LEARNING UNDER IMPLICIT BIAS AND DATA BIAS

A Dissertation

by

JIANGYUAN LI

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond K. W. Wong
Committee Members,	Debdeep Pati
	Xianyang Zhang
	Kurt K. Zhang
Head of Department,	Brani Vidakovic

August 2023

Major Subject: Statistics

ABSTRACT

Modern machine learning tasks often involve the training of over-parameterized models and the challenge of addressing data bias. However, despite recent advances, there remains a significant knowledge gap in these areas. This thesis aims to push the boundaries of our understanding by exploring the implicit bias of neural network training and proposing strategies for mitigating data bias in matrix completion.

In the first result, we study the implicit regularization of gradient descent on a diagonally linear neural network with general depth-N under a realistic setting of noise and correlated designs. We characterize the impact of depth and early stopping and show that for a general depth parameter N, gradient descent with early stopping achieves minimax optimal sparse recovery with sufficiently small initialization and step size. In particular, we show that increasing depth enlarges the scale of working initialization and the early-stopping window so that this implicit sparse regularization effect is more likely to take place.

Continuing our exploration of implicit bias, our second main result introduces a novel neural reparametrization known as the "diagonally grouped linear neural network". This reparametrization exhibits a fascinating property wherein gradient descent, operating on the squared regression loss without explicit regularization, biases towards solutions with a group sparsity structure. In contrast to many existing works in understanding implicit regularization, we prove that our training trajectory cannot be simulated by mirror descent. Compared to existing bounds for implicit sparse regularization using diagonal linear networks, our analysis with the new reparameterization shows improved sample complexity in the general noise setting.

In our third result, we propose a pseudolikelihood approach for matrix completion with informative missing. We focus on a flexible and generally applicable missing mechanism, which contains both ignorable and nonignorable missing as special cases. We show that the regularized pairwise pseudolikelihood estimator can recover the low-rank matrix up to a constant shift and scaling while effectively mitigating the impact of data bias.

DEDICATION

To my parents and beloved ones.

ACKNOWLEDGMENTS

First and foremost, I want to express my deepest gratitude to my advisor Raymond K. W. Wong for his guidance and support throughout my doctoral study. Without his exceptional mentorship, the accomplishments presented in this thesis would not have been possible. His enthusiasm, patience and faith in my abilities served as a constant source of inspiration for me to strive towards becoming a better version of myself. Not only did he teach me how to do good research, but he also taught me how to face the ups and downs in life. The lessons I learned from him will undoubtedly continue to benefit me for the remainder of my life.

I would also like to extend my thanks to Debdeep Pati and Xianyang Zhang, whose profound knowledge of statistics and machine learning has greatly influenced the quality and depth of this thesis. Their insightful comments and feedback have played a significant role in enhancing the overall work.

Furthermore, I am deeply grateful to Kurt K. Zhang, Jianhua Z. Huang, and Yu Ding for their guidance during the initial two years of my graduate study. Their support and encouragement, especially the days I spent working in Dr. Zhang's Lab as a research assistant, have been instrumental in fostering my achievements in interdisciplinary statistical research.

I would like to express my sincere appreciation to Simon Foucart for introducing me to the captivating world of mathematical data science. His knowledge and guidance have been invaluable in shaping my research pursuits. I would also like to thank my long-term collaborators, Chinmay Hegde, Thanh V. Nguyen, and KC Gary Chan, for their invaluable contributions and collaborative efforts. Their insights and expertise have greatly enriched the outcomes of this thesis.

To my beloved parents and friends, I want to express my gratitude for their unconditional love and unwavering support. Life becomes more vibrant and meaningful with your presence. Additionally, I extend my thanks to the dedicated staff in the Department of Statistics for creating a pleasant and nurturing environment during my graduate study.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This was supported by a dissertation committee consisting of Professors Raymond K. W. Wong (advisor), Debdeep Pati and Xianyang Zhang of the Department of Statistics and Professor Kurt K. Zhang of the Department of Nutrition.

All the research work of the dissertation was completed by the student as the first author.

Funding Sources

Graduate study was supported by a teaching assistantship from the Department of Statistics at Texas A&M University.

NOMENCLATURE

GD	Gradient Descent
GF	Gradient Flow
PGD	Projected Gradient Descent
NN	Neural Network
LNN	Linear Neural Network
CNN	Convolutional Neural Network
LCNN	Linear Convolutional Neural Network
DLNN	Diagonal Linear Neural Network
DGLNN	Diagonally Grouped Linear Neural Network
SNR	Signal-to-Noise Ratio
RIP	Restricted Isometry Property
SDP	Semidefinite Programming
NP	Non-deterministic Polynomial-time
KKT	Karush-Kuhn-Tucker
IRMAE	Implicit Rank Minimizing Auto Encoder
GAE	Grouped Auto Encoder
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
MSE	Mean Squared Error
MAE	Mean Absolute Error
SVD	Singular Value Decomposition

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

CONTRIBUTORS AND FUNDING SOURCES v

NOMENCLATURE vi

TABLE OF CONTENTS vii

LIST OF FIGURES x

LIST OF TABLES xii

1. INTRODUCTION 1

2. IMPLICIT REGULARIZATION FOR SPARSITY 4

    2.1 Introduction 4
    2.2 Setup 7
    2.3 Main Results 10
    2.4 Proof Ingredients 15
        2.4.1 A Simplified Analysis 15
        2.4.2 Proof Sketch 16
    2.5 Simulation Study 18
    2.6 Conclusions and Future Work 20
    2.7 Implicit Regularization for Dictionary Sparsity 21

3. IMPLICIT REGULARIZATION FOR GROUP SPARSITY 25

    3.1 Introduction 25
    3.2 Setup 29
    3.3 Analysis of Gradient Flow 31
        3.3.1 First Attempt: Mirror Flow 31
        3.3.2 Layer Balancing and Gradient Flow 33
    3.4 Gradient Descent with Weight Normalization 34
    3.5 Simulation Study 38
    3.6 Discussion 40

4. MATRIX COMPLETION WITH INFORMATIVE MISSING 41

4.1 Introduction 41
4.2 Preliminaries 43
4.3 Main Results 45
4.4 Numerical Experiments 48

5. SUMMARY 52

REFERENCES 54

APPENDIX A. SUPPLEMENTARY MATERIAL FOR CHAPTER II 68

A.1 Proof for Non-negative Signals 68
A.1.1 Setup 68
A.1.2 The Key Propositions 70
A.1.3 Technical Lemmas 72
A.1.4 Proof for Non-negative Signals 73
A.2 Multiplicative Update Sequences with General Order N 77
A.2.1 Error Growth 77
A.2.2 Understanding 1-d Case 79
A.2.2.1 Basic Setting 79
A.2.2.2 Dealing with Bounded Errors $\mathbf{b}_t$ 87
A.2.3 Dealing with Negative Targets 93
A.3 Proof of Propositions and Technical Lemmas 96
A.3.1 Proof of Proposition 1 96
A.3.2 Proof of Proposition 2 98
A.3.3 Proof of Technical Lemmas 99
A.4 Proof of Theorems in Chapter 2.3 100
A.4.1 Proof of Theorem 1 100
A.4.2 Proof of Corollary 1 103
A.4.3 Proof of Theorem 2 103
A.4.4 Proof of Remark 2 105
A.5 Experiments on MNIST 105

APPENDIX B. SUPPLEMENTARY MATERIAL FOR CHAPTER III 107

B.1 Geometric properties of the parametrization 107
B.2 Proof for Analysis of Gradient Flow 109
B.3 Analysis of gradient descent 116
B.3.1 Monotonic updates 117
B.3.2 Updates with bounded perturbations 117
B.3.3 Analysis of perturbations 122
B.3.4 Error analysis outside the support 128
B.4 Proof of Theorems in Chapter 3.4 129
B.4.1 Proof of Theorem 5 129

B.4.2 Proof for Corollary 2 132
B.4.3 Convergence for algorithm 2 132
B.5 More numerical results 135
B.5.1 Stability issue of Algorithm 1 and standard GD 135
B.5.2 Autoencoder with grouping layer 136
B.5.3 Experiments with Gaussian measurements 139

APPENDIX C. SUPPLEMENTARY MATERIAL FOR CHAPTER IV 141

C.1 Proof of Theorem 7 141
C.2 Proof of Corollary 7 144
C.3 Useful lemmas 144

LIST OF FIGURES

FIGURE Page

2.1 The coordinate path with the same initialization $\alpha = 0.005$ and step size $\eta = 0.01$ for $N = 2, 3, 4$. Reprinted with permission from [1]. 14

2.2 Coordinates paths for different choice of $N = 2, 3, 5$ with $\alpha^N = 10^{-6}$ and $\eta = 1/(5N^2)$. Reprinted with permission from [1]. 19

2.3 The effect of N on the initialization α^N with $\eta = 1/(5N^2)$. Reprinted with permission from [1]. Reprinted with permission from [1]. 19

2.4 $\log$-ℓ_2 error of $N = 2, 3, 4$ with the fixed step size $\eta = 0.01$. Reprinted with permission from [1]. 20

2.5 Coordinates paths for $N = 2, 3, 5$. The entries of $\mathbf{w}^\star$ on the support S are now $[1, 2, 3, 4]$. The initialization is $\alpha^N = 10^{-4}$ and the step size is $\eta = 10^{-3}$ for all N. .. 20

2.6 $\log$-ℓ_2 error of $N = 2, 3, 4$ for a ridge regression setting. The ridge regression solution is selected by 5-fold cross validation. Reprinted with permission from [1]. . 21

2.7 Implicit regularization for dictionary sparsity. 23

2.8 Learning to denoise. 24

3.1 An illustration of the two architectures for standard and group sparse regularization. Reprinted with permission from [2]. 27

3.2 Convergence of Algorithm 1. The entries on the support are all 10. Reprinted with permission from [2]. 38

3.3 Convergence of Algorithm 2. The entries on the support are from 5 to 13. Reprinted with permission from [2]. 39

3.4 Comparison with reparameterization using standard sparsity. $n = 100, p = 500$. Reprinted with permission from [2]. 39

3.5 Degenerate case when each group size is 1. The $\log \ell_2$-error plot is repeated 30 times, and the mean is depicted. The shaded area indicates the region between the 25^{th} and 75^{th} percentiles. Reprinted with permission from [2]. 40

4.1 Observation bias. 50

4.2 The recovered entries are left skewed from other methods. 51

A.1 Experiments with different choice depth parameter N. Reprinted with permission from [1]. 106

B.1 Numerical instability of algorithm 1. Reprinted with permission from [2]. 135

B.2 Gradient descent without weight normalization. Reprinted with permission from [2]. 136

B.3 Implicit rank-minimizing autoencoder. Reprinted with permission from [2]. 137

B.4 Implicit rank-minimizing autoencoder with grouping layers. Reprinted with permission from [2]. 137

B.5 Linear interpolations between data points on the MNIST dataset. GAE4/8 stands for grouped autoencoder with 4/8 groups. Reprinted with permission from [2]. 138

B.6 Convergence of algorithm 2 with Gaussian measurements. Reprinted with permission from [2]. 139

B.7 Comparisons with proximal gradient descent and iterative regularization. Reprinted with permission from [2]. 140

LIST OF TABLES

TABLE Page

2.1 Comparisons with closely related recent work. GF/GD: gradient flow/descent, respectively. Reprinted with permission from [1]. 6

3.1 Comparisons to related work on implicit and explicit regularization. Here, GD stands for gradient descent, (D)LNN/CNN for (diagonal) linear/convolutional neural network, and DGLNN for diagonally grouped linear neural network. Reprinted with permission from [2]. 26

4.1 Test root mean squared errors (TRMSE), test mean absolute errors (TMAE). 50

B.1 Number of parameters of hidden layers in latent space. Reprinted with permission from [2]. 139

B.2 Comparisons of MSE (mean squared error) on test set. Reprinted with permission from [2]. 140

1. INTRODUCTION

In this thesis, we study the learning algorithms under implicit bias and data bias. The recent advancement of applications of large-scale models and big data has brought new challenges in understanding the generalization performance of over-parametrized models and mitigating data bias. In the first line of contributions, we study the implicit bias of gradient descent in training linear neural networks and show that it can exhibit (group) sparsity. In the second line of contributions, we propose a pseudolikelihood approach for matrix completion with informative missing.

Deep neural networks [3] are emerging as the dominant choice across several domains, from natural language processing [4, 5], to computer vision [6, 7] and reinforcement learning [8]. Besides that, it has shown promise to be applied in much more tasks such as signal processing [9, 10], time series [11, 12, 13], medical analysis [14, 15, 16] and mathematical physics [17, 18]. Large-scale neural networks are the keys to accomplishing more and more difficult tasks.

The large size of deep neural networks not only requires significant memory and computation costs [19, 20] but also makes them prone to overfitting. One particular characteristic of these architectures is that they generalize well on unseen data despite being over-parametrized [21]. The training of deep neural works relies on gradient descent and its variants, entailing many tricks and hands-on experiences [22]. The efficiency and accuracy would be highly dependent on the optimization algorithms [23, 24, 25]. One consensus in learning theory suggests that one should use a model, just expressive enough to avoid overfitting [26]. The idea goes back to the *Ocacam's Razor* philosophical principle, which states that the model is trained to perform well on the training data, but should be as simple as possible.

Instead of having a small number of parameters, simplicity in neural networks usually refers to minimizing a certain measure of complexity by adding a regularization term to the objective function. For example, weight decay [27] has been commonly used to prevent overfitting. However, over-parametrized neural networks seem to generalize well even when trained without explicit regularization [21]. Therefore, the implicit regularization/bias provided by gradient descent

optimization is widely believed to be one of the keys to deep neural networks' generalization ability [28]. Characterizing such bias has been a subject of extensive research. Many recent theoretical efforts have revisited traditional, well-understood problems such as linear regression [29, 1, 30], matrix factorization [31, 32, 33] and tensor decomposition [34, 35], from the perspective of neural network training. For nonlinear models with squared error loss, [36] and [37] study the implicit bias of gradient descent in wide depth-2 ReLU networks with input dimension 1. Other works [38, 39, 40] show that gradient descent biases the solution towards the max-margin (or minimum ℓ_2-norm) solutions over separable data. Many recent theoretical efforts have revisited traditional, well-understood problems such as linear regression [29, 1, 30], matrix factorization [31, 32, 33] and tensor decomposition [34, 35], from the perspective of neural network training. For nonlinear models with squared error loss, [36] and [37] study the implicit bias of gradient descent in wide depth-2 ReLU networks with input dimension 1. Other works [38, 39, 40] show that gradient descent biases the solution towards the max-margin (or minimum ℓ_2-norm) solutions over separable data.

In Chapter 2 [1] and Chapter 3 [2], we focus on the implicit bias of gradient descent in linear regression, which amounts to linear neural networks. We extend the existing results to general depth N and study how depth affects the gradient dynamics as well as implicit bias. Moreover, we prove a new type of implicit regularization for structured sparsity in linear neural networks. In contrast to many existing works, we show that the training trajectory cannot be simulated by mirror descent. We analyze the gradient dynamics of the corresponding regression problem in the general noise setting and obtain minimax-optimal error rates.

Data bias in machine learning refers to the presence of errors caused by the observed distribution that does not accurately represent the underlying true distribution. When a dataset is biased, it fails to reflect the intended use case of a model, leading to skewed outcomes, reduced accuracy, and analytical errors. The impact of biased algorithms extends beyond model accuracy and can raise concerns related to ethics, fairness, and inclusion. The goal during machine learning model development is to reduce both data bias and data variance as much as possible in order to get the

most accurate outputs. Extensive research has been conducted on learning algorithms that address data bias, including transfer learning [41] and fairness [42].

In Chaper 4, we investigate a contemporary problem in high-dimensional missing data under the influence of data bias. Specifically, we focus on matrix completion with informative missing, where the observed entries are biased. The problem of matrix completion arises in various domains, such as collaborative filtering, multi-class learning, system identification, global positioning, and computer vision. For instance, in computer vision, missing pixels may be encountered in digital images, while collaborative filtering involves predicting user preferences by gathering information from multiple users. Despite notable advancements in the field over the past two decades [43, 44, 45], most existing research on matrix completion primarily addresses scenarios where the missing probability is not dependent on the value. We propose a penalized pairwise pseudolikelihood approach for matrix completion with informative missing. We demonstrate that our method effectively handles data bias under a flexible and widely applicable assumption [46, 47, 48]. The efficacy of our method is validated via numerical experiments.

2. IMPLICIT REGULARIZATION FOR SPARSITY*

2.1 Introduction

Motivation. Central to recent research in learning theory is the insight that the choice of optimization algorithms plays an important role in model generalization [21, 23, 49]. A widely adopted view is that (stochastic) gradient descent — the most popular optimization algorithm in machine learning — exhibits some implicit form of regularization. Indeed for example, in the classical under-determined least squares setting, gradient descent (with small step size) starting from the origin converges to the model with minimum Euclidean norm. Similar implicit biases are also observed in deep neural network training in which the networks typically have many more parameters than the sample size. There, gradient descent without explicit regularization finds solutions that not only interpolate the training data points but also generalize well on test sets [23, 50, 51, 52, 53].

This insight, combined with the empirical success stories of deep learning, has sparked significant interest among theoretical researchers to rigorously understand implicit regularization. The majority of theoretical results focus on well-understood problems such as regression with linear models [54, 55, 56, 29, 30, 57] and matrix factorization [58, 31, 32, 33], and show that the parametrization (or architecture) of the model plays a crucial role. For the latter, Gunasekar et al. [31] conjectured that gradient descent on factorized matrix representations converges to the solution with minimum nuclear norm. The conjecture was partially proved by Li et al. [32] under the Restricted Isometry Property (RIP) and the absence of noise. Arora et al. [33] further show the same nuclear-norm implicit bias using depth-N linear networks (i.e., the matrix variable is factorized into N components).

Parallel work on nonlinear models and classification [51, 55] has shown that gradient descent biases the solution towards the max-margin/minimum ℓ_2-norm solutions over separable data. The

*Reprinted with permission from [1]. This is a joint work with Thanh V. Nguyen, Chinmay Hegde and Raymond K. W. Wong. Copyright 2021 by the authors.

scale of initialization in gradient descent leads to two learning regimes (dubbed "kernel" and "rich") in linear networks [59], shallow ReLU networks [36] and deep linear classifiers [60]. Li et al. [61] showed that depth-2 network requires an exponentially small initialization, whereas depth-N network ($N \geq 3$) only requires a polynomial small initialization, to obtain low-rank solution in matrix factorisation. Woodworth et al. [59] obtained a similar result for high dimensional sparse regression.

The trend in the large majority of the above works has been to capture implicit regularization of gradient descent using some type of norm with respect to the working parametrization [59, 62, 63, 64]. On the other hand, progress on understanding the *trajectory* of gradient descent has been somewhat more modest. [29, 30] study the sparse regression problem using quadratic and Hadamard parametrization respectively and show that gradient descent with small initialization and careful *early stopping* achieves minimax optimal rates for sparse recovery. Unlike [32, 59] that study noiseless settings and require no early stopping, [29, 30] mathematically characterize the role of early stopping and empirically show that it may be necessary to prevent gradient descent from over-fitting to the noise. These works suggest that the inductive bias endowed by gradient descent may be influenced not only by the choice of parametrization, *but also algorithmic choices* such as initialization, learning rate, and the number of iterations. However, our understanding of such gradient dynamics is incomplete, *particularly* in the context of deep architectures; see Table 2.1 for some comparisons.

Contributions. Our focus in this work is the implicit regularization of (standard) gradient descent for high dimensional sparse regression, namely *implicit sparse regularization*. Let us assume a ground-truth sparse linear model and suppose we observe n noisy samples $(\mathbf{x}_i, y_i)$, such that $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$; a more formal setup is given in Section 2.2. Using the samples, we consider gradient descent on a squared loss $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ with no explicit sparsity regularization. Instead, we write the parameter vector $\mathbf{w}$ in the form $\mathbf{w} = \mathbf{u}^N - \mathbf{v}^N$ with $N \geq 2$. Now, the regression function $f(\mathbf{x}, \mathbf{u}, \mathbf{v}) = \langle \mathbf{x}, \mathbf{u}^N - \mathbf{v}^N \rangle$ can be viewed as a depth-N *diagonal linear network* [59]. Minimizing the (now non-convex) loss over $\mathbf{u}$ and $\mathbf{v}$ with gradient descent is then analogous to training this

depth-N network.

Our main contributions are the following. We characterize the impact of both the depth and early stopping for this non-convex optimization problem. Along the way, we also generalize the results of [29] for $N > 2$. We show that under a general depth parameter N and an incoherence assumption on the design matrix, gradient descent with early stopping achieves minimax optimal recovery with sufficiently small initialization $\mathbf{w}_0$ and step size η. The choice of step size is of order $O(1/N^2)$. Moreover, the upper bound of the initialization, as well as the early-stopping window, increase with N, suggesting that depth leads to a more accessible generalizable solution on gradient trajectories.

Table 2.1: Comparisons with closely related recent work. GF/GD: gradient flow/descent, respectively. Reprinted with permission from [1].

	Design Matrix	In Noise	Depth	Early Stopping	GD vs. GF	Remark
Vaskevicius et al. (2020) [29]	RIP	✓	$N = 2$	✓	GD	recovery
Gissin et al. (2020) [57]	uncorrelated	✗	$N = 2, N > 2$	✗	GF	interpolation
Woodworth et al. (2020) [59]	✗	✗	$N = 2, N > 2$	✗	GF	interpolation
This work	μ-coherence	✓	$N > 2$	✓	GD	recovery

Techniques. At a high level, our work continues the line of work on implicit bias initiated in [29, 30, 59] and extends it to the deep setting. Table 2.1 highlights key differences between our work and [29, 57, 59]. Specifically, Woodworth et al. [59] study the *interpolation* given by the gradient flow of the squared-error loss function. Vaskevicius et al. [29] analyze the finite gradient descent and characterize the implicit sparse regularization on the *recovery* of true parameters with $N = 2$. Lastly, Gissin et al. [57] discover the incremental learning dynamic of gradient flow for general N but in an idealistic model setting where $\mathbf{u} \succeq 0, \mathbf{v} = 0, \boldsymbol{\xi} = 0$, uncorrelated design and with infinitely many samples.

At first glance, one could attempt a straightforward extension of the proof techniques in [29] to general settings of $N > 2$. However, this turns out to be very challenging. Consider even the simplified case where the true model $\mathbf{w}^\star$ is non-negative, the design matrix is unitary (i.e.,

$n^{-1}\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$), and the noise is absent ($\boldsymbol{\xi} = 0$); this is the setting studied in [57]. For each entry w_i of $\mathbf{w}$, the t^{th} iterate of gradient descent over the depth-N reparametrized model is given by:

$$w_{i,t+1} = w_{i,t}\left(1 + w_{i,t}^{1-\frac{2}{N}}(w_i^\star - w_{i,t})\right)^N,$$

which is no longer a simple multiplicative update. As pointed out in [57] (see their Appendix C), the recurrence relation is not analytically solvable due to the presence of the (pesky) term $w_{i,t}^{1-\frac{2}{N}}$ when $N > 2$. Moreover, this extra term $w_{i,t}^{1-\frac{2}{N}}$ leads to widely divergent growth rates of weights with different magnitudes, which further complicates analytical bounds. To resolve this and rigorously analyze the dynamics for $N > 2$, we rely on a novel first order, continuous approximation to study growth rates without requiring additional assumptions on gradient flow, and carefully bound the approximation error due to finite step size; see Section 2.4.

2.2 Setup

Sparse regression/recovery. Let $\mathbf{w}^\star \in \mathbb{R}^p$ be a p-dimensional sparse vector with k non-zero entries. Assume that we observe n data points $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ such that $y_i = \langle \mathbf{x}_i, \mathbf{w}^\star \rangle + \xi_i$ for $i = 1, \ldots, n$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ is the noise vector. We do not assume any particular scaling between the number of observations n and the dimension p. Due to the sparsity of $\mathbf{w}^\star$, however, we allow $n \ll p$.

The linear model can be expressed in the matrix-vector form:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \boldsymbol{\xi}, \tag{2.1}$$

with the $n \times p$ design matrix $\mathbf{X} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top]^\top$, where $\mathbf{x}_i$ denotes the i^{th} *row* of $\mathbf{X}$. We also denote $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$, where $\mathbf{X}_i$ denotes the i^{th} *column* of $\mathbf{X}$.

The goal of sparse regression is to estimate the unknown, sparse vector $\mathbf{w}^\star$ from the observations. Over the past two decades, this problem has been a topic of active research in statistics and signal processing [65]. A common approach to sparse regression is penalized least squares with

sparsity-induced regularization such as ℓ_0 or ℓ_1 penalties/constraints, leading to several well-known estimators [65, 66, 67] and algorithms [68, 69]. Multiple estimators enjoy optimal statistical and algorithmic recovery guarantees under some conditions of the design matrix $\mathbf{X}$ (e.g., RIP [70]) and the noise $\boldsymbol{\xi}$.

We deviate from the standard penalized least squares formulation and instead learn $\mathbf{w}^*$ via a polynomial parametrization:

$$\mathbf{w} = \mathbf{u}^N - \mathbf{v}^N, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^p,$$

where $N \geq 2$ and $\mathbf{z}^N = [z_1^N, \ldots, z_p^N]^\top$ for any $\mathbf{z} = [z_1, \ldots, z_N]^\top \in \mathbb{R}^p$. The regression function $f(\mathbf{x}, \mathbf{u}, \mathbf{v}) = \langle \mathbf{x}, \mathbf{u}^N - \mathbf{v}^N \rangle$ induced by such a parametrization is equivalent to a N-layer diagonal linear network [59] with $2p$ hidden neurons and the diagonal weight matrix shared across all layers.

Given the data $\{\mathbf{X}, \mathbf{y}\}$ observed in (2.1), we analyze gradient descent with respect to the new parameters $\mathbf{u}$ and $\mathbf{v}$ over the mean squared error loss without explicit regularization:

$$\mathcal{L}(\mathbf{u}, \mathbf{v}) = \frac{1}{n} \left\| \mathbf{X}(\mathbf{u}^N - \mathbf{v}^N) - \mathbf{y} \right\|_2^2, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^p.$$

Even though the loss function yields the same value for the two parametrizations, $\mathcal{L}(\mathbf{u}, \mathbf{v})$ is non-convex in $\mathbf{u}$ and $\mathbf{v}$. Unlike several recent studies in implicit regularization for matrix factorization and regression [32, 59, 57], we consider the noisy setting, which is more realistic and leads to more insights into the bias induced during the optimization. Because of noise, the loss evaluated at the ground truth (i.e., any $\mathbf{u}, \mathbf{v}$ such that $\mathbf{w}^\star = \mathbf{u}^N - \mathbf{v}^N$) is not necessarily zero or even minimal.

Gradient descent. The standard gradient descent update over $\mathcal{L}(\mathbf{u}, \mathbf{v})$ reads as:

$$\begin{aligned} &\mathbf{u}_0 = \mathbf{v}_0 = \alpha \mathbf{1}, \\ &(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) = (\mathbf{u}_t, \mathbf{v}_t) - \eta \frac{\partial \mathcal{L}(\mathbf{u}_t, \mathbf{v}_t)}{\partial (\mathbf{u}_t, \mathbf{v}_t)}, \quad t = 0, 1, \ldots. \end{aligned} \tag{2.2}$$

Here, $\eta > 0$ is the step size and $\alpha > 0$ is the initialization of $\mathbf{u}, \mathbf{v}$. In general, we analyze the algorithm presented in (2.2), and at each step t, we can estimate the signal of interest by

simply calculating $\mathbf{w}_t = \mathbf{u}_t^N - \mathbf{v}_t^N$. We consider constant initialization for simplicity sake. Our results apply for random initialization concentrating on a small positive region with a probabilistic statement.

Vaskevicius et al. [29] establish the implicit sparse regularization of gradient descent for $N = 2$ and show minimax optimal recovery, provided sufficiently small α and early stopping. Our work aims to generalize that result to $N > 2$ and characterize the role of N in convergence.

Notation. We define $S = \{i \in \{1, \dots, p\} : w_i^\star \neq 0\}$ and $S^c = \{1, \dots, p\} \backslash S$. The largest and smallest absolute value on the support is denoted as $w_{\max}^\star = \max_{i \in S} |w_i^\star|$ and $w_{\min}^\star = \min_{i \in S} |w_i^\star|$. We use $\mathbf{1}$ to denote the vector of all ones and $\mathbf{1}_S$ denotes the vector whose elements on S are all one and 0 otherwise. Also, $\odot$ denotes coordinate-wise multiplication. We denote $\mathbf{s}_t = \mathbf{1}_S \odot \mathbf{w}_t$ and $\mathbf{e}_t = \mathbf{1}_{S^c} \odot \mathbf{w}_t$ meaning the signal part and error part at each time step t. We use $\wedge$ and $\vee$ to denote the pointwise maximum and minimum. The coordinate-wise inequalities are denoted as $\succcurlyeq$. We denote inequalities up to multiplicative absolute constants by $\lesssim$, which means that they do not depend on any parameters of the problem.

Definition 1. *Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a matrix with ℓ_2-normalized columns $\mathbf{X}_1, \dots, \mathbf{X}_p$, i.e., $\|\mathbf{X}_i\|_2 = 1$ for all i. The coherence $\mu = \mu(\mathbf{X})$ of the matrix $\mathbf{X}$ is defined as*

$$\mu \coloneqq \max_{1 \leq i \neq j \leq p} |\langle \mathbf{X}_i, \mathbf{X}_j \rangle|.$$

The matrix $\mathbf{X}$ is said to be satisfying μ-incoherence.

