k)}\big)/n \leq \bigg(1 - \eta_2 \mu L\bigg)^k \cdot \mathcal{L}_{\mathbf{S}}\big(\mathcal{W}_D^{(0)}\big)/n + \mathcal{O}_{\mathbb{P}}(1), \tag{4.13}$$

for $k \geq 0$. Additionally, after $k \geq \Omega\big((\eta_2\mu L)^{-1}\log(n^{3/2})\big)$ iterations of (4.6), for some *constant $C > 0$, we have*

$$\left\| \mathbf{u}_D(k) - \mathbf{H}_L^{\infty}\left(C\mu \cdot \mathcal{I} + \mathbf{H}_L^{\infty} \right)^{-1} \mathbf{y} \right\|_2 \leq \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n}\right), \tag{4.14}$$

where we denote $\mathbf{u}_D(k) := [\widehat{f}_{\mathbf{W}_D^{(k)}}(\mathbf{x}_1), \ldots, \widehat{f}_{\mathbf{W}_D^{(k)}}(\mathbf{x}_n)]^{\top}$.

Several comments are in sequel. Theorem 4.3.10 is, to our knowledge, the first result that rigorously shows the training dynamics of ℓ_2-regularized ReLU DNN in over-parametrized setting. Observe that the first term on the right-hand side of the inequality (4.13) converges linearly to 0, and the second term is some positive constant that is bounded away from 0. This implies that the MSE of regularized DNN is upper-bounded by some positive constant. Note that we only provide the upper bound, but the results of our numerical experiments indicate that the MSE is lower-bounded by $\mathcal{O}_{\mathbb{P}}(1)$ as well. We leave the proof of this conjecture for the future work.

The inequality (4.14) states that the trained dynamics of the regularized neural network can approximate the optimal solution (denoted as $g_{\mu}^{\star}$) of the following kernel ridge regression problem:

$$\min_{f \in \mathcal{H}^{\text{NTK}}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i) \big)^2 + \frac{C\mu}{2} \|f\|_{\mathcal{H}_L^{\text{NTK}}}^2 \right\}, \tag{4.15}$$

where $\| \cdot \|_{\mathcal{H}_L^{\text{NTK}}}$ denotes a NTK-induced RKHS norm. Note that the optimization problem in (4.15) is not normalized by sample size n. The inequality (4.14) states that after approximately $(\eta_2 \mu L)^{-1}$ iterations of (4.6), the error rate becomes $\mathcal{O}_{\mathbb{P}}\big(\frac{1}{n}\big)$. The approximation error is computed at the training data points under ℓ_2 norm. This should be compared with the Theorem 5.1 of [142], where they showed that the similar approximation holds "within" a certain range of algorithm in shallow neural network setting. In contrast, we show that the approximation holds "after" $k \geq \Omega\big((\eta_2 \mu L)^{-1} \log(n^{3/2})\big)$ in deep neural network. It should be noted that the difference of results comes from the regularization scheme, where we penalize the $\sum_{\ell=1}^{L}\|\mathbf{W}_\ell - \mathbf{W}_\ell^{(0)}\|_F^2$, whereas [142] regularized the term $\|\mathbf{W}_1\|_F^2$.

As another important comparison, [165] showed the equivalence of a solution of kernel ridge regression associated with NTK and the first order Taylor expansion of the regularized neural network dynamics; note, however, that the $\mathbf{u}_D(k)$ in (4.14) is a full neural network dynamics. Let $\mathcal{R}(\widehat{f}_{\mathbf{W}_D^{(k)}}, f_\rho^\star)$ be the L_2-prediction risk of the regularized estimator $\widehat{f}_{\mathbf{W}_D^{(k)}}$ via modified GD (4.6). Next theorem states the result of generalization ability of $\widehat{f}_{\mathbf{W}_D^{(k)}}$.

Theorem 4.3.11 ***(Generalization)*** *Let $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$. Suppose the network width $\frac{m}{\log^3(m)} \geq \Omega\big(\frac{L^{20}n^{24}}{\delta^2}\big)$ and model parameters are set as suggested in (4.12). Then, with probability tending to* 1*, we have*

$$\mathcal{R}\big(\widehat{f}_{\mathbf{W}_D^{(k)}}, f_\rho^\star\big) = \mathcal{O}_{\mathbb{P}}\bigg(n^{-\frac{d}{2d-1}}\bigg).$$

The resulting convergence rate is $\mathcal{O}\big(n^{-\frac{d}{2d-1}}\big)$ with respect to the training sample size n. Note that the rate is always faster than $\mathcal{O}\big(n^{-1/2}\big)$ and turns out to be the minimax optimal [166, 167] for recovering $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$ in the following sense:

$$\lim_{r\to 0}\liminf_{n\to\infty}\inf_{\widehat{f}}\sup_{\rho}\mathbb{P}\bigg[\mathcal{R}\big(\widehat{f}, f_\rho^\star\big) > rn^{-\frac{d}{2d-1}}\bigg] = 1, \tag{4.16}$$

where ρ is a data distribution class satisfying the Assumptions (A1), (A2) and $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$, and infimum is taken over all estimators $\mathcal{D} \to \widehat{f}$. It is worth noting that the minimax rate in (4.16) is same with the minimax rate for recovering $f_\rho^\star \in \mathcal{H}_1^{\mathbf{NTK}}$. (i.e., [142]) This result can be derived from the recent discovery of the equivalence between two function spaces , $\mathcal{H}_1^{\mathbf{NTK}} = \mathcal{H}_L^{\mathbf{NTK}}$. See [158] and [157].

Remark 4.3.12 *A particular choice of $\mu = \Theta\big(n^{\frac{d-1}{2d-1}}\big)$ in (4.12) is for obtaining an optimal* *minimax rate for prediction error in Theorem 4.3.11. Specifically, the order of μ determines the L_2 distance between the $f_\rho^\star$ and the kernel regressor $g_\mu^\star$. That is, $\|f_\rho^\star - g_\mu^\star\|_2^2 = \mathcal{O}_{\mathbb{P}}(\frac{\mu}{n})$. With the result $\mathcal{H}_1^{\boldsymbol{NTK}} = \mathcal{H}_L^{\boldsymbol{NTK}}$, the same proof of Lemma D.2. in [142] can be applied for proving this result.*

4.4 Conclusion

We analyze the convergence rate of L_2-prediction error of both the unregularized and the regularized gradient descent for overparameterized DNN with ReLU activation for a regression problem. Under a positivity assumption of NTK, we show that without the adoption of early stopping, the L_2-prediction error of the estimated DNN via vanilla GD is bounded away from 0 (Theorem 2.4.1), whereas the prediction error of the DNN via ℓ_2-regularized GD achieves the optimal minimax rate (Theorem 4.3.11). The minimax rate $\mathcal{O}\big(n^{-\frac{d}{2d-1}}\big)$ is faster than the $\mathcal{O}(n^{-1/2})$ by specifying the complexities of target function and hypothesis space.

CHAPTER 5

APPROXIMATION AND NON-PARAMETRIC ESTIMATION OF FUNCTIONS OVER HIGH-DIMENSIONAL SPHERES VIA DEEP RELU NETWORKS

Neural networks have demonstrated tremendous success in the tasks of image classification [168, 169], pattern recognition [170], natural language processing [171, 172, 173], etc. The datasets used in these real world applications frequently lie in high-dimensional spaces [7], In this chapter, we try to understand the fundamental limits of neural network in the high-dimensional regime through the lens of its approximation power and its generalization error.

Both approximation power and generalization error of neural network can be analyzed through specifying the target function's property such as its smoothness index $r > 0$ and the input space $\mathcal{X}$. In particular, deep feed-forward neural networks (FNNs) with Rectified Linear Unit (ReLU) have been extensively studied when they are used for approximating and estimating functions from general function class such as Sobolev class defined on d-dimensional cube (i.e., $\mathcal{X} := \mathcal{C}^d$), denoted as $W_p^r(\mathcal{C}^d)$ for $1 \leq p \leq \infty$. However, in practice, signals on spherical surface (i.e., $\mathcal{X} := \mathcal{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$) rather than in Euclidean spaces often arise in various situations, such as astrophysics [174, 175], computer vision [176], and medical imaging [177].

Motivated from this, we focus our attention in the cases where deep ReLU FNNs are used for function approximators and estimators, when functions are assumed to be from the Sobolev spaces defined over $\mathcal{S}^{d-1}$; that is $f \in W_\infty^r(\mathcal{S}^{d-1})$. Under this setting, our analysis focuses on how the input dimension d explicitly affects the approximation and estimation rates of $f \in W_\infty^r(\mathcal{S}^{d-1})$. And, at the same time, we show how the scalability of deep ReLU FNNs grows in the high-dimensional regime. Here, the scalability is mainly measured through the three metrics: (1) *the width denoted as* $\mathcal{W}$, (2) *the depth, denoted*

Table 5.1: Notation $\tilde{\mathcal{O}}(\cdot)$ hide the logarithmic factor in n. Note that the upper-bounds for $\mathcal{N}$ in Theorem 5.3.3 (i.e., $\mathcal{N} = \mathcal{O}(Md)$) are from Theorem 5.2.1 with choices $M = \lceil n^{\frac{3d}{3d+4r}} \rceil$ and $n \ll d$.

	Theorem 5.3.3		Theorem 5.3.4
Function class	$W^r_\infty(\mathcal{S}^{d-1})$		$W^r_\infty([0,1]^d)$
Smoothness r	$\mathcal{O}(d)$	$\mathcal{O}(1)$	$\forall r > 0$
Upper-bound on $\mathcal{N}$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}((d+r)^d)$
Estimation error rate	$\tilde{\mathcal{O}}\big(d^C \cdot n^{-\frac{4r}{4r+3d}}\big)$	$\tilde{\mathcal{O}}\big(\big(\frac{6}{\pi e}\big)^{\frac{d}{2}} d^d \cdot n^{-\frac{4r}{4r+3d}}\big)$	$\tilde{\mathcal{O}}\big((d+r)^d \cdot n^{-\frac{2r}{2r+d}}\big)$

as L, and (3) *the number of active parameters, denoted as* $\mathcal{N}$ of the network, [178]. It should be emphasized that we find there exists an interaction with smoothness index $r > 0$ and dimension d, whereas we cannot find one for the case when $f \in W^r_\infty(\mathcal{C}^d)$. We further summarize our detailed findings in the following subsection.

5.0.1 Chapter Road map and Contributions

In Theorem 5.2.1, we provide an approximation bound of deep ReLU FNN (i.e., $\tilde{f}$) for approximating the target functions in Sobolev spaces defined over sphere (i.e., $f \in W^r_\infty(\mathcal{S}^{d-1})$). Notably, in the bound, we track the explicit dependence on data dimension d allowing it tend to infinity. This tracking enables how the three components of network architecture, ***width*** ($\mathcal{W}$), ***depth*** (L), and ***the number of active parameters*** ($\mathcal{N}$), should change as d increases, for obtaining the good approximation error rate.

As a Corollary of Theorem 5.2.1, we show how the order of function smoothness r can have the effect on the scale of network in terms of d. Specifically, when the function smoothness $r = \mathcal{O}(1)$, we show that the constructed network, $\tilde{f}$, requires $\mathcal{W} = \mathcal{O}(d^d)$, $L = \mathcal{O}(d^\gamma \log_2 d)$ for $0 < \gamma < 1$, and $\mathcal{N} = \mathcal{O}(d^{d+1})$ for obtaining $d^{-\mathcal{O}(1)}$ approximation error up to some constant factors independent with d. Furthermore, when $r = \mathcal{O}(d)$, we show that only $\mathcal{W} = \mathcal{O}(d^\alpha)$, $L = \mathcal{O}(d^\gamma \log_2 d)$, and at most $\mathcal{N} = \mathcal{O}(d^2)$ are required for obtaining the sharp approximation rate $\mathcal{O}(d^{-d^\beta})$ for $0 < \alpha, \beta < 1$. See Corollary 5.2.3 for the detailed statement of the result.

Our result implies that for approximating $f \in W_{\infty}^{r}(\mathcal{S}^{d-1})$, the larger the smoothness index r is, the narrower the width of the network should be enough, while the depth of the network can be fixed. Moreover, when r is in the same order of d, the network can avoid the ***curse of dimensionality*** requiring only $\mathcal{O}(d^2)$ number of active parameters. It is interesting to note that the function smoothness index can affect the design of the network, specifically on width, while it has little effect on the design of depth. Admittedly, the condition $r = \mathcal{O}(d)$ is restrictive in a sense that it makes the function space $W_{\infty}^{r}(\mathcal{S}^{d-1})$ small.

Nonetheless, to the best of our knowledge, this finding is not observed in the current approximation theory of neural network literature when $f \in W_{\infty}^{r}(\mathcal{C}^{d})$ where $\mathcal{C}^d$ denotes some d-dimensional cubes, and $\tilde{f}$ is a deep ReLU FNN. Out of the long list of literature to be introduced shortly, we choose the result from [8] for the comparison as it also has the explicit dependence on d in their approximation bound. From their result, it can be seen that the ***curse*** cannot be avoided, even when $r = \mathcal{O}(d)$. The width of their constructed network is lower bounded by $\Omega(r^d \vee e^d)$ and the number of active parameters is upper-bounded by $\mathcal{O}((r+d)^d)$. [1] Note that the bounds on both components grow exponentially in d as r increases. See subsection 5.2.1 for the detailed comparisons.

We further make the comparisons between estimating functions $f \in W_{\infty}^{r}(\mathcal{S}^{d-1})$ (Theorems 5.3.3) versus $f \in W_{\infty}^{r}(\mathcal{C}^{d})$ (Theorems 5.3.4) via deep ReLU FNNs under the non-parametric regression framework. Given n noisy samples, the two Theorems suggest the specific orders of $\mathcal{W}$, L and $\mathcal{N}$ in terms of n, d and r, for which they give the tightest bound on excess risk of respective function estimator from Proposition 5.3.2. Under the high-dimensional setting where $n = \mathcal{O}(d^q)$ for $0 < q < 1$, when $r = \mathcal{O}(1)$, it is shown that the excess risk upper-bounds of both function estimators have d^d in the constant factors. In contrast, when $r = \mathcal{O}(d)$, estimating functions $f \in W_{\infty}^{r}(\mathcal{S}^{d-1})$ has at most $d^{\mathcal{O}(1)}$ factor in the bound, whereas the bound for function estimator of $f \in W_{\infty}^{r}(\mathcal{C}^{d})$ has d^d. See Table 5.1

[1] Interested readers can find the intuitive technical reason for having the exponential dependence in d on width $\mathcal{W}$ and active parameters $\mathcal{N}$ in the Appendix E.1.

and Subsection 5.3.2 for detailed comparisons.

5.0.2 Related works

In this subsection, to aid readers have more clear understandings on the contributions of our paper, we provide the list of relevant works with comparisons on how these works are different from ours.

Approximation of $f \in W^r_\infty(\mathcal{S}^{d-1})$ ***via deep CNN.*** For the approximation theory of $f \in W^r_\infty(\mathcal{S}^{d-1})$, we must refer readers [179] and [180]. But in their works, the convolutional neural network (CNN) is used for the function approximator under fixed d setting.

Approximation of $f \in W^r_\infty(\mathcal{C}^d)$ ***via deep ReLU FNN.*** Approximation theory of deep ReLU FNN for functions $f \in W^r_\infty(\mathcal{C}^d)$ has a lengthy history in the literature. Representatively, [181] showed that f can be approximated uniformly within ε-approximation accuracy with a 1-layer neural network of $\mathcal{O}(\varepsilon^{-d/r})$ neurons and an infinitely differentiable activation function. Later, for deep ReLU networks, [182] showed that the number of active parameters ($\mathcal{N}$) in networks is bounded by $\mathcal{O}(\varepsilon^{-d/r}\log\left(\frac{1}{\varepsilon}\right))$, and the depth has the order $\mathcal{O}(\log(\frac{1}{\varepsilon}))$. He further proved that $\mathcal{N}$ is lower-bounded by the order $\mathcal{O}(\varepsilon^{-d/r})$, which is backed up by the result in [183]. For $f \in W^r_p(\mathcal{C}^d)$ with $1 \leq p \leq \infty$, [184] showed that there exists a deep ReLU network with bounded and quantized weight parameters, with $\mathcal{O}(\varepsilon^{-d/r})$ network size, and with ε-independent depth for achieving the ε-accuracy in the L_p norm. For approximating functions $f \in W^r_\infty(\mathcal{C}^d)$, [8] proved that a network of size $\mathcal{O}(\varepsilon^{-d/r})$ with bounded weight parameters achieves ε-approximation error in the L_∞ norm.

Function spaces with special structures. The result of [182] implies that deep ReLU net cannot escape the curse of dimensionality for approximating $f \in W^r_\infty(\mathcal{C}^d)$. Many papers have demonstrated that the effects of dimension can be either avoided or lessened by considering function spaces different from Sobolev spaces, but defined over $\mathcal{C}^d$. Just to name a few, [185] studied that a function with a compositional structure with regularity r can be approximated by neural network with $\mathcal{O}(\varepsilon^{-2/r})$ neurons within ε accuracy. [186] proved the deep ReLU network with $\mathcal{O}(\varepsilon^{-1/r})$ neurons can avoid the curse for approximating functions in mixed smooth Besov spaces. [187] showed the network size scales as $\mathcal{O}(\varepsilon^{-D/r})$ for approximating C^r functions, when they are defined on a Riemannian manifold isometrically embedded in $\mathbb{R}^d$ with manifold dimension D with $D \ll d$. [188] and [189] showed respectively the deep and shallow ReLU network break the curse for Korobov spaces.

Estimation rates of excess risk under non-parametric framework. Many researchers also have tried to tackle how the neural networks avoid the curse by considering specially designed function spaces under the non-parametric regression framework. We only provide the incomplete list of them. Such structures include additive ridge functions [9], composite function spaces with hierarchical structures [8, 190], mixed-Besov spaces [186], Hölder spaces defined over a lower-dimensional manifold embedded in $\mathbb{R}^d$ [191]. They all showed the function estimators with neural network architectures can lessen the curse by showing the excess risks of the estimators are bounded by $\mathcal{O}(n^{-2r/(2r+D')})$, where n denotes the size of a noisy dataset, and $D' \ll d$ is an intrinsic dimension uniquely determined through the characteristics of function spaces, when they are compared with the minimax risk $\mathcal{O}(n^{-2r/(2r+d)})$ [192] for $f \in W^r_\infty(\mathcal{C}^d)$.

Comparisons. The aforementioned works mainly focused on the approximation and estimation of functions defined on $\mathcal{C}^d$, not $\mathcal{S}^{d-1}$, for the fixed d. Moreover, the introduced papers on approximation theory, except the work of [8], hide the dependence on d in the Big-$\mathcal{O}$ notation of $\mathcal{N}$ in ε-accuracy, even for papers where they consider the function spaces with special structures. Thus, it is not clear how the d affects the approximation bound and the scale of the provided network architecture. Introduced papers on estimation rate for excess risk also follow the same philosophy with papers on approximation theory, as they work on the fixed d setting. In contrast, we work on the $\mathcal{S}^{d-1}$ input space, track the explicit dependence on d in the error bound, and describe how d affects the scale of deep ReLU FNN as $d \to \infty$ with its interactions with function smoothness $r > 0$. Our paper focuses on tracking the dependence on d in the constant factor hidden in the Big-$\mathcal{O}$ notations both in approximation and estimation error rates, rather than paying attentions on reducing the exponential dependence of d with base ε in $\mathcal{N}$ or with base n in excess risk bound.

5.1 Preliminary Definitions

In this section, we provide the mathematical definitions of deep ReLU FNN and Sobolev function spaces on unit sphere.

5.1.1 Definition of Deep ReLU network

For defining the deep ReLU network mathematically, we adopt the notation used in [8]. For $\mathbf{v} = (\mathbf{v_1}, \ldots, \mathbf{v_r}) \in \mathbb{R}^r$, let $\sigma_{\mathbf{v}} : \mathbb{R}^r \to \mathbb{R}^r$ be the shifted ReLU (Rectified Linear Units) activation function as $\sigma_{\mathbf{v}}((y_1, \ldots, y_r)^\top) := \sigma((y_1 - \mathbf{v}_1, \ldots, y_r - \mathbf{v}_r)^\top)$, where $\sigma(x) = max(x, 0)$.

With this notation, the network architecture $(L, \mathbf{p})$ consists of a positive integer L, called *the number of hidden layers*, and *a width vector* $\mathbf{p} := (\mathbf{p}_0, \ldots, \mathbf{p}_{L+1}) \in \mathbb{N}^{L+2}$. A deep ReLU network with architecture $(L, \mathbf{p})$ considered in this work is then any function of the form

$$\tilde{f} : \mathcal{S}^{d-1} \to \mathbb{R}, \quad \mathbf{x} \to f(\mathbf{x}) = W_L \sigma_{\mathbf{v_L}} W_{L-1} \sigma_{\mathbf{v_{L-1}}} \cdots \sigma_{\mathbf{v_1}} W_1 \mathbf{x}, \tag{5.1}$$

where $\mathbf{W}_i \in \mathbb{R}^{p_{i+1} \times p_i}$ is a weight matrix with $\mathbf{p}_0 = d$, $\mathbf{p}_{L+1} = 1$ and $\mathbf{v}_i \in \mathbb{R}^{p_i}$ is a shift vector. Network functions are built by alternating matrix-vector multiplications with the action of the nonlinear activation function σ.

Let $\|\mathbf{W}_j\|_0$ and $|\mathbf{v}_j|_0$ be the number of nonzero entries of $\mathbf{W}_j$ and $\mathbf{v}_j$ in the j^{th} hidden layer. The final form of neural network we consider in this paper is given by

$$\mathcal{F}(L, \mathbf{p}, \mathcal{N}) := \left\{ \tilde{f} \text{ of the form (5.1)} : \sum_{j=1}^{L} \|\mathbf{W}_j\|_0 + |\mathbf{v}_j|_0 \leq \mathcal{N} \right\}. \tag{5.2}$$

The main advantage of using this notation comes from its convenience for tracking the construction process of network $\tilde{f}$ for approximating $f \in W_{\infty}^{r}(\mathcal{S}^{d-1})$. See Section E.2.2 in the Appendix. Now, we define the Sobolev spaces over sphere in the next subsection.

5.1.2 Definition of Sobolev Spaces over Sphere

For $1 \leq p \leq \infty$, we denote $\mathcal{L}_p(\mathcal{S}^{d-1}) = \mathcal{L}_p(\mathcal{S}^{d-1}, \rho_{\mathcal{X}})$ as the $\mathcal{L}_p$-function space defined with respect to the normalized Lebesgue measure $\rho_{\mathcal{X}}$ on $\mathcal{S}^{d-1}$, with norm $\|g\|_p := \big(\int_{\mathcal{S}^{d-1}} |g(\mathbf{x})|^p \rho_{\mathcal{X}}(d\mathbf{x})\big)^{1/p}$.

Let $\mathcal{H}_k^d$ be the space of homogeneous harmonic polynomials of total degree $k \in \mathbb{Z}_+$ restricted on $\mathcal{S}^{d-1} \subset \mathbb{R}^d$. In [193, 194], its dimension for $k \in \mathbb{N}$ is found to be

$$\mathcal{N}(k,d) = \frac{2k+d-2}{k}\binom{k+d-3}{k-1}. \tag{5.3}$$

Note that $\mathcal{L}_2(\mathcal{S}^{d-1})$ is a Hilbert space with inner product $\langle f, g\rangle_{\mathcal{L}_2(\mathcal{S}^{d-1})} := \int_{\mathcal{S}^{d-1}} f(\mathbf{x})g(\mathbf{x})\rho_{\mathcal{X}}(\mathbf{x})$ for $f, g \in \mathcal{L}_2(\mathcal{S}^{d-1})$. The spaces $\mathcal{H}_k^d$, for $k \in \mathbb{Z}_+$, of spherical harmonics are mutually orthogonal with respect to the inner product of $\mathcal{L}_2(\mathcal{S}^{d-1})$. Since the space of spherical polynomials is dense in $\mathcal{L}_2(\mathcal{S}^{d-1})$, every $f \in \mathcal{L}_2(\mathcal{S}^{d-1})$ has a spherical harmonic expansion

$$f = \sum_{k=0}^{\infty} \mathbf{Proj}_k(f) = \sum_{k=0}^{\infty} \sum_{\ell=1}^{\mathcal{N}(k,d)} \widehat{f}_{k,\ell}\mathbf{Y}_{k,\ell} \tag{5.4}$$

converging in the $\mathcal{L}_2(\mathcal{S}^{d-1})$ norm. Hereafter, $\{\mathbf{Y}_{k,\ell}\}_{\ell=1}^{\mathcal{N}(k,d)}$ denotes an orthonormal basis of $\mathcal{H}_k^d$, $\widehat{f}_{k,\ell}$ is the Fourier coefficients of f given by

$$\widehat{f}_{k,\ell} := \langle f, \mathbf{Y}_{k,\ell}\rangle_{\mathcal{L}_2(\mathcal{S}^{d-1})} := \int_{\mathcal{S}^{d-1}} f(\mathbf{x})\mathbf{Y}_{k,\ell}(\mathbf{x})\rho_{\mathcal{X}}(d\mathbf{x}),$$

and $\mathbf{Proj}_k(f)$ denotes the orthogonal projection of $\mathcal{L}_2(\mathcal{S}^{d-1})$ onto $\mathcal{H}_k^d$, which has an integral representation

$$\mathbf{Proj}_k(f)(\mathbf{x}) := \int_{\mathcal{S}^{d-1}} f(\mathbf{y})\mathcal{Z}_k(\mathbf{x},\mathbf{y})\rho_{\mathcal{X}}(d\mathbf{y}), \quad \forall \mathbf{x} \in \mathcal{S}^{d-1},$$

where

$$\mathcal{Z}_k(\mathbf{x},\mathbf{y}) := \sum_{\ell=1}^{\mathcal{N}(k,d)} \mathbf{Y}_{k,\ell}(\mathbf{x})\mathbf{Y}_{k,\ell}(\mathbf{y}), \quad \forall \mathbf{x},\mathbf{y} \in \mathcal{S}^{d-1}.$$

We know that $\mathcal{Z}_k(\mathbf{x},\mathbf{y})$ is a reproducing kernel of $\mathcal{H}_k^d$, independent of the choice of $\big\{\mathbf{Y}_{k,\ell}\big\}_{\ell=1}^{\mathcal{N}(k,d)}$, and with $\lambda_{\mathrm{G}} = \frac{d-2}{2}$,

$$\mathcal{Z}_k(\mathbf{x},\mathbf{y}) := \frac{N+\lambda_{\mathrm{G}}}{\lambda_{\mathrm{G}}}\mathcal{G}_k^{\lambda_{\mathrm{G}}}\big(\langle\mathbf{x},\mathbf{y}\rangle\big), \quad \forall \mathbf{x},\mathbf{y} \in \mathcal{S}^{d-1} \tag{5.5}$$

where $\mathcal{G}_k^{\lambda_{\mathrm{G}}}$ is the Gegenbauer polynomial of degree k with parameter $\lambda_{\mathrm{G}} > -\frac{1}{2}$, see for instance [193]. Denote $u := \langle\mathbf{x},\mathbf{y}\rangle$, the exact expression of $\mathcal{G}_k^{\lambda_{\mathrm{G}}}\big(u\big)$ is given in terms of the Gamma function by

$$\mathcal{G}_k^{\lambda_{\mathrm{G}}}\big(u\big) := \sum_{\ell=0}^{\lfloor \frac{k}{2} \rfloor}(-1)^\ell \frac{\Gamma\big(k-\ell+\lambda_{\mathrm{G}}\big)}{\Gamma\big(\lambda_{\mathrm{G}}\big)\ell!\big(k-2\ell\big)!}\big(2u\big)^{k-2\ell}. \tag{5.6}$$

The space of $\mathcal{H}_k^d$ of spherical harmonics can also be characterized as eigenfunction spaces of the Laplace-Beltrami operator $\Delta_{\mathcal{S}^{d-1}}$ on $\mathcal{S}^{d-1}$. Indeed,

$$\mathcal{H}_k^d = \bigg\{ f \in \mathcal{C}^2\big(\mathcal{S}^{d-1}\big) : \Delta_{\mathcal{S}^{d-1}} f = -\lambda_k f \bigg\},$$

where $\lambda_k = k(k+d-2)$ and $\mathcal{C}^2\big(\mathcal{S}^{d-1}\big)$ denotes the space of all twice continuously differentiable functions on $\mathcal{S}^{d-1}$. In fact, with the identity operator $\mathcal{I}$, we may define the fractional power of $\big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)^\alpha$ of the operator $\big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)$ in a distributional sense for

$\alpha \in \mathbb{R}$:

$$\mathbf{Proj}_k\big(\big(-\Delta_{\mathcal{S}^{d-1}}+\mathcal{I}\big)^{\alpha} f\big)=\big(1+\lambda_k\big)^{\alpha}\mathbf{Proj}_k\big(f\big).$$

Now, we define the Sobolev space $W_p^r(\mathcal{S}^{d-1})$ to be the subspace of $\mathcal{L}_p(\mathcal{S}^{d-1})$ for $1 \leq p \leq \infty$, $r > 0$, with the finite norm

$$\|f\|_{W_p^r(\mathcal{S}^{d-1})}=\left\|\bigg(-\Delta_{\mathcal{S}^{d-1}}+\mathcal{I}\bigg)^{r/2} f\right\|_p<\infty. \tag{5.7}$$

The sphere $\mathbb{S}^{d-1}$ is a smooth Riemannian manifold without boundary. Its nice Laplace-Beltrami operator (i.e.,$\Delta_{\mathcal{S}^{d-1}}$) acting as a Hessian operator of functions on the sphere gives the natural definition of Sobolev spaces $W_\infty^r(\mathbb{S}^{d-1})$ in (5.7); that is, the Sobolev space is a collection of continuous functions defined on sphere $\mathbb{S}^{d-1}$ whose generalized (distributional) derivatives up to order r are essentially bounded. See Equations (16) in [195], (3.4) in [179], (16) in [180], (5.1.9) in [196] for more detailed treatments on $W_\infty^r(\mathbb{S}^{d-1})$. Readers can also refer the definition of $W_\infty^r(\mathcal{C}^d)$ in the Appendix E.1, when $\mathcal{C}^d = [0,1]^d$, for comparison with $W_\infty^r(\mathcal{S}^{d-1})$ and later use in Subsection 5.2.1.

5.2 Approximation error

Now, we present our Theorem on approximating functions $f \in W_\infty^r(\mathcal{S}^{d-1})$ via $\mathcal{F}(L, \mathbf{p}, \mathcal{N})$ in (5.2).

Theorem 5.2.1 *Let* $0 < \alpha < 1, m, N, M \in \mathbb{N}$ *with* $1 \leq N \leq d^\alpha + 1$. *For any function* $f \in W_\infty^r(\mathcal{S}^{d-1})$ *with* $r > 0$*, there exists a network*

$$\tilde{f} \in \mathcal{F}\big(L, \big(d, 22NM, \ldots, 22NM, 1\big), \mathcal{N}\big) \tag{5.8}$$

with depth $L = (m+4)\lceil \log_2(2N) \rceil$ *and number of parameters* $\mathcal{N} \leq M(2d + 404N \cdot (m+$

$3) + 2N + 4) + 1$ *such that*

$$\left\|f - \tilde{f}\right\|_\infty \leq C_\eta^{''} \|f\|_{W_\infty^r(\mathcal{S}^{d-1})} \times \max\left\{N^{-r}, \frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{N+\frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}}, d^{2N}\left(\log_2(2N)\right)^2 2^{-2m}\right\}, \quad (5.9)$$

where $C_\eta^{''}$ is a constant dependent on η, and independent on d, r, N, M or f.

The proof of Theorem 5.2.1 is lengthy and technical. We provide the detailed proof ideas with technical remarks for the Lemmas and Proposition used for the proof of Theorem 5.2.1 in the Appendix E.2. The detailed technical proofs of those Lemmas and Proposition are provided in the Appendix E.3. Here, for conciseness, we provide some important remarks on the Theorem and a simple proof sketch, where it starts with a simple triangle inequality:

$$\left\|f - \tilde{f}\right\|_\infty \leq \|f - L_N(f)\|_\infty + \left\|L_N(f) - \widehat{L}_{N,M}^{\boldsymbol{y}}(f)\right\|_\infty + \left\|\widehat{L}_{N,M}^{\boldsymbol{y}}(f) - \tilde{f}\right\|_\infty. \quad (5.10)$$

Three error terms in (5.9) correspond to the bounds on three terms of right-hand side in the inequality (5.10). We want to emphasize that the constant $C_\eta^{''} > 0$ in (5.9) is independent of d. Furthermore, we track how the bound is explicitly dependent on d allowing it to tend to infinity.

For first term, note that any $f \in W_\infty^r(\mathcal{S}^{d-1})$ is approximated by a weighted sum of $\mathbf{Proj}_k(f)$ for $0 \leq k \leq 2N$, denoted as $L_N(f)$. The corresponding approximation error is small for large enough N and r. Here, importantly, we set the $N = \lceil d^\alpha \rceil$ for $0 < \alpha < 1$, so that the input dimension d grows faster than N.

For second term, notice that the definition of $L_N(f)$ is involved with the integral over the sphere, and the key for approximating the function is to discretize this integral by M random samples $\mathbf{y} = \{\mathbf{y_1}, \ldots, \mathbf{y_M}\}$ independently drawn from $\rho_{\mathcal{X}}$. The discretized version of $L_N(f)$ is denoted as $\widehat{L}_{N,M}^{\boldsymbol{y}}(f)$. As observed in the error bound, the higher degree N the $L_N(f)$ has, the more sampled points M the approximation requires. However, the requirement is ameliorated as r increases. Similar effect can be observed in the constant factor in d. The higher the data dimension d is, the more the sampled point M is required for good approximation, but the requirement is alleviated as the smoothness index r increases. If r increases up to order d, the factor [2] decays exponentially fast as $d \to \infty$, eventually letting $M \geq 1$ to be any integer. This phenomena is further investigated in the Corollary 5.2.3.

The last term corresponds to the error of the neural network $\tilde{f}$ approximating $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)$. For any point $\mathbf{x} \in \mathcal{S}^{d-1}$, the evaluated function value $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)(\mathbf{x})$ is simply a weighted average of $\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$, for the sampled $\mathbf{y} = \{\mathbf{y_1}, \dots, \mathbf{y_M}\}$. Here, $\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$ is a linear combination of $\mathcal{G}_k^{\lambda_{\mathrm{G}}}\big(\langle \mathbf{x}, \mathbf{y}_i \rangle\big)$ in (5.6) for $0 \leq k \leq 2N$. Thus, it is sum of univariate polynomials of degree up to $2N$. We construct sub-networks approximating $\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$ for each $i \in [M]$. This explains the width of $\tilde{f}$ is proportional to NM. The corresponding error bound is dependent on d^{2N}, where it comes from the applications of Stirling's formula on the coefficient factors in $\mathcal{G}_k^{\lambda_{\mathrm{G}}}\big(\langle \mathbf{x}, \mathbf{y}_i \rangle\big)$. The error, $\big(\log_2(2N)\big)^2 2^{-2m}$, comes from approximating $\langle \mathbf{x}, \mathbf{y}_i \rangle^k$ for $0 \leq k \leq 2N$ via neural networks. The larger the m is, the deeper the network becomes as $L = \mathcal{O}(m)$, and the error gets smaller.

5.2.1 Comparison with [8]

In this subsection, we compare the result from Theorem 5.2.1 with the result from [8], where they consider the approximation of $f \in W^r_\infty([0,1]^d)$ via deep ReLU FNN. The Theorem is stated as follows:

Theorem 5.2.2 *[Theorem 5 of [8]] For any function $f \in W^r_\infty([0,1]^d)$ and let $K > 0$ be the radius of Hölder ball. Then, for any integers $m \geq 1$ and $N^H \geq (r+1)^d \vee (K+1)e^d$, there exists a network*

$$\tilde{f}^H \in \mathcal{F}^H\big(L, (d, 6(d + \lceil r \rceil)N^H, \dots, 6(d + \lceil r \rceil)N^H, 1), \mathcal{N}^H\big) \tag{5.11}$$

[2] If $r = \mathcal{O}(d)$, the factor becomes $\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{\lceil d^{\alpha} \rceil}$ for $0 < \alpha < 1$. Here, the exponential decay term $\left(\frac{6}{\pi e}\right)^{\frac{d}{4}}$ is derived from Sobolev embedding Lemma. See Proposition E.2.3 in Appendix E.2.

with depth $L = 8 + (m+5)\big(1 + \lceil \log_2(d \vee r) \rceil\big)$ *and the number of parameters* $\mathcal{N}^H \leq 141(1+d+r)^{3+d} N^H (m+6)$*, such that*

$$\left\| f - \tilde{f}^H \right\|_\infty \leq (2K+1)(1+d^2+r^2)6^d \big(N^H\big) 2^{-m} + K 3^r \big(N^H\big)^{-\frac{r}{d}}. \tag{5.12}$$

To avoid the confusion with the notations used in Theorem 5.2.1, we put the superscript H to a parameter that determines width of the network (i.e., N^H), to the total number of parameters in the network (i.e., $\mathcal{N}^H$), and to the network class (i.e., $\mathcal{F}^H$). It is interesting to note that the exponential growth of the network size in d is observed in the construction of $\mathcal{F}^H$, whereas there exists a flexibility in $\mathcal{F}$ dependent on the choice of M. Specifically, the width of the network in $\mathcal{F}^H$ is exponentially dependent on d as $N^H = \Omega(r^d \vee e^d)$, whereas the width of the network in $\mathcal{F}$ is dependent on two parameters $N = o(d)$ and any integers $M \geq 1$. For the total number of network parameters, we have $\mathcal{N}^H = \mathcal{O}((d+r)^d)$, whereas $\mathcal{N} = \mathcal{O}(Md + Nmd)$.

Analogously, the bound on the approximation error of $\tilde{f}^H$ in (5.12) is dependent on d exponentially, but this exponential dependence in d can be avoided in the error bound of $\tilde{f}$ in (5.9) under two scenarios: ***(1)*** $r = \mathcal{O}(d)$ ***and any integer*** $M \geq 1$ or ***(2)*** $r = \mathcal{O}(1)$ ***and*** $M = \mathcal{O}(d^d)$. In the Corollary presented in the next subsection, we further specify the two scenarios, and describe how the approximation error bound in each scenario converges to 0 in terms of d.

5.2.2 Fast Approximation error in terms of d

Corollary 5.2.3 *Let* $0 < \alpha, \beta, \gamma < 1$ *with* $\gamma > \max\{\alpha, \beta\}$ *and* $N \in \mathbb{N}$ *with* $1 \leq N \leq d^\alpha + 1$*. For any* $f \in W^r_\infty(\mathcal{S}^{d-1})$ *with* $r > 0$*, we have:*

(I) For $\frac{3d-2}{4} - C_1 \leq r \leq \frac{3d-2}{4}$ *with some constant* $C_1 \geq 0$ *independent of d, there exists a network*

$$\tilde{f}^{(1)} \in \mathcal{F}\left(L, (d, 66N, 66N, \ldots, 66N, 1), \mathcal{N}\right)$$

with depth $L = \mathcal{O}(d^{\gamma}\log_2 d)$ *and the number of active parameters* $\mathcal{N} = \mathcal{O}\left(d^{\max\{\alpha+\gamma,1\}}\right)$, *such that* $\left\|f - \tilde{f}^{(1)}\right\|_{\infty} \leq C'_{\eta,\alpha,\beta,\gamma}\|f\|_{W^r_{\infty}(\mathcal{S}^{d-1})}d^{-d^{\beta}}$, *where* $C'_{\eta,\alpha,\beta,\gamma}$ *is a constant depending only on* C_1, η, α, β *and* γ.

(II) For $r = \mathcal{O}(1)$ *and* $M = \mathcal{O}\left(9^d d^{\frac{9}{4}d}\right)$, *there exists a network*

$$\tilde{f}^{(2)} \in \mathcal{F}\big(L, \big(d, 22NM, \ldots, 22NM, 1\big), \mathcal{N}\big)$$

with depth $L = \mathcal{O}(d^{\gamma}\log_2 d)$ *and the number of active parameters* $\mathcal{N} = \mathcal{O}\left(9^d d^{\frac{13}{4}d}\right)$ *such that* $\left\|f - \tilde{f}^{(2)}\right\|_{\infty} \leq C'_{\eta,\alpha,\beta,\gamma}\|f\|_{W^r_{\infty}(\mathcal{S}^{d-1})}d^{-\alpha r}$, *where* $C'_{\eta,\alpha,\beta,\gamma}$ *is a constant depending only on* η, α, β *and* γ.

The detailed proof on Corollary 5.2.3 is deferred in the Appendix E.3.6. The approximation error in scenario (1) decays at a rate $d^{-d^{\beta}}$ for $0 < \beta < 1$, while the required number of active parameters $\mathcal{N}$ is at most $\mathcal{O}(d^2)$. Here, the construction of network $\tilde{f}^{(1)}$ is independent with the choice of M, and we simply choose $M = 3$. In scenario (2), since $r = \mathcal{O}(1)$ and $0 < \alpha < 1$, the approximation error decays to 0 at $d^{-\mathcal{O}(1)}$ rate, which can be slower than $d^{-d^{\beta}}$ for β close to 1. The width of $\tilde{f}^{(2)}$ grows exponentially in d requiring $M = \mathcal{O}(d^d)$. Interestingly, in both scenarios, the depth L has the same order in d as $\mathcal{O}(d^{\gamma}\log_2 d)$ for $0 < \gamma < 1$.

5.3 Statistical risk bound

Let $\mathcal{X} := \mathcal{S}^{d-1}$ and $\mathcal{Y} \subset \mathbb{R}$ be the measureable feature space and output space. We denote ρ as a joint probability measure on the product space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and let $\rho_{\mathcal{X}}$ be the marginal distribution of the feature space $\mathcal{X}$. We assume that the noisy data set $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ are generated from the non-parametric regression model

$$\mathbf{y}_i = f_{\rho}(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{5.13}$$

where the noise ε_i is assumed to be centered sub-gaussian random variable and $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = 0$. Our goal is to estimate the regression function $f_\rho(\mathbf{x})$ with the given noisy data set $\mathcal{D}$. Specifically, it is assumed that the regression function belongs to Sobolev space on d-dimensional sphere; that is $f_\rho \in W^r_\infty(\mathcal{S}^{d-1})$. It is easy to see regression function $f_\rho := \mathbb{E}(\mathbf{y}|\mathbf{x})$ is a minimizer of the following population risk $\mathcal{E}(f)$ defined as:

$$\mathcal{E}(f) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\rho}\left[\big(\mathbf{y} - f(\mathbf{x})\big)^2\right].$$

However, since the joint distribution ρ is unknown, we cannot find f_ρ directly. Instead, we solve a following empirical risk minimization problem induced from the dataset $\mathcal{D}$:

$$\widehat{f}_n = \underset{f\in\mathcal{F}(L,\mathbf{p},\mathcal{N})}{\operatorname{argmin}}\ \mathcal{E}_D(f) := \underset{f\in\mathcal{F}(L,\mathbf{p},\mathcal{N})}{\operatorname{argmin}} \left\{\frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{y}_i - f(\mathbf{x}_i)\big)^2\right\}. \tag{5.14}$$

Note that the function estimator is taken from the feedforward neural network hypothesis space $\mathcal{F}(L, \mathbf{p}, \mathcal{N})$[3] defined in (5.2), and we denote the empirical minimizer of (5.14) as $\widehat{f}_n$. It is assumed that $|\mathbf{y}| \leq B$ almost everywhere and we have $|f_\rho(\mathbf{x})| \leq B$. We project the output function $f : \mathcal{S}^{d-1} \to \mathbb{R}$ onto the interval $[-B, B]$ by a projection operator

$$\pi_B f(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \text{if } -B \leq f(\mathbf{x}) \leq B, \\ B, & \text{if } f(\mathbf{x}) > B, \\ -B, & \text{if } f(\mathbf{x}) < -B. \end{cases} \tag{5.15}$$

We consider the clipped estimator $\pi_B\widehat{f}_n$ for recovering the regression function f_ρ. Note that the clipped estimator has been widely used in statistical learning papers [186, 9, 197]. The quality of $\pi_B\widehat{f}_n$ is measured through the difference between two expected risks (i.e., excess risk) defined as $\mathcal{E}\big(\pi_B\widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big)$.

[3] Henceforth, we will use a shorthand notation of $\mathcal{F}(L, \mathbf{p}, \mathcal{N})$ as $\mathcal{F}$. Dependence on $(L, \mathbf{p}, \mathcal{N})$ should be implicitly understood.

5.3.1 Upper-bound on excess risk

In this Subsection, we provide the upper-bound on the excess risk of the clipped estimator $\pi_B(\widehat{f}_n)$ with respect to the pseudo-dimension (i.e., Pdim$(\mathcal{F})$) and the approximation error (i.e., $\|f - f_\rho\|_\infty$). Before presenting the bound, the definition of Pdim$(\mathcal{F})$ is presented.

Definition 5.3.1 *Denote by Pdim$(\mathcal{F})$, the pseudo-dimension of $\mathcal{F}$, which is the largest integer ℓ, for which there exists $(\xi_1, \ldots, \xi_\ell, \eta_1, \ldots, \eta_\ell) \in \mathcal{X}^\ell \times \mathbb{R}^\ell$ such that for any $(a_1, \ldots, a_\ell) \in \{0, 1\}^\ell$, there is some $f \in \mathcal{F}$ satisfying*

$$\forall i : f(\xi_i) > \eta_i \Longleftrightarrow a_i = 1.$$

For more comprehensive exploration on Pdim$(\mathcal{F})$ can be found in references [178, 198]. We provide the first theorem on the excess risk.

Proposition 5.3.2 *Set $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, we have*

$$\mathcal{E}\big(\pi_B \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big) \leq C_{B,\delta,f} \cdot \left(\frac{\textit{Pdim}(\mathcal{F}) \cdot \log(n)}{n} + \frac{\|f - f_\rho\|_\infty}{\sqrt{n}} + \|f - f_\rho\|_\infty^2 \right), \quad (5.16)$$

where $C_{B,\delta,f}$ is an absolute constant dependent on B, δ, f independent on n, r, d.

A detailed proof of Proposition 5.3.2 is deferred in the Appendix. The excess risk $\mathcal{E}\big(\pi_B \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big)$ is a random quantity over the estimator $\widehat{f}_n$ and the statement in the Theorem holds with probability at least $1 - \delta$. The failure probability $\delta \in (0, 1)$ is hidden in the constant $C_{B,\delta,f}$ logarithmically, i.e., $\log(\frac{1}{\delta})$. In the bound, it should be noted that there is a trade-off between the ***"approximation error"*** (i.e., $\|f - f_\rho\|_\infty$) term and the combinatorial ***"complexity measure"*** term of a neural network class $\mathcal{F}$ (i.e., Pdim$(\mathcal{F}) \cdot \log(n)/n$); that is, the richer the network hypothesis space $\mathcal{F}$ becomes, the finer the approximation result we get. Nonetheless, the arbitrary increase in the hypothesis space $\mathcal{F}$ eventually leads the increase of the bound in excess risk. In the following subsection, we will show how the specifications (i.e., the choices of $(L, \mathbf{p}, \mathcal{N})$) of the network architecture affect the tension between these two terms.

5.3.2 Convergence Rate of Excess Risk

Now we are ready to formally state bounds on the excess risks of $\pi_M \widehat{f}_n$ when $f_\rho \in W^r_\infty(\mathcal{S}^{d-1})$ (i.e., Theorem 5.3.3) and $f_\rho \in W^r_\infty([0,1]^d)$ (i.e., Theorem 5.3.4), respectively.

Theorem 5.3.3 *Suppose $f_\rho \in W^r_\infty(\mathcal{S}^{d-1})$ with $r > 0$. A network $\widehat{f}_n$ from* (5.8) *with choices $N = \lceil n^{\frac{2}{3d+4r}} \rceil$, $M = \lceil n^{\frac{3d}{3d+4r}} \rceil$, and $m = \lceil \frac{r}{3d+4r} \log_2(n) \rceil$ yield the bound on the excess risk with probability at least $1 - \delta$ as follows:*

$$\begin{aligned}&\mathcal{E}\big(\pi_M \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big) \\ &\leq \mathcal{C}_{B,\eta,\delta,f} \cdot \max\left\{1, \frac{6rd}{(3d+4r)^2}(\log_2(n))^4, \big(\frac{6}{\pi e}\big)^{\frac{d}{2}} d^{2N + \frac{3d-4r-2}{4}}, d^{4N}\right\} \cdot n^{-\frac{2r}{2r+1.5d}},\end{aligned} \tag{5.17}$$

where $\mathcal{C}_{B,\eta,\delta,f}$ depends on B, η, δ, f and independent on d, r and n.

Theorem 5.3.4 *Suppose $f_\rho \in W^r_\infty([0,1]^d)$ with $r > 0$. A network $\widehat{f}_n$ from* (5.11) *with choices $N^H = \lceil n^{\frac{d}{2d+r}} \rceil$, and $m^H = \lceil \frac{d+r}{d+2r} \log_2(n) \rceil$ yield the bound on the excess risk with probability at least $1 - \delta$ as follows:*

$$\begin{aligned}&\mathcal{E}\big(\pi_M \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big) \\ &\leq \mathcal{C}_{B,\eta,\delta,K} \cdot \max\left\{\lceil \log_2(d + \lceil r \rceil) \rceil^2 (d+r)^d \cdot (\log_2(n))^3, \big(1 + r^2 + d^2\big)^2 6^{2d} + 3^{2r}\right\} \cdot n^{-\frac{2r}{2r+d}},\end{aligned} \tag{5.18}$$

where $\mathcal{C}_{B,\eta,\delta,K}$ depends on B, η, δ, K and independent on d, r and n.

Detailed proofs on Theorems 5.3.3 and 5.3.4 are deferred in the Appendix E.4.2 and E.4.3. Both proofs are simple applications of Proposition 5.3.2 with results from Theorem 5.2.1 and 5.2.2. For both cases, Pdim($\mathcal{F}$) can be easily computed from Lemma E.5.1 in the Appendix. The parameters that determine the network architectures, N, M, m and N^H, m^H

in two Theorems are chosen in a way that the bound in (5.16) is tight in terms of sample size n. Constant factors $\mathcal{C}_{B,\eta,\delta,f}$ and $\mathcal{C}_{B,\eta,\delta,K}$ are dependent on $\delta \in (0,1)$ as $\log(\frac{1}{\delta})$. The bound in Theorem 5.3.3, $\mathcal{O}_d(n^{-\frac{2r}{2r+1.5d}})$, is sub-optimal in a minimax sense for estimating functions $f_\rho \in W^r_\infty(\mathcal{S}^{d-1})$, where $\mathcal{O}_d$ hides the constant factor in d. The extra $0.5d$ factor in the denominator of exponent comes from the Sobolev embedding Lemma (Lemma E.2.3) and discretization Lemma (Lemma E.2.4). For the constant factor in d, when $r = \mathcal{O}(1)$, the exponential dependence on d can be observed. However, when $r = \mathcal{O}(d)$, the excess bound in (5.17) reduces to $\mathcal{E}\left(\pi_M \widehat{f}_n\right) - \mathcal{E}\left(f_\rho\right) \leq \mathcal{C}_{B,\eta,\delta,f} \cdot \max\left\{(\log_2(n))^4, d^{4N}\right\} \cdot n^{-\frac{2r}{2r+1.5d}}$, with $N = \lceil n^{\frac{2}{3d+4r}} \rceil$. Specifically, in high-dimensional setting where $n = \mathcal{O}(d^q)$ for $0 < q < 1$, the constant in d becomes $d^{d^{\frac{2q}{3d+4r}}}$. Then, as $d, r \to \infty$, the constant d^{4N} becomes $d^{\mathcal{O}(1)}$. In contrast, in (E.44) for estimating functions $f_\rho \in W^r_\infty([0,1]^d)$, the rate $n^{-\frac{2r}{2r+d}}$ is minimax optimal, but we cannot observe the interactions between r and d as we observe in (5.17).

Remark 5.3.5 *From the technical point of view, the result in Theorem 5.3.3 should be compared with the results in the existing literature, i.e., [8, 191, 186], in a sense that our result doesn't require the boundedness of the weight parameters in the network construction. The detailed readings of their proofs reveal that they require the bound on the uniform covering number of $\mathcal{F}$ and it can be bounded by the Lipschitzness of the network output with respect to the weight parameters. Naturally, for the discretizations of the parameter space, the boundedness assumption is required. In contrast, in our result, due from the [198] (See Lemma E.5.1), bounding the complexity measure Pdim$(\mathcal{F})$ doesn't require the parameter boundedness assumption.*

5.4 An Open Question

In this paper, we prove when $r = \mathcal{O}(d)$, deep ReLU FNNs only require at most $\mathcal{N} = \mathcal{O}(d^2)$ parameters to get a sharp approximation rate. However, this condition seems restrictive, and needs further investigation whether it is a necessary and sufficient condition to avoid the curse of dimensionality for approximating $f \in W^r_\infty(\mathcal{S}^{d-1})$. To answer this question, it is essential to study the lower bound of $\mathcal{N}$ with a similar approximation error as stated in Theorem 5.2.1, and see if it has the matching order with the upper-bound we get in d. We conjecture obtaining this result is possible by combining the ideas of using VC-dimension of deep ReLU FNNs [198, 182] and of constructing the packing set on the sphere through the spherical cap [195], while tracking the d-dependency in the constant factor carefully. We leave this for the future research.

Appendices

APPENDIX A
A NETWORK MODEL THAT COMBINES LATENT FACTORS AND SPARSE GRAPHS

A.1 Computation

For the convenience of readers, we rewrite the optimization problem that we want to solve:

$$\min_{\substack{\alpha \in R, S=S^T \\ L \succcurlyeq 0}} -\frac{1}{n} \log \prod_{1 \leq i < j \leq n} \frac{\exp\left(X_{ij}\left(\alpha + L_{ij} + S_{ij}\right)\right)}{1 + \exp\left(\alpha + L_{ij} + S_{ij}\right)} + \delta \|L\|_* + \gamma \|S\|_1. \tag{A.1}$$

We propose a method that takes advantage of the special structure of the L_1 and the nuclear norm by means of the alternating direction method of multiplier (ADMM), which is a method that has recently gained momentum. An examination of the objective function in (A.1) unvails that terms

$$\alpha \sum_{1 \leq i < j \leq n} X_{ij} + \frac{1}{2} X \bullet L + \frac{1}{2} X \bullet S$$

are linear in α, L, and S. The term

$$\sum_{1 \leq i < j \leq n} \log\left(1 + e^{\alpha + L_{ij} + S_{ij}}\right)$$

is convex with respect to α, L, and S. Functions $\|S\|_1$ and $\|L\|_*$ are known to be convex functions. Therefore, the objective function in (A.1) is convex. The above convex optimization problem can be solved via ADMM as follows.

A.1.1 ADMM approach

We give a review of the alternating direction method of multiplier (ADMM). Consider two closed convex functions

$$f : \chi_f \to \mathbb{R} \text{ and } g : \chi_g \to \mathbb{R},$$

where the domain χ_f and χ_g of functions f and g are closed convex subsets of $\mathbb{R}^d$, and $\chi_f \bigcap \chi_g$ is nonempty. Both f and g are possibly non-differentiable. The alternating direction method of multiplier is an iterative algorithm that solves the following generic optimization problem:

$$\min_{x \in \chi_f \bigcap \chi_g} \left\{ f(x) + g(x) \right\},$$

or equivalently

$$\begin{aligned} \min_{x \in \chi_f, z \in \chi_g} \quad & \left\{ f(x) + g(z) \right\}, \\ \text{subject to} \quad & x = z. \end{aligned} \tag{A.2}$$

To describe the algorithm, we will need the following proximal operators

- $\mathbf{P}_{\lambda,f} : \mathbb{R}^d \to \chi_f$ as

$$\mathbf{P}_{\lambda,f}(v) = \text{arg min}_{x \in \chi_f} \left\{ f(x) + \frac{1}{2\lambda} \|x - v\|_2^2 \right\},$$

- and $\mathbf{P}_{\lambda,g} : \mathbb{R}^d \to \chi_g$ as

$$\mathbf{P}_{\lambda,g}(v) = \text{arg min}_{x \in \chi_g} \left\{ g(x) + \frac{1}{2\lambda} \|x - v\|_2^2 \right\},$$

  where $\| \cdot \|_2$ is the usual Euclidean norm on $\mathbb{R}^d$ and λ is a scale parameter that is a fixed positive constant.

The algorithm starts with some initial values $x^0 \in \chi_f, z^0 \in \chi_g, u^0 (= \lambda y^0) \in \mathbb{R}^d$. At the $(m+1)$th iteration, (x^m, z^m, u^m) is updated according to the following steps until convergence

- Step 1: $x^{m+1} = \mathbf{P}_{\lambda,f}(z^m - u^m)$,
- Step 2: $z^{m+1} = \mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$,
- Step 3: $u^{m+1} = u^m + x^{m+1} - z^{m+1}$.

The convergence properties of the algorithm are summarized in the following result as in [37]. Let p^* be the minimal value in (A.2).

Theorem A.1.1 (Boyd et al., 2011) *Assume functions $f : \chi_f \to \mathbb{R}$ and $g : \chi_g \to \mathbb{R}$ are closed convex functions, whose domains χ_f and χ_g are closed convex subsets of $\mathbb{R}^d$ and $\chi_f \bigcap \chi_g \neq \emptyset$. Assume the Lagrangian of* (A.2)

$$L(x, z, y) = f(x) + g(z) + y^T(x - z)$$

has a saddle point, that is, there exists (x^*, z^*, y^*) (not necessarily unique) that $x^* \in \chi_f$ and $z^* \in \chi_g$, for which

$$L(x^*, z^*, y) \leq L(x^*, z^*, y^*) \leq L(x, z, y^*), \qquad \forall x, z, y \in \mathbb{R}^d.$$

Then the ADMM has the following convergence properties.

1. *Residual convergence. $x^m - z^m \to 0$ as $m \to \infty$; i.e., the iterates approach feasibility.*
2. *Objective convergence. $f(x^m) + g(z^m) \to p^*$ as $m \to \infty$; i.e., the objective function of the iterates approaches the optimal value.*
3. *Dual variable convergence. $y^m \to y^*$ as $m \to \infty$, where y^* is a dual optimal point.*

Now we describe how ADMM can be adopted to solve for our penalized likelihood estimation problem in (A.1). We reparameterize $M = L + S$ and let $x = (\alpha, M, L, S)$ (viewed as a vector). We define the following:

$$\begin{aligned}
\chi_f &= \{(\alpha, M, L, S) : \alpha \in \mathbb{R}, M, L, S \in \mathbb{R}^{n\times n}, L \text{ is positive semidefinite}, S \text{ is symmetric}\},\\
f(x) &= -\frac{\alpha}{n}\sum_{1\le i<j\le n} X_{ij} - \frac{1}{2n} X \bullet M + \frac{1}{n}\sum_{1\le i<j\le n} \log\left(1 + e^{\alpha + M_{ij}}\right) + \gamma\|S\|_1 + \delta\|L\|_*,\\
\chi_g &= \{(\alpha, M, L, S) : \alpha \in \mathbb{R}, M, L, S \in \mathbb{R}^{n\times n}, M \text{ is symmetric and } M = L + S\}, \text{ and}\\
g(x) &= 0, \text{ for } x \in \chi_g.
\end{aligned}$$

One can verify that (A.1) can be written as

$$\min_{x\in\chi_f \bigcap \chi_g} \{f(x) + g(x)\}.$$

We now present each of the three steps of the ADMM algorithm and show that the proximal operators $\mathbf{P}_{\lambda,f}$ and $\mathbf{P}_{\lambda,g}$ are easy to evaluate. Let

$$x^m = (x^m_\alpha, x^m_M, x^m_L, x^m_S), \quad z^m = (z^m_\alpha, z^m_M, z^m_L, z^m_S), \quad u^m = (u^m_\alpha, u^m_M, u^m_L, u^m_S).$$

Step 1. We solve $x^{m+1} = \mathbf{P}_{\lambda,f}(z^m - u^m)$. Due to the special structure of $f(\cdot)$, $x^{m+1}_\alpha, x^{m+1}_M, x^{m+1}_L$, and x^{m+1}_S can be updated separately. More precisely, we have

$$\begin{aligned}
x^{m+1}_\alpha, x^{m+1}_M &= \text{arg min}_{\alpha,M} \quad -\frac{\alpha}{n}\sum_{1\le i<j\le n} X_{ij} - \frac{1}{2n} X \bullet M + \frac{1}{n}\sum_{1\le i<j\le n} \log\left(1 + e^{\alpha + M_{ij}}\right)\\
&\quad + \frac{1}{2\lambda}\left[\alpha - (z^m_\alpha - u^m_\alpha)\right]^2 + \frac{1}{2\lambda}\|M - (z^m_M - u^m_M)\|_F^2, \qquad \text{(A.3)}\\
x^{m+1}_L &= \text{arg min}_L \quad \delta\|L\|_* + \frac{1}{2\lambda}\|L - (z^m_L - u^m_L)\|_F^2, \qquad \text{(A.4)}\\
&\quad \text{subject to } L \text{ is positive semidefinite;}\\
x^{m+1}_S &= \text{arg min}_S \quad \gamma\|S\|_1 + \frac{1}{2\lambda}\|S - (z^m_S - u^m_S)\|_F^2, \qquad \text{(A.5)}\\
&\quad \text{subject to } S \text{ is symmetric,}
\end{aligned}$$

where $\|\cdot\|_F$ is the matrix Frobenius norm, defined as $\|M\|_F^2 = \sum_{i,j} m_{ij}^2$ for a matrix $M = \{(m_{ij})\}_{i,j=1}^n$. The problem in (A.3) may not have a closed-form solution. We use a simple gradient descent to solve in this step, setting the step size equal to 0.05 and stopping criteria as $\max\big(|x_{\alpha,m}^{(t+1)} - x_{\alpha,m}^{(t)}|, \|x_{M,m}^{(t+1)} - x_{M,m}^{(t)}\|_\infty\big) \leq 10^{-7}$. Note that there are closed-form solutions to (A.4) and (A.5), while (A.3) is a unconstrained convex optimization problem. More specifically, in (A.4), suppose the eigenvalue decomposition of the symmetric matrix $(z_L^m - u_L^m)$ can be written as

$$z_L^m - u_L^m = T\Lambda T^T,$$

where T is orthogonal ($TT^T = I_n$). Then, for $J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$, we have

$$x_L^{m+1} = J\big(T\text{diag}(\Lambda - \lambda\delta\big)_+ T^T)J^T,$$

and diag$(\Lambda - \lambda\delta)_+$ is a diagonal matrix with the jth diagonal entry being

$$(\Lambda_{jj} - \lambda\delta)_+ = \begin{cases} 0, & \text{if } \Lambda_{jj} < \lambda\delta, \\ \Lambda_{jj} - \lambda\delta, & \text{if } \Lambda_{jj} \geq \lambda\delta. \end{cases}$$

Updating x_L^m requires full eigen-decomposition in each iteration, and this can be computationally expensive step when n is large. We adopt truncated-SVD to speed up this step performing eigen-decomposition for the first $r \ll n$ eigen-vectors of matrix $(z_L^m - u_L^m)$. The adoption of truncated-SVD reduces the computational complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(rn^2)$ in this step.

In (A.5), we have, for $i \neq j$,

$$S_{ij} = \begin{cases} 0, & \text{if } |(z_S^m - u_S^m)_{ij}| < \lambda\gamma, \\ (z_S^m - u_S^m)_{ij} - \lambda\gamma, & \text{if } (z_S^m - u_S^m)_{ij} > \lambda\gamma, \\ (z_S^m - u_S^m)_{ij} + \lambda\gamma, & \text{if } (z_S^m - u_S^m)_{ij} < -\lambda\gamma. \end{cases}$$

Step 2. We solve $z^{m+1} = \mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$. A closed-form solution exists here. Denote $\bar{\alpha} = x_{\alpha}^{m+1} + u_{\alpha}^m$, $\bar{M} = x_M^{m+1} + u_M^m$, $\bar{L} = x_L^{m+1} + u_L^m$, and $\bar{S} = x_S^{m+1} + u_S^m$, then evaluating $\mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$ becomes

$$\begin{array}{ll} \min\limits_{\alpha,M,L,S} & \frac{1}{2}[\alpha - \bar{\alpha}]^2 + \frac{1}{2}\|M - \bar{M}\|_F^2 + \frac{1}{2}\|L - \bar{L}\|_F^2 + \frac{1}{2}\|S - \bar{S}\|_F^2 \\ \text{subject to} & M \text{ is symmetric and } M = L + S. \end{array}$$

The above optimization problem has a close-form solution, which is as follows:

$$\begin{aligned} z_{\alpha}^{m+1} &= \bar{\alpha}, \\ z_M^{m+1} &= \frac{1}{3}\bar{M} + \frac{1}{3}\bar{M}^T + \frac{1}{3}\bar{L} + \frac{1}{3}\bar{S}, \\ z_L^{m+1} &= \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^T + \frac{2}{3}\bar{L} - \frac{1}{3}\bar{S}, \quad \text{and} \\ z_S^{m+1} &= \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^T - \frac{1}{3}\bar{L} + \frac{2}{3}\bar{S}. \end{aligned}$$

Step 3. We solve $u^{m+1} = u^m + x^{m+1} - z^{m+1}$, which is a simple arithmetic.

The most important implementation details of this algorithm are the choice of λ and stopping criterion. In this work, we simply choose $\lambda = 0.5$. We terminate the algorithm when in the mth iteration, we have $\|x_M^m - x_L^m - x_S^m\|_F \leq \varepsilon$, with $\varepsilon = 5 \times 10^{-6}$.

Remark A.1.2 *The convexity and separability of the objective function of the proposed model allow ADMM to be efficient. Consequently, we can apply our model to large size datasets, seeing Subsection 7.3. In some examples in the previous literature (e.g., in [27, 26], where the latent variables are treated as random effects and the MCMC method are used for parameter estimation), significant improvement in numerical efficiency is observed.*

A.2 Synthetic Setting

We describe a set of steps for setting the model parameters, α^*, F^*, D^* and S^* sequentially. We put astroids in the superscripts of parameters to indicate that they are the ground truth. Readers can refer the meaning of each parameter in the model in Section 1 and 3.

1. We draw an intercept term α^* in the logistic regression model from the uniform distribution that is supported on [-11,-10]. In this way, we can make α^* have the least effects in creating edges in the network.