The coherence is a measure for the suitability of the measurement matrix in compressive sensing [71, 72]. In general, the smaller the coherence, the better the recovery algorithms perform. There are multiple ways to construct a sensing matrix with low-incoherence. One of them is based on the fact that sub-Gaussian matrices satisfy low-incoherence property with high probability [73, 74]. In contrast to the coherence, the Restricted Isometry Property (RIP) is a powerful performance measure for guaranteeing sparse recovery and has been widely used in many contexts. However, verifying the RIP for deterministically constructed design matrices is NP-hard. On the

other hand, coherence is a computationally tractable measure and its use in sparse regression is by now classical [74, 75]. Therefore, in contrast with previous results [29] (which assumes RIP), the assumptions made in our main theorems are verifiable in polynomial time.

2.3 Main Results

We now introduce several quantities that are relevant for our main results. First, the condition number $r \coloneqq w^\star_{\max}/w^\star_{\min}$ plays an important role when we work on the incoherence property of the design matrix. Next, we require an upper bound on the initialization α, which depends on the following terms:

$$\Phi(w^\star_{\max}, w^\star_{\min}, \epsilon, N) \coloneqq \left(\frac{1}{8}\right)^{2/(N-2)} \wedge \left(\frac{(w^\star_{\max})^{(N-2)/N}}{\log \frac{w^\star_{\max}}{\epsilon}}\right)^{2/(N-2)} \wedge \left(\frac{(w^\star_{\min})^{(N-2)/N}}{\log \frac{w^\star_{\min}}{\epsilon}}\right)^{4/(N-2)},$$

$$\Psi(w^\star_{\min}, N) \coloneqq (2 - 2^{\frac{N-2}{N}})^{\frac{1}{N-2}} (w^\star_{\min})^{\frac{1}{N}} \wedge 2^{\frac{3}{N}} (2^{\frac{1}{N}} - 1)^{\frac{1}{N-2}} (w^\star_{\min})^{\frac{1}{N}}.$$

Finally, define

$$\zeta \coloneqq \frac{1}{5} w^\star_{\min} \vee \frac{200}{n} \left\| \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty \vee 200\epsilon.$$

We are now ready to state the main theorem:

Theorem 1. *Suppose that $k \geq 1$ and $\mathbf{X}/\sqrt{n}$ satisfies μ-incoherence with $\mu \lesssim 1/kr$. Take any precision $\epsilon > 0$, and let the initialization be such that*

$$0 < \alpha \leq \left(\frac{\epsilon}{p+1}\right)^{4/N} \wedge \Phi(w^\star_{\max}, w^\star_{\min}, \epsilon, N) \wedge \Psi(w^\star_{\min}, N). \tag{2.3}$$

For any iteration t that satisfies

$$T_l(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon) \leq t \leq T_u(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon), \tag{2.4}$$

where $T_l(\cdot)$ and $T_u(\cdot)$ are given in (A.13) *of the Appendix, the gradient descent algorithm* (2.2)

with step size $\eta \leq \frac{\alpha^N}{8N^2(w^\star_{\max})^{(3N-2)/N}}$ *yields the iterate* $\mathbf{w}_t$ *with the following property:*

$$|w_{t,i} - w_i^\star| \lesssim \begin{cases} \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \epsilon & \text{if } i \in S \text{ and } w^\star_{\min} \lesssim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \epsilon, \\ \left|\frac{1}{n}(\mathbf{X}^\mathsf{T}\boldsymbol{\xi})_i\right| \vee k\mu \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi} \odot \mathbf{1}_S\right\|_\infty \vee \epsilon & \text{if } i \in S \text{ and } w^\star_{\min} \gtrsim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \epsilon, \\ \alpha^{N/4} & \text{if } i \notin S. \end{cases} \tag{2.5}$$

In the special case $\mathbf{w}^\star = \mathbf{0}$*, if* $\alpha \leq \left(\frac{\epsilon}{p+1}\right)^{4/N}$*,* $\eta \leq \frac{1}{N(N-1)\zeta\alpha^{(N-2)/2}}$ *and* $t \leq T_u(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon)$*, then we have* $|w_{t,i} - w_i^\star| \leq \alpha^{N/4}, \forall i$*.*

Theorem 1 states the convergence of the gradient descent algorithm (2.2) in ℓ_∞-norm. The exact formula of $T_l(\cdot)$ and $T_u(\cdot)$ is omitted here due to the space limitation. We ensure that $T_u(\cdot) > T_l(\cdot)$ so that there indeed exists some epochs to early stop at. The error bound on the signal is invariant to the choice of $N \geq 2$, and the overall bound generalizes that of [29] for $N = 2$. We also establish the convergence result in ℓ_2-norm in the following corollary:

Corollary 1. *Suppose the noise vector* $\boldsymbol{\xi}$ *has independent* σ^2*-sub-Gaussian entries and*

$$\epsilon = 2\sqrt{\frac{\sigma^2 \log(2p)}{n}}.$$

Under the assumptions of Theorem 1, the gradient descent algorithm (2.2) *would produce iterate* $\mathbf{w}_t$ *satisfying* $\|\mathbf{w}_t - \mathbf{w}^\star\|_2^2 \lesssim (k\sigma^2 \log p)/n$ *with probability at least* $1 - 1/(8p^3)$*.*

Note that the error bound we obtain is minimax-optimal, which is the same as [29] in the $N = 2$ case. However, with some calculation, the sample complexity we obtain here is $n \gtrsim k^2 r^2$, while the sample complexity in [29] is $n \gtrsim k^2 \log^2 r \log p/k$. Although neither our work nor [29] achieved the optimal sample complexity $k \log p/k$, the goal of this work is to understand how the depth parameter affects implicit sparse regularization.

Let us now discuss the implications of Theorem 1 and the role of initialization and early stopping:

(a) Requirement on initialization. To roughly understand the role of initialization and the effect of N, we look at the non-negative case where $\mathbf{w}^\star \succcurlyeq 0$ and $\mathbf{w} = \mathbf{u}^N$. This simplifies our discussion while still capturing the essential insight of the general setting. At each step, the "update" on $\mathbf{w}$ can be translated from the corresponding gradient update of $\mathbf{u} = \mathbf{w}^{1/N}$ as

$$\begin{aligned} \mathbf{w}_0 &= \alpha^N \mathbf{1}, \\ \mathbf{w}_{t+1} &= \mathbf{w}_t \odot \left(\mathbf{1} - \frac{2N\eta}{n} \left(\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{w}_t - \mathbf{w}^*) - \mathbf{X}^\mathsf{T}\boldsymbol{\xi} \right) \odot \mathbf{w}_t^{(N-2)/N} \right)^N . \end{aligned} \tag{2.6}$$

In order to guarantee the convergence, we require the initialization α to be sufficiently small so the error outside the support can be controlled. On the other hand, too small initialization slows down the convergence of the signal. Interestingly, the choice of N affects the allowable initialization α that results in guarantees on the entries inside and outside the support.

Specifically, the role of N is played by the term $\mathbf{w}_t^{(N-2)/N}$ in (2.6), which simply disappears as $N = 2$. Since this term only affects the update of $\mathbf{w}_{t+1}$ entry-wise, we only look at a particular entry w_t of $\mathbf{w}_t$. Let w_t represent an entry outside the support. For $N > 2$, the term $w_t^{(N-2)/N}$ is increasingly small as N increases and $w_t < 1$. Therefore, with a small initialization, it remains true that $w_t < 1$ for the early iterations. Intuitively, this suggests that the requirement on the upper bound of the initialization would become looser when N gets larger. This indeed aligns with the behavior of the upper bound we derive in our theoretical results. Since $\alpha = w_0^{1/N}$ increases naturally with N, we fix $w_0 = \alpha^N$ instead of α to mimic the same initialization in terms of w_0, for the following comparison.

We formalize this insight in Theorem 8 in Appendix A.1 and show the convergence of (2.6) under the special, non-negative case. Note that, in terms of initialization requirement, the only difference from Theorem 1 is that we no longer require the term $\Psi(w^\star_{\min}, N)$ in (2.3).

Remark 1. *We investigate how the depth N influences the requirement on initialization due to the change on gradient dynamics. We rewrite $\Phi(w^\star_{\max}, w^\star_{\min}, \epsilon, N)$ in terms of $w_0 = \alpha^N$, and therefore*

the upper bound for w_0 under the simplified setting of non-negative signals (Theorem 8) is

$$w_0 \leq \left(\frac{1}{8}\right)^{2N/(N-2)} \wedge \left(\frac{(w^\star_{\max})^{(N-2)/N}}{\log \frac{w^\star_{\max}}{\epsilon}}\right)^{2N/(N-2)} \wedge \left(\frac{(w^\star_{\min})^{(N-2)/N}}{\log \frac{w^\star_{\min}}{\epsilon}}\right)^{4N/(N-2)}.$$

We start by analyzing each term in the upper bound. First, we notice that $\left(\frac{1}{8}\right)^{2N/(N-2)}$ is increasing with respect to N. For the second term,

$$\left(\frac{(w^\star_{\max})^{(N-2)/N}}{\log \frac{w^\star_{\max}}{\epsilon}}\right)^{2N/(N-2)} = \frac{(w^\star_{\max})^2}{(\log \frac{w^\star_{\max}}{\epsilon})^{2N/(N-2)}},$$

the denominator gets smaller as N increases when we pick the error tolerance parameter ϵ small. Therefore, we get that the second term is getting larger as N increases. The last term

$$\left(\frac{(w^\star_{\min})^{(N-2)/N}}{\log \frac{w^\star_{\min}}{\epsilon}}\right)^{4N/(N-2)}$$

follows a similar argument. We see that it is possible to pick a larger initialization $w_0 = \alpha^N$ for larger N. We will demonstrate that below in our experiments.

(b) Early stopping. Early stopping is shown to be crucial, if not necessary, for implicit sparse regularization [29, 30]. Interestingly, [57, 59] studied the similar depth-N polynomial parametrization but did not realize the need of early stopping due to an oversimplification in the model. We will discuss this in details in Section 2.4.1. We are able to explicitly characterize the window of the number of iterations that are sufficient to guarantee the optimal result. In particular, we get a lower bound of the window size for early stopping to get a sense of how it changes with different N.

Theorem 2 (Informal)**.** *Define the early stopping window size as*

$$T_u(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon) - T_l(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon),$$

the difference between the upper bound and lower bound of the number of iterations in (2.4) *of Theorem 1. Fixing* α *and* η *for all* N*, the early stopping window size is increasing with* N *under mild conditions.*

We defer the formal argument and proof of Theorem 2 to Appendix A.4.3. We note that the window we obtain in Theorem 1 is not necessarily the largest window that allows the guarantee, and hence the early stopping window size can be effectively regarded a lower bound of that derived from the largest window. We note that a precise characterization of the largest window is difficult. Although we only show that this lower bound increases with N, we see that the conclusion matches empirically with the largest window. We show the coordinate path in Figure 2.1. The black line indicates the early stopping window for different $N = 2, 3, 4$. The blue line is the coordinate path for each entry on the support. The red line indicates the absolute value of the largest entry on the coordinate path outside the support. We use the orange line to indicate the requirement outside the support for early stopping. We can see that as N increases, the early stopping window increases and the error bound captures the time point that needs stopping quite accurately. The experimental details and more experiments about early stopping is presented in Section 2.5.

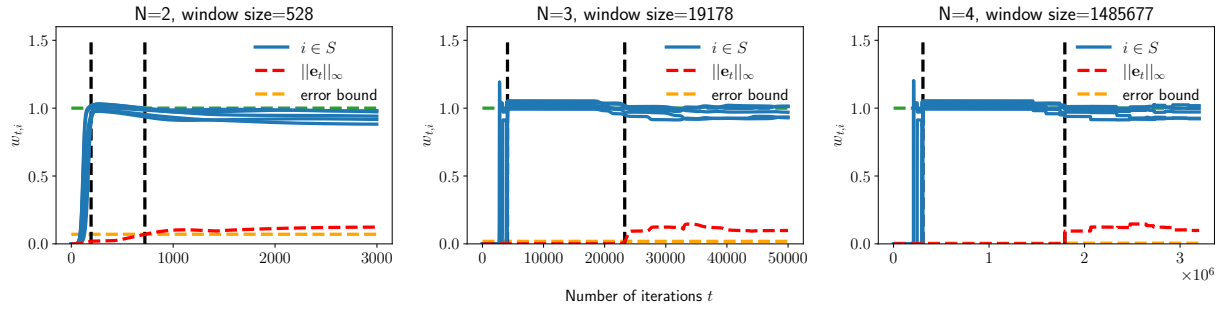


Figure 2.1: The coordinate path with the same initialization $\alpha = 0.005$ and step size $\eta = 0.01$ for $N = 2, 3, 4$. Reprinted with permission from [1].

Remark 2. *Similar to Theorem 2, we look at how initialization scale affects the early stopping window for any fixed $N > 2$. With η fixed, the early stopping window is increasing as the initialization α decreases.*

We defer the detailed calculation to Section A.4.4. This generalizes the finding that vanishing initialization increases the gap between the phase transition times in [58] from $N = 2$ to any $N > 2$.

2.4 Proof Ingredients

The goal of this work is to understand how generalization and gradient dynamics change with different $N > 2$. For $N = 2$, gradient descent yields both statistically and computationally optimal recovery under the RIP assumption [29]. The matrix formulation of the same type of parametrization is considered in the setting of low-rank matrix recovery, and exact recovery can be achieved in the noiseless setting [31, 32]. The key proof ingredient is to reduce the convergence analysis to one-dimensional iterates and differentiate the convergence on the support from the error outside the support. Before we get into that, we conduct a simplified gradient flow analysis.

2.4.1 A Simplified Analysis

Consider a simplified problem where the target signal $\mathbf{w}^\star$ is non-negative, $n^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{I}$ and the noise is absent. We omit the reparametrization of $\mathbf{v}^N$ like before and the gradient descent updates on $\mathbf{u}$ will be independent for each coordinate. The gradient flow dynamics of $\mathbf{w} = \mathbf{u}^N$ is derived as

$$\frac{\partial w_i}{\partial t} = \frac{\partial w_i}{\partial u_i}\frac{\partial u_i}{\partial t} = -\frac{\partial w_i}{\partial u_i}\frac{\partial \mathcal{L}}{\partial u_i} = 2N^2(w_i^\star - w_i)w_i^{2-\frac{2}{N}}, \tag{2.7}$$

for all $i \in \{1, 2, \ldots, p\}$. Notice that w_i increases monotonically and converges to $w_i^\star$ if $w_i^\star$ is positive or otherwise keeps decreasing and converges to 0 if $w_i^\star = 0$. As such, we can easily distinguish the support and non-support. In fact, gradient flow with dynamics as in (2.7) would exhibit a behavior of "incremental learning" — the entries are learned separately, one at a time [57]. However, with the presence of noise and perturbation arising from correlated designs, the gradient flow may end up over-fitting the noise. Therefore, early stopping as well as the choice of

step size is crucial for obtaining the desired solution [29]. We use (2.7) to obtain a gradient descent update:

$$w_{i,t+1} = w_{i,t}(1 + 2N^2\eta(w_i^\star - w_{i,t})w_{i,t}^{1-\frac{2}{N}}). \tag{2.8}$$

The gradient descent with $N = 2$ is analyzed in [29]. However, when $N > 2$, the presence of $w_{i,t}^{1-\frac{2}{N}}$ imposes an asymmetrical effect on the gradient dynamics. The difficulty of analyzing such gradient descent (2.8) is pointed out in [57]. More specifically, the recurrence relation is not solvable. However, gradient descent updates still share similar dynamics with the idealized gradient flow in (2.7). Inspired by this effect, we are able to show that the entries inside the support and those outside the support are learned separately with a practical optimization algorithm shown in (2.2) and (2.12). As a result, we are able to explore how the depth N affects the choice of step size and early stopping criterion.

2.4.2 Proof Sketch

Growth rate of gradient descent. We adopt the same decomposition as illustrated in [29], and define the following error sequences:

$$\mathbf{b}_t = \frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{e}_t - \frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}, \quad \mathbf{p}_t = \left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{I}\right)(\mathbf{s}_t - \mathbf{w}^\star), \tag{2.9}$$

where $\mathbf{e}_t$ and $\mathbf{s}_t$ stand for error and signal accordingly, and the definitions can be found in (A.1) in Appendix. We can then write the updates on $\mathbf{s}_t$ and $\mathbf{e}_t$ as

$$\begin{aligned} \mathbf{s}_{t+1} &= \mathbf{s}_t \odot (\mathbf{1} - 2N\eta(\mathbf{s}_t - \mathbf{w}^\star + \mathbf{p}_t + \mathbf{b}_t) \odot \mathbf{s}_t^{(N-2)/N})^N, \\ \mathbf{e}_{t+1} &= \mathbf{e}_t \odot (\mathbf{1} - 2N\eta(\mathbf{p}_t + \mathbf{b}_t) \odot \mathbf{e}_t^{(N-2)/N})^N. \end{aligned} \tag{2.10}$$

To illustrate the idea, we think of the one-dimensional updates $\{s_t\}_{t\geq 0}$ and $\{e_t\}_{t\geq 0}$, ignore the error perturbations $\mathbf{p}_t$ and $\mathbf{b}_t$ in the signal updates $\{s_t\}_{t\geq 0}$, and treat $\|\mathbf{p}_t + \mathbf{b}_t\|_\infty \leq B$ in the error updates $\{e_t\}_{t\geq 0}$.

$$s_{t+1} = s_t(1 - 2N\eta(s_t - w^\star)s_t^{(N-2)/N})^N, \quad e_{t+1} = e_t(1 - 2N\eta B e_t^{(N-2)/N})^N. \tag{2.11}$$

We use the continuous approximation to study the discrete updates. Therefore, we can borrow many insights from the analysis about gradient flow to overcome the difficulties caused by $w_{i,t}^{1-\frac{2}{N}}$ as pointed out in equation (2.8). With a proper choice of step size η, the number of iterations T_l for s_t converging to $w^\star$ is derived as

$$T_l \le \sum_{t=0}^{T_l-1} \frac{s_{t+1}-s_t}{2N^2\eta(w^\star - s_t)s_t^{(2N-2)/N}} \qquad \le \frac{1}{N^2\eta w^\star}\int_{\alpha^N}^{w^\star} \frac{1}{s^{(2N-2)/N}} ds + \mathcal{O}\left(\frac{w^\star - \alpha^N}{\alpha^{2N-2}}\right).$$

The number of iterations T_u for e_t staying below some threshold $\alpha^{N/4}$ is derived as

$$T_u \ge \sum_{t=0}^{T_u-1} \frac{e_{t+1}-e_t}{4N^2\eta B e_t^{(2N-2)/N}} \ge \frac{1}{4N^2\eta B}\int_{\alpha^N}^{\alpha^{N/4}} \frac{1}{e^{(2N-2)/N}} de.$$

With our choice of coherence μ in Theorem 1, we are able to control B to be small so that T_l is smaller than T_u. This means the entries on the support converge to the true signal while the entries outside the support stay around 0, and we are able to distinguish signals and errors.

Dealing with negative targets. We now illustrate the idea about how to generalize the result about non-negative signals to general signals. The exact gradient descent updates on $\mathbf{u}$ and $\mathbf{v}$ are given by:

$$\begin{aligned}
\mathbf{u}_{t+1} &= \mathbf{u}_t \odot \left(\mathbf{1} - 2N\eta\left(\frac{1}{n}\mathbf{X}^\mathsf{T}(\mathbf{X}(\mathbf{w}_t - \mathbf{w}^\star) - \boldsymbol{\xi}) \odot \mathbf{u}_t^{N-2}\right)\right), \\
\mathbf{v}_{t+1} &= \mathbf{v}_t \odot \left(\mathbf{1} + 2N\eta\left(\frac{1}{n}\mathbf{X}^\mathsf{T}(\mathbf{X}(\mathbf{w}_t - \mathbf{w}^\star) - \boldsymbol{\xi}) \odot \mathbf{v}_t^{N-2}\right)\right).
\end{aligned} \tag{2.12}$$

The basic idea is to show that when $w_i^\star$ is positive, $v_i^\star$ remains small up to the early stopping criterion, and when $w_i^\star$ is negative, $u_i^\star$ remains small up to the early stopping criterion. We turn to studying the gradient flow of such dynamics. Write $\mathbf{r}(t) = \frac{1}{n}\mathbf{X}^\mathsf{T}(\mathbf{X}(\mathbf{w}(t) - \mathbf{w}^\star) - \boldsymbol{\xi}$. It is easy to verify that the gradient flow has a solution:

$$\begin{aligned}
\mathbf{u}(t) &= \left(\alpha^{2-N}\mathbf{1} + 2N(N-2)\eta\int_0^t \mathbf{r}(\upsilon)d\upsilon\right)^{\frac{1}{2-N}}, \\
\mathbf{v}(t) &= \left(\alpha^{2-N}\mathbf{1} - 2N(N-2)\eta\int_0^t \mathbf{r}(\upsilon)d\upsilon\right)^{\frac{1}{2-N}}.
\end{aligned}$$

We may observe some symmetry here, when $u_{i,t}$ is large, $v_{i,t}$ must be small. For the case $w_i > 0$, to ensure the increasing of $u_{i,t}$ and decreasing of $v_{i,t}$ as we desire, the initialization needs to be smaller than w_i, which leads to the extra constraint on initialization $\Psi(w^\star_{\min}, \epsilon)$ with order of $\mathcal{O}(w^\star_{\min})$ as defined before. It remains to build the connection between gradient flow and gradient descent, where again we uses the continuous approximation as before. The detailed derivation is presented in Appendix A.2.3.

2.5 Simulation Study

We conduct a series of simulation experiments to further illuminate our theoretical findings. Our simulation setup is described as follows. The entries of $\mathbf{X}$ are sampled as i.i.d. Rademacher random variables and the entries of the noise vector $\boldsymbol{\xi}$ are i.i.d. $N(0, \sigma^2)$ random variables. We let $\mathbf{w}^\star = \gamma \mathbf{1}_S$. The values for the simulation parameters are: $n = 500$, $p = 3000$, $k = 5$, $\gamma = 1$, $\sigma = 0.5$ unless otherwise specified. For ℓ_2-plots each simulation is repeated 30 times, and the median ℓ_2 error is depicted. The shaded area indicates the region between 25^{th} and 75^{th} percentiles pointwisely.

Convergence results. We start by showing that the general choice of N leads to the sparse recovery, similar to $N = 2$ in [29], as shown in our main theorem. We choose different values of N to illustrate the convergence of the algorithm. The result on simulated data is shown in Figure 2.2, and we defer the result on MNIST to Appendix A.5. Note that the ranges in the x-axes of these figures differ due to different choice of N and η. We observe that as N increases, the number of iterations increases significantly. This is due to the term $\mathbf{u}^{N-2}$ and $\mathbf{v}^{N-2}$ in (2.12), and the step size $\eta \approx \frac{1}{N^2}$. With a very small initialization, it takes a large number of iterations to escape from the small region (close to 0).

Larger initialization. As discussed in Remark 1, the upper bound on initialization gets larger with larger N. We intentionally pick a relatively large $\alpha^N = 2 \times 10^{-3}$ where the algorithm fails to converge for $N = 2$. With the same initialization, the recovery manifests as N increases (Figure 2.3).

Early stopping window size. Apart from the coordinate path shown in Figure 2.1, we obtain

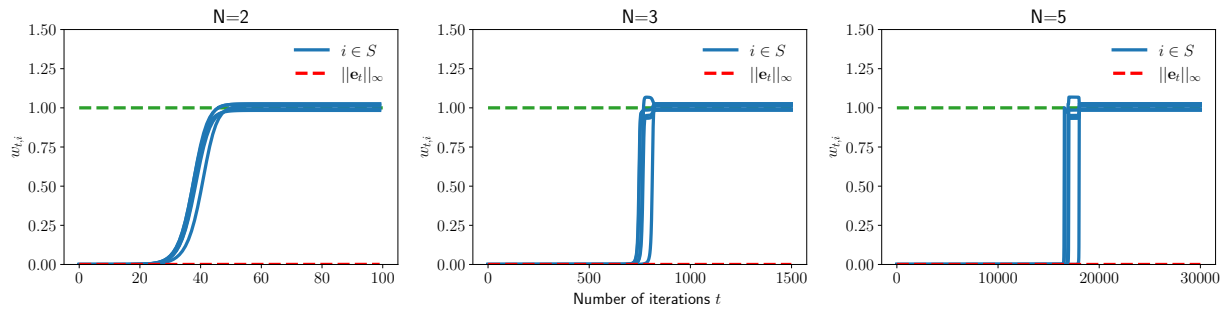


Figure 2.2: Coordinates paths for different choice of $N = 2, 3, 5$ with $\alpha^N = 10^{-6}$ and $\eta = 1/(5N^2)$. Reprinted with permission from [1].

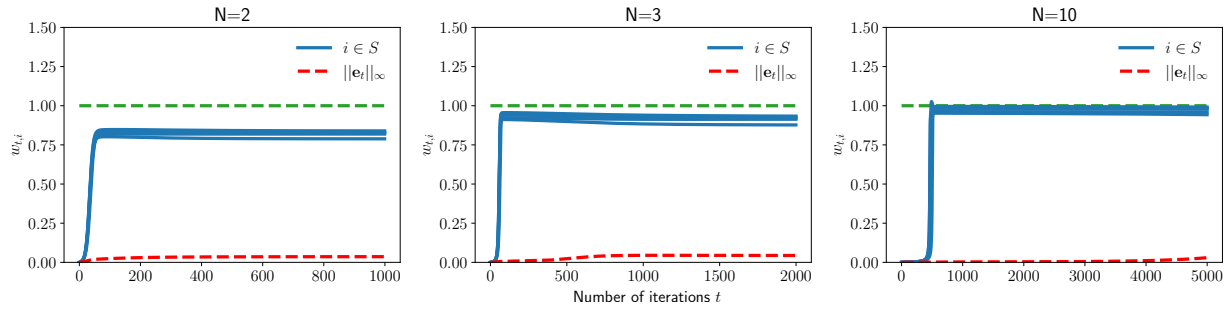


Figure 2.3: The effect of N on the initialization α^N with $\eta = 1/(5N^2)$. Reprinted with permission from [1]. Reprinted with permission from [1].

multiple runs and plot the $\log$-ℓ_2 error (the logarithm of the ℓ_2-error) of the recovered signals to further confirm the increase of early stopping window, as shown in Section 2.3. Note that for both Figures 2.1 and 2.4, we set $n = 100$ and $p = 200$. Since α^N would decrease quickly with N, which would cause the algorithm takes a large number of iterations to escape from the small region. We fix $\alpha^N = 10^{-5}$ instead of fixing α for Figure 2.4.

Incremental learning dynamics. The dynamics of incremental learning for different N is discussed in [57]. The distinct phases of learning are also observed in sparse recovery (Figure 2.5), though we do not provide a theoretical justification. Larger values of N would lead to more distinct learning phases for entries with different magnitudes under the same initialization α^N and step size η.

Kernel regime. As pointed out in [59], the scale of initialization determines whether the

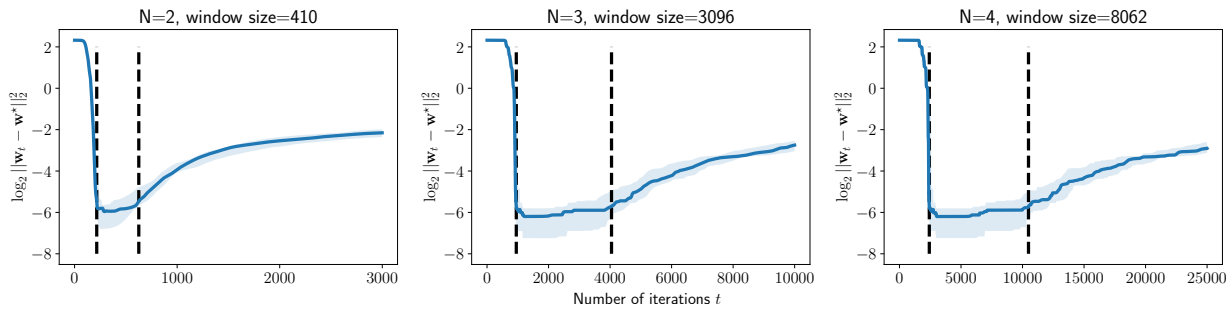


Figure 2.4: $\log$-ℓ_2 error of $N = 2, 3, 4$ with the fixed step size $\eta = 0.01$. Reprinted with permission from [1].

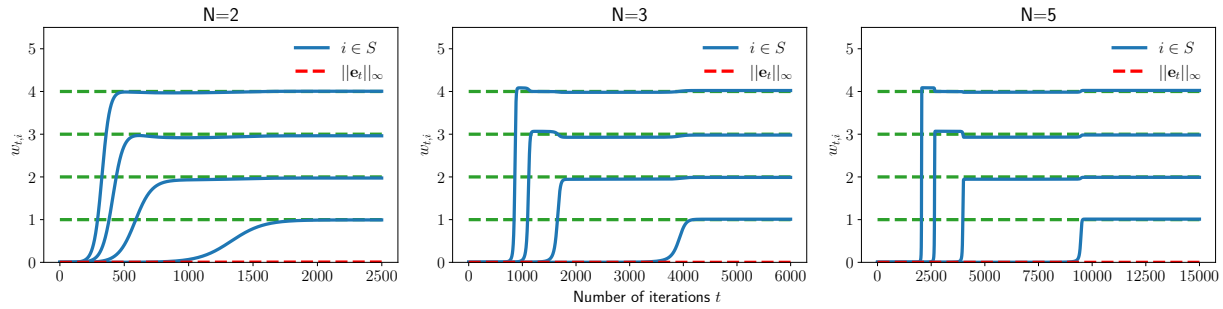


Figure 2.5: Coordinates paths for $N = 2, 3, 5$. The entries of $\mathbf{w}^\star$ on the support S are now $[1, 2, 3, 4]$. The initialization is $\alpha^N = 10^{-4}$ and the step size is $\eta = 10^{-3}$ for all N.

gradient dynamics obey the "kernel" or "rich" regimes for diagonal linear networks. We have carefully analyzed and demonstrated the sparse recovery problem with small initialization, which corresponds to the "rich" regime. To explore the "kernel" regime in a more practical setting, we set $n = 500$, $p = 100$, and the entries of $\mathbf{w}^\star$ are i.i.d. $\mathcal{N}(0, 1)$ random variables. The noise level is $\sigma = 25$, and the initialization and step size is set as $\alpha^N = 1000$ and $\eta = 10^{-7}$ for all N. Note that we are not working in the case $n \ll p$ as [59]. We still observe that the gradient dynamics with large initialization (Figure 2.6) can be connected to ridge regression if early stopping is deployed.

2.6 Conclusions and Future Work

In this work, we extend the implicit regularization results in [29] from $N = 2$ to general $N > 2$, and further study how gradient dynamics and early stopping is affected by different choice N. We

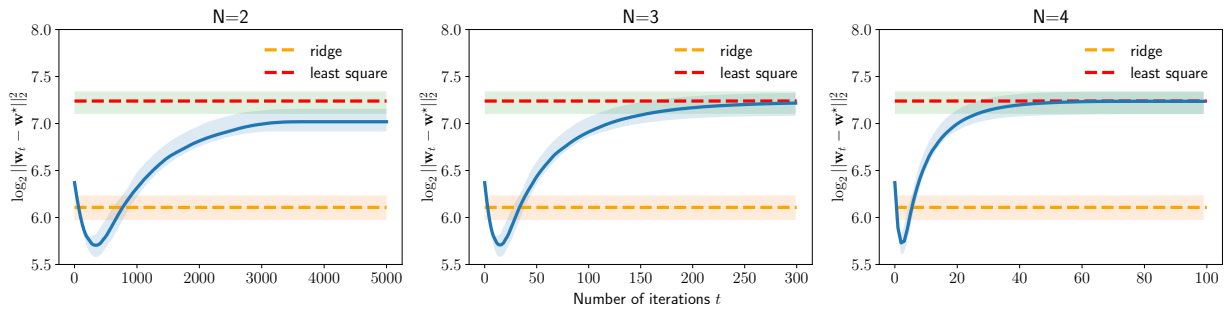


Figure 2.6: log-ℓ_2 error of $N = 2, 3, 4$ for a ridge regression setting. The ridge regression solution is selected by 5-fold cross validation. Reprinted with permission from [1].

show that the error bound is invariant with different choices of N and yields the minimax optimal rate. The step size is of order $\mathcal{O}(1/N^2)$. The initialization and early stopping window gets larger when increasing N due to the changes on gradient dynamics. The incremental learning dynamics and kernel regime of such parametrizations are empirically shown, however not theoretically justified, which is left for future work.

The convergence result can be further improved by relaxing the requirement on the incoherence of design matrix from $\mu \lesssim \frac{1}{kw^\star_{\max}/w^\star_{\min}}$ to $\mu \lesssim \frac{1}{k\log(w^\star_{\max}/w^\star_{\min})}$, similar to [29]. Overall, we believe that such an analysis and associated techniques could be applied for studying other, deeper nonlinear models in more practical settings.

2.7 Implicit Regularization for Dictionary Sparsity

Another type of structured sparsity is called Dictionary sparsity, which means $\mathbf{D}\mathbf{w}^\star \in \mathbb{R}^m$ is sparse for some linear transformation $\mathbf{D} \in \mathbb{R}^{m\times p}$. Such model includes many well-known sparse formulations as special cases including fused lasso [76] and total variation, where for fused lasso, the linear transformation reads

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(p-1)\times p}.$$

When $\mathrm{rank}(\mathbf{D}) = r < p$, we construct an additional matrix $\mathbf{A} \in \mathbb{R}^{(p-r)\times p}$ such that

$$\widetilde{\mathbf{D}} = \begin{bmatrix} \mathbf{D} \\ \mathbf{A} \end{bmatrix} \in \mathbb{R}^{(m+p-r)\times p}, \quad \mathrm{rank}(\widetilde{\mathbf{D}}) = p.$$

Let us denote $\boldsymbol{\theta} = \mathbf{D}\mathbf{w} \in \mathbb{R}^m$, $\boldsymbol{\gamma} = \mathbf{A}\mathbf{w} \in \mathbb{R}^{p-r}$ and $\mathbf{D}^+ \in \mathbb{R}^{p\times(m+p-r)}$ is the Moore-Penrose inverse of $\widetilde{\mathbf{D}}$. Therefore,

$$\mathbf{D}^+ \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\gamma} \end{bmatrix} = \mathbf{D}^+\widetilde{\mathbf{D}}\mathbf{w} = \mathbf{w}.$$

We use $\mathbf{D}_1$ denotes the first m rows of $\mathbf{D}^+$ and $\mathbf{D}_2$ denotes the remaining $p - r$ rows. Therefore, we aim to solve

$$\min_{\boldsymbol{\theta},\boldsymbol{\gamma}} \left\| \mathbf{y} - \mathbf{X}\mathbf{D}^+ \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\gamma} \end{bmatrix} \right\|_2^2 = \|\mathbf{y} - \mathbf{X}(\mathbf{D}_1\boldsymbol{\theta} + \mathbf{D}_2\boldsymbol{\gamma})\|_2^2 . \tag{2.13}$$

One can see the relationship between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{D}_2^\top\mathbf{X}^\top\mathbf{X}\mathbf{D}_2)^-\mathbf{D}_2^\top\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{D}_1\hat{\boldsymbol{\theta}}).$$

Denoting $\mathbf{P} = \mathbf{X}\mathbf{D}_2(\mathbf{D}_2^\top\mathbf{X}^\top\mathbf{X}\mathbf{D}_2)^-\mathbf{D}_2^\top\mathbf{X}^\top$, we rewrite (2.13) as

$$\min_{\boldsymbol{\theta}} \|(\mathbf{I} - \mathbf{P})\mathbf{y} - (\mathbf{I} - \mathbf{P})\mathbf{X}\mathbf{D}_1\boldsymbol{\theta}\|_2^2 .$$

Since we are interested in a sparse solution on $\boldsymbol{\theta}$, we apply the power transformation

$$\boldsymbol{\theta} = \mathbf{u}^{\circ N} - \mathbf{v}^{\circ N}, \quad N \geq 2,$$

and define the loss,

$$\mathcal{L}(\mathbf{u}, \mathbf{v}) = \left\|(\mathbf{I} - \mathbf{P})\mathbf{y} - (\mathbf{I} - \mathbf{P})\mathbf{X}\mathbf{D}_1(\mathbf{u}^{\circ N} - \mathbf{v}^{\circ N})\right\|_2^2 . \tag{2.14}$$

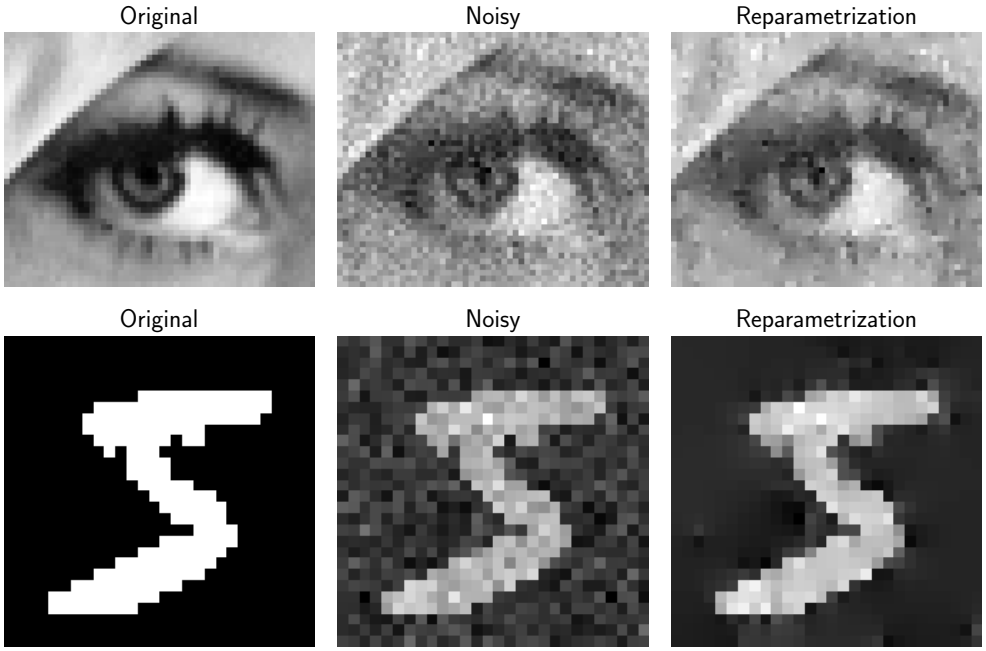


Figure 2.7: Implicit regularization for dictionary sparsity.

The proposed algorithm is to run gradient descent on $\mathcal{L}(\mathbf{u}, \mathbf{v})$ and apply early stopping. When $\mathrm{rank}(\mathbf{D}) = p$, it falls into an easy case where $\mathbf{D}$ is invertible. The additional matrix $\mathbf{A}$ is not needed, and the loss function (2.14) can be directly derived from (2.13) without projection.

We demonstrate the practical usage via image denoising. The image is added with Gaussian noise first. We apply the reparametrization (2.14) at each pixel and run gradient descent without any explicit regularization. The results are shown in Figure 2.7, where the result is similar to the total-variation (L_1) denoising [77]. The proposed algorithm exhibits a learning-to-denoise phenomenon, as shown in Figure 2.8, where the major feature is learned first and then minor features are learned, at the end the tiny details are discarded as noise.

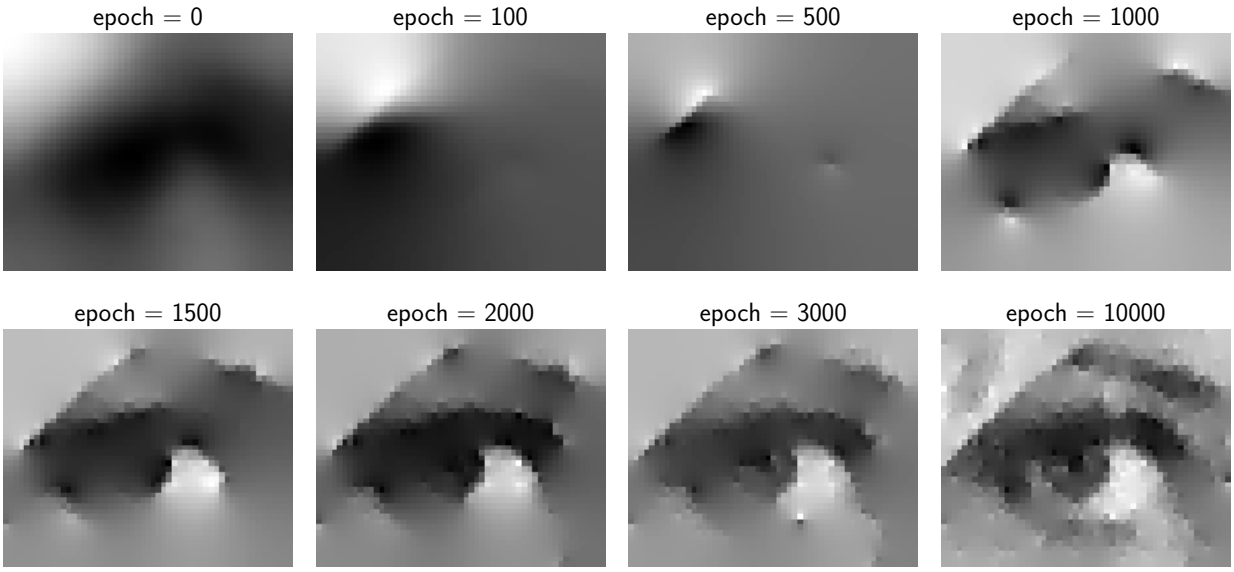


Figure 2.8: Learning to denoise.

3. IMPLICIT REGULARIZATION FOR GROUP SPARSITY*

3.1 Introduction

Motivation. A salient feature of modern deep neural networks is that they are highly overparameterized with many more parameters than available training examples. Surprisingly, however, deep neural networks trained with gradient descent can generalize quite well in practice, even without explicit regularization. One hypothesis is that the dynamics of gradient descent-based training itself induce some form of implicit regularization, biasing toward solutions with low-complexity [23, 50]. Recent research in deep learning theory has validated the hypothesis of such implicit regularization effects. A large body of work, which we survey below, has considered certain (restricted) families of linear neural networks and established two types of implicit regularization — standard sparse regularization and ℓ_2-norm regularization — depending on how gradient descent is initialized.

On the other hand, the role of *network architecture*, or the way the model is parameterized in implicit regularization, is less well-understood. Does there exist a parameterization that promotes implicit regularization of gradient descent towards richer structures beyond standard sparsity?

In this work, we analyze a simple, prototypical hierarchical architecture for which gradient descent induces *group* sparse regularization. Our finding — that finer, *structured* biases can be induced via gradient dynamics — highlights the richness of co-designing neural networks along with optimization methods for producing more sophisticated regularization effects.

Outside of implicit regularization, several other works study the inductive bias of network architectures under *explicit* ℓ_2 regularization on model weights [81, 82]. For multichannel linear convolutional networks, [79] show that ℓ_2-norm minimization of weights leads to a norm regularizer on predictors, where the norm is given by a semidefinite program (SDP). The representation cost in predictor space induced by explicit ℓ_2 regularization on (various different versions of) linear

*Reprinted with permission from [2]. This is a joint work with Thanh V. Nguyen, Chinmay Hegde and Raymond K. W. Wong. Copyright 2023 by the authors.

	NNs	Noise	Implicit vs. Explicit	Regularization
[29]	DLNN	✓	Implicit (GD)	Sparsity
[78]	LNN	✗	Explicit (ℓ_2-penalty)	(Group) Quasi-norm
[79]	LCNN	✗	Explicit (ℓ_2-penalty)	Norm induced by SDP
[80]	DLNN	✗	Implicit	ℓ_2-norm
This work	DGLNN	✓	Implicit (GD)	Structured sparsity

Table 3.1: Comparisons to related work on implicit and explicit regularization. Here, GD stands for gradient descent, (D)LNN/CNN for (diagonal) linear/convolutional neural network, and DGLNN for diagonally grouped linear neural network. Reprinted with permission from [2].

neural networks is studied in [78], which demonstrates several interesting (induced) regularizers on the linear predictors such as ℓ_p quasi-norms and group quasi-norms. However, these results are silent on the behavior of gradient descent-based training *without* explicit regularization. In light of the above results, we ask the following question:

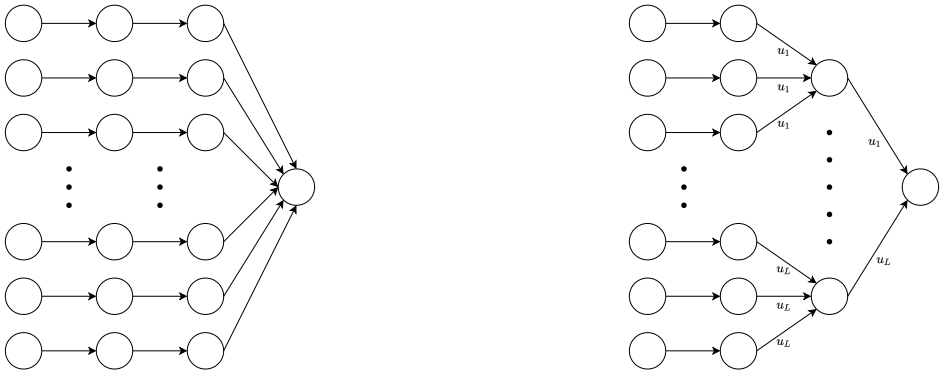
> Beyond ℓ_2-norm, sparsity and low-rankness, can gradient descent induce other forms of implicit regularization?

Our contributions. In this work, we rigorously show that a *diagonally-grouped linear neural network* (see Figure 3.1b) trained by gradient descent with (proper/partial) weight normalization induces *group-sparse* regularization: a form of structured regularization that, to the best of our knowledge, has not been provably established in previous work.

One major approach to understanding implicit regularization of gradient descent is based on its equivalence to a mirror descent (on a different objective function) e.g., [83, 59]. However, we show that, for the diagonally-grouped linear network architecture, the gradient dynamics is beyond mirror descent. We then analyze the convergence of gradient flow with early stopping under orthogonal design with possibly noisy observations, and show that the obtained solution exhibits an implicit regularization effect towards structured (specifically, group) sparsity. In addition, we show that weight normalization can deal with instability related to the choices of learning rates and initialization. With weight normalization, we are able to obtain a similar implicit regularization result

but in more general settings: orthogonal/non-orthogonal designs with possibly noisy observations. Also, the obtained solution can achieve minimax-optimal error rates.

Overall, compared to existing analysis of diagonal linear networks, our model design — that induces structured sparsity — exhibits provably improved sample complexity. In the degenerate case of size-one groups, our bounds coincide with previous results, and our approach can be interpreted as a new algorithm for sparse linear regression.



(a) Diagonal linear NN (DLNN). (b) Diagonally grouped linear NN (DGLNN).

Figure 3.1: An illustration of the two architectures for standard and group sparse regularization. Reprinted with permission from [2].

Our techniques. Our approach is built upon the *power reparameterization* trick, which has been shown to promote model sparsity [84]. Raising the parameters of a linear model element-wisely to the N-th power ($N > 1$) results in that parameters of smaller magnitude receive smaller gradient updates, while parameters of larger magnitude receive larger updates. In essence, this leads to a "rich get richer" phenomenon in gradient-based training. In [57] and [85], the authors analyze the gradient dynamics on a toy example, and call this "incremental learning". Concretely, for a linear predictor $\mathbf{w} \in \mathbb{R}^p$, if we re-parameterize the model as $\mathbf{w} = \mathbf{u}^{\circ N} - \mathbf{v}^{\circ N}$ (where $\mathbf{u}^{\circ N}$ means the N-th element-wise power of $\mathbf{u}$), then gradient descent will bias the training towards sparse solutions. This reparameterization is equivalent to a diagonal linear network, as shown in Figure 3.1a. This is further studied in [59] for interpolating predictors, where they show that a

small enough initialization induces ℓ_1-norm regularization. For noisy settings, [29] and [1] show that gradient descent converges to sparse models with early stopping. In the special case of sparse recovery from under-sampled observations (or compressive sensing), the optimal sample complexity can also be obtained via this reparameterization [86].

Inspired by this approach, we study a novel model reparameterization of the form

$$\mathbf{w} = [\mathbf{w}_1^\top, \ldots, \mathbf{w}_L^\top],$$

where $\mathbf{w}_l = u_l^2 \mathbf{v}_l$ for each group $l \in \{1, \ldots, L\}$. (One way to interpret this model is to think of u_l as the "magnitude" and $\mathbf{v}_l$ as the "direction" of the subvector corresponding to each group; see Section 3.2 for details.) This corresponds to a special type of linear neural network architecture, as shown in Figure 3.1b. A related architecture has also been recently studied in [78], but there the authors have focused on the bias induced by an *explicit* ℓ_2 regularization on the weights and have not investigated the effect of gradient dynamics.

The diagonally linear network parameterization of [59, 1] does not suffer from identifiability issues. In contrast to that, in our setup the "magnitude" parameter u_l of each group interacts with the norm of the "direction", $\|\mathbf{v}_l\|_2$, causing a fundamental problem of identifiability. By leveraging the layer balancing effect [87] in DGLNN, we verify the group regularization effect implicit in gradient flow with early stopping. But gradient flow is idealized; for a more practical algorithm, we use a variant of gradient descent based on *weight normalization*, proposed in [88], and studied in more detail in [80]. Weight normalization has been shown to be particularly helpful in stabilizing the effect of learning rates [89, 90]. With weight normalization, the learning effect is separated into magnitudes and directions. We derive the gradient dynamics on both magnitudes and directions with perturbations. Directions guide magnitude to grow, and as the magnitude grows, the directions get more accurate. Thereby, we are able to establish regularization effect implied by such gradient dynamics.

A remark on grouped architectures. Finally, we remark that grouping layers have been

commonly used in grouped CNN and grouped attention mechanisms [91, 92], which leads to parameter efficiency and better accuracy. Group sparsity is also useful for deep learning models in multi-omics data for survival prediction [93]. We hope our analysis towards diagonally grouped linear NN could lead to more understanding of the inductive biases of grouping-style architectures.

3.2 Setup

Notation. Denotes the set $\{1, 2, \ldots, L\}$ by $[L]$, and the vector ℓ_2 norm by $\|\cdot\|$. We use $\mathbf{1}_p$ and $\mathbf{0}_p$ to denote p-dimensional vectors of all 1s and all 0s correspondingly. Also, $\odot$ represents the entry-wise multiplication whereas $\boldsymbol{\beta}^{\circ N}$ denotes element-wise power N of a vector $\boldsymbol{\beta}$. We use $\mathbf{e}_i$ to denote the i^{th} canonical vector. We write inequalities up to multiplicative constants using the notation $\lesssim$, whereby the constants do not depend on any problem parameter.

Observation model. Suppose that the index set $[p] = \cup_{j=l}^{L} G_l$ is partitioned into L disjoint (i.e., non-overlapping) groups $G_1, G_2, \ldots, G_L$ where $G_i \cap G_j = \emptyset, \forall i \neq j$. The size of G_l is denoted by $p_l = |G_l|$ for $l \in [L]$. Let $\mathbf{w}^\star \in \mathbb{R}^p$ be a p-dimensional vector where the entries of $\mathbf{w}^\star$ are non-zero only on a subset of groups. We posit a linear model of data where observations $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R},\ i \in [n]$ are given such that $y_i = \langle \mathbf{x}_i, \mathbf{w}^\star \rangle + \xi_i$ for $i = 1, \ldots, n$, and $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n]^\top$ is a noise vector. Note that we do not impose any special restriction between n (the number of observations) and p (the dimension). We write the linear model in the following matrix-vector form: $\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \boldsymbol{\xi}$, with the $n \times p$ design matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L]$, where $\mathbf{X}_l \in \mathbb{R}^{n \times p_l}$ represents the features from the l^{th} group G_l, for $l \in [L]$. We make the following assumptions on $\mathbf{X}$:

Assumption 1. *The design matrix* $\mathbf{X}$ *satisfies*

$$\sup_{\|\boldsymbol{\beta}_1\| \leq 1, \|\boldsymbol{\beta}_2\| \leq 1} \left| \left\langle \boldsymbol{\beta}_1, \left(\frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_l - \mathbf{I} \right) \boldsymbol{\beta}_2 \right\rangle \right| \leq \delta_{in}, \quad \textit{where } \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{p_l}, \tag{3.1}$$

$$\sup_{\|\boldsymbol{\beta}_1\| \leq 1, \|\boldsymbol{\beta}_2\| \leq 1} \left| \left\langle \frac{1}{\sqrt{n}} \mathbf{X}_l \boldsymbol{\beta}_1, \frac{1}{\sqrt{n}} \mathbf{X}_{l'} \boldsymbol{\beta}_2 \right\rangle \right| \leq \delta_{out}, \quad \textit{where } \boldsymbol{\beta}_1 \in \mathbb{R}^{p_l}, \boldsymbol{\beta}_2 \in \mathbb{R}^{p_{l'}}, l \neq l', \tag{3.2}$$

for some constants $\delta_{in}, \delta_{out} \in (0, 1)$.

The first part (3.1) is a within-group eigenvalue condition while the second part (3.2) is a between-group block coherence assumption. There are multiple ways to construct a sensing matrix to fulfill these two conditions [94, 95]. One of them is based on the fact that random Gaussian matrices satisfy such conditions with high probability [96].

Reparameterization. Our goal is to learn a parameter $\mathbf{w}$ from the data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with coefficients which obey group structure. Instead of imposing an explicit group-sparsity constraint on $\mathbf{w}$ (e.g., via weight penalization by group), we show that gradient descent on the *unconstrained* regression loss can still learn $\mathbf{w}^\star$, provided we design a special reparameterization. Define a mapping $g(\cdot) : [p] \to [L]$ from each index i to its group $g(i)$. Each parameter is rewritten as $w_i = u_{g(i)}^2 v_i, \forall i \in [p]$. The parameterization $G(\cdot) : \mathbb{R}_+^L \times \mathbb{R}^p \to \mathbb{R}^p$ reads

$$[u_1, \ldots, u_L, v_1, v_2, \ldots, v_p] \to [u_1^2 v_1, u_1^2 v_2, \ldots, u_L^2 v_p].$$

This corresponds to the 2-layer neural network architecture displayed in Figure 3.1b, in which $\mathbf{W}_1 = \mathrm{diag}(v_1, \ldots, v_p)$, and $\mathbf{W}_2$ is "diagonally" tied within each group:

$$\mathbf{W}_2 = \mathrm{diag}(u_1, \ldots, u_1, u_2, \ldots, u_2, \ldots, u_L, \ldots, u_L).$$

Gradient dynamics. We learn $\mathbf{u}$ and $\mathbf{v}$ by minimizing the standard squared loss:

$$\mathcal{L}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \left\| \mathbf{y} - \mathbf{X}[(\mathbf{D}\mathbf{u})^{\circ 2} \odot \mathbf{v}] \right\|^2,$$

where

$$\mathbf{D} = \begin{pmatrix} \mathbf{1}_{p_1} & \mathbf{0}_{p_1} & \ldots & \mathbf{0}_{p_1} \\ \mathbf{0}_{p_2} & \mathbf{1}_{p_2} & \ldots & \mathbf{0}_{p_2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_{p_L} & \mathbf{0}_{p_L} & \ldots & \mathbf{1}_{p_L} \end{pmatrix} \in \mathbb{R}^{p \times L}.$$

By simple algebra, the gradients with respect to $\mathbf{u}$ and $\mathbf{v}$ read as follows:

$$\nabla_{\mathbf{u}} L = 2\mathbf{D}^\top \left(\mathbf{v} \odot \left[\mathbf{X}^\top \mathbf{X}((\mathbf{D}\mathbf{u})^{\circ 2} \odot \mathbf{v} - \mathbf{w}^\star) - \mathbf{X}^\top \boldsymbol{\xi} \right] \odot \mathbf{D}\mathbf{u} \right),$$
$$\nabla_{\mathbf{v}} L = \left[\mathbf{X}^\top \mathbf{X}((\mathbf{D}\mathbf{u})^{\circ 2} \odot \mathbf{v} - \mathbf{w}^\star) - \mathbf{X}^\top \boldsymbol{\xi} \right] \odot (\mathbf{D}\mathbf{u})^{\circ 2}.$$

Denote $\mathbf{r}(t) = \mathbf{y} - \sum_{l'=1}^{L} u_l^2(t) \mathbf{X}_l \mathbf{v}_l(t)$. For each group $l \in [L]$, the gradient flow reads

$$\frac{\partial u_l(t)}{\partial t} = \frac{2}{n} u_l(t) \mathbf{v}_l^\top(t) \mathbf{X}_l^\top \mathbf{r}(t), \quad \frac{\partial \mathbf{v}_l(t)}{\partial t} = \frac{1}{n} u_l^2(t) \mathbf{X}_l^\top \mathbf{r}(t). \tag{3.3}$$

Although we are not able to transform the gradient dynamics back onto $\mathbf{w}(t)$ due to the overparameterization, the extra term $u_l(t)$ on group magnitude leads to "incremental learning" effect.

3.3 Analysis of Gradient Flow

3.3.1 First Attempt: Mirror Flow

Existing results about implicit bias in overparameterized models are mostly based on recasting the training process from the parameter space $\{\mathbf{u}(t), \mathbf{v}(t)\}_{t \geq 0}$ to the predictor space $\{\mathbf{w}(t)\}_{t \geq 0}$ [59, 83]. If properly performed, the (induced) dynamics in the predictor space can now be analyzed by a classical algorithm: mirror descent (or mirror flow). Implicit regularization is demonstrated by showing that the limit point satisfies a KKT (Karush–Kuhn–Tucker) condition with respect to minimizing some regularizer $R(\cdot)$ among all possible solutions.

At first, we were unable to express the gradient dynamics in Eq. (3.3) in terms of $\mathbf{w}(t)$ (i.e., in the predictor space), due to complicated interactions between $\mathbf{u}$ and $\mathbf{v}$. This hints that the training trajectory induced by an overparameterized DGLNN may not be analyzed by mirror flow techniques. In fact, we prove a stronger negative result, and rigorously show that the corresponding dynamics *cannot* be recast as a mirror flow. Therefore, we conclude that our subsequent analysis techniques are necessary and do not follow as a corollary from existing approaches.

We first list two definitions from differential topology below.