2. Recall that the binary factor loading matrix F^* encodes the relation between factors and nodes (i.e., if ith node has kth factor, then we have $F^*_{ki} = 1$, otherwise $F^*_{ki} = 0$). First, we assume that there are n nodes in the network, and K factors are embedded in it. Each of them consists of roughly $\frac{n}{K}$ nodes. This can be expressed in F^* as follows:

$$F^* = \begin{bmatrix} \underbrace{1 \ \cdots \ 1}_{n/K} & & & \\ & \underbrace{1 \ \cdots \ 1}_{n/K} & & \\ & & \ddots & \\ & & & \underbrace{1 \ \cdots \ 1}_{n/K} \end{bmatrix} \in R^{K \times n},$$

   where each row of F^* has $\frac{n}{K}$ 1's and each column has only one 1. Note that the remaining entries of the matrix are filled with zeros.

3. In the aforementioned matrix F^*, each node is associated with only one factor. In this situation, a “perfect” clustering is possible, as each cluster corresponds to one hidden factor. On the other hand, we would like to consider the situation where one node may be related to multiple factors. If a node is associated with l factors, we say that this node has the multiplicity l. For simplicity, we consider only two possible

multiplicity l and m ($1 < l < m \leq K$). We assume that n_l nodes share l factors and n_m nodes share m factors. This can be incorporated in the aforementioned F^* in the following steps:

(a) Pick distinct n_l indices randomly from $\{1, 2, \ldots, n\}$. We will denote the set of the indices as Ω_l.

(b) Choose n_m indices from the set $\{1, 2, \ldots, n\} \setminus \Omega_l$ and denote the set of those indices as Ω_m.

(c) Make the columns of F^* with corresponding indices in set $\Omega_l \cup \Omega_m$ zeros. We use a notation f_j^* to denote the jth column of the matrix F^*. Fill arbitrary l entries of f_j^* for $j \in \Omega_l$ with 1's, and also fill arbitrary m entries of f_j^* for $j \in \Omega_m$ with 1's.

Lastly, we set $F^* = JF^*$ where $J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$. This is due to the discussion around the equation 10 in the paper.

4. Generate the weight coefficients of the factors D_{ii}^* from the uniform distribution that is supported on $[19, 20]$, $\forall 1 \leq i \leq K$. In this way, we can force the nodes that are associated with the same factor cluster together.

5. Recall that the positive entries of S^* can characterize the links in the network, which cannot be accounted by the common factors. We assume that there are $\frac{|S^*|}{\binom{K}{2}}$ edges between any two clusters, where $|S^*|$ denotes the number of non-zero entries of the upper-triangular part of the matrix S^*. This can be implemented via the following steps:

(a) We construct K sets $C_1^*, \ldots, C_K^*$ that are defined as follows:

$$\begin{aligned}
C_1^* &\subseteq \left\{1, 2, \ldots, \frac{n}{K}\right\} \setminus \{\Omega_l \cup \Omega_m\} \\
C_2^* &\subseteq \left\{\frac{n}{K} + 1, \ldots, \frac{2n}{K}\right\} \setminus \{\Omega_l \cup \Omega_m\} \\
&\vdots \\
C_K^* &\subseteq \left\{\frac{(K-1)n}{K} + 1, \ldots, n\right\} \setminus \{\Omega_l \cup \Omega_m\}
\end{aligned}$$

where each of them has arbitrary $\frac{|S^*|}{\binom{K}{2}}$ elements.

(b) Create a set I_{S^*} whose elements are pairs of indices such that

$$I_{S^*} = \bigcup_{1 \leq p < q \leq K} \left\{ (i_r^{p,q}, j_r^{p,q}) : i_r \in C_p^*, j_r \in C_q^*, r = 1, 2, \ldots, \frac{|S^*|}{\binom{K}{2}} \right\},$$

where $(i_r^{p,q}, j_r^{p,q})$ is the rth edge that connects a node in C_p^* and a node in C_q^*. Set I_{S^*} contains the edges that connect nodes in two different clusters.

(c) Draw $S_{ij}^* \sim \text{Unif}[19, 20], \forall (i, j) \in I_{S^*}$.

(d) Lastly, make it symmetric by setting $S_{ji}^* = S_{ij}^*, \forall 1 \leq i < j \leq n$.

6. Create an upper-triangular part of the adjacency matrix X whose each entry X_{ij} follows Bernoulli distribution. The distribution's parameter is parametrized by a probability, $P_{ij}^* = \frac{\exp(\alpha^* + f_i^{*T} D^* f_j^* + S_{ij}^*)}{1 + \exp(\alpha^* + f_i^{*T} D^* f_j^* + S_{ij}^*)}$. After drawing all the entries of X in the upper-triangular part, then make the matrix symmetric by setting $X_{ji} = X_{ij}, \forall 1 \leq i < j \leq n$.

A.3 Proof Theorem 1.

We briefly introduce several notations, including a notion on the decomposability of regularizer, and a useful lemma that is proved in the work [29] (SubsectionA.3.1). Then, we present Lemma A.3.3 and its proof (Subsection A.3.2). Finally, we present the proof of our

Theorem 1 (Subsection A.3.3).

A.3.1 Preliminary

Throughout the proof, we adopt the convenient short-hand notation on projection of matrix P on subspace M as P_M. We use $\langle A, B\rangle$ to denote the trace inner product of two matrices A and B $\big($i.e.,$\langle A, B\rangle = \text{tr}\big(A^T B\big)\big)$. We use $\|A\|_\infty$ to denote the maximum absolute entry of matrix A, and use $\|B\|_{op}$ to denote the largest singular value of matrix B. And we will use the notion of decomposability of L_1 norm with respect to a pair of subspace $(M, M^\perp)$. Given an arbitrary subset $S \subseteq \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$ of matrix indices, M is defined as follows:

$$M(S) := \{U \in \mathbb{R}^{n\times n} | U_{ij} = 0, \forall (i,j) \in S\}$$

and $M^\perp(S) := (M(S))^\perp$. With this in mind, we recall the formal definition of the decomposability of L_1 norm as follows:

Definition A.3.1 *Given a subspace $M \subset \mathbb{R}^{n\times n}$ and its orthogonal complement $M^\perp$, an elementwise L_1 norm is decomposable with respect to $(M, M^\perp)$ if*

$$\|A + B\|_1 = \|A\|_1 + \|B\|_1, \forall A \in M \textit{ and } B \in M^\perp.$$

The notion of decomposability is used to penalize the perturbation from the model subspace M, and to obtain the tightest bound the L_1 norm can achieve. We will also use two results in our proof, which are presented and proved in [29]. For the convenience of readers, we present them here:

Lemma A.3.2 *(**Agarwal, et al** [29]) For any $k = 1, 2, \ldots, n$, there is a decomposition $\widehat{\Delta}^L = \widehat{\Delta}^L_A + \widehat{\Delta}^L_B$ such that:*

1. *The decomposition satisfies*

$$\textit{rank}\big(\widehat{\Delta}^L_A\big) \leq 2k, \quad \textit{and} \quad \big(\widehat{\Delta}^L_A\big)^T \widehat{\Delta}^L_B = \big(\widehat{\Delta}^L_B\big)^T \widehat{\Delta}^L_A = 0 \tag{A.6}$$

2. *The difference* $\mathbb{Q}\big(L^*, S^*\big) - \mathbb{Q}\big(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*\big)$ *is upper-bounded by*

$$\mathbb{Q}\big(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\big) - \mathbb{Q}\big(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\big) + 2\sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 2\frac{\gamma}{\delta}\left\|S^*_{M^\perp}\right\|_1, \tag{A.7}$$

where the notation $\mathbb{Q}(L, S)$ is defined as the weighted combination of the two regularizers for any pair of positive tuning parameters (γ, δ):

$$\mathbb{Q}\left(L, S\right) := \left\|L\right\|_* + \frac{\gamma}{\delta}\left\|S\right\|_1.$$

A.3.2 Lemma A.3.3

Lemma A.3.3 *If a pair of regularization parameters* (δ, γ) *satisfies condition* (15) *in the paper, then for* $\mathbb{Q}\left(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\right)$*, we have*

$$\mathbb{Q}\left(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\right) \leq \left\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\right\|_F + 3\mathbb{Q}\left(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\right) + 4\sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 4\frac{\gamma}{\delta}\left\|S^*_{M^\perp}\right\|_1.$$

Proof. Through the application of basic inequality by using optimality of $\widehat{\Theta}$ and feasibility of Θ^* to convex program (A.1), we have

$$h\big(\widehat{\Theta}\big) - h\big(\Theta^*\big) \leq \delta\mathbb{Q}\big(L^*, S^*\big) - \delta\mathbb{Q}\big(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*\big). \tag{A.8}$$

By using convexity of $h(\Theta)$, we can write

$$\begin{aligned} h\big(\widehat{\Theta}\big) - h\big(\Theta^*\big) \geq \big\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \big\rangle &= -\big\langle \frac{1}{n}(X - P^*), \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T + \widehat{\Delta}^L + \widehat{\Delta}^S \big\rangle \\ &\geq -\frac{1}{n}\|X - P^*\|_{op}\Big(\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_* + \big\|\widehat{\Delta}^L\big\|_*\Big) - \frac{1}{n}\|X - P^*\|_\infty \big\|\widehat{\Delta}^S\big\|_1 \\ &\geq -\frac{\delta}{2}\Big(\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F + \big\|\widehat{\Delta}^L_A\big\|_* + \big\|\widehat{\Delta}^L_B\big\|_*\Big) - \frac{\gamma}{2}\Big(\big\|\widehat{\Delta}^S_M\big\|_1 + \big\|\widehat{\Delta}^S_{M^\perp}\big\|_1\Big). \end{aligned} \tag{A.9}$$

An application of Agarwal et al [29]'s second element of lemma A.3.2, we can get an upper bound of difference $\mathbb{Q}\big(L^*, S^*\big) - \mathbb{Q}\big(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*\big)$ as follows:

$$\mathbb{Q}\big(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\big) - \mathbb{Q}\big(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\big) + 2\sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 2\frac{\gamma}{\delta}\big\|S^*_{M^\perp}\big\|_1. \tag{A.10}$$

By combining inequalities (A.8), (A.9) and (A.10), we can get the upper bound of $\mathbb{Q}\big(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\big)$:

$$\mathbb{Q}\big(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\big) \le \big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F + 3\mathbb{Q}\big(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\big) + 4\sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 4\frac{\gamma}{\delta}\big\|S^*_{M^\perp}\big\|_1.$$

□

A.3.3 Main proof of Theorem 1

Proof. Since $\widehat{\Theta}$ and Θ^* are optimal minimizer and feasible solution respectively for the convex program (A.1), we have

$$h\big(\widehat{\Theta}\big) + \delta\big\|\widehat{L}\big\|_* + \gamma\big\|\widehat{S}\big\|_1 \le h\big(\Theta^*\big) + \delta\big\|L^*\big\|_* + \gamma\big\|S^*\big\|_1. \tag{A.11}$$

Through the assumption of strong convexity on $h(\Theta)$, and by the Taylor expansion, we can get a following lower bound on the term $h\big(\widehat{\Theta}\big) - h\big(\Theta^*\big)$:

$$h\big(\widehat{\Theta}\big) - h\big(\Theta^*\big) \ge \big\langle \nabla_\Theta h\big(\Theta^*\big), \widehat{\Theta} - \Theta^* \big\rangle + \frac{\tau}{2}\big\|\widehat{\Delta}^\Theta\big\|_F^2.$$

By rearranging the term in (A.11) and plugging in above inequality relation, we get:

$$\frac{\tau}{2}\big\|\widehat{\Delta}^\Theta\big\|_F^2 \le -\big\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \big\rangle + \delta\big\|L^*\big\|_* + \gamma\big\|S^*\big\|_1 - \delta\big\|\widehat{L}\big\|_* - \gamma\big\|\widehat{S}\big\|_1. \tag{A.12}$$

Through the definition of $\mathbb{Q}$, we can rewrite (A.12) as follows:

$$\frac{\tau}{2}\big\|\widehat{\Delta}^{\Theta}\big\|_F^2 \le -\big\langle\, \nabla_{\Theta} h(\Theta^*), \widehat{\Theta}-\Theta^* \big\rangle\, + \delta\mathbb{Q}\big(L^*, S^*\big) - \delta\mathbb{Q}\big(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*\big). \tag{A.13}$$

According to Agarwal et al [29]'s second element of lemma A.3.2, the difference $\mathbb{Q}\big(L^*, S^*\big) - \mathbb{Q}\big(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*\big)$ is upper-bounded by

$$\mathbb{Q}\big(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S\big) - \mathbb{Q}\big(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S\big) + 2\sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 2\frac{\gamma}{\delta}\big\|S^*_{M^\perp}\big\|_1. \tag{A.14}$$

First, we want to control upper bound of the term $-\big\langle\, \nabla_{\Theta} h(\Theta^*), \widehat{\Theta}-\Theta^* \big\rangle$ in (A.13).

$$-\big\langle\, \nabla_{\Theta} h(\Theta^*), \widehat{\Theta}-\Theta^* \big\rangle\, = \big\langle\, \frac{1}{n}(X-P^*), \widehat{\Delta}^{\alpha\mathbb{1}\mathbb{1}^T} + \widehat{\Delta}^L + \widehat{\Delta}^S \big\rangle \tag{A.15}$$

$$\le \frac{1}{n}\|X-P^*\|_{op}\bigg(\big\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\big\|_* + \big\|\widehat{\Delta}^L\big\|_*\bigg) + \frac{1}{n}\|X-P^*\|_\infty\big\|\widehat{\Delta}^S\big\|_1$$

$$\le \frac{1}{n}\|X-P^*\|_{op}\bigg(\,\Big\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\Big\|_F + \Big\|\widehat{\Delta}_A^L\Big\|_* + \Big\|\widehat{\Delta}_B^L\Big\|_*\bigg) + \frac{1}{n}\,\|X-P^*\|_\infty\,\bigg(\,\Big\|\widehat{\Delta}_M^S\Big\|_1 + \Big\|\widehat{\Delta}_{M^\perp}^S\Big\|_1\bigg)$$

$$\le \frac{\delta}{2}\bigg(\big\|\hat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\big\|_F + \big\|\widehat{\Delta}_A^L\big\|_* + \big\|\widehat{\Delta}_B^L\big\|_*\bigg) + \frac{\gamma}{2}\bigg(\,\Big\|\widehat{\Delta}_M^S\Big\|_1 + \Big\|\widehat{\Delta}_{M^\perp}^S\Big\|_1\bigg). \tag{A.16}$$

Combining the inequalities (A.14) and (A.16), we can obtain the upper bound of RHS in (A.13) as follows:

$$\frac{\tau}{2}\big\|\widehat{\Delta}^{\Theta}\big\|_F^2 \le \frac{\delta}{2}\big\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\big\|_F + \frac{3\delta}{2}\mathbb{Q}\big(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S\big) + 2\delta\sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 2\gamma\big\|S^*_{M^\perp}\big\|_1. \tag{A.17}$$

Second, we wish to control the lower bound of the term $\frac{\tau}{2}\big\|\widehat{\Delta}^{\Theta}\big\|_F^2$ with respect to $\widehat{\Delta}^{\alpha}, \widehat{\Delta}^L, \widehat{\Delta}^S$.

$$\begin{aligned}\big\|\widehat{\Delta}^{\Theta}\big\|_F^2 &= \big\|\widehat{\Theta}-\Theta^*\big\|_F^2 = \big\|\widehat{\Delta}^{\alpha\mathbb{1}\mathbb{1}^T} + \widehat{\Delta}^L + \widehat{\Delta}^S\big\|_F^2 \\ &= \big\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\big\|_F^2 + \big\|\widehat{\Delta}^L + \widehat{\Delta}^S\big\|_F^2 + 2\big\langle\, \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\big\rangle \\ &= \big\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\big\|_F^2 + \big\|\widehat{\Delta}^L\big\|_F^2 + \big\|\widehat{\Delta}^S\big\|_F^2 + 2\big\langle\, \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\big\rangle\, + 2\big\langle\, \widehat{\Delta}^L, \widehat{\Delta}^S\big\rangle.\end{aligned} \tag{A.18}$$

We want to get the further lower bound on trace inner product terms, $\left\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T \right\rangle$, $\left\langle \widehat{\Delta}^L, \widehat{\Delta}^S \right\rangle$. To control the first trace inner product term, we use the relation $\widehat{\Delta}^L \mathbb{1} = 0$, apply the definition of dual norm on inner product term, apply triangular inequality on $\widehat{\Delta}^\alpha$, and lastly we apply the constraint imposed on $|\alpha|$ stated in Assumption 2. We have

$$\begin{aligned} \left|\left\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T \right\rangle\right| &= \left|\left\langle \widehat{\Delta}^S, \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T \right\rangle\right| \\ &\leq \big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_\infty \big\|\widehat{\Delta}^S\big\|_1 \\ &\leq \Big(|\widehat{\alpha}| + |\alpha^*|\Big) \big\|\widehat{\Delta}^S\big\|_1 \\ &\leq 2C\kappa \big\|\widehat{\Delta}^S\big\|_1. \end{aligned} \tag{A.19}$$

To control the term $\left\langle \widehat{\Delta}^L, \widehat{\Delta}^S \right\rangle$, we first apply the definition of dual norm on trace inner product term, then apply triangular inequality on $\widehat{\Delta}^L$ and spikiness condition. We have

$$\begin{aligned} \left|\left\langle \widehat{\Delta}^L, \widehat{\Delta}^S \right\rangle\right| &\leq \big\|\widehat{\Delta}^L\big\|_\infty \big\|\widehat{\Delta}^S\big\|_1 \\ &\leq \Big(\big\|\widehat{L}\big\|_\infty + \big\|L^*\big\|_\infty\Big) \big\|\widehat{\Delta}^S\big\|_1 \\ &\leq \left(\frac{2\kappa}{n}\right) \big\|\widehat{\Delta}^S\big\|_1. \end{aligned} \tag{A.20}$$

We can combine the inequality (A.18), (A.19) and (A.20). Then applying the assumption on regularization parameter γ, and the fact $\big\|\widehat{\Delta}^L\big\|_* \geq 0$ sequentially, we can get the following,

$$\begin{aligned} \frac{\tau}{2}\big\|\widehat{\Delta}^\Theta\big\|_F^2 &\geq \frac{\tau}{2}\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^L\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^S\big\|_F^2 - \kappa\tau\left(\frac{Cn+1}{n}\right)\big\|\widehat{\Delta}^S\big\|_1 \\ &\geq \frac{\tau}{2}\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^L\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^S\big\|_F^2 - \frac{\gamma}{2}\big\|\widehat{\Delta}^S\big\|_1 \\ &\geq \frac{\tau}{2}\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^L\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^S\big\|_F^2 - \frac{\delta}{2}\mathbb{Q}\big(\widehat{\Delta}^L, \widehat{\Delta}^S\big). \end{aligned} \tag{A.21}$$

By combining the relations (A.17) and (A.21), applying triangular inequality, $\mathbb{Q}\big(\widehat{\Delta}^L, \widehat{\Delta}^S\big) \leq$

$\mathbb{Q}\big(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\big) + \mathbb{Q}\big(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\big)$, and rearranging the term, we can get following inequality,

$$\begin{aligned}
&\frac{\tau}{2}\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^L\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^S\big\|_F^2 \\
&\leq \frac{\delta}{2}\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F + 2\mathbb{Q}\big(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\big) + \frac{\delta}{2}\mathbb{Q}\big(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}\big) + 2\delta \sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 2\gamma\big\|S^*_{M^\perp}\big\|_1.
\end{aligned}$$

Further, by plugging in Lemma 1 to get an upper bound on $\mathbb{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp})$, we can rewrite the above inequality as follows:

$$\frac{\tau}{2}\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^L\big\|_F^2 + \frac{\tau}{2}\big\|\widehat{\Delta}^S\big\|_F^2 - \frac{\delta}{2}\big\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\big\|_F \tag{A.22}$$

$$\leq \frac{7\delta}{2}\mathbb{Q}\big(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\big) + 4\delta \sum_{j=k+1}^{n} \sigma_j\big(L^*\big) + 4\gamma\big\|S^*_{M^\perp}\big\|_1. \tag{A.23}$$

Noting that $\widehat{\Delta}^L_A$ has rank at most $2k$ and that $\widehat{\Delta}^S_M$ lies in the model space M, we find that

$$\begin{aligned}
\delta\mathbb{Q}\big(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M\big) &\leq \sqrt{2k}\delta\big\|\widehat{\Delta}^L_A\big\|_F + \Psi(M)\gamma\big\|\widehat{\Delta}^S_M\big\|_F \\
&\leq \sqrt{2k}\delta\big\|\widehat{\Delta}^L\big\|_F + \Psi(M)\gamma\big\|\widehat{\Delta}^S\big\|_F.
\end{aligned} \tag{A.24}$$

Here $\Psi(M)$ measures the compatibility between Frobenius norm and component-wise L_1 regularizer, where M is an arbitrary subset of matrix indices of cardinality at most s. We have

$$\Psi(M) := \sup_{U \in M, U \neq 0} \frac{\|U\|_1}{\|U\|_F}.$$

Using Cauchy-Schwarz inequality, we can easily check the quantity $\Psi(M)$ is bounded by at most $\sqrt{s}$. Plugging in the relation ($A.24$) into ($A.23$) and rearranging the term relevant with $e^2\big(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}\big)$ yield the claim. □

A.4 Discussions on “Mixed Topics” cluster in Subsection 7.4.

Mixed Topics. The sub-network structure estimated in Step 2 in Subsection 7.4. has a big collection of papers that we refer it as “Mixed Topics” cluster **(Mixed)**. In the cluster, we can see papers with topics on Statistical learning theory, Non-parametric/Semi-parametric statistics, Spatial statistics, Theoretical machine learning, which does not seem to belong to any of the five communities listed above. Additionally, we can identify papers with combination of two or three topics. Papers such as ‘The Bayesian Lasso’ (T. Park, et al. 2008), ‘Coordinate-independent sparse sufficient dimension reduction and variable selection’ (X. Chen, et al. 2010), can be taken as examples. It is also interesting to think about reasons for why papers that seem to have obvious membership in one of aforementioned 5 communities other than Mixed Topic are classified as Mixed Topic. For instance, the paper, ‘On the “degrees of freedom” on the LASSO’ (H. Zou, et al. 2007), is classified as Mixed Topic paper. We can conjecture variable selection has lots of applications in other topics, so it might either cite or have been cited by many papers in other communities. Actually, among 232 papers in the network, the paper has 11 citational relationships with papers from 4 different communities. (i.e., VarSel, Bayes, CovEst, Mixed)

Ad-hoc Edges of “Mixed Topics”. The selected model in Step 2 in Subsection 7.4. has 151 ad-hoc edges. Among those 151 ad-hoc citational relationships, 118 of them are formed by pairs of papers from “Mixed Community”. This result is unexpected since our model is designed to capture ad-hoc edges whose corresponding nodes belong to different communities. One possible explanation on this phenomenon is as follows: we observe that latent vectors of papers classified as “Mixed Topics” are clustered around origin in an L_2 sense. See Figure. A.1 and its caption for details. Roughly speaking, this leads $\widehat{L}_{ij}^{\text{sub}}$ to have small value so that the edges within “Mixed Topic” cluster are captured via positive entries $\widehat{S}_{ij}^{\text{sub}}$, where i, j are indices of papers clustered as “Mixed Topics”. Also note that the remaining

31 ad-hoc edges (except for the 2 pairs of papers from “FuncAn-FuncAn”) are formed by pairs of papers with different communities. (See Table. A.1.) A full list of papers which form the 151 ad-hoc edges is provided in the webpage[1].

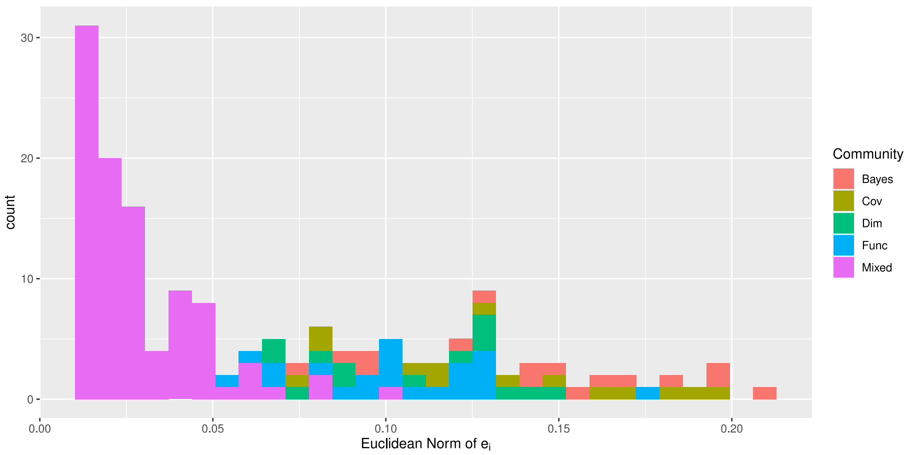


Figure A.1: Distribution of $\|e_i\|_2$ where e_i is the i^{th} row of $\widehat{E}_5^{\text{sub}} \in \mathbb{R}^{162\times 5}$. Note that nodes classified as “Mixed Topics” are highly clustered around origin in a L_2 sense. Recall that the definition of $\widehat{E}_K$ is in Subsection 7.3.

Community - Community	Number of citations
“Mixed - Mixed”	118
“Mixed - CovEst”	12
“Mixed - FuncAn”	8
“Mixed - Bayes”	7
“Mixed - DimRed”	3
“FuncAn - CovEst”	1
“FuncAn - FuncAn”	2
Total	151

Table A.1: Among the 151 ad-hoc citational relationships, 118 of them are formed by pairs of papers from “Mixed Community”. Remaining 31 ad-hoc edges (except for the 2 pairs of papers from “FuncAn-FunAc”) are formed by pairs of papers with different communities.

[1] https://sites.google.com/site/namjoonsuh/publications

APPENDIX B

ASYMPTOTIC THEORY OF ℓ_1-REGULARIZED PDE IDENTIFICATION FROM A SINGLE NOISY TRAJECTORY

B.1 Primal-Dual Witness construction

In this section, we briefly rephrase the explanation of PDW construction in the book [199] for reader's convenience. A primal-dual pair $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{z}}) \in \mathbb{R}^{K \times K}$ is said to be optimal if $\widehat{\boldsymbol{\beta}}$ is a minimizer of (3.7) and $\widehat{\mathbf{z}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$, where $\partial\|\widehat{\boldsymbol{\beta}}\|_1$ denotes a sub-differential set of $\|\cdot\|_1$ evaluated at $\widehat{\boldsymbol{\beta}}$. Any such pair must satisfy the zero-subgradient condition of (3.7), which is as follows:

$$-\frac{1}{NM}\widehat{\mathbf{F}}^T(\widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\widehat{\boldsymbol{\beta}}) + \lambda_N \widehat{\mathbf{z}} = 0 \ , \ \text{for } \widehat{\mathbf{z}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1 \ . \tag{B.1}$$

Recall that we denote the ground-truth support of $\boldsymbol{\beta}^*$ as $\mathcal{S}$, and suppose that we know $\mathcal{S}$ apriori. For the ground-truth support set $\mathcal{S}$ and its complement set $\mathcal{S}^c$, PDW is said to be successful if the constructed tuple, $(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}, \widehat{\boldsymbol{\beta}}_{\mathcal{S}^c}, \widehat{\mathbf{z}}_{\mathcal{S}}, \widehat{\mathbf{z}}_{\mathcal{S}^c})$, is primal-dual optimal, and act as a witness for the fact that the LASSO finds the unique optimal solution with correct support set. We construct the tuple through the following three steps.

1. Set $\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} = 0$.

2. Find $(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}, \widehat{\mathbf{z}}_{\mathcal{S}})$ by solving the s-dimensional oracle sub-problem

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}} \in \underset{\boldsymbol{\beta}_{\mathcal{S}} \in \mathbb{R}^s}{\operatorname{argmin}} \left\{ \frac{1}{2NM} \left\| \widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}} \right\|_2 + \lambda_N \|\boldsymbol{\beta}_{\mathcal{S}}\|_1 \right\},$$

   where s is the cardinality of the set $\mathcal{S}$. Thus $\widehat{\mathbf{z}}_{\mathcal{S}} \in \partial\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1$ satisfies the relation $-\frac{1}{NM}\widehat{\mathbf{F}}_{\mathcal{S}}^T(\widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}_{\mathcal{S}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}}) + \lambda_N \widehat{\mathbf{z}}_{\mathcal{S}} = 0$.

3. Solve for $\widehat{\mathbf{z}}_{\mathcal{S}^c}$ through the zero-subgradient equation (B.1), and check whether or not the *strict dual feasibility* condition $\|\widehat{\mathbf{z}}_{\mathcal{S}}\|_{\infty} < 1$ holds.

B.2 Local-Polynomial estimator : Closed-form solutions

Recall that we want to solve following two optimization problems for constructing $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$, given the noisy observation $\mathcal{D} = \{ \big(X_i, t_n, U_i^n\big) \mid i = 0, \dots, M-1; n = 0, \dots, N-1 \}$.

$$\Big\{ \widehat{b}_j(X_i, t) \Big\}_{j=0,1,2} = \underset{b_j(t) \in \mathbb{R}, 0 \le j \le 2}{\operatorname{argmin}} \sum_{n=0}^{N-1} \bigg(U_i^n - \sum_{j=0}^{2} b_j(t)(t_n - t)^j \bigg)^2 \mathcal{K}_{h_N}\Big(t_n - t \Big) ,$$

$$\text{for } i = 0, 1, \dots, M-1 \; ; \tag{B.2}$$

$$\Big\{ \widehat{c}_j^p(x, t_n) \Big\}_{j=0,1,\dots,p+1} = \underset{c_j(t) \in \mathbb{R}, 0 \le j \le p+1}{\operatorname{argmin}} \sum_{i=0}^{M-1} \bigg(U_i^n - \sum_{j=0}^{p+1} c_j^p(t)(X_i - x)^j \bigg)^2 \mathcal{K}_{w_M}\Big(X_i - x \Big)$$

$$\text{for } n = 0, 1, \dots, N-1 \text{ and } p = 0, 1, \dots, P_{\max}. \tag{B.3}$$

and set $\widehat{u}_t(X_i, t) = \widehat{b}_1(X_i, t)$ and $\widehat{\partial_x^p u}(x, t_n) = p! \widehat{c}_p^p(x, t_n)$. Then, the standard weighted least-square theory leads to the solutions of (B.2) and (B.3), respectively:

$$\widehat{u}_t(X_i, t) = \xi_1^T \big(\mathbf{T_1}^{\mathrm{T}} \mathbf{W_t} \mathbf{T_1} \big)^{-1} \mathbf{T_1}^{\mathrm{T}} \mathbf{W_t} \mathbf{U_i}, \quad \forall i = 0, 1, \dots, M-1,$$

$$\widehat{\partial_x^p u}(x, t_n) = p! \xi_{p,x}^T \big(\mathbf{X_p}^{\mathrm{T}} \mathbf{W_x} \mathbf{X_p} \big)^{-1} \mathbf{X_p}^{\mathrm{T}} \mathbf{W_x} \mathbf{U^n}, \quad \forall p = 0, 1, \dots, P_{\max}, \quad \forall n = 0, 1, \dots, N-1,$$

where $\mathbf{U_i} = [U_i^0, \dots, U_i^{N-1}]^{\mathrm{T}}$ and $\mathbf{U^n} = [U_0^n, \dots, U_{M-1}^n]^{\mathrm{T}}$, and

$$\mathbf{T_1} := \begin{bmatrix} 1 & t_0 - t & \big(t_0 - t\big)^2 \\ 1 & t_1 - t & \big(t_1 - t\big)^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{N-1} - t & \big(t_{N-1} - t\big)^2 \end{bmatrix} , \quad \mathbf{X_p} := \begin{bmatrix} 1 & X_0 - x & \cdots & \big(X_0 - x\big)^{p+1} \\ 1 & X_1 - x & \cdots & \big(X_1 - x\big)^{p+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{M-1} - x & \cdots & \big(X_{M-1} - x\big)^{p+1} \end{bmatrix} ,$$

for $p = 0, \ldots, P_{\max}$, and

$$\mathbf{W_t} := \text{diag}\big\{\mathcal{K}_{h_N}(t_0 - t), \ldots, \mathcal{K}_{h_N}(t_{N-1} - t)\big\},$$
$$\mathbf{W_x} := \text{diag}\big\{\mathcal{K}_{w_M}(X_0 - x), \ldots, \mathcal{K}_{w_M}(X_{M-1} - x)\big\},$$

are $N \times N$ and $M \times M$ diagonal matrices of kernel weights, and ξ_2 is the 3×1 vector having 1 in the 2nd entry and zeros in the other entries, and $\xi_{p,x}$ is the $(p+1) \times 1$ vector having 1 in the pth entry and zeros in the other entries.

B.3 Proof of Proposition 5.1

By the KKT-condition, any minimizer $\check{\boldsymbol{\beta}}$ of (3.7) satisfies:

$$-\frac{1}{NM}\widehat{\mathbf{F}}^T(\widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\check{\boldsymbol{\beta}}) + \lambda_N \check{\mathbf{z}} = 0 \ , \ \text{for } \check{\mathbf{z}} \in \partial\|\check{\boldsymbol{\beta}}\|_1 \ . \tag{B.4}$$

Recall that $\Delta\mathbf{u}_t = \widehat{\mathbf{u}}_t - \mathbf{u}_t$, $\Delta\mathbf{F} = \widehat{\mathbf{F}} - \mathbf{F}$ denote the error terms. By using the ground-truth PDE $u_t = \mathbf{F}\beta^*$ and definitions of $\Delta\mathbf{u}_t$ and $\Delta\mathbf{F}$, we have $\widehat{\mathbf{u}}_t = \widehat{\mathbf{F}}\boldsymbol{\beta}^* - \Delta\mathbf{F}\boldsymbol{\beta}^* + \Delta\mathbf{u}_t$. Thus from (B.4), we get

$$\widehat{\mathbf{F}}^T\widehat{\mathbf{F}}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \widehat{\mathbf{F}}^T(\Delta\mathbf{F}\boldsymbol{\beta}^* - \Delta\mathbf{u}_t) + \lambda_N NM\mathbf{z} = 0 \ . \tag{B.5}$$

We decompose (B.5) as follows:

$$\begin{bmatrix} \widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}} & \widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}^c} \\ \widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}} & \widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}^c} \end{bmatrix} \begin{bmatrix} \check{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}^* \\ 0 \end{bmatrix} + \begin{bmatrix} \widehat{\mathbf{F}}_{\mathcal{S}}^T \\ \widehat{\mathbf{F}}_{\mathcal{S}^c}^T \end{bmatrix} (\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^* - \Delta\mathbf{u}_t) + \lambda_N NM \begin{bmatrix} \check{\mathbf{z}}_{\mathcal{S}} \\ \check{\mathbf{z}}_{\mathcal{S}^c} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \ , \tag{B.6}$$

where we used the fact $\boldsymbol{\beta}^*_{\mathcal{S}^c} = \mathbf{0}$ and $\check{\boldsymbol{\beta}}_{\mathcal{S}^c} = 0$ via PDW construction. Solving (B.6), we have following two equalities:

$$\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big(\check{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\big) + \widehat{\mathbf{F}}_{\mathcal{S}}^T(\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}} - \Delta\mathbf{u}_t) + \lambda_N NM\check{\mathbf{z}}_{\mathcal{S}} = 0 \tag{B.7}$$

$$\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big(\check{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\big) + \widehat{\mathbf{F}}_{\mathcal{S}^c}^T(\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}} - \Delta\mathbf{u}_t) + \lambda_N NM\check{\mathbf{z}}_{\mathcal{S}^c} = 0 \tag{B.8}$$

Using the minimum eigen-value condition in the assumption (A3), from (B.7), we have

$$\check{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}} = \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\bigg(\widehat{\mathbf{F}}_{\mathcal{S}}^T(\Delta\mathbf{u}_t - \Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}) - \lambda_N NM\check{\mathbf{z}}_{\mathcal{S}}\bigg). \tag{B.9}$$

Plugging (B.9) into (B.8) gives:

$$\check{\mathbf{z}}_{\mathcal{S}^c} = \widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}})^{-1}\mathbf{z}_{\mathcal{S}} + \frac{1}{\lambda_N MN}\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\mathbf{\Pi}_{\mathcal{S}^\perp}(\Delta\mathbf{u}_t - \Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}})\ ,$$

where $\mathbf{\Pi}_{\mathcal{S}^\perp} = \mathbf{I} - \widehat{\mathbf{F}}_{\mathcal{S}}(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{S})^{-1}\widehat{\mathbf{F}}_{\mathcal{S}}^T$ is an orthogonal projection operator on the column space of $\widehat{\mathbf{F}}_{\mathcal{S}}$. By the complementary slackness condition, for $j \in \mathcal{S}^c$, $|\check{\mathbf{z}}_j| < 1$ implies $\check{\boldsymbol{\beta}}_j = \mathbf{0}$, which guarantees the proper support recovery. i.e., $\mathcal{S}(\check{\boldsymbol{\beta}}) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$. Now, we can focus on proving that, as $N, M \to \infty$, for μ in (A3), $\mathbb{P}\big[\max_{j\in\mathcal{S}^c}|\widetilde{Z}_j| \geq \mu\big] \to 0$, for $\widetilde{Z}_j = [\widehat{\mathbf{F}}_{\mathcal{S}^c}]_j^T\mathbf{\Pi}_{\mathcal{S}^\perp}\frac{\Delta\mathbf{u}_t - \Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}}{\lambda_N NM}$, $[\widehat{\mathbf{F}}_{\mathcal{S}^c}]_j$ is the j-th column of $\widehat{\mathbf{F}}_{\mathcal{S}^c}$. By the following lemma, we claim that to prove (i) of Proposition 1, it suffices to bound ℓ_∞-norm of the PDE estimation error $\boldsymbol{\tau}$.

Lemma B.3.1 *For any $\varepsilon > 0$:*

$$\mathbb{P}\bigg[\max_{j\in\mathcal{S}^c}\Big|\widetilde{Z}_j\Big| \geq \varepsilon\bigg] \leq \mathbb{P}\bigg[\|\boldsymbol{\tau}\|_\infty \geq \frac{\lambda_N\varepsilon}{\sqrt{K}}\bigg]\ .$$

Proof.

$$\begin{aligned}
\mathbb{P}\left[\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\boldsymbol{\Pi}_{\mathcal{S}^\perp}\frac{\boldsymbol{\tau}}{\lambda_N NM}\right\|_\infty \ge \varepsilon\right] &\le \mathbb{P}\left[\left\|\widehat{\mathbf{F}}^T\boldsymbol{\Pi}_{\mathcal{S}^\perp}\frac{\boldsymbol{\tau}}{\lambda_N NM}\right\|_2 \ge \varepsilon\right] \\
&\le \mathbb{P}\left[\left\|\boldsymbol{\Pi}_{\mathcal{S}^\perp}\big(\widehat{\mathbf{F}}\big)\right\|_2\left\|\frac{\boldsymbol{\tau}}{\lambda_N NM}\right\|_2 \ge \varepsilon\right] \\
&\le \mathbb{P}\left[\left\|\widehat{\mathbf{F}}\right\|_F\left\|\frac{\boldsymbol{\tau}}{\lambda_N NM}\right\|_2 \ge \varepsilon\right] \\
&\le \mathbb{P}\left[\|\boldsymbol{\tau}\|_2 \ge \lambda_N\varepsilon\sqrt{\frac{NM}{K}}\right] \\
&\le \mathbb{P}\left[\|\boldsymbol{\tau}\|_\infty \ge \frac{\lambda_N\varepsilon}{\sqrt{K}}\right].
\end{aligned}$$

In the second inequality, we use the definition of spectral norm of matrix, and in the third inequality, we use the fact $\|\boldsymbol{\Pi}_{\mathcal{S}^\perp}\|_{\mathbf{2}} = \mathbf{1}$. In the fourth inequality, the condition $\frac{1}{\sqrt{NM}}\max_{j=1,\ldots,K}\|\widehat{\mathbf{F}}_j\|_2 \le 1$ is used, giving us $\|\widehat{\mathbf{F}}\|_F \le \sqrt{KNM}$. In the last inequality, we use $\|\boldsymbol{\tau}\|_2 \le \sqrt{NM}\|\boldsymbol{\tau}\|_\infty$. □

B.3.1 Sufficient conditions for bounding $\widehat{\mathbf{u}}_t - \mathbf{u}_t$

Lemma B.3.2 *Let $\mathcal{K}^*_{\max} = \|\mathcal{K}^*\|_\infty$, B_N be an arbitrary increasing sequence $B_N \to \infty$ as $N \to \infty$, and $B_N^{'} = B_N + \|u\|_{L^\infty(\Omega)}$. For any $i = 0, 1, \ldots, M$ and arbitrary real r, there exist finite positive constants $A(X_i), C^*(X_i), a_0, b_0, c_0$, and $d_0(X_i)$ which do not depend on the temporal sample size N, such that for any $\alpha > 1$ and*

$$\begin{aligned}
&\varepsilon_N^*(X_i, r, \alpha) > \\
&\max\left\{3|C^*(X_i)|h_N^2, \frac{6\mathcal{K}^*_{\max}B_N^{'}}{Nh_N^2}, 6\frac{A(X_i)\big(B_N^{'}\big)^{-1}}{h_N}, \frac{6B_N^{'}\mathcal{K}^*_{\max}(a_0\ln N + r)\ln N}{h_N^2 N},\right. \\
&\qquad\left. 12\sqrt{\alpha}d_0(X_i)\sqrt{\frac{\ln 1/h_N}{h_N^3 N}}\right\},
\end{aligned}$$

as long as N is sufficiently large, we have:

$$\mathbb{P}\Big[\sup_{t\in[0,T]}|\Delta u_t(X_i,t)| > \varepsilon_N^*(X_i,r,\alpha)\Big] < 2N\exp\left(-\frac{B_N^2}{2\sigma^2}\right) + b_0\exp\big(-c_0 r\big) + 4\sqrt{2}\eta^4 h_N^\alpha\ .$$

Proof. In the following argument, we fix some $i = 0,\cdots,M-1$ and omit the dependence on X_i in the notations. Let $B_N^{'} = B_N + \|u\|_{L^\infty(\Omega)}$ with B_N being a sequence of increasing positive numbers such that $B_N \to \infty$ as $N \to \infty$, then define the truncated estimate

$$\begin{aligned}\widehat{u_t}^{B_N^{'}}(X_i,t) &= \frac{1}{Nh_N^2}\sum_{n=0}^{N-1}\mathcal{K}^*\left(\frac{t_n-t}{h_N}\right)U_i^n I\{|U_i^n| < B_N^{'}\} \\ &= \frac{1}{h_N^2}\iint_{|y|<B_N^{'}}\mathcal{K}^*\left(\frac{z-t}{h_N}\right)y\, df_N(z,y),\end{aligned} \tag{B.10}$$

where $f_N(\cdot,\cdot) := f_N(\cdot,\cdot|X_i)$ is the empirical distribution of (t_n, U_i^n) conditioned on the space X_i. For any (X_i,t), decomposing the estimation error of the temporal partial derivative as follows

$$\widehat{u}_t - u_t = \underbrace{\left(\widehat{u}_t - \widehat{u}_t^{B_N^{'}} - \mathbb{E}\big(\widehat{u}_t - \widehat{u_t}^{B_N^{'}}\big)\right)}_{\text{Asymptotic deviation on the truncation error}} + \underbrace{\left(\widehat{u_t}^{B_N^{'}} - \mathbb{E}\widehat{u_t}^{B_N^{'}}\right)}_{\substack{\text{Asymptotic deviation of}\\ \text{truncated estimator}}} + \underbrace{\Big(\mathbb{E}\widehat{u}_t - u_t\Big)}_{\substack{\text{Asymptotic bias of}\\ \text{Local-Polynomial estimator}}}\ ,$$

we will prove that the error is bounded (in probability) by showing each component is bounded.

Component 1. Asymptotic deviation on the truncation error: For any $\varepsilon_{0,N} \geq \frac{\mathcal{K}^*_{\max}B_N^{'}}{Nh_N^2}$:

$$\begin{aligned}\mathbb{P}\Big[\sup_t|\widehat{u}_t - \widehat{u_t}^{B_N^{'}}| > \varepsilon_{0,N}\Big] &= \mathbb{P}\Big[\sup_t\left|\frac{1}{Nh_N^2}\sum_{n=0}^{N-1}\mathcal{K}^*\left(\frac{t_n-t}{h_N}\right)U_i^n I\{|U_i^n| \geq B_N^{'}\}\right| > \varepsilon_{0,N}\Big] \\ &\leq \mathbb{P}\Big[\frac{\mathcal{K}^*_{\max}}{Nh_N^2}\sum_{n=0}^{N-1}|U_i^n|I\{|U_i^n| \geq B_N^{'}\} > \varepsilon_{0,N}\Big] \leq \mathbb{P}\Big[\exists n = 0,1,\cdots,N-1,\ |U_i^n| \geq B'_N\Big] \\ &= \mathbb{P}\Big[\max_{n=0,1,\cdots,N-1}|U_i^n| \geq B'_N\Big] \leq \mathbb{P}\Big[\max_{n=0,1,\cdots,N-1}|U_i^n - u_i^n| \geq B_N\Big] \leq 2N\exp\left(-\frac{B_N^2}{2\sigma^2}\right)\end{aligned}$$

where σ denotes the standard deviation of the Gaussian noise added on the data. On the other hand, from Proposition 1 of [97]:

$$\mathbb{E}|\widehat{u}_t - \widehat{u}_t^{B_N^{'}}| \leq \frac{A\big(B_N^{'}\big)^{-1}}{h_N} \ .$$

for $A = \int |\mathcal{K}(\zeta)|\, d\zeta \times \sup_t \int |y| f(t, y|X_i)\, dy$ with $f(\cdot, \cdot|X_i)$ as the distribution of $(t, U(X_i, t))$; hence for any $\varepsilon_{1,N} \geq 2\max\{\frac{\mathcal{K}^*_{\max} B_N}{N h_N^2}, \frac{A\left(B_N^{'}\right)^{-1}}{h_N}\}$, we have:

$$\mathbb{P}\Big[\sup_t |\widehat{u}_t(X_i, t) - \widehat{u}_t^{B_N^{'}}(X_i, t) - (\mathbb{E}(\widehat{u}_t(X_i, t) - \widehat{u}_t^{B_N^{'}}(X_i, t)))| > \varepsilon_{1,N}\Big] \leq 2N \exp\left(- \frac{B_N^2}{2\sigma^2}\right) .$$

Component 2. Asymptotic deviation of truncated estimator: Observe that

$$\begin{aligned}\widehat{u}_t^{B_N^{'}} - \mathbb{E}\big(\widehat{u}_t^{B_N^{'}}\big) &= \frac{1}{\sqrt{N} h_N^2} \int_{z\in\mathbb{R}} \int_{|y|\leq B_N^{'}} \mathcal{K}^*\left(\frac{z-t}{h_N}\right) y d_z d_y \underbrace{\left(\sqrt{N}(f_N(z,y) - f(z,y))\right)}_{:=Z_N(z,y)} \\ &= \frac{1}{\sqrt{N} h_N^2} \int_{z\in\mathbb{R}} \mathcal{K}^*\left(\frac{z-t}{h_N}\right) d_z U^{B_N^{'}}(z), \qquad \text{(B.11)}\end{aligned}$$

where $U^{B_N^{'}}(z)$ is defined by

$$U^{B_N^{'}}(z) := \int_{|y|\leq B_N^{'}} y d_y Z_N(z,y).$$

Let $\mathcal{T} : \mathbb{R}^2 \to [0,1]^2$ be the Rosenblatt transformation [200], defined as,

$$\mathcal{T}(x,y) = \Big(F_X(x), F_{Y|X}(y|x)\Big),$$

and define $\mathcal{B}$ as the 2-dimensional solution path of the Brownian bridge which takes the transformed $\mathcal{T}(z,y)$ as an argument; then we have

$$U^{B_N^{'}}(z) := \int_{|y|\leq B_N^{'}} y d_y \big\{Z_N(z,y) - \mathcal{B}(\mathcal{T}(z,y))\big\} + \int_{|y|\leq B_N^{'}} y d_y \mathcal{B}(\mathcal{T}(z,y)). \qquad \text{(B.12)}$$

Plug in (B.12) to (B.11), we get

$$\begin{aligned}
\widehat{u}_t^{B'_N} - \mathbb{E}\big(\widehat{u}_t^{B'_N}\big) &= \underbrace{\frac{1}{\sqrt{N}h_N^2}\int_{z\in\mathbb{R}}\mathcal{K}^*\bigg(\frac{z-t}{h_N}\bigg)d_z\int_{|y|\le B'_N} y d_y\big\{Z_N(z,y)-\mathcal{B}(\mathcal{T}(z,y))\big\}}_{\gamma_N(t)} \\
&+ \frac{1}{\sqrt{N}}\underbrace{\frac{1}{h_N^2}\int_{z\in\mathbb{R}}\int_{|y|\le B'_N}\mathcal{K}^*\bigg(\frac{z-t}{h_N}\bigg)y d_z d_y \mathcal{B}(\mathcal{T}(z,y))}_{\rho_N(t)} \\
&= \gamma_N(t) + \frac{1}{\sqrt{N}}\rho_N(t).
\end{aligned}$$

In the following, we bound γ_N and $\rho_N(t)/\sqrt{N}$ respectively.

1. *Bound for $\gamma_N(t)$:* Since $\mathcal{K}^*$ has compact support, applying integration by parts on $\gamma_N(t)$ gives

$$\begin{aligned}
\gamma_N(t) &= -\frac{1}{\sqrt{N}h_N^2}\int_{z\in\mathbb{R}}\int_{|y|\le B'_N} y d_y\big\{Z_N(z,y)-\mathcal{B}(\mathcal{T}(z,y))\big\}d_z\mathcal{K}^*\bigg(\frac{z-t}{h_N}\bigg) \\
&\le \frac{2B'_N\mathcal{K}^*_{\max}}{\sqrt{N}h_N^2}\sup_{z,y}\bigg|Z_N(z,y)-\mathcal{B}(\mathcal{T}(z,y))\bigg|. \qquad \text{(B.13)}
\end{aligned}$$

By Tusnady's strong approximation result [98], there exist absolute positive constants a_0, b_0 and c_0 such that

$$\mathbb{P}\bigg[\sup_{z,y}\bigg|Z_N(z,y)-\mathcal{B}(\mathcal{T}(z,y))\bigg| > \frac{\big(a_0\ln N + r\big)\ln N}{\sqrt{N}}\bigg] < b_0\exp(-c_0 r) \qquad \text{(B.14)}$$

holds for any real r. Therefore, if we take $\varepsilon'_{2,N}(r) = \frac{2B'_N\mathcal{K}^*_{\max}(a_0\ln N + r)\ln N}{Nh_N^2}$, combining (B.13) and (B.14) gives

$$\mathbb{P}\bigg[\sup_t|\gamma_N(t)| > \varepsilon'_{2,N}(r)\bigg] < b_0\exp(-c_0 r). \qquad \text{(B.15)}$$

2. *Bound for $\rho_N(t)/\sqrt{N}$:* Similarly to (7) of [97], we have

$$\frac{h_N^{3/2}\sup_t|\rho_N(t)|}{\sqrt{\ln\frac{1}{h_N}}} \leq \underbrace{16(\ln V)^{1/2}S^{1/2}\Big(\ln\frac{1}{h_N}\Big)^{-1/2}\int|\zeta|^{1/2}\,|d\mathcal{K}^*(\zeta)|}_{:=Q_{1,N}}$$
$$+\underbrace{16\sqrt{2}h_N^{-1/2}\Big(\ln\frac{1}{h_N}\Big)^{-1/2}\int q(Sh_N|\zeta|)\,|d\mathcal{K}^*(\zeta)|}_{:=Q_{2,N}},$$

where V is a random variable satisfying $\mathbb{E}V \leq 4\sqrt{2}\eta^4$ (recall that $\eta^2 := \max_{i,n}\mathbb{E}(U_i^n)^2$), $q(r) := \int_0^r \frac{1}{2}(\frac{1}{y}\ln\frac{1}{y})^{1/2}\,dy$, $S := \sup_z\int y^2 f(z,y)\,dy$. Let $d_0 = 16\sqrt{2}S^{1/2}\int|\zeta|^{1/2}|d\mathcal{K}^*(\zeta)|$, which is a positive number independent of either N or M. Consider the following inequality for an arbitrary ε

$$\begin{aligned}\mathbb{P}\Bigg(\frac{h_N^{3/2}\sup_t|\rho_N(t)|}{\sqrt{\ln\frac{1}{h_N}}}\geq\varepsilon\Bigg) &\leq \mathbb{P}\Big(Q_{1,N}\geq\frac{\varepsilon}{2}\Big)+\mathbb{P}\Big(Q_{2,N}\geq\frac{\varepsilon}{2}\Big)\\ &\leq \mathbb{P}\Bigg(\big(\ln V\big)^{1/2}\geq\frac{\varepsilon\big(\ln\frac{1}{h_N}\big)^{1/2}}{2d_0}\Bigg)+\mathbb{P}\Big(Q_{2,N}\geq\frac{\varepsilon}{2}\Big)\\ &\leq 4\sqrt{2}\eta^4\exp\Bigg(-\frac{\varepsilon^2\big(\ln\frac{1}{h_N}\big)}{4d_0^2}\Bigg)+\mathbb{P}\Big(Q_{2,N}\geq\frac{\varepsilon}{2}\Big),\end{aligned} \tag{B.16}$$

where the Markov Inequality is used in the last inequality. Setting $\varepsilon''_{2,N} = \varepsilon\sqrt{\frac{\ln\frac{1}{h_N}}{Nh_N^3}}$ gives

$$\mathbb{P}\Bigg(\frac{\sup_t|\rho_N(t)|}{\sqrt{N}}\geq\varepsilon''_{2,N}\Bigg)\leq 4\sqrt{2}\eta^4\exp\Bigg(-\frac{\varepsilon^2\big(\ln\frac{1}{h_N}\big)}{4d_0^2}\Bigg)+\mathbb{P}\Big(Q_{2,N}\geq\frac{\varepsilon}{2}\Big).$$

Notice that $Q_{2,N}$ converges to d_0 by Silverman [101]. For any arbitrary $\alpha > 1$, if $\varepsilon = 2\sqrt{\alpha}d_0$, there exists a positive integer $N(\alpha)$ such that as long as $N > N(\alpha)$, we have $Q_{2,N} < \sqrt{\alpha}d_0$; hence the second probability in (B.16) becomes 0. Considering that $\varepsilon''_{2,N}$ now depends on α, we write it as $\varepsilon''_{2,N}(\alpha)$, and for sufficiently large N

$(N > N(\alpha))$, we obtain

$$\mathbb{P}\left(\frac{\sup_t |\rho_N(t)|}{\sqrt{N}} \geq \varepsilon''_{2,N}(\alpha)\right) \leq 4\sqrt{2}\eta^4 h_N^\alpha \ . \tag{B.17}$$

Now if we take $\varepsilon_{2,N}(r,\alpha) = 2\max\{\varepsilon'_{2,N}(r), \varepsilon''_{2,N}(\alpha)\}$ and combine (B.15) with (B.17), we have

$$\mathbb{P}\left(\sup_t |\widehat{u}_t^{B'_N} - \mathbb{E}\big(\widehat{u}_t^{B'_N}\big)| > \varepsilon_{2,N}(r,\alpha)\right) < b_0\exp(-c_0 r) + 4\sqrt{2}\eta^4 h_N^\alpha$$

Component 3. Asymptotic bias: From [67], the asymptotic bias of the estimator directly follows

$$\mathbb{E}\big(\widehat{u}_t\big) - u_t = C^* h_N^2 \ .$$

for some constant C^* independent of N. Specifically, since we fit a degree 2 polynomial to obtain $\widehat{u}_t(X_i, \cdot)$, we plug $p = 2$ and $\nu = 1$ in the expression of asymptotic bias of the estimator. See page 83 of the paper [67] for the expression. Taking $\varepsilon_{3,N} = |C^*| h_N^2$, we have $\mathbb{P}\left(|\mathbb{E}\big(\widehat{u}_t\big) - u_t| > \varepsilon_{3,N}\right) = 0$.

Combining all the three components above and taking $\varepsilon_N^*(r,\alpha) > 3\max\{\varepsilon_{1,N}, \varepsilon_{2,N}(r,\alpha), \varepsilon_{3,N}\}$ gives the desired result. □

B.3.2 Sufficient conditions for bounding $(\widehat{\mathbf{F}} - \mathbf{F})\beta^*$

For the p-th order partial derivative estimators with respect to x, we have results similarly to Lemma B.3.2.

Lemma B.3.3 *Fix an order $p \geq 0$, and let B_M be an arbitrary increasing sequence $B_M \to \infty$ as $M \to \infty$, and $B'_M = B_M + \|u\|_{L^\infty(\Omega)}$. For any $n = 0, 1, \ldots, N-1$ and arbitrary r, there exist finite positive constants $A_p(t_n), C^*(t_n), a_0, b_0, c_0$, and $d_0(t_n)$ which do not* *depend on the spacial sample size M, such that for any $\alpha > 1$ and*

$$\varepsilon^*_{M,p}(t_n, r, \alpha) > \max\left\{ 3|C^*(t_n)|w_M^2, \frac{6p!\mathcal{K}^*_{\max} B_M^{'}}{M w_M^{1+p}}, 6\frac{p! A_p(t_n)(B_M^{'})^{-1}}{w_M^p}, \frac{6p! B_M^{'}(a_0 \ln M + r)\ln M}{w_M^{1+p} M}, 12p!\sqrt{\alpha} d_0(t_n)\sqrt{\frac{\ln 1/w_M}{w_M^{2p+1} M}} \right\},$$

as long as $M > M(\alpha)$ for some positive integer $M(\alpha)$, we have:

$$\mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\widehat{\partial_x^p u}(x,t_n) - \partial_x^p u(x,t_n)| > \varepsilon^*_{M,p} \Big] < 2M \exp\left(-\frac{B_M^2}{2\sigma^2} \right) + b_0 \exp(-c_0 r) + 4\sqrt{2}\eta^4 w_M^{\alpha} \ .$$

Proof. Notice that for any fixed temporal point t_n, $n = 0, 1, \dots, N-1$, the estimation for the p-th order partial derivative takes the form

$$\widehat{\partial_x^p u}(x,t_n) = \frac{p!}{M w_M^{p+1}} \sum_{i=1}^{M} \mathcal{K}^*\left(\frac{X_i - x}{w_M} \right) U_i^n \tag{B.18}$$

with probability 1 [68]. Hence, we can prove the desired result by substituting h_N^2 with $w_M^{p+1}/p!$ in (B.10) and follow the proof of Lemma B.3.2 and keeping in mind that the constants now depend on t_n and not on M. Notice that the kernel $\mathcal{K}$ used for the spacial dimension may be different from that used for the temporal; this can be addressed by taking $\mathcal{K}^*_{\max}$ to be the larger value between their ℓ_∞-norms. Finally, given any fixed t_n, the asymptotic bias takes the form

$$\mathbb{E}\big(\widehat{\partial_x^p u}\big) - \partial_x^p u = C_p^* w_M^2$$

where $C_p^* \leq \max_{p=0,1,\dots,P_{\max}} \left\{ \int z^{p+1}\mathcal{K}_p^*(z)\,dz \right\} \frac{p!}{(p+2)!}\partial_x^{p+1} u := C^*$ for any $0 \leq p \leq P_{\max}$. Here, since we fit the Local-Polynomial with degree $\ell + 1$ to obtain $\widehat{\partial_x^{\ell} u}$, we plug $p = \ell + 1$

and $\nu = \ell$ in the expression of asymptotic bias in [67]. □

As for the product terms:

Lemma B.3.4 *Fix any two orders $p, q \geq 0$, and let B_M be an arbitrary increasing sequence $B_M \to \infty$ as $M \to \infty$, and $B_M^{'} = B_M + \|u\|_{L^\infty(\Omega)}$. For any $n = 0, 1, \ldots, N-1$ and arbitrary r, there exist finite positive constants $A(t_n), C^*(t_n), a_0, b_0, c_0,$ and $d_0(t_n)$ which do not depend on the spacial sample size M, such that for any $\alpha > 1$ and*

$$\varepsilon_{M,p,q}^{**} > \max\{3\|\partial_x^p u(\cdot,t_n)\|_\infty \varepsilon_{M,p}^*,\ 3\|\partial_x^q u(\cdot,t_n)\|_\infty \varepsilon_{M,q}^*,\ 3(\varepsilon_{M,p}^*)^2,\ 3(\varepsilon_{M,q}^*)^2\}$$

as long as $M > M(\alpha)$ for some positive integer $M(\alpha)$, we have:

$$\begin{aligned}&\frac{1}{4}\mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)| > \varepsilon_{M,p,q}^{**}\Big]\\ &< 2M\exp(-\frac{B_M^2}{2\sigma^2}) + b_0\exp(-c_0 r) + 4\sqrt{2}\eta^4 w_M^\alpha\ ,\end{aligned}$$

Here $\varepsilon_{M,p}^*$ and $\varepsilon_{M,q}^*$ (depending on B_M') are the thresholds in Lemma B.3.3 for the sup-norm bound of the estimator $\widehat{\partial_x^p u}$ and $\widehat{\partial_x^q u}$, respectively,

Proof. Notice that for any $\varepsilon > 0$, we can bound the probability:

$$\begin{aligned}&\mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)| > \varepsilon\Big]\\ &\leq \mathbb{P}\Big[\|\partial_x^p u(\cdot,t_n)\|_\infty \sup_{x\in[0,X_{\max})} |\Delta\partial_x^q u(x,t_n)| > \varepsilon/3\Big]\\ &+ \mathbb{P}\Big[\|\partial_x^q u(\cdot,t_n)\|_\infty \sup_{x\in[0,X_{\max})} |\Delta\partial_x^p u(x,t_n)| > \varepsilon/3\Big]\\ &+ \mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\Delta\partial_x^p u(x,t_n)| > \sqrt{\frac{\varepsilon}{3}}\Big] + \mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\Delta\partial_x^q u(x,t_n)| > \sqrt{\frac{\varepsilon}{3}}\Big],\end{aligned}$$

hence the results follow from Lemma B.3.3. □

As for higher degree terms, we can take the similar approach to obtain general results but with more complicated notations. In this work, we focus on demonstrating the essence without involving more indices.

B.3.3 Simplification on the Probability Bounds

Before proceeding further, we simplify the expressions for ε_N^* as well as the probability bounds in Lemma B.3.2 by considering the window width h_N and the diverging sequence B_N as follows

$$h_N = \frac{1}{N^a},\ B_N = N^b\ .$$

Here $a, b > 0$ are positive coefficients to be determined.

Consequently, we update the expressions of the five terms whose maximum defines the threshold ε_N^*

$$E_1(N) = \frac{3|C^*(X_i)|}{N^{2a}}, \quad E_2(N) = \frac{6\mathcal{K}^*_{\max}(N^b + \|u\|_{L^\infty(\Omega)})}{N^{1-2a}}, \quad E_3(N) = \frac{6A(X_i)}{N^{-a}\big(N^b + \|u\|_{L^\infty(\Omega)}\big)}$$
$$E_4(N) = \frac{6\mathcal{K}^*_{\max}(N^b + \|u\|_{L^\infty(\Omega)})(a_0 \ln N + r)\ln N}{N^{1-2a}}, \quad E_5(N) = 12\sqrt{\alpha}d_0(X_i)\sqrt{\frac{a\ln N}{N^{1-3a}}}\ .$$

When N is sufficiently large, to determine ε_N^*, we only need to focus on comparing the powers of N in $E_i(N)$, $i = 1, 2, \cdots, 5$; this immediately leads to:

$$E_2(N) = \mathcal{O}\left(E_4(N)\right),$$

hence it's sufficient to only consider $E_1(N)$, $E_2(N)$, $E_4(N)$, and $E_5(N)$. The optimal choice of a and b is determined by requiring

$$\begin{cases} 2a = 1 - b - 2a \\ 2a = \frac{1-3a}{2} \end{cases} \Longrightarrow \begin{cases} a = \frac{1}{7} \\ b = \frac{3}{7} \end{cases}$$

To summarize the discussion above, we have

Corollary B.3.5 *Let $h_N = N^{-1/7}$. For any $i = 0, 1, \ldots, M$ and arbitrary real r, there exist finite positive constants $C^*(X_i), a_0, b_0, c_0,$ and $d_0(X_i)$ which do not depend on the temporal sample size N, such that for N sufficiently large, any $\alpha > 1$, and*

$$\varepsilon_N^*(X_i, r, \alpha) > N^{-\frac{2}{7}} \max\left\{3|C^*(X_i)|, 6(a_0 \ln N + r)\ln N, 12\sqrt{\alpha} d_0(X_i)\sqrt{\frac{\ln N}{7}}\right\} ,$$

we have:

$$\mathbb{P}\Big[\sup_{t\in[0,T]} |\Delta u_t(X_i, t)| > \varepsilon_N^*(X_i, r, \alpha)\Big] < 2N \exp\left(-\frac{N^{6/7}}{2\sigma^2}\right) + b_0 \exp(-c_0 r) + 4\sqrt{2}\eta^4 N^{-\alpha/7} ,$$

Similarly, we can obtain optimal $w_M = M^{-1/(2p+5)}$ and $B_M = M^{(p+2)/(2p+5)}$ for the estimation of p-th partial derivative of u. Consequently, the threshold lower bound in Lemma B.3.3 becomes

$$\varepsilon_{M,p}^*(t_n, r, \alpha) > M^{-2/(2p+5)} \max\left\{3|C^*(t_n)|, 6p!(a_0 \ln M + r)\ln M, 12p!\sqrt{\alpha} d_0(t_n)\sqrt{\frac{\ln M}{2p+5}}\right\} .$$

Notice that the right hand side of the inequality above is non-decreasing with respect to $p \geq 0$. Moreover, note that for sufficiently large M, if the probability bound in Lemma B.3.3 holds for some w_M, then it holds for any smaller window width $w'_M < w_M$. Therefore, we have the following simplified result

Corollary B.3.6 *Let $w_M = M^{-1/7}$. For any $n = 0, 1, \ldots, N-1$ and arbitrary r, there exist finite positive constants $C^*(t_n), a_0, b_0, c_0,$ and $d_0(t_n)$ which do not depend on the spacial* *sample size M, such that for M sufficiently large, any $\alpha > 1$, and*

$$\varepsilon_M^*(t_n, r, \alpha) >$$
$$M^{-\frac{2}{2P_{\max}+5}} \max\left\{3|C^*(t_n)|, 6P_{\max}!(a_0 \ln M + r)\ln M, 12P_{\max}!\sqrt{\alpha}d_0(t_n)\sqrt{\frac{\ln M}{2P_{\max}+5}}\right\},$$

we have:

$$\mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\widehat{\partial_x^p u}(x,t_n) - \partial_x^p u(x,t_n)| > \varepsilon_M^*\Big]$$
$$< 2M\exp\left(-\frac{M^{(2P_{\max}+4)/(2P_{\max}+5)}}{2\sigma^2}\right) + b_0\exp(-c_0 r) + 4\sqrt{2}\eta^4 M^{-\alpha/(2P_{\max}+5)}$$

for any order $0 \le p \le P_{\max}$.