Definition 2. *Let M be a smooth submanifold of $\mathbb{R}^D$. Given two C^1 vector fields of X, Y on M, we define the* Lie Bracket *of X and Y as $[X, Y](x) \coloneqq \partial Y(x)X(x) - \partial X(x)Y(x)$.*

Definition 3. *Let M be a smooth submanifold of $\mathbb{R}^D$. A C^2 parameterization $G : M \to \mathbb{R}^d$ is said to be commuting iff for any $i, j \in [d]$, the Lie Bracket $[\nabla G_i, \nabla G_j](x) = 0$ for all $x \in M$.*

The parameterization studied in most existing works on diagonal networks is separable, meaning that each parameter only affects one coordinate in the predictor space. In DGLNN, the parameterization is not separable, due to the shared parameter $\mathbf{u}$ within each group. We formally show that it is indeed not commuting.

Lemma 1. *$G(\cdot)$ is not a commuting parameterization.*

Non-commutativity of the parameterization implies that moving along $-\nabla G_i$ and then $-\nabla G_j$ is different with moving with $-\nabla G_j$ first and then $-\nabla G_i$. This causes extra difficulty in analyzing the gradient dynamics. [97] study the equivalence between gradient flow on reparameterized models and mirror flow, and show that a commuting parameterization is a sufficient condition for when a gradient flow with certain parameterization simulates a mirror flow. A complementary necessary condition is also established on the Lie algebra generated by the gradients of coordinate functions of G with order higher than 2. We show that the parameterization $G(\cdot)$ violates this necessary condition.

Theorem 3. *There exists an initialization $[\mathbf{u}_{init}^\top, \mathbf{v}_{init}^\top] \in \mathbb{R}_+^L \times \mathbb{R}^p$ and a time-dependent loss L_t such that gradient flow under $L_t \odot G$ starting from $[\mathbf{u}_{init}^\top, \mathbf{v}_{init}^\top]$ cannot be written as a mirror flow with respect to any Legendre function R under the loss L_t.*

The detailed proof is deferred to the Appendix. Theorem 3 shows that the gradient dynamics implied in DGLNN cannot be emulated by mirror descent. Therefore, a different technique is needed to analyze the gradient dynamics and any associated implicit regularization effect.

3.3.2 Layer Balancing and Gradient Flow

Let us first introduce relevant quantities. Following our reparameterization, we rewrite the true parameters for each group l as

$$\mathbf{w}_l^\star = (u_l^\star)^2 \mathbf{v}_l^\star, \quad \|\mathbf{v}_l^\star\|_2 = 1, \quad \mathbf{v}_l^\star \in \mathbb{R}^{p_l}.$$

The support is defined on the group level, where $S = \{l \in [L] : u_l^\star > 0\}$ and the support size is defined as $s = |S|$. We denote $u_{max}^\star = \max\{u_l^\star | l \in S\}$, and $u_{min}^\star = \min\{u_l^\star | l \in S\}$.

The gradient dynamics in our reparameterization does not preserve $\|\mathbf{v}_l(t)\|_2 = 1$, which causes difficulty to identify the magnitude of each u_l and $\|\mathbf{v}_l(t)\|_2$. [87] and [98] show that the gradient flow of multi-layer homogeneous functions effectively enforces the differences between squared norms across different layers to remain invariant. Following the same idea, we discover a similar balancing effect in DGLNN between the parameter $\mathbf{u}$ and $\mathbf{v}$.

Lemma 2. *For any $l \in [L]$, we have*

$$\frac{d}{dt}\left(\frac{1}{2}u_l^2 - \|\mathbf{v}_l\|^2\right) = 0.$$

The balancing result eliminates the identifiability issue on the magnitudes. As the coordinates within one group affect each other, the direction which controls the growth rate of both $\mathbf{u}$ and $\mathbf{v}$ need to be determined as well.

Lemma 3. *If the initialization $\mathbf{v}_l(0)$ is proportional to $\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$, then*

$$\left\langle \frac{\mathbf{v}_l(0)}{\|\mathbf{v}_l(0)\|}, \mathbf{v}_l^\star \right\rangle \geq 1 - \left(\delta_{in} + L\delta_{out} + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2 / (u_l^\star)^2\right)^2.$$

Note that this initialization can be obtained by a single step of gradient descent with $\mathbf{0}$ initialization. Lemma 3 suggests the direction is close to the truth at the initialization. We can further normalize it to be $\|\mathbf{v}_l(0)\|_2^2 = \frac{1}{2}u_l^2(0)$ based on the balancing criterion. The magnitude equality, $\|\mathbf{v}_l(t)\|_2^2 = \frac{1}{2}u_l^2(t)$, is preserved by Lemma 2. However, ensuring the closeness of the direction throughout the gradient flow presents significant technical difficulties. That said, we are able to present a meaningful implicit regularization result of the gradient flow under orthogonal (and noisy) settings.

Theorem 4. *Fix $\epsilon > 0$. Consider the case where $\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_l = \mathbf{I}$, $\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_{l'} = \mathbf{O}, l \neq l'$, the initialization $u_l(0) = \theta < \frac{\epsilon}{2(u^\star_{max})^2}$ and $\mathbf{v}_l(0) = \eta_l \frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$ with $\|\mathbf{v}_l(0)\|_2^2 = \frac{1}{2}\theta^2, \forall l \in [L]$, there exists an lower bound and upper bound of the time $T_l < T_u$ in the gradient flow in Eq.* (3.3)*, such that for any $T_l \leq t \leq T_u$ we have*

$$\left\|u_l^2(t)\mathbf{v}_l(t) - \mathbf{w}_l^\star\right\|_\infty \lesssim \begin{cases} \left\|\frac{1}{n}\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty \vee \epsilon, & \text{if } l \in S. \\ \theta^{3/2}, & \text{if } l \notin S. \end{cases}$$

Theorem 4 states the error bounds for the estimation of the *true* weights $\mathbf{w}^\star$. For entries outside the (true) support, the error is controlled by $\theta^{3/2}$. When θ is small, the algorithm keeps all non-supported entries to be close to zero through iterations while maintaining the guarantee for supported entries. Theorem 4 shows that under the assumption of orthogonal design, gradient flow with early stopping is able to obtain the solution with group sparsity.

3.4 Gradient Descent with Weight Normalization

We now seek a more practical algorithm with more general assumptions and requirements on initialization. To speed up the presentation, we will directly discuss the corresponding variant of (the more practical) gradient descent instead of gradient flow. When standard gradient descent is applied on DGLNN, initialization for directions is very crucial; The algorithm may fail even with a very small initialization when the direction is not accurate, as shown in Appendix B.5. The balancing effect (Lemma 2) is sensitive to the step size, and errors may accumulate [87].

Algorithm 1 Gradient descent with weight normalization

Initialize: $\mathbf{u}(0) = \alpha\mathbf{1}$, unit norm initialization $\mathbf{v}_l(0)$ for each $l \in [L]$, $\eta_{l,t} = \frac{1}{u_l^4(t)}$.
for $t = 0$ to T **do**
  $\mathbf{z}(t+1) = \mathbf{v}(t) - \eta_{l,t}\nabla_{\mathbf{v}}\mathcal{L}(\mathbf{u}(t), \mathbf{v}(t))$
  $\mathbf{v}_l(t+1) = \frac{\mathbf{z}_l(t+1)}{\|\mathbf{z}_l(t+1)\|_2}, \forall l \in [L]$
  $\mathbf{u}(t+1) = \mathbf{u}(t) - \gamma\nabla_{\mathbf{u}}\mathcal{L}(\mathbf{u}(t), \mathbf{v}(t+1))$
  if the early stopping criterion is satisfied **then**
    stop
  end if
end for

Weight normalization as a commonly used training technique has been shown to be helpful in stabilizing the training process. The identifiability of the magnitude is naturally resolved by weight normalization on each $\mathbf{v}_l$. Moreover, weight normalization allows for a larger step size on $\mathbf{v}$, which makes the direction estimation at each step behave like that at the origin point. This removes the restrictive assumption of orthogonal design. With these intuitions in mind, we study the gradient descent algorithm with weight normalization on $\mathbf{v}$ summarized in Algorithm 1. One advantage of our algorithm is that it converges with *any* unit norm initialization $\mathbf{v}_l(0)$. The step size on $\mathbf{u}(t)$ is chosen to be small enough in order to enable the incremental learning, whereas the step size on $\mathbf{v}(t)$ is chosen as $\eta_{l,t} = \frac{1}{u_l^4(t)}$ as prescribed by our theoretical investigation. For convenience, we define $\zeta = 80\left(\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \vee \epsilon\right)$, for a precision parameter $\epsilon > 0$. The convergence of Algorithm 1 is formalized as follows:

Theorem 5. *Fix $\epsilon > 0$. Consider Algorithm 1 with*

$$u_l(0) = \alpha < \frac{\epsilon^4 \wedge 1}{(u^\star_{max})^8} \wedge \frac{1}{80L}(u^\star_{min})^2 \wedge \frac{\epsilon}{L}, \quad \forall l \in [L],$$

any unit-norm initialization on $\mathbf{v}_l$ for each $l \in [L]$ and $\gamma \leq \frac{1}{20(u^\star_{max})^2}$. Suppose Assumption 1 is satisfied with $\delta_{in} \leq \frac{(u^\star_{min})^2}{120(u^\star_{max})^2}$ and $\delta_{out} \leq \frac{(u^\star_{min})^2}{120s(u^\star_{max})^2}$. There exist a lower bound on the number of

iterations

$$T_{lb} = \frac{\log \frac{(u^\star_{max})^2}{2\alpha^2}}{2\log(1+\frac{\gamma}{2}(\zeta \vee (u^\star_{min})^2))} + \left\lfloor \log_2 \frac{(u^\star_{max})^2}{\zeta} \right\rfloor \frac{5}{2\gamma(\zeta \vee (u^\star_{min})^2)},$$

and an upper bound

$$T_{ub} \geq \frac{5}{16\gamma(\zeta \vee (u^\star_{min})^2)} \log \frac{1}{\alpha^4},$$

such that $T_{lb} \leq T_{ub}$ *and for any* $T_{lb} \leq t \leq T_{ub}$*,*

$$\left\| u_l^2(t)\mathbf{v}_l(t) - \mathbf{w}_l^\star \right\|_\infty \lesssim \begin{cases} \left\| \frac{1}{n}\mathbf{X}^\top \boldsymbol{\xi} \right\|_\infty \vee \epsilon, & \text{if } l \in S \\ \alpha, & \text{if } l \notin S \end{cases}.$$

Similarly as Theorem 4, Theorem 5 states the error bounds for the estimation of the *true* weights $\mathbf{w}^\star$. When α is small, the algorithm keeps all non-supported entries to be close to zero through iterations while maintaining the guarantee for supported entries. Compared to the works on implicit (unstructured) sparse regularization [29, 86], our assumption on the incoherence parameter δ_{out} scales with $1/s$, where s is the number of non-zero groups, instead of the total number of non-zero entries. Therefore, the relaxed bound on δ_{out} implies an improved sample complexity, which is also observed experimentally in Figure 3.4. We now state a corollary in a common setting with independent random noise, where (asymptotic) recovery of $\mathbf{w}^\star$ is possible.

Definition 4. *A random variable* Y *is* σ*-sub-Gaussian if for all* $t \in \mathbb{R}$ *there exists* $\sigma > 0$ *such that*

$$\mathbb{E}e^{tY} \leq e^{\sigma^2 t^2/2}.$$

Corollary 2. *Suppose the noise vector* $\boldsymbol{\xi}$ *has independent* σ^2*-sub-Gaussian entries and*

$$\epsilon = 2\sqrt{\frac{\sigma^2 \log(2p)}{n}}.$$

Under the assumptions of Theorem 5, Algorithm 1 produces $\mathbf{w}(t) = (\mathbf{D}\mathbf{u}(t))^{\circ 2} \odot \mathbf{v}(t)$ *that satisfies* $\|\mathbf{w}(t) - \mathbf{w}^\star\|_2^2 \lesssim (s\sigma^2 \log p)/n$ *with probability at least* $1 - 1/(8p^3)$ *for any* t *such that* $T_{lb} \le t \le T_{ub}$.

Note that the error bound we obtain is minimax-optimal. Despite these appealing properties of Algorithm 1, our theoretical results require a large step size on each $\mathbf{v}_l(t)$, which may cause instability at later stages of learning. We observe this instability numrerically (see Figure B.1, Appendix B.5). Although the estimation error of $\mathbf{w}^\star$ remains small (which aligns with our theoretical result), individual entries in $\mathbf{v}$ may fluctuate considerably. Indeed, the large step size is mainly introduced to maintain a strong directional information extracted from the gradient of $\mathbf{v}_l(t)$ so as to stabilize the updates of $\mathbf{u}(t)$ at the early iterations. Therefore, we also propose Algorithm 2, a variant of Algorithm 1, where we decrease the step size after a certain number of iterations.

Algorithm 2. *Run Algorithm 1 with the same setup till each* $u_l(t), l \in [L]$ *gets roughly accurate, set* $\eta_{l,t} = \eta$. *Continue Algorithm 1 until early stopping criterion is satisfied.*

Theorem 6. *Under the assumptions of Theorem 5 with replacing the condition on* δ*'s by* $\delta_{in} \le \frac{\sqrt{\zeta}(u^\star_{min})^2}{120(u^\star_{max})^3}$ *and* $\delta_{out} \le \frac{\sqrt{\zeta}(u^\star_{min})^2}{120s(u^\star_{max})^3}$, *we apply Algorithm 2 with* $\eta_{l,t} = \frac{1}{u^4(t)}$ *at the beginning, and* $\eta_{l,t} = \eta \le \frac{4}{9(u^\star_{max})^2}$ *after* $\forall l \in [L], u_l^2(t) \ge \frac{1}{2}(u_l^\star)^2$, *then with the same* T_{lb} *and* T_{ub}, *we have that for any* $T_{lb} \le t \le T_{ub}$,

$$\left\|u_l^2(t)\mathbf{v}_l(t) - \mathbf{w}_l^\star\right\|_\infty \lesssim \begin{cases} \left\|\frac{1}{n}\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty \vee \epsilon, & \text{if } l \in S. \\ \alpha, & \text{if } l \notin S. \end{cases}$$

In Theorem 6, the criterion to decrease the step size is: $u_l^2(t) \ge \frac{1}{2}(u_l^\star)^2, \forall l \in [L]$. Once this criterion is satisfied, our proof indeed ensures that it would hold for at least up to the early stopping time T_{ub} specified in the theorem. In practice, since $u_l^\star$'s are unknown, we can switch to a more practical criterion: $\max_{l\in[L]}\{|u_l(t+1) - u_l(t)|/|u_l(t) + \varepsilon|\} < \tau$ for some pre-specified tolerance $\tau > 0$ and small value $\varepsilon > 0$ as the criterion for changing the step size. The motivation of this criterion is further discussed in Appendix B.4. The error bound remains the same as Theorem 5. The change in

step size requires a new way to study the gradient dynamics of directions with perturbations. With our proof technique, Theorem 6 requires a smaller bound on δ's (see Lemma 31 versus Lemma 24 in Appendix B.3 for details). We believe it is a proof artifact and leave the improvement for future work.

Connection to standard sparsity. Consider the degenerate case where each group size is 1. Our reparameterization, together with the normalization step, can roughly be interpreted as $w_i \approx u_i^2 \operatorname{sgn}(v_i)$, which is different from the power-reparameterization $w_i = u_i^N - v_i^N, N \geq 2$ in [29] and [1]. This also shows why a large step size on v_i is needed at the beginning. If the initialization on v_i is incorrect, the sign of v_i may not move with a small step size.

3.5 Simulation Study

We conduct various experiments on simulated data to support our theory. Following the model in Section 3.2, we sample the entries of $\mathbf{X}$ i.i.d. using Rademacher random variables and the entries of the noise vector $\boldsymbol{\xi}$ i.i.d. under $N(0, \sigma^2)$. We set $\sigma = 0.5$ throughout the experiments.

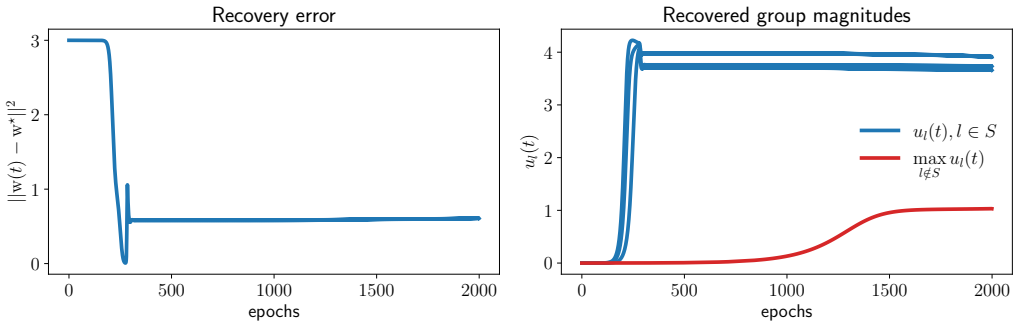


Figure 3.2: Convergence of Algorithm 1. The entries on the support are all 10. Reprinted with permission from [2].

The effectiveness of our algorithms. We start by demonstrating the convergence of the two proposed algorithms. In this experiment, we set $n = 150$ and $p = 300$. The number of non-zero entries is 9, divided into 3 groups of size 3. We run both Algorithms 1 and 2 with the same

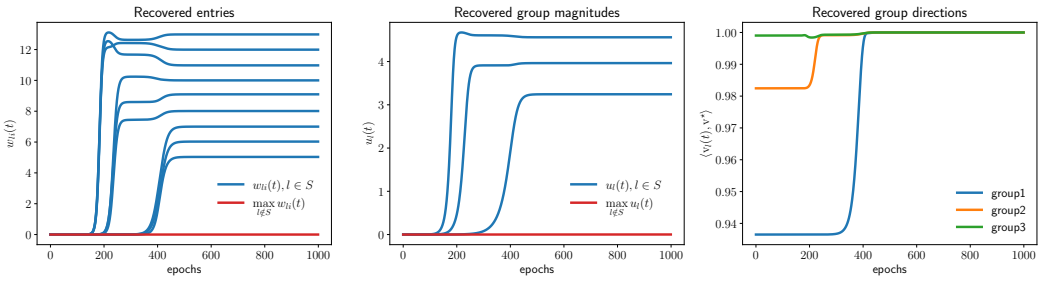


Figure 3.3: Convergence of Algorithm 2. The entries on the support are from 5 to 13. Reprinted with permission from [2].

initialization $\alpha = 10^{-6}$. The step size γ on $\mathbf{u}$ and decreased step size η on $\mathbf{v}$ are both 10^{-3}. In Figure 3.2, we present the recovery error of $\mathbf{w}^\star$ on the left, and recovered group magnitudes on the right. As we can see, early stopping is crucial for reaching the structured sparse solution. In Figure 3.3, we present the recovered entries, recovered group magnitudes and recovered directions for each group from left to right. In addition to convergence, we also observe an incremental learning effect.

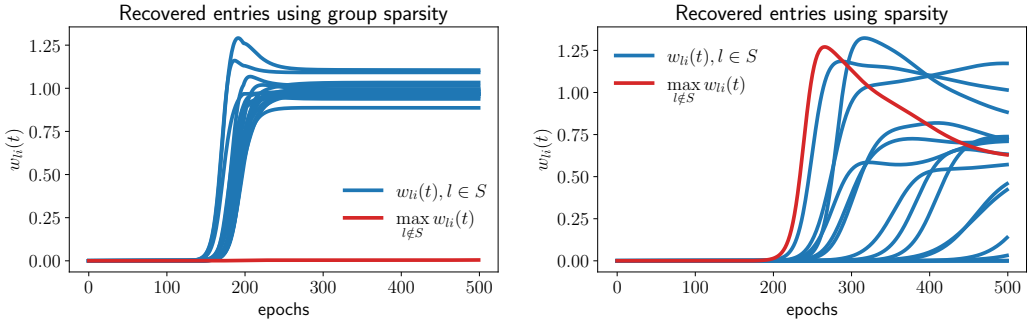


Figure 3.4: Comparison with reparameterization using standard sparsity. $n = 100, p = 500$. Reprinted with permission from [2].

Structured sparsity versus standard sparsity. From our theory, we see that the block incoherence parameter scales with the number of non-zero groups, as opposed to the number of non-zero entries. As such, we can expect an improved sample complexity over the estimators based

on unstructured sparse regularization. We choose a larger support size of 16. The entries on the support are all 1 for simplicity. We apply our Algorithm 2 with group size 4. The result is shown in Figure 3.4 (left). We compare with the method in [29] with parameterization $\mathbf{w} = \mathbf{u}^{\circ 2} - \mathbf{v}^{\circ 2}$, designed for unstructured sparsity. We display the result in the right figure, where interestingly, that algorithm fails to converge because of an insufficient number of samples.

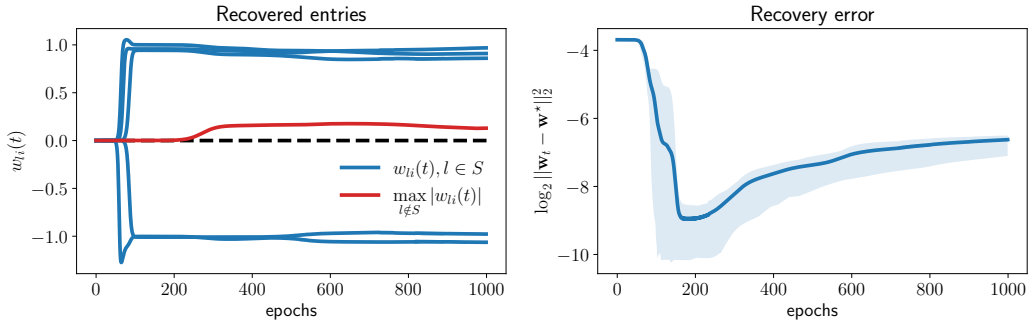


Figure 3.5: Degenerate case when each group size is 1. The $\log \ell_2$-error plot is repeated 30 times, and the mean is depicted. The shaded area indicates the region between the 25^{th} and 75^{th} percentiles. Reprinted with permission from [2].

Degenerate case. In the degenerate case where each group is of size 1, our reparameterization takes a simpler form $w_i \approx u_i^2 \mathrm{sgn}(v)$, i.e., due to weight normalization, our method normalizes v to 1 or -1 after each step. We demonstrate the efficacy of our algorithms even in the degenerate case. We set $n = 80$ and $p = 200$. The entries on the support are $[1, -1, 1, -1, 1]$ with both positive and negative entries. We present the coordinate plot and the recovery error in Figure 3.5.

3.6 Discussion

In this work, we show that implicit regularization for group-structured sparsity can be obtained by gradient descent (with weight normalization) for a certain, specially designed network architecture. Overall, we hope that such analysis further enhances our understanding of neural network training. Future work includes relaxing the assumptions on δ's in Theorem 2, and rigorous analysis of modern grouping architectures as well as power parametrizations.

4. MATRIX COMPLETION WITH INFORMATIVE MISSING

4.1 Introduction

Matrix completion is the study of recovering an underlying matrix from its noisy and partial observations, where we can also view it as a modern high-dimensional missing data problem. Despite various significant breakthroughs made in the last two decades [99, 44], most work on matrix completion still focus on missing-at-random mechanism. While such an assumption is doubtful in many real-life applications, there are very few available options, especially those with theoretical guarantees. This work aims to provide a principled alternative method for missing-not-at-random settings. In particular, we focus on a pairwise pseudo-likelihood approach [46].

A usual assumption that allows succeeding matrix completion is to suppose that the unknown matrix has low rank or has an approximately low rank. The noiseless setting was first studied in [43] using nuclear norm minimization. The vast majority of existing theories on matrix completion assume that entries are revealed with the sample probability independently [100, 101]. Recent approaches to handling entries being revealed with non-uniform probabilities have shown the strength to improve matrix completion accuracy substantially [45, 102, 103]. However, such approaches require the knowledge of missing mechanisms to estimate the propensity score/weights to de-bias the existing matrix completion methods [104]. There is another line of work that assumed that the matrix follows some form of a latent variable model [105, 106]. In this work, we focus on the recovery of a single matrix without other covariate information available [107, 108]. Although the non-uniform missing mechanism is quite flexible, it is fundamentally different from the missing-not-at-random mechanism. The key difference would be whether the missing probability is dependent on the noisy observations, which we will highlight below. In the missing-not-at-ramdom case, the methods developed for the non-uniform missing mechanism are not applicable in general. However, the method we propose in this work for informative missing is applicable for non-uniform missing as well.

The generalized low-rank model has received extensive attention within the matrix completion literature for its efficacy in modeling non-Gaussian data, particularly discrete data. Notably, researchers have investigated its application in specific scenarios such as one-bit matrix completion[109] and multinomial matrix completion [110, 111]. The application of the generalized low-rank model also extends to accommodating unbounded non-Gaussian observations, including Poisson matrix completion [112] and exponential family matrix completion [113, 114].

Within the missing data literature, likelihood-based methods commonly involve specifying a parametric distribution of the missing data mechanism. However, this assumption should be dealt with caution, as it is highly sensitive and may easily induce a misspecified model, resulting in biased estimation and inaccurate results. To circumvent such issues, it is preferable to adopt an assumption as flexible and generally applicable as possible. This type of assumption, often referred to as an unspecified missing data mechanism [48], avoids explicitly specifying a parametric model. Instead, it allows for the derivation of a nonregular likelihood [46], which serves as the foundation for subsequent estimation. Such nonstandard likelihood approaches have been used in regression analysis [115] and variable selection when confronted with informative missing [48]. One disadvantage of this approach is not all the unknown parameters are estimable due to the non-identification issue [116, 46].

In this work, we adapt the pairwise pseudo-likelihood approach [46] to matrix completion with a mild separable/decomposable assumption on the missing mechanism to deal with informative missing under a generalized trace-regression framework. The assumption is very flexible and generally applicable. While not all the parameters are estimable, we can identify the dispersion-scaled matrix values up to a constant shift without suffering from the informative missing, which shows promises to be applied in practice, for example, recommendation systems.

4.2 Preliminaries

Let $\mathbf{A}_\star = (A_{\star,ij})_{i,j=1}^{m_1,m_2} \in \mathbb{R}^{m_1 \times m_2}$ be the matrix of interest, and the observations $(Y, \boldsymbol{X})$ satisfying a generalized trace regression model

$$f(Y|\boldsymbol{X}; \mathbf{A}_\star) = \mathrm{Exp}_{h,G,\phi}(\langle \boldsymbol{X}, \mathbf{A}_\star \rangle) \coloneqq h(Y)\exp\left(\frac{\langle \boldsymbol{X}, \mathbf{A}_\star \rangle Y - G(\langle \boldsymbol{X}, \mathbf{A}_\star \rangle)}{\phi}\right),$$

where h and G are the base measure and log partition functions associated to the canonical representation, ϕ is a constant corresponding to the dispersion parameter. Assume the design matrices $\boldsymbol{X}_k$ are i.i.d. copies of a random matrix $\boldsymbol{X}$ having distribution Π on the set

$$\mathcal{X} = \{\mathbf{e}_i(m_1)\mathbf{e}_j^\top(m_2), 1 \le i \le m_1, 1 \le j \le m_2\},$$

where $\mathbf{e}_i(m)$ are the canonical basis vectors in $\mathbb{R}^m$. For simplicity, we consider a specific sampling model

$$\mathbb{P}(\boldsymbol{X} = \mathbf{e}_i(m_1)\mathbf{e}_j^\top(m_2)) = \frac{1}{m_1 m_2},$$

which shares the similarity with Uniform Sampling at Random (USR) matrix completion. However, we further consider a second-stage missing mechanism. For each data pair $(Y, \boldsymbol{X})$, let T be the indicator variable with value 1 if $(Y, \boldsymbol{X})$ is observed and 0 otherwise. Note that in this case the probability of observation location (i, j) is $\Pr(\boldsymbol{X} = \mathbf{e}_{ij}|T = 1)$, which heavily depends on the missing mechanism and goes beyond USR matrix completion. Note that our analysis below is not restricted to uniform sampling model.

Assumption 2. *The observation probability is separable, i.e.,*

$$\Pr(T = 1|Y, \boldsymbol{X}) = s(Y)t(\boldsymbol{X})$$

for some functions $s(\cdot)$ and $t(\cdot)$.

Note that we do not require any specific form of $s(\cdot)$ and $t(\cdot)$. This assumption is very flexible

and generally applicable. It includes the non-uniform missing mechanism as a special case, where we set $s(Y) = 1$ and leave $t(\boldsymbol{X})$ to account for the non-uniform missing.

Applying the Bayes formula, we now rewrite the conditional distribution as

$$\begin{aligned} p(Y|\boldsymbol{X}, T=1) &= \frac{\Pr(T=1|Y,\boldsymbol{X})f(Y|\boldsymbol{X})}{\Pr(T=1|\boldsymbol{X})} = s(Y)\frac{t(\boldsymbol{X})}{\Pr(T=1|\boldsymbol{X})}f(Y|\boldsymbol{X}) \\ &= s(Y)b(\boldsymbol{X})f(Y|\boldsymbol{X}), \end{aligned}$$

where $b(\boldsymbol{X}) = \frac{t(\boldsymbol{X})}{\mathbb{P}(T=1|\boldsymbol{X})}$.

We may see the conditional likelihood of the observations is complicated, which makes the estimation of the matrix $\mathbf{A}$ intractable. To address this issue, we adopt the procedure called statistical chromatography [46, 47] in matrix completion with informative missing to extract information on $\mathbf{A}$.