Similarly, for the product terms, we have

Corollary B.3.7 *Let $w_M = M^{-1/7}$. For any $n = 0, 1, \dots, N-1$ and arbitrary r, there exist finite positive constants $C^*(t_n), a_0, b_0, c_0$, and $d_0(t_n)$ which do not depend on the spacial sample size M, such that for M sufficiently large, any $\alpha > 1$, and*

$$\varepsilon_M^{**} > \max\{3\|u(\cdot,t_n)\|_{P_{\max},\infty}\varepsilon_M^*,\ 3(\varepsilon_M^*)^2\}$$

where $\|u(\cdot,t_n)\|_{P_{\max},\infty} = \sum_{0\le k\le P_{\max}} \|\partial_x^k u(\cdot,t_n)\|_\infty$, we have

$$\frac{1}{4}\mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)| > \varepsilon_M^{**}\Big]$$
$$< 2M\exp\left(-\frac{M^{(2P_{\max}+4)/(2P_{\max}+5)}}{2\sigma^2}\right) + b_0\exp(-c_0 r) + 4\sqrt{2}\eta^4 M^{-\alpha/(2P_{\max}+5)}$$

for any orders $0 \le p, q \le P_{\max}$.

B.3.4 ℓ_∞ Bound for the PDE Estimation Error τ

Notice that in the previous results, although the constants $C^*(X_i)$ and $d_0(X_i)$ are independent of N, they show dependence on the spacial point X_i. Similarly, $C^*(t_n)$ and $d_0(t_n)$ are independent of M, yet their values may depend on N. To guarantee that as both $N, M \to \infty$, these constants are uniformly bounded, we prove the following lemma.

Lemma B.3.8 *For any integer $M \geq 1$, and any $i = 0, 1, \cdots, M-1$, $|C^*(X_i)|$ and $d_0(X_i)$ in Corollary B.3.5 are bounded by constants that are independent of M. That is, there exist constants $C^*, d_0 > 0$ such that for any $M \geq 1$*

$$\max_{i=0,\cdots,M-1} |C^*(X_i)| \leq C^* \|\partial_t^3 u\|_\infty, \textit{ and } \max_{i=0,\cdots,M-1} d_0(X_i) \leq d_0 \ .$$

Proof. From (3.7) in the Theorem 3.1 of [68], we have

$$|C^*(X_i)| \leq C^* \|\partial_t^3 u\|_\infty < \infty$$

where C^* only depends on the choice of the kernel function and the order of the Local-Polynomial. Recalling that $d_0(X_i) = 16 S^{1/2} \int |\zeta|^{1/2} |d\mathcal{K}^*(\zeta)|$ where $S = \sup_z \int y^2 f(z, y|X_i)\, dy$. For a general real number s, we know that

$$\begin{aligned}
&\sup_{z\in[0,T_{\max}]} \int |y|^s f(z, y|X_i)\, dy = \sup_{z\in[0,T_{\max}]} \int |y|^s \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y - u(X_i, z))^2}{2\sigma^2} \right) dy \\
&= \sup_{z\in[0,T_{\max}]} \sigma^s 2^{s/2} \frac{\Gamma\left(\frac{1+s}{2}\right)}{\sqrt{\pi}} {}_1F_1\left(-\frac{s}{2}, \frac{1}{2}, -\frac{1}{2}\left(\frac{u(X_i, z)}{\sigma}\right)^2 \right)
\end{aligned}$$

where ${}_1F_1(p, q, w)$ is Kummer's confluent hyper-geometric function of $w \in \mathbb{C}$ with parameters $p, q \in \mathbb{C}$ (See, e.g.[201]) and Γ is the Gamma function. Since ${}_1F_1(-\frac{s}{2}, \frac{1}{2}, \cdot)$ is an entire function for fixed parameters,

$$
\begin{aligned}
&\sup_{z\in[0,T_{\max}]}\int |y|^s f(z,y|X_i)\,dy \leq \\
&\qquad \sup_{z\in[0,T_{\max}]} \sigma^s 2^{s/2}\frac{\Gamma\left(\frac{1+s}{2}\right)}{\sqrt{\pi}} \sup_{w\in[-\frac{\max_{x\in\Omega}u^2(x,z)}{2\sigma^2},-\frac{\min_{x\in\Omega}u^2(x,z)}{2\sigma^2}]} {}_1F_1(-\frac{s}{2},\frac{1}{2},w) < \infty
\end{aligned}
$$

which clearly does not depend on M. Taking $s = 2$, we can obtain that $d_0(X_i) \leq d_0$ for some d_0 that only depends on the choice of kernel $\mathcal{K}$, underlying function $\|u\|_{L^\infty(\Omega)}$, and noise level σ. $\square$

Note that the same proof can derive that the constants in Lemma B.3.3 and Lemma B.3.4 are also bounded by N-independent constants. This technical lemma allows us to state

Proposition B.3.9 *Take $h_N = N^{-1/7}$ in the temporal direction and $w_M = M^{-1/7}$ in the space direction. There exist constants C, a_0, b_0, and c_0 which do not depend on N nor M such that for N and M sufficiently large, any r, $\alpha > 1$, and*

$$
\begin{aligned}
&\varepsilon_{N,M}(r,\alpha) > \\
&C\max\left\{\frac{(a_0\ln N + r)\ln N}{N^{2/7}}, \frac{\sqrt{\alpha\ln N}}{N^{2/7}}, \frac{(a_0\ln M + r)\ln M}{M^{2/(2P_{\max}+5)}}, \sqrt{\frac{\alpha\ln M}{(2P_{\max}+5)M^{4/(2P_{\max}+5)}}}\right\}
\end{aligned}
$$

we have

$$
\begin{aligned}
&\mathbb{P}\Big[\|\boldsymbol{\tau}\|_\infty > \varepsilon_{N,M}\Big] < \\
&2NM\exp\left(-\frac{N^{6/7}}{2\sigma^2}\right) + b_0\exp(-c_0 r)M + 4\sqrt{2}\eta^4 MN^{-\alpha/7} + \\
&8sNM\exp\left(-\frac{M^{(2P_{\max}+4)/(2P_{\max}+5)}}{2\sigma^2}\right) + 4sb_0\exp(-c_0 r)N + 16\sqrt{2}\eta^4 sNM^{-\alpha/(2P_{\max}+5)}
\end{aligned}
$$

Here K is the number of feature variables in the dictionary.

Proof. By triangle inequality, the ℓ_∞-norm of PDE estimation error τ (3.6) can be bounded by

$$\|\boldsymbol{\tau}\|_\infty \le \|\Delta\mathbf{F}\boldsymbol{\beta}^*\|_\infty + \|\Delta\mathbf{u}_t\|_\infty \,.$$

By Corollary B.3.5 and Lemma B.3.8, there exists a constant C_1 independent of N and M such that with sufficiently large N and any $\varepsilon_N(r,\alpha) > C_1 N^{-2/7}\max\{(a_0\ln N + r)\ln N, \sqrt{\alpha\ln N}\}$, we have

$$\begin{aligned}
\mathbb{P}\Big[\|\Delta\mathbf{u}_t\|_\infty > \varepsilon_N(r,\alpha)\Big] &\le \mathbb{P}\Big[\max_{i=0,1,\cdots,M-1}\sup_{t\in[0,T_{\max}]}|\Delta u_t(X_i,t)| > \varepsilon_N(r,\alpha)\Big] \\
&\le \sum_{i=0}^{M-1}\mathbb{P}\Big[\sup_{t\in[0,T_{\max}]}|\Delta u_t(X_i,t)| > \varepsilon_N(r,\alpha)\Big] \\
&< 2NM\exp\left(-\frac{N^{6/7}}{2\sigma^2}\right) + b_0\exp(-c_0 r)M + 4\sqrt{2}\eta^4 M N^{-\alpha/7} \,.
\end{aligned}$$

On the other hand, if we denote $\Delta F_k(x,t)$ as the approximation error of the k-th feature variable at time t and space x, we have

$$\|\Delta\mathbf{F}\boldsymbol{\beta}^*\|_\infty \le \max_{n=0,1,\cdots,N}\|\boldsymbol{\beta}^*\|_\infty \sup_{x\in[0,X_{\max})}\sum_{k=1}^{s}|\Delta F_k(x,t_n)| \,.$$

By Corollary B.3.6 and B.3.7, there exists a constant C_2 independent of N and M such that with sufficiently large M and any $\varepsilon_{K,M}(r,\alpha) > C_2 P_{\max}! K\|\boldsymbol{\beta}^*\|_\infty M^{-2/(2P_{\max}+5)}\max\{(a_0\ln M + r)\ln M, \sqrt{\frac{\alpha\ln M}{2P_{\max}+5}}\}$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\|\Delta\mathbf{F}\boldsymbol{\beta}^*\|_\infty > \varepsilon_M(r,\alpha)\Big] \le \sum_{n=0}^{N-1}\sum_{k=1}^{s}\mathbb{P}\Big[\sup_{x\in[0,X_{\max})}|\Delta F_k(x,t_n)| > \frac{\varepsilon_M(r,\alpha)}{s\|\boldsymbol{\beta}^*\|_\infty}\Big] \\
&< 8NMs\exp\left(-\frac{M^{(2P_{\max}+4)/(2P_{\max}+5)}}{2\sigma^2}\right) + 4b_0\exp(-c_0 r)Ns + 16\sqrt{2}\eta^4 NsM^{-\alpha/(2P_{\max}+5)} \,.
\end{aligned}$$

Taking $C = \max\{2C_1, 2s\|\boldsymbol{\beta}^*\|_\infty C_2 P_{\max}!\}$ proves the theorem. □

B.3.5 Further Simplification

We further simplify our result by taking $M = N^b$ for some coefficient $b > 0$. Since r and α are arbitrary, we can vary them as we increase M, N by taking $r = N^c$ and $\alpha = N^d$ for some positive coefficients $c > 0$ and $d > 0$, respectively. Consequently, we have the lower bound for $\varepsilon_{N,M}$ in Proposition B.3.9 becoming

$$\varepsilon_{N,M}(r,\alpha) > \\ C \max\left\{ \frac{(a_0 \ln N + N^c)\ln N}{N^{2/7}}, \frac{\sqrt{\ln N}}{N^{2/7-d/2}}, \frac{b(a_0 b \ln N + N^c)\ln N}{N^{2b/(2P_{\max}+5)}}, \sqrt{\frac{b \ln N}{(2P_{\max}+5)N^{4b/(2P_{\max}+5)-d}}} \right\}, \tag{B.19}$$

To guarantee that the lower bound (B.19) converges to 0 as $N \to \infty$, we have the following constraints on positive coefficients b, c, and d

$$\begin{cases} 0 < c < 2/7 \\ 2/7 - d/2 > 0 \\ c < 2b/(2P_{\max}+5) \\ 4b/(2P_{\max}+5) - d > 0 \end{cases}$$

Furthermore, we take $d = 2c$ so that

$$\frac{\sqrt{\ln N}}{N^{2/7-d/2}} = \mathcal{O}\left(\frac{(a_0 \ln N + N^c)\ln N}{N^{2/7}}\right), \quad \sqrt{\frac{b \ln N}{N^{4b/(2P_{\max}+5)-d}}} = \mathcal{O}\left(\frac{b(a_0 b \ln N + N^c)\ln N}{N^{2b/(2P_{\max}+5)}}\right).$$

and we can focus on the second and fourth term in (B.19). As a result, the optimal choice for b is computed by $2/7 = 2b/(2P_{\max}+5) \implies b = (2P_{\max}+5)/7$. Based on the set-ups above, we obtain that for N sufficiently large, with

$$\varepsilon_N(c) > C\frac{\ln N}{N^{2/7-c}}$$

for any $0 < c < 2/7$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\|\boldsymbol{\tau}\|_\infty > \varepsilon_N(c)\Big] < \\
&\quad 2N^{(2P_{\max}+12)/7}\exp\left(-\frac{N^{6/7}}{2\sigma^2}\right) + b_0\exp(-c_0N^c)N^{(2P_{\max}+5)/7} + 4\sqrt{2}\eta^4N^{-N^{2c}/7} + \\
&\quad 8N^{(2P_{\max}+12)/7}K\exp\left(-\frac{N^{(2P_{\max}+5)/7}}{2\sigma^2}\right) + 4b_0\exp(-c_0N^c)NK + 16\sqrt{2}\eta^4KN^{-N^{2c}/7} \\
&\quad = \mathcal{O}\left(N^{\frac{2P_{\max}+5}{7}}\exp\Big(-\frac{1}{6}N^c\Big)\right),
\end{aligned}$$

where in the last equality, we plug $b_0 = 2$ and $c_0 = \frac{1}{6}$ from [202]. Combining this with Lemma B.3.1 proves the first part of the Proposition 1.

B.3.6 Proof of ℓ_∞ bound in (4.1)

Recall that in (B.9), we have

$$\check{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}} = \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\Big(\widehat{\mathbf{F}}_{\mathcal{S}}^T(\Delta\mathbf{u}_t - \Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}}) - \lambda_N NM\check{\mathbf{z}}_{\mathcal{S}}\Big).$$

Now, we are ready to bound the $\left\|\widehat{\boldsymbol{\beta}}^\lambda_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\right\|_\infty$ bound in (4.1) as follows:

$$\begin{aligned}
\max_{k\in\mathcal{S}}|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*_k| &\le \left\|\Big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\Big)^{-1}\right\|_2\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\boldsymbol{\tau}\|_\infty + \lambda_N NM\left\|\Big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\Big)^{-1}\right\|_2 \\
&\le \left\|\Big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}/(NM)\Big)^{-1}\right\|_2\Big(\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\boldsymbol{\tau}\|_\infty/(NM) + \lambda_N\Big) \\
&\overset{\text{(A3)}}{\le} \sqrt{K}C_{\min}\Big(\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\boldsymbol{\tau}\|_\infty/(NM) + \lambda_N\Big) \\
&\le \sqrt{K}C_{\min}\left(\|\boldsymbol{\tau}\|_\infty\frac{\left\|\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_{\infty,\infty}}{NM} + \lambda_N\right) \\
&\le \sqrt{K}C_{\min}\left(\|\boldsymbol{\tau}\|_\infty\frac{\left\|\widehat{\mathbf{F}}\right\|_F}{\sqrt{NM}} + \lambda_N\right) \le \sqrt{K}C_{\min}\Big(K\|\boldsymbol{\tau}\|_\infty + \lambda_N\Big),
\end{aligned}$$

where we use normalized columns of $\widehat{\mathbf{F}}$ in the last inequality. Following the set-ups from Proposition 1 gives the desired result.

B.4 Proofs of Corollaries B.4.1, B.4.2, and Lemma 6.1

Corollary B.4.1 *Fix any four orders $p, q, k \geq 0$, and let B_M be an arbitrary increasing sequence $B_M \to \infty$ as $M \to \infty$, and $B_M' = B_M + \|u\|_{L^\infty(\Omega)}$. For any $n = 0, 1, \ldots, N-1$ and arbitrary r, there exist finite positive constants $A(t_n), C^*(t_n), a_0, b_0, c_0,$ and $d_0(t_n)$ which do not depend on the spacial sample size M, such that for any $\alpha > 1$ and*

$$\varepsilon^{***}_{M,p,q,k} > \max\left\{3\|\partial_x^k u(\cdot, t_n)\|_\infty \varepsilon^{**}_{M,p,q},\ 3\|\partial_x^p u(\cdot, t_n)\partial_x^q u(\cdot, t_n)\|_\infty \varepsilon^{**}_{M,k},\ 3(\varepsilon^{**}_{M,p,q})^2,\ 3(\varepsilon^{**}_{M,k})^2\right\}$$

as long as $M > M(\alpha)$ for some positive integer $M(\alpha)$, we have:

$$\frac{1}{4}\mathbb{P}\left[\sup_{x\in[0,X_{\max})}\left|\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n)\widehat{\partial_x^k u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)\partial_x^k u(x,t_n)\right| > \varepsilon^{***}_{M,p,q,k}\right]$$
$$< 8M\exp\left(-\frac{M^{(2P_{\max}+4)/(2P_{\max}+5)}}{2\sigma^2}\right) + 4b_0\exp(-c_0 r) + 16\sqrt{2}\eta^4 M^{-\alpha/(2P_{\max}+5)}\ ,$$

Here $\varepsilon^{**}_{M,p,q}$ and $\varepsilon^{**}_{M,k,l}$ (depending on B_M') are the thresholds in Corollary B.3.7 for the sup-norm bound of the estimator $\widehat{\partial_x^p u \partial_x^q u}$ and $\widehat{\partial_x^k u \partial_x^l u}$, respectively,

Proof. Notice that for any $\varepsilon > 0$, we can bound the probability:

$$\mathbb{P}\left[\sup_{x\in[0,X_{\max})}\left|\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n)\widehat{\partial_x^k u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)\partial_x^k u(x,t_n)\right| > \varepsilon\right]$$
$$\leq \mathbb{P}\left[\|\partial_x^k u(\cdot,t_n)\|_\infty \sup_{x\in[0,X_{\max})}\left|\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)\right| > \varepsilon/3\right]$$
$$+ \mathbb{P}\left[\|\partial_x^p u(\cdot,t_n)\partial_x^q u(\cdot,t_n)\|_\infty \sup_{x\in[0,X_{\max})}\left|\widehat{\partial_x^k u}(x,t_n) - \partial_x^k u(x,t_n)\right| > \varepsilon/3\right]$$
$$+ \mathbb{P}\Big[\sup_{x\in[0,X_{\max})}\left|\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)\right| > \sqrt{\frac{\varepsilon}{3}}\Big]$$
$$+ \mathbb{P}\Big[\sup_{x\in[0,X_{\max})}\left|\widehat{\partial_x^k u}(x,t_n) - \partial_x^k u(x,t_n)\right| > \sqrt{\frac{\varepsilon}{3}}\Big],$$

hence the results follow from corolloary B.3.7. □

Corollary B.4.2 *Fix any four orders $p, q, k, l \geq 0$, and let B_M be an arbitrary increasing sequence $B_M \to \infty$ as $M \to \infty$, and $B_M^{'} = B_M + \|u\|_{L^\infty(\Omega)}$. For any $n = 0, 1, \ldots, N-1$ and arbitrary r, there exist finite positive constants $A(t_n), C^*(t_n), a_0, b_0, c_0,$ and $d_0(t_n)$ which do not depend on the spacial sample size M, such that for any $\alpha > 1$ and*

$$\varepsilon^{****}_{M,p,q,k,l} >$$

$$\max\left\{3\|\partial_x^p u(\cdot,t_n)\partial_x^q u(\cdot,t_n)\|_\infty \varepsilon^{**}_{M,p,q},\ 3\|\partial_x^k u(\cdot,t_n)\partial_x^l u(\cdot,t_n)\|_\infty \varepsilon^{**}_{M,k,l},\ 3(\varepsilon^{**}_{M,p,q})^2,\ 3(\varepsilon^{**}_{M,k,l})^2\right\}$$

as long as $M > M(\alpha)$ for some positive integer $M(\alpha)$, we have:

$$\frac{1}{4}\mathbb{P}\Bigg[\sup_{x\in[0,X_{\max})}\Big|\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n)\widehat{\partial_x^k u}(x,t_n)\widehat{\partial_x^l u}(x,t_n)$$

$$-\partial_x^p u(x,t_n)\partial_x^q u(x,t_n)\partial_x^k u(x,t_n)\partial_x^l u(x,t_n)\Big| > \varepsilon^{****}_{M,p,q,k,l}\Bigg]$$

$$< 8M\exp\left(-\frac{M^{(2P_{\max}+4)/(2P_{\max}+5)}}{2\sigma^2}\right) + 4b_0\exp(-c_0 r) + 16\sqrt{2}\eta^4 M^{-\alpha/(2P_{\max}+5)}\ ,$$

Here $\varepsilon^{**}_{M,p,q}$ and $\varepsilon^{**}_{M,k,l}$ (depending on B'_M) are the thresholds in Corollary B.3.7 for the sup-norm bound of the estimator $\widehat{\partial_x^p u}\widehat{\partial_x^q u}$ and $\widehat{\partial_x^k u}\widehat{\partial_x^l u}$, respectively,

Proof. Proof of this Corollary is similar with that of the Corollary B.4.1. We omit the proof for simplicity. □

Lemma B.4.3 ***(Lemma 2.6.1 in the main paper)*** *Let $\varepsilon^*_M, \varepsilon^{**}_M, \varepsilon^{***}_M, \varepsilon^{****}_M$ be the thresholds defined in corollaries B.3.6, B.3.7, B.4.1, and B.4.2. Then for any $\varepsilon_M^{\mathit{max}'}$ such that*

$$\varepsilon_M^{\mathit{max}'} > \sqrt{s(K-s)}\max\left\{\varepsilon^*_M, \varepsilon^{**}_M, \varepsilon^{***}_M, \varepsilon^{****}_M\right\},$$

then, for $0 < c < \frac{2}{7}$, and for sufficiently large enough N, we have

$$\mathbb{P}\left[\frac{1}{NM}\left\|\widehat{\mathbf{F}}^T_{\mathcal{S}^c}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}^T_{\mathcal{S}^c}\mathbf{F}_{\mathcal{S}}\right\|_2 > \varepsilon_M^{\mathit{max}'}\right] \leq \mathcal{O}\Big(N\exp\big(-\frac{1}{6}N^c\big)\Big).$$

Proof.

$$\begin{aligned}
&\mathbb{P}\Bigg[\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\right\|_2>{\varepsilon_M^{\text{max}}}'\Bigg]\\
&\quad\leq\mathbb{P}\Bigg[\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\right\|_{\text{F}}>NM{\varepsilon_M^{\text{max}}}'\Bigg]\\
&\quad\leq\mathbb{P}\Bigg[\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\right\|_{\infty,\infty}>NM\frac{{\varepsilon_M^{\text{max}}}'}{\sqrt{s(K-s)}}\Bigg]\\
&\quad\leq\mathbb{P}\Bigg[\max_{n=0,\dots,N-1}\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}}_i(x,t_n)\widehat{\mathbf{F}_j}(x,t_n)-\mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right|>\frac{{\varepsilon_M^{\text{max}}}'}{\sqrt{s(K-s)}}\Bigg]\\
&\quad\leq\sum_{n=0}^{N-1}\mathbb{P}\Bigg[\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}}_i(x,t_n)\widehat{\mathbf{F}_j}(x,t_n)-\mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right|>\frac{{\varepsilon_M^{\text{max}}}'}{\sqrt{s(K-s)}}\Bigg]\\
&\quad\leq\mathcal{O}\bigg(N\exp\big(-\frac{1}{6}N^c\big)\bigg)\;,
\end{aligned}$$

where we use the results from corollaries B.3.6, B.3.7, B.4.1, and B.4.2, and simplication argument used in the B.3.5 in the last inequality. □

B.4.1 Proof of Lemma 6.1

Proof. Observe that we can write:

$$\begin{aligned}
\Lambda_{\text{min}}\bigg(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\bigg)&:=\frac{1}{NM}\min_{\|x\|_2=1}x^T\bigg(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\bigg)x\\
&=\frac{1}{NM}\min_{\|x\|_2=1}\bigg\{x^T\bigg(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\bigg)x+x^T\bigg(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}-\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\bigg)x\bigg\}\\
&\leq\frac{1}{NM}\bigg\{y^T\bigg(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\bigg)y+y^T\bigg(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}-\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\bigg)y\bigg\}
\end{aligned}$$

where $y\in\mathbb{R}^K$ is a unit-norm minimal eigen-vector of $\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}$. Therefore, we can write,

$$\begin{aligned}
\Lambda_{\text{min}}\bigg(\frac{1}{NM}\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\bigg)&\geq\Lambda_{\text{min}}\bigg(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\bigg)-\frac{1}{NM}\left\|\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}-\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2\\
&\geq C_{\text{min}}-\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right\|_2.
\end{aligned}$$

By using a similar argument used in Lemma 2.6.1, we can prove $\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right\|_2 \to 0$ with high-probability as $N\to\infty$. For any $\varepsilon_M^{\max}$ such that,

$$\varepsilon_M^{\max} > s\max\left\{\varepsilon_M^{*},\varepsilon_M^{**},\varepsilon_M^{***},\varepsilon_M^{****}\right\},$$

Then, we can bound the probability as follows:

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right\|_2 > \varepsilon_M^{\max}\right]\\
&\quad\le \mathbb{P}\left[\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right\|_{\mathrm{F}} > NM\varepsilon_M^{\max}\right] \le \mathbb{P}\left[\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right\|_{\infty,\infty} > NM\frac{\varepsilon_M^{\max}}{s}\right]\\
&\quad\le \mathbb{P}\left[\max_{n=0,\dots,N-1}\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}}_i(x,t_n)\widehat{\mathbf{F}_j}(x,t_n)-\mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right| > \frac{\varepsilon_M^{\max}}{s}\right]\\
&\quad\le \sum_{n=0}^{N-1}\mathbb{P}\left[\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}}_i(x,t_n)\widehat{\mathbf{F}_j}(x,t_n)-\mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right| > \frac{\varepsilon_M^{\max}}{s}\right]\\
&\quad\le \mathcal{O}\left(N\exp\left(-\frac{1}{6}N^c\right)\right).
\end{aligned}$$

□

B.4.2 Proof of Lemma 6.2

Proof. Motviated from [75], we begin the proof by decomposing the matrix $\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)\left(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}$ into four parts:

$$\begin{aligned}
\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)\left(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} &= \underbrace{\mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\left(\left(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}-\left(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right)^{-1}\right)}_{:=\mathbf{T_1}} + \underbrace{\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right)^{-1}}_{:=\mathbf{T_2}}\\
&\quad+\underbrace{\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\right)\left(\left(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}-\left(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right)^{-1}\right)}_{:=\mathbf{T_3}}\\
&\quad+\underbrace{\left(\mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\right)^{-1}}_{:=\mathbf{T_4}}.
\end{aligned}$$

Since we know $\|\mathbf{T_4}\|_\infty \leq 1-\mu$ for some $\mu \in (0,1]$, the decomposition reduces the proof showing $\|\mathbf{T_i}\|_\infty \to \mathbf{0}$ with probability $1-\mathcal{O}(N\exp(-\frac{1}{6}N^c))$ for $i = 1,2,3$.

1. Control of $\mathbf{T_1}$: Observe that we can re-factorize $\mathbf{T_1}$ as follows:

$$\mathbf{T_1} = \big(\mathbf{F}^{\mathbf{T}}_{\mathcal{S}^c}\mathbf{F}_{\mathcal{S}}\big)\big(\mathbf{F}^{\mathbf{T}}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}\big)^{\mathbf{-1}}\big[\mathbf{F}^{\mathbf{T}}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}^{\mathbf{T}}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\big]\big(\widehat{\mathbf{F}}^{\mathbf{T}}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{\mathbf{-1}}.$$

Then, by taking the advantage of sub-multiplicative property $\|AB\|_\infty \leq \|A\|_\infty\|B\|_\infty$ and the fact $\|\mathbf{T_4}\|_\infty \leq 1-\mu$ and $\|C\|_\infty \leq \sqrt{N}\|C\|_2$ for $C \in \mathbb{R}^{M\times N}$, we can bound $\|\mathbf{T_1}\|_\infty$ as follows:

$$\begin{aligned}
\|\mathbf{T_1}\|_\infty &\leq \left\|\big(\mathbf{F}^{T}_{\mathcal{S}^c}\mathbf{F}_{\mathcal{S}}\big)\big(\mathbf{F}^{T}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}\big)^{-1}\right\|_\infty \left\|\mathbf{F}^{T}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}^{T}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_\infty \left\|\big(\widehat{\mathbf{F}}^{T}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\right\|_\infty \\
&\leq s(1-\mu)\bigg(\frac{1}{NM}\left\|\mathbf{F}^{T}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}^{T}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2\bigg)\bigg(NM\left\|\big(\widehat{\mathbf{F}}^{T}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\right\|_2\bigg) \\
&\leq \frac{s(1-\mu)}{C_{\text{min}}}\bigg(\frac{1}{NM}\left\|\mathbf{F}^{T}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}^{T}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2\bigg).
\end{aligned}$$

Note that we use $\|\big(\widehat{\mathbf{F}}^{T}_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\|_2 \leq \frac{1}{NMC_{min}}$ with probability $1-\mathcal{O}(N\exp(-\frac{1}{6}N^c))$ in the last inequality from Lemma 6.1.

2. Control of $\mathbf{T_2}$: With similar techniques employed for controlling $\|\mathbf{T_1}\|_\infty$, we can bound $\|\mathbf{T_2}\|_\infty$ as follows:

$$\begin{aligned}
\|\mathbf{T_2}\|_\infty &\leq \left\|\widehat{\mathbf{F}}^{T}_{\mathcal{S}^c}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}^{T}_{\mathcal{S}^c}\mathbf{F}_{\mathcal{S}}\right\|_\infty \left\|\big(\mathbf{F}^{T}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}\big)^{-1}\right\|_\infty \\
&\leq s\left\|\widehat{\mathbf{F}}^{T}_{\mathcal{S}^c}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}^{T}_{\mathcal{S}^c}\mathbf{F}_{\mathcal{S}}\right\|_2 \left\|\big(\mathbf{F}^{T}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}\big)^{-1}\right\|_2 \\
&= s\bigg(\frac{1}{NM}\left\|\widehat{\mathbf{F}}^{T}_{\mathcal{S}^c}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}^{T}_{\mathcal{S}^c}\mathbf{F}_{\mathcal{S}}\right\|_2\bigg)\bigg(NM\left\|\big(\widehat{\mathbf{F}}^{T}_{\mathcal{S}}\mathbf{F}_{\mathcal{S}}\big)^{-1}\right\|_2\bigg) \\
&\leq \frac{s}{C_{\text{min}}}\bigg(\frac{1}{NM}\left\|\widehat{\mathbf{F}}^{T}_{\mathcal{S}^c}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}^{T}_{\mathcal{S}^c}\mathbf{F}_{\mathcal{S}}\right\|_2\bigg).
\end{aligned}$$

3. Control of $\mathbf{T_3}$: To bound $\|\mathbf{T_3}\|_\infty$, we re-factorize the second argument of product in $\mathbf{T_3}$:

$$\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} - \big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1} = \big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\big[\big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big) - \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)\big]\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}$$

With the factorization, we bound $\|\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} - \big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\|_\infty$ by using sub-multiplicative property and the fact $\|C\|_\infty \leq \sqrt{N}\|C\|_2$ for any $C \in \mathbb{R}^{M\times N}$ again:

$$\begin{aligned}
\Big\|\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} - \big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\Big\|_\infty &= \Big\|\big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\big[\big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big) - \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)\big]\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\Big\|_\infty \\
&\leq \sqrt{s}\Big\|\big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\big[\big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big) - \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)\big]\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\Big\|_2 \\
&\leq \sqrt{s}\Big\|\big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\Big\|_2\Big\|\big[\big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big) - \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)\big]\Big\|_2\Big\|\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}\Big\|_2 \\
&\leq \frac{\sqrt{s}}{NMC_{\text{min}}^2}\bigg(\frac{1}{NM}\Big\|\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\Big\|_2\bigg). \qquad \text{(B.20)}
\end{aligned}$$

In the last inequality, we use the result of Lemma 6.1. Now we can bound $\|\mathbf{T_3}\|_\infty$ as follows:

$$\begin{aligned}
\|\mathbf{T_3}\|_\infty &= \bigg\|\Big(\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\Big)\Big(\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} - \big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\Big)\bigg\|_\infty \\
&\leq \Big\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\Big\|_\infty\Big\|\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1} - \big(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}\big)^{-1}\Big\|_\infty \\
&\leq \frac{s}{C_{\text{min}}}\bigg(\frac{1}{NM}\Big\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^T\mathbf{F}_{\mathcal{S}}\Big\|_2\bigg)\bigg(\frac{1}{NM}\Big\|\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\Big\|_2\bigg),
\end{aligned}$$

where in the last inequality, we use (2.14) and $\|C\|_\infty \leq \sqrt{N}\|C\|_2$ for any $C \in \mathbb{R}^{M\times N}$. Take $\varepsilon_M^{\text{max}''}$ such that, for $\varepsilon_M^{\text{max}'}$ and $\varepsilon_M^{\text{max}}$ in Lemma 2.6.1 and Lemma 6.1 respectively: $\varepsilon_M^{\text{max}''} > \max\left\{\frac{C_{\text{min}}}{s(1-\mu)}\varepsilon_M^{\text{max}}, \frac{C_{\text{min}}}{s}\varepsilon_M^{\text{max}'}\right\}$, for large enough N, we have

$$\mathbb{P}\bigg[\forall i = 1, 2, 3 : \|\mathbf{T_i}\|_\infty > \varepsilon_M^{\text{max}''}\bigg] \leq \mathcal{O}\Big(N\exp\big(-\frac{1}{6}N^c\big)\Big)\ .$$

□

APPENDIX C
HIGH-DIMENSIONAL MULTIVARIATE LINEAR REGRESSION WITH WEIGHTED NUCLEAR NORM REGULARIZATION

This appendix provides the technical details for the main paper. Sections C.1 to C.6 provides the proofs of Lemma 3.2.1, Theorem 3.2.2, Proposition 3.3.1, Lemma 3.3.2, Theorem 3.3.3, and Proposition 3.4.1. Section C.7 shows an extension of the proposed WMVR-ADMM algorithm for solving trace regression problem with weighted nuclear norm penalization.

C.1 Proof of Lemma 3.2.1

For simplicity, denote $\boldsymbol{B}^{(k)} := -\boldsymbol{\Lambda}^{(k)} + \rho \cdot \boldsymbol{\Gamma}^{(k)}$, then we can solve the optimization problem in **Step 1** as follows:

$$\begin{aligned}
\boldsymbol{\Theta}^{(k+1)} &= \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}}\ \mathcal{L}_\rho\big(\boldsymbol{\Theta}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) \\
&= \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} \left\{ f(\boldsymbol{\Theta}) + \mathbf{tr}\big(\boldsymbol{\Lambda}^{(k)\top}\boldsymbol{\Theta}\big) + \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}}^2 \right\} \\
&= \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} \left\{ \sum_{j=1}^{p} \left(\frac{\rho}{2}\sigma_j(\boldsymbol{\Theta})^2 + \boldsymbol{\lambda}_n \omega_j \cdot \sigma_j(\boldsymbol{\Theta}) \right) - \mathbf{tr}\big(\boldsymbol{B}^{(k)\top}\boldsymbol{\Theta}\big) \right\}. \qquad \text{(C.1)}
\end{aligned}$$

We plugged-in $f(\boldsymbol{\Theta}) = \boldsymbol{\lambda}_n\|\boldsymbol{\Theta}\|_{\boldsymbol{\omega},\star}$, used $\mathbf{tr}\big(\boldsymbol{\Theta}\boldsymbol{\Theta}^\top\big) = \sum_{j=1}^{p}\sigma_j\big(\boldsymbol{\Theta}\big)^2$ and the definition of $\boldsymbol{B}^{(k)}$ for deriving the last equality. For further convenience of notation, let $\{d_j\}_{j=1}^p := \{\sigma_j\big(\boldsymbol{\Theta}\big)\}_{j=1}^p$ and denote $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right singular matrices of $\boldsymbol{\Theta}$ and $\boldsymbol{D} := \operatorname{diag}\big(\{d_1, d_2, \ldots, d_p\}\big)$. Note that the entries in $\boldsymbol{D}$ are in non-increasing order. (i.e. $d_1 \geq d_2 \cdots \geq d_p \geq 0$) Then, we can rewrite the optimization problem in (C.1) of WMVR-ADMM algorithm as follows:

$$\boldsymbol{\Theta}^{(k+1)} = \underset{d_1 \geq d_2 \geq \cdots \geq d_p \geq 0}{\operatorname{argmin}} \left\{ \sum_{j=1}^{p} \left(\frac{\rho}{2} d_j^2 + \boldsymbol{\lambda}_n \omega_j d_j \right) - \max_{\boldsymbol{U}^\top \boldsymbol{U} = \mathcal{I}_{d_1}, \boldsymbol{V}^\top \boldsymbol{V} = \mathcal{I}_{d_2}} \mathbf{tr}\big(\boldsymbol{B}^{(k)\top} \boldsymbol{\Theta} \big) \right\} \tag{C.2}$$

The maximum of second term in (C.2) can be achieved when $\boldsymbol{U}$ and $\boldsymbol{V}$ coincide with left and right singular matrices of $\boldsymbol{B}^{(k)}$ respectively, giving us the maximized value as $\sum_{j=1}^{p} \sigma_j(\boldsymbol{B}^{(k)}) d_j$. This is a well-known Von Neumann's trace inequality. See [203, 204]. Then, the final form of the optimization problem (C.2) reduces to obtaining the diagonal entries of the matrix $\boldsymbol{D}$ by minimizing the following :

$$\min_{d_1 \geq d_2 \geq \cdots \geq d_p \geq 0} \left\{ \sum_{j=1}^{p} \left(\frac{\rho}{2} d_j^2 + \big(\boldsymbol{\lambda}_n \omega_j - \sigma_j(\boldsymbol{B}^{(k)}) \big) d_j \right) \right\}. \tag{C.3}$$

The objective function (C.3) is completely decompsable coordinate-wise and is minimized at $d_j = \max\left\{ \frac{1}{\rho} \big(\sigma_j(\boldsymbol{B}^{(k)}) - \boldsymbol{\lambda}_n \omega_j \big), 0 \right\}$ for $j = 1, \ldots, p$. Since $\sigma_1(\boldsymbol{B}^{(k)}) \geq \sigma_2(\boldsymbol{B}^{(k)}) \cdots \geq \sigma_p(\boldsymbol{B}^{(k)})$ and $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_p$, the solution is feasible. Furthermore, we have an unique minimizer due to the equality condition of von-Neumann's trace inequality when $\boldsymbol{B}^{(k)}$ has distinct non-zero singular values, and the uniqueness of strict convex optimization of (C.3) in d_j for $j = 1, \ldots, p$. □

C.2 Proof of Theorem 3.2.2

To prove Theorem 3.2.2 in the main paper about the convergence of WMVR-ADMM algorithm converges globally, we need to use the following two lemmas whose proofs are given subsequently.

Lemma C.2.1 *Set* $\rho > 2L_{\nabla g}$ *with* $L_{\nabla g} := \sigma_1\big(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\big)$. *Then, the iterates* $\{(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})\}_{k \geq 1}$ *generated from WMVR-ADMM satisfy the following conditions:*

a. $\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big)$ *is lower-bounded and non-increasing over* $k \geq 1$.

b. $\{(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})\}_{k\geq 1}$ *is bounded.*

c. $\left\|\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}\right\|_F \to 0$ *and* $\left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_F \to 0,$ *as* $k \to \infty$*.*

Proof of Lemma B.1: By the result of Lemma 2.1 in the main paper, we have

$$\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) - \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) \geq 0. \tag{C.4}$$

Now, we control the following difference term.

$$\begin{aligned}
&\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) - \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k)}\big) \\
&\quad = g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}\big(\boldsymbol{\Lambda}^{(k)\top}\big(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\big)\big) \\
&\qquad\quad - \rho \cdot \mathbf{tr}\big(\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big)^\top\big(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\big)\big) + \frac{\rho}{2}\left\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\right\|_F^2 \\
&\quad = g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}\big(\boldsymbol{\Lambda}^{(k+1)\top}\big(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\big)\big) + \frac{\rho}{2}\left\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\right\|_F^2.
\end{aligned} \tag{C.5}$$

Note that we use $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big)$ in the last equality. Recall the definition of $\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big)$ from equation (5) in the main paper and $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big)$. Then, we have

$$\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k)}\big) - \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}\big) = -\frac{1}{\rho}\left\|\boldsymbol{\Lambda}^{(k)} - \boldsymbol{\Lambda}^{(k+1)}\right\|_F^2. \tag{C.6}$$

By combining (C.5) and (C.6), we have

$$\begin{aligned}
&\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) - \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}\big) \\
&\quad = g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}\big(\boldsymbol{\Lambda}^{(k+1)\top}\big(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\big)\big) \\
&\qquad\qquad\qquad + \frac{\rho}{2}\left\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\right\|_F^2 - \frac{1}{\rho}\left\|\boldsymbol{\Lambda}^{(k)} - \boldsymbol{\Lambda}^{(k+1)}\right\|_F^2.
\end{aligned} \tag{C.7}$$

Recall the definition of $\boldsymbol{\Gamma}^{(k+1)}$ from **Step 2** of WMVR-ADMM Algorithm.

$$\boldsymbol{\Gamma}^{(k+1)} = \underset{\boldsymbol{\Gamma}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} \left\{ g(\boldsymbol{\Gamma}) - \mathbf{tr}\big(\boldsymbol{\Lambda}^{(k)\top}\boldsymbol{\Gamma}\big) + \frac{\rho}{2}\left\|\boldsymbol{\Gamma} - \boldsymbol{\Theta}^{(k+1)}\right\|_F^2 \right\}.$$

Since $\boldsymbol{\Gamma}^{(k+1)}$ is a stationary point of the above optimization problem, we have

$$\nabla g\big(\boldsymbol{\Gamma}^{(k+1)}\big) = \boldsymbol{\Lambda}^{(k)} + \rho\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big) = \boldsymbol{\Lambda}^{(k+1)},$$

where $\nabla g(\cdot)$ is a gradient of g. Likewise, we get $\nabla g\big(\boldsymbol{\Gamma}^{(k)}\big) = \boldsymbol{\Lambda}^{(k)}$. Recall the definition of $g(\cdot)$, then we can easily have

$$\left\|\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}\right\|_F = \left\|\nabla g\big(\boldsymbol{\Gamma}^{(k+1)}\big) - \nabla g\big(\boldsymbol{\Gamma}^{(k)}\big)\right\|_F \le \sigma_1\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)\cdot\left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_F. \tag{C.8}$$

Function g is Lipschitz smooth with constant $L_{\nabla g} := \sigma_1\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)$. Then, we have

$$g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}\left(\nabla g\big(\boldsymbol{\Gamma}^{(k+1)}\big)^\top\big(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\big)\right) \ge -\frac{L_{\nabla g}}{2}\left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_F^2. \tag{C.9}$$

Recall $\nabla g\big(\boldsymbol{\Gamma}^{(k+1)}\big) = \boldsymbol{\Lambda}^{(k+1)}$, combining (C.4), (C.7), (C.8), and (C.9) yields

$$\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) - \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}\big) \ge \left(-\frac{L_{\nabla g}}{2} - \frac{1}{\rho}L_{\nabla g}^2 + \frac{\rho}{2}\right)\cdot\left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_F^2.$$

Setting $\rho > 2L_{\nabla g}$ makes $C_1 := -\frac{L_{\nabla g}}{2} - \frac{1}{\rho}L_{\nabla g}^2 + \frac{\rho}{2} > 0$, which implies that $\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big)$ is non-increasing over $k \in \mathbb{R} \cup \{0\}$. Now, we will prove $\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big)$ is bounded below over $k \in \mathbb{N} \cup \{0\}$.

$$\begin{aligned}
\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\big) &= f(\mathbf{\Theta}^{(k)}) + g(\mathbf{\Gamma}^{(k)}) + \mathbf{tr}\big(\mathbf{\Lambda}^{(k)\top}\big(\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\big)\big) + \frac{\rho}{2}\|\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\|_{\mathrm{F}}^2 \\
&= f(\mathbf{\Theta}^{(k)}) + g(\mathbf{\Gamma}^{(k)}) + \mathbf{tr}\big(\nabla g(\mathbf{\Gamma}^{(k)})^\top\big(\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\big)\big) + \frac{\rho}{2}\|\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\|_{\mathrm{F}}^2 \\
&\geq g(\mathbf{\Gamma}^{(k)}) + \mathbf{tr}\big(\nabla g(\mathbf{\Gamma}^{(k)})^\top\big(\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\big)\big) + \frac{\rho}{2}\|\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\|_{\mathrm{F}}^2 \\
&\geq g(\mathbf{\Theta}^{(k)}) - \frac{L_{\nabla g}}{2}\|\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\|_{\mathrm{F}}^2 + \frac{\rho}{2}\|\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\|_{\mathrm{F}}^2 \\
&\geq g(\mathbf{\Theta}^{(k)}) := \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Theta}^{(k)}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

In the first inequality, $f(\mathbf{\Theta}^{(k)}) \geq 0$ is used. In the second inequality, Lipschitz smoothness of g with constant $L_{\nabla g}$ is used, and in the last inequality, the choice on $\rho > 2L_{\nabla g}$ is used. It is obvious that $g(\mathbf{\Theta}^{(k)})$ is bounded below from 0.

As long as $\{(\mathbf{\Theta}^{(0)}, \mathbf{\Gamma}^{(0)}, \mathbf{\Lambda}^{(0)})\}$ is bounded, it is easy to see the generated sequence $\{\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\}_{k\geq 1}$ is bounded as well. Since the minimizers of **Step 1.** and **Step 2.** of WMVR-ADMM Algorithm have explicit closed form solution, the pair $\{\mathbf{\Theta}^{(1)}, \mathbf{\Gamma}^{(1)}\}$ is bounded, and by **Step 3.** of WMVR-ADMM algorithm, the boundedness of $\mathbf{\Lambda}^{(1)}$ is automatically ensured. Applying the same logic over the $k \geq 2$ yields the claim.

Lemma C.2.2 *For $k \geq 1$, there exist a constant $C_2 > 0$ and $p^{(k+1)} \in \partial\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}, \mathbf{\Lambda}^{(k+1)}\big)$ such that $\|p^{(k+1)}\|_F \leq C_2\|\mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)}\|_F$.*

Proof of Lemma B.2: Let us define the partial derivative of $\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}, \mathbf{\Lambda}^{(k+1)}\big)$ as

$$\partial\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}, \mathbf{\Lambda}^{(k+1)}\big) := \big(\partial_\Theta\mathcal{L}_\rho, \nabla_\Gamma\mathcal{L}_\rho, \nabla_\Lambda\mathcal{L}_\rho\big)\big(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}, \mathbf{\Lambda}^{(k+1)}\big).$$

for technical convenience. It is easy to see followings:

$$\begin{aligned}
\nabla_{\Gamma}\mathcal{L}_{\rho}\big(\boldsymbol{\Theta}^{(k+1)},\boldsymbol{\Gamma}^{(k+1)},\boldsymbol{\Lambda}^{(k+1)}\big) &= \boldsymbol{\Lambda}^{(k+1)}-\boldsymbol{\Lambda}^{(k)}\\
\nabla_{\Lambda}\mathcal{L}_{\rho}\big(\boldsymbol{\Theta}^{(k+1)},\boldsymbol{\Gamma}^{(k+1)},\boldsymbol{\Lambda}^{(k+1)}\big) &= \frac{1}{\rho}\big(\boldsymbol{\Lambda}^{(k+1)}-\boldsymbol{\Lambda}^{(k)}\big).
\end{aligned}$$

Since $\boldsymbol{\Theta}^{(k+1)}$ is a minimizer of **Step 1.**, it satisfies the following stationary condition.

$$\mathbf{0} \in \partial f(\boldsymbol{\Theta}^{(k+1)}) + \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}). \tag{C.10}$$

Then, we are interested in getting a subdifferential of $\mathcal{L}_{\rho}\big(\boldsymbol{\Theta}^{(k+1)},\boldsymbol{\Gamma}^{(k+1)},\boldsymbol{\Lambda}^{(k+1)}\big)$ with respect to $\boldsymbol{\Theta}$, which can be calculated as follows:

$$\begin{aligned}
&\partial_{\boldsymbol{\Theta}}\mathcal{L}_{\rho}\big(\boldsymbol{\Theta}^{(k+1)},\boldsymbol{\Gamma}^{(k+1)},\boldsymbol{\Lambda}^{(k+1)}\big)\\
&= \partial f(\boldsymbol{\Theta}^{(k+1)}) + \boldsymbol{\Lambda}^{(k+1)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)})\\
&= \partial f(\boldsymbol{\Theta}^{(k+1)}) + \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}) + (\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}) + \rho(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}).
\end{aligned}$$

Then, by (C.10), we have

$$(\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}) + \rho(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}) \in \partial_{\boldsymbol{\Theta}}\mathcal{L}_{\rho}\big(\boldsymbol{\Theta}^{(k+1)},\boldsymbol{\Gamma}^{(k+1)},\boldsymbol{\Lambda}^{(k+1)}\big).$$

If we define $p^{(k+1)}$ as

$$p^{(k+1)} := \Big((\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}) + \rho(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}), \boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}, \frac{1}{\rho}\big(\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}\big)\Big), \tag{C.11}$$

then $p^{(k+1)} \in \partial\mathcal{L}_{\rho}\big(\boldsymbol{\Theta}^{(k+1)},\boldsymbol{\Gamma}^{(k+1)},\boldsymbol{\Lambda}^{(k+1)}\big)$. Furthermore, its Frobenious norm can be bounded by combining (C.8) and (C.11) as follows:

$$\left\|p^{(k+1)}\right\|_{\mathrm{F}} \leq \left(\rho + \big(2 + \frac{1}{\rho}\big) L_{\nabla g}\right) \cdot \left\|\mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)}\right\|_{\mathrm{F}}.$$

Setting $C_2 := \rho + \big(2 + \frac{1}{\rho}\big) L_{\nabla g}$ completes the proof of Lemma B.2.

Main Proof of Theorem 3.2.2: By Bolzano-Weierstrass threorem, we know the bounded sequence $\{\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\}_{k\geq 0}$ has a convergent subsequence $\{\mathbf{\Theta}^{(k_s)}, \mathbf{\Gamma}^{(k_s)}, \mathbf{\Lambda}^{(k_s)}\}_{s\geq 1}$, and denote its limit point as $(\mathbf{\Theta}^*, \mathbf{\Gamma}^*, \mathbf{\Lambda}^*)$. From Lemma B.1, we know the augmented lagrangian function $\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\big)$ is non-increasing and bounded from below. This implies the sequence $\{\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\big)\}_{k\geq 0}$ converges. By continuity of $\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\big)$ and results from **Step 1** and **Step 2** of WMVR-ADMM algorithm in the main paper, we have

$$\lim_{k\to\infty} \mathcal{L}_\rho\big(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\big) = \lim_{s\to\infty} \mathcal{L}_\rho\big(\mathbf{\Theta}^{(k_s)}, \mathbf{\Gamma}^{(k_s)}, \mathbf{\Lambda}^{(k_s)}\big) = \mathcal{L}_\rho\big(\mathbf{\Theta}^*, \mathbf{\Gamma}^*, \mathbf{\Lambda}^*\big).$$

By the result of Lemma B.2, there exists $p^{(k+1)} \in \partial\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}, \mathbf{\Lambda}^{(k+1)}\big)$ such that $\left\|p^{(k+1)}\right\|_{\mathrm{F}} \to 0$ as $k \to \infty$. Consequently, we conclude the following:

$$p^{(k+1)} \in \partial\mathcal{L}_\rho\big(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}, \mathbf{\Lambda}^{(k+1)}\big) \to \mathbf{0} \in \partial\mathcal{L}_\rho\big(\mathbf{\Theta}^*, \mathbf{\Gamma}^*, \mathbf{\Lambda}^*\big), \quad \text{as} k \to \infty.$$

It remains to prove $\mathcal{L}_\rho$ is a Kurdyka-Lojasiewicz (KL) function [205] for ensuring the generated sequence $\{(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)})\}_{k\geq 0}$ converges globally to the unique point $\{\mathbf{\Theta}^*, \mathbf{\Gamma}^*, \mathbf{\Lambda}^*\}$. This can be proved by applying Proposition 2 in [123] and Theorem 2.9 in [205]: If a function is semi-algebraic [206], then it is known to be a KL function. Since $\frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Gamma}\right\|_{\mathrm{F}}^2 + \mathbf{tr}\big(\mathbf{\Lambda}^\top\big(\mathbf{\Theta} - \mathbf{\Gamma}\big)\big)$ is a real-polynomial function, it is semi-algebraic. Since the finite sum of semi-algebraic functions are semi-algebraic [205], it remains to prove $\lambda_n\|\mathbf{\Theta}\|_{\omega,*}$ is a semi-algebraic. In proposition 3 of [206], it is proved that each singular value of the matrix $\mathbf{\Theta}$, $\sigma_j(\mathbf{\Theta})$, is a semi-algebraic. Therefore, the weighted singular value $\omega_j\sigma_j(\mathbf{\Theta})$ is also a semi-algebraic function, and finally, summing rule gives that $\lambda_n \|\boldsymbol{\Theta}\|_{\omega,*}$ is a semi-algebraic function, so $\mathcal{L}_\rho$ is a KL function, which completes the proof of Theorem 2.2. □

C.3 Proof of Proposition 3.3.1

The derivation on the closed-form solution of $\widehat{\boldsymbol{\Theta}}$ is exactly same with that of Lemma 1 in [118], under the orthogonal design assumption. So we omit the proof. We only focus on controlling the distance between singular values of $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^\star$. With the equality $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta}^\star + \boldsymbol{E}$ and $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_{d_1 \times d_1}$, we have

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{\Theta}^\star + \frac{\boldsymbol{X}^\top \boldsymbol{E}}{n}. \tag{C.12}$$

By the corollary of Weyl's Theorem and the equality (C.12), inequality

$$\max_{j=1,\ldots,p} \left| \sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big) - \sigma_j\big(\boldsymbol{\Theta}^\star\big) \right| \leq \sigma_1\left(\frac{\boldsymbol{X}^\top \boldsymbol{E}}{n}\right). \tag{C.13}$$

can be obtained. Recall from our problem setting that the rows of $\boldsymbol{E}$ are independent from $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_{d_2 \times d_2})$. Therefore, we know each entry of $\boldsymbol{X}^\top \boldsymbol{E}/\sigma\sqrt{n}$ follows $\mathcal{N}(0,1)$ and is independent with each other. Following Chapter 6 in [7], with probability at least $1 - 2\exp(-(\sqrt{d_1} + \sqrt{d_2})^2/2)$, the right hand side of (C.13) satisfies

$$\sigma_1\left(\frac{\boldsymbol{X}^\top \boldsymbol{E}}{n}\right) \leq 2\sigma\sqrt{\frac{d_1 + d_2}{n}}. \tag{C.14}$$

Because $\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big) = \sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big) - \boldsymbol{\lambda}_n w_j > 0$ for $j = 1, \ldots, \widehat{r}$, with further combing (C.13) and (C.14), we know for $j \in \{1, \ldots, \widehat{r}\}$

$$\begin{aligned}\Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big) - \sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\Big| &= \Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big) - \boldsymbol{\lambda}_n\omega_j - \sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\Big| \\ &\leq \Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big) - \sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\Big| + \boldsymbol{\lambda}_n\omega_j \\ &\leq \sigma_1\bigg(\frac{\boldsymbol{X}^{\top}\boldsymbol{E}}{n}\bigg) + \boldsymbol{\lambda}_n\omega_j \leq \max\big(4\sigma, 2\omega_j\big) \cdot \sqrt{\frac{d_1 + d_2}{n}},\end{aligned}$$

where in the last inequality, we use (C.14) and choose $\boldsymbol{\lambda}_n = \sqrt{\frac{d_1+d_2}{n}}$. For $j \in \{\widehat{r}+1, \ldots, p\}$ such that $\sigma_j\big(\boldsymbol{\Theta}^{\star}\big) > 0$, the following inequalities hold

$$\begin{aligned}\Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big) - \sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\Big| &\leq \Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big) - \sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\Big| + \Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big)\Big| \\ &\leq \sigma_1\bigg(\frac{\boldsymbol{X}^{\top}\boldsymbol{E}}{n}\bigg) + \boldsymbol{\lambda}_n\omega_j \leq \max\big(4\sigma, 2\omega_j\big) \cdot \sqrt{\frac{d_1 + d_2}{n}},\end{aligned}$$

where in the second inequality, we use (C.13) and $|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big)| \leq \boldsymbol{\lambda}_n\omega_j$ for $j \in \{\widehat{r}+1, \ldots, p\}$. For $j \in \{\widehat{r} + 1, \ldots, p\}$ such that $\sigma_j\big(\boldsymbol{\Theta}^{\star}\big) = 0$, we have the following result:

$$\Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}\big) - \sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\Big| \leq \Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big)\Big| \leq \boldsymbol{\lambda}_n\omega_j. \tag{C.15}$$

Using that the three inequalities (C.13), (C.14), and (C.15) should hold at the same time, we can conclude the proof. □

C.4 Proof of Lemma 3.3.2

Since $\widehat{\boldsymbol{\Theta}}$ is a minimizer and $\boldsymbol{\Theta}^{\star}$ is a feasible solution of the optimization problem in equation (2) in the main paper, we have the following basic inequality:

$$\frac{1}{2n}\Big\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}\Big\|_{\mathrm{F}}^{2} + \boldsymbol{\lambda}_n\|\widehat{\boldsymbol{\Theta}}\|_{w,\star} \leq \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}^{\star}\right\|_{\mathrm{F}}^{2} + \boldsymbol{\lambda}_n\|\boldsymbol{\Theta}^{\star}\|_{w,\star}. \tag{C.16}$$

Plugging in $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta}^\star + \boldsymbol{E}$ in the (C.16) yields

$$\frac{1}{2n}\left\|\boldsymbol{X}\big(\boldsymbol{\Theta}^\star - \widehat{\boldsymbol{\Theta}}\big)\right\|_{\mathrm{F}}^2 \le \frac{1}{n}\mathbf{tr}\big((\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star)^\top \boldsymbol{X}^\top \boldsymbol{E}\big) + \boldsymbol{\lambda}_n\big(\|\boldsymbol{\Theta}^\star\|_{w,\star} - \|\widehat{\boldsymbol{\Theta}}\|_{w,\star}\big). \qquad \text{(C.17)}$$

By denoting $\widehat{\boldsymbol{\Delta}} \equiv \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star$ and since left-hand side of (C.17) is ≥ 0, we have

$$0 \le \frac{1}{n}\mathbf{tr}\big(\widehat{\boldsymbol{\Delta}}^\top \boldsymbol{X}^\top \boldsymbol{E}\big) + \boldsymbol{\lambda}_n\big(\|\boldsymbol{\Theta}^\star\|_{w,\star} - \|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w,\star}\big). \qquad \text{(C.18)}$$

First, we will control the upper-bound on the second term of the (C.18). By the definition of the weighted nuclear norm, we can re-write the term as follows:

$$\begin{aligned}
\|\boldsymbol{\Theta}^\star\|_{w,\star} - \|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w,\star} &= \sum_{j=1}^p w_j\sigma_j\big(\boldsymbol{\Theta}^\star\big) - \sum_{j=1}^p w_j\sigma_j\big(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\big) \\
&= \left[w_p\sum_{j=1}^p \sigma_j\big(\boldsymbol{\Theta}^\star\big) - \sum_{j=1}^p (w_p - w_j)\sigma_j\big(\boldsymbol{\Theta}^\star\big)\right] \\
&\quad - \left[w_p\sum_{j=1}^p \sigma_j\big(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\big) - \sum_{j=1}^p (w_p - w_j)\sigma_j\big(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\big)\right] \\
&= w_p\left[\sum_{j=1}^p \sigma_j\big(\boldsymbol{\Theta}^\star\big) - \sum_{j=1}^p \sigma_j\big(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\big)\right] + \left[\sum_{j=1}^p (w_p - w_j)\big\{\sigma_j\big(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\big) - \sigma_j\big(\boldsymbol{\Theta}^\star\big)\big\}\right] \\
&= \underbrace{w_p\big(\|\boldsymbol{\Theta}^\star\|_\star - \|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_\star\big)}_{:=I} + \underbrace{\big(\|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w_p - w,\star} - \|\boldsymbol{\Theta}^\star\|_{w_p - w,\star}\big)}_{:=II}, \qquad \text{(C.19)}
\end{aligned}$$

where $\|\Theta\|_{w_p - w,\star} = \sum_{j=1}^p (w_p - w_j)\sigma_j(\boldsymbol{\Theta})$.

Recall the definitions of the two subspaces $\mathcal{M}_r$ and $\overline{\mathcal{M}}_r^\perp$ in subsection 3.2. For any $r \le p$, we have

$$\boldsymbol{\Theta}^\star = \Pi_{\mathcal{M}_r}\big(\boldsymbol{\Theta}^\star\big) + \Pi_{\overline{\mathcal{M}}_r^\perp}\big(\boldsymbol{\Theta}^\star\big). \qquad \text{(C.20)}$$

Recall that $\widehat{\boldsymbol{\Delta}}'' \in \Pi_{\overline{\mathcal{M}}_r^\perp}\big(\widehat{\boldsymbol{\Delta}}\big)$ and $\widehat{\boldsymbol{\Delta}}' = \widehat{\boldsymbol{\Delta}} - \widehat{\boldsymbol{\Delta}}''$. Then, we can control the term $\|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_\star$

as follows:

$$\begin{aligned}
\|\widehat{\mathbf{\Delta}}+\mathbf{\Theta}^\star\|_\star &= \|\widehat{\mathbf{\Delta}}'+\widehat{\mathbf{\Delta}}''+\Pi_{\mathcal{M}_r}(\mathbf{\Theta}^\star)+\Pi_{\overline{\mathcal{M}}_r^\perp}(\mathbf{\Theta}^\star)\|_\star \\
&\geq \|\widehat{\mathbf{\Delta}}''+\Pi_{\mathcal{M}_r}(\mathbf{\Theta}^\star)\|_\star - \{\|\widehat{\mathbf{\Delta}}'\|_\star + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\mathbf{\Theta}^\star)\|_\star\} \\
&= \|\widehat{\mathbf{\Delta}}''\|_\star + \|\Pi_{\mathcal{M}_r}(\mathbf{\Theta}^\star)\|_\star - \{\|\widehat{\mathbf{\Delta}}'\|_\star + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\mathbf{\Theta}^\star)\|_\star\}, \qquad \text{(C.21)}
\end{aligned}$$

where in the first inequality, we used the triangle inequality of $\|\cdot\|_\star$ and in the last equality, the decomposability of $\|\cdot\|_\star$ with respect to a pair of subspaces $(\mathcal{M}_r, \overline{\mathcal{M}}_r^\perp)$ is used. With (C.21), we are ready to control the term I in (C.19).

$$\begin{aligned}
& w_p\bigg(\|\mathbf{\Theta}^\star\|_\star - \|\widehat{\mathbf{\Delta}}+\mathbf{\Theta}^\star\|_\star\bigg) \\
&\quad \leq w_p \cdot \Bigg\{ \Big(\|\Pi_{\mathcal{M}_r}(\mathbf{\Theta}^\star)\|_\star + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\mathbf{\Theta}^\star)\|_\star\Big) \\
&\qquad\qquad - \Big(\|\widehat{\mathbf{\Delta}}''\|_\star + \|\Pi_{\mathcal{M}_r}(\mathbf{\Theta}^\star)\|_\star - \{\|\widehat{\mathbf{\Delta}}'\|_\star + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\mathbf{\Theta}^\star)\|_\star\}\Big) \Bigg\} \\
&\quad = w_p \cdot \Big\{2\|\Pi_{\overline{\mathcal{M}}_r^\perp}(\mathbf{\Theta}^\star)\|_\star + \|\widehat{\mathbf{\Delta}}'\|_\star - \|\widehat{\mathbf{\Delta}}''\|_\star\Big\} \qquad \text{(C.22)}
\end{aligned}$$

Note that the equality $\|\Theta^\star\|_\star = \|\Pi_{\mathcal{M}_r}(\mathbf{\Theta}^\star)\|_\star + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\mathbf{\Theta}^\star)\|_\star$ is used in the first inequality due to (C.20).

Now the term II in (C.19) needs to be controlled. First, we need to see the norm $\|\cdot\|_{w_p-w,\star} = \sum_{j=1}^p (w_p - w_j)\sigma_j(\cdot)$ with respect to any pair of matrices: $(A,B) \in (\mathcal{M}_r, \overline{\mathcal{M}}_r^\perp)$ satisfies the decomposability, meaning $\|A+B\|_{w_p-w,\star} = \|A\|_{w_p-w,\star} + \|B\|_{w_p-w,\star}$. By definition of the subspace pair $(\mathcal{M}_r, \overline{\mathcal{M}}_r^\perp)$, we can write A and B as

$$A = \boldsymbol{U}\begin{bmatrix} \boldsymbol{T}_{1,1} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix}\boldsymbol{V}^\top, \qquad B = \boldsymbol{U}\begin{bmatrix} \mathbf{0}_{r\times r} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \boldsymbol{T}_{2,2} \end{bmatrix}\boldsymbol{V}^\top,$$

where $\boldsymbol{T}_{1,1} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{T}_{2,2} \in \mathbb{R}^{(p-r)\times(p-r)}$ are arbitrary matrices. Define two diagonal matrices $\boldsymbol{W}_1 := \mathrm{diag}(w_p - w_1, \ldots, w_p - w_r)$ and $\boldsymbol{W}_2 := \mathrm{diag}(w_p - w_{r+1}, \ldots, w_p - w_p)$. Then, we have

$$\begin{aligned}
\|A + B\|_{w_p - w,\star} &= \left\| \begin{bmatrix} \boldsymbol{W}_1 \boldsymbol{T}_{1,1} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{r\times r} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \boldsymbol{W}_2 \boldsymbol{T}_{2,2} \end{bmatrix} \right\|_\star \\
&= \left\| \begin{bmatrix} \boldsymbol{W}_1 \boldsymbol{T}_{1,1} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix} \right\|_\star + \left\| \begin{bmatrix} \mathbf{0}_{r\times r} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \boldsymbol{W}_2 \boldsymbol{T}_{2,2} \end{bmatrix} \right\|_\star \\
&= \|A\|_{w_p - w,\star} + \|B\|_{w_p - w,\star}.
\end{aligned}$$

In the first equality, the definition of $\|\cdot\|_{w_p - w,\star}$ and the invariance of the nuclear norm to orthogonal transformation to multiplication by the matrices $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ are used.

Using this fact, similarly with (C.21) and (C.22), we get the upper-bound on II in the equality (C.19):

$$\begin{aligned}
&\|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w_p - w,\star} - \|\boldsymbol{\Theta}^\star\|_{w_p - w,\star} \\
&\Leftrightarrow \|\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^\star) + \Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star) + \widehat{\boldsymbol{\Delta}}' + \widehat{\boldsymbol{\Delta}}''\|_{w_p - w,\star} - \|\boldsymbol{\Theta}^\star\|_{w_p - w,\star} \\
&\leq \|\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^\star) + \widehat{\boldsymbol{\Delta}}''\|_{w_p - w,\star} + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_{w_p - w,\star} + \|\widehat{\boldsymbol{\Delta}}'\|_{w_p - w,\star} - \|\boldsymbol{\Theta}^\star\|_{w_p - w,\star} \\
&= \Big\{ \|\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^\star)\|_{w_p - w,\star} + \|\widehat{\boldsymbol{\Delta}}''\|_{w_p - w,\star} + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_{w_p - w,\star} + \\
&\qquad \|\widehat{\boldsymbol{\Delta}}'\|_{w_p - w,\star} \Big\} - \Big\{ \|\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^\star)\|_{w_p - w,\star} + \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_{w_p - w,\star} \Big\} \\
&= \|\widehat{\boldsymbol{\Delta}}''\|_{w_p - w,\star} + \|\widehat{\boldsymbol{\Delta}}'\|_{w_p - w,\star}. \qquad \text{(C.23)}
\end{aligned}$$

By combining the inequalities (C.22) and (C.23), we can obtain an upper-bound on the

Eq. (C.19);

$$\begin{aligned}
&\|\boldsymbol{\Theta}^\star\|_{w,\star} - \|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w,\star} \\
&\quad = w_p(\|\boldsymbol{\Theta}^\star\|_\star - \|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_\star) + (\|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w_p-w,\star} - \|\boldsymbol{\Theta}^\star\|_{w_p-w,\star}) \\
&\quad \le w_p\left\{2\|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_\star + \|\widehat{\boldsymbol{\Delta}}'\|_\star - \|\widehat{\boldsymbol{\Delta}}''\|_\star\right\} + \left\{\|\widehat{\boldsymbol{\Delta}}''\|_{w_p-w,\star} + \|\widehat{\boldsymbol{\Delta}}'\|_{w_p-w,\star}\right\}.
\end{aligned}$$

Now, we control the first term of right-hand side in (C.18) as follows:

$$\left|\frac{1}{n}\mathbf{tr}\big(\widehat{\boldsymbol{\Delta}}^\top \boldsymbol{X}^\top \boldsymbol{E}\big)\right| \le \left\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{E}\right\|_{\text{op}} \|\widehat{\boldsymbol{\Delta}}\|_\star \le \frac{\boldsymbol{\lambda}_n}{2}\|\widehat{\boldsymbol{\Delta}}\|_\star. \tag{C.24}$$

In the first inequality, we used Hölder's inequality and in the second inequality the condition $\boldsymbol{\lambda}_n \ge \frac{2}{n}\left\|\boldsymbol{X}^\top \boldsymbol{E}\right\|_{\text{op}}$ is used. Combining everything, we finally have a bound on Eq. (C.18):

$$\begin{aligned}
0 &\le \frac{1}{n}\mathbf{tr}\big(\widehat{\boldsymbol{\Delta}}^\top \boldsymbol{X}^\top \boldsymbol{E}\big) + \boldsymbol{\lambda}_n\left\{\|\boldsymbol{\Theta}^\star\|_{w,\star} - \|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w,\star}\right\} \\
&\le \boldsymbol{\lambda}_n\left\{\frac{1}{2}\|\widehat{\boldsymbol{\Delta}}\|_\star + w_p\left\{2\|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_\star + \|\widehat{\boldsymbol{\Delta}}'\|_\star - \|\widehat{\boldsymbol{\Delta}}''\|_\star\right\} + \left\{\|\widehat{\boldsymbol{\Delta}}''\|_{w_p-w,\star} + \|\widehat{\boldsymbol{\Delta}}'\|_{w_p-w,\star}\right\}\right\} \\
&\le \boldsymbol{\lambda}_n\left\{\frac{1}{2}\|\widehat{\boldsymbol{\Delta}}'\|_\star + \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}''\|_\star + w_p\left\{2\|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_\star + \|\widehat{\boldsymbol{\Delta}}'\|_\star - \|\widehat{\boldsymbol{\Delta}}''\|_\star\right\}\right. \qquad \text{(C.25)} \\
&\qquad \left. + \left\{\|\widehat{\boldsymbol{\Delta}}''\|_{w_p-w,\star} + \|\widehat{\boldsymbol{\Delta}}'\|_{w_p-w,\star}\right\}\right\} \\
&= \boldsymbol{\lambda}_n\left\{2w_p\|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_\star + \|\widehat{\boldsymbol{\Delta}}'\|_{2w_p-w+\frac{1}{2},\star} - \|\widehat{\boldsymbol{\Delta}}''\|_{w-\frac{1}{2},\star}\right\}. \qquad \text{(C.26)}
\end{aligned}$$

Note the norm denotes $\|\cdot\|_{2w_p-w+\frac{1}{2},\star} := \sum_{j=1}^p (2w_p - w_j + \frac{1}{2})\sigma_j\big(\cdot\big)$. The inequality (C.26) implies

$$\sum_{j=1}^p \left(w_j - \frac{1}{2}\right)\sigma_j\big(\widehat{\boldsymbol{\Delta}}''\big) \le 2w_p \cdot \|\Pi_{\overline{\mathcal{M}}_r^\perp}(\boldsymbol{\Theta}^\star)\|_\star + \sum_{j=1}^{2r}\left(2w_p - w_j + \frac{1}{2}\right)\sigma_j\big(\widehat{\boldsymbol{\Delta}}'\big). \tag{C.27}$$

In (C.27), we use the fact rank$(\widehat{\boldsymbol{\Delta}}') \le 2r$. See the proof of Lemma 1 in [5]. Because $(w_1 - \frac{1}{2})\sum_{j=1}^p \sigma_j\big(\widehat{\boldsymbol{\Delta}}''\big) \le \sum_{j=1}^p \big(w_j - \frac{1}{2}\big)\sigma_j\big(\widehat{\boldsymbol{\Delta}}''\big)$, and similarly, $\sum_{j=1}^{2r}\big(2w_p - w_j +$

$\frac{1}{2})\sigma_j(\widehat{\boldsymbol{\Delta}}') \leq (2w_p - w_1 + \frac{1}{2})\sum_{j=1}^{2r}\sigma_j(\widehat{\boldsymbol{\Delta}}')$, the inequality (C.27) implies

$$\|\widehat{\boldsymbol{\Delta}}''\|_\star \leq \frac{2w_p}{w_1 - \frac{1}{2}}\sum_{j=r+1}^{p}\sigma_j(\boldsymbol{\Theta}^\star) + \frac{2w_p - w_1 + \frac{1}{2}}{w_1 - \frac{1}{2}} \cdot \|\widehat{\boldsymbol{\Delta}}'\|_\star. \tag{C.28}$$

□

C.5 Proof of Theorem 3.3.3

First, recall the basic inequality (C.17), transformation of weighted nuclear norm (C.19) and duality of operator and nuclear norm (C.24). Then, we have

$$\begin{aligned}
\frac{1}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2 &\leq \frac{1}{n}\mathbf{tr}(\widehat{\boldsymbol{\Delta}}^\top\boldsymbol{X}^\top\boldsymbol{E}) + \boldsymbol{\lambda}_n(\|\boldsymbol{\Theta}^\star\|_{w,\star} - \|\widehat{\boldsymbol{\Theta}}\|_{w,\star}) \\
&\leq \boldsymbol{\lambda}_n\Big\{\frac{1}{2}\|\widehat{\boldsymbol{\Delta}}'\|_\star + \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}''\|_\star + w_p(\|\boldsymbol{\Theta}^\star\|_\star - \|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_\star) \\
&\qquad\qquad + (\|\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star\|_{w_p - w,\star} - \|\boldsymbol{\Theta}^\star\|_{w_p-w,\star})\Big\} \\
&\leq \boldsymbol{\lambda}_n\Big\{\frac{1}{2}\|\widehat{\boldsymbol{\Delta}}'\|_\star + \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}''\|_\star + w_p(\|\widehat{\boldsymbol{\Delta}}'\|_\star + \|\widehat{\boldsymbol{\Delta}}''\|_\star) \\
&\qquad\qquad + (\|\widehat{\boldsymbol{\Delta}}'\|_{w_p-w,\star} + \|\widehat{\boldsymbol{\Delta}}''\|_{w_p-w,\star})\Big\} \\
&= \boldsymbol{\lambda}_n\Big\{\sum_{j=1}^{2r}\Big(2w_p - w_j + \frac{1}{2}\Big)\sigma_j(\widehat{\boldsymbol{\Delta}}') + \sum_{j=1}^{p}\Big(2w_p - w_j + \frac{1}{2}\Big)\sigma_j(\widehat{\boldsymbol{\Delta}}'')\Big\} \\
&\leq \boldsymbol{\lambda}_n\Big(2w_p - w_1 + \frac{1}{2}\Big)\Big(\left\|\widehat{\boldsymbol{\Delta}}'\right\|_\star + \left\|\widehat{\boldsymbol{\Delta}}''\right\|_\star\Big),
\end{aligned} \tag{C.29}$$

where in the third inequality, triangle inequality of norms $\|\cdot\|_\star$ and $\|\cdot\|_{w_p-w,\star}$ is applied twice. In the last inequality, we used $\sum_{j=1}^{2r}(2w_p - w_j + \frac{1}{2})\sigma_j(\widehat{\boldsymbol{\Delta}}') \leq (2w_p - w_1 + \frac{1}{2})\sum_{j=1}^{2r}\sigma_j(\widehat{\boldsymbol{\Delta}}')$ and $\sum_{j=1}^{p}(2w_p - w_j + \frac{1}{2})\sigma_j(\widehat{\boldsymbol{\Delta}}'') \leq (2w_p - w_1 + \frac{1}{2})\sum_{j=1}^{p}\sigma_j(\widehat{\boldsymbol{\Delta}}'')$.

By the RSC condition, there exists a constant $\kappa > 0$ such that $\kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 \leq \frac{1}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2$.

Then, by (C.27) and (C.29), with some straightforward calculations, we have

$$\begin{aligned}
\kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 &\leq \boldsymbol{\lambda}_n \frac{2w_p\big(2w_p - w_1 + \frac{1}{2}\big)}{w_1 - \frac{1}{2}} \cdot \bigg(\Big\|\widehat{\boldsymbol{\Delta}}'\Big\|_\star + \sum_{j=r+1}^{p} \sigma_j\big(\boldsymbol{\Theta}^\star\big)\bigg) \\
&\leq \boldsymbol{\lambda}_n \frac{2w_p\big(2w_p - w_1 + \frac{1}{2}\big)}{w_1 - \frac{1}{2}} \cdot \bigg(2\sqrt{r}\Big\|\widehat{\boldsymbol{\Delta}}\Big\|_{\mathrm{F}} + \sum_{j=r+1}^{p} \sigma_j\big(\boldsymbol{\Theta}^\star\big)\bigg) \\
&\leq \boldsymbol{\lambda}_n \frac{w_p\big(2w_p - w_1 + \frac{1}{2}\big)}{w_1 - \frac{1}{2}} \cdot \max\bigg\{ 8\sqrt{r}\Big\|\widehat{\boldsymbol{\Delta}}\Big\|_{\mathrm{F}}, 4\sum_{j=r+1}^{p} \sigma_j\big(\boldsymbol{\Theta}^\star\big)\bigg\},
\end{aligned}$$

where in the second inequality, we used the fact $\|\widehat{\boldsymbol{\Delta}}'\|_\star \leq \sqrt{2r}\|\widehat{\boldsymbol{\Delta}}'\|_{\mathrm{F}} \leq 2\sqrt{r}\|\widehat{\boldsymbol{\Delta}}\|_F$, and in the last inequality, the inequality $a + b \leq \max\{2a, 2b\}$ for $a, b \geq 0$ is used. Let us denote $\mathcal{W} := \frac{w_p\left(2w_p - w_1 + \frac{1}{2}\right)}{w_1 - \frac{1}{2}}$. Then, we obtain the final bound:

$$\Big\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\Big\|_{\mathrm{F}} \leq \max\bigg\{ 8\mathcal{W} \cdot \frac{\boldsymbol{\lambda}_n\sqrt{r}}{\kappa}, \bigg[4\mathcal{W} \cdot \frac{\boldsymbol{\lambda}_n \sum_{j=r+1}^{p} \sigma_j\big(\boldsymbol{\Theta}^\star\big)}{\kappa}\bigg]^{1/2}\bigg\}. \tag{C.30}$$

Let us construct a set of indices whose corresponding eigenvalues are greater than a threshold $\tau > 0$, and denote it as $\mathcal{K}$ and its complement as $\mathcal{K}^c$.

$$\mathcal{K} := \Big\{ j \in \{1, \ldots, p\} : \sigma_j\big(\boldsymbol{\Theta}^\star\big) > \tau \Big\}, \qquad \mathcal{K}^c := \Big\{ j \in \{1, \ldots, p\} : \sigma_j\big(\boldsymbol{\Theta}^\star\big) \leq \tau \Big\}.$$

Since it is assumed that $\boldsymbol{\Theta}^\star \in \mathbb{B}_q(r^\star)$, for $q \in [0, 1]$, we have the following inequality:

$$r^\star \geq \sum_{j=1}^{p} \big|\sigma_j\big(\boldsymbol{\Theta}^\star\big)\big|^q \geq |\mathcal{K}| \cdot \tau^q, \tag{C.31}$$

where $|\mathcal{K}|$ denotes a cardinality of the set $\mathcal{K}$. Similarly, by using the definition of set $\mathcal{K}^c$, for $q \in [0, 1]$, we have the following inequality:

$$\sum_{j \in \mathcal{K}^c} \sigma_j\big(\boldsymbol{\Theta}^\star\big) = \tau \sum_{j=|\mathcal{K}|+1}^{p} \frac{\sigma_j\big(\boldsymbol{\Theta}^\star\big)}{\tau} \leq \tau \sum_{j=|\mathcal{K}|+1}^{p} \bigg(\frac{\sigma_j\big(\boldsymbol{\Theta}^\star\big)}{\tau}\bigg)^q \leq r^\star \cdot \big(\tau^{1-q}\big). \tag{C.32}$$

Set $r = |\mathcal{K}|$ and plugging (C.31) and (C.32) in (C.30) yields:

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^{\star}\right\|_{\mathrm{F}} \leq \max\left\{8\mathcal{W} \cdot \frac{\boldsymbol{\lambda}_n\sqrt{r^{\star}} \cdot \left(\tau^{-q/2}\right)}{\kappa}, \left[4\mathcal{W} \cdot \frac{\boldsymbol{\lambda}_n r^{\star} \cdot \left(\tau^{1-q}\right)}{\kappa}\right]^{1/2}\right\}. \tag{C.33}$$

Setting $\tau = \boldsymbol{\lambda}_n/\kappa$ yields that

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^{\star}\right\|_{\mathrm{F}} \leq 8\mathcal{W} \cdot \sqrt{r^{\star}}\left(\frac{\boldsymbol{\lambda}_n}{\kappa}\right)^{1-q/2}. \tag{C.34}$$

Recall that we choose $\boldsymbol{\lambda}_n$ such that $\boldsymbol{\lambda}_n \geq \frac{2}{n}\left\|\boldsymbol{X}^{\top}\boldsymbol{E}\right\|_{\mathrm{op}}$. [5] proved that $\frac{2}{n}\left\|\boldsymbol{X}^{\top}\boldsymbol{E}\right\|_{\mathrm{op}} \leq 10\sigma\|\mathbf{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{d_1+d_2}{n}}$ holds with high probability in Lemma 3 of their paper. We formally re-state the Lemma in the following.

Lemma C.5.1 *[Negahban and Wainwright [2011]] There are universal constants* $c_1, c_2 > 0$ *such that*

$$\mathbb{P}\left\{\left|\frac{1}{n}\left\|\boldsymbol{X}^{\top}\boldsymbol{E}\right\|_{op}\right| \geq 5\sigma\|\mathbf{\Sigma}\|_{op}\sqrt{\frac{d_1+d_2}{n}}\right\} \leq c_1\exp\Big(-c_2\big(d_1+d_2\big)\Big).$$

Then, it remains us to determine the constant term κ, which satisfies the RSC property. Readers can refer Lemma 2 in [5] for the following result.

Lemma C.5.2 *[Negahban and Wainwright [2011]] Let* $\boldsymbol{X} \in \mathbb{R}^{n\times d_1}$ *be a random matrix with i.i.d. rows sampled from a* d_1*- variate* $\mathcal{N}(0,\Sigma)$ *distribution. Then for* $n \geq 2d_1$*, we have*

$$\mathbb{P}\left\{\sigma_{min}\left(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right) \geq \frac{\sigma_{min}(\mathbf{\Sigma})}{9}\right\} \geq 1 - 4\exp\Big(-\frac{n}{2}\Big).$$

With the result from Lemma E.2, some algebra shows that we can easily establish the lower bound on the quantity $\frac{1}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2$ as follows:

$$\frac{1}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2 = \frac{1}{2n}\sum_{j=1}^{d_2}\left\|\big(\boldsymbol{X}\widehat{\boldsymbol{\Delta}}\big)_j\right\|_2^2 \geq \frac{1}{2n}\sigma_{\min}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2 \geq \frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathrm{F}}^2.$$

This shows that the RSC property holds with probability at least $1-4\exp(-n/2)$ with the constant $\kappa = \frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}$. Plugging $\boldsymbol{\lambda}_n = 10\sigma\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{d_1+d_2}{n}}$ and $\kappa = \frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}$ in (C.34) yields the following inequality:

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^\star\right\|_{\mathrm{F}}^2 &\leq 64\mathcal{W}^2\cdot r^\star\left(10\sigma\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{d_1+d_2}{n}}\frac{18}{\sigma_{\min}(\boldsymbol{\Sigma})}\right)^{2-q}\\ &= c_1\mathcal{W}^2\left(\frac{\sigma^2\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^2}{\sigma_{\min}^2(\boldsymbol{\Sigma})}\right)^{1-q/2} r^\star\left(\frac{d_1+d_2}{n}\right)^{1-q/2}.\end{aligned}$$

□

C.6 Proof of Proposition 3.4.1

Let $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ be the SVD of $\boldsymbol{\Theta}$. Then, we have

$$\mathbf{tr}\big(\boldsymbol{\Theta}^\top\boldsymbol{K}\boldsymbol{\Theta}\big) = \mathbf{tr}\big(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{K}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\big) = \mathbf{tr}\big(\boldsymbol{D}^2\boldsymbol{U}^\top\boldsymbol{K}\boldsymbol{U}\big) = \mathbf{tr}\big(\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top\boldsymbol{K}\big).$$

Let $\boldsymbol{A} := \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top$ and $\boldsymbol{B} := \boldsymbol{K}$. As proved in [207], for two positive-semi definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\mathbf{tr}\big(\boldsymbol{A}^\top\boldsymbol{B}\big) \geq \sum_{j=1}^{p}\sigma_j\big(\boldsymbol{A}\big)\sigma_{p+1-j}\big(\boldsymbol{B}\big), \tag{C.35}$$

where $\sigma_1\big(\,\cdot\,\big) \geq \sigma_2\big(\,\cdot\,\big) \geq \cdots \geq \sigma_p\big(\,\cdot\,\big) \geq 0$. Denote $d_j := \sigma_j\big(\boldsymbol{\Theta}\big)$ for $j \in \{1, \dots, p\}$. Given $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_p$, it is easy to see that

$$\sum_{j=1}^{p} \sigma_j\big(\boldsymbol{A}\big)\sigma_{p+1-j}\big(\boldsymbol{B}\big) = \sum_{j=1}^{\widehat{r}} \frac{\omega_j}{\widehat{d}_j} d_j^2. \tag{C.36}$$

Recalling the assumption $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_{d_1 \times d_1}$ and a simple fact $\big(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big)^\top \boldsymbol{X} = 0$, we can rewrite the cost function in equation (14) of the main paper as follows:

$$\frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\right\|_{\text{F}}^2 = \frac{1}{2n}\mathbf{tr}\big(\big(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big)^\top\big(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big)\big) + \frac{1}{2}\mathbf{tr}\big(\big(\widehat{\boldsymbol{\Theta}}^{\text{LS}} - \boldsymbol{\Theta}\big)^\top\big(\widehat{\boldsymbol{\Theta}}^{\text{LS}} - \boldsymbol{\Theta}\big)\big). \tag{C.37}$$

By combining (C.36) and (C.37), we can obtain the lower bound of the objective function (14) in the main paper as follows:

$$\frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\Theta\right\|_{\text{F}}^2 + \frac{\boldsymbol{\lambda}_n}{2}\mathbf{tr}\big(\boldsymbol{\Theta}^\top \boldsymbol{K}\boldsymbol{\Theta}\big) \geq \frac{1}{2}\sum_{j=1}^{p} d_j^2 - \sum_{j=1}^{p}\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big)d_j + \frac{\boldsymbol{\lambda}_n}{2}\sum_{j=1}^{\widehat{r}}\frac{\omega_j}{\widehat{d}_j}d_j^2.$$

We use the equality $\mathbf{tr}\big(\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\big) = \sum_{j=1}^{p} d_j^2$ and the inequality (C.35) to get a lower bound. It also should be noted that the equality in the above lower bound holds when $\boldsymbol{U} = \widehat{\boldsymbol{U}}^{\text{LS}}$ and $\boldsymbol{V} = \widehat{\boldsymbol{V}}^{\text{LS}}$. Solving the quadratic equation yields the followings:

$$\widehat{\boldsymbol{D}}_{jj}^{\text{SR}} = \begin{cases} \frac{\widehat{d}_j\sigma_j\left(\widehat{\boldsymbol{\Theta}}^{\text{LS}}\right)}{\widehat{d}_j + \boldsymbol{\lambda}_n\omega_j} & j = 1, \dots, \widehat{r}, \\ \sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big) & j = \widehat{r}+1, \dots, p. \end{cases} \tag{C.38}$$

Recall that $\widehat{d}_j = \sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\text{LS}}\big) - \boldsymbol{\lambda}_n\omega_j$ for $j = 1, \dots, \widehat{r}$, and plugging this equality in (C.38) for $j \in \{1, \dots, \widehat{r}\}$ yields the claim. □

C.7 Extension of WMVR-ADMM to Trace Regression Model

let us consider the following trace regression problem : $y_i = \mathbf{tr}(\boldsymbol{X}_i^T\boldsymbol{\Theta}^\star)+\varepsilon_i$, $i = 1, \cdots, n$, where $\boldsymbol{X}_i \in \boldsymbol{R}^{d_1\times d_2}$ is a known measurement matrix for $i = 1, \cdots, n$ and $\big\{\varepsilon_i\big\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{N}\big(0, \sigma^2\big)$. In this section, we present an extension of WMVR-ADMM algorithm for solving the following optimization problem.

$$\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{2n}\sum_{i=1}^n (y_i - \mathbf{tr}(\boldsymbol{X}_i^T\boldsymbol{\Theta}))^2 + \boldsymbol{\lambda}_n \|\boldsymbol{\Theta}\|_{\boldsymbol{\omega},\star} \right\}.$$

First, let $v(M) \in \mathbb{R}^{d_1d_2}$ be the vectorized version of matrix M concatenating columns of $M \in \mathbb{R}^{d_1\times d_2}$ into one column vector, and let us also define an inverse operator $\mathbf{Mat}\big[v(M)\big] := M$. With this notation, the algorithm is summarized in the Algorithm 2. The presented algorithm can be derived easily by using exactly the same techniques employed in Section 2.1 of the main paper by plugging $f(\boldsymbol{\Theta}) := -\frac{1}{n}\mathbf{tr}\Big(\big(\sum_{j=1}^n y_i\boldsymbol{X}_i\big)^\top\boldsymbol{\Theta}\Big) + \boldsymbol{\lambda}_n\|\boldsymbol{\Theta}\|_{\boldsymbol{\omega},\star}$ and $g(\boldsymbol{\Gamma}) = \frac{1}{2n}\sum_{j=1}^n \mathbf{tr}\big(\boldsymbol{X}_{\boldsymbol{i}}^\top\boldsymbol{\Gamma}\big)^2$ in equation (4) of the main paper.

```
Input : $\big\{\boldsymbol{X}_i, y_i\big\}_{i=1}^n$, $\boldsymbol{\lambda}_n \geq 0$.
Prelimiaries : $\boldsymbol{M_y x} := \sum_{i=1}^n y_i\boldsymbol{X}_i$, and
$\mathcal{A} := \frac{1}{n}\sum_{i=1}^n v(\boldsymbol{X}_i)v(\boldsymbol{X}_i)^\top + \rho\cdot\mathcal{I}_{d_1d_2\times d_1d_2}$.
Initialization : $\boldsymbol{\Theta}^{(0)} = \mathbf{0}$, $\boldsymbol{\Gamma}^{(0)} = \mathbf{0}$, $\boldsymbol{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{d_1\times d_2}$.
    Repeat following Steps :
        Step 1. Let $\boldsymbol{B^{(k)}} := \frac{1}{n}\boldsymbol{M_y x} - \boldsymbol{\Lambda}^{(k)} + \rho\cdot\boldsymbol{\Gamma}^{(k)}$.
$\boldsymbol{B^{(k)}} = \boldsymbol{U}^{\mathbf{B}}\boldsymbol{D}^{\mathbf{B}}\big(\boldsymbol{V}^{\mathbf{B}}\big)^\top$. (SVD)
            Set $\mathcal{S}_{\boldsymbol{\lambda}_n\omega}\big(\boldsymbol{D}^{\mathbf{B}}\big) = \mathbf{diag}\Bigg\{ \max\bigg\{\frac{1}{\rho}\big(\sigma_j(\boldsymbol{B}^{(k)}) - \boldsymbol{\lambda}_n w_j\big), 0\bigg\}$ for
$j = 1, \ldots, p\Bigg\}$.
            $\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{U}^{\mathbf{B}}\mathcal{S}_{\boldsymbol{\lambda}_n\omega}\big(\boldsymbol{D}^{\mathbf{B}}\big)\big(\boldsymbol{V}^{\mathbf{B}}\big)^\top$.
        Step 2. $\boldsymbol{\Gamma}^{(k+1)} = \mathbf{Mat}\big[\mathcal{A}^{-1}\big(\rho v(\boldsymbol{\Theta}^{(k+1)}) - v(\boldsymbol{\Lambda}^{(k)})\big)\big]$.
        Step 3. $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big)$.
    Until $\|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\|_F \leq 10^{-7}$ and $\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\|_F \leq 10^{-7}$.
Output : $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(k+1)}$.
```

Algorithm 2: ADMM for weighted Trace Regression. (WTR-ADMM)

APPENDIX D

A NON-PARAMETRIC REGRESSION VIEWPOINT : GENERALIZATION OF OVERPARAMETRIZED DEEP RELU NETWORK UNDER NOISY OBSERVATIONS

D.1 Numerical illustrations

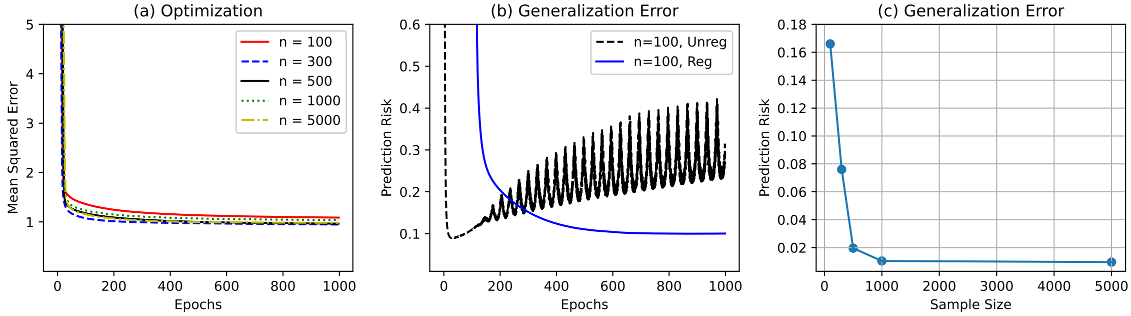


Figure D.1: Results on synthetic data.

In this section, we use synthetic data to corroborate our theoretical findings. We use the He initialization [151] and employ (ℓ_2-regularized) GD as introduced in subsection 4.2.2. For the experiments, we run 1000 epochs of GD and use a fixed step size, setting $\eta_1 = \eta_2 = 0.001$. We uniformly generate n feature data $\mathbf{x_i}^{\text{train}}$ from $\mathcal{S}^{d-1}$ with $d = 2$ and generate $\mathbf{y}_i$ from $f_\rho^\star(\mathbf{x}_i^{\text{train}})$ with $\varepsilon_i \sim \mathcal{N}(0,1)$. To create a function $f_\rho^\star \in \mathcal{H}_L^{\mathbf{NTK}}$, we use the definition in (4.10) with $\alpha \in \textbf{Unif}(\mathcal{S}^{p-1})$ and with p fixed points $\{\tilde{\mathbf{x}}_j\}_{j=1}^{p} \subset \textbf{Unif}(\mathcal{S}^{d-1})$, where p is simply set as 1. Note that $\textbf{Ker}(\cdot,\cdot)$ in (4.10) can be calculated via the formulas (4.8) and (4.9) with specified network depth L. We consider a scenario where we have a network with depth $L = 8$ and width $m = 2000$. The variance parameter of the output layer (ω) is set as 1 for unregularized and 0.001 for regularized cases.

In Figure D.1.(a), we record the training errors of regularized networks over the GD epochs

$k \leq 1000$, where we have $n \in \{100, 300, 500, 1000, 5000\}$ training samples. This aims to verify the inequality (4.13) that the MSE of regularized network is bounded away from 0 by some constant. In Figure D.1.(b), the prediction risks of both unregularized and regularized networks are displayed. We approximate the risk with $\frac{1}{500}\sum_{j=1}^{500}\big(\widehat{f}_k(\mathbf{x}_j^{\text{test}}) - f_\rho^\star(\mathbf{x}_j^{\text{test}})\big)^2$ with a new test data set $\{\mathbf{x}_j^{\text{test}}, f_\rho^\star(\mathbf{x}_j^{\text{test}})\}_{j=1}^{500}$ over $k \leq 1000$ for both unregularized and regularized cases. In both cases, they reach the same minimal risks, but the risk of unregularized network increase after it hits the minimal point, whereas the risk of regularized network stays stable. Theorem 4.3.8 tells us that for the iteration less than the order $\mathcal{O}\big(\frac{1}{\eta m \omega L}\big)$, the prediction error is bounded away from 0. In the experiment for unregularized case, we set $\eta = 0.01$, $m = 2000$, $L = 8$, and $\omega = 1$. Plugging in these parameters in the bound says that the minimum can be achieved within a very few iterations. Note that the optimal risk is non-zero as long as we have finite sample sizes n, but converges to 0 at the rate $\mathcal{O}\big(n^{-\frac{d}{2d-1}}\big)$. In Figure D.1.(c), we verify that the more training sample sizes we have, the closer the risks of the regularized networks get to 0. The risk is evaluated at the sample sizes $n = \{100, 300, 500, 1000, 5000\}$.

We have to acknowledge that there is a discrepancy between our experiment setting and theory. Specifically, due to the limited computing power, we could not run the experiment under the regime of width $\frac{m}{\log^3(m)} \geq \Omega\big(\frac{\omega^7 n^8 L^{18}}{\lambda_\infty^8 \delta^2}\big)$. But the prediction risk behaves similarly as expected by our theorems, which can be a partial evidence that the statement in theorems still holds in the narrower width of the network.

D.2 Preliminary Notations

Before presenting the formal proofs of Lemmas and main results, we introduce several notations used frequently throughout the proofs. First, we denote $\mathbf{x}_{\ell,i}$ the output of the ℓth hidden layer with the input data $\mathbf{x}_i$ after applying entry-wise ReLU activation function.

$$\mathbf{x}_{\ell,i} = \sigma\big(\mathbf{W}_\ell \sigma\big(\mathbf{W}_{\ell-1} \cdots \sigma\big(\mathbf{W}_1 \mathbf{x}_i\big) \cdots\big)\big).$$

Denote $f_{\mathbf{W}(k)}(x)$ a value of neural network (4.2) evaluated at the collection of network parameters $\mathcal{W}^{(k)} := \big\{\mathbf{W}_\ell^{(k)}\big\}_{\ell=1,\ldots,L}$ and $\mathbf{W}_\ell^{(k)}$ denotes the ℓth hidden layer parameter updated by kth GD iterations.

Partial gradient of $f_{\mathbf{W}(k)}(x)$**.** We employ the following matrix product notation which was used in several other papers [137, 140]:

$$\prod_{r=\ell_1}^{\ell_2} A_r := \begin{cases} A_{\ell_2} A_{\ell_2 - 1} \cdots A_{\ell_1} \text{ if } \quad \ell_1 \leq \ell_2, \\ \mathcal{I} \qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{D.1}$$

Then, the partial gradient of $f_{\mathbf{W}(k)}(x)$ with respect to $\mathbf{W}_\ell^{(k)}$ for $1 \leq \ell \leq L$ has a following form: for $i \in \{1, \ldots, n\}$,

$$\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_i)\big] = \sqrt{m} \cdot \left[\mathbf{x}_{\ell-1,i}^{(k)} \mathbf{v}^{\mathrm{T}} \bigg(\prod_{r=\ell+1}^{L} \mathbf{\Sigma}_{r,i}^{(k)} \mathbf{W}_r^{(k)}\bigg) \mathbf{\Sigma}_{\ell,i}^{(k)}\right]^{\top}, \qquad \ell \in [L],$$

where $\mathbf{\Sigma}_{\ell,i}^{(k)} := \text{Diag}\big(\mathbb{1}\big(\langle \mathbf{w}_{\ell,1}^{(k)}, \mathbf{x}_{\ell-1,i}^{(k)} \rangle \geq 0\big), \ldots, \mathbb{1}\big(\langle \mathbf{w}_{\ell,m}^{(k)}, \mathbf{x}_{\ell-1,i}^{(k)} \rangle \geq 0\big)\big) \in \mathbb{R}^{m \times m}$ and $\mathbf{w}_{\ell,j}^{(k)}$ denotes jth column of the matrix $\mathbf{W}_\ell^{(k)}$.

Gram matrix $\mathbf{H}(k)$**.** Each entries of empirical gram matrix evaluated at the kth GD update are defined as follows:

$$\mathbf{H}_{i,j}(k) = \frac{1}{m} \sum_{\ell=1}^{L} \big\langle \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_i)\big], \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_j)\big] \big\rangle_{\mathrm{Tr}}.$$

Note that $\mathbf{H}(0) \to \mathbf{H}_L^{\infty}$ as $m \to \infty$ which is proved in [131, 153, 132, 130].

Perturbation region of weight matrices. Consider a collection of weight matrices $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_\ell\}_{\ell=1,\ldots,L}$ such that

$$\widetilde{\mathbf{W}} \in \mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big) := \Big\{\widetilde{\mathbf{W}}_\ell : \|\widetilde{\mathbf{W}}_\ell - \mathbf{W}_\ell^{(0)}\|_2 \leq \tau, \quad \forall \ell \in [L]\Big\}. \tag{D.2}$$

For all $i \in \{1, \ldots, n\}$ and $\ell = 1, \ldots, L$, we denote $\mathbf{x}_{\ell,i}$ and $\widetilde{\mathbf{x}}_{\ell,i}$ as the outputs of the ℓ-th layer of the neural network with weight matrices $\mathbf{W}^{(0)}$ and $\widetilde{\mathbf{W}}$, and $\Sigma_{\ell,i}$ and $\widetilde{\Sigma}_{\ell,i}$ are diagonal matrices with $\big(\Sigma_{\ell,i}\big)_{jj} = \mathbb{1}\big(\langle \mathbf{w}_{\ell,j}^{(0)}, \mathbf{x}_{\ell-1,i}\rangle \geq 0\big)$ and $\big(\widetilde{\Sigma}_{\ell,i}\big)_{jj} = \mathbb{1}\big(\langle \widetilde{\mathbf{w}}_{\ell,j}, \widetilde{\mathbf{x}}_{\ell-1,i}\rangle \geq 0\big)$, respectively.

D.3 Why is it hard to prove $\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\|_2 \leq \mathcal{O}(1)$?

In this subsection, we provide a heuristic argument on why it is hard to prove $\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\|_2 \leq \mathcal{O}(1)$, where $\mathbf{W}_{D,\ell}^{(k)}$ is the model parameter of ℓth layer in kth iteration of algorithm. Here, we regularize solely on the model parameter, instead on the relative to the initialization. In this case, we can write the update rule as follows :

$$\mathbf{W}_{D,\ell}^{(k)} = \big(1 - \eta_2\mu\big)\mathbf{W}_{D,\ell}^{(k-1)} - \eta_1 \nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D^{(k-1)}\big)\big], \quad \forall 1 \leq \ell \leq L \quad \text{and} \quad \forall k \geq 1. \tag{D.3}$$

By recursively applying above equation (4.3), we can write $\mathbf{W}_{D,\ell}^{(k)}$ with respect to $\mathbf{W}_{D,\ell}^{(0)}$ as follows:

$$\mathbf{W}_{D,\ell}^{(k)} = (1 - \eta_2\mu)^k \mathbf{W}_{D,\ell}^{(0)} - \eta_1 \sum_{\ell=0}^{k-1} (1 - \eta_2\mu)^\ell \nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D^{(k-\ell-1)}\big)\big].$$

Then, we can control the bound as follows:

$$\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\|_2 \leq \Big(1 - \big(1-\eta_2\mu\big)^k\Big) \Big\|\mathbf{W}_{D,\ell}^{(0)}\Big\|_2 + \frac{\eta_1}{\eta_2\mu} \max_{\ell=0,\ldots,k-1} \Big\|\nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D^{(k-\ell-1)}\big)\big]\Big\|_2.$$

We know under the initialization setting in our paper, $\|\mathbf{W}_{D,\ell}^{(k)}\|_2 \leq \mathcal{O}(1)$ with high-probability (see [208]), and as long as we can prove the ℓ_2-norm of gradient is bounded, then we can conclude $\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\|_2 \leq \mathcal{O}\big(1\big)$. However, we are not aware of works in which they control the size of $\|\nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D^{(k-\ell-1)}\big)\big]\|_2$ where the non-convex interactions between model parameters across the hidden layers are allowed. To the best of our knowledge, we know the work [209] deals with the three layer case under this setting. But we need further investigations on whether the techniques employed in their paper can be generalized to arbitrary L-hidden layer setting.

D.4 Useful Lemmas

A simple fact. Suppose $\mathbf{v}_j \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \frac{\omega}{m})$ for $j \in [m]$. Then, with probability at least $1 - \exp\left[-\Omega(m)\right]$, $\|\mathbf{v}\|_2^2 \leq \mathcal{O}(\omega)$.

Proof. Since $\big\|\mathbf{v}_j^2\big\|_{\Psi_1} \leq \mathcal{O}\big(\frac{\omega}{m}\big)$ for $j \in [m]$, where $\|\cdot\|_{\Psi_1}$ denotes a sub-exponential norm, Bernstein's inequality for i.i.d. centered sub-exponential random variables can be employed : For any $t \geq 0$,

$$\mathbb{P}\bigg(\bigg|\sum_{j=1}^{m}\bigg(\mathbf{v}_j^2 - \frac{\omega}{m}\bigg)\bigg| \geq t\bigg) \leq 2\exp\bigg(-c\min\bigg(\frac{t^2}{\sum_{j=1}^{m}\big\|\mathbf{v}_j^2\big\|_{\Psi_1}^2}, \frac{t}{\max_j \big\|\mathbf{v}_j^2\big\|_{\Psi_1}}\bigg)\bigg), \tag{D.4}$$

where $c > 0$ is an absolute constant. Note that we used the fact centering does not hurt the sub-exponentiality of random variable. Choosing $t = \mathcal{O}\big(\omega\big)$ concludes the proof. □

Lemma D.4.1 (Lemma 7.1. [136]) *With probability at least* $1 - \mathcal{O}(nL)\cdot\exp[-\Omega(m/L)]$, $3/4 \leq \|\mathbf{x}_{\ell,i}^{(0)}\|_2 \leq 5/4$ *for all* $i \in \{1,\ldots,n\}$ *and* $\ell \in \{1,\ldots,L\}$.

Lemma D.4.2 (Lemma B.1. [140]) *If $\tau \leq \mathcal{O}\big(L^{-9/2}[\log(m)]^{-3}\big)$, then with probability at least $1 - \mathcal{O}(nL) \cdot \exp[-\Omega(m\tau^{2/3}L)]$, $1/2 \leq \|\widetilde{\mathbf{x}}_{\ell,i}\|_2 \leq 3/2$ for all $\widetilde{\mathbf{W}} \in \mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big)$, $i \in \{1, \ldots, n\}$ and $\ell \in \{1, \ldots, L\}$.*

Lemma D.4.3 ([136]) *Uniformly over $i \in \{1, \ldots, n\}$ and $1 \leq \ell_1 \leq \ell_2 \leq L$, the following results hold:*

1. *(**Lemma.7.3,** [136]) Suppose $m \geq \Omega(nL\log(nL))$, then with probability at least, $1 - \mathcal{O}\big(nL^2\big) \cdot \exp[-\Omega(m\tau^{2/3}L)]$,*

$$\left\|\prod_{r=\ell_1}^{\ell_2} \mathbf{\Sigma}_{r,i}^{(0)} \mathbf{W}_r^{(0)}\right\|_2 \leq \mathcal{O}\big(\sqrt{L}\big).$$

2. *(**Lemma.7.4,** [136]) Suppose $m \geq \Omega(nL\log(nL))$, then with probability at least, $1 - \mathcal{O}\big(nL\big) \cdot \exp[-\Omega(m/L)]$,*

$$\left\|\mathbf{v}^\top\Big(\prod_{r=\ell_1}^{L} \mathbf{\Sigma}_{r,i}^{(0)} \mathbf{W}_r^{(0)}\Big)\right\|_2 \leq \mathcal{O}\big(\sqrt{w}\big).$$

3. *(**Lemma.8.2,** [136]) Suppose $\tau \leq \mathcal{O}\big(L^{-9/2}[\log(m)]^{-3}\big)$. For all $\widetilde{\mathbf{W}} \in \mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big)$, with probability at least, $1 - \mathcal{O}\big(nL^2\big) \cdot \exp[-\Omega(m\tau^{2/3}L)]$,*

$$\left\|\widetilde{\mathbf{x}}_{\ell_1,i} - \mathbf{x}_{\ell_1,i}^{(0)}\right\|_2 \leq \mathcal{O}\big(\tau L^{5/2}\sqrt{\log(m)}\big).$$

4. *(**Corollary.8.4,** [136]) Suppose $\tau \leq \mathcal{O}\big(L^{-9/2}[\log(m)]^{-3}\big)$, then with probability at least, $1 - \mathcal{O}\big(nL^2\big) \cdot \exp[-\Omega(m\tau^{2/3}L)]$,*

$$\left\|\widetilde{\mathbf{\Sigma}}_{\ell_1,i} - \mathbf{\Sigma}_{\ell_1,i}^{(0)}\right\|_0 \leq \mathcal{O}\big(m\tau^{2/3}L\big).$$

5. *(**Lemma.8.7,** [136]) For all $\ell \in [L]$, let $\mathbf{\Sigma}''_{\ell,i} \in [-3, 3]^{m\times m}$ be the diagonal matrices with at most $s = \mathcal{O}(m\tau^{2/3}L)$ non-zero entries. For all $\widetilde{\mathbf{W}} \in \mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big)$, where*

$\tau = \mathcal{O}\big(\frac{1}{L^{1.5}}\big)$*, with probability at least* $1 - \mathcal{O}\big(nL\big) \cdot \exp[-\Omega(s\log(m))]$,

$$\left\|\mathbf{v}^{T}\Big(\prod_{r=\ell_1+1}^{L}(\boldsymbol{\Sigma}''_{r,i} + \boldsymbol{\Sigma}^{(0)}_{r,i})\widetilde{\mathbf{W}}_{r,i}\Big)\big(\boldsymbol{\Sigma}''_{\ell_1,i} + \boldsymbol{\Sigma}^{(0)}_{r,i}\big) - \mathbf{v}^{T}\Big(\prod_{r=\ell_1+1}^{L}\boldsymbol{\Sigma}^{(0)}_{r,i}\mathbf{W}^{(0)}_{r,i}\Big)\boldsymbol{\Sigma}^{(0)}_{\ell_1,i}\right\|_2$$
$$\leq \mathcal{O}\big(\tau^{1/3}L^2\sqrt{\omega\log(m)}\big).$$

Lemma D.4.4 (Lemma B.3. [140]) *There exists an absolute constant* κ *such that, with probability at least* $1 - \mathcal{O}\big(nL^2\big) \cdot \exp[-\Omega(m\tau^{2/3}L)]$, $i \in 1,\dots,n$ *and* $\ell \in 1,\dots,L$ *and for all* $\widetilde{\mathbf{W}} \in \mathcal{B}\big(\mathcal{W}^{(0)},\tau\big)$*, with* $\tau \leq \kappa L^{-6}[\log(m)]^{-3}$*, it holds uniformly that*

$$\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)\big]\big\|_2 \leq \mathcal{O}\big(\sqrt{\omega m}\big).$$

Lemma D.4.5 *Suppose* $\widetilde{\mathbf{W}} \in \mathcal{B}\big(\mathcal{W}^{(0)},\tau\big)$ *and* $\tau \leq \mathcal{O}\big(L^{-9/2}[\log(m)]^{-3}\big)$*. For all* $u \in \mathbb{R}^m$ *with a cardinality* $\|u\|_0 \leq s$*, for any* $1 \leq \ell \leq L$ *and* $i \in \{1,\dots,n\}$*, with probability at least* $1 - \mathcal{O}(nL)\cdot\exp\big(-\Omega(s\log(m))\big) - \mathcal{O}(nL)\cdot\exp\big(-\Omega(m\tau^{2/3}L)\big)$,

$$\left|\mathbf{v}^{\top}\Big(\prod_{r=\ell}^{L}\widetilde{\boldsymbol{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_{r,i}\Big)u\right| \leq \sqrt{\frac{\omega s\log(m)}{m}}\cdot\mathcal{O}\big(\|u\|_2\big).$$

Proof. Recall Lemma D.4.2. For any fixed vector $u \in \mathbb{R}^m$, with probability at least $1 - \mathcal{O}(nL)\cdot\exp[-\Omega(m\tau^{2/3}L)]$ for $\tau \leq \mathcal{O}\big(L^{-9/2}[\log(m)]^{-3}\big)$, for any $1 \leq \ell \leq L$ and $i \in \{1,\dots,n\}$, we have the event $\mathcal{T}$,

$$\left\|\Big(\prod_{r=\ell}^{L}\widetilde{\boldsymbol{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_{r,i}\Big)u\right\|_2 \leq 3\left\|u\right\|_2. \tag{D.5}$$

Conditioned on this event happens, it is easy to see the random variable $\mathbf{v}^{\top}\big(\prod_{r=a}^{L}\widetilde{\boldsymbol{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_{r,i}\big)u \sim$

$SG\big(\frac{9\omega}{m}\|u\|_2^2\big)$. Based on this observation, we have the probability,

$$\begin{aligned}\mathbb{P}\bigg(\bigg|\mathbf{v}^\top\Big(\prod_{r=\ell}^{L}\widetilde{\boldsymbol{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_{r,i}\Big)u\bigg| \geq \sqrt{\frac{\omega s\log(m)}{m}}\cdot\mathcal{O}\big(\|u\|_2\big)\bigg) \\ \leq \mathbb{P}\bigg(\bigg|\mathbf{v}^\top\Big(\prod_{r=\ell}^{L}\widetilde{\boldsymbol{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_{r,i}\Big)u\bigg| \geq \sqrt{\frac{\omega s\log(m)}{m}}\cdot\mathcal{O}\big(\|u\|_2\big)\;\Big|\;\mathcal{T}\bigg) + \mathbb{P}(\mathcal{T}^c) \\ \leq \mathcal{O}(nL)\cdot\exp\big(-\Omega(s\log(m))\big) + \mathcal{O}(nL)\cdot\exp\big(-\Omega(m\tau^{2/3}L)\big),\end{aligned}$$

where in the last inequality, union bounds over the indices ℓ and i, and over the vector $u \in \mathbb{R}^m$ with $\|u\|_0 \leq s$ are taken. □

Lemma D.4.6 *Suppose $\tau \leq \frac{1}{CL^{9/2}[\log(m)]^3}$ for some constant $C > 0$. Then, for all $i \in [n]$ and $\ell \in [L]$, with probability at least $1 - \mathcal{O}\big(nL\big)\cdot\exp[-\Omega(m\tau^{2/3}L)]$, we have*

$$\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}}(\mathbf{x}_i)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\big]\big\|_2 \leq \mathcal{O}\bigg(\tau^{1/3}L^2\sqrt{\omega m\log(m)}\bigg).$$

Proof. By using the results from Lemma D.4.3, we can control the term :

$$\begin{aligned}&\|\nabla_{\mathbf{W}_\ell}\;\big[f_{\mathbf{W}^{(k)}}(\mathbf{x}_i)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\big]\;\|_2 \\ &= \sqrt{m}\cdot\bigg\|\mathbf{x}_{\ell-1}^{(k)}\mathbf{v}^{\mathrm{T}}\Big(\prod_{r=\ell+1}^{L}\boldsymbol{\Sigma}_r^{(k)}\mathbf{W}_r^{(k)}\Big)\boldsymbol{\Sigma}_\ell^{(k)} - \mathbf{x}_{\ell-1}^{(0)}\mathbf{v}^{\mathrm{T}}\Big(\prod_{r=\ell+1}^{L}\boldsymbol{\Sigma}_r^{(0)}\mathbf{W}_r^{(0)}\Big)\boldsymbol{\Sigma}_\ell^{(0)}\bigg\|_2 \\ &\leq \sqrt{m}\cdot\underbrace{\|\mathbf{x}_{\ell-1}^{(k)} - \mathbf{x}_{\ell-1}^{(0)}\;\|_2}_{\leq\mathcal{O}(\tau L^{5/2}\sqrt{\log(m)})}\cdot\underbrace{\bigg\|\mathbf{v}^{\mathrm{T}}\Big(\prod_{r=\ell+1}^{L}\boldsymbol{\Sigma}_r^{(k)}\mathbf{W}_r^{(k)}\Big)\boldsymbol{\Sigma}_\ell^{(k)}\bigg\|_2}_{\leq\mathcal{O}(\sqrt{\omega})} \\ &\qquad + \sqrt{m}\cdot\underbrace{\Big\|\mathbf{x}_{\ell-1}^{(0)}\Big\|_2}_{\leq\mathcal{O}(1)}\cdot\underbrace{\bigg\|\mathbf{v}^{\mathrm{T}}\Big(\prod_{r=\ell+1}^{L}\boldsymbol{\Sigma}_r^{(k)}\mathbf{W}_r^{(k)}\Big)\boldsymbol{\Sigma}_\ell^{(k)} - \mathbf{v}^{\mathrm{T}}\Big(\prod_{r=\ell+1}^{L}\boldsymbol{\Sigma}_r^{(0)}\mathbf{W}_r^{(0)}\Big)\boldsymbol{\Sigma}_\ell^{(0)}\bigg\|_2}_{\leq\mathcal{O}(\tau^{1/3}L^2\sqrt{\omega\log(m)})} \\ &\leq \mathcal{O}\bigg(\tau^{1/3}L^2\sqrt{\omega m\log(m)}\bigg),\end{aligned}$$

where, in the last inequality, we used the condition on $\tau \leq \frac{1}{CL^{9/2}[\log(m)]^3} < 1$. □

Remark D.4.7 *Note that the results in Lemmas* 6.3 *(second and fifth items),* 6.4*,* 6.5*,* 6.6 *are in the setting of* $v_j \sim \mathcal{N}(0, \frac{\omega}{m})$ *for* $j \in [m]$.

For the notational convenience, in following Lemmas we denote $f_{\mathbf{W}(k)}(\mathbf{x}_i)$ as $\mathbf{u}_i(k)$ and let $\mathbf{u}(k) := [\mathbf{u}_1(k), \ldots, \mathbf{u}_n(k)]^\top$ for $k \geq 0$.

Lemma D.4.8 *For some* $\delta \in [0, 1]$*, if* $m \geq \Omega\big(L \log(nL/\delta)\big)$*, then with probability at least* $1 - \delta$*,* $\|\mathbf{u}(k)\|_2 \leq \mathcal{O}\big(\frac{\sqrt{n\omega}}{\delta}\big)$ *for any* $k \geq 0$.

Proof. Recall the Lemma D.4.2 stating that $\left\|\mathbf{x}_{L,i}^{(k)}\right\|_2 = \mathcal{O}(1)$ for any input data $\mathbf{x}_i$ for $i \in [n]$. Also recall that $\mathbf{v}_j \sim \mathcal{N}(0, \frac{\omega}{m})$ for $j \in [m]$, $\mathbf{x}_{L,i} \in \mathbb{R}^m$ and $\mathbf{u}_i(k) = \sqrt{m}\mathbf{v}^\top \mathbf{x}_{L,i} \sim \mathcal{N}\big(0, \mathcal{O}(\omega)\big)$. Then, we have a following via simple Markov inequality: for any $t \geq 0$,

$$\mathbb{P}\bigg(\ \|\mathbf{u}(k)\|_2 \geq t\bigg) \leq \frac{\mathbb{E}\big[\ \|\mathbf{u}(k)\|_2\ \big]}{t} \leq \frac{\sqrt{\mathbb{E}\big[\ \|\mathbf{u}(k)\|_2^2\ \big]}}{t} \leq \frac{\mathcal{O}\big(\sqrt{n\omega}\big)}{t}.$$

□

Lemma D.4.9 *For some* $\delta \in [0, 1]$*, if* $m \geq \Omega\big(L \log(nL/\delta)\big)$*, then with probability at least* $1 - \delta$*, we have*

$$\|\mathbf{u}(0) - \mathbf{y}\|_2 \leq \mathcal{O}\bigg(\sqrt{\frac{n}{\delta}}\bigg).$$

Proof. By Markov's inequality, for any $t \geq 0$,

$$\mathbb{P}\bigg(\ \|\mathbf{u}(0) - \mathbf{y}\|_2 \geq t\bigg) \leq \frac{\mathbb{E}_{\varepsilon, \mathbf{W}(0), \mathbf{v}}\Big[\ \|\mathbf{u}(0) - \mathbf{y}\|_2^2\Big]}{t^2}. \tag{D.6}$$

Note that the expectation in the nominator of (D.6) is taken over the random noise ε and initialized parameter $\mathbf{W}(0)$, $\mathbf{v}$. We can expand the nominator as follows:

$$\mathbb{E}_{\varepsilon, \mathbf{W}(0), \mathbf{v}}\Big[\ \|\mathbf{u}(0) - \mathbf{y}\|_2^2\Big] = \mathbb{E}_{\mathbf{W}(0), \mathbf{v}}\|\mathbf{u}(0)\|_2^2 + \mathbb{E}_\varepsilon\|\mathbf{y}\|_2^2 - 2\mathbb{E}_{\varepsilon, \mathbf{W}(0), \mathbf{v}}\Big[\mathbf{y}^\top \mathbf{u}(0)\Big]. \tag{D.7}$$

For the convenience of notation, let $\mathbf{y}^* := [f_\rho^\star(\mathbf{x}_1), \dots, f_\rho^\star(\mathbf{x}_n)]^\top$ and $\varepsilon := [\varepsilon_1, \dots, \varepsilon_n]^\top$. Recall that we have $\mathbf{y} = \mathbf{y}^* + \varepsilon$, and $\|\mathbf{y}^*\|_2^2 = \mathcal{O}(n)$. Also note that by Lemma D.4.1, with probability at least $1 - \mathcal{O}(nL) \cdot \exp[-\Omega(m/L)]$, for any $i = 1, \dots, n$, $\|\mathbf{x}_{L,i}^{(0)}\|_2^2 = \mathcal{O}(1)$. Then, we have a random variable $\mathbf{u}_i(0) = \sqrt{m}\mathbf{v}^\top \mathbf{x}_{L,i} \sim \mathcal{N}(0, \mathcal{O}(\omega))$. Now, we are ready to derive the orders of three terms on the RHS of (D.7). $\mathbb{E}_{\mathbf{W}(0),\mathbf{v}}\|\mathbf{u}(0)\|_2^2 = \mathcal{O}(n), \mathbb{E}_\varepsilon\|\mathbf{y}\|_2^2 = \mathbb{E}_\varepsilon\Big[\|\mathbf{y}^*\|_2^2 + \|\varepsilon\|_2^2 - 2\mathbf{y}^\top\varepsilon\Big] = \mathcal{O}(n), \mathbb{E}_{\varepsilon,\mathbf{W}(0),\mathbf{v}}\Big[\mathbf{y}^\top\mathbf{u}(0)\Big] = \mathbb{E}_{\varepsilon,\mathbf{W}(0),\mathbf{v}}\Big[\big(\mathbf{y}^* + \varepsilon\big)^\top\mathbf{u}(0)\Big] = 0.$ Combining the above three equalities, we conclude the proof. □

Lemma D.4.10 *Suppose* $\tau = \mathcal{O}\big(\frac{n\sqrt{\omega}}{\sqrt{m\delta}\lambda_\infty}\big)$. *For some* $\delta \in [0,1]$ *such that* $\delta \geq \mathcal{O}(nL) \cdot \exp[-\Omega(m\tau^{2/3}L)]$*, then with probability at least* $1-\delta$*, we have*

$$\|\mathbf{H}(k) - \mathbf{H}(0)\|_2 \leq \mathcal{O}\left(\omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\delta\lambda_\infty^2}}\right).$$

Proof. By the definition of gram matrix $\mathbf{H}_{i,j}(k)$ for any $k \geq 0$, we have

$$\begin{aligned}
&|\mathbf{H}_{i,j}(k) - \mathbf{H}_{i,j}(0)| \\
&\quad = \left|\frac{1}{m}\sum_{\ell=1}^{L}\Big\langle \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_i)\big], \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_j)\big]\Big\rangle_{\mathrm{Tr}} - \Big\langle \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_i)\big], \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_j)\big]\Big\rangle_{\mathrm{Tr}}\right| \\
&\quad \leq \frac{1}{m}\sum_{\ell=1}^{L}\bigg\{\left|\Big\langle \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_i)\big], \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_j)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_j)\big]\Big\rangle_{\mathrm{Tr}}\right| \\
&\qquad\qquad + \left|\Big\langle \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_j)\big], \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_i)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_i)\big]\Big\rangle_{\mathrm{Tr}}\right|\bigg\} \\
&\quad \leq \frac{1}{m}\sum_{\ell=1}^{L}\bigg\{\underbrace{\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_i)\big]\big\|_2}_{\leq\mathcal{O}\left(\sqrt{\omega m}\right)} \cdot \underbrace{\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_j)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_j)\big]\big\|_2}_{\leq\mathcal{O}\left(\tau^{1/3}L^2\sqrt{\omega m\log(m)}\right)} \\
&\qquad\qquad + \underbrace{\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_j)\big]\big\|_2}_{\leq\mathcal{O}\left(\sqrt{\omega m}\right)} \cdot \underbrace{\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_i)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_i)\big]\big\|_2}_{\leq\mathcal{O}\left(\tau^{1/3}L^2\sqrt{\omega m\log(m)}\right)}\bigg\} \\
&\quad \leq \mathcal{O}\left(\omega^{7/6} n^{1/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\delta\lambda_\infty^2}}\right).
\end{aligned}$$

In the second inequality, Lemmas D.4.4 and D.4.6 are used, and in the last inequality, $\tau = \mathcal{O}\big(\frac{n\sqrt{\omega}}{\sqrt{m\delta}\lambda_\infty}\big)$ is plugged in. With this, using the fact that Frobenius norm of a matrix is bigger than the operator norm, we bound the term $\|\mathbf{H}(k) - \mathbf{H}(0)\|_2$ as follows:

$$\|\mathbf{H}(k) - \mathbf{H}(0)\|_2 \leq \|\mathbf{H}(k) - \mathbf{H}(0)\|_F \leq \mathcal{O}\left(\omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\delta\lambda_\infty^2}}\right).$$

□

Lemma D.4.11 *For some $\delta \in [0, 1]$, with probability at least $1 - \delta$,*

$$\|\mathbf{H}_L^\infty - \mathbf{H}(0)\|_2 \leq \mathcal{O}\left(\omega n L^{5/2} \sqrt[4]{\tfrac{\log(nL/\delta)}{m}}\right)$$

Proof. For some $\delta^{'} \in [0, 1]$, set $\varepsilon = L^{3/2}\sqrt[4]{\frac{\log(L/\delta^{'})}{m}}$ from Theorem 3.1. of [130]. For any fixed $i, j \in [n]$, we have

$$\mathbb{P}\left[\left|\mathbf{H}_{i,j}^\infty - \mathbf{H}_{i,j}(0)\right| \leq \mathcal{O}\left(\omega L^{5/2}\sqrt[4]{\tfrac{\log(L/\delta^{'})}{m}}\right)\right] \geq 1 - \delta^{'}.$$

After applying the union bound over all $i, j \in [n]$, setting $\delta = \frac{\delta^{'}}{n^2}$, and using the fact that Frobenius norm of a matrix is bigger than the operator norm, we conclude the proof. □

For two positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$, if we write $\mathbf{A} \succeq \mathbf{B}$, then it means $\mathbf{A} - \mathbf{B}$ is positive semi-definite matrix. Similarly, if we write $\mathbf{A} \succ \mathbf{B}$, then it means $\mathbf{A} - \mathbf{B}$ is positive definite matrix. With these notations, we introduce a following Lemma.

Lemma D.4.12 (Lemma D.6. [142]) *For two positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$,*

1. *Suppose $\mathbf{A}$ is non-singular, then $\mathbf{A} \succeq \mathbf{B} \iff \lambda_{max}(\mathbf{B}\mathbf{A}^{-1}) \leq 1$ and $\mathbf{A} \succ \mathbf{B} \iff \lambda_{max}(\mathbf{B}\mathbf{A}^{-1}) < 1$, where $\lambda_{max}(\cdot)$ denotes the maximum eigenvalue of the input matrix.*

2. *Suppose $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{Q}$ are positive definite matrices, $\mathbf{A}$ and $\mathbf{B}$ are exchangeable, then $\mathbf{A} \succeq \mathbf{B} \Longrightarrow \mathbf{AQA} \succeq \mathbf{BQB}$.*

D.5 Proof of Theorem 4.3.5

Denote $u_i(k) = f_{\mathbf{W}(k)}(\mathbf{x}_i)$ and let $\mathbf{u}(k) = \big[u_1(k), u_2(k), \ldots, u_n(k)\big]^\top$. In order to achieve linear convergence rate of the training error, $\|\mathbf{u}(k) - y\|_2^2$, we decompose the term as follows:

$$\begin{aligned}\|\mathbf{u}(k+1) - y\|_2^2 &= \big\|\mathbf{u}(k) - y + \big(\mathbf{u}(k+1) - \mathbf{u}(k)\big)\big\|_2^2 \\ &= \|\mathbf{u}(k) - y\|_2^2 - 2\big(\mathbf{u}(k) - y\big)^\top\big(\mathbf{u}(k+1) - \mathbf{u}(k)\big) + \|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2 .\end{aligned}$$

Equipped with this decomposition, the proof consists of the following steps:

1. Similarly with [152], a term $\big(\mathbf{u}(k+1) - \mathbf{u}(k)\big)$ is decomposed into two terms, where we denote them as $\mathbf{I}_1^{(k)}$ and $\mathbf{I}_2^{(k)}$, respectively. We note that the first term $\mathbf{I}_1^{(k)}$ is related with a gram matrix $\mathbf{H}(k)$ and a second term $\mathbf{I}_2^{(k)}$ can be controlled small enough in ℓ_2 sense with proper choices of the step size and the width of network.

2. A term $\|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2$ needs to be controlled small enough to ensure $2\big(\mathbf{u}(k) - y\big)^\top\big(\mathbf{u}(k+1) - \mathbf{u}(k)\big) > \|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2$ so that the loss decreases.

3. It is shown that the distance between the gram matrix $\mathbf{H}(k)$ and the NTK matrix $\mathbf{H}_L^\infty$ is close enough in terms of operator norm.

4. Lastly, we inductively show that the weights generated from gradient descent stay within a perturbation region $\mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big)$, irrespective with the number of iterations of algorithm,

We start the proof by analyzing the term $\mathbf{u}(k+1) - \mathbf{u}(k)$.