Suppose we get n data points $(Y_k, \boldsymbol{X}_k)$ from the distribution of $(Y, \boldsymbol{X})$ given $T = 1$, we decompose $\mathbf{Y} = (Y_1, \ldots, Y_n)$ into $\mathbf{R} = (R_1, \ldots, R_n)$ and $\mathbf{Y}_{(\cdot)} = (Y_{(1)}, \ldots, Y_{(n)})$, which denote the rank and order statistics of $\mathbf{Y}$, respectively. Denote $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, we have that

$$\begin{aligned} \Pr(\mathbf{R}|\mathbf{X}, \mathbf{Y}_{(\cdot)}, \mathbf{T}=\mathbf{1}; \mathbf{A}) &= \frac{\prod_{k=1}^n s(Y_k)t(\boldsymbol{X}_k)f(Y_k|\boldsymbol{X}_k;\mathbf{A})}{\sum_{\pi\in\Xi}\prod_{k=1}^n s(Y_{\pi(k)})t(\boldsymbol{X}_k)f(Y_{\pi(k)}|\boldsymbol{X}_k;\mathbf{A})} \\ &= \frac{\prod_{k=1}^n \exp(\langle \boldsymbol{X}_k, \mathbf{A}\rangle Y_k/\phi)}{\sum_{\pi\in\Xi}\prod_{k=1}^n \exp(\langle \boldsymbol{X}_k, \mathbf{A}\rangle Y_{\pi(k)}/\phi)} \\ &= \Pr(\mathbf{R}|\mathbf{X}, \mathbf{Y}_{(\cdot)}; \mathbf{A}), \end{aligned}$$

where Ξ is the set of all one-to-one maps from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$. We notice that this conditional likelihood does not involve the missing probabilities $s(\cdot)$ and $t(\cdot)$ due to the separable property. It is also not dependent on the base measure $h(\cdot)$ and the log partition function $G(\cdot)$. More importantly, this pseudolikelihood on observed data is identical to that on the complete data, which indicates that this procedure does not suffer from informative missing.

However, the full pseudolikelihood conditioned on the rank statistics is computationally intensive due to the combinatorial nature of permutations, we consider a surrogate of $\Pr(\mathbf{R}|\mathbf{X}, \mathbf{Y}_{(\cdot)}; \mathbf{A})$

using second-order information. For any k and k', let $\mathbf{R}^L_{kk'}$ denote the local rank statistic of Y_k and $Y_{k'}$ among the paper $(Y_k, Y_{k'})$. Instead of considering the full conditional probability, we study the product of all possible combinations of the local rank conditional probability

$$\begin{aligned}
&\prod_{k<k'} \Pr(\mathbf{R}^L_{kk'} = \mathbf{r}^L_{kk'} | \boldsymbol{X}_k, \boldsymbol{X}_{k'}, \mathbf{Y}^L_{(k,k')}, T_k = T_{k'} = 1; \mathbf{A}) \\
&= \prod_{k<k'} \frac{\exp\left(\frac{\langle \boldsymbol{X}_k, \mathbf{A}\rangle Y_k + \langle \boldsymbol{X}_{k'}, \mathbf{A}\rangle Y_{k'}}{\phi}\right)}{\exp\left(\frac{\langle \boldsymbol{X}_k, \mathbf{A}\rangle Y_k + \langle \boldsymbol{X}_{k'}, \mathbf{A}\rangle Y_{k'}}{\phi}\right) + \exp\left(\frac{\langle \boldsymbol{X}_k, \mathbf{A}\rangle Y_{k'} + \langle \boldsymbol{X}_{k'}, \mathbf{A}\rangle Y_k}{\phi}\right)} \\
&= \prod_{k<k'} \frac{1}{1 + \exp(-(Y_k - Y_{k'})\langle \mathbf{A}, \boldsymbol{X}_k - \boldsymbol{X}_{k'}\rangle/\phi)} \\
&= \prod_{k<k'} \Pr(\mathbf{R}^L_{kk'} = \mathbf{r}^L_{kk'} | \boldsymbol{X}_k, \boldsymbol{X}_{k'}, \mathbf{Y}^L_{(k,k')}; \mathbf{A}).
\end{aligned}$$

Taking the negative part after logarithmic transformation, we obtain the function

$$\ell(\mathbf{A}/\phi) = \binom{n}{2}^{-1} \sum_{1 \le k < k' \le n} \log(1 + R_{kk'}(\mathbf{A}/\phi)),$$

where $R_{kk'}(\mathbf{A}/\phi) = \exp\{-(Y_k - Y_{k'})\langle \mathbf{A}/\phi, \boldsymbol{X}_k - \boldsymbol{X}_{k'}\rangle\}$. We notice that the effect of $\mathbf{A}$ and ϕ can not be separated from $\frac{\mathbf{A}}{\phi}$. Therefore, we aim to estimate a dispersion-scaled matrix $\frac{\mathbf{A}_\star}{\phi}$. After removing the unwanted quantities from this pairwise pseudolikelihood, some information about $\mathbf{A}$ is also missed. We also notice that $\ell(\mathbf{A}) = \ell(\mathbf{A} + c)$ for any constant c. Denote $\mathbf{J}$ as all one's matrix. The penalized pairwise pseudolikelihood estimator we consider is

$$\widehat{\mathbf{A}} \in \underset{\langle \mathbf{J}, \mathbf{A}\rangle = 0, \|\mathbf{A}\|_\infty \le a}{\operatorname{argmin}} \ell(\mathbf{A}) + \lambda \|\mathbf{A}\|_\star .$$

4.3 Main Results

We denote some convenient constants for dimensions, i.e.

$$m = \min\{m_1, m_2\}, M = \min\{m_1, m_2\}, d = m_1 + m_2.$$

Assumption 3. *We assume the following holds*

(C1) The rank of the centered, dispersion-scaled true matrix is bounded by r, i.e.,

$$\operatorname{rank}\left(\frac{\mathbf{A}_\star}{\phi} - \left\langle \frac{\mathbf{A}_\star}{\phi}, \mathbf{J} \right\rangle \mathbf{J}\right) \leq r.$$

(C2) There exists $\tau_l > 0$ and $\tau_u > 0$ such that

$$\frac{1}{\tau_l m_1 m_2} \leq \Pr(\boldsymbol{X} = \mathbf{e}_{ij} | T = 1) \leq \frac{\tau_u}{m_1 m_2}, \quad \forall i \in [m_1], j \in [m_2].$$

(C3) Denote $\pi_r = \sum_{1 \leq j \leq m_2} \pi_{rj}$ and $\pi_c = \sum_{1 \leq i \leq m_1} \pi_{ic}$. We assume there is a positive constant $L \geq 1$ s.t.

$$\max_{r \in [m_1], c \in [m_2]} (\pi_r, \pi_c) \leq L/m.$$

(C4) The observation Y is bounded by some constant B almost surely.

(C5) Denote

$$Z_{kk'} = (Y_k - Y_{k'}) \frac{\exp((Y_k - Y_{k'})\langle \boldsymbol{X}_k - \boldsymbol{X}_{k'}, \mathbf{A} \rangle / 2)}{1 + \exp((Y_k - Y_{k'})\langle \boldsymbol{X}_k - \boldsymbol{X}_{k'}, \mathbf{A} \rangle)},$$

where $\|\mathbf{A}\|_\infty \leq M$. There exists some constant $\kappa_M > 0$ s.t. $\mathbb{E}(Z_{kk'}^2 | \boldsymbol{X}_k, \boldsymbol{X}_{k'}) \geq \kappa_M$.

Condition (C1) is the low-rank assumption for the underlying unknown matrix of interest. Instead of making assumptions on the original matrix, we assume the centered and dispersion-scaled matrix is low-rank. These two assumptions are very similar since it is easy to check that one implies the other. With a slight abuse of notation, we may denote $\mathbf{A}_\star$ as $\frac{\mathbf{A}_\star}{\phi} - \left\langle \frac{\mathbf{A}_\star}{\phi}, \mathbf{J} \right\rangle \mathbf{J}$. Condition (C2) and (C3) are to avoid some specific entries/rows/columns being sampled with very high probability, where the trace-norm penalization fails to work [117, 118]. Condition (C4) is a technical assumption for analyzing the concentration inequalities of the involved U-statistics in pairwise pseudolikelihood. Note that this does not violate the parametric assumption on the distribution of Y. For example, truncated normal distribution satisfies both. We leave the extension

to a light tail type of assumption for future work. Condition (C5) is about the second moment of the difference between two observations, which is bounded away from zero. Intuitively, the larger the difference, the more effective the pairwise pseudolikelihood.

Let $\{\mathbf{u}_k \in \mathbb{R}^{m_1}\}$ and $\{\mathbf{v}_k \in \mathbb{R}^{m_2}\}$ be the left and right singular vectors of $\mathbf{A}_\star$ respectively. Let the column and row span of $\mathbf{A}_\star$ be $\mathbf{U}_\star = \mathrm{col}(\mathbf{A}_\star) = \mathrm{span}\{\mathbf{u}_k\}$ and $\mathbf{V}_\star = \mathrm{row}(\mathbf{A}_\star) = \mathrm{span}\{\mathbf{v}_k\}$. Define

$$\mathcal{M} \coloneqq \{\mathbf{A} : \mathrm{row}(\mathbf{A}) \subseteq \mathbf{V}_\star, \mathrm{col}(\mathbf{A}) \subseteq \mathbf{U}_\star\}$$
$$\overline{\mathcal{M}}^\perp \coloneqq \{\mathbf{A} : \mathrm{row}(\mathbf{A}) \subseteq \mathbf{V}_\star^\perp, \mathrm{col}(\mathbf{A}) \subseteq \mathbf{U}_\star^\perp\}.$$

It is easy to see that $\mathcal{M} \subseteq \overline{\mathcal{M}}$, but $\mathcal{M} \neq \overline{\mathcal{M}}$. The subspace compatibility of $\overline{\mathcal{M}}$ is upper bounded by $\sqrt{2r}$, i.e.

$$\psi(\overline{\mathcal{M}}) = \sup_{\mathbf{A} \in \overline{\mathcal{M}} \setminus \{0\}} \frac{\|\mathbf{A}\|_\star}{\|\mathbf{A}\|_F} \leq \sqrt{2r}.$$

The pairwise pseudo likelihood in can be written as

$$\ell(\mathbf{A}) = \frac{2}{n(n-1)} \sum_{1 \leq k < k' \leq n} \{\psi(Y_{k\backslash k'} \langle \boldsymbol{X}_{k\backslash k'}, \mathbf{A} \rangle) - Y_{k\backslash k'} \langle \boldsymbol{X}_{k\backslash k'}, \mathbf{A} \rangle\},$$

where $\psi(t) = \log(1 + \exp(t))$, and hence $\psi'(t) = \frac{\exp(t)}{1+\exp(t)}$, $\psi''(t) = \frac{\exp(t)}{(1+\exp(t))^2}$ and $\psi'''(t) = \frac{\exp(t)(1-\exp(t))}{(1+\exp(t))^3}$. After some algebra, we have that $|\psi'(t)| \leq 1$, $|\psi''(t)| \leq 0.25$ and $|\psi'''(t)| \leq 0.1$. Therefore, its first and second order derivatives are

$$\nabla \ell(\mathbf{A}) = \frac{2}{n(n-1)} \sum_{1 \leq k < k' \leq n} \{\psi'(Y_{k\backslash k'} \langle \boldsymbol{X}_{k\backslash k'}, \mathbf{A} \rangle) Y_{k\backslash k'} \boldsymbol{X}_{k\backslash k'} - Y_{k\backslash k'} \boldsymbol{X}_{k\backslash k'}\},$$

and

$$\nabla^2 \ell(\mathbf{A}) = \frac{2}{n(n-1)} \sum_{1 \leq k < k' \leq n} \{\psi''(Y_{k\backslash k'} \langle \boldsymbol{X}_{k\backslash k'}, \mathbf{A} \rangle) Y_{k\backslash k'}^2 \mathrm{vec}(\boldsymbol{X}_{k\backslash k'})^{\otimes 2}\},$$

where $\mathrm{vec}(\boldsymbol{X})$ is the standard vectorization of matrix $\boldsymbol{X}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^\top$.

Denote

$$\Sigma_{g,R} = \sum_{kk' \in g} \frac{2}{n} \varepsilon_{kk'} Z_{kk'} (\boldsymbol{X}_k - \boldsymbol{X}_{k'})$$

where g is any group of non-overlapped pairs in $[n]$ and $\varepsilon_{kk'}$ is i.i.d. Radamacher random variables.

Theorem 7. *Suppose $\langle \mathbf{A}_\star, \mathbf{J} \rangle = 0$, $\|\mathbf{A}_\star\|_\infty \leq a$ and all the conditions in Assumption 3 hold, there exists an absolute constant $c > 0$ such that with probability at least $1 - \frac{9}{d}$, the following holds*

$$\frac{1}{m_1 m_2} \left\| \widehat{\mathbf{A}} - \frac{\mathbf{A}_\star}{\phi} \right\|_F^2 \leq c \max \Bigg\{ \frac{1}{\kappa_M^2} r \tau_l^4 m_1 m_2 \max\{\lambda^2, B^2 a^2 (\mathbb{E} \|\Sigma_{g,R}\|)^2\}, \frac{\tau_l^2}{\kappa_M} a^2 \sqrt{\frac{\log(d) + \log(n)}{n}} \Bigg\}.$$

Corollary 3. *Under the same condition as Theorem 7, there exists absolute constants $C^\star, c, c' > 0$ such that choosing $\lambda = \sqrt{C^\star L B^2 \frac{\log(d)+\log(n)}{mn}}$ and when $\frac{3c}{128} L m \log^3(d) \leq n \leq m_1 m_2$, with probability at least $1 - \frac{9}{d}$, the following holds*

$$\frac{1}{m_1 m_2} \left\| \widehat{\mathbf{A}} - \frac{\mathbf{A}_\star}{\phi} \right\|_F^2 \leq c' \max \Bigg\{ \frac{r \tau_l^4 B^2 L}{\kappa_M^2} \frac{M \log(d)}{n} \max\{1, a^2\}, \frac{\tau_l^2 a^2}{\kappa_M} \sqrt{\frac{\log(d)}{n}} \Bigg\}.$$

Our result implies that the penalized pairwise pseudolikelihood approach can estimate the remaining components of $\mathbf{A}_\star$ after excluding the non-identifiable parts, i.e., a constant shift and scaling. Compared with [44], the error rate shown above is nearly the same up to a logarithmic factor. However, our method is able to handle the informative missing, which contains the non-uniform missing as a special case.

4.4 Numerical Experiments

We propose an efficient algorithm to solve the optimization problem related to the penalized pairwise pseudolikelihood. Note that the objective function and the constraint set are both convex.

and the constraint set is a closed convex set. We adapt the projected gradient descent algorithm to solve the optimization problem, shown below.

Algorithm 2 Projected gradient descent

Initialize: $\mathbf{A}(0) = \mathbf{0}$, set learning rate η.

for $t = 0$ to T **do**

$\mathbf{G}(t) = \nabla \ell(\mathbf{A}(t)) - \frac{1}{m_1 m_2} \langle \mathbf{J}, \nabla \ell(\mathbf{A}(t)) \rangle$

$\mathbf{K}(t) = \mathbf{A}(t) - \eta \mathbf{G}(t)$

$\mathbf{Q}(t) = \mathcal{S}_\lambda(\mathbf{K}(t))$

$\mathbf{A}(t+1) = \mathrm{POCS}(\mathbf{Q}(t))$

if the early stopping criterion is satisfied **then**

stop

end if

end for

Here, POCS is the projection onto the intersection of two convex sets $\{\mathbf{A} | \langle \mathbf{J}, \mathbf{A} \rangle = 0\}$ and $\{\mathbf{A} | \, \|\mathbf{A}\|_\infty \leq a\}$. The notation $\mathcal{S}_\lambda(\cdot)$ is the soft-thresholding operator,

$$\mathcal{S}_\lambda(\mathbf{A}) = \mathbf{U} \mathbf{D}_\lambda \mathbf{V}^\top, \quad \text{with} \quad \mathbf{D}_\lambda = \mathrm{diag}[(d_1 - \lambda)_+, \ldots, (d_r - \lambda)_+],$$

where $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ is the SVD of $\mathbf{A}$, and $t_x = \max(t, 0)$.

We use the following simulation study to demonstrate the efficacy of the proposed method. We generate a 50×50 matrix $\mathbf{A}_\star$ with rank $r = 5$, and the observations $\mathbf{Y}$ are generated from a Gaussian distribution with noise level $\sigma = 1$. The probability of each entry being observed is related to the value of entry itself

$$\mathbb{P}(T = 1 | Y, \boldsymbol{X} = \mathbf{e}_{ij}) = \frac{1}{1 + \exp(3Y)}.$$

Therefore, the observed entries are more likely to be small in magnitude, as shown in Figure 4.1.

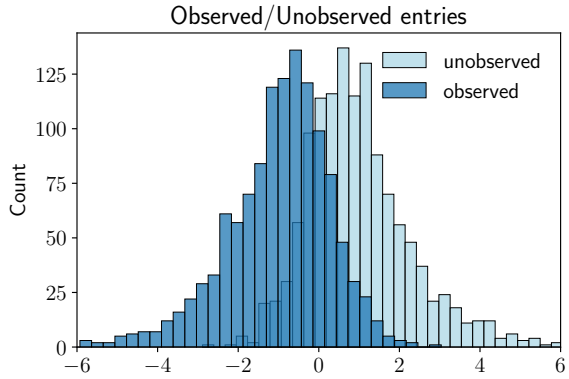


Figure 4.1: Observation bias.

We use the observed entries as training data, and equally split the unobserved data as validation and test data. We compare the proposed method with [119, 99, 45, 106]. The distribution of recovered entries is shown in Figure 4.2. It shows that only our method is able to mitigate the observational bias and exhibits a symmetric distribution, while the distribution of other methods is left skewed. As there exists some identification issue with our method, we run a simple linear regression using validation data for all methods and report the test error in Table 4.1, where our method achieves the best performance.

Table 4.1: Test root mean squared errors (TRMSE), test mean absolute errors (TMAE).

	Informative Missing	
Method	TRMSE	TMAE
SoftImpute [119]	1.0687	0.8149
CZ [99]	1.1193	0.8580
MFW [45]	1.0554	0.8036
SNN [106]	1.3110	0.9737
Our method	0.9462	0.7127

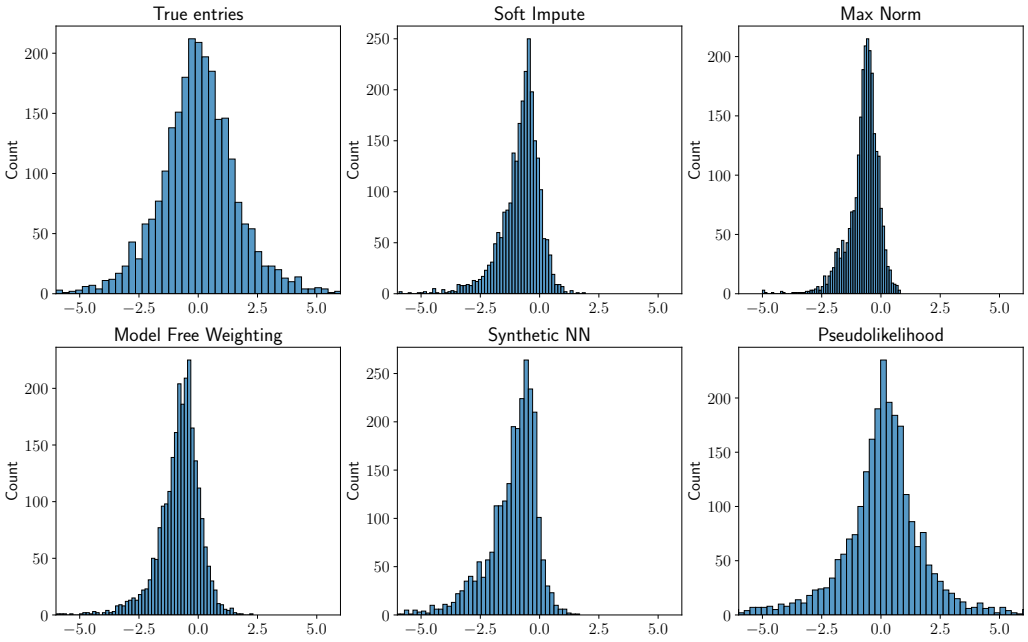


Figure 4.2: The recovered entries are left skewed from other methods.

5. SUMMARY

This dissertation presents several contributions towards learning algorithms under implicit bias and data bias. The research focuses on understanding the implicit bias of training algorithms in neural networks and developing effective algorithms for mitigating data bias.

We specifically study the training trajectories of diagonally linear neural networks with general depth-N [1]. We theoretically prove the implicit sparse regularization effect under the general correlated and noisy setting. We characterize the impact of depth and early stopping, and show that increasing depth enlarges the scale of working initialization and the early-stopping window so that this implicit sparse regularization effect is more likely to take place. Such insights and the associated techniques shed light on studying the training trajectories of more general deep models.

To further delve into the implicit regularization of gradient descent towards structured sparsity, we propose a novel neural reparameterization [2]. Through this approach, we uncover an intriguing property: gradient descent on the squared regression loss, without explicit regularization, exhibits a bias towards solutions with a group sparsity structure. In contrast to many existing works in understanding implicit regularization, we prove that our training trajectory cannot be simulated by mirror descent. We analyze the gradient dynamics of the corresponding regression problem in the general noise setting. This discovery not only advances our understanding of implicit regularization but also opens up new possibilities for leveraging this property to promote structured sparsity in various machine learning tasks.

Additionally, we explore the problem of matrix completion with informative missing. We propose a penalized pairwise pseudolikelihood approach to mitigate data bias in matrix completion. The proposed method does not suffer from data bias under a flexible and generally applicable assumption. We provide a near-optimal error bound after excluding the non-identifiable components and demonstrate the efficacy of the proposed method through numerical experiments.

Moving forward, we envision several promising future directions. Building upon our exploration of deeper models and richer implicit regularization effects, we aim to understand the learn-

ing algorithms under implicit bias for richer, deeper, and more general models. By discovering additional intriguing and fundamental implicit regularization effects, we anticipate gaining deeper insights into neural network training and generalization performance, leading to the development of more effective, reliable, and trustworthy artificial intelligence models. Given the ubiquity of data bias in real-world applications, stemming from human reporting and selection biases, as well as algorithmic and interpretation biases, we also aim to develop more efficient and powerful algorithms for data bias mitigation. Furthermore, we intend to extend the insights gained from our study of matrix completion with informative missing to other machine learning models, addressing data bias challenges across diverse applications.

By addressing these fundamental aspects, this thesis not only advances our understanding of implicit bias, and data bias mitigation but also offers valuable insights and techniques for overcoming challenges in modern machine learning. Through these contributions, we aim to drive the field forward, foster the development of more effective algorithms, and promote fairness and reliability in practical machine learning applications.

REFERENCES

[1] J. Li, T. Nguyen, C. Hegde, and R. K. W. Wong, "Implicit sparse regularization: The impact of depth and early stopping," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[2] J. Li, T. V. Nguyen, C. Hegde, and R. K. Wong, "Implicit regularization for group sparsity," in *The Eleventh International Conference on Learning Representations*, 2023.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[5] X. Wang, L. Zhang, and D. Klabjan, "Keyword-based topic modeling and keyword selection," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1148–1154, IEEE, 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[7] S. Shan, Y. Li, and J. B. Oliva, "Meta-neighborhoods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5047–5057, 2020.

[8] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[9] S. Shan, L. Hantrakul, J. Chen, M. Avent, and D. Trevelyan, "Differentiable wavetable synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), pp. 4598–4602, IEEE, 2022.

[10] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[11] J. Li and M. Armandpour, "Deep spatio-temporal wind power forecasting," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4138–4142, IEEE, 2022.

[12] S. Shan, Y. Li, and J. B. Oliva, "Nrtsi: Non-recurrent time series imputation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[13] L. Zhang, A. Ebrahimi, and D. Klabjan, "Layer flexible adaptive computation time," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, IEEE, 2021.

[14] L. Zhang, X. Chen, T. Chen, Z. Wang, and B. J. Mortazavi, "Dynehr: Dynamic adaptation of models with data heterogeneity in electronic health records," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–4, IEEE, 2021.

[15] K. K. Zhang, J. Li, M. Jeon, and K. S. Ramos, "Single-cell mrna sequencing into precision medicine: Promise and challenges," in *Reference Module in Biomedical Sciences*, Elsevier, 2023.

[16] L. Zhang, N. C. Hurley, B. Ibrahim, E. Spatz, H. M. Krumholz, R. Jafari, and M. J. Bobak, "Developing personalized models of blood pressure estimation from wearable sensors data using minimally-trained domain adversarial neural networks," in *Machine Learning for Healthcare Conference*, pp. 97–120, PMLR, 2020.

[17] L. Lu, X. Meng, Z. Mao, and G. E. Karniadakis, "Deepxde: A deep learning library for solving differential equations," *SIAM review*, vol. 63, no. 1, pp. 208–228, 2021.

[18] Y. Li, H. Yi, C. Bender, S. Shan, and J. B. Oliva, "Exchangeable neural ode for set modeling," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6936–6946, 2020.

[19] B. Lei, D. Xu, R. Zhang, S. He, and B. K. Mallick, "Balance is essence: Accelerating sparse training via adaptive gradient correction," *arXiv preprint arXiv:2301.03573*, 2023.

[20] G. Bellec, D. Kappel, W. Maass, and R. Legenstein, "Deep rewiring: Training very sparse deep networks," *arXiv preprint arXiv:1711.05136*, 2017.

[21] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International Conference on Machine Learning*, pp. 1225–1234, PMLR, 2016.

[24] L. Zhang, J. Zhang, B. Lei, S. Mukherjee, X. Pan, B. Zhao, C. Ding, Y. Li, and D. Xu, "Accelerating dataset distillation via model augmentation," *arXiv preprint arXiv:2212.06152*, 2022.

[25] B. Lei, R. Zhang, D. Xu, and B. Mallick, "Calibrating the rigged lottery: Making all tickets reliable," *arXiv preprint arXiv:2302.09369*, 2023.

[26] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[28] G. Vardi, "On the implicit bias in deep-learning algorithms," *Communications of the ACM*, vol. 66, no. 6, pp. 86–93, 2023.

[29] T. Vaskevicius, V. Kanade, and P. Rebeschini, "Implicit regularization for optimal sparse recovery," in *Advances in Neural Information Processing Systems*, pp. 2972–2983, 2019.

[30] P. Zhao, Y. Yang, and Q.-C. He, "Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression," *arXiv preprint arXiv:1903.09367*, 2019.

[31] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," in *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10, IEEE, 2018.

[32] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," in *Conference On Learning Theory*, pp. 2–47, PMLR, 2018.

[33] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," *arXiv preprint arXiv:1905.13655*, 2019.

[34] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," *arXiv preprint arXiv:1711.00501*, 2017.

[35] X. Wang, C. Wu, J. D. Lee, T. Ma, and R. Ge, "Beyond lazy training for over-parameterized tensor decomposition," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21934–21944, 2020.

[36] F. Williams, M. Trager, C. Silva, D. Panozzo, D. Zorin, and J. Bruna, "Gradient dynamics of shallow univariate relu networks," *arXiv preprint arXiv:1906.07842*, vol. 32, 2019.

[37] H. Jin and G. Montúfar, "Implicit bias of gradient descent for mean squared error regression with wide neural networks," *arXiv preprint arXiv:2006.07356*, 2020.

[38] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[39] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[40] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry, "Convergence of gradient descent on separable data," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428, PMLR, 2019.

[41] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[42] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[43] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[44] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[45] J. Wang, R. K. Wong, X. Mao, and K. C. G. Chan, "Matrix completion with model-free weighting," *arXiv preprint arXiv:2106.05850*, 2021.

[46] K.-Y. Liang and J. Qin, "Regression analysis under non-standard situations: a pairwise pseudolikelihood approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 4, pp. 773–786, 2000.

[47] Y. Ning, T. Zhao, and H. Liu, "A likelihood ratio framework for high-dimensional semiparametric regression," *The Annals of Statistics*, vol. 45, no. 6, pp. 2299–2327, 2017.

[48] J. Zhao, Y. Yang, and Y. Ning, "Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data," *Statistica Sinica*, vol. 28, no. 4, pp. 2125–2148, 2018.

[49] M. Li, M. Soltanolkotabi, and S. Oymak, "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks," in *International conference on artificial intelligence and statistics*, pp. 4313–4324, PMLR, 2020.

[50] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, "Geometry of optimization and implicit regularization in deep learning," *arXiv preprint arXiv:1705.03071*, 2017.

[51] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[52] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias-variance trade-off," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 32, pp. 15849–15854, 2019.

[53] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 67–83, 2020.

[54] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[55] S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry, "Implicit bias of gradient descent on linear convolutional networks," *Advances in Neural Information Processing Systems*, vol. 2018, pp. 9461–9471, 2018.

[56] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *International Conference on Machine Learning*, pp. 1832–1841, PMLR, 2018.

[57] D. Gissin, S. Shalev-Shwartz, and A. Daniely, "The implicit bias of depth: How incremental learning drives generalization," *arXiv preprint arXiv:1909.12051*, 2019.

[58] G. Gidel, F. Bach, and S. Lacoste-Julien, "Implicit regularization of discrete gradient dynamics in linear neural networks," *arXiv preprint arXiv:1904.13262*, 2019.

[59] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, "Kernel and rich regimes in overparametrized models," in *Conference on Learning Theory*, pp. 3635–3673, PMLR, 2020.

[60] E. Moroshko, B. E. Woodworth, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry, "Implicit bias in deep linear classification: Initialization scale vs training accuracy," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 22182–22193, Curran Associates, Inc., 2020.