<u>Step 1.</u> Control on $\mathbf{u}(\mathbf{k}+\mathbf{1}) - \mathbf{u}(\mathbf{k})$. Recall $\big(\boldsymbol{\Sigma}_{\ell,i}^{(k)}\big)_{jj} = \mathbb{1}\big(\langle \mathbf{w}_{\ell,j}^{(k)}, \mathbf{x}_{\ell-1,i}^{(k)}\rangle \geq 0\big)$ and we introduce a diagonal matrix $\widetilde{\boldsymbol{\Sigma}}_{\ell,i}^{(k)}$, whose jth entry is defined as follows:

$$\big(\widetilde{\boldsymbol{\Sigma}}_{\ell,i}^{(k)}\big)_{jj} = \big(\boldsymbol{\Sigma}_{\ell,i}^{(k+1)} - \boldsymbol{\Sigma}_{\ell,i}^{(k)}\big)_{jj} \cdot \frac{\big\langle \mathbf{w}_{\ell,j}^{(k+1)}, \mathbf{x}_{\ell-1,i}^{(k+1)} \big\rangle}{\big\langle \mathbf{w}_{\ell,j}^{(k+1)}, \mathbf{x}_{\ell-1,i}^{(k+1)} \big\rangle - \big\langle \mathbf{w}_{\ell,j}^{(k)}, \mathbf{x}_{\ell-1,i}^{(k)} \big\rangle}.$$

With this notation, the difference $\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}$ can be rewritten via the recursive applications of $\widetilde{\boldsymbol{\Sigma}}_{\ell,i}^{(k)}$:

$$\begin{aligned}
\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)} &= \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\big(\mathbf{W}_L^{(k+1)}\mathbf{x}_{L-1,i}^{(k+1)} - \mathbf{W}_L^{(k)}\mathbf{x}_{L-1,i}^{(k)}\big) \\
&= \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\mathbf{W}_L^{(k+1)}\big(\mathbf{x}_{L-1,i}^{(k+1)} - \mathbf{x}_{L-1,i}^{(k)}\big) \\
&\quad + \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\big(\mathbf{W}_L^{(k+1)} - \mathbf{W}_L^{(k)}\big)\mathbf{x}_{L-1,i}^{(k)} \\
&= \sum_{\ell=1}^{L} \Bigg(\prod_{r=\ell+1}^{L} \big(\boldsymbol{\Sigma}_{r,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{r,i}^{(k)}\big)\mathbf{W}_r^{(k+1)} \Bigg)\big(\boldsymbol{\Sigma}_{\ell,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{\ell,i}^{(k)}\big)\big(\mathbf{W}_\ell^{(k+1)} - \mathbf{W}_\ell^{(k)}\big)\mathbf{x}_{\ell-1,i}^{(k)}
\end{aligned} \tag{D.8}$$

Then, we introduce following notations :

$$\mathbf{D}_{\ell,i}^{(k)} = \Bigg(\prod_{r=\ell+1}^{L} \boldsymbol{\Sigma}_{r,i}^{(k)}\mathbf{W}_r^{(k)} \Bigg)\boldsymbol{\Sigma}_{\ell,i}^{(k)}, \qquad \widetilde{\mathbf{D}}_{\ell,i}^{(k)} = \Bigg(\prod_{r=\ell+1}^{L} \big(\boldsymbol{\Sigma}_{r,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{r,i}^{(k)}\big)\mathbf{W}_r^{(k+1)} \Bigg)\big(\boldsymbol{\Sigma}_{\ell,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{\ell,i}^{(k)}\big).$$

Now, we can write $u_i(k+1) - u_i(k)$ by noting that $u_i(k) = \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}}\mathbf{x}_{L,i}^{(k)}$:

$$\begin{aligned}
u_i(k+1) - u_i(k) &= \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}}\big(\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\big) \\
&= \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \widetilde{\mathbf{D}}_{\ell,i}^{(k)} \Big(\mathbf{W}_\ell^{(k+1)} - \mathbf{W}_\ell^{(k)} \Big)\mathbf{x}_{\ell-1,i}^{(k)} \qquad \text{(D.9)} \\
&= \underbrace{-\eta\sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \mathbf{D}_{\ell,i}^{(k)} \nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}^{(k)}\big)\big]\mathbf{x}_{\ell-1,i}^{(k)}}_{\mathbf{I}_{1,i}^{(k)}} \\
&\qquad \underbrace{-\eta\sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \Big(\widetilde{\mathbf{D}}_{\ell,i}^{(k)} - \mathbf{D}_{\ell,i}^{(k)} \Big) \nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}^{(k)}\big)\big]\mathbf{x}_{\ell-1,i}^{(k)}}_{\mathbf{I}_{2,i}^{(k)}}
\end{aligned}$$

Here, $\mathbf{I}_{1,i}^{(k)}$ can be rewritten as follows:

$$\begin{aligned}
\mathbf{I}_{1,i}^{(k)} &= -\eta\sqrt{m}\cdot\mathbf{v}^{\mathrm{T}}\sum_{\ell=1}^{L}\mathbf{D}_{\ell,i}^{(k)}\sum_{j=1}^{n}\big(u_j(k)-y_j\big)\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}}(\mathbf{x}_j)\big]\mathbf{x}_{\ell-1,i}^{(k)}\\
&= -\eta\cdot\sum_{j=1}^{n}\big(u_j(k)-y_j\big)\cdot\Bigg(\sqrt{m}\sum_{\ell=1}^{L}\mathbf{v}^{\mathrm{T}}\mathbf{D}_{\ell,i}^{(k)}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}}(\mathbf{x}_j)\big]\mathbf{x}_{\ell-1,i}^{(k)}\Bigg)\\
&= -m\eta\cdot\sum_{j=1}^{n}\big(u_j(k)-y_j\big)\cdot\frac{1}{m}\sum_{\ell=1}^{L}\Bigg\langle\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}}(\mathbf{x}_i)\big],\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}}(\mathbf{x}_j)\big]\Bigg\rangle_{\mathrm{Tr}}\\
&= -m\eta\cdot\sum_{j=1}^{n}\big(u_j(k)-y_j\big)\cdot\mathbf{H}_{i,j}(k).
\end{aligned}$$

For $\mathbf{I}_{2,i}^{(k)}$, we need a more careful control. First, we pay our attention on bounding the term $\|\mathbf{v}^\top(\widetilde{\mathbf{D}}_{\ell,i}^{(k)}-\mathbf{D}_{\ell,i}^{(k)})\|_2$ as follows: By triangle inequality, we have

$$\left\|\mathbf{v}^\top\Big(\widetilde{\mathbf{D}}_{\ell,i}^{(k)}-\mathbf{D}_{\ell,i}^{(k)}\Big)\right\|_2 \leq \left\|\mathbf{v}^\top\Big(\mathbf{D}_{\ell,i}^{(k)}-\mathbf{D}_{\ell,i}^{(0)}\Big)\right\|_2+\left\|\mathbf{v}^\top\Big(\widetilde{\mathbf{D}}_{\ell,i}^{(k)}-\mathbf{D}_{\ell,i}^{(0)}\Big)\right\|_2. \qquad \text{(D.10)}$$

We control the first term of the right-hand side (R.H.S) in (D.10). By the fourth item of the Lemma D.4.3, we know $\|\mathbf{\Sigma}_{r,i}^{(k)}-\mathbf{\Sigma}_{r,i}^{(0)}\|_0\leq\mathcal{O}\big(m\tau^{2/3}L\big)$ and $|\big(\mathbf{\Sigma}_{r,i}^{(k)}-\mathbf{\Sigma}_{r,i}^{(0)}\big)_{j,j}|\leq 1$ for $j\in[m]$. Then, we can plug $\mathbf{\Sigma}_{r,i}''=\mathbf{\Sigma}_{r,i}^{(k)}-\mathbf{\Sigma}_{r,i}^{(0)}$ in the inequality of the fifth item of Lemma D.4.3. So, the first term of the R.H.S in (D.10) can be bounded by $\mathcal{O}\big(\tau^{1/3}L^2\sqrt{\omega\log(m)}\big)$.

The second term of the R.H.S in (D.10) can be similarly controlled as the first term. Observe that $|\big(\mathbf{\Sigma}_{r,i}^{(k)}+\widetilde{\mathbf{\Sigma}}_{r,i}^{(k)}\big)_{jj}|\leq 1$, then we have $|\big(\mathbf{\Sigma}_{r,i}^{(k)}+\widetilde{\mathbf{\Sigma}}_{r,i}^{(k)}-\mathbf{\Sigma}_{r,i}^{(0)}\big)_{j,j}|\leq 2$ for all $j\in[m]$. Note that by the definition of $\widetilde{\mathbf{\Sigma}}_{r,i}^{(k)}$, we have $\|\widetilde{\mathbf{\Sigma}}_{r,i}^{(k)}\|_0=\|\mathbf{\Sigma}_{r,i}^{(k+1)}-\mathbf{\Sigma}_{r,i}^{(k)}\|_0\leq\|\mathbf{\Sigma}_{r,i}^{(k+1)}-\mathbf{\Sigma}_{r,i}^{(0)}\|_0+\|\mathbf{\Sigma}_{r,i}^{(k)}-\mathbf{\Sigma}_{r,i}^{(0)}\|_0\leq\mathcal{O}\big(m\tau^{2/3}L\big)$. Thus, by the triangle inequality, we have $\|\mathbf{\Sigma}_{r,i}^{(k)}+\widetilde{\mathbf{\Sigma}}_{r,i}^{(k)}-\mathbf{\Sigma}_{r,i}^{(0)}\|_0\leq\mathcal{O}\big(m\tau^{2/3}L\big)$. These observations enable us to plug $\mathbf{\Sigma}_{r,i}''=\mathbf{\Sigma}_{r,i}^{(k)}+\widetilde{\mathbf{\Sigma}}_{r,i}^{(k)}-\mathbf{\Sigma}_{r,i}^{(0)}$ in the inequality of the fifth item of Lemma D.4.3, and give the bound on the second term as $\mathcal{O}\big(\tau^{1/3}L^2\sqrt{\omega\log(m)}\big)$.

We have $\|\mathbf{v}^\top(\widetilde{\mathbf{D}}_{\ell,i}^{(k)} - \mathbf{D}_{\ell,i}^{(k)})\|_2 \leq \mathcal{O}(\tau^{1/3}L^2\sqrt{\omega\log(m)})$. Now, we control the ℓ_2-norm of the $\mathbf{I}_2^{(k)}$ as follows:

$$\begin{aligned}
\left\|\mathbf{I}_2^{(k)}\right\|_2 &\leq \sum_{i=1}^{n}\left|\mathbf{I}_{2,i}^{(k)}\right| \\
&\leq \eta\sqrt{m}\cdot\sum_{i=1}^{n}\left[\sum_{\ell=1}^{L}\underbrace{\left\|\mathbf{v}^\top\left(\widetilde{\mathbf{D}}_{\ell,i}^{(k)} - \mathbf{D}_{\ell,i}^{(k)}\right)\right\|_2}_{\leq\mathcal{O}(L^2\tau^{1/3}\sqrt{\omega\log(m)})}\cdot\left\|\nabla_{\mathbf{W}_\ell}\left[\mathcal{L}_{\mathbf{S}}(\mathbf{W}^{(k)})\right]\right\|_2\cdot\underbrace{\left\|\mathbf{x}_{\ell-1,i}^{(k)}\right\|_2}_{\leq\mathcal{O}(1)\textbf{ : Lemma D.4.2}}\right] \\
&\leq \mathcal{O}\left(\eta nL^2\tau^{1/3}\sqrt{\omega m\log(m)}\right)\sum_{\ell=1}^{L}\left\|\nabla_{\mathbf{W}_\ell}\left[\mathcal{L}_{\mathbf{S}}(\mathbf{W}^{(k)})\right]\right\|_2 \\
&\leq \mathcal{O}\left(\eta nL^{5/2}\tau^{1/3}\sqrt{\omega m\log(m)}\right)\sqrt{\sum_{\ell=1}^{L}\left\|\nabla_{\mathbf{W}_\ell}\left[\mathcal{L}_{\mathbf{S}}(\mathbf{W}^{(k)})\right]\right\|_F^2} \\
&\leq \mathcal{O}\left(\eta nL^{5/2}\tau^{1/3}\sqrt{\omega m\log(m)}\right)\sqrt{\sum_{j=1}^{n}\left(u_j(k)-y_j\right)^2\sum_{\ell=1}^{L}\left\|\nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}^{(k)}}(\mathbf{x}_j)\right]\right\|_F^2} \\
&\leq \mathcal{O}\left(\eta nL^3\tau^{1/3}\omega m\sqrt{\log(m)}\right)\left\|\mathbf{u}(k)-y\right\|_2. \qquad \text{(D.11)}
\end{aligned}$$

Step 2. Control on $\|\mathbf{u}(\mathbf{k+1})-\mathbf{u}(\mathbf{k})\|_2^2$**.** Recall that by (D.9), $\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}$ can be written as follows:

$$\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)} = \sum_{\ell=1}^{L}\widetilde{\mathbf{D}}_{\ell,i}^{(k)}\left(\mathbf{W}_\ell^{(k+1)} - \mathbf{W}_\ell^{(k)}\right)\mathbf{x}_{\ell-1,i}^{(k)} = -\eta\cdot\sum_{\ell=1}^{L}\widetilde{\mathbf{D}}_{\ell,i}^{(k)}\nabla_{\mathbf{W}_\ell}\left[\mathcal{L}_{\mathbf{S}}(\mathbf{W}^{(k)})\right]\mathbf{x}_{\ell-1,i}^{(k)}.$$

It is worth noting that,

$$\begin{aligned}
\left\|\nabla_{\mathbf{W}_\ell}\left[\mathcal{L}_{\mathbf{S}}(\mathbf{W}^{(k)})\right]\right\|_2^2 &= \left\|\sum_{j=1}^{n}\left(u_j(k)-y_j\right)\nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}^{(k)}}(\mathbf{x}_j)\right]\right\|_2^2 \\
&\leq \sum_{j=1}^{n}\left(u_j(k)-y_j\right)^2\sum_{j=1}^{n}\left\|\nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}^{(k)}}(\mathbf{x}_j)\right]\right\|_2^2 \\
&\leq \mathcal{O}(nm\omega)\left\|\mathbf{u}(k)-y\right\|_2^2. \qquad \text{(D.12)}
\end{aligned}$$

Also, observe that $|\big(\mathbf{\Sigma}_{r,i}^{(k)} + \widetilde{\mathbf{\Sigma}}_{r,i}^{(k)}\big)_{jj}| \leq 1$ for all $j \in [m]$, so by Lemma A.3 of [138], we know $\|\widetilde{\mathbf{D}}_{\ell,i}^{(k)}\|_2 \leq \mathcal{O}\big(\sqrt{L}\big)$. Combining all the facts, we can conclude:

$$\begin{aligned}
\|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2 &= m \cdot \sum_{i=1}^{n} \left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_{L,i}^{(k+1)} - \mathbf{v}^{\mathrm{T}}\mathbf{x}_{L,i}^{(k)}\right)^2 \\
&\leq m \cdot \|\mathbf{v}\|_2^2 \sum_{i=1}^{n} \left\|\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\right\|_2^2 \\
&\leq \eta^2 m \cdot \|\mathbf{v}\|_2^2 \sum_{i=1}^{n} \left[\sum_{\ell=1}^{L} \left\|\widetilde{\mathbf{D}}_{\ell,i}^{(k)}\right\|_2^2 \cdot \left\|\nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}^{(k)}\big)\big]\right\|_2^2 \cdot \left\|\mathbf{x}_{\ell-1,i}^{(k)}\right\|_2^2\right] \\
&\leq \mathcal{O}\big(\eta^2 n^2 L^2 m^2 \omega^2\big) \|\mathbf{u}(k) - y\|_2^2 \\
&\leq \mathcal{O}\big(\eta^2 n^2 L^2 m^2\big) \|\mathbf{u}(k) - y\|_2^2, \qquad \text{(D.13)}
\end{aligned}$$

where in the third inequality, we additionally used the fact $\|\mathbf{v}\|_2^2 = \mathcal{O}(\omega)$ with probability at least $1 - \exp(-\Omega(m))$, and the inequality (D.12). In the last inequality, we used the assumption $\omega \leq 1$.

Step 3. ***$\lambda_{\textit{min}}\big(H(k)\big) \geq \frac{\lambda_\infty}{2}$ with sufficiently large m.*** Denote $\rho(A)$ as a sprectral radius of a matrix A. Then, we have

$$\begin{aligned}
\|\mathbf{H}(k) - \mathbf{H}_L^\infty\|_2 &\geq \rho\big(\mathbf{H}(k) - \mathbf{H}_L^\infty\big) \\
&\geq -\lambda_{\text{min}}\big(\mathbf{H}(k) - \mathbf{H}_L^\infty\big) \\
&\geq \lambda_{\text{min}}\big(\mathbf{H}_L^\infty\big) - \lambda_{\text{min}}\big(\mathbf{H}(k)\big) \\
&\geq \lambda_\infty - \lambda_{\text{min}}\big(\mathbf{H}(k)\big), \qquad \text{(D.14)}
\end{aligned}$$

where, in the second inequality, we used a triangle inequality, $\lambda_{\text{min}}\big(\mathbf{H}(k) - \mathbf{H}_L^\infty\big) + \lambda_{\text{min}}\big(\mathbf{H}_L^\infty\big) \leq \lambda_{\text{min}}\big(\mathbf{H}(k)\big)$. By Lemmas D.4.10 and D.4.11, let $m \geq \Omega\left(\omega^7 n^8 L^{18} \frac{\log^3(m)}{\lambda_\infty^8 \delta}\right)$ and $\tilde{\mathcal{O}}\left(\frac{\lambda_\infty^{4/3}\delta^{1/3}}{n^{4/3}L^4}\right) \leq$

$\omega \leq 1$, we have

$$\begin{aligned}
\|\mathbf{H}(k) - \mathbf{H}_L^\infty\|_2 &\leq \|\mathbf{H}(k) - \mathbf{H}(0)\|_2 + \|\mathbf{H}(0) - \mathbf{H}_L^\infty\|_2 \\
&\leq \mathcal{O}\left(\omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\delta\lambda_\infty^2}}\right) + \mathcal{O}\left(\omega n^2 L^{5/2} \sqrt[4]{\frac{\log(nL/\delta)}{m}}\right) \\
&\leq \mathcal{O}\left(\omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\delta\lambda_\infty^2}}\right) \\
&\leq \frac{\lambda_\infty}{2}. \qquad \text{(D.15)}
\end{aligned}$$

Thus, combining (D.14) and (D.15) yields that $\lambda_{\text{min}}\big(H(k)\big) \geq \frac{\lambda_\infty}{2}$.

Step 4. Concluding the proof. Recall that $\mathbf{I}_1^{(k)} = -m\eta \cdot \mathbf{H}(k)(\mathbf{u}(k) - y)$. Then observe that

$$\begin{aligned}
(\mathbf{u}(k) - y)^\top \mathbf{I}_1^{(k)} &= -\eta m \cdot (\mathbf{u}(k) - y)^\top \mathbf{H}(k)(\mathbf{u}(k) - y) \\
&\leq -\eta m \cdot \lambda_{\text{min}}\big(\mathbf{H}(k)\big) \|\mathbf{u}(k) - y\|_2^2 \\
&\leq -\eta m \cdot \frac{\lambda_\infty}{2} \|\mathbf{u}(k) - y\|_2^2 . \qquad \text{(D.16)}
\end{aligned}$$

We set the step size η, radius of perturbation region τ and network width m as follows,

$$\begin{aligned}
\eta &= \Omega\left(\frac{\lambda_\infty}{n^2 L^2 m}\right), \\
\tau &= \mathcal{O}\left(\frac{n\sqrt{\omega}}{\sqrt{m\delta}\lambda_\infty}\right), \\
m &\geq \Omega\left(\omega^7 n^8 L^{18} \frac{\log^3(m)}{\lambda_\infty^8 \delta}\right).
\end{aligned}$$

With the above settings, we can control the $\|\mathbf{u}(k+1) - y\|_2^2$ by combining (E.16), (E.10) and (E.12) as follows,

$$\begin{aligned}
&\|\mathbf{u}(k+1) - y\|_2^2 \\
&= \left\|\mathbf{u}(k) - y + \big(\mathbf{u}(k+1) - \mathbf{u}(k)\big)\right\|_2^2 \\
&= \|\mathbf{u}(k) - y\|_2^2 - 2\eta m \cdot \big(\mathbf{u}(k) - y\big)^\top \mathbf{H}(k)\big(\mathbf{u}(k) - y\big) \\
&\qquad\qquad - \big(\mathbf{u}(k) - y\big)^\top \mathbf{I}_2^{(k)} + \|\mathbf{u}(k+1) - \mathbf{u}(k)\|_2^2 \\
&\le \left(1 - \eta m \lambda_\infty + \mathcal{O}\big(\eta n L^3 \tau^{1/3} m\omega \sqrt{\log(m)}\big) + \mathcal{O}\big(\eta^2 n^2 L^2 m^2\big)\right) \|\mathbf{u}(k) - y\|_2^2 \\
&\le \left(1 - \frac{\eta m \lambda_\infty}{2}\right) \|\mathbf{u}(k) - y\|_2^2 .
\end{aligned}$$

So far, we have shown from Step 1 to Step 4 that given the radius of perturbation region τ has the order $\mathcal{O}\big(\frac{n\sqrt{\omega}}{\sqrt{m}\delta\lambda_\infty}\big)$, then we can show the training error drops linearly to 0 with the discount factor $(1 - \frac{\eta m \lambda_\infty}{2})$ along with the proper choices of η and m. It remains us to prove the iterates $\mathbf{W}_\ell^{(k)}$ for all $\ell \in [L]$ generated by GD algorithm indeed stay in the perturbation region $\mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big)$ over $k \ge 0$ with $\tau = \mathcal{O}\big(\frac{n\sqrt{\omega}}{\sqrt{m}\delta\lambda_\infty}\big)$.

Step 5. ***The order of the radius of perturbation region.*** We employ the induction process for the proof. The induction hypothesis is : $\forall s \in [k+1]$,

$$\left\|\mathbf{W}_\ell^{(s)} - \mathbf{W}_\ell^{(0)}\right\|_2 \le \eta \mathcal{O}\left(n\sqrt{\frac{m\omega}{\delta}}\right) \sum_{t=0}^{s-1} \left(1 - \frac{\eta m \lambda_\infty}{2}\right)^{\frac{t}{2}} \le \mathcal{O}\left(\frac{n\sqrt{\omega}}{\sqrt{m}\delta\lambda_\infty}\right). \tag{D.17}$$

First, it is easy to see it holds for $s = 0$. Now, suppose it holds for $s = 0, \dots, k$, we consider $s = k+1$.

$$\begin{aligned}
\left\|\mathbf{W}_\ell^{(k+1)} - \mathbf{W}_\ell^{(k)}\right\|_2 &\le \eta \cdot \mathcal{O}\Big(\sqrt{nm\omega}\Big) \sqrt{2\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}^{(k)}\big)} \\
&\le \eta \cdot \mathcal{O}\Big(\sqrt{nm\omega}\Big) \left(1 - \frac{\eta m \lambda_\infty}{2}\right)^{\frac{k}{2}} \mathcal{O}\left(\sqrt{\frac{n}{\delta}}\right),
\end{aligned} \tag{D.18}$$

where in the third inequality, we used Lemmas D.4.4. Note that since it is assumed that

$\mathbf{W}_\ell^{(k)} \in \mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big)$, the Lemma is applicable with $m \geq \Omega\Big(\omega^7 n^8 L^{18} \frac{\log^3(m)}{\lambda_\infty^8 \delta}\Big)$. Similarly, since it is assumed that the induction hypothesis holds for $s = 0, \ldots, k$, we can see $\|\mathbf{u}(k) - y\|_2^2 \leq \big(1 - \frac{\eta m \lambda_\infty}{2}\big)^k \|\mathbf{u}(0) - y\|_2^2$. This inequality is plugged in the last inequality with Lemma D.4.9.

By combining the inequalities (D.17) for $s \in [k]$ and (D.18), and triangle inequality, we conclude the proof:

$$\Big\|\mathbf{W}_\ell^{(k+1)} - \mathbf{W}_\ell^{(0)}\Big\|_2 \leq \eta \cdot \mathcal{O}\Big(\sqrt{nm\omega}\Big) \sum_{t=0}^{k} \Big(1 - \frac{\eta m \lambda_\infty}{2}\Big)^{\frac{t}{2}} \mathcal{O}\Big(\sqrt{\frac{n}{\delta}}\Big) \leq \mathcal{O}\Big(\frac{n\sqrt{\omega}}{\sqrt{m\delta}\lambda_\infty}\Big).$$

Proposition D.5.1 *Let $\delta \in [0,1]$, set the width of the network as $m \geq \Omega\big(\omega^7 n^8 L^{18} \frac{\log^3(m)}{\lambda_\infty^8 \delta^2}\big)$, and the step-size of gradient descent as $\eta = \mathcal{O}\big(\frac{\lambda_\infty}{n^2 L^2 m}\big)$. Then, with probability at least $1 - \delta$ over the randomness of initialized parameters $\mathcal{W}^{(0)}$, we have for $k = 0, 1, 2, \ldots,$*

$$\mathbf{u}(k) - y = \big(\mathbf{I} - \eta m \mathbf{H}_L^\infty\big)^k \big(\mathbf{u}(0) - y\big) + \xi(k),$$

where

$$\|\xi(k)\|_2 = k\Bigg(1 - \frac{\eta m \lambda_\infty}{2}\Bigg)^{k-1} \mathcal{O}\Bigg(\eta m \cdot \omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\lambda_\infty^2 \delta}}\Bigg) \|\mathbf{y} - \mathbf{u}(0)\|_2 .$$

Proof. Define $u_i(k) := f_{\mathbf{W}(k)}(\mathbf{x}_i)$, then we have

$$\begin{aligned}
\mathbf{u}(k+1) - \mathbf{u}(k) &= -\eta m \cdot \mathbf{H}(k)\big(\mathbf{u}(k) - y\big) + \mathbf{I}_1^{(k)} \\
&= -\eta m \cdot \mathbf{H}_L^\infty\big(\mathbf{u}(k) - y\big) - \eta m \cdot \big(\mathbf{H}(k) - \mathbf{H}_L^\infty\big)\big(\mathbf{u}(k) - y\big) + \mathbf{I}_1^{(k)} \\
&= -\eta m \cdot \mathbf{H}_L^\infty\big(\mathbf{u}(k) - y\big) + \mathbf{e}(k).
\end{aligned}$$

By recursively applying the above equality, we can easily derive a following for any $k \geq 0$,

$$\mathbf{u}(k) - y = \big(\mathbf{I} - \eta m \mathbf{H}_L^{\infty}\big)^k \big(\mathbf{u}(0) - y\big) + \underbrace{\sum_{t=0}^{k-1} \big(\mathbf{I} - \eta m \mathbf{H}_L^{\infty}\big)^t \mathbf{e}(k-1-t)}_{=\xi(k)}. \tag{D.19}$$

Now, we want to show $\xi(k)$ can be controlled in arbitrarily small number. First, $e(k)$ needs to be bounded in an ℓ_2 norm:

$$\begin{aligned} \|e(k)\|_2 &\leq \eta m \left\|\mathbf{H}_L^{\infty} - \mathbf{H}(k)\right\|_2 \|\mathbf{u}(k) - y\|_2 + \left\|\mathbf{I}_2^{(k)}\right\|_2 \\ &\leq \eta m \cdot \mathcal{O}\left(\omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\lambda_\infty^2 \delta}}\right) \|\mathbf{u}(k) - y\|_2 \,, \end{aligned}$$

where, in the second inequality, $\tau = \mathcal{O}\left(\frac{n\sqrt{\omega}}{\sqrt{m\delta}\lambda_\infty}\right)$ is plugged in (E.16). Equipped with the bound on $\|e(k)\|_2$, we can easily bound the $\|\xi(k)\|_2$ as follows:

$$\begin{aligned} &\left\|\sum_{t=0}^{k-1} \big(\mathbf{I} - \eta m \mathbf{H}_L^{\infty}\big)^t \mathbf{e}(k-1-t)\right\|_2 \\ &\quad\leq \sum_{t=0}^{k-1} \left\|\mathbf{I} - \eta m \mathbf{H}_L^{\infty}\right\|_2^t \|\mathbf{e}(k-1-t)\|_2 \\ &\quad\leq \sum_{t=0}^{k-1} \Big(1 - \eta m \lambda_\infty\Big)^t \mathcal{O}\left(\eta m \cdot \omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\lambda_\infty^2 \delta}}\right) \|\mathbf{u}(k-1-t) - y\|_2 \\ &\quad\leq \sum_{t=0}^{k-1} \Big(1 - \eta m \lambda_\infty\Big)^t \mathcal{O}\left(\eta m \cdot \omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\lambda_\infty^2 \delta}}\right) \left(1 - \frac{\eta m \lambda_\infty}{2}\right)^{k-1-t} \|\mathbf{u}(0) - y\|_2 \\ &\quad= k\left(1 - \frac{\eta m \lambda_\infty}{2}\right)^{k-1} \mathcal{O}\left(\eta m \cdot \omega^{7/6} n^{4/3} L^3 \sqrt[6]{\frac{\log^3(m)}{m\lambda_\infty^2 \delta^2}}\right) \|\mathbf{u}(0) - y\|_2 \,. \end{aligned} \tag{D.20}$$

Note that in the third inequality, we used the result from Theorem 1. □

D.6 Proof of Theorem 4.3.8

We begin the proof by decomposing the error $\widehat{f}_{\mathbf{W}^{(k)}}(x) - f^{*}(x)$ for any fixed $x \in \mathbf{Unif}(\mathcal{S}^{d-1})$ into two terms as follows:

$$\widehat{f}_{\mathbf{W}^{(k)}}(x) - f^{*}(x) = \underbrace{\big(\widehat{f}_{\mathbf{W}^{(k)}}(x) - g^{*}(x)\big)}_{\Delta_1} + \underbrace{\big(g^{*}(x) - f^{*}(x)\big)}_{\Delta_2}. \tag{D.21}$$

Here, we denote the solution of kernel regression with kernel $\mathbf{H}_L^{\infty}$ as $g^{*}(x)$, which is a minimum RKHS norm interpolant of the noise-free data set $\{\mathbf{x}_i, f_{\rho}^{\star}(\mathbf{x}_i)\}_{i=1}^{n}$. To avoid the confusion of the notation, we write $\mathbf{Ker}(x, \mathbf{X}) = \big(\mathbf{H}_L^{\infty}(x, \mathbf{x}_1), \ldots, \mathbf{H}_L^{\infty}(x, \mathbf{x}_n)\big)_{i=1}^{n} \in \mathbb{R}^n$ and let $\mathbf{y}^{*} = [f_{\rho}^{\star}(\mathbf{x}_1), \ldots, f_{\rho}^{\star}(\mathbf{x}_n)]^{\top}$. Then, we have a following closed form solution $g^{*}(x)$ as, $g^{*}(x) := \mathbf{Ker}(x, \mathbf{X})\big(\mathbf{H}_L^{\infty}\big)^{-1}\mathbf{y}^{*}$.

With the decomposition (D.21), the proof sketch of Theorem 3.3.3 is as follows.

1. Note that for any $\ell \in [L]$, we have $\widehat{f}_{\mathbf{W}^{(k)}}(x) = \langle \mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big), \mathbf{vec}\big(\mathbf{W}_{\ell}^{(k)}\big)\rangle$. We can write the term $\mathbf{vec}\big(\mathbf{W}_{\ell}^{(k)}\big)$ with respect to $\mathbf{vec}\big(\mathbf{W}_{\ell}^{(0)}\big)$, $\mathbf{H}_L^{\infty}$ and the residual term via recursive applications of GD update rule and the result from proposition 2.1. Readers can refer (D.22). Using the equality (D.22), we can further decompose Δ_1 into three terms. That is, $\Delta_1 = \Delta_{11} + \Delta_{12} + \Delta_{13}$. Then, using the boundedness of ℓ_2-norm of network gradient and the fact that the size of $\|\xi(k)\|_2$ can be controlled with wide enough network, we can control the size of $\|\Delta_{12}\|_2$ and $\|\Delta_{13}\|_2$ aribtarily small.

2. In the term Δ_2, the $g^{\star}$ is an interpolant based on noiseless data. For large enough data points, $g^{\star}$ converges fastly to $f^{\star}$ at the rate $\mathcal{O}_{\mathbb{P}}(\frac{1}{\sqrt{n}})$.

3. Lastly, the Δ_{11} is the only term that is involved with random error ε, and we show that $\|\Delta_{11}\|_2$ is bounded away from 0 for small and large GD iteration index k.

Step 1. ***Control on*** Δ_1. For n data points $\big(\mathbf{x}_1, \ldots, \mathbf{x}_n\big)$ and for the k^{th} updated parameter

$\mathbf{W}(k)$, denote:

$$\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{X})\big] = \Bigg[\mathbf{vec}\Big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_1)\big]\Big), \cdots, \mathbf{vec}\Big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{x}_n)\big]\Big)\Bigg].$$

Note that when $\ell = 1$, $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{X})\big] \in \mathbb{R}^{md\times n}$ and when $\ell = 2, \dots, L$, $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{X})\big] \in \mathbb{R}^{m^2\times n}$. With this notation, we can rewrite the Gradient Descent update rule as $\mathbf{vec}\big(\mathbf{W}_\ell^{(k+1)}\big) = \mathbf{vec}\big(\mathbf{W}_\ell^{(k)}\big) - \eta\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{X})\big]\big(\mathbf{u}(k) - \mathbf{y}\big), \quad k \geq 0$. Applying Proposition 3.3.3, we get :

$$\begin{aligned}
&\mathbf{vec}\big(\mathbf{W}_\ell^{(k)}\big) - \mathbf{vec}\big(\mathbf{W}_\ell^{(0)}\big) = -\eta\cdot\sum_{j=0}^{k-1}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(j)}(\mathbf{X})\big]\big(\mathbf{u}(j) - \mathbf{y}\big) \\
&\quad = \eta\cdot\sum_{j=0}^{k-1}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(j)}(\mathbf{X})\big]\big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^j\big(\mathbf{y} - \mathbf{u}(0)\big) - \eta\cdot\sum_{j=0}^{k-1}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{X})\big]\xi(j) \\
&\quad = \eta\cdot\sum_{j=0}^{k-1}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]\big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^j\big(y - \mathbf{u}(0)\big) - \eta\cdot\sum_{j=0}^{k-1}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(\mathbf{X})\big]\xi(j) \\
&\qquad\qquad + \eta\cdot\sum_{j=0}^{k-1}\Bigg(\Bigg[\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(j)}(\mathbf{X})\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]\Bigg]\big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^j\big(y - \mathbf{u}(0)\big)\Bigg) \\
&\quad = \eta\cdot\sum_{j=0}^{k-1}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]\big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^j\big(y - \mathbf{u}(0)\big) + \xi^{'}(k). \qquad \text{(D.22)}
\end{aligned}$$

First, we control ℓ_2-norm of the first term of $\xi^{'}(k)$ as follows: Note that $\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(j)}(\mathbf{X})\big]\|_F \leq \mathcal{O}\big(\sqrt{nm\omega}\big)$ by Lemma D.4.4 for $0 \leq j \leq k-1$. Then, we have

$$\begin{aligned}
&\Bigg\|\eta\cdot\sum_{j=0}^{k-1}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(j)}(\mathbf{X})\big]\xi(j)\Bigg\|_2 \\
&\quad \leq \sum_{j=0}^{k-1}\mathcal{O}\Big(\eta\sqrt{nm\omega}\Big)\mathcal{O}\Bigg(j\Big(1 - \frac{\eta m\lambda_\infty}{2}\Big)^{j-1}\Bigg)\mathcal{O}\Bigg(\eta m\cdot\omega^{7/6}n^{4/3}L^3\sqrt[6]{\frac{\log^3(m)}{m\lambda_\infty^2\delta}}\Bigg)\|\mathbf{y} - \mathbf{u}(0)\|_2 \\
&\quad \leq \mathcal{O}\Bigg(\frac{n^{11/6}L^3\omega^{5/3}}{m^{2/3}\lambda_\infty^{7/3}\delta^{1/6}}\sqrt{\log(m)}\Bigg)\|\mathbf{y} - \mathbf{u}(0)\|_2. \qquad \text{(D.23)}
\end{aligned}$$

In the second inequality, $\sum_{j=1}^\infty j\big(1 - \frac{\eta m\lambda_\infty}{2}\big)^j = \mathcal{O}\big(\frac{1}{\eta^2m^2\lambda_\infty^2}\big)$ is used. Then, we control

ℓ_2-norm of the second term of $\xi^{'}(k)$ as follows:

$$
\begin{aligned}
&\left\|\eta \cdot \sum_{j=0}^{k-1}\left[\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(j)}(\mathbf{X})\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]\right]\big(\mathbf{I} - \eta m \mathbf{H}_L^{\infty}\big)^{j}\big(y - \mathbf{u}(0)\big)\right\|_2 \\
&\leq \sum_{j=0}^{k-1} \eta \left\|\mathbf{I} - \eta m \mathbf{H}_L^{\infty}\right\|_2^{j} \left\|\mathbf{y} - \mathbf{u}(0)\right\|_2 \sqrt{\sum_{i=1}^{n}\left\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(j)}(\mathbf{x}_i)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_i)\big]\right\|_2^2} \\
&\leq \sum_{j=0}^{k-1} \eta\big(1 - \eta m \lambda_\infty\big)^{j} \mathcal{O}\left(\frac{n^{1/3} m^{1/3} L^2 \omega^{2/3}}{\lambda_\infty^{1/3}\delta^{1/6}}\sqrt{\log(m)}\right)\mathcal{O}\big(\sqrt{n}\big)\left\|\mathbf{y} - \mathbf{u}(0)\right\|_2 \\
&\leq \mathcal{O}\left(\frac{n^{5/6} L^2 \omega^{2/3}}{m^{2/3}\lambda_\infty^{4/3}\delta^{1/6}}\sqrt{\log(m)}\right)\left\|\mathbf{y} - \mathbf{u}(0)\right\|_2, \qquad \text{(D.24)}
\end{aligned}
$$

where in the second inequality, we used Lemmas D.4.6 with $\tau = \mathcal{O}\big(\frac{n\sqrt{\omega}}{\sqrt{m\delta\lambda_\infty}}\big)$.

Now, we are ready to control Δ_1 term. By using the equality (D.22), we can decompose the term Δ_1 as follows: Let us denote $G_k = \sum_{j=0}^{k-1} \eta m \big(\mathbf{I} - \eta m \mathbf{H}_L^{\infty}\big)^{j}$. Note that for any $\ell \in [L]$, $\widehat{f}_{\mathbf{W}^{(k)}}(x) = \langle \mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big), \mathbf{vec}\big(\mathbf{W}_\ell^{(k)}\big)\rangle$ and recall that $\mathbf{y} = \mathbf{y}^* + \varepsilon$. Then, for any fixed $\ell^{'} \in [L]$, we have:

$$
\begin{aligned}
\Delta_1 = &\underbrace{\left[\mathbf{Ker}(x, \mathbf{X})\big[G_k - \big(\mathbf{H}_L^{\infty}\big)^{-1}\big]\mathbf{y}^* + \mathbf{Ker}(x, \mathbf{X}) G_k \varepsilon\right]}_{=\Delta_{11}} \\
&\underbrace{\begin{aligned}+ &\left[\frac{1}{m}\sum_{\ell=1}^{L} \mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big)^{\top} \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big] - \mathbf{Ker}(x, \mathbf{X})\right] G_k \mathbf{y} \\
&- \frac{1}{m}\sum_{\ell:\ell \neq \ell'} \mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big)^{\top} \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big] G_k \mathbf{y}\end{aligned}}_{=\Delta_{12}} \\
&\underbrace{\begin{aligned}+ &\left[\left\langle \mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\big[f_{\mathbf{W}(k)}(x)\big]\big), \mathbf{vec}\big(\mathbf{W}_{\ell'}^{(0)}\big)\right\rangle + \mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\big[f_{\mathbf{W}(k)}(x)\big]\big)^{\top} \xi^{'}(k)\right. \\
&\left.- \frac{1}{m}\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\big[f_{\mathbf{W}(k)}(x)\big]\big)^{\top} \nabla_{\mathbf{W}_{\ell'}}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big] G_k \mathbf{u}(\mathbf{0})\right].\end{aligned}}_{=\Delta_{13}}
\end{aligned}
$$

(D.25)

Our goal in this step is to control $\|\Delta_{12}\|_2$ and $\|\Delta_{13}\|_2$. Then, in the third step, we will show $\|\Delta_{11}\|_2$ is the term, which governs the behavior of the prediction risk with respect to algorithm iteration k. First, we bound the ℓ_2 norm of the first term in Δ_{12} as:

$$\begin{aligned}
&\left\|\left[\frac{1}{m}\sum_{\ell=1}^{L}\mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big)^\top\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]-\mathbf{Ker}(x,\mathbf{X})\right]G_k\mathbf{y}\right\|_2 \\
&\quad\le\frac{1}{mL}\sum_{\ell=1}^{L}\underbrace{\big\|\mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big)-\mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(x)\big]\big)\big\|_2}_{\le\mathcal{O}\left(\tau^{1/3}L^2\sqrt{\omega m\log(m)}\right)\text{ : \textbf{Lemma D.4.6}}}\underbrace{\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]\big\|_F}_{\le\mathcal{O}\left(\sqrt{\omega nm}\right)\text{ : \textbf{Lemma D.4.4}}}\|G_k\mathbf{y}\|_2 \\
&\qquad+\frac{1}{L}\sqrt{\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{\ell=1}^{L}\big\langle\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(x)\big],\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{x}_i)\big]\big\rangle_{\mathrm{Tr}}-\mathbf{Ker}(x,\mathbf{x}_i)\right)^2}\;\|G_k\mathbf{y}\|_2 \\
&\quad\le\left\{\mathcal{O}\left(\frac{n^{5/6}L^2\omega^{7/6}}{m^{1/6}\delta^{1/6}\lambda_\infty^{1/3}}\sqrt{\log(m)}\right)+\mathcal{O}\left(\omega n^{1/2}L^{3/2}\sqrt[4]{\tfrac{\log(nL/\delta)}{m}}\right)\right\}\|G_k\|_2\,\|\mathbf{y}\|_2 \\
&\quad\le\mathcal{O}\left(\frac{n^{5/6}L^2\omega^{7/6}}{m^{1/6}\delta^{1/6}\lambda_\infty^{4/3}}\sqrt{\log(m)}\cdot\|\mathbf{y}\|_2\right)+\mathcal{O}\left(\frac{\omega n^{1/2}L^{3/2}}{\lambda_\infty}\sqrt[4]{\tfrac{\log(nL/\delta)}{m}}\cdot\|\mathbf{y}\|_2\right),
\end{aligned} \tag{D.26}$$

where, in the second inequality, we plugged $\tau=\mathcal{O}\big(\frac{n\sqrt{\omega}}{\sqrt{m\delta}\lambda_\infty}\big)$ in the result of Lemma D.4.6 and used Lemma D.4.11. In the last inequality, we used $\|G_k\|_2\le\mathcal{O}\big(\frac{1}{\lambda_\infty}\big)$. Similarly, we can control the ℓ_2 norm of the second term in Δ_{12} as follows:

$$\begin{aligned}
&\left\|\frac{1}{m}\sum_{\ell:\ell\neq\ell'}\mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big)^\top\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]G_k\mathbf{y}\right\|_2 \\
&\quad\le\frac{1}{m}\sum_{\ell:\ell\neq\ell'}\underbrace{\|\mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big)\|_2}_{\le\mathcal{O}\left(\sqrt{\omega m}\right)}\cdot\underbrace{\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(0)}(\mathbf{X})\big]\|_F}_{\le\mathcal{O}\left(\sqrt{\omega mn}\right)}\cdot\underbrace{\|G_k\|_2}_{\le\mathcal{O}\left(\frac{1}{\lambda_\infty}\right)}\|\mathbf{y}\|_2 \\
&\quad\le\mathcal{O}\left(\frac{\omega L\sqrt{n}}{\lambda_\infty}\right)\cdot\|\mathbf{y}\|_2.
\end{aligned} \tag{D.27}$$

We turn our attention to controlling $\|\Delta_{13}\|_2$. The first term in Δ_{13};

Recall that $\left\|\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\big[f_{\mathbf{W}(k)}(x)\big]\big)\right\|_2\le\mathcal{O}\big(\sqrt{m\omega}\big)$ by Lemma D.4.4. Then, the random variable $\mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}(k)}(x)\big]\big)^\top\mathbf{vec}\big(\mathbf{W}_\ell^{(0)}\big)$ is simply a $\mathcal{N}\big(0,\mathcal{O}(\omega)\big)$ for $1\le\ell\le L$. A straightforward application of Chernoff bound for normal random variable and taking union bound over the layer $1 \leq \ell \leq L$ yield that: with probability at least $1-\delta$,

$$\left|\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(k)}(x)\right]\big)^{\top}\mathbf{vec}\big(\mathbf{W}_{\ell'}^{(0)}\big)\right| \leq \mathcal{O}\left(\sqrt{\omega\log\left(\frac{L}{\delta}\right)}\right). \tag{D.28}$$

The ℓ_2 norm of the third term in Δ_{13} can be bounded as follows:

$$\begin{aligned}
&\left\|\frac{1}{m}\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(k)}(x)\right]\big)^{\top}\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(0)}(\mathbf{X})\right]G_k\mathbf{u}(\mathbf{0})\right\|_2 \\
&\quad\leq \frac{1}{m}\underbrace{\left\|\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(k)}(x)\right]\big)\right\|_2}_{\leq\mathcal{O}\left(\sqrt{m\omega}\right)}\underbrace{\left\|\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(0)}(\mathbf{X})\right]\right\|_F}_{\leq\mathcal{O}\left(\sqrt{\omega mn}\right)}\underbrace{\|G_k\mathbf{u}(\mathbf{0})\|_2}_{\leq\mathcal{O}\left(\frac{\sqrt{n\omega}}{\lambda_\infty\delta}\right)} \leq \mathcal{O}\left(\frac{n\omega^{3/2}}{\lambda_\infty\delta}\right).
\end{aligned} \tag{D.29}$$

In the last inequality, we used the Lemma D.4.8 and $\|G_k\|_2 \leq \mathcal{O}\big(\frac{1}{\lambda_\infty}\big)$. By combining (D.23), (D.24), (D.28), (D.29) with $\left\|\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(0)}(x)\right]\right\|_F \leq \mathcal{O}\big(\sqrt{m\omega}\big)$, we have a following :

$$\begin{aligned}
\|\Delta_{13}\|_2 &\leq \left\|\left\langle\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(k)}(x)\right]\big),\mathbf{vec}\big(\mathbf{W}_{\ell'}^{(0)}\big)\right\rangle\right\|_2 + \left\|\big(\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(0)}(x)\right]\big)\big)^{\top}\xi'(k)\right\|_2 \\
&\qquad + \left\|\frac{1}{m}\big(\mathbf{vec}\big(\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(0)}(x)\right]\big)^{\top}\nabla_{\mathbf{W}_{\ell'}}\left[f_{\mathbf{W}(0)}(\mathbf{X})\right]G_k\mathbf{u}(\mathbf{0})\right\|_2 \\
&\leq \mathcal{O}\left(\sqrt{\omega\log\left(\frac{L}{\delta}\right)}\right) + \mathcal{O}\left(\frac{n^{11/6}L^3\omega^{13/6}\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2}{m^{1/6}\lambda_\infty^{4/3}\delta^{1/6}}\sqrt{\log(m)}\right) \\
&\qquad + \mathcal{O}\left(\frac{n^{5/6}L^2\omega^{7/6}\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2}{m^{1/6}\lambda_\infty^{7/3}\delta^{1/6}}\sqrt{\log(m)}\right) + \mathcal{O}\left(\frac{n\omega^{3/2}}{\lambda_\infty\delta}\right) \\
&= \mathcal{O}\left(\frac{n^{11/6}L^3\omega^{13/6}\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2}{m^{1/6}\lambda_\infty^{4/3}\delta^{1/6}}\sqrt{\log(m)}\right) + \mathcal{O}\left(\frac{n^{5/6}L^2\omega^{7/6}\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2}{m^{1/6}\lambda_\infty^{7/3}\delta^{1/6}}\sqrt{\log(m)}\right) \\
&\qquad + \mathcal{O}\left(\frac{n\omega^{3/2}}{\lambda_\infty\delta}\right).
\end{aligned} \tag{D.30}$$

Step 2. Control on Δ_2**.** First, note that there is a recent finding that the reproducing kernel Hilbert spaces of NTKs with any number of layers (i.e., $L \geq 1$) have the same set of functions, if kernels are defined on $\mathcal{S}^{d-1}$. See [157]. Along with this result, we can apply the proof used in Lemma.D.2. in [142] for proving a following :

$$\|\Delta_2\|_2 = \mathcal{O}_{\mathbb{P}}\bigg(\frac{1}{\sqrt{n}}\bigg). \tag{D.31}$$

Step 3. The behavior of L_2 risk is characterized by the term Δ_{11}. Recall the decompositions (D.21) and (D.25), then we have:

$$\widehat{f}_{\mathbf{W}^{(k)}}(x) - f^*(x) = \Delta_{11} + \big(\Delta_{12} + \Delta_{13} + \Delta_2\big) := \Delta_{11} + \Theta. \tag{D.32}$$

Our goal in this step is mainly two-folded: (1) Control $\mathbb{E}_\varepsilon \|\Theta\|_2^2$ arbitrarily small with proper choices of step-size of GD η and width of the network m. (2) Show that how $\mathbb{E}_\varepsilon \|\Delta_{11}\|_2^2$ affect the behavior of prediction risk over the GD iterations k. First, note that we have

$$\mathbb{E}_\varepsilon \|\mathbf{y}\|_2^2 = \mathbb{E}_\varepsilon \|\mathbf{y}^* + \varepsilon\|_2^2 \leq 2\big(\mathbf{y}^*\big)^\top \mathbf{y}^* + 2\mathbb{E}_\varepsilon \|\varepsilon\|_2^2 = \mathcal{O}(n). \tag{D.33}$$

Second, recall Lemma D.4.9 and note that over the random initialization, with probability at least $1-\delta$, $\mathbb{E}_\varepsilon \|\mathbf{y} - \mathbf{u}(0)\|_2^2 \leq \mathcal{O}\Big(\frac{n}{\delta}\Big)$. Now, by combining the bounds (D.26), (D.30) and (D.31), we have

$$\begin{aligned}
\mathbb{E}_\varepsilon \|\Theta\|_2^2 &\leq 3\mathbb{E}_\varepsilon\big(\|\Delta_{12}\|_2^2 + \|\Delta_{13}\|_2^2 + \|\Delta_2\|_2^2 \big) \\
&\leq \mathbb{E}_\varepsilon\Bigg[\mathcal{O}\Bigg(\frac{n^{5/3}L^4\omega^{7/3}}{m^{1/3}\lambda_\infty^{8/3}\delta^{1/3}}\log(m)\cdot\|\mathbf{y}\|_2^2\Bigg) + \mathcal{O}\Bigg(\frac{\omega^2 n L^3}{\lambda_\infty^2}\sqrt{\frac{\log(nL/\delta)}{m}}\cdot\|\mathbf{y}\|_2^2\Bigg) \\
&\qquad + \mathcal{O}\Bigg(\frac{n^{11/3}L^6\omega^{13/3}\|\mathbf{y}-\mathbf{u}(0)\|_2^2}{m^{1/3}\lambda_\infty^{8/3}\delta^{1/3}}\log(m)\Bigg) \\
&\qquad + \mathcal{O}\Bigg(\frac{n^{5/3}L^4\omega^{7/3}\|\mathbf{y}-\mathbf{u}(0)\|_2^2}{m^{1/3}\lambda_\infty^{14/3}\delta^{1/3}}\log(m)\Bigg)\Bigg] + \mathcal{O}\Bigg(\frac{n^2\omega^3}{\lambda_\infty^2\delta^2}\Bigg) + \mathcal{O}\bigg(\frac{1}{n}\bigg) \\
&\leq \mathcal{O}\Bigg(\frac{\omega^2 n^2 L^3}{\lambda_\infty^2}\sqrt{\frac{\log(nL/\delta)}{m}}\Bigg) + \mathcal{O}\Bigg(\frac{n^{14/3}L^6\omega^{13/3}}{m^{1/3}\lambda_\infty^{8/3}\delta^{4/3}}\log(m)\Bigg) + \mathcal{O}\Bigg(\frac{n^{8/3}L^4\omega^{7/3}}{m^{1/3}\lambda_\infty^{14/3}\delta^{4/3}}\log(m)\Bigg) \\
&\qquad + \mathcal{O}\Bigg(\frac{n^2\omega^3}{\lambda_\infty^2\delta^2}\Bigg) + \mathcal{O}\bigg(\frac{1}{n}\bigg),
\end{aligned} \tag{D.34}$$

where in the third inequality, we used (D.33) and (D.6).

<u>Case 1.</u> When k is large, the L_2 risk is bounded away from zero by some constant.

Now we control $\mathbb{E}_\varepsilon \left\|\Delta_{11}\right\|_2^2$. Recall the definitions $\|f\|_2^2 := \int_{\mathbf{x}\in\mathcal{S}^{d-1}} |f(\mathbf{x})|^2 d\mathbf{x}$ and $G_k = \sum_{j=0}^{k-1} \eta m \big(\mathbf{I} - \eta m \mathbf{H}_L^\infty\big)^j$. Let us denote $\mathbf{S} = \mathbf{y}^*\mathbf{y}^{*\top}$. Then, we have

$$\begin{aligned}\mathbb{E}_\varepsilon \left\|\Delta_{11}\right\|_2^2 &= \int_{x\in\mathcal{S}^{d-1}} \mathbf{Ker}(x,\mathbf{X})\bigg[\bigg(G_k - \big(\mathbf{H}_L^\infty\big)^{-1}\bigg)\mathbf{y}^*\mathbf{y}^{*\top}\bigg(G_k - \big(\mathbf{H}_L^\infty\big)^{-1}\bigg) + G_k^2\bigg]\mathbf{Ker}(\mathbf{X},x)dx \\ &= \int_{x\in\mathcal{S}^{d-1}} \mathbf{Ker}(x,\mathbf{X})\big(\mathbf{H}_L^\infty\big)^{-1} M_k \big(\mathbf{H}_L^\infty\big)^{-1}\mathbf{Ker}(\mathbf{X},x)dx\end{aligned}$$

where

$$\begin{aligned}M_k &= \big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^k \mathbf{S}\big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^k + \big(\mathbf{I} - \big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^k\big)^2 \\ &= \big[\big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^k - \big(\mathbf{S}+\mathcal{I}\big)^{-1}\big]\big(\mathbf{S}+\mathcal{I}\big)\big[\big(\mathbf{I} - \eta m\mathbf{H}_L^\infty\big)^k - \big(\mathbf{S}+\mathcal{I}\big)^{-1}\big] + \mathcal{I} - \big(\mathbf{S}+\mathcal{I}\big)^{-1}.\end{aligned}$$

For the algorithm iterations $k \geq \big(\frac{\log(n)}{\eta m\lambda_\infty}\big)C_0$ with some constant $C_0 > 1$, we have

$$\big(\mathcal{I} - \eta m\mathbf{H}_L^\infty\big)^k \preceq \big(1 - \eta m\lambda_\infty\big)^k \cdot \mathcal{I} \preceq \exp(-\eta m\lambda_\infty k)\cdot\mathcal{I} \preceq \exp(-C_0\log(n)) = \frac{1}{n^{C_0}}\cdot\mathcal{I}.$$

Since $1 + \|\mathbf{y}\|_2^2 \leq C_1 n$ for some constant C_1, we have

$$\lambda_{\max}\bigg(\frac{1}{n^{C_0}}\cdot\big(\mathbf{S}+\mathcal{I}\big)\bigg) = \frac{1+\|\mathbf{y}\|_2^2}{n^{C_0}} \leq \frac{C_1}{n^{C_0-1}} < 1. \tag{D.35}$$

Using the first item of Lemma (D.4.12) with the inequality (D.35), we have

$$\big(\mathcal{I} - \eta m\mathbf{H}_L^\infty\big)^k \preceq \frac{1}{n^{C_0}}\cdot\mathcal{I} \prec \big(\mathbf{S}+\mathcal{I}\big)^{-1}. \tag{D.36}$$

The above inequality (D.36) lead to a following result :

$$\big(\mathbf{S}+\mathcal{I}\big)^{-1}-\big(\mathcal{I}-\eta m\mathbf{H}_L^{\infty}\big)^{k}\succeq\big(\mathbf{S}+\mathcal{I}\big)^{-1}-\frac{1}{n^{C_0}}\cdot\mathcal{I}. \tag{D.37}$$

It is obvious that both $\big(\mathbf{S}+\mathcal{I}\big)^{-1}-\big(\mathcal{I}-\eta m\mathbf{H}_L^{\infty}\big)^{k}$ and $\big(\mathbf{S}+\mathcal{I}\big)^{-1}-\frac{1}{n^{C_0}}\cdot\mathcal{I}$ are positive definite matrices due to (D.37), and it is also easy to see that they are exchangeable. By using the second item of Lemma (D.4.12), we have

$$\begin{aligned}M_k&=\big[\big(\mathbf{I}-\eta m\mathbf{H}_L^{\infty}\big)^{k}-\big(\mathbf{S}+\mathcal{I}\big)^{-1}\big]\big(\mathbf{S}+\mathcal{I}\big)\big[\big(\mathbf{I}-\eta m\mathbf{H}_L^{\infty}\big)^{k}-\big(\mathbf{S}+\mathcal{I}\big)^{-1}\big]+\mathcal{I}-\big(\mathbf{S}+\mathcal{I}\big)^{-1}\\&\succeq\big[\big(\mathbf{S}+\mathcal{I}\big)^{-1}-\frac{1}{n^{C_0}}\cdot\mathcal{I}\big]\big(\mathbf{S}+\mathcal{I}\big)\big[\big(\mathbf{S}+\mathcal{I}\big)^{-1}-\frac{1}{n^{C_0}}\cdot\mathcal{I}\big]+\mathcal{I}-\big(\mathbf{S}+\mathcal{I}\big)^{-1}\\&=\frac{1}{n^{2C_0}}\mathbf{S}+\left(1-\frac{1}{n^{C_0}}\right)^{2}\cdot\mathcal{I}.\end{aligned}$$

Then, we have $\mathbb{E}_{\varepsilon}\left\|\Delta_{11}\right\|_2^2\succeq\frac{1}{n^{2C_0}}\mathcal{A}+\left(1-\frac{1}{n^{C_0}}\right)^{2}\mathcal{B}\succeq c_0\mathcal{B}$, where $c_0\in(0,1)$ is a constant and

$$\mathcal{A}=\int_{x\in\mathcal{S}^{d-1}}\big[\mathbf{Ker}(x,\mathbf{X})\big(\mathbf{H}_L^{\infty}\big)^{-1}\mathbf{y}^*\big]^2dx,\quad\text{and}\quad\mathcal{B}=\int_{x\in\mathcal{S}^{d-1}}\big[\mathbf{Ker}(x,\mathbf{X})\big(\mathbf{H}_L^{\infty}\big)^{-1}\big]^2dx. \tag{D.38}$$

By triangle inequality with the decomposition (D.32) and the bound on $\mathbb{E}_{\varepsilon}\left\|\Theta\right\|_2^2$ in (D.34), we have:

$$\begin{aligned}\mathbb{E}_{\varepsilon}\left\|\widehat{f}_{\mathbf{W}^{(k)}}-f^*\right\|_2^2&=\mathbb{E}_{\varepsilon}\left\|\Delta_{11}+\Theta\right\|_2^2\\&\geq\frac{1}{2}\mathbb{E}_{\varepsilon}\left\|\Delta_{11}\right\|_2-\mathbb{E}_{\varepsilon}\left\|\Theta\right\|_2^2\\&\geq\frac{c_0}{2}\mathcal{B}-\mathcal{O}\bigg(\frac{1}{n}\bigg)-\mathcal{O}\bigg(\frac{n^2\omega^3}{\lambda_\infty^2\delta^2}\bigg)-\tilde{\mathcal{O}}\bigg(\frac{1}{m^{1/3}}\text{poly}\bigg(\omega,n,L,\frac{1}{\lambda_\infty},\frac{1}{\delta}\bigg)\bigg).\end{aligned} \tag{D.39}$$

For the third term in (D.39), we can choose $\omega\leq C_2\big(\frac{\lambda_\infty\delta}{n}\big)^{2/3}$ for some constant $C_2>$

0 such that the term can be bounded by $\frac{c_0}{8}\left\|\mathbf{Ker}(\cdot,\mathbf{X})\big(\mathbf{H}_L^{\infty}\big)^{-1}\right\|_2^2$. Similarly, the width m can be chosen large enough such that the fourth term in (D.39) is upper-bounded by $\frac{c_0}{8}\|\mathbf{Ker}(\cdot,\mathbf{X})\big(\mathbf{H}_L^{\infty}\big)^{-1}\|_2^2$. Using the above choices of k, ω, and m, we can further bound (D.39):

$$\mathbb{E}_{\varepsilon}\left\|f_{\mathbf{W}(k)}-f^*\right\|_2^2 \geq \frac{c_0}{4}\left\|\mathbf{Ker}(\cdot,\mathbf{X})\big(\mathbf{H}_L^{\infty}\big)^{-1}\right\|_2^2-\mathcal{O}\left(\frac{1}{n}\right). \tag{D.40}$$

Note that $\mathbb{E}_{\varepsilon}\|\widehat{f}_{\infty}-g^*\|_2^2=\|\mathbf{Ker}(\cdot,\mathbf{X})\big(\mathbf{H}_L^{\infty}\big)^{-1}\|_2^2$ where $g^*:=0$ and $\widehat{f}_{\infty}$ denotes the noise interpolator. Then, by Theorem 4.2. of [142], we know that $\mathbb{E}_{\varepsilon}\|\widehat{f}_{\infty}-g^*\|_2^2\geq c_1$ for some constant $c_1>0$. Then, we can take n large enough such that the term $\mathcal{O}\big(\frac{1}{n}\big)$ is upper-bounded by $\frac{c_0c_1}{8}$, and finish the proof.

<u>Case 2.</u> When k is small, the L_2 risk is bounded away from zero by some constant.

Recall the definition of Δ_{11} in the decomposition (D.25),

$$\begin{aligned}\Delta_{11} &:= \mathbf{Ker}(x,\mathbf{X})G_k\big[\mathbf{y}^*+\varepsilon\big]-\mathbf{Ker}(x,\mathbf{X})\mathbf{H}_L^{\infty}\mathbf{y}^* \\ &:= \Delta_{11}^*-\mathbf{Ker}(x,\mathbf{X})\mathbf{H}_L^{\infty}\mathbf{y}^*. \end{aligned} \tag{D.41}$$

We denote the eigen-decomposition of the matrix $\mathbf{H}_L^{\infty}:=\sum_{i=1}^n\lambda_i\mathbf{v_i}\mathbf{v_i}^{\top}$, then we can easily see a following:

$$G_k=\eta m\sum_{j=0}^{k-1}\left(\sum_{i=1}^n(1-\eta m\lambda_i)^j\mathbf{v_i}\mathbf{v_i}^{\top}\right)\preceq\eta m\sum_{j=0}^{k-1}\sum_{i=1}^n\mathbf{v_i}\mathbf{v_i}^{\top}\preceq\eta mk\cdot\mathcal{I}.$$

By using the above inequality, we have

$$\begin{aligned}\mathbb{E}_{\varepsilon}\left\|\Delta_{11}^*\right\|_2^2 &= \int_{x\in\mathcal{S}^{d-1}}\mathbf{Ker}(x,\mathbf{X})G_k\big(\mathbf{S}+\mathcal{I}\big)G_k\mathbf{Ker}(\mathbf{X},x)dx \\ &\leq \eta^2m^2k^2\left(\int_{x\in\mathcal{S}^{d-1}}\big[\mathbf{Ker}(x,\mathbf{X})\mathbf{y}^*\big]^2dx+\|\mathbf{Ker}(\cdot,\mathbf{X})\|_2^2\right)=\mathcal{O}\Big(\eta^2m^2k^2\omega^2n^2L^2\Big).\end{aligned}$$

Recall the decompositions (D.21) and (D.25), then we have:

$$\begin{aligned}\mathbb{E}_\varepsilon \left\| \widehat{f}_{\mathbf{W}^{(k)}} - f^* \right\|_2^2 &= \mathbb{E}_\varepsilon \left\| \Delta_{11}^* + \Theta - \mathbf{Ker}(\cdot, \mathbf{X}) \mathbf{H}_L^\infty \mathbf{y}^* \right\|_2^2 \\ &\geq \frac{1}{2} \left\| \mathbf{Ker}(\cdot, \mathbf{X}) \mathbf{H}_L^\infty \mathbf{y}^* \right\|_2^2 - \mathbb{E}_\varepsilon \left\| \Delta_{11}^* + \Theta \right\|_2^2 \\ &\geq \frac{1}{2} \left\| \mathbf{Ker}(\cdot, \mathbf{X}) \mathbf{H}_L^\infty \mathbf{y}^* \right\|_2^2 - 2\mathbb{E}_\varepsilon \left\| \Delta_{11}^* \right\|_2^2 - 2\mathbb{E}_\varepsilon \left\| \Theta \right\|_2^2 \\ &\geq \frac{1}{2} \left\| \mathbf{Ker}(\cdot, \mathbf{X}) \mathbf{H}_L^\infty \mathbf{y}^* \right\|_2^2 - \mathcal{O}\left(\eta^2 m^2 k^2 \omega^2 n^2 L^2 \right) - \mathcal{O}\left(\frac{1}{n} \right) \\ &\qquad - \mathcal{O}\left(\frac{n^2 \omega^3}{\lambda_\infty^2 \delta^2} \right) - \tilde{\mathcal{O}}\left(\frac{1}{m^{1/3}} \mathrm{poly}\left(\omega, n, L, \frac{1}{\lambda_\infty}, \frac{1}{\delta} \right) \right). \end{aligned} \tag{D.42}$$

For some constant $C_1^{'} > 0$, let $k \leq C_1^{'} \left(\frac{1}{\eta m n \omega L} \right)$ such that the second term in the bound (D.42) can be bounded by $\frac{1}{8} \| \mathbf{Ker}(\cdot, \mathbf{X}) \left(\mathbf{H}_L^\infty \right)^{-1} \mathbf{y}^* \|_2^2$. For the fourth term in (D.42), we can choose $\omega \leq C_2^{'} \left(\frac{\lambda_\infty \delta}{n} \right)^{2/3}$ for some constant $C_2^{'} > 0$ such that the term can be bounded by $\frac{1}{8}$ $\left\| \mathbf{Ker}(\cdot, \mathbf{X}) \left(\mathbf{H}_L^\infty \right)^{-1} \mathbf{y}^* \right\|_2^2$. Similarly, the width m can be chosen large enough such that the fifth term in (D.42) is upper-bounded by $\frac{1}{8} \| \mathbf{Ker}(\cdot, \mathbf{X}) \left(\mathbf{H}_L^\infty \right)^{-1} \mathbf{y}^* \|_2^2$. Using the above choices of k, ω, and m, we can further bound (D.42):

$$\mathbb{E}_\varepsilon \left\| f_{\mathbf{W}^{(k)}} - f^* \right\|_2^2 \geq \frac{1}{4} \left\| \mathbf{Ker}(\cdot, \mathbf{X}) \left(\mathbf{H}_L^\infty \right)^{-1} \mathbf{y}^* \right\|_2^2 - \mathcal{O}\left(\frac{1}{n} \right) \geq C_3^{'} \left\| f_\rho^\star \right\|_2^2 - \mathcal{O}\left(\frac{1}{n} \right). \tag{D.43}$$

In the second inequality, we used (D.31) with triangle inequality. In the third inequality, we can take n large enough such that the term $\mathcal{O}\left(\frac{1}{n} \right)$ is upper-bounded by $\frac{C_3^{'}}{2} \left\| f_\rho^\star \right\|_2^2$. Lastly, by using the assumption that $f_\rho^\star$ is a square-integrable function, we finish the proof.