[61] Z. Li, Y. Luo, and K. Lyu, "Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning," *arXiv preprint arXiv:2012.09839*, 2020.

[62] G. Neu and L. Rosasco, "Iterate averaging as regularization for stochastic gradient descent," in *Conference On Learning Theory*, pp. 3222–3242, PMLR, 2018.

[63] 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.

[64] A. Suggala, A. Prasad, and P. K. Ravikumar, "Connecting optimization and regularization paths," *Advances in Neural Information Processing Systems*, vol. 31, pp. 10608–10619, 2018.

[65] 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.

[66] 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.

[67] E. Candes, T. Tao, *et al.*, "The dantzig selector: Statistical estimation when p is much larger than n," *Annals of statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[68] K. Bredies and D. A. Lorenz, "Linear convergence of iterative soft-thresholding," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 813–837, 2008.

[69] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence of gradient methods for high-dimensional statistical recovery," *The Annals of Statistics*, pp. 2452–2482, 2012.

[70] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[71] S. Foucart and H. Rauhut, "An invitation to compressive sensing," in *A mathematical introduction to compressive sensing*, pp. 1–39, Springer, 2013.

[72] S. Foucart and J. Li, "Sparse recovery from inaccurate saturated measurements," *Acta Applicandae Mathematicae*, vol. 158, no. 1, pp. 49–66, 2018.

[73] L. Carin, D. Liu, and B. Guo, "Coherence, compressive sensing, and random sensor arrays," *IEEE Antennas and Propagation Magazine*, vol. 53, no. 4, pp. 28–39, 2011.

[74] 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.

[75] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.

[76] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[77] C. R. Vogel and M. E. Oman, "Iterative methods for total variation denoising," *SIAM Journal on Scientific Computing*, vol. 17, no. 1, pp. 227–238, 1996.

[78] Z. Dai, M. Karzand, and N. Srebro, "Representation costs of linear neural networks: Analysis and design," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[79] M. Jagadeesan, I. Razenshteyn, and S. Gunasekar, "Inductive bias of multi-channel linear convolutional networks with bounded weight norm," *arXiv preprint arXiv:2102.12238*, 2021.

[80] X. Wu, E. Dobriban, T. Ren, S. Wu, Z. Li, S. Gunasekar, R. Ward, and Q. Liu, "Implicit regularization and convergence for weight normalization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2835–2847, 2020.

[81] M. Pilanci and T. Ergen, "Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks," in *International Conference on Machine Learning*, pp. 7695–7705, PMLR, 2020.

[82] A. Sahiner, T. Ergen, J. Pauly, and M. Pilanci, "Vector-output relu neural network problems are copositive programs: Convex analysis of two layer networks and polynomial-time algorithms," *arXiv preprint arXiv:2012.13329*, 2020.

[83] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *International Conference on Machine Learning*, pp. 1832–1841, PMLR, 2018.

[84] J. Schwarz, S. Jayakumar, R. Pascanu, P. Latham, and Y. Teh, "Powerpropagation: A sparsity inducing weight reparameterisation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[85] R. Berthier, "Incremental learning in diagonal linear networks," *arXiv preprint arXiv:2208.14673*, 2022.

[86] H.-H. Chou, J. Maly, and H. Rauhut, "More is less: Inducing sparsity via overparameterization," *arXiv preprint arXiv:2112.11027*, 2021.

[87] S. S. Du, W. Hu, and J. D. Lee, "Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[88] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[89] D. Morwani and H. G. Ramaswamy, "Inductive bias of gradient descent for weight normalized smooth homogeneous neural nets," in *International Conference on Algorithmic Learning Theory*, pp. 827–880, PMLR, 2022.

[90] T. Van Laarhoven, "L2 regularization versus batch and weight normalization," *arXiv preprint arXiv:1706.05350*, 2017.

[91] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[92] X. Wu, Z. Zhang, W. Zhang, Y. Yi, C. Zhang, and Q. Xu, "A convolutional neural network based on grouping structure for scene classification," *Remote Sensing*, vol. 13, no. 13, p. 2457, 2021.

[93] G. Xie, C. Dong, Y. Kong, J. F. Zhong, M. Li, and K. Wang, "Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features," *Genes*, vol. 10, no. 3, p. 240, 2019.

[94] Y. C. Eldar and H. Bolcskei, "Block-sparsity: Coherence and efficient recovery," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2885–2888, IEEE, 2009.

[95] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on information theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[96] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3075–3085, 2009.

[97] Z. Li, T. Wang, J. Lee, and S. Arora, "Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent," *arXiv preprint arXiv:2207.04036*, 2022.

[98] S. Arora, N. Cohen, and E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization," in *International Conference on Machine Learning*, pp. 244–253, PMLR, 2018.

[99] T. T. Cai and W.-X. Zhou, "Matrix completion via max-norm constrained optimization," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1493–1525, 2016.

[100] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *The Annals of Statistics*, pp. 177–214, 2015.

[101] D. Song, C. E. Lee, Y. Li, and D. Shah, "Blind regression: Nonparametric regression for latent variable models via collaborative filtering," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[102] X. Mao, R. K. Wong, and S. X. Chen, "Matrix completion under low-rank missing mechanism," *arXiv preprint arXiv:1812.07813*, 2018.

[103] W. Ma and G. H. Chen, "Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption," *Advances in neural information processing systems*, vol. 32, pp. 14871–14880, 2019.

[104] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *international conference on machine learning*, pp. 1670–1679, PMLR, 2016.

[105] M. Udell and A. Townsend, "Why are big data matrices approximately low rank?," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 1, pp. 144–160, 2019.

[106] A. Agarwal, M. Dahleh, D. Shah, and D. Shen, "Causal matrix completion," *arXiv preprint arXiv:2109.15154*, 2021.

[107] X. Mao, S. X. Chen, and R. K. Wong, "Matrix completion with covariate information," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 198–210, 2019.

[108] H. Jin, Y. Ma, and F. Jiang, "Matrix completion with covariate information and informative missingness," *Journal of Machine Learning Research*, vol. 23, no. 180, pp. 1–62, 2022.

[109] 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.

[110] T. Cai and W.-X. Zhou, "A max-norm constrained minimization approach to 1-bit matrix completion.," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3619–3647, 2013.

[111] O. Klopp, K. Lounici, and A. B. Tsybakov, "Robust matrix completion," *Probability Theory and Related Fields*, vol. 169, pp. 523–564, 2017.

[112] Y. Cao and Y. Xie, "Poisson matrix recovery and completion," *IEEE Transactions on Signal Processing*, vol. 64, no. 6, pp. 1609–1620, 2015.

[113] S. Gunasekar, P. Ravikumar, and J. Ghosh, "Exponential family matrix completion under structural constraints," in *International Conference on Machine Learning*, pp. 1917–1925, PMLR, 2014.

[114] J. Lafond, "Low rank matrix completion with exponential family noise," in *Conference on Learning Theory*, pp. 1224–1243, PMLR, 2015.

[115] G. Tang, R. J. Little, and T. E. Raghunathan, "Analysis of multivariate missing data with nonignorable nonresponse," *Biometrika*, vol. 90, no. 4, pp. 747–764, 2003.

[116] J. D. Kalbfleisch, "Likelihood methods and nonparametric tests," *Journal of the American Statistical Association*, vol. 73, no. 361, pp. 167–170, 1978.

[117] R. Foygel, O. Shamir, N. Srebro, and R. R. Salakhutdinov, "Learning with the weighted trace-norm under arbitrary sampling distributions," *Advances in neural information processing systems*, vol. 24, 2011.

[118] N. Srebro and R. R. Salakhutdinov, "Collaborative filtering in a non-uniform world: Learning with the weighted trace norm," *Advances in neural information processing systems*, vol. 23, 2010.

[119] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[120] T. K. Lee, W. J. Baddar, S. T. Kim, and Y. M. Ro, "Convolution with logarithmic filter groups for efficient shallow cnn," in *International Conference on Multimedia Modeling*, pp. 117–129, Springer, 2018.

[121] L. Jing, J. Zbontar, *et al.*, "Implicit rank-minimizing autoencoder," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14736–14746, 2020.

[122] I. Carmichael, T. Keefe, N. Giertych, and J. P. Williams, "yaglm: a python package for fitting and tuning generalized linear models that supports structured, adaptive and non-convex penalties," *arXiv preprint arXiv:2110.05567*, 2021.

[123] C. Molinari, M. Massias, L. Rosasco, and S. Villa, "Iterative regularization for convex regularizers," in *International conference on artificial intelligence and statistics*, pp. 1684–1692, PMLR, 2021.

[124] T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, *et al.*, "Regulation of gene expression in the mammalian eye and its relevance to eye disease," *Proceedings of the National Academy of Sciences*, vol. 103, no. 39, pp. 14429–14434, 2006.

[125] Y. Yang and H. Zou, "A fast unified algorithm for solving group-lasso penalize learning problems," *Statistics and Computing*, vol. 25, pp. 1129–1141, 2015.

[126] J. Wang, R. K. Wong, and X. Zhang, "Low-rank covariance function estimation for multidimensional functional data," *Journal of the American Statistical Association*, vol. 117, no. 538, pp. 809–822, 2022.

[127] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, pp. 389–434, 2012.

[128] W. N. Anderson Jr and T. D. Morley, "Eigenvalues of the laplacian of a graph," *Linear and multilinear algebra*, vol. 18, no. 2, pp. 141–145, 1985.

[129] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[130] O. Klopp, "Rank penalized estimators for high-dimensional matrices," *Electronic Journal of Statistics*, vol. 5, pp. 1161–1183, 2011.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER II

The appendix is organized as follows.

In Appendix A.1, we present a simplied theorem about non-negative signals and illustrate the idea behind the proof.

In Appendix A.2, we study the multiplicative updates and build connections to its continuous approximation, which will be used next.

In Appendix A.3, we provide the proof of propositions and technical lemmas in Appendix A.1.

In Appendix A.4, we prove the main results stated in the paper.

In Appendix A.5, we provide the experimental results on real-world datasets to illustrate the effectiveness of the proposed algorithm.

A.1 Proof for Non-negative Signals

We mainly follow the proof structure from [29] to obtain the convergence of similar gradient descent algorithm for the case $N = 2$, which is a limiting case of ours. We will demonstrate how gradient dynamics changes with $N > 2$, which requires us to study the growth rate of error and convergence rate more carefully.

In this section, we will start with the general set up and provide a simplified version of Theorem 1 about non-negative signals.

A.1.1 Setup

The gradients of $\mathcal{L}(\mathbf{u}, \mathbf{v})$ with respect to $\mathbf{u}, \mathbf{v}$ read as

$$\nabla_{\mathbf{u}}\mathcal{L}(\mathbf{w}) = \frac{2N}{n}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) \odot \mathbf{u}^{N-1}$$
$$\nabla_{\mathbf{v}}\mathcal{L}(\mathbf{w}) = -\frac{2N}{n}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) \odot \mathbf{v}^{N-1}.$$

With the step size η, the gradient descent updates on $\mathbf{u}_t$ and $\mathbf{v}_t$ simply are

$$\mathbf{u}_{t+1} = \mathbf{u}_t \odot \left(\mathbf{1} - 2N\eta\left(\frac{1}{n}\mathbf{X}^\mathsf{T}(\mathbf{X}(\mathbf{w}_t - \mathbf{w}^\star) - \boldsymbol{\xi}) \odot \mathbf{u}_t^{N-2}\right)\right),$$
$$\mathbf{v}_{t+1} = \mathbf{v}_t \odot \left(\mathbf{1} + 2N\eta\left(\frac{1}{n}\mathbf{X}^\mathsf{T}(\mathbf{X}(\mathbf{w}_t - \mathbf{w}^\star) - \boldsymbol{\xi}) \odot \mathbf{v}_t^{N-2}\right)\right).$$

Let $\mathbf{w}_t = \mathbf{w}_t^+ - \mathbf{w}_t^-$ where $\mathbf{w}_t^+ \coloneqq \mathbf{u}_t^N$ and $\mathbf{w}_t^- \coloneqq \mathbf{v}_t^N$ with the power taken element-wisely. We denote S as the support of $\mathbf{w}^\star$, and let $S^+ = \{i | w_i^\star > 0\}$ denote the index set of coordinates with positive values, and $S^- = \{i | w_i^\star < 0\}$ denote the index set of coordinates with negative values. Therefore $S = S^+ \cup S^-$ and $S^+ \cap S^- = \emptyset$. Then define the following signal and noise-related quantities:

$$\begin{aligned}
\mathbf{s}_t &\coloneqq \mathbf{1}_{S^+} \odot \mathbf{w}_t^+ - \mathbf{1}_{S^-} \odot \mathbf{w}_t^-, \\
\mathbf{e}_t &\coloneqq \mathbf{1}_{S^c} \odot \mathbf{w}_t + \mathbf{1}_{S^-} \odot \mathbf{w}_t^+ - \mathbf{1}_{S^+} \odot \mathbf{w}_t^-, \\
\mathbf{b}_t &\coloneqq \frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{e}_t - \frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}, \\
\mathbf{p}_t &\coloneqq \left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{I}\right)(\mathbf{s}_t - \mathbf{w}^\star).
\end{aligned} \tag{A.1}$$

Let α^N be the initial value for each entry of $\mathbf{w}$ and rewrite the updates on $\mathbf{w}_t$, $\mathbf{w}_t^+$ and $\mathbf{w}_t^-$ in a more succinct way:

$$\begin{aligned}
\mathbf{w}_0^+ &= \mathbf{w}_0^- = \alpha^N, \\
\mathbf{w}_t &= \mathbf{w}_t^+ - \mathbf{w}_t^-, \\
\mathbf{w}_{t+1}^+ &= \mathbf{w}_t^+ \odot \left(\mathbf{1} - 2N\eta\left(\mathbf{s}_t - \mathbf{w}^\star + \mathbf{p}_t + \mathbf{b}_t\right) \odot \left(\mathbf{w}_t^+\right)^{(N-2)/N}\right)^N, \\
\mathbf{w}_{t+1}^- &= \mathbf{w}_t^- \odot \left(\mathbf{1} + 2N\eta\left(\mathbf{s}_t - \mathbf{w}^\star + \mathbf{p}_t + \mathbf{b}_t\right) \odot \left(\mathbf{w}_t^-\right)^{(N-2)/N}\right)^N.
\end{aligned} \tag{A.2}$$

When our target $\mathbf{w}^\star$ is with non-negative entries, the design of $\mathbf{v}_t$ is no longer needed and the

algorithm could be simplied to the following form.

$$\begin{aligned}
\mathbf{w}_0^+ &= \mathbf{u}_0^N = \alpha^N, \\
\mathbf{w}_t^+ &= \mathbf{u}_t^N, \\
\mathbf{w}_{t+1}^+ &= \mathbf{w}_t^+ \odot \left(\mathbf{1} - 2N\eta \left(\mathbf{s}_t - \mathbf{w}^\star + \mathbf{p}_t + \mathbf{b}_t\right) \odot \left(\mathbf{w}_t^+\right)^{(N-2)/N}\right)^N
\end{aligned} \tag{A.3}$$

The results in this section are all about updates in equation (A.3), and will be generalized to updates in equation (A.2) in Section A.4.

A.1.2 The Key Propositions

Starting from $t = 0$, we have $\|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty \lesssim \mathcal{O}(w^\star_{\max})$ and $\|\mathbf{e}_0\|_\infty \leq \alpha^N$. The idea of proposition 1 is to show that after some certain number of iterations t, we obtain $\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty \lesssim \mathcal{O}(w^\star_{\min})$ and $\|\mathbf{e}_t\|_\infty \leq \alpha^{N/2}$. Proposition 2 further reduces the approximation error from $\mathcal{O}(w^\star_{\min})$ to $\mathcal{O}(\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty)$ if possible, while still maintaining $\|\mathbf{e}_t\|_\infty \leq \alpha^{N/4}$.

Proposition 1. *Consider the updates in equations* (A.3)*. Fix any* $0 < \zeta \leq w^\star_{\max}$ *and let* $\gamma = C_\gamma \frac{w^\star_{\min}}{w^\star_{\max}}$ *where* C_γ *is some small enough absolute constant. Suppose the error sequences* $(\mathbf{b}_t)_{t\geq 0}$ *and* $(\mathbf{p}_t)_{t\geq 0}$ *for any* $t \geq 0$ *satisfy the following:*

$$\begin{aligned}
\|\mathbf{b}_t\|_\infty &\leq C_b\zeta - \alpha^{N/4}, \\
\|\mathbf{p}_t\|_\infty &\leq \gamma \|\mathbf{s}_t - \mathbf{w}^\star\|_\infty,
\end{aligned}$$

where C_b *is some small enough absolute constants. If the initialization satisfies*

$$\alpha \leq \left(\frac{1}{8}\right)^{2/(N-2)} \wedge \left(\frac{(w^\star_{\max})^{(N-2)/N}}{\log \frac{w^\star_{\max}}{\epsilon}}\right)^{2/(N-2)},$$

and the step size $\eta \leq \frac{\alpha^N}{8N^2\zeta^{(3N-2)/N}}$*, then for any* $T_1 \leq T \leq T_2$ *where*

$$T_1 = \frac{75}{16\eta N^2 \zeta^{(2N-2)/N}} \log \frac{|w^\star_{\max} - \alpha^N|}{\epsilon} + \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}},$$
$$T_2 = \frac{5}{N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}}\right),$$

and any $0 \leq t \leq T$*, we have*

$$\|\mathbf{s}_T - \mathbf{w}^\star\|_\infty \leq \zeta,$$
$$\|\mathbf{e}_t\|_\infty \leq \alpha^{N/2}.$$

Note that the requirement on $\|\mathbf{b}_t\|_\infty \leq C_b\zeta - \alpha^{N/4}$ can be relaxed to $\|\mathbf{b}_t\|_\infty \leq C_b\zeta$ when we just consider the updates in equation (A.3). However, we still consider the stronger requirement in order to further generalize to updates in equation (A.2) later.

Proposition 2. *Consider the updates in equations* (A.3)*. Fix any* $0 < \zeta \leq w^\star_{\max}$ *and suppose that the error sequences* $(\mathbf{b}_t)_{t\geq 0}$ *and* $(\mathbf{p}_t)_{t\geq 0}$ *for any* $t \geq 0$ *satisfy*

$$B = \|\mathbf{b}_t\|_\infty + \|\mathbf{p}_t\|_\infty \leq \frac{1}{200} w^\star_{\min}$$
$$\|\mathbf{b}_t \odot \mathbf{1}_i\|_\infty \leq B_i \leq \frac{1}{10} w^\star_{\min},$$
$$\|\mathbf{p}_t\|_\infty \leq \frac{1}{20}\|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty.$$

Suppose that

$$\alpha \le \left(\frac{1}{4}\right)^{2/(N-2)} \wedge \left(\frac{(w^\star_{\min})^{(N-2)/N}}{\log \frac{w^\star_{\min}}{\epsilon}}\right)^{4/(N-2)},$$
$$\|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty \le \frac{1}{5} w^\star_{\min},$$
$$\|\mathbf{e}_0\| \le \alpha^{N/2}.$$

Let the step size satisfy $\eta \le \frac{\alpha^N}{8N^2(w^\star_{\min})^{(3N-2)/N}}$. *Then for any* $T_3 \le t \le T_4$,

$$T_3 = \frac{6}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon},$$
$$T_4 = \frac{25}{N(N-1)\eta w^\star_{\min}} \left(\frac{1}{\alpha^{(N-2)/2}} - \frac{1}{\alpha^{(N-2)/4}}\right),$$

and any $i \in S$ *we have*

$$|s_{i,t} - w^\star_i| \lesssim k\mu \max_{j \in S} B_j \vee B_i \vee \epsilon,$$
$$\|\mathbf{e}_t\|_\infty \le \alpha^{N/4}.$$

A.1.3 Technical Lemmas

There are several lemmas, which are about the coherence of the design matrices and the upper bound of subGaussian noise term.

Lemma 4. *Suppose that* $\frac{1}{\sqrt{n}}\mathbf{X}$ *is a* $n \times p$ *matrix with* ℓ_2*-normalized columns and satisfies* μ*-coherence with* $0 \le \mu \le 1$. *Then for any vector* $\mathbf{z} \in \mathbb{R}^p$ *we have*

$$\left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{z} \right\|_\infty \le p \|\mathbf{z}\|_\infty .$$

Lemma 5. *Suppose that* $\frac{1}{\sqrt{n}}\mathbf{X}$ *is a* $n \times p$ ℓ_2*-normalized matrix satisfying* μ*-incoherence; that is* $\frac{1}{n}|\mathbf{X}_i^\top \mathbf{X}_j| \le \mu, i \neq j$*. For* k*-sparse vector* $\mathbf{z} \in \mathbb{R}^p$*, we have:*

$$\left\| \left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{I} \right) \mathbf{z} \right\|_\infty \le k\mu \left\| \mathbf{z} \right\|_\infty .$$

Lemma 6. *Let* $\frac{1}{\sqrt{n}}\mathbf{X}$ *be a* $n \times p$ *matrix with* ℓ_2*-normalized columns. Let* $\boldsymbol{\xi} \in \mathbb{R}^n$ *be a vector of independent* σ^2*-sub-Gaussian random variables. Then, with probability at least* $1 - \frac{1}{8p^3}$

$$\left\| \frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi} \right\|_\infty \lesssim \sqrt{\frac{\sigma^2 \log p}{n}}.$$

A.1.4 Proof for Non-negative Signals

Recall the notation

$$\Phi(w^\star_{\max}, w^\star_{\min}, \epsilon, N) \coloneqq \left(\frac{1}{8}\right)^{2/(N-2)} \wedge \left(\frac{(w^\star_{\max})^{(N-2)/N}}{\log \frac{w^\star_{\max}}{\epsilon}} \right)^{2/(N-2)} \wedge \left(\frac{(w^\star_{\min})^{(N-2)/N}}{\log \frac{w^\star_{\min}}{\epsilon}} \right)^{4/(N-2)},$$

and

$$\zeta \coloneqq \frac{1}{5} w^\star_{\min} \vee \frac{200}{n} \left\| \mathbf{X}^\mathsf{T}\boldsymbol{\xi} \right\|_\infty \vee 200\epsilon.$$

Theorem 8. *Suppose that* $\mathbf{w}^\star \succcurlyeq 0$ *with* $k \ge 1$ *and* $\mathbf{X}/\sqrt{n}$ *satisfies* μ*-incoherence with* $\mu \le C_\gamma / kr$*, where* C_γ *is some small enough constant. Take any precision* $\epsilon > 0$*, and let the initialization be such that*

$$0 < \alpha \le \left(\frac{\epsilon}{p+1} \right)^{4/N} \wedge \Phi(w^\star_{\max}, w^\star_{\min}, \epsilon, N)$$

For any iteration t *that satisfies*

$$\frac{1}{\eta N^2 \zeta^{(2N-2)/N} \alpha^{N-2}} \lesssim t \lesssim \frac{1}{\eta N^2 \tau} \left(\frac{1}{\alpha^{N-2}} - \frac{1}{\zeta^{(N-2)/2}} \right),$$

the gradient descent algorithm (A.3) *with step size* $\eta \le \frac{\alpha^N}{8N^2 (w^\star_{\max})^{(3N-2)/N}}$ *yields the iterate* $\mathbf{w}_t$ *with*

the following property:

$$|w_{t,i} - w_i^\star| \lesssim \begin{cases} \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \epsilon & \text{if } i \in S \text{ and } w_{\min}^\star \lesssim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \epsilon, \\ \left|\frac{1}{n}(\mathbf{X}^\mathsf{T}\boldsymbol{\xi})_i\right| \vee k\mu \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi} \odot \mathbf{1}_S\right\|_\infty \vee \epsilon & \text{if } i \in S \text{ and } w_{\min}^\star \gtrsim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \epsilon, \\ \alpha^{N/4} & \text{if } i \notin S. \end{cases} \tag{A.4}$$

Proof. Let

$$\zeta \coloneqq \frac{1}{5}w_{\min}^\star \vee \frac{2}{C_b}\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \frac{2}{C_b}\epsilon,$$

where C_b is some small enough positive constant that will be explicitly derived later. Also by the requirement of the coherence of the design matrix, we have

$$\|\mathbf{p}_t\|_\infty \leq \frac{C_\gamma}{w_{\max}^\star / w_{\min}^\star}\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty .$$

Setting

$$\alpha \leq \left(\frac{\epsilon}{p+1}\right)^{4/N} \wedge \left(\frac{1}{8}\right)^{2/(N-2)} \wedge \left(\frac{(w_{\max}^\star)^{(N-2)/N}}{\log\frac{w_{\max}^\star}{\epsilon}}\right)^{2/(N-2)} \wedge \left(\frac{(w_{\min}^\star)^{(N-2)/N}}{\log\frac{w_{\min}^\star}{\epsilon}}\right)^{4/(N-2)}.$$

As long as $\|\mathbf{e}_t\|_\infty \leq \alpha^{N/4}$ we have

$$\begin{aligned} \|\mathbf{b}_t\|_\infty + \alpha^{N/4} &\leq \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\epsilon\right\|_\infty + \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{e}_t\right\|_\infty + \alpha^{N/4} \\ &\leq 2\left(\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\epsilon\right\|_\infty \vee (p\|\mathbf{e}_t\|_\infty\right) + \alpha^{N/4}) \\ &\leq 2\left(\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\epsilon\right\|_\infty \vee (p+1)\alpha^{N/4}\right) \\ &\leq C_b\frac{2}{C_b}\left(\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\epsilon\right\|_\infty \vee \epsilon\right) \\ &\leq C_b\zeta. \end{aligned}$$

where the second inequality is from Lemma 4. Further by Lemma 5, we also have

$$\|\mathbf{p}_t\|_\infty \leq \frac{C_\gamma}{w^\star_{\max}/w^\star_{\min}} \|\mathbf{s}_t - \mathbf{w}^\star\|_\infty .$$

Therefore, both sequences $(\mathbf{b}_t)_{t\geq 0}$ and $(\mathbf{p}_t)_{t\geq 0}$ satisfy the assumptions of Proposition 1 conditionally on $\|\mathbf{e}_t\|_\infty$ staying below $\alpha^{N/4}$. If $\zeta \geq w^\star_{\max}$, at $t = 0$, we have already have

$$\|\mathbf{s}_0 - \mathbf{w}^\star\| \leq \zeta.$$

Otherwise, applying Proposition 1, after

$$T_1 = \frac{75}{16\eta N^2 \zeta^{(2N-2)/N}} \log \frac{|w^\star_{\max} - \alpha^N|}{\epsilon} + \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}},$$

iterations and before

$$T_2 = \frac{5}{N(N-1)\eta\zeta} \left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}} \right)$$

iterations, we have

$$\|\mathbf{s}_{T_1} - \mathbf{w}^\star\| \leq \zeta,$$
$$\|\mathbf{e}_{T_1}\|_\infty \leq \alpha^{N/2}.$$

If $\frac{1}{5} w^\star_{\min} \leq \frac{2}{C_b} \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty \vee \frac{2}{C_b} \epsilon$, then we are done.

If $\frac{1}{5} w^\star_{\min} > \frac{2}{C_b} \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty \vee \frac{2}{C_b} \epsilon$, we have $\zeta = \frac{1}{5} w^\star_{\min}$. Choose $C_b + C_\gamma \leq \frac{1}{40}$ as we have in Proposition 1. After T_1 iterations, we have

$$\|\mathbf{b}_t\|_\infty + \|\mathbf{p}_t\|_\infty \leq C_b \frac{1}{5} w^\star_{\min} + \frac{C_\gamma}{w^\star_{\max}/w^\star_{\min}} \frac{1}{5} w^\star_{\min} \leq (C_b + C_\gamma) \frac{1}{5} w^\star_{\min} \leq \frac{1}{200} w^\star_{\min}.$$

Now all the assumptions of Proposition 2 are satisfied. To further reduce $\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty$ from

$\frac{1}{5}w^\star_{\min}$ to $\mathcal{O}(\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|)$, we apply Proposition 2 and obtain that after

$$T_3 = \frac{6}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon}$$

iterations and before

$$T_4 = \frac{25}{N(N-1)\eta w^\star_{\min}} \left(\frac{1}{\alpha^{(N-2)/2}} - \frac{1}{\alpha^{(N-2)/4}} \right)$$

iterations, we have for any $i \in S$,

$$|s_{t,i} - w^\star_i| \lesssim k\mu \max_{j \in S} B_j \vee B_i \vee \epsilon,$$
$$\|\mathbf{e}_t\|_\infty \leq \alpha^{N/4}.$$

We use $\mathbb{1}\{\cdot\}$ to denote the indicator function. Therefore, the total number of iterations needed is

$$\begin{aligned} T_1 + T_3 = {} & \frac{75}{16\eta N^2 \zeta^{(2N-2)/N}} \log \frac{|w^\star_{\max} - \alpha^N|}{\epsilon} + \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}} \\ & + \frac{6}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon} \mathbb{1}\left\{ \frac{1}{5} w^\star_{\min} > \frac{2}{C_b} \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty \vee \frac{2}{C_b} \epsilon \right\} \end{aligned} \tag{A.5}$$

and the upper bound for the total number of iterations would be

$$\begin{aligned} T_2 + T_4 = {} & \frac{5}{N(N-1)\eta\zeta} \left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}} \right) \\ & + \frac{25}{N(N-1)\eta w^\star_{\min}} \left(\frac{1}{\alpha^{(N-2)/2}} - \frac{1}{\alpha^{(N-2)/4}} \right) \mathbb{1}\left\{ \frac{1}{5} w^\star_{\min} > \frac{2}{C_b} \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty \vee \frac{2}{C_b} \epsilon \right\} \end{aligned} \tag{A.6}$$

□

A.2 Multiplicative Update Sequences with General Order N

In this section, we analyze the one-dimensional updates that exhibits the similar dynamics to our gradient descent algorithm. The lemmas we derive will be assembled together to prove Proposition 1 and 2. The whole framework is similar to [29]. However, the continuous approximation plays an important role to deal with $N > 2$, and the detailed derivation differs from [29] a lot, especially for Lemma 8, 11 and 18.