D.7 Proof of Theorem 4.3.10-Training error

Denote $u_{D,i}(k) = f_{\mathbf{W}_D^{(k)}}(\mathbf{x}_i)$ and let $\mathbf{u}_D(k) = \big[u_{1,D}(k), \ldots, u_{n,D}(k)\big]^\top$. In order to analyze the training error of ℓ_2-regularized estimator, $\|\mathbf{u}_D(k) - y\|_2^2$, we decompose the term as follows:

$$
\begin{aligned}
&\|\mathbf{u}_D(k+1) - y\|_2^2 && \text{(D.44)}\\
&= \|\mathbf{u}_D(k+1) - (1-\eta_2\mu L)\mathbf{u}_D(k)\|_2^2 + \|(1-\eta_2\mu L)\mathbf{u}_D(k) - y\|_2^2 \\
&\qquad - 2\big(y - (1-\eta_2\mu L)\mathbf{u}_D(k)\big)^\top \big(\mathbf{u}_D(k+1) - (1-\eta_2\mu L)\mathbf{u}_D(k)\big) && \text{(D.45)}
\end{aligned}
$$

Equipped with this decomposition, the proof consists of the following steps:

1. We decompose the decayed prediction difference $\mathbf{u}_D(k+1) - (1-\eta_2\mu L)\mathbf{u}_D(k)$ into two terms. We note that the first term is related with a gram matrix $\mathbf{H}_D(k)$ and denote a second term as $\mathbf{I}_D^{(k)}$.

2. The term $\mathbf{I}_D^{(k)}$ can be further decomposed into three terms, where we denote them as $\mathbf{I}_{2,D}^{(k)}$, $\mathbf{I}_{3,D}^{(k)}$ and $\mathbf{I}_{5,D}^{(k)}$. The crux for controlling the ℓ_2-norm of the above three terms is to utilize the results from the Appendix $D.4$. The applications of Lemmas in the Appendix $D.4$ is possible, since we can inductively guarantee that $\|W_{D,\ell}^{(k)} - W_{D,\ell}^{(0)}\|_2$ is sufficiently small enough for large enough m.

3. Given the decomposition (D.44), we further decompose it into four terms as follows:

$$
\begin{aligned}
(D.44) = &\underbrace{\|(1-\eta_2\mu L)\mathbf{u}_D(k) - y\|_2^2}_{:=\mathbf{T}_1} + \underbrace{\|\mathbf{u}_D(k+1) - (1-\eta_2\mu L)\mathbf{u}_D(k)\|_2^2}_{:=\mathbf{T}_2} \\
&+ \underbrace{2m\eta_1\big(y - (1-\eta_2\mu L)\mathbf{u}_D(k)\big)^\top \mathbf{H}_D(k)\big(\mathbf{u}_D(k) - y\big)}_{:=\mathbf{T}_3} \\
&\underbrace{-2\big(y - (1-\eta_2\mu L)\mathbf{u}_D(k)\big)^\top \mathbf{I}_D^{(k)}}_{:=\mathbf{T}_4}. \qquad \text{(D.46)}
\end{aligned}
$$

In this step, we obtain the upper-bound of $\|\mathbf{T}_i\|_2$ for $i = 1, 2, 3, 4$ obtained in Step 4.

4. We combine the upper-bounds of $\|\mathbf{T}_i\|_2$ for $i = 1, 2, 3, 4$ in step 3 and obtain the bound on $\|\mathbf{u}_D(k+1) - y\|_2^2$ with respect to $\|\mathbf{u}_D(k) - y\|_2^2$ and $\|y\|_2$.

5. Lastly, we inductively show that the weights generated from regularized gradient descent stay within a perturbation region $\mathcal{B}\big(\mathcal{W}^{(0)}, \tau\big)$, irrespective with the number of iterations of algorithm.

We start the proof by analyzing the term $\mathbf{u}(k+1) - (1 - \eta_2\mu L)\mathbf{u}(k)$.

Step 1. ***Dynamics of*** $\mathbf{u_D}(\mathbf{k+1}) - (1 - \eta_2\mu L) \cdot \mathbf{u_D}(\mathbf{k})$**.** Recall $\big(\mathbf{\Sigma}_{D,\ell,i}^{(k)}\big)_{jj} = \mathbb{1}\big(\langle \mathbf{w}_{D,\ell,j}^{(k)}, \mathbf{x}_{D,\ell-1,i}^{(k)} \rangle \geq 0\big)$ and we introduce a diagonal matrix $\widetilde{\mathbf{\Sigma}}_{D,\ell,i}^{(k)}$, whose jth entry is defined as follows:

$$\big(\widetilde{\mathbf{\Sigma}}_{D,\ell,i}^{(k)}\big)_{jj} = \big(\mathbf{\Sigma}_{D,\ell,i}^{(k+1)} - \mathbf{\Sigma}_{D,\ell,i}^{(k)}\big)_{jj} \cdot \frac{\langle \mathbf{w}_{D,\ell,j}^{(k+1)}, \mathbf{x}_{D,\ell-1,i}^{(k+1)} \rangle}{\langle \mathbf{w}_{D,\ell,j}^{(k+1)}, \mathbf{x}_{D,\ell-1,i}^{(k+1)} \rangle - \langle \mathbf{w}_{D,\ell,j}^{(k)}, \mathbf{x}_{D,\ell-1,i}^{(k)} \rangle}.$$

With this notation, the difference $\mathbf{x}_{D,L,i}^{(k+1)} - \mathbf{x}_{D,L,i}^{(k)}$ can be rewritten via the recursive applications of $\widetilde{\mathbf{\Sigma}}_{D,\ell,i}^{(k)}$: Then, we introduce following notations :

$$\begin{aligned}
\mathbf{D}_{D,\ell,i}^{(k)} &= \bigg(\prod_{r=\ell+1}^{L} \mathbf{\Sigma}_{D,r,i}^{(k)} \mathbf{W}_{D,r}^{(k)}\bigg) \mathbf{\Sigma}_{D,\ell,i}^{(k)}, \\
\widetilde{\mathbf{D}}_{D,\ell,i}^{(k)} &= \bigg(\prod_{r=\ell+1}^{L} \big(\mathbf{\Sigma}_{D,r,i}^{(k)} + \widetilde{\mathbf{\Sigma}}_{D,r,i}^{(k)}\big) \mathbf{W}_{D,r}^{(k+1)}\bigg) \big(\mathbf{\Sigma}_{D,\ell,i}^{(k)} + \widetilde{\mathbf{\Sigma}}_{D,\ell,i}^{(k)}\big).
\end{aligned}$$

Now, we can write $u_{D,i}(k+1) - u_{D,i}(k)$ by noting that $u_{D,i}(k) = \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \mathbf{x}_{D,L,i}^{(k)}$:

$$\begin{aligned}
&u_{D,i}(k+1) - u_{D,i}(k) \\
&\quad= \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \big(\mathbf{x}_{D,L,i}^{(k+1)} - \mathbf{x}_{D,L,i}^{(k)} \big) \\
&\quad= \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \widetilde{\mathbf{D}}_{D,\ell,i}^{(k)} \left(\mathbf{W}_{D,\ell}^{(k+1)} - \mathbf{W}_{D,\ell}^{(k)} \right) \mathbf{x}_{D,\ell-1,i}^{(k)} \\
&\quad= \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \widetilde{\mathbf{D}}_{D,\ell,i}^{(k)} \left(-\eta_1 \nabla_{\mathbf{W}_\ell} \big[\mathcal{L}_{\mathbf{S}} \big(\mathbf{W}_D^{(k)} \big) \big] - \eta_2 \mu \mathbf{W}_{D,\ell}^{(k)} + \eta_2 \mu \mathbf{W}_{D,\ell}^{(0)} \right) \mathbf{x}_{D,\ell-1,i}^{(k)} \\
&\quad= \underbrace{-\eta_1 \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \mathbf{D}_{D,\ell,i}^{(k)} \nabla_{\mathbf{W}_\ell} \big[\mathcal{L}_{\mathbf{S}} \big(\mathbf{W}_D^{(k)} \big) \big] \mathbf{x}_{D,\ell-1,i}^{(k)}}_{\mathbf{I}_{1,D,i}^{(k)}} \\
&\qquad\qquad \underbrace{-\eta_1 \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \left(\widetilde{\mathbf{D}}_{D,\ell,i}^{(k)} - \mathbf{D}_{D,\ell,i}^{(k)} \right) \nabla_{\mathbf{W}_\ell} \big[\mathcal{L}_{\mathbf{S}} \big(\mathbf{W}_D^{(k)} \big) \big] \mathbf{x}_{D,\ell-1,i}^{(k)}}_{\mathbf{I}_{2,D,i}^{(k)}} \\
&\qquad\qquad \underbrace{-\eta_2 \mu \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \left(\widetilde{\mathbf{D}}_{D,\ell,i}^{(k)} - \mathbf{D}_{D,\ell,i}^{(k)} \right) \left(\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)} \right) \mathbf{x}_{D,\ell-1,i}^{(k)}}_{\mathbf{I}_{3,D,i}^{(k)}} \\
&\qquad\qquad \underbrace{-\eta_2 \mu \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \mathbf{D}_{D,\ell,i}^{(k)} \mathbf{W}_{D,\ell}^{(k)} \mathbf{x}_{D,\ell-1,i}^{(k)}}_{\mathbf{I}_{4,D,i}^{(k)}} \\
&\qquad\qquad \underbrace{+\eta_2 \mu \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \mathbf{D}_{D,\ell,i}^{(k)} \mathbf{W}_{D,\ell}^{(0)} \mathbf{x}_{D,\ell-1,i}^{(k)}}_{\mathbf{I}_{5,D,i}^{(k)}} \qquad \text{(D.47)}
\end{aligned}$$

where in the second equality, we used the recursive relation (D.8), and in the third equality, modified GD update rule (4.6) is applied.

Furthermore, $\mathbf{I}^{(k)}_{1,D,i}$ can be rewritten as follows:

$$\begin{aligned}
\mathbf{I}^{(k)}_{1,D,i} &= -\eta_1\sqrt{m}\cdot\mathbf{v}^{\mathrm{T}}\sum_{\ell=1}^{L}\mathbf{D}^{(k)}_{D,\ell,i}\sum_{j=1}^{n}\big(u_{D,j}(k)-y_j\big)\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}_D}(\mathbf{x}_j)\big]\mathbf{x}^{(k)}_{D,\ell-1,i}\\
&= -\eta_1\cdot\sum_{j=1}^{n}\big(u_{D,j}(k)-y_j\big)\cdot\left(\sqrt{m}\sum_{\ell=1}^{L}\mathbf{v}^{\mathrm{T}}\mathbf{D}^{(k)}_{D,\ell,i}\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}_D}(\mathbf{x}_j)\big]\mathbf{x}^{(k)}_{D,\ell-1,i}\right)\\
&= -m\eta_1\cdot\sum_{j=1}^{n}\big(u_{D,j}(k)-y_j\big)\cdot\frac{1}{m}\sum_{\ell=1}^{L}\left\langle\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}_D}(\mathbf{x}_i)\big],\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}^{(k)}_D}(\mathbf{x}_j)\big]\right\rangle_{\mathrm{Tr}}\\
&= -m\eta_1\cdot\sum_{j=1}^{n}\big(u_{D,j}(k)-y_j\big)\cdot\mathbf{H}_{D,i,j}(k). \qquad\qquad \text{(D.48)}
\end{aligned}$$

With $\mathbf{I}^{(k)}_{4,i}=(-\eta_2\mu L)\cdot u_{D,i}(k)$ and (D.48), we can rewrite (D.47) as follows:

$$u_{D,i}(k+1)-(1-\eta_2\mu L)u_{D,i}(k) = -m\eta_1\cdot\sum_{j=1}^{n}\big(u_{D,j}(k)-y_j\big)\cdot\mathbf{H}_{D,i,j}(k)+\mathbf{I}^{(k)}_{2,D,i}+\mathbf{I}^{(k)}_{3,D,i}+\mathbf{I}^{(k)}_{5,D,i}. \tag{D.49}$$

Step 2. Control of the size $\left\|\mathbf{I}^{(k)}_{D}\right\|_2$.

Let $\mathbf{I}^{(k)}_{D}=[\mathbf{I}^{(k)}_{2,D,1}+\mathbf{I}^{(k)}_{3,D,1}+\mathbf{I}^{(k)}_{5,D,1},\ldots,\mathbf{I}^{(k)}_{2,D,n}+\mathbf{I}^{(k)}_{3,D,n}+\mathbf{I}^{(k)}_{5,D,1}]^{\top}$. Now, we control the bound on the $\left\|\mathbf{I}^{(k)}_{D}\right\|_2^2$. Recall that in Eq. (E.16), we have

$$\left\|\mathbf{I}^{(k)}_{2,D}\right\|_2\le\mathcal{O}\bigg(\eta_1 nL^3\tau^{1/3}\omega m\sqrt{\log(m)}\bigg)\left\|\mathbf{u}_D(k)-y\right\|_2. \tag{D.50}$$

Similarly, $\left\|\mathbf{I}^{(k)}_{3,D}\right\|_2$ can be bounded:

$$\begin{aligned}
\left\|\mathbf{I}^{(k)}_{3,D}\right\|_2 &\le\sum_{i=1}^{n}\left|\mathbf{I}^{(k)}_{3,D,i}\right|\le\eta_2\mu\sqrt{m}\cdot\sum_{i=1}^{n}\left[\sum_{\ell=1}^{L}\underbrace{\left\|\mathbf{v}^{\top}\left(\widetilde{\mathbf{D}}^{(k)}_{D,\ell,i}-\mathbf{D}^{(k)}_{D,\ell,i}\right)\right\|_2}_{\le\mathcal{O}\left(L^2\tau^{1/3}\sqrt{\omega\log(m)}\right)}\cdot\underbrace{\left\|\mathbf{W}^{(k)}_{D,\ell}-\mathbf{W}^{(0)}_{D,\ell}\right\|_2}_{\le\tau}\cdot\underbrace{\left\|\mathbf{x}^{(k)}_{D,\ell-1,i}\right\|_2}_{\le\mathcal{O}(1)}\right]\\
&\le\mathcal{O}\bigg(\eta_2\mu nL^3\tau^{4/3}\sqrt{\omega m\log(m)}\bigg). \qquad\qquad \text{(D.51)}
\end{aligned}$$

Lastly $\left\|\mathbf{I}_{5,D}^{(k)}\right\|_2$ can be bounded:

$$
\begin{aligned}
&\left\|\mathbf{I}_{5,D}^{(k)}\right\|_2 \leq \sum_{i=1}^{n} \left|\mathbf{I}_{5,D,i}^{(k)}\right| \\
&\leq \sum_{i=1}^{n} \left| \eta_2 \mu \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \mathbf{D}_{D,\ell,i}^{(k)} \mathbf{W}_{D,\ell}^{(k)} \mathbf{x}_{D,\ell-1,i}^{(k)} \right| + \sum_{i=1}^{n} \left| \eta_2 \mu \sqrt{m} \cdot \mathbf{v}^{\mathrm{T}} \sum_{\ell=1}^{L} \mathbf{D}_{D,\ell,i}^{(k)} \left(\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)} \right) \mathbf{x}_{D,\ell-1,i}^{(k)} \right| \\
&\leq \eta_2 \mu L \cdot \sum_{i=1}^{n} |\mathbf{u}_{i,D}(k)| + \eta_2 \mu \sqrt{m} \cdot \sum_{i=1}^{n} \left[\sum_{\ell=1}^{L} \underbrace{\|\mathbf{v}\|_2}_{\leq \mathcal{O}(\sqrt{\omega})} \cdot \underbrace{\left\|\mathbf{D}_{D,\ell,i}^{(k)}\right\|_2}_{\leq \mathcal{O}(\sqrt{L})} \cdot \underbrace{\left\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\right\|_2}_{\leq \tau} \cdot \underbrace{\left\|\mathbf{x}_{D,\ell-1,i}^{(k)}\right\|_2}_{\leq \mathcal{O}(1)} \right] \\
&\leq \mathcal{O}\left(\eta_2 \mu n L \sqrt{\omega \log(L/\delta)} \right) + \mathcal{O}\left(\eta_2 \mu n L^{3/2} \tau \sqrt{m\omega} \right), \qquad \text{(D.52)}
\end{aligned}
$$

where in the last inequality, we employed the same logic used in (D.28) with the Lemma D.4.2 to obtain the upper-bound on the $|\mathbf{u}_{i,D}(k)|$. We set the orders of the parameters μ, η_1, η_2, τ, and ω as follows:

$$
\begin{aligned}
\mu = \Theta\left(n^{\frac{d-1}{2d-1}} \right), \quad \eta_1 = \Theta\left(\frac{1}{m} n^{-\frac{3d-2}{2d-1}} \right), \quad \eta_2 = \Theta\left(\frac{1}{L} n^{-\frac{3d-2}{2d-1}} \right), \\
\tau = \mathcal{O}\left(\frac{L\sqrt{\omega}}{\sqrt{m}\delta} n^{\frac{d}{2d-1}} \right), \quad \omega = \mathcal{O}\left(\frac{1}{L^{3/2}} n^{-\frac{5d-2}{2d-1}} \right). \qquad \text{(D.53)}
\end{aligned}
$$

Plugging the choices of parameters (D.53) with sufficiently large m in (D.50), (D.51) and (D.52) yields

$$
\left\|\mathbf{I}_{D}^{(k)}\right\|_2 \leq \mathcal{O}\left(L^{37/12} n^{-\frac{9d-8}{12d-6}} \frac{\sqrt{\log(m)}}{m^{1/6}\delta^{1/3}} \right) \cdot \|\mathbf{u}_D(k) - y\|_2 + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n^2} \right). \qquad \text{(D.54)}
$$

Step 3. ***Upper-bound of*** $\|\mathbf{T}_i\|_2$ ***on*** $i = 1, 2, 3, 4$.

First, we work on getting the upper-bound on $\lambda_{\max}\big(\mathbf{H}_D(k)\big)$. By the Gershgorin's circle theorem [210], we know the maximum eigenvalue of symmetric positive semi-definite matrix is upper-bounded by the maximum absolute column sum of the matrix. Using this fact, we can bound the $\lambda_{\max}\big(\mathbf{H}_D(k)\big)$ as :

$$\begin{aligned}
\lambda_{\max}\big(\mathbf{H}_D(k)\big) &\leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |\mathbf{H}_{D,i,j}(k)| \\
&\leq \max_{i=1,\dots,n} \sum_{j=1}^{n} \left| \frac{1}{m} \sum_{\ell=1}^{L} \left\langle \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{x}_i)\big], \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{x}_j)\big] \right\rangle_{\text{Tr}} \right| \\
&\leq \max_{i=1,\dots,n} \sum_{j=1}^{n} \frac{1}{m} \sum_{\ell=1}^{L} \underbrace{\left\| \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{x}_i)\big] \right\|_F}_{\leq \mathcal{O}(\sqrt{m\omega})} \underbrace{\left\| \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{x}_j)\big] \right\|_F}_{\leq \mathcal{O}(\sqrt{m\omega})} \\
&\leq \mathcal{O}\big(nL\omega\big).
\end{aligned} \tag{D.55}$$

Recall the decomposition (D.46). Our goal is to obtain the upper-bound on $\mathbf{T}_i$ for $i = 1, 2, 3, 4$.

Control on $\mathbf{T}_1$. By using the inequality $2\eta_2\mu L(1-\eta_2\mu L)y^\top\big(y - \mathbf{u}_D(k)\big) \leq \eta_2\mu L \|y\|_2^2 + \eta_2\mu L(1-\eta_2\mu L)^2 \|y - \mathbf{u}_D(k)\|_2^2$, we have

$$\begin{aligned}
&\|y - (1-\eta_2\mu L)\mathbf{u}_D(k)\|_2^2 \\
&= \big\|(1-\eta_2\mu L)\big(y - \mathbf{u}_D(k)\big) + \eta_2\mu L y\big\|_2^2 \\
&= (1-\eta_2\mu L)^2 \|y - \mathbf{u}_D(k)\|_2^2 + \eta_2^2\mu^2L^2 \|y\|_2^2 \\
&\qquad + 2\eta_2\mu L(1-\eta_2\mu L)y^\top\big(y - \mathbf{u}_D(k)\big) \\
&\leq (\eta_2\mu L + \eta_2^2\mu^2L^2)\|y\|_2^2 + \big(1+\eta_2\mu L\big)\big(1-\eta_2\mu L\big)^2 \|y - \mathbf{u}_D(k)\|_2^2.
\end{aligned} \tag{D.56}$$

Control on T_2**.** Recall the equality (D.49). Then, through applications of the Young's inequality $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ for $a, b \in \mathbb{R}^n$, we have

$$\begin{aligned}\|\mathbf{u}_D(k+1) - (1-\eta_2\mu L)\mathbf{u}_D(k)\|_2^2 &= \left\| -m\eta_1 \cdot \mathbf{H}_D(k)\big(\mathbf{u}_D(k) - y\big) + \mathbf{I}_D^{(k)} \right\|_2^2 \\ &\le 2m^2\eta_1^2\lambda_{\text{max}}\big(\mathbf{H}_D(k)\big)^2 \|y - \mathbf{u}_D(k)\|_2^2 + 2\left\|\mathbf{I}_D^{(k)}\right\|_2^2 .\end{aligned} \tag{D.57}$$

Similarly with $\mathbf{T}_1$ and $\mathbf{T}_2$, we can control $\mathbf{T}_3$ and $\mathbf{T}_4$ as follows:

Control on T_3**.** Recall $\mathbf{H}_D(k)$ is a Gram matrix by definition. Then, by using the fact $\lambda_{\text{min}}\big(\mathbf{H}_D(k)\big) \ge 0$ and Cauchy-Schwarz inequality, we have

$$\begin{aligned}&2m\eta_1\big(y - (1-\eta_2\mu L)\mathbf{u}_D(k)\big)^\top \mathbf{H}_D(k)\big(\mathbf{u}_D(k) - y\big) \\ &\quad = -2m\eta_1(1-\eta_2\mu L)\big(y - \mathbf{u}_D(k)\big)^\top \mathbf{H}_D(k)\big(y - \mathbf{u}_D(k)\big) \\ &\qquad + \big(2m\eta_1\eta_2\mu L\big) \cdot y^\top \mathbf{H}_D(k)\big(\mathbf{u}_D(k) - y\big) \\ &\quad \le \big(2m\eta_1\eta_2\mu L\big) \cdot \lambda_{\text{max}}\big(\mathbf{H}_D(k)\big) \|y - \mathbf{u}_D(k)\|_2^2 \\ &\qquad + \big(2m\eta_1\eta_2\mu L\big) \cdot \big(\lambda_{\text{max}}\big(\mathbf{H}_D(k)\big) \|y\|_2 \|y - \mathbf{u}_D(k)\|_2\big) \\ &\qquad - 2m\eta_1\lambda_{\text{min}}\big(\mathbf{H}_D(k)\big) \|y - \mathbf{u}_D(k)\|_2^2 \\ &\quad = \big(4m\eta_1\eta_2\mu L\big) \cdot \lambda_{\text{max}}\big(\mathbf{H}_D(k)\big) \|y - \mathbf{u}_D(k)\|_2^2 + \big(4m\eta_1\eta_2\mu L\big) \cdot \lambda_{\text{max}}\big(\mathbf{H}_D(k)\big) \|y\|_2^2 .\end{aligned} \tag{D.58}$$

Control on T_4**.** By a simple Cauchy-Schwarz and Young's inequality, we have

$$\begin{aligned}&-2\big(y - (1-\eta_2\mu L)\mathbf{u}_D(k)\big)^\top \mathbf{I}_D^{(k)} \\ &\quad = -2(1-\eta_2\mu L)\big(y - \mathbf{u}_D(k)\big)^\top \mathbf{I}_D^{(k)} + 2\eta_2\mu L \cdot y^\top \mathbf{I}_D(k) \\ &\quad \le 2\big(1-\eta_2\mu L\big) \|y - \mathbf{u}_D(k)\|_2 \left\|\mathbf{I}_D^{(k)}\right\|_2 + \eta_2\mu L \|y\|_2^2 + \eta_2\mu L \left\|\mathbf{I}_D^{(k)}\right\|_2^2\end{aligned} \tag{D.59}$$

Step 4. Upper-bound of the decomposition on training error (D.46).

Before getting the upper bound of the decomposition (D.46), we first work on obtaining the bound of (D.60). Set $\kappa = \mathcal{O}\big(\frac{1}{n^2}\big)$ and notice $\eta_2 \mu L = \mathcal{O}\big(\frac{1}{n}\big)$ by (D.53), then we have

$$\begin{aligned}
&2\left\|\mathbf{I}_D^{(k)}\right\|_2^2 + 2\big(1-\eta_2\mu L\big)\left\|y-\mathbf{u}_D(k)\right\|_2 \left\|\mathbf{I}_D^{(k)}\right\|_2 + \eta_2\mu L\left\|\mathbf{I}_D^{(k)}\right\|_2^2 \qquad \text{(D.60)}\\
&\quad \leq \left(2+\eta_2\mu L+\frac{1}{\kappa^2}\right)\left\|\mathbf{I}_D^{(k)}\right\|_2^2 + \kappa^2\big(1-\eta_2\mu L\big)^2\left\|y-\mathbf{u}_D(k)\right\|_2^2\\
&\quad = \frac{1}{\kappa^2}\cdot\left\|\mathbf{I}_D^{(k)}\right\|_2^2 + \kappa^2\big(1-\eta_2\mu L\big)^2\left\|y-\mathbf{u}_D(k)\right\|_2^2\\
&\quad \leq \left\{\frac{1}{\kappa^2}\cdot\mathcal{O}\left(L^{37/6}n^{-\frac{9d-8}{6d-3}}\frac{\log(m)}{m^{1/3}\delta^{2/3}}\right)+\kappa^2\big(1-\eta_2\mu L\big)^2\right\}\cdot\left\|y-\mathbf{u}_D(k)\right\|_2^2 + \frac{1}{\kappa^2}\cdot\mathcal{O}_{\mathbb{P}}\left(\frac{1}{n^4}\right)\\
&\quad \leq \big(\eta_2\mu L\big)^4\big(1-\eta_2\mu L\big)^2\cdot\left\|y-\mathbf{u}_D(k)\right\|_2^2 + \eta_2\mu L\cdot\|y\|_2^2, \qquad \text{(D.61)}
\end{aligned}$$

where in the second inequality, the Eq. (D.54) is used with $(a+b)^2 \leq 2a^2+2b^2$ for $a, b \in \mathbb{R}$, and in the last inequality, we used $\|y\|_2^2 = \mathcal{O}(n)$ and the sufficiently large m to control the order of the coefficient terms of $\|y-\mathbf{u}_D(k)\|_2^2$. Specifically, we choose $m \geq \Omega\big(L^{19}n^{20}\frac{\log^3(m)}{\delta^2}\big)$. Now, by combining the inequalities (D.56), (D.57), (D.58), (D.59), (D.55) and (D.61), we obtain the upper-bound on the decomposition (D.46);

$$\begin{aligned}
&\left\|\mathbf{u}_D(k+1)-y\right\|_2^2\\
&\leq \left\{3\eta_2\mu L+\eta_2^2\mu^2L^2+\mathcal{O}\left(\omega m n\eta_1\eta_2\mu L^2\right)\right\}\cdot\|y\|_2^2\\
&\qquad+\left\{\big(1+\eta_2\mu L+\eta_2^4\mu^4L^4\big)\big(1-\eta_2\mu L\big)^2+\mathcal{O}\left(\omega^2m^2n^2\eta_1^2L^2\right)+\mathcal{O}\left(\omega m n\eta_1\eta_2\mu L^2\right)\right\}\\
&\qquad\cdot\left\|y-\mathbf{u}_D(k)\right\|_2^2\\
&:= \mathcal{A}\cdot\|y\|_2^2+(1-\mathcal{B})\cdot\left\|y-\mathbf{u}_D(k)\right\|_2^2. \qquad \text{(D.62)}
\end{aligned}$$

With the order choices of μ, η_1 and η_2 as in (D.53), it is easy to see the leading terms of both $\mathcal{A}$ and $\mathcal{B}$ are same as $\eta_2 \mu L = o(\frac{1}{n})$. Then, by recursively applying the inequality (D.62), we can get the upper-bound on the training error.

$$\begin{aligned} \|y - \mathbf{u}_D(k+1)\|_2^2 &\leq \mathcal{A} \cdot \|y\|_2^2 + (1-\mathcal{B}) \cdot \|y - \mathbf{u}_D(k)\|_2^2 \\ &\leq \mathcal{A}\|y\|_2^2 \cdot \left(\sum_{j=0}^{k} (1-\mathcal{B})^j \right) + (1-\mathcal{B})^{k+1} \cdot \|y - \mathbf{u}_D(0)\|_2^2 \\ &\leq \frac{\mathcal{A}}{\mathcal{B}} \cdot \|y\|_2^2 + (1-\mathcal{B})^{k+1} \cdot \|y - \mathbf{u}_D(0)\|_2^2 \\ &\leq \mathcal{O}\big(n\big) + (1-\eta_2\mu L)^{k+1} \cdot \|y - \mathbf{u}_D(0)\|_2^2 . \end{aligned} \tag{D.63}$$

In the last inequality, we used $\frac{\mathcal{A}}{\mathcal{B}} = o(1)$, $\mathcal{B} \geq \eta_2 \mu L$ and $\|y\|_2^2 = \mathcal{O}(n)$.

Step 5. The order of the radius of perturbation region. It remains us to prove the radius of perturbation region τ has the order $\mathcal{O}_{\mathbb{P}}\left(\frac{L\sqrt{\omega}}{\sqrt{m}} n^{\frac{d}{2d-1}} \right)$. First, recall that the ℓ_2-regularized GD update rule is as:

$$\mathbf{W}_{D,\ell}^{(k)} = \big(1-\eta_2\mu\big)\mathbf{W}_{D,\ell}^{(k-1)} - \eta_1 \nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D^{(k-1)}\big)\big] + \eta_2\mu\mathbf{W}_{D,\ell}^{(0)}, \quad \forall 1 \leq \ell \leq L \quad \text{and} \quad \forall k \geq 1. \tag{D.64}$$

Similarly with the proof in the Theorem 2.4.1, we employ the induction process for the proof. The induction hypothesis is

$$\left\| \mathbf{W}_{D,\ell}^{(s)} - \mathbf{W}_{D,\ell}^{(0)} \right\|_2 \leq \mathcal{O}\left(\frac{\eta_1 n \sqrt{m\omega}}{\sqrt{\delta}\eta_2\mu} \right), \qquad \forall s \in [k+1]. \tag{D.65}$$

It is easy to see it holds for $s = 0$, and suppose it holds for $s = 0, 1, \ldots, k$, we consider

$k+1$. Using the update rule (D.64), we have

$$\begin{aligned}
\left\|\mathbf{W}_{D,\ell}^{(k+1)} - \mathbf{W}_{D,\ell}^{(k)}\right\|_2 &\leq \eta_2\mu\left\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\right\|_2 + \eta_1\left\|\nabla_{\mathbf{W}_\ell}\big[\mathcal{L}_{\mathbf{S}}\big(\mathbf{W}_D^{(k)}\big)\big]\right\|_2 \\
&= \eta_2\mu\left\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\right\|_2 + \eta_1\left\|\sum_{i=1}^{n}\Big(y_i - \mathbf{u}_{D,i}(k)\Big)\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k)}}(\mathbf{x}_i)\big]\right\|_2 \\
&\leq \mathcal{O}\bigg(\frac{\eta_1 n\sqrt{m\omega}}{\sqrt{\delta}}\bigg) + \mathcal{O}\Big(\eta_1\sqrt{nm\omega}\Big)\cdot\left\|y - \mathbf{u}_D(k)\right\|_2 \\
&\leq \mathcal{O}\bigg(\frac{\eta_1 n\sqrt{m\omega}}{\sqrt{\delta}}\bigg) + \mathcal{O}\Big(\eta_1\sqrt{nm\omega}\Big)\cdot\bigg\{\mathcal{O}\big(\sqrt{n}\big) + (1-\eta_2\mu L)^{\frac{k}{2}}\mathcal{O}\bigg(\sqrt{\frac{n}{\delta}}\bigg)\bigg\} \\
&\leq \mathcal{O}\bigg(\frac{\eta_1 n\sqrt{m\omega}}{\sqrt{\delta}\eta_2\mu}\bigg).
\end{aligned}$$

In the first inequality, we use the induction hypothesis for $s = k$, and Lemma D.4.4. In the second inequality, since the induction hypothesis holds for $s = 0, 1, \ldots, k$, we employ $\left\|y - \mathbf{u}_D(k)\right\|_2 \leq \mathcal{O}\big(\sqrt{n}\big) + (1-\eta_2\mu L)^{\frac{k}{2}}\left\|y - \mathbf{u}_D(0)\right\|_2$ with the Lemma D.4.9. In the last inequality, we use $\eta_2\mu < 1$. By triangle inequality, the induction holds for $s = k+1$. Plugging the proper choices of η_1, η_2 and μ as suggested in (D.53) to $\mathcal{O}\bigg(\frac{\eta_1 n\sqrt{m\omega}}{\sqrt{\delta}\eta_2\mu}\bigg)$ yields $\|\mathbf{W}_{D,\ell}^{(k)} - \mathbf{W}_{D,\ell}^{(0)}\|_2 \leq \mathcal{O}_{\mathbb{P}}\bigg(\frac{L\sqrt{\omega}}{\sqrt{m}}n^{\frac{d}{2d-1}}\bigg)$.

D.8 Proof of Theorem 4.3.10-Kernel ridge regressor approximation

We present a following proof sketch on the approximation of regularized DNN estimator to kernel ridge regressor.

1. The key idea for proving the second result in Theorem 3.8 is to write the distance between $\mathbf{u}_{i,D}(k)$ (where D is to denote the prediction is obtained from regularized GD rule) and kernel regressor $\mathbf{B} := \mathbf{H}_L^{\infty}\big(C\mu\cdot\mathcal{I} + \mathbf{H}_L^{\infty}\big)^{-1}\mathbf{y}$ in terms of NTK matrix $\mathbf{H}_L^{\infty}$, which is as follows:

$$\mathbf{u}_D(k) - \mathbf{B} = \Big(\big(1-\eta_2\mu L\big)\cdot\mathcal{I} - m\eta_1\mathbf{H}_L^{\infty}\Big)^k\big(\mathbf{u}_D(0) - \mathbf{B}\big) + \mathbf{e}_D(k).$$

Above equality describes how the regularized estimator evolves to fit the kernel regressor as iteration of algorithm goes by.

2. We can bound the ℓ_2-norm of residual term $\mathbf{e}_D(k)$ as $\mathcal{O}(1/n)$, and show that the ℓ_2 norm of the first term on the RHS of equation (4.3) decays at the rate $\mathcal{O}\big(\sqrt{n}\big(1-\eta_2\mu L\big)^k\big)$. Here the $\sqrt{n}$ comes from the bound $\|\mathbf{B}\|_2 \leq \mathcal{O}(\sqrt{n})$, since we know $\|\mathbf{u}(0)\|_2$ has $\mathcal{O}(\sqrt{n\omega})$ with small $\omega \leq 1$. This yields the claim.

Recall the equality (D.49). Then, we have

$$\begin{aligned}
&\mathbf{u}_D(k+1) - (1-\eta_2\mu L)\mathbf{u}_D(k) \\
&\qquad = -m\eta_1 \cdot \mathbf{H}_D(k)\big(\mathbf{u}_D(k) - \mathbf{y}\big) + \mathbf{I}_{2,D}^{(k)} + \mathbf{I}_{3,D}^{(k)} + \mathbf{I}_{5,D}^{(k)} \\
&\qquad = -m\eta_1 \cdot \mathbf{H}_L^{\infty}\big(\mathbf{u}_D(k) - \mathbf{y}\big) \\
&\qquad\qquad - m\eta_1 \cdot \big(\mathbf{H}_D(k) - \mathbf{H}_L^{\infty}\big)\big(\mathbf{u}_D(k) - \mathbf{y}\big) + \mathbf{I}_{2,D}^{(k)} + \mathbf{I}_{3,D}^{(k)} + \mathbf{I}_{5,D}^{(k)} \\
&\qquad = -m\eta_1 \cdot \mathbf{H}_L^{\infty}\big(\mathbf{u}_D(k) - \mathbf{y}\big) + \xi_D(k). \qquad\qquad \text{(D.66)}
\end{aligned}$$

With $\tau = \mathcal{O}\Bigg(\frac{L\sqrt{\omega}}{\sqrt{m\delta}} n^{\frac{d}{2d-1}}\Bigg)$, similarly with Lemma D.4.10 and a direct employment of the result from Lemma D.4.11, we can control the distance from $\mathbf{H}_D(k)$ to $\mathbf{H}_L^{\infty}$ under operator norm as follows:

$$\begin{aligned}
\|\mathbf{H}_D(k) - \mathbf{H}_L^{\infty}\|_2 &\leq \|\mathbf{H}_D(k) - \mathbf{H}(0)\|_2 + \|\mathbf{H}(0) - \mathbf{H}_L^{\infty}\|_2 \\
&\leq \mathcal{O}\Bigg(\omega^{7/6} L^{10/3} n^{\frac{7d-3}{6d-3}} \sqrt[6]{\frac{\log^3(m)}{m\delta^2}}\Bigg) + \mathcal{O}\Bigg(\omega L^{5/2} n \sqrt[4]{\frac{\log(nL/\delta)}{m}}\Bigg) \\
&\leq \mathcal{O}\Bigg(L^{19/12} n^{-\frac{21d-8}{12d-6}} \sqrt[6]{\frac{\log^3(m)}{m\delta^2}}\Bigg) + \mathcal{O}\Bigg(L n^{-\frac{18d-6}{12d-6}} \sqrt[4]{\frac{\log(nL/\delta)}{m}}\Bigg) \\
&\leq \mathcal{O}\Bigg(L^{19/12} n^{-\frac{21d-8}{12d-6}} \sqrt[6]{\frac{\log^3(m)}{m\delta^2}}\Bigg), \qquad\qquad \text{(D.67)}
\end{aligned}$$

where in the third inequality, $\omega = \mathcal{O}\big(\frac{1}{L^{3/2}} n^{-\frac{5d-2}{2d-1}}\big)$ is plugged-in. The last inequality holds with $d \geq 2$ with large enough n and the condition on width $m \geq \Omega\big(L^{19}n^{20}\frac{\log^3(m)}{\delta^2}\big)$. Then, the ℓ_2 norm of $\xi_D(k)$ can be bounded as:

$$\begin{aligned}
\|\xi_D(k)\|_2 &\leq m\eta_1 \cdot \|\mathbf{H}_L^{\infty} - \mathbf{H}_D(k)\|_2 \, \|\mathbf{u}_D(k) - y\|_2 + \left\|\mathbf{I}_D^{(k)}\right\|_2 \\
&\leq \mathcal{O}\left(L^{19/12} n^{-\frac{12d-5}{6d-3}} \frac{\sqrt{\log(m)}}{m^{1/6}\delta^{1/3}}\right) \cdot \underbrace{\|\mathbf{u}_D(k) - y\|_2}_{\leq \mathcal{O}(\sqrt{n/\delta})} + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n^2}\right) \\
&\leq \mathcal{O}\left(L^{19/12} n^{-\frac{18d-7}{12d-6}} \frac{\sqrt{\log(m)}}{m^{1/6}\delta^{5/6}}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n^2}\right) = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n^2}\right), \qquad \text{(D.68)}
\end{aligned}$$

where in the second inequality, we used (D.67) with $\eta_1 = \mathcal{O}\left(\frac{1}{m}n^{-\frac{3d-2}{2d-1}}\right)$ to control the first term and employed Eq. (D.54) to control the second term. In the last equality, we used $m \geq \Omega\big(L^{19}n^{20}\frac{\log^3(m)}{\delta^2}\big)$. Now, by setting $\mathbf{B} := \left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^{\infty}\right)^{-1}\mathbf{H}_L^{\infty}\mathbf{y}$, we can easily convert the equality (D.66) as follows: for $k \geq 1$,

$$\mathbf{u}_D(k) - \mathbf{B} = \Big(\big(1 - \eta_2\mu L\big) \cdot \mathcal{I} - m\eta_1\mathbf{H}_L^{\infty}\Big)\big(\mathbf{u}_D(k-1) - \mathbf{B}\big) + \xi_D(k-1). \qquad \text{(D.69)}$$

The recursive applications of the equality (D.69) yields

$$\begin{aligned}
\mathbf{u}_D(k) - \mathbf{B} &= \Big(\big(1 - \eta_2\mu L\big) \cdot \mathcal{I} - m\eta_1\mathbf{H}_L^{\infty}\Big)^k \big(\mathbf{u}_D(0) - \mathbf{B}\big) \\
&\qquad + \sum_{j=0}^{k} \Big(\big(1 - \eta_2\mu L\big) \cdot \mathcal{I} - m\eta_1\mathbf{H}_L^{\infty}\Big)^j \xi_D(k-j-1) \\
&= \Big(\big(1 - \eta_2\mu L\big) \cdot \mathcal{I} - m\eta_1\mathbf{H}_L^{\infty}\Big)^k \big(\mathbf{u}_D(0) - \mathbf{B}\big) + \mathbf{e}_D(k). \qquad \text{(D.70)}
\end{aligned}$$

Now, we bound the ℓ_2 norm of $\mathbf{e}_D(k)$ in (D.70):

$$\begin{aligned}
\|\mathbf{e}_D(k)\|_2 &= \left\| \sum_{j=0}^{k} \Big(\big(1-\eta_2\mu L\big)\cdot \mathcal{I} - m\eta_1 \mathbf{H}_L^\infty \Big)^j \xi_D(k-j-1) \right\|_2 \\
&\leq \sum_{j=0}^{k} \left\| \big(1-\eta_2\mu L\big)\cdot \mathcal{I} - m\eta_1 \mathbf{H}_L^\infty \right\|_2^j \|\xi_D(k-j-1)\|_2 \\
&\leq \sum_{j=0}^{k} \big(1-\eta_2\mu L\big)^j \|\xi_D(k-j-1)\|_2 = \mathcal{O}\bigg(\frac{1}{n}\bigg), \qquad \text{(D.71)}
\end{aligned}$$

in the last inequality, we used $\eta_2\mu L = \mathcal{O}\big(\frac{1}{n}\big)$ and Eq. (D.68). Now, we control the ℓ_2-norm of the first term in (D.70) as:

$$\begin{aligned}
\left\| \Big(\big(1-\eta_2\mu L\big)\cdot \mathcal{I} - m\eta_1 \mathbf{H}_L^\infty \Big)^k \big(\mathbf{u}_D(0) - \mathbf{B}\big) \right\|_2 &\leq \big(1-\eta_2\mu L\big)^k \|\mathbf{u}_D(0) - \mathbf{B}\|_2 \\
&\leq \mathcal{O}\bigg(\sqrt{n}\big(1-\eta_2\mu L\big)^k \bigg), \qquad \text{(D.72)}
\end{aligned}$$

where in the second inequality, we used $\|\mathbf{u}_D(0)\|_2 \leq \mathcal{O}(\sqrt{n\omega}/\delta)$ and the fact that

$$\|\mathbf{B}\|_2 \leq \left\| \bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \bigg)^{-1} \mathbf{H}_L^\infty \right\|_2 \cdot \|\mathbf{y}\|_2 \leq \mathcal{O}(\sqrt{n}).$$

By combining (D.71) and (E.4) and using a fact $(1-\eta_2\mu L)^k \leq \exp(-\eta_2\mu L k)$, we conclude that after $k \geq \Omega\big((\eta_2\mu L)^{-1}\log(n^{3/2})\big)$, the error $\|\mathbf{u}_D(k) - \mathbf{B}\|_2$ decays at the rate $\mathcal{O}\big(\frac{1}{n}\big)$.

D.9 Proof of Theorem 4.3.11

We begin the proof by decomposing the error $\widehat{f}_{\mathbf{W}_D^{(k)}}(x) - f^*(x)$ for any fixed $x \in$ **Unif**$(\mathcal{S}^{d-1})$ into two terms as follows:

$$\widehat{f}_{\mathbf{W}_D^{(k)}}(x) - f^*(x) = \underbrace{\big(\widehat{f}_{\mathbf{W}_D^{(k)}}(x) - g_\mu^*(x)\big)}_{\Delta_{D,1}} + \underbrace{\big(g_\mu^*(x) - f^*(x)\big)}_{\Delta_{D,2}}. \qquad \text{(D.73)}$$

Here, we devise a solution of kernel ridge regression $g_{\mu}^{*}(x)$ in the decomposition (D.73):

$$g_{\mu}^{*}(x) := \mathbf{Ker}(x, \mathbf{X})\big(C\mu \cdot \mathcal{I} + \mathbf{H}_{L}^{\infty}\big)^{-1}\mathbf{y},$$

for some constant $C > 0$. Specifically, in the proof to follow, we choose η_1 and η_2 such that $C = \frac{\eta_2 L}{\eta_1 m}$ for the theoretical convenience. Our goal is to show that all the terms $\|\Delta_{D,1}\|_2^2$, and $\|\Delta_{D,2}\|_2^2$ have the order either equal to or smaller than $\mathcal{O}\big(n^{-\frac{d}{2d-1}}\big)$ with the proper choices on m, μ, η_1 and η_2. Since the high-level proof idea is similar with that of Theorem 3.3.3, we omit the step-by-step proof sketch of Theorem 4.3.11. The most notable difference between the proof strategies of the two theorems is that the regularized DNN approximate the kernel ridge regressor of noisy data, whereas in Theorem 3.3.3, unregularized DNN approximate the interpolant based on noiseless data.

Step 1. ***Control on*** $\Delta_{D,2}$**.** First, note that there is a recent finding that the reproducing kernel Hilbert spaces induced from NTKs with any number of layers (i.e., $L \geq 1$) have the same set of functions, if kernels are defined on $\mathcal{S}^{d-1}$. See [157]. Along with this result, under the choice of model parameters as suggested in (D.53), we can apply exactly the same proof used in Theorem.3.2 in [142] for proving a following :

$$\|\Delta_{D,2}\|_2^2 := \big\|g_{\mu}^{*} - f^{*}\big\|_2^2 = \mathcal{O}_{\mathbb{P}}\bigg(n^{-\frac{d}{2d-1}}\bigg), \qquad \big\|g_{\mu}^{*}\big\|_{\mathcal{H}}^2 = \mathcal{O}_{\mathbb{P}}(1). \tag{D.74}$$

Step 2. ***Control on*** $\Delta_{D,1}$**.** For n data points $\big(\mathbf{x}_1, \dots, \mathbf{x}_n\big)$ and for the k^{th} updated parameter $\mathbf{W}_D^{(k)}$, denote:

$$\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{X})\big] = \Bigg[\mathbf{vec}\bigg(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{x}_1)\big]\bigg), \cdots, \mathbf{vec}\bigg(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{x}_n)\big]\bigg)\Bigg].$$

Note that when $\ell = 1$, $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{X})\big] \in \mathbb{R}^{md\times n}$ and when $\ell = 2, \dots, L$, $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(\mathbf{X})\big] \in \mathbb{R}^{m^2\times n}$.

With this notation, we can write the vectorized version of the update rule (D.64) as:

$$\mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(k)}\big) = \mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(0)}\big) - \eta_1 \sum_{j=0}^{k-1} \big(1-\eta_2\mu\big)^j \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(\mathbf{X})\big]\Big(\mathbf{u}_D(k-j-1) - y\Big),$$

$\forall 1 \leq \ell \leq L$ and $\forall k \geq 1$. Using the equality, we can get the decomposition :

$$\mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(k)}\big) = \underbrace{\mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(0)}\big)}_{:=\mathbf{E}_1} \underbrace{-\eta_1 \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{X})\big] \sum_{j=0}^{k-1} \big(1-\eta_2\mu\big)^j \Big(\mathbf{u}_D(k-j-1) - y\Big)}_{:=\mathbf{E}_2}$$
$$\underbrace{-\eta_1 \sum_{j=0}^{k-1} \big(1-\eta_2\mu\big)^j \Bigg[\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(\mathbf{X})\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{X})\big]\Bigg]\Big(\mathbf{u}_D(k-j-1) - y\Big).}_{:=\mathbf{E}_3} \tag{D.75}$$

Let $z_{D,k}(x) := \mathbf{vec}\big(\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D^{(k)}}(x)\big]\big)$, and note that $f_{\mathbf{W}_D^{(k)}}(x) = \langle z_{D,k}(x), \mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(k)}\big)\rangle$. Then, by the definition of $\Delta_{D,1}$ and the decomposition (D.75), we have

$$\begin{aligned}\Delta_{D,1} &= \frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_1+\mathbf{E}_2+\mathbf{E}_3\rangle - \mathbf{Ker}(x,\mathbf{X})\bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty\bigg)^{-1}\mathbf{y} \\ &= \frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_1\rangle + \frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_3\rangle \\ &\quad + \underbrace{\frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_2\rangle - \mathbf{Ker}(x,\mathbf{X})\bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty\bigg)^{-1}\mathbf{y}}_{:=\mathcal{C}}\end{aligned} \tag{D.76}$$

First, we focus on controlling the ℓ_2 bound on the first two terms in (D.76). Observe that the first term can be bounded as:

$$\left|\frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_1\rangle\right|^2 \leq \frac{1}{L}\sum_{\ell=1}^{L}\left|\langle z_{D,k}(x), \mathbf{E}_1\rangle\right|^2. \tag{D.77}$$

Recall that $\|z_{D,k}(x)\|_2 \leq \mathcal{O}\big(\sqrt{m\omega}\big)$ by Lemma D.4.4. Then, the random variable $z_{D,k}(x)^\top$

$\mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(0)}\big) \mid z_{D,k}(x)$ is simply a $\mathcal{N}\big(0, \mathcal{O}(\omega)\big)$ for $1 \leq \ell \leq L$. A straightforward application of Chernoff bound for normal random variable and taking union bound over the layer $1 \leq \ell \leq L$ yield that: with probability at least $1-\delta$,

$$\frac{1}{L}\sum_{\ell=1}^{L}\Big|z_{D,k}(x)^\top \mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(0)}\big)\Big|^2 \leq \mathcal{O}\Bigg(\omega \log\bigg(\frac{L}{\delta}\bigg)\Bigg). \tag{D.78}$$

The ℓ_2 norm of the second term in (D.76) can be similarly bounded as (D.77) in addition with the Cauchy-Schwarz inequality:

$$\Bigg|\frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_3\rangle\Bigg|^2 \leq \frac{1}{L}\sum_{\ell=1}^{L}|\langle z_{D,k}(x), \mathbf{E}_3\rangle|^2 \leq \frac{1}{L}\sum_{\ell=1}^{L}\|z_{D,k}(x)\|_2^2\,\|\mathbf{E}_3\|_2^2. \tag{D.79}$$

The $\|\mathbf{E}_3\|_2$ is bounded as :

$$\begin{aligned}
\|\mathbf{E}_3\|_2 &= \Bigg\|\eta_1\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\bigg[\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(\mathbf{X})\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{X})\big]\bigg]\Big(\mathbf{u}_D(k-j-1)-y\Big)\Bigg\|_2 \\
&\leq \eta_1\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\cdot\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(\mathbf{X})\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{X})\big]\big\|_2\,\|\mathbf{u}_D(k-j-1)-y\|_2 \\
&\leq \eta_1\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\cdot\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(\mathbf{X})\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{X})\big]\big\|_F\,\|\mathbf{u}_D(k-j-1)-y\|_2 \\
&= \eta_1\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\cdot\sqrt{\sum_{i=1}^{n}\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(\mathbf{x}_i)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{x}_i)\big]\big\|_F^2}\,\|\mathbf{u}_D(k-j-1)-y\|_2 \\
&\leq \eta_1\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\cdot\sqrt{2\sum_{i=1}^{n}\big\|\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(\mathbf{x}_i)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{x}_i)\big]\big\|_2^2}\,\|\mathbf{u}_D(k-j-1)-y\|_2 \\
&\leq \frac{\eta_1}{\eta_2\mu}\cdot\mathcal{O}\bigg(\tau^{1/3}L^2\sqrt{\omega m n\log(m)}\bigg)\cdot\mathcal{O}\big(\sqrt{n}\big) \leq \mathcal{O}\bigg(\frac{L^{10/3}\omega^{1/6}}{m^{2/3}\delta^{1/3}}n^{\frac{4d}{6d-3}}\sqrt{\log(m)}\bigg).
\end{aligned} \tag{D.80}$$

In the first, second and third inequalities, we used a simple fact that for the matrix $A \in \mathbb{R}^{d_1\times d_2}$ with rank r, then $\|A\|_2 \leq \|A\|_F \leq \sqrt{r}\|A\|_2$. Recall that the rank of the matrix $\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(k-j-1)}}(x)\big] - \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(x)\big]$ is at most 2. In the second to the last inequality, we use the result of Lemma D.4.6 and the $\|\mathbf{u}_D(i) - \mathbf{y}\|_2 \leq \mathcal{O}(\sqrt{n})$ for any $i \geq 1$. In the last inequality, we plug the correct orders as set in (D.53) to τ, η_1, η_2 and μ. Back to the inequality (D.79), using the $\left\|z_{D,k}(x)\right\|_2 \leq \mathcal{O}\big(\sqrt{m\omega}\big)$ and (D.80), we can get

$$\frac{1}{L}\sum_{\ell=1}^{L} \|z_{D,k}(x)\|_2^2 \, \|\mathbf{E}_3\|_2^2 \leq \mathcal{O}_{\mathbb{P}}\Bigg(\frac{L^{20/3}\omega^{4/3}}{m^{1/3}} n^{\frac{8d}{6d-3}} \log(m)\Bigg). \tag{D.81}$$

Before controlling the ℓ_2 norm of $\mathcal{C}$ in (D.76), recall that we set $\mathbf{B} := \left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty\right)^{-1}\mathbf{H}_L^\infty\mathbf{y}$ and the dynamics of $\mathbf{u}_D(k) - \mathbf{B}$ can be expressed in terms of $\mathbf{H}_L^\infty$ as follows: For any $k \geq 1$,

$$\mathbf{u}_D(k) - \mathbf{B} = \bigg(\big(1 - \eta_2\mu L\big)\cdot\mathcal{I} - m\eta_1\mathbf{H}_L^\infty\bigg)^k \big(\mathbf{u}_D(0) - \mathbf{B}\big) + \mathbf{e}_D(k), \tag{D.82}$$

with $\left\|\mathbf{e}_D(k)\right\|_2 \leq \mathcal{O}\big(\frac{1}{n}\big)$. Using (D.82), we can further decompose the term $\mathbf{E}_2$ in (D.75) as:

$$\begin{aligned}
\mathbf{E}_2 &:= -\eta_1\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big]\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\bigg(\mathbf{u}_D(k-j-1) - y\bigg) \\
&= \eta_1\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big]\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\bigg(\big(1-\eta_2\mu L\big)\cdot\mathcal{I} - m\eta_1\mathbf{H}_L^\infty\bigg)^{k-j-1}\mathbf{B} \\
&\qquad - \eta_1\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big]\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\bigg(\big(1-\eta_2\mu L\big)\cdot\mathcal{I} - m\eta_1\mathbf{H}_L^\infty\bigg)^{k-j-1}\mathbf{u}_D(0) \\
&\qquad - \eta_1\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big]\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\mathbf{e}_D(k-j-1) \\
&\qquad - \eta_1\nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big]\sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\bigg(\mathbf{B} - y\bigg) \\
&= \mathbf{E}_{2,1} + \mathbf{E}_{2,2} + \mathbf{E}_{2,3} + \mathbf{E}_{2,4}.
\end{aligned} \tag{D.83}$$

Then, we can re-write the error term $\mathcal{C}$ in (D.76) as:

$$\mathcal{C} = \frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_{2,1}\rangle + \frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_{2,2}\rangle + \frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_{2,3}\rangle$$
$$+ \underbrace{\left\{\frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_{2,4}\rangle - \mathbf{Ker}(x, \mathbf{X})\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\}}_{:=\mathcal{D}}. \qquad \text{(D.84)}$$

Our goal is to control the ℓ_2 norm of each summand in the equality (D.84). For the first three terms in (D.84), a simple Cauchy-Schwarz inequality can be applied: for $i = 1, 2, 3$:

$$\left|\frac{1}{L}\sum_{\ell=1}^{L}\langle z_{D,k}(x), \mathbf{E}_{2,i}\rangle\right|^2 \le \frac{1}{L}\sum_{\ell=1}^{L}\left|\langle z_{D,k}(x), \mathbf{E}_{2,i}\rangle\right|^2 \le \frac{1}{L}\sum_{\ell=1}^{L}\|z_{D,k}(x)\|_2^2 \cdot \|\mathbf{E}_{2,i}\|_2^2 .$$

We work on obtaining the bound of $\sum_{\ell=1}^{L}\|\mathbf{E}_{2,1}\|_2^2$. Let $\mathcal{T}_k$ be defined as

$$\mathcal{T}_k := \sum_{j=0}^{k-1}\left(1-\eta_2\mu\right)^j\left(\left(1-\eta_2\mu L\right)\cdot\mathcal{I} - m\eta_1\mathbf{H}_L^\infty\right)^{k-j-1}.$$

Then, we have

$$\begin{aligned}\sum_{\ell=1}^{L}\|\mathbf{E}_{2,1}\|_2^2 &= \eta_1^2\sum_{\ell=1}^{L}\left(\mathbf{B}^\top\mathcal{T}_k^\top\nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}_D(0)}(\mathbf{X})\right]^\top\nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}_D(0)}(\mathbf{X})\right]\mathcal{T}_k\mathbf{B}\right)\\ &= m\eta_1^2\mathbf{B}^\top\mathcal{T}_k^\top\mathbf{H}(0)\mathcal{T}_k\mathbf{B}\\ &= m\eta_1^2\mathbf{B}^\top\mathcal{T}_k^\top\left(\mathbf{H}(0)-\mathbf{H}_L^\infty\right)\mathcal{T}_k\mathbf{B} + m\eta_1^2\mathbf{B}^\top\mathcal{T}_k^\top\mathbf{H}_L^\infty\mathcal{T}_k\mathbf{B}\\ &\le m\eta_1^2\left\|\mathbf{H}(0)-\mathbf{H}_L^\infty\right\|_2\cdot\mathbf{B}^\top\mathcal{T}_k^2\mathbf{B} + m\eta_1^2\mathbf{B}^\top\mathcal{T}_k^\top\mathbf{H}_L^\infty\mathcal{T}_k\mathbf{B}. \qquad \text{(D.85)}\end{aligned}$$

To obtain the upper-bound on (D.85), we need to control the terms $\mathcal{T}_k^\top\mathbf{H}_L^\infty\mathcal{T}_k$ and $\mathbf{B}^\top\mathcal{T}_k^2\mathbf{B}$. Let us denote $\mathbf{H}_L^\infty = \sum_{i=1}^{n}\lambda_i v_i v_i^\top$ be the eigen-decomposition of $\mathbf{H}_L^\infty$. Using $1-\eta_2\mu L \le$

$1-\eta_2\mu$, note that

$$\begin{aligned}\mathcal{T}_k &= \sum_{j=0}^{k-1}\big(1-\eta_2\mu\big)^j\big(1-\eta_2\mu L\big)^{k-j-1}\bigg(\mathcal{I}-\frac{m\eta_1}{1-\eta_2\mu L}\mathbf{H}_L^\infty\bigg)^{k-j-1} \\ &\preceq \big(1-\eta_2\mu\big)^{k-1}\sum_{i=0}^{k-1}\bigg(\mathcal{I}-\frac{m\eta_1}{1-\eta_2\mu}\mathbf{H}_L^\infty\bigg)^{i} \\ &= \big(1-\eta_2\mu\big)^{k-1}\sum_{j=0}^{n}\Bigg(\frac{1-\big(1-\frac{m\eta_1}{1-\eta_2\mu}\lambda_j\big)^k}{\frac{m\eta_1}{1-\eta_2\mu}\lambda_j}\Bigg)v_jv_j^\top \preceq \frac{\big(1-\eta_2\mu\big)^k}{m\eta_1\lambda_\infty}\cdot\mathcal{I}. \end{aligned} \tag{D.86}$$

A similar logic can be applied to bound $\mathcal{T}_k^\top\mathbf{H}_L^\infty\mathcal{T}_k$:

$$\begin{aligned}\mathcal{T}_k^\top\mathbf{H}_L^\infty\mathcal{T}_k &\preceq \big(1-\eta_2\mu\big)^{k-1}\sum_{j=0}^{n}\Bigg(\frac{1-\big(1-\frac{m\eta_1}{1-\eta_2\mu}\lambda_j\big)^k}{\frac{m\eta_1}{1-\eta_2\mu}\lambda_j}\Bigg)^2\lambda_jv_jv_j^\top \\ &\preceq \frac{\big(1-\eta_2\mu\big)^{2k}}{m^2\eta_1^2}\cdot\big(\mathbf{H}_L^\infty\big)^{-1}. \end{aligned} \tag{D.87}$$

Recall the definition of the notation $\mathbf{B} := \mathbf{H}_L^\infty\bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\bigg)^{-1}\mathbf{y}$. Then, we can bound the term $\mathbf{B}^\top\mathcal{T}_k^\top\mathbf{H}_L^\infty\mathcal{T}_k\mathbf{B}$:

$$\begin{aligned}\mathbf{B}^\top\mathcal{T}_k^\top\mathbf{H}_L^\infty\mathcal{T}_k\mathbf{B} &\le \frac{\big(1-\eta_2\mu\big)^{2k}}{m^2\eta_1^2}\cdot\mathbf{B}^\top\big(\mathbf{H}_L^\infty\big)^{-1}\mathbf{B} \\ &= \frac{\big(1-\eta_2\mu\big)^{2k}}{m^2\eta_1^2}\cdot\mathbf{y}^\top\bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\bigg)^{-1}\mathbf{H}_L^\infty\bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\bigg)^{-1}\mathbf{y} \\ &= \mathcal{O}\Bigg(\frac{\big(1-\eta_2\mu\big)^{2k}}{m^2\eta_1^2}\Bigg), \end{aligned} \tag{D.88}$$

where in the last equality, we used $\big\|g_\mu^*\big\|_{\mathcal{H}}^2=\mathcal{O}_{\mathbb{P}}(1)$ in (D.74). Now we turn our attention to bound the term $\mathbf{B}^\top\mathcal{T}_k^2\mathbf{B}$,

$$\begin{aligned}\mathbf{B}^\top\mathcal{T}_k^2\mathbf{B} &\le \frac{\big(1-\eta_2\mu\big)^{2k}}{m^2\eta_1^2\lambda_\infty^2}\mathbf{y}^\top\bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\bigg)^{-1}\Big(\mathbf{H}_L^\infty\Big)^2\bigg(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\bigg)^{-1}\mathbf{y} \\ &= \mathcal{O}\Bigg(\frac{\big(1-\eta_2\mu\big)^{2k}n}{m^2\eta_1^2\lambda_\infty^2}\Bigg), \end{aligned} \tag{D.89}$$

where we used $\|y\|_2^2 = \mathcal{O}(n)$ in the last inequality. Combining the bounds (D.88), (D.89) and the result from Lemma D.4.11, we can further bound (D.85) and have:

$$\sum_{\ell=1}^{L} \|\mathbf{E}_{2,1}\|_2^2 \leq \mathcal{O}\bigg(\omega \frac{\big(1-\eta_2\mu\big)^{2k}}{m\lambda_\infty^2} n^2 L^{5/2} \sqrt[4]{\tfrac{\log(nL/\delta)}{m}} + \frac{\big(1-\eta_2\mu\big)^{2k}}{m}\bigg) \leq \mathcal{O}\bigg(\frac{\big(1-\eta_2\mu\big)^{2k}}{m}\bigg), \tag{D.90}$$

where in the second inequality, we used $m \geq \Omega\bigg(L^{19} n^{20} \frac{\log^3(m)}{\delta^2}\bigg)$. Similarly, we can bound $\sum_{\ell=1}^{L} \|\mathbf{E}_{2,2}\|_2^2$:

$$\begin{aligned}
\sum_{\ell=1}^{L} \|\mathbf{E}_{2,2}\|_2^2 &= \eta_1^2 \sum_{\ell=1}^{L} \bigg(\mathbf{u}_D(0)^\top \mathcal{T}_k^\top \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{X})\big]^\top \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_{D(0)}}(\mathbf{X})\big] \mathcal{T}_k \mathbf{u}_D(0)\bigg) \\
&= m\eta_1^2 \mathbf{u}_D(0)^\top \mathcal{T}_k^\top \mathbf{H}(0) \mathcal{T}_k \mathbf{u}_D(0) \\
&= m\eta_1^2 \mathbf{u}_D(0)^\top \mathcal{T}_k^\top \bigg(\mathbf{H}(0) - \mathbf{H}_L^\infty\bigg) \mathcal{T}_k \mathbf{u}_D(0) + m\eta_1^2 \mathbf{u}_D(0)^\top \mathcal{T}_k^\top \mathbf{H}_L^\infty \mathcal{T}_k \mathbf{u}_D(0) \\
&\leq m\eta_1^2 \|\mathbf{H}(0) - \mathbf{H}_L^\infty\|_2 \cdot \mathbf{u}_D(0)^\top \mathcal{T}_k^2 \mathbf{u}_D(0) + m\eta_1^2 \mathbf{u}_D(0)^\top \mathcal{T}_k^\top \mathbf{H}_L^\infty \mathcal{T}_k \mathbf{u}_D(0) \\
&\leq m\eta_1^2 \frac{\big(1-\eta_2\mu\big)^{2k}}{m^2\eta_1^2\lambda_\infty^2} \mathcal{O}\bigg(\omega n L^{5/2} \sqrt[4]{\tfrac{\log(nL/\delta)}{m}}\bigg) \|\mathbf{u}_D(0)\|_2^2 \\
&\qquad + m\eta_1^2 \frac{\big(1-\eta_2\mu\big)^{2k}}{m^2\eta_1^2} \mathbf{u}_D(0)^\top \big(\mathbf{H}_L^\infty\big)^{-1} \mathbf{u}_D(0) \\
&\leq \mathcal{O}\bigg(\frac{\big(1-\eta_2\mu\big)^{2k} n^2\omega^2 L^{5/2}}{m\lambda_\infty^2\delta^2} \sqrt[4]{\tfrac{\log(nL/\delta)}{m}} + \frac{n\omega\big(1-\eta_2\mu\big)^{2k}}{m\lambda_\infty\delta^2}\bigg) \\
&= \mathcal{O}_{\mathbb{P}}\bigg(\frac{n\omega\big(1-\eta_2\mu\big)^{2k}}{m\lambda_\infty}\bigg).
\end{aligned} \tag{D.91}$$

Here, in the second inequality, we used the inequalities (D.86) and (D.87) and Lemma D.4.11. In the third inequality, we used the Lemma D.4.8, $\|\mathbf{u}(0)\|_2 = \mathcal{O}\big(\frac{\sqrt{n\omega}}{\delta}\big)$ with probability at least $1-\delta$. In the last equality, we used $m \geq \Omega\bigg(L^{19} n^{20} \frac{\log^3(m)}{\delta^2}\bigg)$.

Next, we bound $\sum_{\ell=1}^{L} \|\mathbf{E}_{2,3}\|_2^2$ as:

$$\sum_{\ell=1}^{L} \|\mathbf{E}_{2,3}\|_2^2 = m\eta_1^2 \cdot \Big(\sum_{j=0}^{k-1} \big(1-\eta_2\mu\big)^j \big(\mathbf{e}_{k-j-1}\big) \Big)^\top \mathbf{H}_D(0) \Big(\sum_{j=0}^{k-1} \big(1-\eta_2\mu\big)^j \big(\mathbf{e}_{k-j-1}\big) \Big)$$
$$\leq \frac{m\eta_1^2}{\eta_2^2\mu^2} \cdot \lambda_{\max}\big(\mathbf{H}_D(k)\big) \cdot \|\mathbf{e}_{k-j-1}\|_2^2 \leq \frac{m\eta_1^2}{\eta_2^2\mu^2} \cdot \mathcal{O}\big(\omega nL\big) \cdot \mathcal{O}\Big(\frac{1}{n^2}\Big) = \mathcal{O}\Big(\frac{L^3}{m}\omega \cdot n^{-\frac{4d-3}{2d-1}}\Big). \tag{D.92}$$

Now, we focus on obtaining the ℓ_2 norm bound on $\mathcal{D}$ in (D.84). Recall the definition of the notation $\mathbf{B} := \mathbf{H}_L^\infty \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y}$. A simple calculation yields that

$$\mathbf{B} - \mathbf{y} = \mathbf{H}_L^\infty \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y} - \mathbf{y} = -\frac{\eta_2\mu L}{m\eta_1} \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y}.$$

Then, we can re-write the expression of the $\mathcal{D}$ as :

$$\begin{aligned}
\mathcal{D} := {} & \Big(\frac{\eta_2\mu L}{m\eta_1} \Big) \cdot \eta_1 \frac{1}{L} \sum_{\ell=1}^{L} \big\langle z_{D,k}(x), \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big] \big\rangle \sum_{j=0}^{k-1} \big(1-\eta_2\mu\big)^j \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y} \\
& - \mathbf{Ker}(x, \mathbf{X}) \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y} \\
= {} & \Big(\frac{1}{m} \sum_{\ell=1}^{L} \big\langle z_{D,k}(x), \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big] \big\rangle - \mathbf{Ker}(x, \mathbf{X}) \Big) \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y} \\
& - \big(1-\eta_2\mu\big)^k \frac{1}{m} \sum_{\ell=1}^{L} \big\langle z_{D,k}(x), \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big] \big\rangle \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y} \\
= {} & \Big(\frac{1}{m} \sum_{\ell=1}^{L} \big\langle z_{0,k}(x), \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big] \big\rangle - \mathbf{Ker}(x, \mathbf{X}) \Big) \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y} \\
& - \big(1-\eta_2\mu\big)^k \frac{1}{m} \sum_{\ell=1}^{L} \big\langle z_{0,k}(x), \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big] \big\rangle \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y} \\
& + \big(1 - \big(1-\eta_2\mu\big)^k\big) \Big(\frac{1}{m} \sum_{\ell=1}^{L} \big\langle z_{D,k}(x) - z_{D,0}(x), \nabla_{\mathbf{W}_\ell}\big[f_{\mathbf{W}_D(0)}(\mathbf{X})\big] \big\rangle \Big) \Big(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty \Big)^{-1} \mathbf{y},
\end{aligned} \tag{D.93}$$

where in the second equality, $\sum_{j=0}^{k-1}(1-\eta_2\mu)^j = \frac{1-(1-\eta_2\mu)^k}{\eta_2\mu}$ is used. The ℓ_2 norm of first term in the (D.93) can be bounded as:

$$\begin{aligned}
&\left\|\left(\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_{D,k}(x), \nabla_{\mathbf{w}_\ell}\left[f_{\mathbf{w}_D(0)}(\mathbf{X})\right]\right\rangle - \mathbf{Ker}(x,\mathbf{X})\right)\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\le \left\|\left(\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_{D,k}(x), \nabla_{\mathbf{w}_\ell}\left[f_{\mathbf{w}_D(0)}(\mathbf{X})\right]\right\rangle - \mathbf{Ker}(x,\mathbf{X})\right)\right\|_2 \cdot \left\|\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&= \sqrt{\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_0(x), \nabla_{\mathbf{w}_\ell}\left[f_{\mathbf{w}_D(0)}(\mathbf{x}_i)\right]\right\rangle - \mathbf{Ker}(x,\mathbf{x}_i)\right)^2} \cdot \left\|\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\le \mathcal{O}\left(\frac{\omega\eta_1 mn}{\eta_2\mu L}L^{5/2}\sqrt[4]{\frac{\log(nL/\delta)}{m}}\right) = \mathcal{O}\left(\omega L^{5/2} n^{\frac{d}{2d-1}}\sqrt[4]{\frac{\log(nL/\delta)}{m}}\right), \qquad \text{(D.94)}
\end{aligned}$$

where, in the second inequality, we used Lemma D.4.11, and also we used

$$\left\|\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \le \sqrt{\mathbf{y}^\top\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-2}\mathbf{y}} \le \sqrt{\frac{\eta_1^2 m^2}{\eta_2^2\mu^2 L^2}\cdot\|\mathbf{y}\|_2^2} = \mathcal{O}\left(\frac{\eta_1 m}{\eta_2\mu L}\sqrt{n}\right). \qquad \text{(D.95)}$$

The ℓ_2 norm of the second term in (D.93) can be easily bounded as:

$$\begin{aligned}
&\left\|\left(1-\eta_2\mu\right)^k\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_0(x), \nabla_{\mathbf{w}_\ell}\left[f_{\mathbf{w}_D(0)}(\mathbf{X})\right]\right\rangle\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\qquad \le \left\|\left(1-\eta_2\mu\right)^k\left(\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_0(x), \nabla_{\mathbf{w}_\ell}\left[f_{\mathbf{w}_D(0)}(\mathbf{X})\right]\right\rangle - \mathbf{Ker}(x,\mathbf{X})\right)\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\qquad\qquad + \left\|\left(1-\eta_2\mu\right)^k\mathbf{Ker}(x,\mathbf{X})\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\qquad \le \left(1-\eta_2\mu\right)^k\left\|\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_0(x), \nabla_{\mathbf{w}_\ell}\left[f_{\mathbf{w}_D(0)}(\mathbf{X})\right]\right\rangle - \mathbf{Ker}(x,\mathbf{X})\right\|_2 \cdot \left\|\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\qquad\qquad + \left(1-\eta_2\mu\right)^k\left\|\mathbf{Ker}(x,\mathbf{X})\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\qquad \le \left(1-\eta_2\mu\right)^k\cdot\mathcal{O}\left(\omega\sqrt{n}L^{3/2}\sqrt[4]{\frac{\log(nL/\delta)}{m}}\right)\cdot\mathcal{O}\left(\frac{\eta_1 m}{\eta_2\mu L}\sqrt{n}\right) + \mathcal{O}\left(\left(1-\eta_2\mu\right)^k\right) \\
&\qquad \le \left(1-\eta_2\mu\right)^k\cdot\mathcal{O}\left(\omega L^{3/2}n^{\frac{d}{2d-1}}\sqrt[4]{\frac{\log(nL/\delta)}{m}}\right) + \mathcal{O}\left(\left(1-\eta_2\mu\right)^k\right). \qquad \text{(D.96)}
\end{aligned}$$

Lastly, the ℓ_2 norm of the third term in (D.93) is bounded as:

$$\begin{aligned}
&\left\| \left(1-\left(1-\eta_2\mu\right)^k\right)\left(\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_{D,k}(x)-z_{D,0}(x), \nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}_D(0)}(\mathbf{X})\right]\right\rangle\right)\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\quad\le \left(1-\left(1-\eta_2\mu\right)^k\right)\cdot\left\|\frac{1}{m}\sum_{\ell=1}^{L}\left\langle z_{D,k}(x)-z_{D,0}(x), \nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}_D(0)}(\mathbf{X})\right]\right\rangle\right\|_2 \\
&\qquad\cdot\left\|\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\quad\le \left(1-\left(1-\eta_2\mu\right)^k\right)\cdot\left(\frac{1}{m}\sum_{\ell=1}^{L}\left\|z_{D,k}(x)-z_{D,0}(x)\right\|_F\left\|\nabla_{\mathbf{W}_\ell}\left[f_{\mathbf{W}_D(0)}(\mathbf{X})\right]\right\|_F\right) \qquad \text{(D.97)} \\
&\qquad\cdot\left\|\left(\frac{\eta_2\mu L}{\eta_1 m}\mathcal{I}+\mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2 \\
&\quad\le \left(1-\left(1-\eta_2\mu\right)^k\right)\cdot\left(\frac{L}{m}\mathcal{O}\left(\tau^{1/3}L^2\sqrt{\omega m\log(m)}\right)\cdot\mathcal{O}\left(\sqrt{\omega m n}\right)\right)\cdot\mathcal{O}\left(\frac{\eta_1 m}{\eta_2\mu L}\sqrt{n}\right) \\
&\quad\le \left(1-\left(1-\eta_2\mu\right)^k\right)\cdot\mathcal{O}\left(\omega^{7/6}L^{10/3}n^{\frac{4d}{6d-3}}\frac{\sqrt{\log(m)}}{m^{1/6}\delta^{1/3}}\right) \qquad \text{(D.98)} \\
&\le \mathcal{O}\left(\omega^{7/6}L^{10/3}n^{\frac{4d}{6d-3}}\frac{\sqrt{\log(m)}}{m^{1/6}\delta^{1/3}}\right), \qquad \text{(D.99)}
\end{aligned}$$

where in the fourth inequality, $\tau=\mathcal{O}_{\mathbb{P}}\left(\frac{L\sqrt{\omega}}{\sqrt{m}}n^{\frac{d}{2d-1}}\right)$ is plugged in. Combining the inequalities (D.94), (D.96) and (D.99), we get the bound on $\|\mathcal{D}\|_2$ in (D.93):

$$\begin{aligned}
\|\mathcal{D}\|_2^2 &\le \mathcal{O}\left(\omega^2L^5n^{\frac{2d}{2d-1}}\sqrt{\frac{\log(nL/\delta)}{m}}\right)+\left(1-\eta_2\mu\right)^{2k}\mathcal{O}\left(\omega^2L^3n^{\frac{2d}{2d-1}}\sqrt{\frac{\log(nL/\delta)}{m}}\right) \qquad \text{(D.100)} \\
&\quad+\mathcal{O}\left(\left(1-\eta_2\mu\right)^{2k}\right) \\
&\quad+\mathcal{O}\left(\omega^{7/3}L^{20/3}n^{\frac{8d}{6d-3}}\frac{\log(m)}{m^{1/3}\delta^{2/3}}\right) \\
&\le \mathcal{O}\left(\omega^{7/3}L^{20/3}n^{\frac{8d}{6d-3}}\frac{\log(m)}{m^{1/3}\delta^{2/3}}\right)+\mathcal{O}\left(\left(1-\eta_2\mu\right)^{2k}\right). \qquad \text{(D.101)}
\end{aligned}$$

Step 3. Combining all pieces. Recall $\|z_{D,k}(x)\|_2\le\mathcal{O}\left(\sqrt{m\omega}\right)$. With this fact, combining the bounds (D.78), (D.81), (D.90), (D.91), (D.92) and (D.101), we can bound the $\|\Delta_{D,1}\|_2^2$ via the decomposition (D.76) as follows:

$$\begin{aligned}
\|\Delta_{D,1}\|_2^2 \leq \frac{1}{L}\sum_{\ell=1}^{L} \left| z_{D,k}(x)^\top \mathbf{vec}\big(\mathbf{W}_{D,\ell}^{(0)}\big)\right|^2 &+ \frac{1}{L}\sum_{\ell=1}^{L} \|z_{D,k}(x)\|_2^2 \, \|\mathbf{E}_3\|_2^2 \\
&+ \frac{1}{L}\sum_{\ell=1}^{L} \|z_{D,k}(x)\|_2^2 \, \|\mathbf{E}_{2,1}\|_2^2 + \frac{1}{L}\sum_{\ell=1}^{L} \|z_{D,k}(x)\|_2^2 \, \|\mathbf{E}_{2,2}\|_2^2 \\
&+ \frac{1}{L}\sum_{\ell=1}^{L} \|z_{D,k}(x)\|_2^2 \, \|\mathbf{E}_{2.3}\|_2^2 \\
&+ \left\| \frac{1}{L}\sum_{\ell=1}^{L} \langle z_{D,k}(x), \mathbf{E}_{2,4}\rangle - \mathbf{Ker}(x,\mathbf{X})\left(\frac{\eta_2 \mu L}{\eta_1 m}\mathcal{I} + \mathbf{H}_L^\infty\right)^{-1}\mathbf{y}\right\|_2^2 \\
\leq \mathcal{O}\Bigg(\omega \log\left(\frac{L}{\delta}\right)\Bigg) &+ \mathcal{O}_{\mathbb{P}}\bigg(\frac{L^{20/3}\omega^{4/3}}{m^{1/3}} n^{\frac{8d}{6d-3}} \log(m)\bigg) \\
&+ \mathcal{O}_{\mathbb{P}}\left(\frac{\omega\big(1-\eta_2\mu\big)^{2k}}{L}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{n\omega^2\big(1-\eta_2\mu\big)^{2k}}{L\lambda_\infty}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{L^2}{m}\omega \cdot n^{-\frac{4d-3}{2d-1}}\right) \\
&+ \mathcal{O}\left(\omega^{7/3} L^{20/3} n^{\frac{8d}{6d-3}} \frac{\log(m)}{m^{1/3}\delta^{2/3}}\right) + \mathcal{O}_{\mathbb{P}}\left(\big(1-\eta_2\mu\big)^{2k}\right) \\
\leq \mathcal{O}_{\mathbb{P}}\left(n^{-\frac{d}{2d-1}}\right).
\end{aligned}$$

APPENDIX E

APPROXIMATION AND NON-PARAMETRIC ESTIMATION OF FUNCTIONS OVER HIGH-DIMENSIONAL SPHERES VIA DEEP RELU NETWORKS

E.1 d^d-dependent constant in $\mathcal{N}$ for approximating $f \in W^r_\infty([0,1]^d)$

First, we define the function space $W^r_\infty([0,1]^d)$ on the d-dimensional unit cube. For $r = n + \sigma$ where $n \in \mathbb{N}_0$ and $\sigma \in (0,1]$, a function has Hölder smoothness index r if all partial derivatives up to order n exist and are bounded and the partial derivatives of order n are σ Hölder. Formally, the ball of r-Hölder functions with radius $\mathcal{Q}$ is then defined as

$$W^r_\infty([0,1]^d) = \\ \left\{ f : [0,1]^d \to \mathbb{R} : \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\leq n} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=n} \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in[0,1]^d \\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|^\sigma_\infty} \leq \mathcal{Q} \right\}.$$

where $\partial^{\boldsymbol{\alpha}} f := \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial^{\alpha_1}\ldots\partial^{\alpha_d}} f$ for the multi-index notation, $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_d)$.

The fundamental ideas for approximating functions $f \in W^r_\infty([0,1]^d)$ in the existing literature rely on a local Taylor approximation technique. The technique dicretizes d-dimensional input cube into a sub-cube set whose size is $(K+1)^d$ where $(K+1)$ is the grid size of each coordinate. For any $\mathbf{x}$ in the input cube, the function f is approximated by using the closest 2^d grid points to $\mathbf{x}$ via Taylor expansion of f up to the degree $\lfloor r \rfloor$, where we denote the largest integer less than or equal to $u > 0$ as $\lfloor u \rfloor$. Therefore, the total number of active parameters for the net is at least more than the total number of coefficients of partial derivatives $\partial^\alpha f := \frac{\partial^{|\alpha|}}{\partial^{\alpha_1}\ldots\partial^{\alpha_d}} f$ for $|\alpha| = |\alpha_1| + \cdots + |\alpha_d| \leq \lfloor r \rfloor$. This yields the lower bound on the active parameters for the network via local Taylor approximation as $(K+1)^d \cdot \sum_{i=0}^{\lfloor r \rfloor} \binom{d+i-1}{d}$.

E.2 Roadmaps for proof of Theorem 5.2.1

In this section of the Appendix, we provide the definitions of $L_N(f)$ and $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)$ along with the overall picture for the proof of Theorem 5.2.1. Recall we have the following decomposition:

$$\left\|f - \tilde{f}\right\|_\infty \leq \underbrace{\|f - L_N(f)\|_\infty}_{:=\textbf{(I)}} + \underbrace{\left\|L_N(f) - \widehat{L}^{\boldsymbol{y}}_{N,M}(f)\right\|_\infty}_{:=\textbf{(II)}} + \underbrace{\left\|\widehat{L}^{\boldsymbol{y}}_{N,M}(f) - \tilde{f}\right\|_\infty}_{:=\textbf{(III)}}. \tag{E.1}$$

In Subsection B.1, we provide the idea for bounding **(I)** and **(II)**. In Subsection B.2, the construction of neural network $\tilde{f}$ for approximating $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)$ is described. In this section, no proofs on Propositions and Lemmas are included, but only key ideas for the proofs and technical comparisons with other literature are provided. All the detailed proofs of technical statements in this section are deferred in the Appendix $C.1$.

E.2.1 Error bounds for **(I)** and **(II)**

A function $f \in W^r_\infty(\mathcal{S}^{d-1})$ is approximated by a linear scheme L_N defined as follows.

Definition E.2.1 *Given a $C^\infty([0,\infty])$ function η with $\eta(t) = 1$ for $0 \leq t \leq 1$ and $\eta(t) = 0$ for $t \geq 2$, we define a sequence of linear operator L_N, $N \in \mathbb{N}$, on $\mathcal{L}_p(\mathcal{S}^{d-1})$ with $1 \leq p \leq \infty$ by*

$$L_N(f)(\mathbf{x}) := \sum_{k=0}^{2N} \eta\left(\frac{k}{N}\right) \boldsymbol{Proj}_k(f)(\mathbf{x}) = \int_{\mathcal{S}^{d-1}} f(\mathbf{y}) \ell_{k,d}(\langle \mathbf{x}, \mathbf{y} \rangle) \rho_{\mathcal{X}}(d\mathbf{y}), \quad \mathbf{x} \in \mathcal{S}^{d-1}, \tag{E.2}$$

where with $\lambda_G = \frac{d-2}{2}$, $\ell_{N,d}$ is a kernel given by

$$\ell_{N,d}(t) := \sum_{k=0}^{2N} \eta\left(\frac{k}{N}\right) \frac{k + \lambda_G}{\lambda_G} \mathcal{G}^{\lambda_G}_k(t), \quad t \in [-1, 1]. \tag{E.3}$$

It can be found in [193] (Chapter 4) that L_N is near best, achieving the order of best approximation for $f \in W_p^r(\mathcal{S}^{d-1})$.

Lemma E.2.2 (Lemma 1 in [179]) *For $N \in \mathbb{N}$, $1 \leq p \leq \infty$, $r > 0$, and $f \in W_p^r(\mathcal{S}^{d-1})$, there holds*

$$\|f - L_N(f)\|_p \leq C_\eta N^{-r} \cdot \|f\|_{W_\infty^r(\mathcal{S}^{d-1})}, \tag{E.4}$$

where C_η is a constant depending only on the function η in defining L_N.

Note that $\big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)^{r/2}$ is a self-adjoint operator. For $\mathbf{x} \in \mathcal{S}^{d-1}$, recalling the definition of $L_N(f)$, we have

$$\begin{aligned} L_N(f)(\mathbf{x}) &= \langle f, \ell_{N,d}(\langle \mathbf{x}, \cdot \rangle)\rangle_{\mathcal{L}_2(\mathcal{S}^{d-1})} \\ &= \Big\langle \big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)^{r/2} f, \big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)^{-r/2} \ell_{N,d}(\langle \mathbf{x}, \cdot \rangle)\Big\rangle_{\mathcal{L}_2(\mathcal{S}^{d-1})} \\ &= \int_{\mathcal{S}^{d-1}} F_r(\mathbf{y}) \cdot \xi_{N,r}(\langle \mathbf{x}, \mathbf{y}\rangle) \rho_{\mathcal{X}}(d\mathbf{y}). \end{aligned} \tag{E.5}$$

Hereafter, we denote $F_r = \big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)^{r/2} f$ and $\xi_{N,r}(\langle \mathbf{x}, \cdot \rangle) = \big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)^{-r/2} \ell_{N,d}(\langle \mathbf{x}, \cdot \rangle)$. By the fractional power of the operator $\big(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I}\big)^{-r/2}$ in a distributional sense, $\xi_{N,r}(\cdot)$ is a polynomial of degree at most $2N$ written as:

$$\xi_{N,r}(t) = \sum_{k=0}^{2N} \big(1 + \lambda_k\big)^{-r/2} \eta\Big(\frac{k}{N}\Big) \frac{k + \lambda_{\mathrm{G}}}{\lambda_{\mathrm{G}}} \mathcal{G}_k^{\lambda_{\mathrm{G}}}(t), \quad t \in [-1, 1]. \tag{E.6}$$

The fractional power of $(-\Delta_{\mathcal{S}^{d-1}} + \mathcal{I})$ caused by the regularity $f \in W_\infty^r(\mathcal{S}^{d-1})$ enables r-dependent error bound for discretizing $L_N(f)$: the larger the regularity r becomes, the smaller the bound for approximation error gets.

Following [179], the key idea for a constructing neural network that approximates $L_N(f)$ is to discretize the integral form (E.5) by M random samples $\mathbf{y} = \{\mathbf{y_1}, \ldots, \mathbf{y_M}\}$

independently drawn from $\rho_{\mathcal{X}}$. We write the discretized version of (E.5) as :

$$\widehat{L}_{N,M}^{\boldsymbol{y}}(f)(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} F_r(\mathbf{y}_i)\cdot \xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i\rangle), \quad \forall \mathbf{x} \in \mathcal{S}^{d-1}. \tag{E.7}$$

Before estimating the distance between $L_N(f)$ and $\widehat{L}_{N,M}^{\boldsymbol{y}}(f)$, we need a Sobolev embedding property.

Proposition E.2.3 *For $d \geq 5$, $1 \leq p \leq \infty$, and $s \geq \frac{3d-2}{4}$, the Sobolev space $W_p^s(\mathcal{S}^{d-1})$ is continuously embedded into $C(\mathcal{S}^{d-1})$, the space of continuous functions on $\mathcal{S}^{d-1}$, which implies*

$$\|f\|_\infty \leq c_0\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} \cdot \|f\|_{W_p^s(\mathcal{S}^{d-1})}, \qquad f \in W_p^s(\mathcal{S}^{d-1}),$$

where c_0 is an absolute constant independent of r, d, s, and f.

Proposition E.2.3 is motivated from Eq.(14) in [195], where they proved $\|f\|_\infty \leq C_{s,d}\,\|f\|_{W_p^s(\mathcal{S}^{d-1})}$, $f \in W_p^s(\mathcal{S}^{d-1})$ for $s \geq \frac{d-1}{2}$. The constant obtained in [195] is $C_{s,d} := \big(\frac{1}{\omega_d}\sum_{k=0}^{\infty}\frac{\mathcal{N}(k,d)}{(k+\frac{d-2}{2})^{2s}}\big)^{1/2}$, where ω_d is the surface of d-dimensional sphere. For large enough d, $(1/\omega_d)^{1/2}$ grows in the order of $\mathcal{O}\big(\big(\frac{d}{2\pi e}\big)^{d/4}\big)$. Then, by choosing $s \geq \frac{3d-2}{4}$, $(1/\omega_d)^{1/2}$ can be absorbed into the infinite sum making the constant $C_{s,d}$ converge in an asymptotic regime of d. It should be noted that the threshold on smoothness index $(i.e., s \geq \frac{3d-2}{4})$ is larger than that from [195] $(i.e., s \geq \frac{d-1}{2})$, where they consider the fixed d. See Appendix E.3.1 for the proof. Next, we state the discretization lemma which provides a probabilistic bound on the difference $L_N(f) - \widehat{L}_{N,M}^{\boldsymbol{y}}(f)$.

Lemma E.2.4 *Let $r \leq \frac{3d-2}{4}$ and $0 < \alpha < 1$. If $f \in W_\infty^r(\mathcal{S}^{d-1})$, then for any $M \in \mathbb{N}$ and $1 \leq N \leq d^\alpha + 1$, there exist $\mathbf{y} = \{y_1, y_2, \ldots, y_M\} \subset \mathcal{S}^{d-1}$ such that*

$$\left\|L_N(f) - \widehat{L}_{N,M}^{\boldsymbol{y}}(f)\right\|_\infty \leq \frac{6\cdot C''\left(\frac{6}{\pi e}\right)^{\frac{d}{4}}\|f\|_{W_\infty^r(\mathcal{S}^{d-1})}\, d^{N+\frac{3d-4r-2}{8}}(2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}},$$

where $C^{''} > 0$ is a constant depending on α but independent of r, f, N, M, and d.

Lemma E.2.4 is motivated by Lemma 2 in [179]. The main framework of the proof is based on the Sobolev embedding property in Proposition E.2.3 and the concentration inequality for random variables with values in a Hilbert space, which can be found in [211]. For the application of the concentration inequality, the random variable $\xi(\mathbf{y}_i) := F_r(\mathbf{y}_i)\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$ in (E.7) needs to be bounded in $\| \cdot \|_{W_2^s(\mathcal{S}^{d-1})}$ norm for $s \geq \frac{3d-2}{4}$. See Appendix E.3.2 for the proof of the Lemma.

When compared with the technical proof of Lemma 2 from [179], the most notable difference comes from tracking the explicit dependency on d in the constant factor. Specifically, under the fixed d setting, [179] did not explicitly express how the constant $c_{s,r,d}$ (see the statement in their Lemma) depends on d. However, in our paper, since the main focus is how the approximation error behaves under $d \to \infty$, we need to keep tracking on how d explicitly affects the bound. The result of Proposition 4.1 in our paper serves as an important role for this tracking. Note that the constant c_0 is independent of s, d, r, f in the bound of Proposition 4.1, and we obtain the bound decays at the rate $\left(\frac{6}{\pi e}\right)^{d/4}$. However, in [179], they utilized the result from [195]; that is, $\|f\|_\infty \leq C_{s,d} \|f\|_{W_p^s(\mathbb{S}^{d-1})}$, $f \in W_p^s(\mathbb{S}^{d-1})$ for $s \geq \frac{d-1}{2}$. Here, note that the constant $C_{s,d}$ is a function of d, and since they work under the fixed d setting, they didn't pay much attention to the dependency. Of course, since we are in an asymptotic setting, we use the Stirling's formula to see behaviors of $\mathcal{N}(k,d)$ as $d \to \infty$, whereas [179] just used a simple calculation $\mathcal{N}(k,d) \leq c_d' k^{d-2}$, for some c_d' dependent on d.

E.2.2 Construction of Deep ReLU Networks : Error bound for **(III)**

In this section, several useful tools for the construction of neural network for approximating function $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)$ are introduced. Then, the full proof of our main theorem is presented.

The first key lemma is from [182] wherein the neural network that approximates the quadratic function x^2 for $x \in [0,1]$ is constructed.

Lemma E.2.5 (Proposition 2 in [182]) *For any positive integer $m \geq 1$, there exists a deep ReLU network*

$$\tilde{f}_m \in \mathcal{F}\big(m, \big(1, 5, \ldots, 5, 1\big)\big),$$

such that $\tilde{f}_m \in [0,1]$ and $\Big|\tilde{f}_m(x) - x^2\Big| \leq 2^{-2m-2}$, for all $x \in [0,1]$.

The main idea of Lemma E.2.5 is to approximate the quadratic function via $\tilde{f}_m(x) := x - \sum_{s=1}^{m} \frac{g_s(x)}{2^{2s}}$. Here, $g_s(x)$ is a s-compositions of sawtooth functions defined as

$$g(x) = 2\sigma(x) - 4\sigma(x - 1/2) + 2\sigma(x - 1).$$

Note that $g(x)$ can be implemented by a single layer ReLU network. Then, we can easily construct a ReLU network $\tilde{f}_m$, which belongs to $\mathcal{F}(m, (1, 5, \ldots, 5, 1))$.

Next lemma states that we can construct a neural network that can implement the multiplication operator.

Lemma E.2.6 *For any positive integer $m \geq 1$, there exists a deep ReLU network*

$$\textit{Mult}_m \in \mathcal{F}\big(m + 3, \big(2, 10, \ldots, 10, 1\big)\big),$$

such that $\textit{Mult}_m(x, y) \in [0, 1]$ and

$$|\textit{Mult}_m(x, y) - xy| \leq 2^{-2m-1},$$

for all $x, y \in [0, 1]$. Moreover, $\textit{Mult}_m(x, 0) = \textit{Mult}_m(0, y) = 0$.

The key idea for constructing $\text{Mult}_m(x, y)$ is to invoke the identity $xy = \frac{1}{4}((x + y)^2 - (x - y)^2)$. The first two hidden layers in the network are used to compute $|\frac{x+y}{2}| \in [0, 1]$ and $|\frac{x-y}{2}| \in [0, 1]$ via $|x| = \sigma(x) + \sigma(-x)$. Given the values $|\frac{x+y}{2}|$ and $|\frac{x-y}{2}|$ as inputs, $\tilde{f}_m$

in Lemma E.2.5 is used for approximating $\frac{1}{4}(x+y)^2$ and $\frac{1}{4}(x-y)^2$ in the identity. See Appendix E.3.3 for the detailed proof.

The final key ingredient is to construct a deep ReLU network that approximates univariate polynomial functions of degree $k \in \mathbb{N}$, that is x^k for $x \in [0,1]$.

Lemma E.2.7 *For any positive integer $m \geq 1, N \geq 2$ and for $P = \lceil \log_2(N) \rceil$, there exists a deep ReLU network*

$$\mathit{Poly}_m^{\{N\}} \in \mathcal{F}\big(L, \big(1, 11N, \ldots, 11N, 2^P\big), \mathcal{N}\big),$$

with the depth $L = m + (m+4)\big(\lceil \log_2(N) \rceil - 1\big)$ and the number of parameters $\mathcal{N} \leq 202N \cdot (m+3)$ such that $\mathit{Poly}_m^{\{N\}}(x) \in [0,1]^{2^P}$ and

$$\big|\mathit{Poly}_m^j(x) - x^j\big| \leq P^2 \cdot 2^{-2m-1} \quad \textit{for all} \quad j \in \{1, \ldots, 2^p\}$$

for all $x \in [0,1]$.

Note that the network $\mathrm{Poly}_m^{\{\mathrm{N}\}}(x) := \{\mathrm{Poly}_m^1(x), \ldots, \mathrm{Poly}_m^{2^P}(x)\}$ with $P = \log_2(N)\rceil$ provides approximations to monomials x^j of degree up to $2N$ for $x \in [0,1]$ at its final output.

The key idea for the construction is to employ a tree structure; that is, the width of the network at $((m+1)+(m+4)\cdot j)^{\text{th}}$ hidden layer is doubled from that at $((m+1)+(m+4)\cdot(j-1))^{\text{th}}$ hidden layer for $j \in \{1, \ldots, p-1\}$ as

$$\begin{aligned} &\underbrace{\big\{\mathrm{Poly}_m^1(x), \ldots, \mathrm{Poly}_m^{2^{j-1}}(x)\big\}}_{((m+1)+(m+4)\cdot(j-1))^{\text{th}}\text{layer}} \\ &\quad \to \underbrace{\big\{\mathrm{Poly}_m^1(x), \ldots, \mathrm{Poly}_m^{2^{j-1}}(x), \mathrm{Mult}_m(x, \mathrm{Poly}_m^{2^{j-1}}(x)), \ldots, \tilde{f}_m\big(\mathrm{Poly}_m^{2^{j-1}}(x)\big)\big\}}_{((m+1)+(m+4)\cdot j)^{\text{th}}\text{ layer}}. \end{aligned} \tag{E.8}$$

The first 2^{j-1} input values in $((m+1)+(m+4)\cdot j)^{\text{th}}$ hidden layer is exactly copied from input values at the $((m+1)+(m+4)\cdot(j-1))^{\text{th}}$ hidden layer. The remaining 2^{j-1} input values in $((m+1)+(m+4)\cdot j)^{\text{th}}$ hidden layer approximates monomials $\{x^{2^{j-1}+1}, \dots, x^{2^j}\}$ through $\tilde{f}_m$ and Mult_m operations in Lemmas E.2.5 and E.2.6. The approximation error can be obtained via proof by induction. Readers can find the detailed proof in the Appendix E.3.4 with the exact descriptions on the construction of $\text{Poly}_m^{\{\text{N}\}}$.

Finally, we are ready to state Proposition E.2.8 on the construction of network $\tilde{f}$ which approximates $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)$.

Proposition E.2.8 *Let* $0 < \alpha < 1, m, N, M \in \mathbb{N}$ *with* $1 \leq N \leq d^{\alpha} + 1$. *For any function* $f \in W^r_{\infty}(\mathcal{S}^{d-1})$ *with* $r > 0$*, define* $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)$ *in* (E.7). *Then, there exists a network*

$$\tilde{f} \in \mathcal{F}\big(L, \big(d, 22NM, \dots, 22NM, 1\big), \mathcal{N}\big)$$

with depth $L = (m+4)\lceil \log_2(2N) \rceil$ *and number of parameters* $\mathcal{N} \leq M(2d + 404N \cdot (m + 3) + 2N + 4) + 1$ *such that*

$$\left\| \widehat{L}^{\boldsymbol{y}}_{N,M}(f) - \tilde{f} \right\|_{\infty} \leq C'_{\eta} \cdot \|f\|_{W^r_{\infty}(\mathcal{S}^{d-1})} \, d^{2N} \big(\log_2(2N)\big)^2 2^{-2m}, \tag{E.9}$$

where C'_{η} *is a positive constant depending on* η *and* α*, but not on* d, r, m, N, M *or* f.

A detailed proof for Proposition E.2.8 is deferred in the Appendix E.3.5.

Given the input data $\mathbf{x} \in \mathcal{S}^{d-1}$, recall the definition of $\widehat{L}^{\boldsymbol{y}}_{N,M}(f)(\mathbf{x})$ in (E.7). The crux of the whole construction procedure of our network is to build the sub-network which approximates $\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$ for each $i \in [M]$. The key observation is that $\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$ is the weighted sum of univariate polynomials of degree up to $2N$. Let $u_i = \langle \mathbf{x}, \mathbf{y}_i \rangle$. With the properly defined constant $\alpha_{i,q}$ (see its definition in the Appendix E.3.5), $\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$ can be re-written as $\xi_{N,r}(u_i) := \sum_{q=0}^{2N} \alpha_{i,q} |u_i|^q$. Since $|u_i| \in [0, 1]$, with the help of network constructed in Lemma E.2.7 with $P = \lceil \log_2(2N) \rceil$, the sub-network that approximates $\xi_{N,r}(u_i)$ is easily constructed. Recall this is enabled through the reproducing property of the kernel of $\mathcal{H}^d_k$ for $0 \leq K \leq 2N$.

E.3 Proofs of Statements in Appendix B and Corollary 5.2.3

E.3.1 Proof of Proposition E.2.3

Proposition E.3.1 *For $d \geq 5$, $1 \leq p \leq \infty$, and $s \geq \frac{3d-2}{4}$, the Sobolev space $W_p^s(\mathcal{S}^{d-1})$ is continuously embedded into $C(\mathcal{S}^{d-1})$, the space of continuous functions on $\mathcal{S}^{d-1}$, which implies*

$$\|f\|_\infty \leq c_0 \left(\frac{6}{\pi e}\right)^{\frac{d}{4}} \cdot \|f\|_{W_p^s(\mathcal{S}^{d-1})}, \qquad f \in W_p^s(\mathcal{S}^{d-1}),$$

where c_0 is an absolute constant independent of r, d, s, and f.

Proof. For $f \in W_p^s(\mathcal{S}^{d-1})$, by Sobolev embedding Lemma (see [195] Eq. 14, p. 420), the infinity norm can be bounded by the Sobolev norm as

$$\|f\|_\infty \leq C_{s,d} \cdot \|f\|_{W_2^s(\mathcal{S}^{d-1})}, \tag{E.10}$$

where the constant $C_{s,d}$ is defined with its square as

$$C_{s,d}^2 := \frac{1}{\omega_d} \sum_{k=0}^{\infty} \frac{\mathcal{N}(k,d)}{(k + \frac{d-2}{2})^{2s}} \tag{E.11}$$

with $\omega_d = 2\pi^{\frac{d}{2}}/\Gamma\left(\frac{d}{2}\right)$. Recalling (5.3), it is easy to see that by Stirling's formula, for large d, $\mathcal{N}(k,d) = (k + \frac{d-2}{2})^{d-2}\left(1 + \mathcal{O}\left(\frac{1}{d}\right)\right)$. Also, we have

$$\Gamma\left(\frac{d}{2}\right) = \frac{2}{d}\Gamma\left(\frac{d}{2}+1\right) = 2\sqrt{\frac{\pi}{d}}\left(\frac{d}{2e}\right)^{\frac{d}{2}}\left(1 + \mathcal{O}\left(\frac{1}{d}\right)\right). \tag{E.12}$$

When $s > \frac{d-1}{2}$, we have

$$\sum_{k=0}^{\infty}\left(k + \frac{d-2}{2}\right)^{d-2-2s} \leq \int_{\frac{d-2}{2}-1}^{\infty} t^{d-2-2s} dt = \frac{1}{2s+1-d}\left(\frac{d-2}{2} - 1\right)^{d-1-2s}.$$

Observe that $d \geq 5$, we have $\frac{d-2}{2} - 1 \geq \frac{d}{12}$. Thus, when $s \geq \frac{3d-2}{4}$, we have $2s+1-d \geq d/2$ and thereby (E.11) is bounded as

$$C_{s,d}^2 \leq \sqrt{\frac{\pi}{d}}\left(\frac{d}{2\pi e}\right)^{\frac{d}{2}}\frac{2}{d}\left(\frac{d}{12}\right)^{-\frac{d}{2}}\left(1+\mathcal{O}\left(\frac{1}{d}\right)\right) = \frac{2\sqrt{\pi}}{d\sqrt{d}}\left(\frac{6}{\pi e}\right)^{\frac{d}{2}}\left(1+\mathcal{O}\left(\frac{1}{d}\right)\right).$$

Then, there exists an absolute constant c_0 such that

$$C_{s,d}^2 \leq c_0^2\left(\frac{6}{\pi e}\right)^{\frac{d}{2}}, \qquad \forall d \geq 5.$$

This yields the claim. □

E.3.2 Proof of Lemma E.2.4

Lemma E.3.2 *Let $0 < r \leq \frac{3d-2}{4}$ and $0 < \alpha < 1$. If $f \in W_\infty^r(\mathcal{S}^{d-1})$, then for any $M \in \mathbb{N}$ and $1 \leq N \leq d^\alpha + 1$, there exist $\mathbf{y} = \{y_1, y_2, \ldots, y_M\} \subset \mathcal{S}^{d-1}$ such that*

$$\left\| L_N(f) - \widehat{L}_{N,M}^{\boldsymbol{y}}(f) \right\|_\infty \leq \frac{6 \cdot C'' \left(\frac{6}{\pi e}\right)^{\frac{d}{4}} \|f\|_{W_\infty^r(\mathcal{S}^{d-1})}\, d^{N+\frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}},$$

where $C'' > 0$ is a constant depending on α but independent of r, f, N, M, and d.

Proof. We recall the following probability inequality for random variables with values in a Hilbert space which can be found in [211].

Lemma E.3.3 *Let $(H, \|\cdot\|)$ be a Hilbert space and ξ be a random variable on $(Y, \rho_{\mathcal{X}})$ with values in H. Assume $\|\xi\| \leq \mathcal{M} < \infty$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}(\|\xi\|^2)$. Let $\{y_i\}_{i=1}^M$ be independent samples from $\rho_{\mathcal{X}}$. Then for any $0 < \delta < 1$, we have with probability at least $1 - \delta$,*

$$\left\| \frac{1}{M}\sum_{i=1}^{M} \xi(y_i) - \mathbb{E}(\xi) \right\|_H \leq \frac{2\mathcal{M}\log\left(\frac{2}{\delta}\right)}{M} + \sqrt{\frac{2\sigma^2(\xi)\log\left(\frac{2}{\delta}\right)}{M}}. \tag{E.13}$$

Let us define the random variable ξ on $(\mathcal{S}^{d-1}, \rho_{\mathcal{X}})$ with values in H given by

$$\xi(y) = F_r(y) \sum_{k=0}^{2N} (1+\lambda_k)^{-r/2} \eta\left(\frac{k}{N}\right) Z_k(y,\cdot), \qquad y \in \mathcal{S}^{d-1}. \tag{E.14}$$

To bound the norm $\|\xi\| = \|\xi(y)\|^2_{W_2^s}$, we set $s = \frac{3d-2}{4}$ and recall the norm of $W_2^s(\mathcal{S}^{d-1})$ given with $p = 2$ and for $y \in \mathcal{S}^{d-1}$,

$$\|\xi(y)\|_{W_2^s(\mathcal{S}^{d-1})} = \left\| F_r(y) \sum_{k=0}^{2N} (1+\lambda_k)^{\frac{s-r}{2}} \eta\left(\frac{k}{N}\right) Z_k(y,\cdot) \right\|_{L_2(\mathcal{S}^{d-1})}. \tag{E.15}$$

Recall $\lambda_k = k(k+d-2)$. Then, for $0 \le k \le 2N$, $d \ge 3$, we have $k^2 < 1 + \lambda_k \le dk^2$. We find $(1+\lambda_k)^{s-r} \le d^{s-r} k^{2(s-r)}$ by $s = \frac{3d-2}{4} \ge r$ ($\because$ $s - r \ge 0$). Also note that $0 \le \eta(t) \le 1$ for $t \in [0,2]$. Employing Stirling's formula $d! = \sqrt{2\pi d}\big(\frac{d}{e}\big)^d \big(1 + \mathcal{O}(1/d)\big)$ in the expression (5.3) for $\mathcal{N}(k,d)$ yields $\mathcal{N}(k,d) \le Cd^k$ for $0 \le k \le 2N$ and some constant C depending on α but independent of d. By using the identity $Z_k(y,y) = \mathcal{N}(k,d)$ (see Corollary 1.2.7. in [193]), $\|\xi\|^2_{W_2^s(\mathcal{S}^{d-1})}$ can be bounded as

$$\begin{aligned} F_r(y)^2 \cdot \sum_{k=0}^{2N} \big(1+\lambda_k\big)^{s-r} \eta^2\left(\frac{k}{N}\right) \mathcal{N}(k,d) &= F_r(y)^2 \cdot \left(1 + \sum_{k=1}^{2N} \big(1+\lambda_k\big)^{s-r} \eta^2\left(\frac{k}{N}\right) \mathcal{N}(k,d)\right) \\ &\le F_r(y)^2 \cdot \left(1 + C \cdot d^{s-r} \cdot \sum_{k=1}^{2N} k^{2(s-r)} d^k\right) \\ &\le F_r(y)^2 \cdot \left(1 + C \cdot d^{2N+s-r} \cdot \sum_{k=1}^{2N} k^{2(s-r)}\right), \end{aligned}$$

while the term $\sum_{k=1}^{2N} k^{2(s-r)}$ with $s - r \ge 0$ can be bounded as

$$\sum_{k=1}^{2N} k^{2(s-r)} \le \int_1^{2N+1} x^{2(s-r)} dx \le \frac{1}{2(s-r)+1} (2N+1)^{2(s-r)+1}.$$

Combining this with the definitions of the norm $\|f\|_{W_\infty^r(\mathcal{S}^{d-1})}$, we know that $\|\xi(y)\|^2_{W_2^s}$ can be bounded as

$$\|\xi(y)\|_{W_2^s}^2 \leq C'^2 \|f\|_{W_\infty^r(\mathcal{S}^{d-1})}^2 \cdot d^{2N+s-r}(2N+1)^{2(s-r)+1},$$

where C' is a constant depending on α but independent of r, s, f, N, and d. Thus the random variable ξ satisfies the condition $\|\xi\| \leq \mathcal{M} < \infty$ in Lemma E.3.3 with $\mathcal{M} = C' \|f\|_{W_\infty^r(\mathcal{S}^{d-1})} d^{N+\frac{s-r}{2}}(2N+1)^{(s-r)+\frac{1}{2}}$. So by Lemma E.3.3, with $\delta = \frac{1}{2}$ and $\sigma^2(\xi) \leq \mathcal{M}^2$, we know from the positive measure of the sample set that there exists a set of points $\mathbf{y} = \{y_i\}_{i=1}^M \in \mathcal{S}^{d-1}$ such that

$$\begin{aligned}\left\|\frac{1}{M}\sum_{i=1}^{M}\xi(y_i) - \mathbb{E}(\xi)\right\|_{\mathrm{H}} &= \left\|L_N(f) - \widehat{L}_{N,M}^{\boldsymbol{y}}(f)\right\|_{W_2^s(\mathcal{S}^{d-1})} \\ &\leq \frac{6 \cdot C' \|f\|_{W_\infty^r(\mathcal{S}^{d-1})} d^{N+\frac{s-r}{2}}(2N+1)^{(s-r)+\frac{1}{2}}}{\sqrt{M}}. \end{aligned} \tag{E.16}$$

Since $s = \frac{3d-2}{4}$, combining the result from Proposition E.2.3 with (E.16) yields

$$\left\|L_N(f) - \widehat{L}_{N,M}^{\boldsymbol{y}}(f)\right\|_\infty \leq \frac{6 \cdot C''\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} \|f\|_{W_\infty^r(\mathcal{S}^{d-1})} d^{N+\frac{3d-4r-2}{8}}(2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}},$$

where $C^{''} > 0$ is a constant depending on α but independent of r, f, N, M, and d. □

E.3.3 Proof of Lemma E.2.6

Lemma E.3.4 *For any positive integer $m \geq 1$, there exists a deep ReLU network*

$$\mathit{Mult}_m \in \mathcal{F}\big(m+3, \big(2, 10, \ldots, 10, 1\big)\big),$$

such that $\mathit{Mult}_m(x, y) \in [0, 1]$ and

$$|\mathit{Mult}_m(x, y) - xy| \leq 2^{-2m-1},$$

for all $x, y \in [0,1]$. *Moreover,* $\mathit{Mult}_m(x,0) = \mathit{Mult}_m(0,y) = 0$.

Proof. Given input (x,y), the network $\mathrm{Mult}_m(x,y)$ computes in the first hidden layer

$$(x,y) \to \left\{\sigma\Big(\frac{x+y}{2}\Big), \sigma\Big(-\Big(\frac{x+y}{2}\Big)\Big), \sigma\Big(\frac{x-y}{2}\Big), \sigma\Big(-\Big(\frac{x-y}{2}\Big)\Big)\right\}.$$

By using the equality $|x| = \sigma(x) + \sigma(-x)$ for $x \in [0,1]$, the network computes in the second hidden layer

$$(x,y) \to \left\{\sigma\Big(\Big|\frac{x+y}{2}\Big|\Big), \sigma\Big(\Big|\frac{x-y}{2}\Big|\Big)\right\}.$$

Note $\sigma\big(\big|\frac{x+y}{2}\big|\big), \sigma\big(\big|\frac{x-y}{2}\big|\big) \in [0,1]$, and $\sigma\big(\big|\frac{x+y}{2}\big|\big) = \big|\frac{x+y}{2}\big|, \sigma\big(\big|\frac{x-y}{2}\big|\big) = \big|\frac{x-y}{2}\big|$. We apply the network $\tilde{f}_m$ on the two components respectively. This gives a network of $(m+2)$ hidden layers with width vector $(2, 10, \ldots, 10, 2)$ that computes

$$(x,y) \to \left\{\sigma\Big(\tilde{f}_m\Big(\Big|\frac{x+y}{2}\Big|\Big)\Big), \sigma\Big(\tilde{f}_m\Big(\Big|\frac{x-y}{2}\Big|\Big)\Big)\right\}. \tag{E.17}$$

The network Mult_m computes (E.17) in the $(m+3)^{\text{th}}$ hidden layer. Since $\tilde{f}_m \in [0,1]$, $\sigma\big(\tilde{f}_m(x)\big) = \tilde{f}_m(x)$. In the output layer, the network value is computed as

$$\mathrm{Mult}_m(x,y) := \tilde{f}_m\Big(\Big|\frac{x+y}{2}\Big|\Big) - \tilde{f}_m\Big(\Big|\frac{x-y}{2}\Big|\Big). \tag{E.18}$$

Since $\tilde{f}_m$ is an increasing function in argument, $\mathrm{Mult}_m(x,y) \geq 0$, and since $\tilde{f}_m \in [0,1]$, $\mathrm{Mult}_m(x,y) \leq 1$. By identity, $xy = \big|\frac{x+y}{2}\big|^2 - \big|\frac{x-y}{2}\big|^2$, and Lemma E.2.5, the error is computed as follows:

$$\begin{aligned}|\mathrm{Mult}_m(x,y) - xy| &\leq \left|\tilde{f}_m\Big(\Big|\frac{x+y}{2}\Big|\Big) - \Big(\Big|\frac{x+y}{2}\Big|\Big)^2\right| + \left|\tilde{f}_m\Big(\Big|\frac{x-y}{2}\Big|\Big) - \Big(\Big|\frac{x-y}{2}\Big|\Big)^2\right| \\ &\leq 2^{-2m-1}.\end{aligned}$$

If either $x = 0$ or $y = 0$, by the definition of (E.18), we have $\text{Mult}_m(x, 0) = \text{Mult}_m(0, y) = 0$. □

E.3.4 Proof of Lemma E.2.7

Lemma E.3.5 *For any positive integer $m \geq 1, N \geq 2$ and for $P = \lceil \log_2(N) \rceil$, there exists a deep ReLU network*

$$\textit{Poly}_m^{\{N\}} \in \mathcal{F}\big(L, \big(1, 11N, \ldots, 11N, 2^P\big), \mathcal{N}\big),$$

with the depth $L = m + (m+4)\big(\lceil \log_2(N) \rceil - 1\big)$ and the number of parameters $\mathcal{N} \leq 202N \cdot (m+3)$ such that $\textit{Poly}_m^{\{N\}}(x) \in [0,1]^{2^P}$ and

$$\big|\textit{Poly}_m^j(x) - x^j\big| \leq P^2 \cdot 2^{-2m-1} \quad \textit{for all} \quad j \in \{1, \ldots, 2^P\}$$

for all $x \in [0,1]$.

Proof. Let us describe the construction of the network $\text{Poly}_m^{\{\text{N}\}}$. With the application of Lemma E.2.5, in the $(m+1)^{\text{th}}$ hidden layer, the network computes

$$x \to \big\{\sigma(x), \sigma(\tilde{f}_m(x))\big\}$$

with the width $\mathbf{p} = (1, 5, \ldots, 5, 2)$. For approximating x^3, the network Mult_m is applied on the pair $(\sigma(x), \sigma(\tilde{f}_m(x)))$, and for approximating x^4, the network $\tilde{f}_m$ is applied on the $\sigma(\tilde{f}_m(x))$. Therefore, in the $\{(m+1) + (m+4)\}^{\text{th}}$ hidden layer, the network $\text{Poly}_m^{\{\text{N}\}}$ computes

$$x \to \Big\{\sigma(x), \sigma(\tilde{f}_m(x)), \sigma\big(\text{Mult}_m(x, \tilde{f}_m(x))\big), \sigma\big(\tilde{f}_m\big(\tilde{f}_m(x)\big)\big)\Big\}. \tag{E.19}$$

Note that each component in the hidden layer is in $[0,1]$ by Lemmas E.2.5 and E.3.4. This procedure is continued until a following vector is in the final output layer,

$$x \to \left\{ \mathrm{Poly}_m^1(x), \ldots, \mathrm{Poly}_m^{2^{P-1}}(x), \mathrm{Mult}_m(x, \mathrm{Poly}_m^{2^{P-1}}(x)), \ldots, \tilde{f}_m\big(\tilde{f}_m \ldots \big(\tilde{f}_m(x)\big)\big) \right\} \in [0,1]^{2^P}.$$

The resulting network is referred as $\mathrm{Poly}_m^{\{\mathrm{N}\}}$ and has $m + (m+4)\big(\lceil \log_2(N) \rceil - 1\big)$ hidden layers. Recall $P = \lceil \log_2(N) \rceil$. By the construction procedure of the network, we can compute the upper bound of maximum width as,

$$2^{\lceil \log_2(N) \rceil - 1} + \left\{ 10 \cdot \left(2^{\lceil \log_2(N) \rceil - 1} - 1 \right) + 5 \right\} \leq 11 \cdot 2^{\lceil \log_2(N) \rceil - 1} \leq 11N, \tag{E.20}$$

where we use $\lceil \log_2(N) \rceil \leq \log_2(N) + 1$ in the second inequality. Now, we need to count the number of active parameters in the network. For $k \in \{1, \ldots, \lceil \log_2(N) \rceil\}$, we compute the upper bound on the total number of active parameters in-between following hidden layers:

$$\left\{ \mathrm{Poly}_m^1(x), \ldots, \mathrm{Poly}_m^{2^{k-1}}(x) \right\} \to \left\{ \mathrm{Poly}_m^1(x), \ldots, \mathrm{Poly}_m^{2^{k-1}}(x), \mathrm{Poly}_m^{2^{k-1}+1}(x), \ldots, \mathrm{Poly}_m^{2^k}(x) \right\}. \tag{E.21}$$

Think of a network which takes the hidden layer in the left hand side of (E.21) as an input, and gives the hidden layer in the right hand side of (E.21) as an output. It is easy to count the number of active parameters in input, hidden, and output layers, separately as follows:

$$\begin{cases} \text{Input layer} & : 2^{k-1} + 1 + 2 \cdot \big(2^{k-1} - 1\big) = 3 \cdot 2^{k-1} - 1. \\ \text{Hidden layers} & : (m+2) \cdot 2^{k-1} + 100 \cdot (m+2) \cdot (2^{k-1} - 1) + 25 \cdot (m+2) \\ & \qquad\qquad = (m+2)(101 \cdot 2^{k-1} - 75). \\ \text{Output layer} & : 2^{k-1} + 10 \cdot (2^{k-1} - 1) + 5 = 11 \cdot 2^{k-1} - 5. \end{cases}$$

Since the k runs over $\{1, \ldots, \lceil \log_2(N) \rceil\}$, the total number of active parameters can be bounded as:

$$
\begin{aligned}
&\sum_{k=1}^{\lceil \log_2(N)} \left\{ \left(m+2 \right) \left(101 \cdot 2^{k-1} - 75 \right) + \left(14 \cdot 2^{k-1} - 6 \right) \right\} \\
&\leq (m+2) \cdot 101 \sum_{k=1}^{\lceil \log_2(N)} 2^{k-1} + 14 \cdot \sum_{k=1}^{\lceil \log_2(N)} 2^{k-1} \\
&\leq 202N \cdot (m+3).
\end{aligned}
$$

The approximation error is proved via induction on the number of iterated multiplications $P = \lceil \log_2(N)$. For $P = 1$, that is $N = 2$, we have

$$\left| x^2 - \tilde{f}_m(x) \right| \leq 2^{-2m-1}$$

by Lemma E.2.5. For the convenience of notation, denote $\tilde{x}^a := \text{Poly}_m^a(x)$ for some positive integer a. For $P = k-1$, assume a following holds

$$\left| x^j - \tilde{x}^j \right| \leq 3^{k-2} \cdot 2^{-2m-1} \quad \text{for} \quad j \in \{1, \ldots, 2^{k-1}\}.$$

Then, for $P = k$, we want to prove

$$\left| x^j - \tilde{x}^j \right| \leq 3^{k-1} \cdot 2^{-2m-1} \quad \text{for} \quad j \in \{1, \ldots, 2^{k}\}.$$

By the construction of neural network and induction assumption, for $j \in \{1, \ldots, 2^{k-1}\}$, we have $|x^j - \tilde{x}^j| \leq 3^{k-2} \cdot 2^{-2m-1} \leq 3^{k-1} \cdot 2^{-2m-1}$. For any $j \in \{2^{k-1}+1, \ldots, 2^k\}$, find any $a, b \in \{1, \ldots, 2^{k-1}\}$ such that $j = a + b$. Then, for $x \in [0, 1]$,

$$
\begin{aligned}
\left| x^{a+b} - \text{Mult}_m\left(\tilde{x}^a, \tilde{x}^b\right) \right| &\leq \left| x^{a+b} - \tilde{x}^a \cdot \tilde{x}^b \right| + \left| \tilde{x}^a \cdot \tilde{x}^b - \text{Mult}_m\left(\tilde{x}^a, \tilde{x}^b\right) \right| \\
&\leq x^a \left| x^b - \tilde{x}^b \right| + \tilde{x}^b \left| x^a - \tilde{x}^a \right| + \left| \tilde{x}^a \cdot \tilde{x}^b - \text{Mult}_m\left(\tilde{x}^a, \tilde{x}^b\right) \right| \\
&\leq 3^{k-2} \cdot 2^{-2m-1} + 3^{k-2} \cdot 2^{-2m-1} + 2^{-2m-1} \leq 3^{k-1} \cdot 2^{-2m-1}.
\end{aligned}
$$

By using the fact $\log_2(3) < 2$, we can deduce $3^{k-1} < P^2$ and conclude the proof. □

E.3.5 Proof of Proposition E.2.8

Proposition E.3.6 *Let* $0 < \alpha < 1, m, N, M \in \mathbb{N}$ *with* $1 \leq N \leq d^\alpha + 1$. *For any function* $f \in W_\infty^r(\mathcal{S}^{d-1})$ *with* $r > 0$, *define* $\widehat{L}_{N,M}^{\boldsymbol{y}}(f)$ *in* (E.7). *Then, there exists a network*

$$\tilde{f} \in \mathcal{F}\big(L, \big(d, 22NM, \ldots, 22NM, 1\big), \mathcal{N}\big)$$

with depth $L = (m+4)\lceil \log_2(2N)$ *and number of parameters* $\mathcal{N} \leq M(2d + 404N \cdot (m + 3) + 2N + 4) + 1$ *such that*

$$\left\| \widehat{L}_{N,M}^{\boldsymbol{y}}(f) - \tilde{f} \right\|_\infty \leq C_\eta' \cdot \|f\|_{W_\infty^r(\mathcal{S}^{d-1})} \, d^{2N} \big(\log_2(2N)\big)^2 2^{-2m}, \tag{E.22}$$

where C_η' *is a positive constant depending on* η *and* α, *but not on* d, r, m, N, M *or* f.

Proof. We adopt the shorthand notation denoting $[n] := \{1, 2, \ldots, n\}$ and $[n]_0 := \{0, 1, \ldots, n\}$ for $n \in \mathbb{N}$ in the proof.

Given the input data $\mathbf{x} \in \mathcal{S}^{d-1}$, recall the definition of $\widehat{L}_{N,M}^{\boldsymbol{y}}(f)(\mathbf{x})$ in (E.7). The crux of the whole construction procedure is to build the the sub-network which approximates $\xi_{N,r}(\langle \mathbf{x}, \mathbf{y}_i \rangle)$ for each $i \in [M]$. First, observe that, by (5.6) and (E.6), $\xi_{N,r}(u_i)$ can be written as:

$$\xi_{N,r}(u_i) = \sum_{k=0}^{2N} (1 + \lambda_k)^{-\frac{r}{2}} \eta\left(\frac{k}{N}\right) \left\{ \frac{k + \lambda_{\mathrm{G}}}{\lambda_{\mathrm{G}}} \sum_{\ell=0}^{\lfloor \frac{k}{2} \rfloor} (-1)^\ell \frac{\Gamma\big(k - \ell + \lambda_{\mathrm{G}}\big)}{\Gamma\big(\lambda_{\mathrm{G}}\big)\ell!\big(k - 2\ell\big)!} \big(2u_i\big)^{k-2\ell} \right\}, \tag{E.23}$$

for $i \in [M]$. The key observation is that Eq. (E.23) is the weighted sum of univariate polynomials of degree up to $2N$. We define a constant $c_{k,\ell,\eta,\lambda_k,r,d}$ as

$$c_{k,\ell,\eta,\lambda_k,r,d} := (1+\lambda_k)^{-\frac{r}{2}}\eta\left(\frac{k}{N}\right)\frac{k+\lambda_{\mathrm{G}}}{\lambda_{\mathrm{G}}}\frac{(-1)^{\ell}\Gamma\big(k-\ell+\lambda_{\mathrm{G}}\big)2^{k-2\ell}}{\Gamma\big(\lambda_{\mathrm{G}}\big)\ell!\big(k-2\ell\big)!}. \tag{E.24}$$

For $i \in \{1,\dots,M\}$, set $\alpha_{i,q}$ as

$$\alpha_{i,q} = \begin{cases} \sum_{(k,\ell)\in\mathcal{A}_q}\big(-c_{k,\ell,\eta,\lambda_k,r,d}\big) & \text{if } u_i < 0 \text{ and } q \text{ is odd,} \\ \sum_{(k,\ell)\in\mathcal{A}_q}\big(c_{k,\ell,\eta,\lambda_k,r,d}\big) & \text{otherwise,} \end{cases} \tag{E.25}$$

where for each $q \in \{0,\dots,2N\}$, the set $\mathcal{A}_q$ is given by $\mathcal{A}_q := \{(k,\ell) \in [2N]_0 \times [\lfloor k/2 \rfloor]_0 : k-2\ell = q\}$. Then, (E.23) can be re-written as $\xi_{N,r}(u_i) := \sum_{q=0}^{2N}\alpha_{i,q}|u_i|^q$.

1. The Network Construction. Now, we are ready for the construction of $\tilde{f}$. Through Lemma E.2.4, we know that there exists $\boldsymbol{y} = \{\mathbf{y}_1,\dots,\mathbf{y}_M\}$ that satisfies the bound (E.2.4). Then, for each $i \in [M]$, we put $\mathbf{y}_i \in \mathcal{S}^{d-1}$ as a weight vector that connects input $\mathbf{x}$ to the $(2i-1)^{\text{th}}$ and $(2i)^{\text{th}}$ nodes in the first hidden layer. Through this, $\tilde{f}$ computes in its first hidden layer $\mathbf{x} \to \big\{\sigma\big(\langle\mathbf{x},\mathbf{y}_1\rangle\big),\sigma\big(-\langle\mathbf{x},\mathbf{y}_1\rangle\big),\dots,\sigma\big(\langle\mathbf{x},\mathbf{y}_M\rangle\big),\sigma\big(-\langle\mathbf{x},\mathbf{y}_M\rangle\big)\big\} \in [0,1]^{2M}$. Then, by the identity $|x| = \sigma(x)+\sigma(-x)$ for $x \in \mathbb{R}$, the network computes in its second hidden layer $\mathbf{x} \to \big\{\sigma\big(|u_1|\big),\sigma\big(|u_2|\big),\dots,\sigma\big(|u_M|\big)\big\} \in [0,1]^M$, where $u_i := \langle\mathbf{x},\mathbf{y}_i\rangle \in [-1,1]$ for $i \in [M]$. Since $\sigma(|u_i|) = |u_i| \in [0,1]$, $\text{Poly}_m^{\{2\text{N}\}}$ with $P = \lceil\log_2(2N)$ is applicable for each $\{|u_i|\}_{i=1}^M$, and it generates $\text{Poly}_m^q(|u_i|)$ with q at most $4N$. Set $\mathcal{B}_{\max} := \max_{i=1,\dots,M}\Big|\sum_{q=0}^{2N}\alpha_{i,q}\cdot\text{Poly}_m^q(|u_i|)\Big|$. Using the definition of the constant $\alpha_{i,q}$, the network $\tilde{f}$ computes in the $(m+4)\lceil\log_2(2N)$th hidden layer $\{\sigma(\sum_{q=0}^{2N}\alpha_{1,q}\text{Poly}_m^q(|u_1|)+2\mathcal{B}_{\max}),\dots,\sigma(\sum_{q=0}^{2N}\alpha_{M,q}\text{Poly}_m^q(|u_M|)+2\mathcal{B}_{\max})\} \in \mathbb{R}^M$. By the definition of $\mathcal{B}_{\max}$, it is easy to see each component in the hidden layer is positive. Set the weight of output layer as $\{\frac{1}{M}F_r(\mathbf{y}_i)\}_{i=1}^M$. Define $\mathcal{L}(|u_i|) := \sum_{q=0}^{2N}\alpha_{i,q}\cdot\text{Poly}_m^q(|u_i|)+2\cdot\mathcal{B}_{\max}$. Then, given the data $\mathbf{y} = \{\mathbf{y}_1,\dots,\mathbf{y}_M\}$, the network $\tilde{f}$ computes its final output as $\tilde{f}(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^M F_r(\mathbf{y}_j)\cdot\big(\mathcal{L}(|\langle\mathbf{x},\mathbf{y}_j\rangle|)-2\mathcal{B}_{\max}\big) := \frac{1}{M}\sum_{i=1}^M F_r(\mathbf{y}_{\mathrm{i}})\cdot\mathcal{L}\big(\xi_{N,r}\big)\big(\langle\mathbf{x},\mathbf{y}_{\mathrm{i}}\rangle\big)$.

2. The Width and Number of Active Parameters of $\tilde{f}$. By the construction of network $\tilde{f}$ and the result of Lemma E.2.7, it is easy to see the maximum width of the network is $22NM$. Now, we work on counting the number of active parameters in the network as

$$\begin{cases} \text{From Input to } 2^{\text{nd}} \text{ hidden layer} & : \quad 2Md + 2M. \\ \text{From } 2^{\text{nd}} \text{ to } \big((m+4)\lceil \log_2(2N) - 1\big)^{\text{th}} \text{ hidden layer} & : \quad 404NM \cdot (m+3). \\ \text{From } \big((m+4)\lceil \log_2(2N) - 1\big)^{\text{th}} \text{ hidden layer to output layer} & : \quad (2N+1)M + M + 1. \end{cases}$$

Summing up the total number yields the desired result.