A.2.1 Error Growth

Lemma 7. *Consider the setting of updates given in equations* (A.2)*. Suppose that* $\|\mathbf{e}_t\|_\infty \leq \frac{1}{8} w^\star_{min}$ *and there exists some* $B \in \mathbb{R}$ *such that for all* t *we have* $\|\mathbf{b}_t\|_\infty + \|\mathbf{p}_t\|_\infty \leq B$*. Then, if* $\eta \leq \frac{1}{12(w^\star_{\max}+B)}$ *for any* $t \geq 0$ *we have*

$$\|\mathbf{e}_t\|_\infty \leq \|\mathbf{e}_0\|_\infty \prod_{i=1}^{t-1} (1 + 2N\eta(\|\mathbf{b}_i\|_\infty + \|\mathbf{p}_i\|_\infty) \|\mathbf{e}_i\|_\infty^{(N-2)/N})^N$$

or in the other form,

$$\|\mathbf{e}_{t+1}\|_\infty \leq \|\mathbf{e}_t\|_\infty (1 + 2N\eta(\|\mathbf{b}_t\|_\infty + \|\mathbf{p}_t\|_\infty) \|\mathbf{e}_t\|_\infty^{(N-2)/N})^N.$$

Proof. From the equations above, we get

$$\begin{aligned}
\mathbf{1}_{S^c} \odot \mathbf{e}_{t+1} &= \mathbf{1}_{S^c} \odot \mathbf{w}_t \odot (\mathbf{1} - 2N\eta(\mathbf{s}_t - \mathbf{w}^\star + \mathbf{p}_t + \mathbf{b}_t) \odot \mathbf{w}_t^{(N-2)/N})^N \\
&= \mathbf{1}_{S^c} \odot \mathbf{e}_t \odot (\mathbf{1}_{S^c} - \mathbf{1}_{S^c} 2N\eta(\mathbf{s}_t - \mathbf{w}^\star + \mathbf{p}_t + \mathbf{b}_t) \odot \mathbf{e}_t^{(N-2)/N})^N \\
&= \mathbf{1}_{S^c} \odot \mathbf{e}_t \odot (\mathbf{1} - 2N\eta(\mathbf{p}_t + \mathbf{b}_t) \odot \mathbf{e}_t^{(N-2)/N})^N
\end{aligned}$$

and hence

$$\|\mathbf{1}_{S^c} \odot \mathbf{e}_{t+1}\|_\infty \leq \|\mathbf{e}_t\|_\infty (1 + 2N\eta(\|\mathbf{b}_t\|_\infty + \|\mathbf{p}_t\|_\infty) \|\mathbf{e}_t\|_\infty^{(N-2)/N})^N.$$

□

When we have the bound for $\|\mathbf{b}_t\|_\infty + \|\mathbf{p}_t\|_\infty$, we can control the size of $\|\mathbf{e}_t\|_\infty$ by the following lemma.

Lemma 8. *Let $(b_t)_{t\geq 0}$ be a sequence such that for $t \geq 0$ we have $|b_t| \leq B$ for some $B > 0$. Let the step size satisfy $\eta \leq \frac{1}{4N(N-1)Bx_0^{(N-2)/(2N)}}$ and consider a one-dimensional sequence $(x_t)_{t\geq 0}$ given by*

$$\begin{aligned} 0 <& x_0 < 1, \\ x_{t+1} =& x_t(1 + 2N\eta b_t x_t^{(N-2)/N})^N. \end{aligned}$$

Then for any $t < \frac{1}{8N(N-1)\eta B}\left(\frac{1}{x_0^{(N-2)/N}} - \frac{1}{x_0^{(N-2)/2N}}\right)$ we have

$$x_t \leq \sqrt{x_0}.$$

Proof. We start with studying the larger increasing rate of the updates,

$$\begin{aligned} x_{t+1} &= x_t(1 + 2N\eta b_t x_t^{(N-2)/N})^N \\ &\leq x_t(1 + 2N\eta B x_t^{(N-2)/N})^N \\ &\leq x_t\left(1 + \frac{2N^2\eta B x_t^{(N-2)/N}}{1 - 2(N-1)N\eta x_t^{(N-2)/N}}\right) \\ &\leq x_t(1 + 4N^2\eta B x_t^{(N-2)/N}), \end{aligned}$$

where the second inequality is obtained by $(1+x)^r \leq 1 + \frac{rx}{1-(r-1)x}$ for $x \in (0, \frac{1}{r-1})$, and the last inequality is by the requirement of step size η. Therefore, to achieve to some value x_T, the number

of iterations needed is lower bounded as

$$T \geq \sum_{t=0}^{T-1} \frac{x_{t+1} - x_t}{4N^2\eta B x_t^{(2N-2)/N}}.$$

We aim at the number of iterations for $\sqrt{x_0}$, and we denote T as the maximal number of iterations, i.e. $x_T < \sqrt{x_0}$ and $x_{T+1} \geq \sqrt{x_0}$. Therefore,

$$\frac{\sqrt{x_0} - x_T}{4N^2\eta B x_T^{(2N-2)/N}} \leq \frac{x_{T+1} - x_T}{4N^2\eta B x_T^{(2N-2)/N}} \leq 1.$$

And for T, we derive the lower bound as

$$\begin{aligned}
T \geq \sum_{t=0}^{T-1} \frac{x_{t+1} - x_t}{4N^2\eta B x_t^{(2N-2)/N}} &\geq \frac{1}{4N^2\eta B} \sum_{t=0}^{T-1} \int_{x_t}^{x_{t+1}} \frac{1}{x^{(2N-2)/N}} dx \\
&\geq \frac{1}{4N^2\eta B} \int_{x_0}^{x_T} \frac{1}{x^{(2N-2)/N}} dx \\
&\geq \frac{1}{4N^2\eta B} \int_{x_0}^{\sqrt{x_0}} \frac{1}{x^{(2N-2)/N}} dx - \frac{1}{4N^2\eta B} \int_{x_T}^{\sqrt{x_0}} \frac{1}{x^{(2N-2)/N}} dx \\
&> \frac{1}{4N^2\eta B} \left(-\frac{N}{2N-2} \frac{1}{x^{(N-2)/N}} \right) \Bigg|_{x_0}^{\sqrt{x_0}} - 1 \\
&= \frac{1}{8N(N-1)\eta B} \left(\frac{1}{x_0^{(N-2)/N}} - \frac{1}{x_0^{(N-2)/2N}} \right) - 1.
\end{aligned}$$

Therefore, we know that for any $t \leq \frac{1}{8N(N-1)\eta B} \left(\frac{1}{x_0^{(N-2)/N}} - \frac{1}{x_0^{(N-2)/2N}} \right) - 1$, we have $x_t \leq \sqrt{x_0}$. Since in practice t is chosen as an integer, without loss of generality, we simply the requirement as $t < \frac{1}{8N(N-1)\eta B} \left(\frac{1}{x_0^{(N-2)/N}} - \frac{1}{x_0^{(N-2)/2N}} \right)$. □

A.2.2 Understanding 1-d Case

A.2.2.1 Basic Setting

In this subsection we analyze one-dimensional sequences with positive target corresponding to gradient descent updates without any perturbations. That is, $\mathbf{w}_t = \mathbf{u}_t^N$, $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$ and ignoring the error sequences $(\mathbf{b}_t)_{t\geq 0}$ and $(\mathbf{p}_t)_{t\geq 0}$. Hence, we will look at one-dimensional sequences of the

form

$$\begin{aligned} 0 < x_0 = \alpha^N < x^\star \\ x_{t+1} = x_t(1 - 2N\eta(x_t - x^\star)x_t^{(N-2)/N})^N. \end{aligned} \tag{A.7}$$

Lemma 9 (Iterates behave monotonically)**.** *Let $\eta > 0$ be the step size and suppose the updates are given by*

$$x_{t+1} = x_t(1 - 2N\eta(x_t - x^\star)x_t^{(N-2)/N})^N.$$

Then the following holds

1. *If $0 < x_0 \leq x^\star$ and $\eta \leq \frac{1}{2N(2N-2)(x^\star)^{(2N-2)/N}}$ then for any $t > 0$ we have $x_0 \leq x_{t-1} \leq x_t \leq x^\star$.*

2. *If $x^\star \leq x_0 \leq \frac{3}{2}x^\star$ and $\eta \leq \frac{1}{6N^2(x^\star)^{(2N-2)/N}}$ then for any $t \geq 0$ we have $x^\star \leq x_t \leq x_{t-1} \leq \frac{3}{2}x^\star$.*

Proof. Note that if $x_0 \leq x_t \leq x^\star$ then $x_t - x^\star \leq 0$ and hence $x_{t+1} \geq x_t$. Thus for the first part it is enough to show that for all $t \geq 0$ we have $x_t \leq x \leq x^\star$.

Assume for a contradiction that exists t such that

$$\begin{aligned} x_0 \leq x_t \leq x^\star, \\ x_{t+1} > x^\star. \end{aligned}$$

Plugging in the update rule for x_{t+1} we can rewrite the above as

$$\begin{aligned} x_t &\leq x^\star \\ &< x_t(1 - 2N\eta(x_t - x^\star)x_t^{(N-2)/N})^N \\ &\leq x_t\left(1 + \frac{1}{2N-2} - \frac{x_t^{(2N-2)/N}}{(2N-2)(x^\star)^{(2N-2)/N}}\right)^N \end{aligned}$$

Letting $\lambda = \left(\frac{x_t}{x^\star}\right)^{(2N-2)/N}$, by our assumption we have $0 < \lambda \leq 1$. The above inequality gives us

$$\left(\frac{1}{\lambda}\right)^{\frac{1}{2N-2}} < 1 + \frac{1}{2N-2} - \frac{1}{2N-2}\lambda.$$

And hence for $0 < \lambda \leq 1$ we have $f(\lambda) \coloneqq \left(\frac{1}{\lambda}\right)^{\frac{1}{2N-2}} + \frac{1}{2N-2}\lambda < 1 + 1/(2N-2)$. Since for $0 < \lambda < 1$ we also have

$$f'(\lambda) = \frac{1}{2N-2} - \frac{1}{2N-2}\left(\frac{1}{\lambda}\right)^{\frac{1}{2N-2}+1} < 0,$$

so $f(\lambda) \geq f(1) = 1 + 1/(2N-2)$. This gives us the desired contradiction and concludes our proof for the first part.

We will now prove the second part. Similarly to the first part, we just need to show that for all $t \geq 0$ we have $x_t \geq x^\star$. Suppose that $x^\star \leq x_t \leq \frac{3}{2}x^\star$ and hence we can write $x_t = x^\star(1+\gamma)$ for some $\gamma \in [0, \frac{1}{2}]$. Then we have

$$\begin{aligned}
x_{t+1} &= (1+\gamma)x^\star(1 - 2N\eta\gamma x^\star x_t^{(N-2)/N})^N \\
&\geq (1+\gamma)x^\star(1 - 3N\eta\gamma (x^\star)^{(N-2)/N})^N \\
&\geq x^\star(1+\gamma)\left(1 - \frac{1}{2N}\gamma\right)^N \\
&\geq x^\star.
\end{aligned}$$

The last inequality is obtained by letting $f(\gamma) \coloneqq (1+\gamma)\left(1 - \frac{1}{2N}\gamma\right)^N$, we could get that

$$\begin{aligned}
f'(\gamma) &= \left(1 - \frac{1}{2N}\gamma\right)^N - \frac{1}{2}(1+\gamma)\left(1 - \frac{1}{2N}\gamma\right)^{N-1} \\
&= \left(1 - \frac{1}{2N}\gamma\right)^{N-1}\left(\frac{1}{2} - \frac{1}{2}\gamma\right) > 0.
\end{aligned}$$

Hence, $f(\gamma) \geq f(0) = 1$ when $\gamma \in [0, \frac{1}{2}]$, which finishes the second part of our proof. □

Lemma 10 (Iterates behaviour near convergence)**.** *Consider the same setting as before. Let* $x^\star > 0$

and suppose that $|x_0 - x^\star| \leq \frac{1}{2}x^\star$. *Then the following holds.*

1. *If* $x_0 \leq x^\star$ *and* $\eta \leq \frac{1}{2N(2N-2)(x^\star)^{(2N-2)/N}}$, *then for any* $t \geq \frac{2}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}}$ *we have*

$$0 \leq x^\star - x_t \leq \frac{1}{2}|x_0 - x^\star|.$$

2. *If* $x^\star \leq x_0 \leq \frac{3}{2}x^\star$ *and* $\eta \leq \frac{1}{6N^2(x^\star)^{(2N-2)/N}}$ *then for any* $t \geq \frac{1}{2N^2\eta(x^\star)^{(2N-2)/N}}$ *we have*

$$0 \leq x_t - x^\star \leq \frac{1}{2}|x_0 - x^\star|.$$

Proof. Let us write $|x_0 - x^\star| = \gamma x^\star$ where $\gamma \in [0, \frac{1}{2}]$.

For the first part, we have $x_0 = (1-\gamma)x^\star$, we want to know how many steps t are needed to halve the error, i.e.,

$$x_t(1 - 2N\eta(x_t - x^\star)x_t^{\frac{N-2}{N}}))^N \geq (1 - \frac{\gamma}{2})x^\star.$$

We have that

$$\begin{aligned} x_t(1 - 2N\eta(x_t - x^\star)x_t^{\frac{N-2}{N}}))^N &\geq x_t(1 + 2N\eta\frac{\gamma}{2}x^\star((1-\gamma)x^\star)^{\frac{N-2}{N}}))^N \\ &\geq x_0(1 + N\eta\gamma(1-\gamma)^{\frac{N-2}{N}}(x^\star)^{\frac{2N-2}{N}}))^{Nt} \end{aligned}$$

It is enough to have

$$\begin{aligned} & x_0(1 + N\eta\gamma(1-\gamma)^{\frac{N-2}{N}}(x^\star)^{\frac{2N-2}{N}}))^{Nt} \geq (1 - \frac{\gamma}{2})x^\star \\ \Rightarrow & (1-\gamma)(1 + tN^2\eta\gamma(1-\gamma)^{\frac{N-2}{N}}(x^\star)^{\frac{2N-2}{N}})) \geq (1 - \frac{\gamma}{2}) \\ \Rightarrow & t \geq \left(\frac{1 - \frac{\gamma}{2}}{1-\gamma} - 1\right) \frac{1}{N^2\eta\gamma(1-\gamma)^{\frac{N-2}{N}}(x^\star)^{\frac{2N-2}{N}}} \\ \Rightarrow & t \geq \frac{1}{2(1-\gamma)^{\frac{2N-2}{N}}N^2\eta(x^\star)^{\frac{2N-2}{N}}} \\ \Rightarrow & t \geq \frac{2}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \end{aligned}$$

The last step is by $\gamma \in [0, \frac{1}{2}]$, we could obtain that $\frac{1}{2(1-\gamma)^{\frac{2N-2}{N}}} \leq \frac{1}{2(1/2)^{\frac{2N-2}{N}}} \leq \frac{1}{2(1/2)^2} \leq 2$. Therefore after $t \geq \frac{2}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}}$, the error is halved.

To deal with the second part, we write $x_0 = x^\star(1+\gamma)$. We will use a similar approach as the one in the first part. If for some x_t we have $x_t \leq (1+\gamma/2)x^\star$ we would be done. If $x_t > x^\star(1+\gamma/2)$ we have $x_{t+1} \leq x_t(1 - 2N\eta\frac{\gamma}{2}x^\star(x^\star)^{(N-2)/N})^N$. Therefore,

$$
\begin{aligned}
& x_0(1 - 2N\eta\frac{\gamma}{2}x^\star(x^\star)^{(N-2)/N})^{Nt} \leq x^\star(1+\gamma/2) \\
\iff & Nt\log(1 - N\eta\gamma(x^\star)^{(2N-2)/N}) \leq \log\frac{x^\star(1+\gamma/2)}{x_0} \\
\iff & t \geq \frac{1}{N}\frac{\log\frac{x^\star(1+\gamma/2)}{x_0}}{\log(1 - N\eta\gamma(x^\star)^{(2N-2)/N})}.
\end{aligned}
$$

We can deal with the term on the right hand side by noting that

$$
\begin{aligned}
\frac{1}{N}\frac{\log\frac{x^\star(1+\gamma/2)}{x_0}}{\log(1 - N\eta\gamma(x^\star)^{(2N-2)/N})} &= \frac{1}{N}\frac{\log\frac{1+\gamma/2}{1+\gamma}}{\log(1 - N\eta\gamma(x^\star)^{(2N-2)/N})} \\
&\leq \frac{1}{N}\frac{\left(\frac{1+\gamma/2}{1+\gamma} - 1\right) / \left(\frac{1+\gamma/2}{1+\gamma}\right)}{-N\eta\gamma(x^\star)^{(2N-2)/N}} \\
&= \frac{1}{N}\frac{-\frac{\gamma}{2}/(1+\frac{\gamma}{2})}{-N\eta\gamma(x^\star)^{(2N-2)/N}} \\
&\leq \frac{1}{2N^2\eta(x^\star)^{(2N-2)/N}}
\end{aligned}
$$

where the second line used $\log x \leq x - 1$ and $\log x \geq \frac{x-1}{x}$. Note that both logarithms are negative.

□

Lemma 11 (Iterates at the beginning)**.** *Consider the same setting as before. If* $0 < x_0 \leq \frac{1}{2}x^\star$ *and* $\eta \leq \frac{x_0}{2N(2N-4)(x^\star)^{(3N-2)/N}}$*, for any* $t \geq \frac{3}{2N(N-2)\eta x^\star x_0^{(N-2)/N}}$*, we will have* $\frac{1}{2}x^\star \leq x_t \leq x^\star$*.*

Proof. We need to find a lower-bound on time T which ensures that $x_T \geq \frac{x^\star}{2}$. At any time t, we

have

$$x_{t+1} = x_t(1 - 2N\eta(x_t - x^\star)x_t^{(N-2)/N})^N \geq x_t(1 - 2N^2\eta(x_t - x^\star)x_t^{(N-2)/N}).$$

$$x_{t+1} - x_t \geq -2N^2\eta(x_t - x^\star)x_t^{(2N-2)/N}$$

$$\frac{x_{t+1} - x_t}{2N^2\eta(x^\star - x_t)x_t^{(2N-2)/N}} \geq 1$$

$$\sum_{t=0}^{T-1} \frac{x_{t+1} - x_t}{2N^2\eta(x^\star - x_t)x_t^{(2N-2)/N}} \geq \sum_{t=0}^{T-1} 1 = T.$$

Therefore, for t that is larger than the left hand side, we have $x_t \geq \frac{1}{2}x^\star$.

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{x_{t+1} - x_t}{2N^2\eta(x^\star - x_t)x_t^{(2N-2)/N}} &\leq \frac{1}{N^2\eta x^\star} \sum_{t=0}^{T-1} \frac{x_{t+1} - x_t}{x_t^{(2N-2)/N}} \\
&= \frac{1}{N^2\eta x^\star} \sum_{t=0}^{T-1} \int_{x_t}^{x_{t+1}} \frac{1}{x^{(2N-2)/N}} + \left(\frac{1}{x_t^{(2N-2)/N}} - \frac{1}{x^{(2N-2)/N}}\right) dx \\
&\leq \frac{1}{N^2\eta x^\star} \sum_{t=0}^{T-1} \int_{x_t}^{x_{t+1}} \frac{1}{x^{(2N-2)/N}} dx \\
&+ \frac{1}{N^2\eta x^\star} \max_{0\leq t\leq T-1} \left(\frac{1}{x_t^{(2N-2)/N}} - \frac{1}{x_{t+1}^{(2N-2)/N}}\right)(x_T - x_0) \\
&\leq \frac{1}{N^2\eta x^\star} \int_{x_0}^{\frac{1}{2}x^\star} \frac{1}{x^{(2N-2)/N}} dx \qquad \text{(A.8)} \\
&+ \frac{1}{N^2\eta x^\star} \max_{0\leq t\leq T-1} \left(\frac{1}{x_t^{(2N-2)/N}} - \frac{1}{x_{t+1}^{(2N-2)/N}}\right)\left(\frac{1}{2}x^\star - x_0\right) \\
& \qquad \text{(A.9)} \\
&+ \frac{1}{N^2\eta x^\star} \frac{1}{(\frac{1}{2}x^\star)^{(2N-2)/N}} \left(x_T - \frac{1}{2}x^\star\right) \qquad \text{(A.10)}
\end{aligned}$$

For equation (A.8),

$$\begin{aligned}\frac{1}{N^2\eta x^\star}\int_{x_0}^{\frac{1}{2}x^\star}\frac{1}{x^{(2N-2)/N}}dx &\leq \frac{1}{N^2\eta x^\star}\left(-\frac{N}{N-2}\frac{1}{x^{(N-2)/N}}\Big|_{x_0}^{\frac{1}{2}x^\star}\right)\\ &= \frac{1}{N^2\eta x^\star}\left(-\frac{N}{N-2}\frac{1}{(\frac{1}{2}x^\star)^{(N-2)/N}}+-\frac{N}{N-2}\frac{1}{x_0^{(N-2)/N}}\right)\\ &= \frac{1}{N(N-2)\eta x^\star}\left(\frac{1}{x_0^{(N-2)/N}}-\frac{2^{(N-2)/N}}{(x^\star)^{(N-2)/N}}\right).\end{aligned}$$

For equation (A.9), we first focus on

$$\frac{1}{x_t^{(2N-2)/N}}-\frac{1}{x_{t+1}^{(2N-2)/N}}.$$

We have that

$$\begin{aligned}x_{t+1} &= x_t(1-2N\eta(x_t-x^\star)x_t^{(N-2)/N})^N,\\ \Rightarrow x_{t+1}^{(2N-2)/N} &= x_t^{(2N-2)/N}(1-2N\eta(x_t-x^\star)x_t^{(N-2)/N})^{2N-2}.\end{aligned}$$

To deal with the multiplicative coefficient, with $\eta \leq \frac{1}{2N(2N-3)(x^\star)^{(2N-2)/N}}$ using the inequality $(1+x)^r \leq 1+\frac{rx}{1-(r-1)x}$ where $x \in (0, \frac{1}{r-1})$, we obtain that

$$\begin{aligned}(1-2N\eta(x_t-x^\star)x_t^{(N-2)/N})^{2N-2} &\leq (1+2N\eta(x^\star)^{(2N-2)/N})^{(2N-2)}\\ &\leq 1+\frac{2N(2N-2)\eta(x^\star)^{(2N-2)/N}}{1-2N(2N-3)\eta(x^\star)^{(2N-2)/N}}\\ &= \frac{1-2N\eta(x^\star)^{(2N-2)/N}}{1-2N(2N-3)\eta(x^\star)^{(2N-2)/N}}.\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{x_t^{(2N-2)/N}} - \frac{1}{x_{t+1}^{(2N-2)/N}} &= \frac{1}{x_t^{(2N-2)/N}} - \frac{1}{x_t^{(2N-2)/N}(1-2N\eta(x_t-x^\star)x_t^{(N-2)/N})^{2N-2}} \\
&= \frac{1}{x_t^{(2N-2)/N}}\left(1 - \frac{1}{(1-2N\eta(x_t-x^\star)x_t^{(N-2)/N})^{2N-2}}\right) \\
&\leq \frac{1}{x_t^{(2N-2)/N}}\left(1 - \frac{1-2N(2N-3)\eta(x^\star)^{(2N-2)/N}}{1-2N\eta(x^\star)^{(2N-2)/N}}\right) \\
&\leq \frac{1}{x_t^{(2N-2)/N}}\frac{2N(2N-4)\eta(x^\star)^{(2N-2)/N}}{1-2N\eta(x^\star)^{(2N-2)/N}} \\
&\leq \frac{1}{x_t^{(2N-2)/N}}2N(2N-4)\eta(x^\star)^{(2N-2)/N} \\
&\leq \frac{1}{x_0^{(2N-2)/N}}2N(2N-4)\eta(x^\star)^{(2N-2)/N}.
\end{aligned}$$

If we further require the step size satisfies $\eta \leq \frac{x_0}{2N(2N-4)(x^\star)^{(3N-2)/N}}$, we have for equation (A.9),

$$\begin{aligned}
\frac{1}{N^2\eta x^\star}\max_{0\leq t\leq T-1}\left(\frac{1}{x_t^{(2N-2)/N}} - \frac{1}{x_{t+1}^{(2N-2)/N}}\right)\left(\frac{1}{2}x^\star - x_0\right) &\leq \frac{1}{N^2\eta x^\star}\frac{1}{x_0^{(N-2)/N}x^\star}\left(\frac{1}{2}x^\star - x_0\right) \\
&\leq \frac{1}{2N^2\eta x^\star}\frac{1}{x_0^{(N-2)/N}},
\end{aligned}$$

which is with the same order with the result of equation (A.8).

Combining the results from equations (A.8), (A.9), (A.10), we obtain that

$$\begin{aligned}
T &\leq \frac{1}{N(N-2)\eta x^\star}\left(\frac{1}{x_0^{(N-2)/N}} - \frac{2^{(N-2)/N}}{(x^\star)^{(N-2)/N}}\right) + \frac{1}{2N^2\eta x^\star}\frac{1}{x_0^{(N-2)/N}} \\
&\quad + \frac{1}{N^2\eta x^\star}\frac{1}{(\frac{1}{2}x^\star)^{(2N-2)/N}}\left(x_T - \frac{1}{2}x^\star\right) \\
&\leq \frac{1}{N(N-2)\eta x^\star}\left(\frac{1}{x_0^{(N-2)/N}} - \frac{2^{(N-2)/N}}{(x^\star)^{(N-2)/N}} + \frac{1}{2x_0^{(N-2)/N}} + \frac{1}{(\frac{1}{2}x^\star)^{(N-2)/N}}\right) \\
&\leq \frac{3}{2N(N-2)\eta x^\star x_0^{(N-2)/N}}.
\end{aligned}$$

□

Lemma 12 (Overall iterates)**.** *Consider the same setting as before. Fix any* $\epsilon > 0$.

1. *If* $\epsilon < |x^\star - x_0| \leq \frac{1}{2}x^\star$ *and* $\eta \leq \frac{1}{6N^2(x^\star)^{(2N-2)/N}}$ *then for any* $t \geq \frac{3}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon}$ *we have*

$$|x^\star - x_t| \leq \epsilon.$$

2. *If* $0 < x_0 \leq \frac{1}{2}x^\star$ *and* $\eta \leq \frac{x_0}{2N(2N-4)(x^\star)^{(3N-2)/N}}$ *then for any*

$$t \geq \frac{3}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon} + \frac{3}{2N(N-2)\eta x^\star x_0^{(N-2)/N}}$$

we have

$$x^\star - \epsilon \leq x_t \leq x^\star.$$

Proof. 1. To prove the first part we simply need apply Lemma 10 $\lceil \log_2 \frac{|x^\star - x_0|}{\epsilon} \rceil$ times. Hence after

$$\frac{2\log_2 e}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon} \leq \frac{3}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon}$$

iterations we are done.