3. Approximation Error Computation. A remaining thing is to calculate the approximation error:

$$\begin{aligned} \left\| \widehat{L}^{\boldsymbol{y}}_{N,M}(f) - \tilde{f} \right\|_\infty &= \sup_{\mathbf{x} \in \mathcal{S}^{d-1}} \left| \frac{1}{M} \sum_{i=1}^{M} F_r(\mathbf{y_i}) \cdot \xi_{N,r}(\langle \mathbf{x}, \mathbf{y_i} \rangle) - \frac{1}{M} \sum_{i=1}^{M} F_r(\mathbf{y_i}) \cdot \mathcal{L}\big(\xi_{N,r}\big)\big(\langle \mathbf{x}, \mathbf{y_i} \rangle\big) \right| \\ &\le \|f\|_{W^r_\infty(\mathcal{S}^{d-1})} \cdot \big\| \xi_{N,r} - \mathcal{L}\big(\xi_{N,r}\big) \big\|_\infty . \end{aligned} \tag{E.26}$$

Recall the definition of $\alpha_{i,q}$ in (E.25). Using Stirling's Formula, $\Gamma(n+1) = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + \mathcal{O}(1/n))$, we observe the behavior of Gegenbauer coefficient in (5.6) where $\lambda_{\text{G}} = \frac{d-2}{2} \gg d^\alpha + 1 \ge N$, and find that it can be bounded as $C \cdot \lambda_{\text{G}}^{k-\ell} \cdot 2^{k-2\ell}(1 + \mathcal{O}(1/d))$, where $C > 0$ is a constant independent of d.

For $k \in \{0, 1, \dots, 2N\}$, combining the facts $(1 + \lambda_k)^{-\frac{r}{2}} < 1$, $\eta(\cdot) \le 1$, $\frac{k + \lambda_{\text{G}}}{\lambda_{\text{G}}} \le 2$ for $k \le 2N \le 2(d^\alpha + 1)$ with $\lambda_{\text{G}} = \frac{d-2}{2}$ yields

$$\left| c_{k,\ell,\eta,\lambda_k,r,d} \right| \le C'_\eta \cdot 2^{-\ell} \cdot d^{k-\ell}, \tag{E.27}$$

where $C'_\eta > 0$ is a constant dependent on α and η. Recall $\mathcal{L}\big(\xi_{N,r}\big)\big(\langle \mathbf{x}, \mathbf{y}_j \rangle\big) := \sum_{q=0}^{2N} \alpha_{j,q} \cdot$

$\text{Poly}_m^q(|\langle \mathbf{x}, \mathbf{y}_j\rangle|)$ and note that $\sum_{q=0}^{2N} |\alpha_{j,q}| = \sum_{k=0}^{2N} \sum_{\ell=0}^{\lfloor \frac{k}{2} \rfloor} \left|c_{k,\ell,\eta,\lambda_k,r,d}\right|$. Then, we have

$$\begin{aligned}\left\|\xi_{N,r} - \mathcal{L}\big(\xi_{N,r}\big)\right\|_\infty &\leq \left(\sum_{k=0}^{2N} \sum_{\ell=0}^{\lfloor \frac{k}{2} \rfloor} \left|c_{k,\ell,\eta,\lambda_k,r,d}\right|\right) \cdot \left(\sup_{u\in[0,1]} \max_{q\in\{0,\dots,2N\}} |u^q - \text{Poly}_m^q(u)|\right) \\ &\leq C_\eta^{'} \cdot \left(\sum_{k=0}^{2N} d^k \sum_{\ell=0}^{\lfloor \frac{k}{2} \rfloor} \frac{1}{(2d)^\ell}\right) \cdot \left(\big(\log_2(2N)\big)^2 \cdot 2^{-2m-1}\right)\end{aligned}$$

where we used the result from Lemma E.2.7 and (E.27) in the second inequality. Using $\sum_{\ell=0}^{\lfloor \frac{k}{2} \rfloor} \frac{1}{(2d)^\ell} \leq 2$ in the last inequality yields the claim. □

E.3.6 Proof of Corollary 5.2.3

Corollary E.3.7 *Let* $0 < \alpha, \beta, \gamma < 1$ *with* $\gamma > \max\{\alpha, \beta\}$ *and* $N \in \mathbb{N}$ *with* $1 \leq N \leq d^\alpha + 1$. *For any* $f \in W_\infty^r(\mathcal{S}^{d-1})$ *with* $r > 0$, *we have :*

(I) For $\frac{3d-2}{4} - C_1 \leq r \leq \frac{3d-2}{4}$ *with some constant* $C_1 \geq 0$ *independent of* d, *there exists a network*

$$\tilde{f}^{(I)} \in \mathcal{F}\left(L, (d, 66N, 66N, \dots, 66N, 1), \mathcal{N}\right)$$

with depth $L = \mathcal{O}\left(d^\gamma \log_2 d\right)$ *and the number of active parameters* $\mathcal{N} = \mathcal{O}\left(d^{\max\{\alpha+\gamma,1\}}\right)$, *such that* $\left\|f - \tilde{f}^{(I)}\right\|_\infty \leq C'_{\eta,\alpha,\beta,\gamma} \|f\|_{W_\infty^r(\mathcal{S}^{d-1})} d^{-d^\beta}$, *where* $C'_{\eta,\alpha,\beta,\gamma}$ *is a constant depending only on* $C_1, \eta, \alpha, \beta, \gamma$.

(II) For $r = \mathcal{O}(1)$ *and* $M = \mathcal{O}\left(9^d d^{\frac{9}{4}d}\right)$, *there exists a network*

$$\tilde{f}^{(II)} \in \mathcal{F}\big(L, \big(d, 22NM, \dots, 22NM, 1\big), \mathcal{N}\big)$$

with depth $L = \mathcal{O}\left(d^\gamma \log_2 d\right)$ *and the number of active parameters* $\mathcal{N} = \mathcal{O}\left(9^d d^{\frac{13}{4}d}\right)$ *such that* $\left\|f - \tilde{f}^{(II)}\right\|_\infty \leq C'_{\eta,\alpha,\beta,\gamma} \|f\|_{W_\infty^r(\mathcal{S}^{d-1})} d^{-\alpha r}$, *where* $C'_{\eta,\alpha,\beta,\gamma}$ *is a constant depending only on* $\eta, \alpha, \beta, \gamma$.

Proof. By the results of Theorem 5.2.1, for $1 \leq N \leq d^\alpha + 1$, we have the following inequality on the approximation error

$$\left\|\tilde{f} - f\right\|_{\infty} \leq C_{\eta}^{''} \left\|f\right\|_{W_{\infty}^{r}(\mathcal{S}^{d-1})} \times \\ \max\left\{N^{-r}, \frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{N+\frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}}, d^{2N}\big(\log_2(2N)\big)^2 2^{-2m}\right\}, \tag{E.28}$$

where $C_{\eta}^{''}$ is a constant dependent on η, and independent on d, r, N, M or f. We divide the proof into two cases.

(I) $r = \mathcal{O}(d)$ ***and any integer*** $M \geq 1$

For the first term in (E.28), since $N = \lceil d^{\alpha}$, we know that $N^{-r} = \mathcal{O}(d^{-\alpha r}) = \mathcal{O}(d^{-d^{\beta}})$ with any $0 < \beta < 1$. This is due to the assumption that $\frac{3d-2}{4} - C_1 \leq r \leq \frac{3d-2}{4}$, which implies $d = \mathcal{O}(r)$ and $d^{\beta} = o(r)$.

For the second term in (E.28), since $N = \lceil d^{\alpha}$ with $0 < \alpha < 1$, we know that it is bounded by

$$\frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{N+\frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}} \leq \frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{d^{\alpha}+\frac{3d-4r+6}{8}} (3d^{\alpha})^{\frac{3d-4r}{4}}}{\sqrt{M}}. \tag{E.29}$$

As $\frac{3d-2}{4} - C_1 \leq r \leq \frac{3d-2}{4}$, we know the term on the right hand side of (E.29) can be written as $\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{d^{\alpha}+\mathcal{O}(1)} 3^{\mathcal{O}(1)}/\sqrt{M}$. To show that the bound is of order $\mathcal{O}(d^{-d^{\beta}})$, we multiply the bound by $d^{d^{\beta}}$, take the logarithm, and find that for any $0 < \alpha, \beta < 1$,

$$\log\left(\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{d^{\alpha}+d^{\beta}+\mathcal{O}(1)}\right) \leq \frac{d}{4}\log\left(\frac{6}{\pi e}\right) + \big(d^{\alpha} + d^{\beta} + \mathcal{O}(1)\big)\log(d) \to -\infty,$$

as $d \to \infty$. Hence, there exists a constant $C_{\alpha,\beta} > 0$ depending only on C_1, α, β such that

$$\frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{N+\frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}} \leq C_{\alpha,\beta} d^{-d^{\beta}},$$

for any fixed $M \in \mathbb{N}$. In our proof, we simply choose $M = 3$. For the third term in (E.28), take $m = \lceil d^\gamma \rceil$ with $\max\{\alpha, \beta\} < \gamma < 1$, then there exists a constant $C_{\alpha,\beta,\gamma}$ depending on α, β, γ such that

$$d^{2N}\big(\log_2(2N)\big)^2 2^{-2m} < d^{2d^\alpha+2}\big(2+\log_2(d)\big)^2 2^{-2m} \le d^{3d^\alpha} 2^{-2m} \le C_{\alpha,\beta,\gamma} d^{-d^\beta},$$

where $\log_2(2d^\alpha + 2) \le \log_2(4d^\alpha) < 2 + \log_2(d)$ is used in the first inequality, and the last inequality follows from the same argument as above, of multiplying with d^{d^β} and taking the logarithm. Combining all the analysis above, we have

$$\left\|\tilde{f} - f\right\|_\infty \le C'_{\eta,\alpha,\beta,\gamma} \|f\|_{W^r_\infty(\mathcal{S}^{d-1})} d^{-d^\beta},$$

where $C'_{\eta,\alpha,\beta,\gamma} > 0$ is a constant dependent on $\eta, \alpha, \beta, \gamma$, and C_1.

Recall from Proposition E.2.8, $\tilde{f}$ is a network with depth $L = (m+4)\lceil \log_2(2N) \rceil$ and number of parameters $\mathcal{N} \le M(2d + 404N \cdot (m+3) + 2N + 4) + 1$. By simply plugging-in $m = \lceil d^\gamma \rceil$, $N = \lceil d^\alpha \rceil$ and $M = 3$, we have $L = \mathcal{O}(d^\gamma \log_2(d))$ and $\mathcal{N} = \mathcal{O}\left(d^{\max\{\alpha+\gamma, 1\}}\right)$.

(II) $r = \mathcal{O}(1)$ ***and*** $M = \mathcal{O}(d^d)$.

For the first term in (E.28), since $N = \lceil d^\alpha \rceil$, we know that $N^{-r} = \mathcal{O}(d^{-\alpha r})$.

For the second term in (E.28), since $N = \lceil d^\alpha \rceil$ with $0 < \alpha < 1$, we know that it is bounded by

$$\frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{N+\frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}} \le \frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{d^\alpha+\frac{3d-4r+6}{8}} (3d^\alpha)^{\frac{3d-4r}{4}}}{\sqrt{M}} \le \frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{d^\alpha+\frac{9}{8}d} 3^d}{\sqrt{M}}. \tag{E.30}$$

Take $M = \mathcal{O}(9^d d^{\frac{9}{4}d})$, multiply the bound (E.30) by d^{d^β}, take the logarithm, and find that for any $0 < \alpha, \beta < 1$,

$$\log\left(\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{d^\alpha + d^\beta + \mathcal{O}(1)}\right) \le \frac{d}{4}\log\left(\frac{6}{\pi e}\right) + \left(d^\alpha + d^\beta + \mathcal{O}(1)\right)\log(d) \to -\infty,$$

as $d \to \infty$. Hence, there exists a constant $C_{\alpha,\beta} > 0$ depending only on C_1, α, β such that

$$\frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{N+\frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}} \leq C_{\alpha,\beta} d^{-d^{\beta}} \leq C_{\alpha,\beta} d^{-\alpha r},$$

for $M = \mathcal{O}(9^d d^{\frac{9}{4}d})$. For the third term in (E.28), take $m = \lceil d^{\gamma} \rceil$ with $\max\{\alpha, \beta\} < \gamma < 1$, then there exists a constant $C_{\alpha,\beta,\gamma}$ depending on α, β, γ such that

$$d^{2N}\big(\log_2(2N)\big)^2 2^{-2m} < d^{2d^{\alpha}+2}\big(2 + \log_2(d)\big)^2 2^{-2m} \leq d^{3d^{\alpha}} 2^{-2m} \leq C_{\alpha,\beta,\gamma} d^{-d^{\beta}} \leq C_{\alpha,\beta,\gamma} d^{-\alpha r},$$

where $\log_2(2d^{\alpha} + 2) \leq \log_2(4d^{\alpha}) < 2 + \log_2(d)$ is used in the first inequality, and the last inequality follows from the same argument as above, of multiplying with $d^{d^{\beta}}$ and taking the logarithm. Combining all the analysis above, we have

$$\left\|\tilde{f} - f\right\|_{\infty} \leq C'_{\eta,\alpha,\beta,\gamma} \|f\|_{W^r_{\infty}(\mathcal{S}^{d-1})} d^{-\alpha r},$$

where $C'_{\eta,\alpha,\beta,\gamma} > 0$ is a constant dependent on $\eta, \alpha, \beta, \gamma$, and C_1.

Recall from Proposition E.2.8, $\tilde{f}$ is a network with depth $L = (m+4)\lceil \log_2(2N) \rceil$ and number of parameters $\mathcal{N} \leq M(2d + 404N \cdot (m+3) + 2N + 4) + 1$. By simply plugging-in $m = \lceil d^{\gamma} \rceil$, $N = \lceil d^{\alpha} \rceil$ and $M = \mathcal{O}(9^d d^{\frac{9}{4}d})$, we have $L = \mathcal{O}(d^{\gamma} \log_2(d))$ and $\mathcal{N} = \mathcal{O}\left(9^d d^{\frac{13}{4}d}\right)$. □

E.4 Proofs of Proposition 5.3.2, Theorem 5.3.3 and Theorem 5.3.4

E.4.1 Proof of Proposition 5.3.2

Proposition E.4.1 *Set $\delta \in (0,1)$. Then, with probability at least $1 - \delta$, we have*

$$\mathcal{E}\big(\pi_B \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big) \leq C_{B,\delta,f} \cdot \left(\frac{\textit{Pdim}(\mathcal{F}) \cdot \log(n)}{n} + \frac{\|f - f_\rho\|_\infty}{\sqrt{n}} + \|f - f_\rho\|_\infty^2\right), \quad \text{(E.31)}$$

where $C_{B,\delta,f}$ is an absolute constant dependent on B, δ, f independent on n, r, d.

Proof. Since $\widehat{f}_n$ is an empirical risk minimizer in (5.14), we have $\mathcal{E}_D(\widehat{f}_n) \leq \mathcal{E}_D(f)$ for any fixed $f \in \mathcal{F}$ and $\mathcal{E}_D(\pi_B\widehat{f}_n) \leq \mathcal{E}_D(\widehat{f}_n)$. Then, we have a following decomposition:

$$\begin{aligned}
\mathcal{E}\big(\pi_B\widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big) &= \Bigg(\big\{\mathcal{E}\big(\pi_B\widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big)\big\} - \big\{\mathcal{E}_D\big(\pi_B\widehat{f}_n\big) - \mathcal{E}_D\big(f_\rho\big)\big\}\Bigg) \\
&\quad + \Bigg(\big\{\mathcal{E}_D\big(\pi_B\widehat{f}_n\big) - \mathcal{E}_D\big(f_\rho\big)\big\} - \big\{\mathcal{E}_D\big(f\big) - \mathcal{E}_D\big(f_\rho\big)\big\}\Bigg) \\
&\quad + \Bigg(\big\{\mathcal{E}_D\big(f\big) - \mathcal{E}_D\big(f_\rho\big)\big\} - \big\{\mathcal{E}\big(f\big) - \mathcal{E}\big(f_\rho\big)\big\}\Bigg) + \Bigg(\mathcal{E}\big(f\big) - \mathcal{E}\big(f_\rho\big)\Bigg) \\
&\leq \Bigg(\big\{\mathcal{E}\big(\pi_B\widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big)\big\} - \big\{\mathcal{E}_D\big(\pi_B\widehat{f}_n\big) - \mathcal{E}_D\big(f_\rho\big)\big\}\Bigg) \qquad \text{(E.32)} \\
&\quad + \Bigg(\big\{\mathcal{E}_D\big(f\big) - \mathcal{E}_D\big(f_\rho\big)\big\} - \big\{\mathcal{E}\big(f\big) - \mathcal{E}\big(f_\rho\big)\big\}\Bigg) + \Bigg(\mathcal{E}\big(f\big) - \mathcal{E}\big(f_\rho\big)\Bigg).
\end{aligned}$$

Let $\mathcal{F}_B := \{\pi_B f : \forall f \in \mathcal{F}\}$ and define two quantities:

$$\begin{aligned}
\mathcal{S}_1(n, \mathcal{F}_B) &:= \big\{\mathcal{E}\big(f\big) - \mathcal{E}\big(f_\rho\big)\big\} - \big\{\mathcal{E}_D\big(f\big) - \mathcal{E}_D\big(f_\rho\big)\big\} \quad \forall f \in \mathcal{F}_B, \\
\mathcal{S}_2(n, \mathcal{F}) &:= \big\{\mathcal{E}_D\big(f\big) - \mathcal{E}_D\big(f_\rho\big)\big\} - \big\{\mathcal{E}\big(f\big) - \mathcal{E}\big(f_\rho\big)\big\} \quad \forall f \in \mathcal{F}.
\end{aligned}$$

Step 1 : Control $\mathcal{S}_1(n, \mathcal{F}_B)$. The following concentration inequality is needed for controlling the term.

Lemma E.4.2 *[Theorem 11.4 of [212]] Assume $|y| \leq B$ almost surely and $B \geq 1$. Let $\alpha, \beta > 0$ and $0 < \varepsilon \leq 1/2$. If $\mathcal{F}^{'}$ is a set of functions $f : \mathbb{R}^d \to [-B, B]$, then for any $f \in \mathcal{F}^{'}$, we have*

$$\begin{aligned}
\mathbb{P}\Bigg(\mathcal{S}_1(n, \mathcal{F}^{'}) \leq \varepsilon(\alpha + \beta + \mathcal{E}(f) - \mathcal{E}(f_\rho))\Bigg) \\
\geq 1 - \sup_{\mathcal{D}} \mathcal{N}\Bigg(\frac{\beta\varepsilon}{20B}, \mathcal{F}^{'}, \|\cdot\|_{L_1(D)}\Bigg) \exp\Bigg(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)B^4}\Bigg).
\end{aligned}$$

Lemma E.4.3 *[Theorem 6 of [213]] Let $B > 0$ and $\mathcal{F}^{'}$ be a set of functions $f : \mathcal{X} \to$*

$[-B, B]$. *Then for any* $\varepsilon \in (0, B]$*, there holds*

$$\mathcal{M}(\varepsilon, \mathcal{F}^{'}, \| \cdot \|_{L_1(D)})) \leq 2\bigg(\frac{2eB}{\varepsilon}\log\frac{2eB}{\varepsilon}\bigg)^{Pdim\big(\mathcal{F}^{'}\big)}. \tag{E.33}$$

Recall a classical relation between ε-packing number and ε-covering number that asserts

$$\mathcal{M}(2\varepsilon, \mathcal{F}, \| \cdot \|_{L_1(D)})) \leq \mathcal{N}(\varepsilon, \mathcal{F}, \| \cdot \|_{L_1(D)})) \leq \mathcal{M}(\varepsilon, \mathcal{F}, \| \cdot \|_{L_1(D)})), \tag{E.34}$$

for any $\varepsilon > 0$. Combining (E.33), (E.34), the facts $\log x < x, \forall x > 0$, and Pdim$(\mathcal{F}_B) \leq$ Pdim$(\mathcal{F})$ (See [214], page 297), we have the upper-bound on $\mathcal{N}(\varepsilon, \mathcal{F}_B, \| \cdot \|_{L_1(D)}))$ as follows:

$$\mathcal{N}(\varepsilon, \mathcal{F}_B, \| \cdot \|_{L_1(D)})) \leq 2\bigg(\frac{2eB}{\varepsilon}\log\frac{2eB}{\varepsilon}\bigg)^{\text{Pdim}\big(\mathcal{F}_B\big)} \leq 2\bigg(\frac{2eB}{\varepsilon}\bigg)^{\text{2Pdim}\big(\mathcal{F}\big)}. \tag{E.35}$$

Then, taking $\varepsilon = \frac{1}{2}$, $\beta = \frac{1}{n}$ in Lemma E.4.2, using the upper-bound on covering number in (E.35) yields the lower bound for the confidence level in Lemma (E.4.2) as follows:

$$\begin{aligned} &1 - \sup_{\mathcal{D}} \mathcal{N}\bigg(\frac{1}{40Bn}, \mathcal{F}_B, \| \cdot \|_{L_1(D)}\bigg) \exp\bigg(- \frac{\alpha n}{2568B^4}\bigg) \\ &\geq 1 - C_B \cdot \exp\bigg(2 \cdot \text{Pdim}(\mathcal{F}) \cdot \log(n) - \frac{\alpha n}{2568B^4}\bigg), \end{aligned} \tag{E.36}$$

where $C_B > 0$ is some absolute constants dependent on B. Choosing α in (E.36) such that

$$\alpha = C_{B,\delta} \cdot \frac{\text{Pdim}(\mathcal{F}) \cdot \log(n)}{n}$$

with a properly chosen $C_{B,\delta} > 0$ absolute constant dependent on B and δ yields the probability of following event is at least $1-\frac{\delta}{2}$:

$$\mathcal{S}_1(n,\mathcal{F}_B) \le \frac{1}{2}\bigg(C_{B,\delta}\cdot\frac{\mathrm{Pdim}(\mathcal{F})\cdot\log(n)}{n}+\frac{1}{n}+\mathcal{E}(\pi_B\widehat{f}_n)-\mathcal{E}(f_\rho)\bigg). \tag{E.37}$$

<u>**Step 2 : Control $\mathcal{S}_2(n,\mathcal{F})$.**</u> Define a random variable η on $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ to be

$$\eta(z)=(y-f(x))^2-(y-f_\rho(x))^2.$$

Since $|\eta(z)|\le(3B+\|f\|_\infty)^2$, then $|\eta(z)-\mathbb{E}[\eta(z)]|\le 2(3B+\|f\|_\infty)^2$. It is also easy to see $\sigma^2\le\mathbb{E}\big[\eta^2\big]\le\big(3B+\|f\|_\infty\big)^2\|f-f_\rho\|_\infty^2$. Then, by the one-side Bernstein's inequality (see Lemma E.5.2), we have

$$\mathbb{P}\bigg(\mathcal{S}_2(n,\mathcal{F})<\varepsilon\bigg)\ge 1-\exp\left\{-\frac{n\varepsilon^2}{2\big(3B+\|f\|_\infty\big)^2\big(\|f-f_\rho\|_\infty^2+\frac{2}{3}\varepsilon\big)}\right\}.$$

Taking $\frac{\delta}{2}=\exp\left\{-\frac{n\varepsilon^2}{2\big(3B+\|f\|_\infty\big)^2\big(\|f-f_\rho\|_\infty^2+\frac{2}{3}\varepsilon\big)}\right\}$, $\mathcal{A}:=2(3B+\|f\|_\infty)^2$, $\mathcal{B}:=\|f-f_\rho\|_\infty^2$ and solving the quadratic equation with respect to ε yield the following inequalities with some absolute constant $C_0^{''}>0$:

$$\begin{aligned}\varepsilon&=\frac{\mathcal{A}\log\left(\frac{2}{\delta}\right)+\sqrt{\mathcal{A}^2\log\left(\frac{2}{\delta}\right)+9n\mathcal{A}\mathcal{B}\log\left(\frac{2}{\delta}\right)}}{3n}\\&\le\frac{2\mathcal{A}\log\left(\frac{2}{\delta}\right)}{3n}+\sqrt{\mathcal{A}\mathcal{B}\cdot\frac{\log\left(\frac{2}{\delta}\right)}{n}}\\&\le C_{B,f,\delta}\cdot\frac{\|f-f_\rho\|_\infty}{\sqrt{n}},\end{aligned}$$

where in the first inequality, the facts $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b>0$ is used, and $C_{B,f,\delta}$ is a constant dependent on C,B and f. Then, with probability at least $1-\frac{\delta}{2}$, we have

$$\mathcal{S}_2(n,\mathcal{F})\le C_{B,f,\delta}\cdot\frac{\|f-f_\rho\|_\infty}{\sqrt{n}}. \tag{E.38}$$

Step 3 : Combining Everything. Note $\mathcal{E}\big(f\big) - \mathcal{E}\big(f_\rho\big) = \|f - f_\rho\|^2_{\rho_\mathcal{X}} \leq \|f - f_\rho\|^2_\infty$. Then, plugging the (E.37) and (E.38) in (E.32) yields the claim. □

E.4.2 Proof of Theorem 5.3.3

Theorem E.4.4 *Suppose $f_\rho \in W^r_\infty(\mathcal{S}^{d-1})$ with $r > 0$. A network $\widehat{f}_n$ from* (5.8) *with choices $N = \lceil n^{\frac{2}{3d+4r}} \rceil$, $M = \lceil n^{\frac{3d}{3d+4r}} \rceil$, and $m = \lceil \frac{r}{3d+4r} \log_2(n) \rceil$ yield the bound on the excess risk with probability at least $1 - \delta$ as follows:*

$$\begin{aligned} &\mathcal{E}\big(\pi_M \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big) \\ &\leq \mathcal{C}_{B,\eta,\delta,f} \cdot \max\left\{1, \frac{6rd}{(3d+4r)^2}(\log_2(n))^4, \big(\frac{6}{\pi e}\big)^{\frac{d}{2}} d^{2N + \frac{3d-4r-2}{4}}, d^{4N}\right\} \cdot n^{-\frac{2r}{2r+1.5d}}, \end{aligned} \tag{E.39}$$

where $\mathcal{C}_{B,\eta,\delta,f}$ depends on B, η, δ, f and independent on d, r and n.

Proof. Let $0 < \alpha < 1, m, N, M \in \mathbb{N}$ with $1 \leq N \leq d^\alpha + 1$. Then, for $f_\rho \in W^r_\infty(\mathcal{S}^{d-1})$, recall from Theorem E.2.8 that there exists a network

$$\tilde{f} \in \mathcal{F}\big(L, \big(d, 22NM, \ldots, 22NM, 1\big), \mathcal{N}\big) \tag{E.40}$$

with depth $L = (m+4)\lceil \log_2(2N) \rceil$ and number of parameters $\mathcal{N} \leq M(2d + 404N \cdot (m + 3) + 2N + 4) + 1$ such that the corresponding network's approximation error is bounded as:

$$\begin{aligned} \left\|\tilde{f} - f_\rho\right\|_\infty &\leq C''_\eta \, \|f\|_{W^r_\infty(\mathcal{S}^{d-1})} \times \\ &\max\left\{N^{-r}, \frac{\big(\frac{6}{\pi e}\big)^{\frac{d}{4}} d^{N + \frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}}, d^{2N}\big(\log_2(2N)\big)^2 2^{-2m}\right\}, \end{aligned} \tag{E.41}$$

where C''_η is a constant dependent on η, and independent on d, r, N, M and f. Since the network width is $22NM$, the total number of units across the L-hidden layers (i.e., $\mathcal{U}$) of $\tilde{f}$

is bounded as

$$\mathcal{U} \leq 22NM \cdot (m+4)\lceil \log_2(2N) \rceil.$$

If $Nm = o(d)$, it is easy to see $\mathcal{N} \leq \mathcal{O}(Md)$. Recall from the result of Lemma E.5.1, the pseudo-dimension of function class $\mathcal{F}$ in (E.40) is bounded as follows: for some universal constants $C > 0$:

$$\text{Pdim}(\mathcal{F}) \leq C \cdot \big(mMd \cdot \lceil \log_2(N) \rceil \cdot \log \big(mMN \lceil \log_2(N) \rceil \big) \big). \tag{E.42}$$

Plug the (E.41), (E.42) in (E.31) from Proposition 5.3.2.

$$\begin{aligned}
\mathcal{E}\big(\pi_M \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big) \leq \mathcal{C}_{B,\eta,\delta,f} \times & \\
& \Bigg\{ \underbrace{\frac{mMd}{n} \log(n) \cdot \lceil \log_2(N) \rceil \cdot \log \big(mMN \lceil \log_2(N) \rceil \big)}_{\textbf{Bound for } \text{Pdim}(\mathcal{F}) \cdot \log(n)/n} \\
& + \underbrace{\max \Bigg\{ N^{-r}, \frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{4}} d^{N + \frac{3d-4r-2}{8}} (2N+1)^{\frac{3d-4r}{4}}}{\sqrt{M}}, d^{2N} \big(\log_2(2N) \big)^2 2^{-2m} \Bigg\} / \sqrt{n}}_{\textbf{Bound for } \| \tilde{f} - f_\rho \|_\infty / \sqrt{n}} \\
& + \underbrace{\max \Bigg\{ N^{-2r}, \frac{\left(\frac{6}{\pi e}\right)^{\frac{d}{2}} d^{2N + \frac{3d-4r-2}{4}} (2N+1)^{\frac{3d-4r}{2}}}{M}, d^{4N} \big(\log_2(2N) \big)^4 2^{-4m} \Bigg\}}_{\textbf{Bound for } \| \tilde{f} - f_\rho \|_\infty^2} \Bigg\},
\end{aligned} \tag{E.43}$$

where $\mathcal{C}_{B,\eta,\delta,f}$ depends on B, η, δ, f and independent on d, r and n. Then, under the regime $1 \leq N \leq d^\alpha + 1$ for some $0 < \alpha < 1$ and $n \ll d$, choices of $m = \lceil \frac{r}{3d+4r} \log_2(n) \rceil$, $N = \lceil n^{\frac{2}{3d+4r}} \rceil$ and $M = \lceil n^{\frac{3d}{3d+4r}} \rceil$ make the fraction of the first term in (E.43) simple as follows:

$$\begin{aligned}\lceil\log_2(N)\cdot\log\big(mMN\lceil\log_2(N)\rceil\big) &\leq \frac{2}{3d+4r}\log_2(n)\log\Bigg(\log_2(n)n^{\frac{3d+2}{3d+4r}}\frac{2r}{(3d+4r)^2}\lceil\log_2(n)\rceil\Bigg)\\ &\leq \frac{6}{3d+4r}\big(\log_2(n)\big)^2.\end{aligned}$$

Then, with the same choices of m, N, M as above, we obtain the bound on the excess risk as :

$$\begin{aligned}&\mathcal{E}\big(\pi_M\widehat{f}_n\big)-\mathcal{E}\big(f_\rho\big)\\ &\leq \mathcal{C}_{B,\eta,\delta,f}\cdot\max\Bigg\{1,\frac{6rd}{(3d+4r)^2}(\log_2(n))^4,\big(\frac{6}{\pi e}\big)^{\frac{d}{2}}d^{2N+\frac{3d-4r-2}{4}},d^{4N}\Bigg\}\cdot n^{-\frac{4r}{4r+3d}}.\end{aligned}$$

This conlcudes the proof. □

E.4.3 Proof of Theorem 5.3.4

Theorem E.4.5 *Suppose $f_\rho \in W^r_\infty([0,1]^d)$ with $r > 0$. A network $\widehat{f}_n$ from* (5.11) *with choices $N^H = \lceil n^{\frac{d}{2d+r}}\rceil$, and $m^H = \lceil\frac{d+r}{d+2r}\log_2(n)\rceil$ yield the bound on the excess risk with probability at least $1-\delta$ as follows:*

$$\begin{aligned}&\mathcal{E}\big(\pi_M\widehat{f}_n\big)-\mathcal{E}\big(f_\rho\big) \qquad\qquad (E.44)\\ &\leq \mathcal{C}_{B,\eta,\delta,K}\cdot\max\Bigg\{\big\lceil\log_2(d+\lceil r\rceil)\big\rceil^2(d+r)^d\cdot(\log_2(n))^3,\big(1+r^2+d^2\big)^2 6^{2d}+3^{2r}\Bigg\}\cdot n^{-\frac{2r}{2r+d}},\end{aligned}$$

where $\mathcal{C}_{B,\eta,\delta,K}$ depends on B, η, δ, K and independent on d, r and n.

Proof. From Theorem 5 of [8], for any function $f_\rho \in \mathcal{C}^r_d([0,1]^d, K)$ and any integers $m \geq 1$ and $N \geq (r+1)^d \vee (K+1)e^d$, there exists a network

$$\tilde{f} \in \mathcal{F}\big(L,(d,6(d+\lceil r\rceil)N,\ldots,6(d+\lceil r\rceil)N,1),\mathcal{N},\infty\big) \qquad (E.45)$$

with depth $L = 8+(m+5)\big(1+\lceil\log_2(d\vee r)\rceil\big)$ and the number of parameters $\mathcal{N} \leq$

$141(1+d+r)^{3+d}N(m+6)$, such that

$$\left\|\tilde{f}-f_\rho\right\|_\infty \le (2K+1)(1+d^2+r^2)6^dN2^{-m}+K3^rN^{-\frac{r}{d}}. \tag{E.46}$$

Then, similarly with the proof in Theorem 5.3.3, by the result of Lemma E.5.1, the pseudo-dimension of $\mathcal{F}$ in (E.45) can be bounded as

$$\text{Pdim}(\mathcal{F}) \le C \cdot \left(m^2N(d+r)^d\lceil\log_2(d\vee r)\rceil \log\left((d+\lceil r\rceil)mN\lceil\log_2(d\vee r)\rceil\right)\right), \tag{E.47}$$

for some universal constants $C>0$.

Plug the (E.46) and (E.47) in (E.31) from Proposition 5.3.2. Then, we obtain the bound on the excess risk as follows:

$$\begin{aligned}\mathcal{E}\big(\pi_M\widehat{f}_n\big)-\mathcal{E}\big(f_\rho\big) \le \mathcal{C}_{B,\delta,K}\times &\\ &\left\{\underbrace{\frac{m^2N}{n}\log(n)\cdot(d+r)^d\lceil\log_2(d\vee r)\rceil\log\left((d+\lceil r\rceil)mN\lceil\log_2(d\vee r)\rceil\right)}_{\textbf{Bound for } \text{Pdim}(\mathcal{F})\cdot\log(n)/n}\right.\\ &+\underbrace{\left((1+d^2+r^2)6^dN2^{-m}+3^rN^{-\frac{r}{d}}\right)/\sqrt{n}}_{\textbf{Bound for } \left\|\tilde{f}-f_\rho\right\|_\infty/\sqrt{n}}\\ &\left.+\underbrace{\left((1+d^2+r^2)^26^{2d}N^22^{-2m}+3^{2r}N^{-\frac{2r}{d}}\right)}_{\textbf{Bound for } \left\|\tilde{f}-f_\rho\right\|_\infty^2}\right\},\end{aligned} \tag{E.48}$$

where $\mathcal{C}_{B,\delta,K}$ depends on B, δ, K and independent on d, r and n. Note that we use $(a+b)^2 \le 2a^2+2b^2$ for all $a,b\in\mathbb{R}$ for getting the bound on $\left\|\tilde{f}-f_\rho\right\|_\infty^2$. We choose the $N=\lceil n^{\frac{d}{2r+d}}\rceil$ and $m=\lceil\frac{d+r}{2r+d}\log_2(n)\rceil$. Then, a fraction of the first term in (E.48) can be bounded as:

$$\begin{aligned}\log\left((d+\lceil r\rceil)\cdot\frac{d+r}{2r+d}\cdot\log_2(n)\cdot n^{\frac{d}{2r+d}}\cdot\log_2(d\vee r)\right) &\le \log_2\left((d+\lceil r\rceil)^2n^2\right)\\ &\le 4\cdot\lceil\log_2(d+\lceil r\rceil)\rceil,\end{aligned}$$

where $n \ll d$ is used in the second inequality. Then, we obtain the bound on the excess risk as :

$$\begin{aligned}&\mathcal{E}\big(\pi_M \widehat{f}_n\big) - \mathcal{E}\big(f_\rho\big)\\&\leq \mathcal{C}_{B,\eta,\delta,K} \cdot \max\left\{ \big\lceil \log_2(d + \lceil r \rceil) \big\rceil^2 (d+r)^d \cdot (\log_2(n))^3, \big(1 + r^2 + d^2\big)^2 6^{2d} + 3^{2r} \right\} \cdot n^{-\frac{2r}{2r+d}}.\end{aligned}$$

This conlcudes the proof. □

E.5 Useful Lemmas

Lemma E.5.1 *[Theorem 6 of [198]] Consider the function class $\mathcal{F}$ computed by a feed-forward neural network architecture with $\mathcal{N}$ parameters and $\mathcal{U}$ computation units arranged across L layers. Suppose that all non-ouput units have piecewise-polynomial activation functions with $p+1$ pieces and degree no more than d, and the output unit has the identity function as its activation function. Then the VC-dimension and pseudo-dimension of class $\mathcal{F}$ is upper bounded by*

$$\textit{VCdim}(\mathcal{F}), \textit{Pdim}(\mathcal{F}) \leq C \cdot \big(L\mathcal{N}\log(p \cdot \mathcal{U}) + L^2\mathcal{N}\log(d)\big),$$

with some universal constants $C > 0$.

Lemma E.5.2 *[Theorem 2.8.4 of [208]] Let η be a random variable on a probability space $\mathcal{Z}$ with mean $\mathbb{E}(\eta) = \mu$, variance $\sigma^2(\eta) = \sigma^2$, and satisfying $|\eta(z) - \mathbb{E}(\eta)| \leq B_\eta$ for almost $z \in \mathcal{Z}$. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}\left\{ \frac{1}{n}\sum_{i=1}^n \eta(z_i) - \mu < \varepsilon \right\} \geq 1 - \exp\left\{ -\frac{n\varepsilon^2}{2\big(\sigma^2 + \frac{1}{3}B_\eta \varepsilon\big)} \right\}.$$

REFERENCES

[1] D. L. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” *IEEE transactions on information theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[2] D. L. Donoho, M. Elad, and V. N. Temlyakov, “Stable recovery of sparse overcomplete representations in the presence of noise,” *IEEE Transactions on information theory*, vol. 52, no. 1, pp. 6–18, 2005.

[3] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[4] M. J. Wainwright, “Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso),” *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[5] S. Negahban and M. J. Wainwright, “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *The Annals of Statistics*, pp. 1069–1097, 2011.

[6] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,” *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[7] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[8] J. Schmidt-Hieber, “Nonparametric regression using deep neural networks with ReLU activation function,” *Annals of Statistics*, vol. 48, no. 4, pp. 1875–1897, 2020.

[9] Z. Fang and G. Cheng, “Optimal learning rates of deep convolutional neural networks: Additive ridge functions,” *arXiv preprint arXiv:2202.12119*, 2022.

[10] P. J. Bickel and A. Chen, “A nonparametric view of network models and newman–girvan and other modularities,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 50, pp. 21 068–21 073, 2009.

[11] K. Chaudhuri, F. Chung, and A. Tsiatas, “Spectral clustering of graphs with general degrees in the extended planted partition model,” in *Conference on Learning Theory*, 2012, pp. 35–1.

[12] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[13] S. Zhang and H. Zhao, "Community identification in networks with unbalanced structure," *Physical Review E*, vol. 85, no. 6, p. 066 114, 2012.

[14] P. Ravikumar, G. Raskutti, M. J. Wainwright, and B. Yu, "Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle.," in *NIPS*, 2008, pp. 1329–1336.

[15] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, pp. 1418–1429, 2006.

[16] K. Chen, H. Dong, and K.-S. Chan, "Reduced rank regression via adaptive nuclear norm penalization," *Biometrika*, vol. 100, pp. 901–920, 2013.

[17] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, "Stochastic blockmodels with a growing number of classes," *Biometrika*, vol. 99, no. 2, pp. 273–284, 2012.

[18] P. Hoff, "Modeling homophily and stochastic equivalence in symmetric relational data," in *Advances in neural information processing systems*, 2008, pp. 657–664.

[19] J. Jin, "Fast community detection by score," *The Annals of Statistics*, vol. 43, no. 1, pp. 57–89, 2015.

[20] Y. Zhao, E. Levina, and J. Zhu, "Consistency of community detection in networks under degree-corrected stochastic block models," *The Annals of Statistics*, vol. 40, no. 4, pp. 2266–2292, 2012.

[21] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social networks*, vol. 31, no. 3, pp. 204–213, 2009.

[22] Z. Ma, Z. Ma, and H. Yuan, "Universal latent space model fitting for large networks with edge covariates," *Journal of Machine Learning Research*, vol. 21, no. 4, pp. 1–67, 2020.

[23] Y.-J. Wu, E. Levina, and J. Zhu, "Generalized linear models with low rank effects for network data," *arXiv preprint arXiv:1705.06772*, 2017.

[24] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[25] P. D. Hoff, A. E. Raftery, and M. S. Handcock, “Latent space approaches to social network analysis,” *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1090–1098, 2002.

[26] P. D. Hoff, “Bilinear mixed-effects models for dyadic data,” *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 286–295, 2005.

[27] P. D. Hoff, *Random effects models for network data*. na, 2003.

[28] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[29] A. Agarwal, S. Negahban, and M. J. Wainwright, “Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions,” *The Annals of Statistics*, vol. 40, no. 2, pp. 1171–1197, 2012.

[30] M. A. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters, “1-bit matrix completion,” *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 189–223, 2014.

[31] Y. Chen, X. Li, J. Liu, and Z. Ying, “A fused latent and graphical model for multivariate binary data,” *arXiv preprint arXiv:1606.08925*, 2016.

[32] S. B. Eysenck, H. J. Eysenck, and P. Barrett, “A revised version of the psychoticism scale,” *Personality and individual differences*, vol. 6, no. 1, pp. 21–29, 1985.

[33] M. D. Reckase, “Multidimensional item response theory models,” in *Multidimensional item response theory*, Springer, 2009, pp. 79–112.

[34] E. Ising, “Beitrag zur theorie des ferromagnetismus,” *Zeitschrift für Physik*, vol. 31, no. 1, pp. 253–258, 1925.

[35] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, “Latent variable graphical model selection via convex optimization,” in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2010, pp. 1610–1613.

[36] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, “Stable principal component pursuit,” in *2010 IEEE international symposium on information theory*, IEEE, 2010, pp. 1518–1522.

[37] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[38] D. Gabay and B. Mercier, *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d'informatique et d'automatique, 1975.

[39] R. Glowinski and A. Marrocco, "On the solution of a class of nonlinear Dirichlet problems by a penalty-duality method and finite elements of order one," in *Optimization Techniques IFIP Technical Conference*, Springer, 1975, pp. 327–333.

[40] H. H. Harman, *Modern factor analysis*. Univ. of Chicago Press, 1960.

[41] K. G. Jöreskog, "A general approach to confirmatory maximum likelihood factor analysis," *Psychometrika*, vol. 34, no. 2, pp. 183–202, 1969.

[42] K. G. Jöreskog, "A general method for estimating a linear structural equation system," *ETS Research Bulletin Series*, vol. 1970, no. 2, pp. i–41, 1970.

[43] F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. IAP, 2008.

[44] R. P. McDonald, *Factor analysis and related methods*. Psychology Press, 2014.

[45] G. Rasch, "Probabilistic models for some intelligence and attainment tests. 1960," *Copenhagen, Denmark: Danish Institute for Educational Research*, 1980.

[46] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American control conference*, Citeseer, vol. 6, 2001, pp. 4734–4739.

[47] F. R. Bach, "Consistency of trace norm minimization," *Journal of Machine Learning Research*, vol. 9, no. Jun, pp. 1019–1048, 2008.

[48] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[49] P. Ji and J. Jin, "Coauthorship and citation networks for statisticians," *The Annals of Applied Statistics*, vol. 10, no. 4, pp. 1779–1812, 2016.

[50] K. Chen and J. Lei, "Network cross-validation for determining the number of communities in network data," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 241–251, 2018.

[51] A. B. Owen and P. O. Perry, "Bi-cross-validation of the svd and the nonnegative matrix factorization," *The Annals of Applied Statistics*, vol. 3, no. 2, pp. 564–594, 2009.

[52] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[53] W. W. Zachary, “An information flow model for conflict and fission in small groups,” *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[54] J. Fan, Y. Fan, X. Han, and J. Lv, “Simple: Statistical inference on membership profiles in large networks,” *arXiv preprint arXiv:1910.01734*, 2019.

[55] J. Jin, Z. T. Ke, and S. Luo, “Estimating network memberships by simplex vertex hunting,” *arXiv preprint arXiv:1708.07852*, 2017.

[56] P.-S. Koutsourelakis and T. Eliassi-Rad, “Finding mixed-memberships in social networks.,” in *AAAI spring symposium: Social information processing*, 2008, pp. 48–53.

[57] L. A. Adamic and N. Glance, “The political blogosphere and the 2004 us election: Divided they blog,” in *Proceedings of the 3rd international workshop on Link discovery*, 2005, pp. 36–43.

[58] Y. Zhang, E. Levina, and J. Zhu, “Detecting overlapping communities in networks using spectral methods,” *arXiv preprint arXiv:1412.3432*, 2014.

[59] B. A. Prakash, A. Sridharan, M. Seshadri, S. Machiraju, and C. Faloutsos, “Eigenspokes: Surprising patterns and scalable community chipping in large graphs,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2010, pp. 435–448.

[60] E. Schrödinger, “An undulatory theory of the mechanics of atoms and molecules,” *Physical review*, vol. 28, no. 6, p. 1049, 1926.

[61] F. Black and M. Scholes, “The pricing of options and corporate liabilities,” *Journal of political economy*, vol. 81, no. 3, pp. 637–654, 1973.

[62] J. Haskovec, L. M. Kreusser, and P. Markowich, “ODE and PDE based modeling of biological transportation networks,” *arXiv preprint arXiv:1805.08526*, 2018.

[63] Y. Achdou, F. J. Buera, J.-M. Lasry, P.-L. Lions, and B. Moll, “Partial differential equation models in macroeconomics,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2028, p. 20 130 397, 2014.

[64] T. Musha and H. Higuchi, “Traffic current fluctuation and the Burgers equation,” *Japanese journal of applied physics*, vol. 17, no. 5, p. 811, 1978.

[65] V. Tikhomirov, “A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem,” in *Selected works of AN Kolmogorov*, Springer, 1991, pp. 242–270.

[66] A. C. Newell, *Solitons in mathematics and physics*. Siam, 1985, vol. 48.

[67] J. Fan, T. Gasser, I. Gijbels, M. Brockmann, and J. Engel, “Local polynomial regression: Optimal kernels and asymptotic minimax efficiency,” *Annals of the Institute of Statistical Mathematics*, vol. 49, no. 1, pp. 79–99, 1997.

[68] J. Fan, *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.

[69] H. Liang and H. Wu, “Parameter estimation for differential equation models using a framework of measurement error in regression models,” *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1570–1583, 2008.

[70] J. Chen and H. Wu, “Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics,” *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 369–384, 2008.

[71] J. Chen and H. Wu, “Estimation of time-varying parameters in deterministic dynamic models,” *Statistica Sinica*, vol. 18, no. 3, pp. 987–1006, 2008.

[72] M. Bär, R. Hegger, and H. Kantz, “Fitting partial differential equations to space-time dynamics,” *Physical Review E*, vol. 59, no. 1, p. 337, 1999.

[73] J. Jia, K. Rohe, and B. Yu, “The Lasso under poisson-like heteroscedasticity,” *Statistica Sinica*, pp. 99–118, 2013.

[74] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[75] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, “High-dimensional Ising model selection using ℓ_1-regularized logistic regression,” *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

[76] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman, “Sparse additive models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 5, pp. 1009–1030, 2009.

[77] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, “High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence,” *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[78] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Union support recovery in high-dimensional multivariate regression," in *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, IEEE, 2008, pp. 21–26.

[79] W. Wang, Y. Liang, and E. Xing, "Block regularized Lasso for multivariate multi-response linear regression," in *Artificial Intelligence and Statistics*, 2013, pp. 608–617.

[80] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar, "A dirty model for multi-task learning," in *Advances in neural information processing systems*, 2010, pp. 964–972.

[81] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[82] A. Cortiella, K.-C. Park, and A. Doostan, "Sparse identification of nonlinear dynamical systems via reweighted ℓ_1-regularized least squares," *Computer Methods in Applied Mechanics and Engineering*, vol. 376, p. 113 620, 2021.

[83] S. H. Kang, W. Liao, and Y. Liu, "Ident: Identifying differential equations with numerical time evolution," *Journal of Scientific Computing*, vol. 87, no. 1, pp. 1–27, 2021.

[84] H. Schaeffer, "Learning partial differential equations via data discovery and sparse optimization," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2197, p. 20 160 446, 2017.

[85] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Data-driven discovery of partial differential equations," *Science Advances*, vol. 3, no. 4, e1602614, 2017.

[86] H. Schaeffer, G. Tran, and R. Ward, "Extracting sparse high-dimensional dynamics from limited data," *SIAM Journal on Applied Mathematics*, vol. 78, no. 6, pp. 3279–3295, 2018.

[87] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[88] A. Feuer and A. Nemirovski, "On sparse representation in pairs of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1579–1581, 2003.

[89] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[90] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[91] K. Knight and W. Fu, "Asymptotics for Lasso-type estimators," *The Annals of statistics*, pp. 1356–1378, 2000.

[92] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE transactions on information theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[93] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine learning research*, vol. 7, no. Nov, pp. 2541–2563, 2006.

[94] J.-J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3601–3608, 2005.

[95] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *The Annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[96] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, p. 101, 2010.

[97] Y.-p. Mack and B. W. Silverman, "Weak and strong uniform consistency of kernel regression estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 61, no. 3, pp. 405–415, 1982.

[98] G. Tusnády, "A remark on the approximation of the sample df in the multidimensional case," *Periodica Mathematica Hungarica*, vol. 8, no. 1, pp. 53–55, 1977.

[99] E. Masry, "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *Journal of Time Series Analysis*, vol. 17, no. 6, pp. 571–599, 1996.

[100] Y. Li and T. Hsing, "Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data," *The Annals of Statistics*, vol. 38, no. 6, pp. 3321–3351, 2010.

[101] B. W. Silverman, "Weak and strong uniform consistency of the kernel estimate of a density and its derivatives," *The Annals of Statistics*, pp. 177–184, 1978.

[102] M. P. Bonkile, A. Awasthi, C. Lakshmi, V. Mukundan, and V. Aswin, "A systematic literature review of Burgers' equation with recent advances," *Pramana*, vol. 90, no. 6, p. 69, 2018.

[103] O. Rudenko and S. Soluian, “The theoretical principles of nonlinear acoustics,” *MoIzN*, 1975.

[104] K. Sawada and T. Kotera, “A method for finding n-soliton solutions of the KdV equation and KdV-like equation,” *Progress of Theoretical Physics*, vol. 51, no. 5, pp. 1355–1367, 1974.

[105] J. Boussinesq, *Essai sur la théorie des eaux courantes*. Impr. nationale, 1877.

[106] J.-Y. Audibert and A. B. Tsybakov, “Fast learning rates for plug-in classifiers,” *The Annals of statistics*, vol. 35, no. 2, pp. 608–633, 2007.

[107] D. Ruppert, S. J. Sheather, and M. P. Wand, “An effective bandwidth selector for local least squares regression,” *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1257–1270, 1995.

[108] J. Fan and J. Lv, “Sure independence screening for ultrahigh dimensional feature space,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[109] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[110] A. Javanmard and A. Montanari, “Confidence intervals and hypothesis testing for high-dimensional regression,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[111] Y. He, S. h. Kang, W. Liao, H. Liu, and Y. Liu, “Robust PDE identification from noisy data,” *arXiv preprint arXiv:2006.06557*, 2020.

[112] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang, “Weighted nuclear norm minimization and its applications to low level vision,” *International journal of computer vision*, vol. 121, pp. 183–208, 2017.

[113] N. Yair and T. Michaeli, “Multi-scale weighted nuclear norm image restoration,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3165–3174.

[114] S. Liu, Q. Hu, P. Li, J. Zhao, M. Liu, and Z. Zhu, “Speckle suppression based on weighted nuclear norm minimization and grey theory,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 2700–2708, 2018.

[115] G. Kim, J. Cho, and M. Kang, “Cauchy noise removal by weighted nuclear norm minimization,” *Journal of Scientific Computing*, vol. 83, pp. 1–21, 2020.

[116] Y. Zhang, X. Lei, Y. Pan, and W. Pedrycz, "Prediction of disease-associated circrnas via circrna–disease pair graph and weighted nuclear norm minimization," *Knowledge-Based Systems*, vol. 214, p. 106 694, 2021.

[117] M. Vanidevi and N. Selvaganesan, "Channel estimation for finite scatterers massive multi-user mimo system," *Circuits, Systems, and Signal Processing*, vol. 36, pp. 3761–3777, 2017.

[118] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, "Dimension reduction and coefficient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B*, vol. 69, pp. 329–346, 2007.

[119] J. D. Lee, Y. Sun, and J. E. Taylor, "On model selection consistency of regularized m-estimators," *Electronic Journal of Statistics*, vol. 9, pp. 608–642, 2015.

[120] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, pp. 2302–2329, 2011.

[121] J. Fan, W. Gong, and Z. Zhu, "Generalized high-dimensional trace regression via nuclear norm regularization," *Journal of econometrics*, vol. 212, pp. 177–202, 2019.

[122] J. Fan, W. Wang, and Z. Zhu, "A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery," *Annals of statistics*, vol. 49, p. 1239, 2021.

[123] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, pp. 29–63, 2019.

[124] A. Rohde and A. B. Tsybakov, "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, vol. 39, pp. 887–930, 2011.

[125] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215–223, 1979.

[126] A. Isenmann, "Modern multivariate statistical techniques," *Regression, Classification and Manifold Learning*, vol. 25, p. 733, 2008.

[127] R. G. Brereton, *Chemometrics: data analysis for the laboratory and chemical plant*. John Wiley & Sons, 2003.

[128] P. Shang and L. Kong, "Regularization parameter selection for the low rank matrix recovery," *Journal of Optimization Theory and Applications*, vol. 189, pp. 772–792, 2021.

[129] M. Law, Y. Ritov, R. Zhang, and Z. Zhu, “Rank-constrained least-squares: Prediction and inference,” *arXiv preprint arXiv:2111.14287*, 2021.

[130] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, “On exact computation with an infinitely wide neural net,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 8141–8150, 2019.

[131] A. Jacot, C. Hongler, and F. Gabriel, “Neural tangent kernel: Convergence and generalization in neural networks,” in *NeurIPS*, 2018.

[132] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, “Deep neural networks as Gaussian processes,” in *International Conference on Learning Representations*, 2018.

[133] L. Chizat and F. Bach, “On the global convergence of gradient descent for over-parameterized models using optimal transport,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 3040–3050.

[134] Y. Li and Y. Liang, “Learning overparameterized neural networks via stochastic gradient descent on structured data,” in *NeurIPS*, 2018.

[135] S. S. Du, X. Zhai, B. Poczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” in *International Conference on Learning Representations*, 2018.

[136] Z. Allen-Zhu, Y. Li, and Z. Song, “A convergence theory for deep learning via over-parameterization,” *arXiv e-prints*, arXiv–1811, 2018.

[137] D. Zou, Y. Cao, D. Zhou, and Q. Gu, “Stochastic gradient descent optimizes over-parameterized deep ReLU networks. arxiv e-prints, art,” *arXiv preprint arXiv:1811.08888*, 2018.

[138] D. Zou, Y. Cao, D. Zhou, and Q. Gu, “Gradient descent optimizes over-parameterized deep ReLU networks,” *Machine Learning*, vol. 109, no. 3, pp. 467–492, 2020.

[139] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, “Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks,” in *International Conference on Machine Learning*, PMLR, 2019, pp. 322–332.

[140] Y. Cao and Q. Gu, “Generalization bounds of stochastic gradient descent for wide and deep neural networks,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 10 836–10 846, 2019.

[141] A. Nitanda and T. Suzuki, "Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime," in *International Conference on Learning Representations*, 2020.

[142] T. Hu, W. Wang, C. Lin, and G. Cheng, "Regularization matters: A nonparametric perspective on overparametrized neural network," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, pp. 829–837.

[143] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[144] C. Wei, J. D. Lee, Q. Liu, and T. Ma, "Regularization matters: Generalization and optimization of neural nets vs their induced kernel," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[145] B. Bauer and M. Kohler, "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Annals of Statistics*, vol. 47, no. 4, pp. 2261–2285, 2019.

[146] R. Liu, B. Boukai, and Z. Shang, "Optimal nonparametric inference via deep neural network," *arXiv preprint arXiv:1902.01687*, 2019.

[147] Y. Kim, I. Ohn, and D. Kim, "Fast convergence rates of deep neural networks for classification," *Neural Networks*, vol. 138, pp. 179–197, 2021.

[148] M. Kohler and A. Krzyzak, "Over-parametrized deep neural networks do not generalize well," *arXiv preprint arXiv:1912.03925*, 2019.

[149] I. Kuzborskij and C. Szepesvári, "Nonparametric regression with shallow overparameterized neural networks trained by GD with early stopping," in *Conference on Learning Theory*, PMLR, 2021, pp. 2853–2890.

[150] Z. Ji, J. Li, and M. Telgarsky, "Early-stopped neural networks are consistent," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[151] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[152] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *International Conference on Machine Learning*, PMLR, 2019, pp. 1675–1685.

[153] G. Yang, “Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation,” *arXiv preprint arXiv:1902.04760*, 2019.

[154] Y. Cho and L. Saul, “Kernel methods for deep learning,” *Advances in Neural Information Processing Systems*, vol. 22, pp. 342–350, 2009.

[155] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, “When do neural networks outperform kernel methods?” *arXiv preprint arXiv:2006.13409*, 2020.

[156] R. Basri, M. Galun, A. Geifman, D. Jacobs, Y. Kasten, and S. Kritchman, “Frequency bias in neural networks for input of non-uniform density,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 685–694.

[157] L. Chen and S. Xu, “Deep neural tangent kernel and Laplace kernel have the same RKHS,” in *International Conference on Learning Representations*, 2020.

[158] A. Geifman, A. Yadav, Y. Kasten, M. Galun, D. Jacobs, and R. Basri, “On the similarity between the Laplace and neural tangent kernels,” *arXiv preprint arXiv:2007.01580*, 2020.

[159] A. Bietti and F. Bach, “Deep equals shallow for ReLU networks in kernel regimes,” in *ICLR 2021-International Conference on Learning Representations*, 2021, pp. 1–22.

[160] A. Bietti and J. Mairal, “On the inductive bias of neural tangent kernels,” in *NeurIPS 2019-Thirty-third Conference on Neural Information Processing Systems*, vol. 32, 2019, pp. 12 873–12 884.

[161] G. Raskutti, M. J. Wainwright, and B. Yu, “Early stopping and non-parametric regression: An optimal data-dependent stopping rule,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 335–366, 2014.

[162] M. Yuan and D.-X. Zhou, “Minimax optimal rates of estimation in high dimensional additive models,” *Annals of Statistics*, vol. 44, no. 6, pp. 2564–2593, 2016.

[163] D. Zou and Q. Gu, “An improved analysis of training over-parameterized deep neural networks,” *Advances in neural information processing systems*, 2019.

[164] P. L. Bartlett, A. Montanari, and A. Rakhlin, “Deep learning: A statistical viewpoint,” *Acta Numerica*, vol. 30, pp. 87–201, 2021.

[165] W. Hu, Z. Li, and D. Yu, “Simple and effective regularization methods for training on noisily labeled data with generalization guarantee,” in *International Conference on Learning Representations*, 2019.

[166] A. Caponnetto and E. De Vito, “Optimal rates for the regularized least-squares algorithm,” *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2007.

[167] G. Blanchard and N. Mücke, “Optimal rates for regularization of statistical inverse learning problems,” *Foundations of Computational Mathematics*, vol. 18, no. 4, pp. 971–1013, 2018.

[168] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.

[169] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[170] D. Silver *et al.*, “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[171] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 6645–6649.

[172] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *International Conference on Learning Representations*, 2015.

[173] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[174] J.-L. Starck, Y. Moudden, P. Abrial, and M. Nguyen, “Wavelets, ridgelets and curvelets on the sphere,” *Astronomy & Astrophysics*, vol. 446, no. 3, pp. 1191–1204, 2006.

[175] Y. Wiaux, L. Jacques, and P. Vandergheynst, “Correspondence principle between spherical and euclidean wavelets,” *The Astrophysical Journal*, vol. 632, no. 1, p. 15, 2005.

[176] C. Brechbühler, G. Gerig, and O. Kübler, “Parametrization of closed surfaces for 3-d shape description,” *Computer vision and image understanding*, vol. 61, no. 2, pp. 154–170, 1995.

[177] P. Yu *et al.*, “Cortical surface shape analysis based on spherical wavelets,” *IEEE transactions on medical imaging*, vol. 26, no. 4, pp. 582–597, 2007.

[178] M. Anthony and P. L. Bartlett, *Neural network learning: Theoretical foundations*. Cambridge University Press, Cambridge, UK, 1999, vol. 9.

[179] Z. Fang, H. Feng, S. Huang, and D.-X. Zhou, “Theory of deep convolutional neural networks II: Spherical analysis,” *Neural Networks*, vol. 131, pp. 154–162, 2020.

[180] H. Feng, S. Huang, and D.-X. Zhou, “Generalization analysis of CNNs for classification on spheres,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[181] H. N. Mhaskar, “Neural networks for optimal approximation of smooth and analytic functions,” *Neural computation*, vol. 8, no. 1, pp. 164–177, 1996.

[182] D. Yarotsky, “Error bounds for approximations with deep ReLU networks,” *Neural Networks*, vol. 94, pp. 103–114, 2017.

[183] R. A. DeVore, R. Howard, and C. Micchelli, “Optimal nonlinear approximation,” *Manuscripta Mathematica*, vol. 63, no. 4, pp. 469–478, 1989.

[184] P. Petersen and F. Voigtlaender, “Optimal approximation of piecewise smooth functions using deep ReLU neural networks,” *Neural Networks*, vol. 108, pp. 296–330, 2018.

[185] H. Mhaskar, Q. Liao, and T. Poggio, “Learning functions: When is deep better than shallow,” *arXiv preprint arXiv:1603.00988*, 2016.

[186] T. Suzuki, “Adaptivity of deep ReLU network for learning in besov and mixed smooth besov spaces: Optimal rate and curse of dimensionality,” in *International Conference on Learning Representations*, 2018.

[187] M. Chen, H. Jiang, W. Liao, and T. Zhao, “Efficient approximation of deep ReLU networks for functions on low dimensional manifolds,” *Advances in neural information processing systems*, vol. 32, 2019.

[188] H. Montanelli and Q. Du, “New error bounds for deep ReLU networks using sparse grids,” *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 1, pp. 78–92, 2019.

[189] M. Blanchard and M. A. Bennouna, “Shallow and deep networks are near-optimal approximators of Korobov functions,” in *International Conference on Learning Representations*, 2022.

[190] Z. Han, S. Yu, S.-B. Lin, and D.-X. Zhou, “Depth selection for deep relu nets in feature extraction and generalization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2022), 1853-1868.

[191] M. Chen, H. Jiang, W. Liao, and T. Zhao, “Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery,” *To appear in Information and Inference: a Journal of the IMA*, 2019.

[192] D. L. Donoho and I. M. Johnstone, “Minimax estimation via wavelet shrinkage,” *The annals of Statistics*, vol. 26, no. 3, pp. 879–921, 1998.

[193] F. Dai and Y. Xu, *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013, vol. 23.

[194] C. Efthimiou and C. Frye, *Spherical harmonics in p dimensions*. World Scientific, 2014.

[195] K. Hesse, “A lower bound for the worst-case cubature error on spheres of arbitrary dimension,” *Numerische Mathematik*, vol. 103, no. 3, pp. 413–433, 2006.

[196] W. Freeden, T. Gervens, and M. Schreiner, *Constructive approximation on the sphere with applications to geomathematics*. Oxford University Press on Demand, 1998.

[197] K. Oono and T. Suzuki, “Approximation and non-parametric estimation of resnet-type convolutional neural networks,” in *International conference on machine learning*, PMLR, 2019, pp. 4922–4931.

[198] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, “Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks,” *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 2285–2301, 2019.

[199] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the Lasso and generalizations*. CRC press, 2015.

[200] M. Rosenblatt, “Remarks on a multivariate transformation,” *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 470–472, 1952.

[201] A. Winkelbauer, “Moments and absolute moments of the normal distribution,” *arXiv preprint arXiv:1209.4340*, 2012.

[202] J. Bretagnolle and P. Massart, “Hungarian constructions from the nonasymptotic viewpoint,” *The Annals of Probability*, pp. 239–256, 1989.

[203] J. Von Neumann, *Some matrix-inequalities and metrization of matric space*. JSTOR, 1937.

[204] L. Mirsky, “A trace inequality of John von Neumann,” *Monatshefte für mathematik*, vol. 79, pp. 303–306, 1975.

[205] H. Attouch, J. Bolte, and B. F. Svaiter, “Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods,” *Mathematical Programming*, vol. 137, pp. 91–129, 2013.

[206] T. Sun, H. Jiang, and L. Cheng, “Global convergence of proximal iteratively reweighted algorithm,” *Journal of Global Optimization*, vol. 68, pp. 815–826, 2017.

[207] A. Ruhe, “Perturbation bounds for means of eigenvalues and invariant subspaces,” *BIT Numerical Mathematics*, vol. 10, pp. 343–354, 1970.

[208] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[209] Z. Allen-Zhu, Y. Li, and Y. Liang, “Learning and generalization in overparameterized neural networks, going beyond two layers,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 6158–6169, 2019.

[210] R. S. Varga, “Geršgorin-type theorems for partitioned matrices,” in *Geršgorin and His Circles*, Springer, 2004, pp. 155–187.

[211] S. Smale and D.-X. Zhou, “Learning theory estimates via integral operators and their approximations,” *Constructive approximation*, vol. 26, no. 2, pp. 153–172, 2007.

[212] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, *et al.*, *A distribution-free theory of nonparametric regression*. Springer, 2002, vol. 1.

[213] D. Haussler, “Decision theoretic generalizations of the pac model for neural net and other learning applications,” in *The Mathematics of Generalization*, CRC Press, 2018, pp. 37–116.

[214] V. Maiorov and J. Ratsaby, “On the degree of approximation by manifolds of finite pseudo-dimension,” *Constructive approximation*, vol. 15, no. 2, pp. 291–300, 1999.