2. For the second part, we simply combine the results from the first part and Lemma 11, it is enough to choose t larger than or equal to

$$\frac{3}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon} + \frac{3}{2N(N-2)\eta x^\star x_0^{(N-2)/N}}.$$

□

A.2.2.2 Dealing with Bounded Errors $\mathbf{b}_t$

In this subsection we extend the previous setting to handle bounded error sequences $(\mathbf{b}_t)_{t\geq 0}$ such that for any $t \geq 0$ we have $\|\mathbf{b}_t\|_\infty \leq B$ for some $B \in \mathbb{R}$. That is, we look at the following updates

$$x_{t+1} = x_t(1 - 2N\eta(x_t - x^\star + b_t)x_t^{(N-2)/N})^N.$$

Surely, if $B \geq x^\star$, the convergence to $x^\star$ is not possible. Hence, we will require B to be small enough, with a particular choice $B \leq \frac{1}{5}x^\star$. For a given B, we can only expect the sequence $(x_t)_{t\geq 0}$ to converge to $x^\star$ up to precision B. We would consider two extreme scenarios,

$$x_{t+1}^+ = x_t^+(1 - 2N\eta(x_t^+ - (x^\star - B))(x_t^+)^{(N-2)/N})^N,$$
$$x_{t+1}^- = x_t^-(1 - 2N\eta(x_t^- - (x^\star + B))(x_t^-)^{(N-2)/N})^N.$$

Lemma 13 (Squeezing iterates with bounded errors)**.** *Consider the sequences* $(x_t^-)_{t\geq 0}, (x_t)_{t\geq 0}$ *and* $(x_t^+)_{t\geq 0}$ *as defined above with*

$$0 < x_0^- = x_0^+ = x_0 \leq x^\star + B$$

If $\eta \leq \frac{1}{8N^2(x^\star)^{(2N-2)/N}}$ *then for all* $t \geq 0$

$$0 \leq x_t^- \leq x_t \leq x_t^+ \leq x^\star + B.$$

Proof. We will prove the claim by induction. The claim holds trivially for $t = 0$. If $x_t^+ \geq x_t$, we

have

$$
\begin{aligned}
x_{t+1}^{+} &= x_t^{+}(1-2N\eta(x_t^{+}-(x^\star+B))(x_t^{+})^{\frac{N-2}{N}})^N \\
&\geq x_t^{+}(1-2N\eta(x_t^{+}-(x^\star+B))x_t^{\frac{N-2}{N}})^N \\
(\triangle = x_t^{+}-x_t) \qquad &= (x_t+\triangle)(1-2N\eta(x_t-x^\star+b_t)x_t^{\frac{N-2}{N}} \\
&\quad + 2N\eta(x_t^{+}-x_t-B-b_t)x_t^{\frac{N-2}{N}})^N \\
(m_t = 1-2N\eta(x_t-x^\star+b_t)x_t^{\frac{N-2}{N}}) \qquad &\geq (x_t+\triangle)(m_t-2N\eta\triangle x_t^{\frac{N-2}{N}})^N \\
&\geq (x_t+\triangle)(m_t-2N\eta\triangle x_t^{\frac{N-2}{N}})^N \\
&= x_t m_t^N + (x_t+\triangle)(m_t-2N\eta\triangle x_t^{\frac{N-2}{N}})^N - x_t m_t^N \\
&= x_t m_t^N + (x_t+\triangle)m_t^N\left(1-\frac{2N\eta\triangle x_t^{\frac{N-2}{N}}}{m_t}\right)^N - x_t m_t^N.
\end{aligned}
$$

We aimed to show that $(x_t+\triangle)m_t^N(1-2N\eta\triangle x_t^{\frac{N-2}{N}}/m_t)^N - x_t m_t^N$ is positive. With $\eta \leq \frac{1}{4N(x^\star+B)(x^\star)^{(N-2)/N}}$, we can see $m_t \geq 1/2$ for all t and

$$
\begin{aligned}
(x_t+\triangle)m_t^N(1-2N\eta\triangle x_t^{\frac{N-2}{N}}/m_t)^N - x_t m_t^N &\geq (x_t+\triangle)m_t^N(1-4N\eta\triangle x_t^{\frac{N-2}{N}})^N - x_t m_t^N \\
&\geq (x_t+\triangle)m_t^N(1-4N^2\eta\triangle x_t^{\frac{N-2}{N}}) - x_t m_t^N.
\end{aligned}
$$

The last inequality is obtained via $(1-x)^n \geq 1-nx$. If we further require $\eta \leq \frac{1}{8N^2(x^\star)^{(2N-2)/N}}$, we obtain that

$$
\begin{aligned}
(x_t+\triangle)m_t^N(1-4N^2\eta\triangle x_t^{\frac{N-2}{N}}) - x_t m_t^N &\geq (x_t+\triangle)m_t^N\left(1-\frac{1}{2x^\star}\triangle\right) - x_t m_t^N \\
&\geq m_t^N\left(x_t+\triangle-\frac{x_t}{2x^\star}\triangle-\frac{1}{2x^\star}\triangle^2-x_t\right) \\
&\geq m_t^N\triangle\left(1-\frac{x_t}{2x^\star}-\frac{\triangle}{2x^\star}\right) \\
&\geq m_t^N\triangle\left(1-\frac{1}{2}-\frac{1}{2}\right) \geq 0.
\end{aligned}
$$

Therefore, we obtain that

$$x_{t+1}^{+} \geq x_t m_t^N = x_{t+1}.$$

For x_t^{-}, it follows a similar proof. □

Lemma 14 (Iterates with bounded errors monotonic behaviour)**.** *Consider the previous setting with* $B \leq \frac{1}{5}x^\star$, $\eta \leq \frac{1}{6N^2(x^\star)^{(2N-2)/N}}$. *Then the following holds*

1. *If* $|x_t - x^\star| > B$ *then* $|x_{t+1} - x^\star| < |x_t - x^\star|$.

2. *If* $|x_t - x^\star| \leq B$ *then* $|x_{t+1} - x^\star| \leq B$.

Proof. The choice of B and step size η ensures us to apply Lemma 9 and Lemma 13 to the sequences $(x_t^{-})_{t\geq 0}$ and $(x_t^{+})_{t\geq 0}$. □

Lemma 15 (Iterates with B near convergence)**.** *Consider the setting as before. Then the following holds:*

1. *If* $\frac{1}{2}(x^\star - B) \leq x_0 \leq x^\star - 5B$ *then for any* $t \geq \frac{2}{\eta N^2(x^\star)^{\frac{2N-2}{N}}}$ *we have*

$$|x^\star - x_t| \leq \frac{1}{2}|x_0 - x^\star|.$$

2. *If* $x^\star + 4B < x_0 < \frac{6}{5}x^\star$ *then for any* $t \geq \frac{4}{\eta N^2(x^\star)^{\frac{2N-2}{N}}}$ *we have*

$$|x^\star - x_t| \leq \frac{1}{2}|x_0 - x^\star|.$$

Proof. 1. To prove the first part, let us first apply Lemma 10 on x_t^{-} twice, therefore for all

$$t \geq \frac{25}{4\eta N^2(x^\star)^{\frac{2N-2}{N}}} \geq 2\frac{2}{\eta N^2(x^\star - B)^{\frac{2N-2}{N}}}$$

we have

$$\begin{aligned} 0 &\leq (x^\star - B) - x_t^- \\ &\leq \frac{1}{4}|x_0 - (x^\star - B)| \\ &\leq \frac{1}{4}|x_0 - x^\star| + \frac{1}{4}B. \end{aligned}$$

When $x_t \leq x^\star$, from Lemma 13 we have

$$\begin{aligned} 0 &\leq x^\star - x_t \\ &\leq x^\star - x_t^- \\ &\frac{1}{4}|x_0 - x^\star| + \frac{5}{4}B \\ &\leq \frac{1}{2}|x_0 - x^\star|. \end{aligned}$$

When $x_t \geq x^\star$ then by Lemma 13 we have

$$0 \leq x_t - x^\star \leq B \leq \frac{1}{5}|x_0 - x^\star|,$$

where both last inequalities are from $x_0 \leq x^\star - 5B$.

2. The second part follows a very similar proof for x_t^+, the number of iterations would be

$$t \geq \frac{4}{\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \geq 2\frac{2}{\eta N^2 (x^\star + B)^{\frac{2N-2}{N}}}.$$

□

Lemma 16 (Overall iterates with B)**.** *Consider the same setting as before. Fix any $\epsilon > 0$, then the following holds*

1. *If $B + \epsilon < |x^\star - x_0| \leq \frac{1}{5}x^\star$ then for any $t \geq \frac{15}{4\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon}$ iterations we have $|x^\star - x_t| \leq B + \epsilon$.*

2. *If* $0 < x_0 \leq x^\star - B - \epsilon$ *then for any*

$$t \geq \frac{75}{16\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon} + \frac{15}{8N(N-2)\eta x^\star x_0^{(N-2)/N}}$$

we have $x^\star - B - \epsilon \leq x_t \leq x^\star + B$.

Proof. 1. If $x_0 > x^\star + B$ then by Lemma 13 and Lemma 14 we only need to show that $(x_t^+)_{t\geq 0}$ hits $x^\star + B + \epsilon$ within the desired number of iterations. From the first part of Lemma 12, we see that

$$\frac{3}{\eta N^2 (x^\star + B)^{\frac{2N-2}{N}}} \log \frac{|x^\star + B - x_0|}{\epsilon} \leq \frac{15}{4\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon}$$

iterations are enough, where we require $\frac{|x^\star - x_0|}{\epsilon} \geq \frac{5}{2}$.

2. The upper bound is obtained immediately from Lemma 13. For lower bound, we simply apply the second part of Lemma 12 to the sequence $(x_t^-)_{t\geq 0}$ to get

$$\begin{aligned} t &\geq \frac{75}{16\eta N^2 (x^\star)^{\frac{2N-2}{N}}} \log \frac{|x^\star - x_0|}{\epsilon} + \frac{15}{8N(N-2)\eta x^\star x_0^{(N-2)/N}} \\ &\geq \frac{3}{\eta N^2 (x^\star - B)^{\frac{2N-2}{N}}} \log \frac{|x^\star - B - x_0|}{\epsilon} + \frac{3}{2N(N-2)\eta (x^\star - B) x_0^{(N-2)/N}} \end{aligned}$$

to ensure the results we wanted.

□

Lemma 17. *Suppose the error sequences* $(\mathbf{b}_t)_{t\geq 0}$ *and* $(\mathbf{p}_t)_{t\geq 0}$ *satisfy the following for any* $t \geq 0$*:*

$$\begin{aligned} \|\mathbf{b}_t \odot \mathbf{1}_S\| &\leq B, \\ \|\mathbf{p}_t\|_\infty &\leq \frac{1}{20} \|\mathbf{s}_t - \mathbf{w}^\star\|_\infty . \end{aligned}$$

Suppose that

$$20B < \|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty \leq \frac{1}{5} w^\star_{\min}.$$

Then for $\eta \leq \frac{1}{6N^2 (w^\star_{\max})^{(2N-2)/N}}$ *and any* $t \geq \frac{2}{\eta N^2 (w^\star_{\max})^{(2N-2)/N}}$ *we have*

$$\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty \leq \frac{1}{2} \|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty .$$

Proof. Note that $\|\mathbf{b}_0\|_\infty + \|\mathbf{p}_t\|_\infty \leq \frac{1}{10} \|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty$. For any i such that $|s_{0,i} - w^\star_i| \leq \frac{1}{2} \|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty$, Lemma 14 guarantees that for any $t \geq 0$ we have $|s_{t,i} - w^\star_i| \leq \frac{1}{2} \|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty$. On the other hand, for any i such that $|s_{0,i} - w^\star_i| > \frac{1}{2} \|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty$ by Lemma 15 we have $|s_{0,i} - w^\star_i| \leq \frac{1}{2} \|\mathbf{s}_0 - \mathbf{w}^\star\|_\infty$ for any $t \geq \frac{2}{\eta N^2 (w^\star_{\max})^{(2N-2)/N}}$ which concludes the proof. □

A.2.3 Dealing with Negative Targets

Lemma 18. *Let* $x_t = u^N - v^N$ *and* $x^\star \in \mathbb{R}$ *be the target such that* $|x^\star| > 0$*. Suppose the sequences* $(u_t)_{t \geq 0}$ *and* $(v_t)_{t \geq 0}$ *evolve as follows*

$$0 < u_0 = \alpha, \quad u_{t+1} = u_t (1 - 2N\eta (x_t - x^\star + b_t) u_t^{N-2}),$$
$$0 < v_0 = \alpha, \quad v_{t+1} = v_t (1 + 2N\eta (x_t - x^\star + b_t) v_t^{N-2}),$$

where $\alpha \leq (2 - 2^{\frac{N-2}{N}})^{\frac{1}{N-2}} |x^\star|^{1/N}$ *and there exists* $B > 0$ *such that* $|b_t| \leq B$ *and* $\eta \leq \frac{\alpha}{4N(N-2)(x^\star + B)x^\star}$*. Then the following holds: For any* $t \geq 0$ *we have*

- *If* $x^\star > 0$ *and* $u_t^N \geq x^\star$*, then* $v_t^N \leq \frac{1}{2}\alpha^N$*.*
- *If* $x^\star < 0$ *and* $v_t^N \geq |x^\star|$*, then* $u_t^N \leq \frac{1}{2}\alpha^N$*.*

Proof. Let us assume $x^\star > 0$ first and prove the first statement. From the updating equation, we obtain that

$$\frac{u_{t+1} - u_t}{u_t^{N-1}} = -2N\eta (x_t - x^\star + b_t).$$

Therefore,

$$\begin{aligned}\sum_{i=0}^{t} -2N\eta(x_i - x^\star + b_i) &= \sum_{i=0}^{t} \frac{u_{i+1} - u_i}{u_i^{N-1}} \\ &\geq \sum_{i=0}^{t} \int_{u_i}^{u_{i+1}} \frac{1}{u^{N-1}} du \\ &= \int_{u_0}^{u_t} \frac{1}{u^{N-1}} du \\ &= (2-N)(u_t^{2-N} - u_0^{2-N}).\end{aligned}$$

When $u_t^N \geq x^\star$, we have that $u_t^{2-N} \leq (x^\star)^{(2-N)/N}$. Therefore,

$$\sum_{i=1}^{t} -2N\eta(x_i - x^\star + b_i) \geq (2-N)(u_t^{2-N} - u_0^{2-N}).$$

Similarly for v_t, we have

$$\begin{aligned}\sum_{i=1}^{t} 2N\eta(x_i - x^\star + b_i) &= \sum_{i=0}^{t} \frac{v_{i+1} - v_i}{v_i^{N-1}} \\ &\geq (2-N)(v_t^{2-N} - v_0^{2-N}).\end{aligned}$$

Therefore, we have that

$$
\begin{aligned}
&(N-2)((x^\star)^{\frac{2-N}{N}} - \alpha^{2-N}) \geq (2-N)(v_t^{2-N} - \alpha^{2-N}).\\
&\implies (\alpha^{2-N} - (x^\star)^{\frac{2-N}{N}}) + \alpha^{2-N} \leq v_t^{2-N}.\\
&\implies v_t \leq \left(\frac{1}{2\alpha^{2-N} - (x^\star)^{\frac{2-N}{N}}}\right)^{\frac{1}{N-2}}\\
&\implies v_t \leq \left(\frac{1}{2 - \alpha^{N-2}/(x^\star)^{\frac{N-2}{N}}}\right)^{\frac{1}{N-2}} \alpha\\
&\implies v_t \leq \left(\frac{1}{2 - (2 - 2^{\frac{N-2}{N}})}\right)^{\frac{1}{N-2}} \alpha\\
&\implies v_t \leq 2^{\frac{1}{N}}\alpha.
\end{aligned}
$$

For $x^\star < 0$, we obtain a similar result by symmetry. □

Lemma 19. *Let $x_t = x_t^+ - x_t^-$ and $x^\star \in \mathbb{R}$ be the target such that $|x^\star| > 0$. Suppose the sequences $(x_t^+)_{t\geq 0}$ and $(x_t^-)_{t\geq 0}$ evolve as follows*

$$
\begin{aligned}
&0 < x_0^+ = \alpha^N \leq 2^3(2^{\frac{1}{N}} - 1)^{\frac{N}{N-2}}|x^\star|, \quad x_{t+1}^+ = x_t^+(1 - 2N\eta(x_t - x^\star + b_t)(x_t^+)^{(N-2)/N})^N,\\
&0 < x_0^- = \alpha^N \leq 2^3(2^{\frac{1}{N}} - 1)^{\frac{N}{N-2}}|x^\star|, \quad x_{t+1}^- = x_t^-(1 + 2N\eta(x_t - x^\star + b_t)(x_t^-)^{(N-2)/N})^N,
\end{aligned}
$$

and that there exists $B > 0$ such that $|b_t| \leq B$ and $\eta \leq \frac{1}{8N(x^\star+B)(x^\star)^{(N-2)/N}}$. Then the following holds: For any $t \geq 0$ we have

- *If $x^\star > 0$ then $x_t^- \leq \alpha^N \Pi_{i=0}^{t-1}(1 + 2N\eta|b_t|(x_i^-)^{(N-2)/N})^N$.*
- *If $x^\star < 0$ then $x_t^+ \leq \alpha^N \Pi_{i=0}^{t-1}(1 + 2N\eta|b_t|(x_i^+)^{(N-2)/N})^N$.*

Proof. Assume $x^\star > 0$ and fix any $t \geq 0$. Let $0 \leq s \leq t$ be the largest s such that $x_s^+ > x^\star$. If no such s exists we are done immediately. If $s = t$ then by the first part we have $x_t^- \leq \alpha^N$ and we are done.

If $s < t$, by Lemma 18, we have $x_s^- \leq \frac{1}{2}\alpha^N$. From the requirement of initialization, we have

$$(1 + 2N\eta(x_s^+ - x_s^- - x^\star + b_s)(x_s^-)^{(N-2)/N})^N \leq \left(1 + \frac{(\frac{1}{2}\alpha^N)^{\frac{N-2}{N}}}{4(x^\star)^{\frac{N-2}{N}}}\right)^N \leq (1 + 2^{\frac{1}{N}} - 1)^N = 2.$$

Therefore

$$\begin{aligned} x_t^- &= x_s^- \prod_{i=s}^{t-1}(1 + 2N\eta(x_i^+ - x_i^- - x^\star + b_i)(x_i^-)^{(N-2)/N})^N \\ &= \frac{1}{2}\alpha^N \cdot 2 \prod_{i=s+1}^{t-1}(1 + 2N\eta(x_i^+ - x_i^- - x^\star + b_i)(x_i^-)^{(N-2)/N})^N \\ &\leq \alpha^N \prod_{i=s+1}^{t-1}(1 + 2N\eta|b_i|(x_i^-)^{(N-2)/N})^N. \end{aligned}$$

This completes the proof for $x^\star > 0$. It follows a similar proof for the case $x^\star < 0$. □

A.3 Proof of Propositions and Technical Lemmas

In this section, we provide the proof for the propositions and technical lemmas mentioned in Appendix A.1.

A.3.1 Proof of Proposition 1

By the assumptions on $(\mathbf{b}_t)_{t\geq 0}$ and $(\mathbf{p}_t)_{t\geq 0}$, we obtain that

$$\begin{aligned} &\|\mathbf{b}_t\|_\infty \leq C_b\zeta - \alpha^{N/4}, \\ &\|\mathbf{p}_t\|_\infty \leq \frac{C_\gamma}{w_{\max}^\star/\zeta}\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty \leq \frac{C_\gamma}{w_{\max}^\star/\zeta} w_{\max}^\star \leq C_\gamma\zeta. \end{aligned}$$

Choose C_b and C_γ such that $C_b + C_\gamma \leq 1/40$. Therefore, we have

$$B \leq \|\mathbf{b}_t\|_\infty + \|\mathbf{p}_t\|_\infty + \alpha^{N/4} \leq (C_b + C_\gamma)\zeta \leq \frac{1}{40}\zeta.$$

For any j such that $w_j^\star \geq \frac{1}{2}\zeta$, we have that $B \leq \frac{1}{20} w_j^\star$. Therefore, by applying Lemma 16, we know when

$$t \geq \frac{75}{16\eta N^2 \zeta^{(2N-2)/N}} \log \frac{|w_{\max}^\star - \alpha^N|}{\epsilon} + \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}} = T_1,$$

we have $|w_{j,t} - w_j^\star| \leq \zeta$.

On the other hand, for any j such that $w_j^\star \leq \frac{1}{2}\zeta$, $w_{j,t}$ will stay in $(0, w_j^\star + \frac{1}{40}\zeta]$ maintaining $|w_{j,t} - w_j^\star| \leq \zeta$ as required.

By Lemma 8, we have that $\|\mathbf{e}_t\|_\infty \leq \alpha^{N/2}$ up to

$$T_2 = \frac{5}{N(N-1)\eta\zeta} \left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}} \right).$$

From our choice of initialization α, we can see that $T_1 \leq T_2$ is ensured. To see this,

$$\begin{aligned}
&\alpha \leq \left(\frac{1}{8}\right)^{2/(N-2)} \wedge \left(\frac{\zeta^{(N-2)/N}}{\log \frac{w_{\max}^\star}{\epsilon}} \right)^{2/(N-2)} \\
\Longrightarrow &\alpha^{(N-2)/2} \leq \frac{1}{8} \wedge \frac{16\zeta^{(N-2)/N}}{15 \log \frac{w_{\max}^\star}{\epsilon}} \\
\Longrightarrow &4\alpha^{(N-2)/2} \left(\frac{15}{16} \alpha^{(N-2)/2} \log \frac{w_{\max}^\star}{\epsilon} + \zeta^{(N-2)/N} \right) \leq \zeta^{(N-2)/N} \\
\Longrightarrow &\frac{15}{2\zeta^{(N-2)/N}} \log \frac{w_{\max}^\star}{\epsilon} + \frac{6}{\alpha^{(N-2)}} \leq 8 \left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}} \right) \\
\Longrightarrow &\frac{75}{16\eta N^2 \zeta^{(2N-2)/N}} \log \frac{|w_{\max}^\star - \alpha^N|}{\epsilon} + \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}} \\
&\quad \leq \frac{5}{N(N-1)\eta\zeta} \left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}} \right) \\
\Longrightarrow &T_1 \leq T_2.
\end{aligned} \tag{A.11}$$

□

A.3.2 Proof of Proposition 2

By Lemma 8, with the choice of $B = \frac{1}{200} w^\star_{\min}$, we can maintain $\|\mathbf{e}_t\|_\infty \le \alpha^{N/4}$ for at least another

$$t \le \frac{25}{N(N-1)\eta w^\star_{\min}} \left(\frac{1}{\alpha^{(N-2)/2}} - \frac{1}{\alpha^{(N-2)/4}} \right) = T_4.$$

Now we consider to further reduce $\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty$ from $\frac{1}{5} w^\star_{\min}$ to $\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee \epsilon$. Let $B_i \coloneqq (\mathbf{b}_t)_i$ and $B \coloneqq \max_{j \in S} B_j$.

We first apply Lemma 17 for $\log_2 \frac{w^\star_{\min}}{100(B \vee \epsilon)}$ times, the total number of iterations for this step would be

$$\frac{2}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log_2 \frac{w^\star_{\min}}{100(B \vee \epsilon)}.$$

After that we have $\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty < 20(B \vee \epsilon)$ and so $\|\mathbf{p}_t\|_\infty < k\mu \cdot 20(B \vee \epsilon)$. Hence, for any $i \in S$ we have

$$\|\mathbf{b}_t \odot \mathbf{1}_i\|_\infty + \|\mathbf{p}_t\|_\infty \le B_i + k\mu 20(B \vee \epsilon).$$

Then we further apply Lemma 16 for each coordinate $i \in S$ to obtain that

$$|w_{i,t} - w^\star_i| \lesssim \left| \frac{1}{n} (\mathbf{X}^\mathsf{T}\boldsymbol{\xi})_i \right| \vee k\mu \left\| \frac{1}{n} \mathbf{X}^\mathsf{T}\boldsymbol{\xi} \odot \mathbf{1}_S \right\|_\infty \vee \epsilon.$$

the number of iterations needed for this step is $\frac{15}{4\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon}$.

Therefore the total number of iterations needed to further reduce $\|\mathbf{s}_t - \mathbf{w}^\star\|_\infty$ is

$$\begin{aligned} T_3 &= \frac{6}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon} \\ &\ge \frac{2}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log_2 \frac{w^\star_{\min}}{100(B \vee \epsilon)} + \frac{15}{4\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon}. \end{aligned}$$

Since T_3 is no longer related to α, we can easily ensure $T_3 \le T_4$ with some mild upper bound on $\alpha^{(N-2)/4} \le \frac{(w^\star_{\min})^{(N-2)/N}}{\log \frac{w^\star_{\min}}{\epsilon}} \wedge 1/2$.

$$
\begin{aligned}
&\alpha^{(N-2)/4} \leq \frac{(w^\star_{\min})^{(N-2)/N}}{\log \frac{w^\star_{\min}}{\epsilon}} \wedge 1/2 \\
\Longrightarrow &\alpha^{(N-2)/4}\left(\alpha^{(N-2)/4}\log\frac{w^\star_{\min}}{\epsilon} + (w^\star_{\min})^{(N-2)/N}\right) \leq (w^\star_{\min})^{(N-2)/N} \\
\Longrightarrow &\alpha^{(N-2)/2}\log\frac{w^\star_{\min}}{\epsilon} \leq (w^\star_{\min})^{(N-2)/N} - \alpha^{(N-2)/4}(w^\star_{\min})^{(N-2)/N} \\
\Longrightarrow &\frac{1}{(w^\star_{\min})^{(N-2)/N}}\log\frac{w^\star_{\min}}{\epsilon} \leq \frac{1}{\alpha^{(N-2)/2}} - \frac{1}{\alpha^{(N-2)/4}} \\
\Longrightarrow &\frac{6}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}}\log\frac{w^\star_{\min}}{\epsilon} \leq \frac{25}{\eta N(N-1)w^\star_{\min}}\left(\frac{1}{\alpha^{(N-2)/2}} - \frac{1}{\alpha^{(N-2)/4}}\right) \\
\Longrightarrow &T_3 \leq T_4.
\end{aligned}
\tag{A.12}
$$

□

A.3.3 Proof of Technical Lemmas

Proof of Lemma 4. Since $\frac{1}{\sqrt{n}}\mathbf{X}$ is with ℓ_2-normalized columns and satisfies μ-coherence, where $0 \leq \mu \leq 1$,

$$
\left|\left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}\right)_{i,j}\right| = \left|\left(\frac{1}{\sqrt{n}}\mathbf{X}_i\right)^\top\left(\frac{1}{\sqrt{n}}\mathbf{X}_j\right)\right| \leq \max\{1,\mu\} \leq 1.
$$

Therefore, for any $\mathbf{z} \in \mathbb{R}^p$,

$$
\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{z}\right\|_\infty \leq p\left\|\mathbf{z}\right\|_\infty.
$$

□

Proof of Lemma 5. It is straightforward to verify that for any $i \in \{1,\ldots,p\}$,

$$
\left|\left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{z}\right)_i - \mathbf{z}_i\right| \leq k\mu\left\|\mathbf{z}\right\|_\infty.
$$

Therefore,

$$
\left\|\left(\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{I}\right)\mathbf{z}\right\|_\infty \leq k\mu\left\|\mathbf{z}\right\|_\infty.
$$

□

Proof of Lemma 6. Since the vector $\boldsymbol{\xi}$ are made of independent σ^2-subGaussian random vari-

ables and any column $\mathbf{X}_i$ of $\mathbf{X}$ is ℓ_2-normalized, i.e. $\left\|\frac{1}{\sqrt{n}}\mathbf{X}_i\right\| = 1$, the random variable $\frac{1}{\sqrt{n}}(\mathbf{X}^\mathsf{T}\boldsymbol{\xi})_i$ is still σ^2-subGaussian.

It is a standard result that for any $\epsilon > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{n}}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty > \epsilon\right) \leq 2p\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Setting $\epsilon = 2\sqrt{2\sigma^2\log(2p)}$, with probability at least $1 - \frac{1}{8p^3}$ we have

$$\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \leq \frac{1}{\sqrt{n}}2\sqrt{\sigma^2\log(2p)} \lesssim \sqrt{\frac{\sigma^2\log p}{n}}.$$

□

A.4 Proof of Theorems in Chapter 2.3

In this section, we provide the proof for all results we mentioned in Section 2.3.

A.4.1 Proof of Theorem 1

Proof. Now let us consider the updates in equation (A.2). The major idea is to show that the results in Theorem 8 can be easily generalized with the lemmas we developed in Section A.2.3.

Let us denote

$$\Psi(w^\star_{\min}, N) \coloneqq (2 - 2^{\frac{N-2}{N}})^{\frac{1}{N-2}}(w^\star_{\min})^{\frac{1}{N}} \wedge 2^{\frac{3}{N}}(2^{\frac{1}{N}} - 1)^{\frac{1}{N-2}}(w^\star_{\min})^{\frac{1}{N}}.$$

We set

$$\alpha \leq \left(\frac{\epsilon}{p+1}\right)^{4/N} \wedge \Phi(w^\star_{\max}, w^\star_{\min}, \epsilon, N) \wedge \Psi(w^\star_{\min}, N).$$

Under the same requirements on other parameters with Theorem 8, we satisfy the conditions

of Lemma 7, Lemma 8 and Lemma 19. From these lemmas, we could maintain that

$$w_j^\star > 0 \Longrightarrow 0 \leq w_t^- \leq \alpha^{N/4},$$
$$w_j^\star < 0 \Longrightarrow 0 \leq w_t^+ \leq \alpha^{N/4},$$

up to $T_2 + T_4$ as defined in Proposition 1 and 2.

Consequently, for $w_j^\star > 0$ we can ignore $(w_{j,t}^-)_{t\geq 0}$ by treating as a part of bounded error b_t. The same holds for sequence $(w_{j,t}^+)_{t\geq 0}$ when $w_j^\star < 0$. Then, for $w_j^\star > 0$ the sequence $(w_{j,t}^+)$ evolves as follows

$$w_{j,t+1}^+ = w_{j,t}^+(1 - 2N\eta(w_{j,t}^+ - w_j^\star + (b_{j,t} - w_{j,t}^-) + p_{j,t})(w_{j,t}^+)^{(N-2)/2})^N.$$

The $b_{j,t} - w_{j,t}^-$ explains why we need $\|\mathbf{b}_t\|_\infty + \alpha^{N/4} \leq C_b\zeta$ in Proposition 1. For $w_j^\star > 0$, we follow the exact proof structure with Theorem 8 with treating $(w_{j,t}^-)_{t\geq 0}$ as a part of bounded error. For $w_j^\star < 0$ it follows the same argument by switching w_t^+ and w_t^-.

Therefore, we could closely follow the proof of Theorem 8 to generalize the result from non-negative signals to general signals. The result remains unchanged as well as the number of iterations requirement in equation (A.5) and (A.6). With the choice of $C_b = \frac{1}{100}$ in the proof of Theorem 8, recall that

$$\zeta = \frac{1}{5}w_{\min}^\star \vee 200\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee 200\epsilon,$$

and define the indicator function with A as the event $\{\frac{1}{5}w_{\min}^\star > 200\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee 200\epsilon\}$,

$$\mathbb{1}(A) = \begin{cases} 1, & \text{when} \quad \frac{1}{5}w_{\min}^\star > 200\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee 200\epsilon, \\ 0, & \text{when} \quad \frac{1}{5}w_{\min}^\star \leq 200\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\boldsymbol{\xi}\right\|_\infty \vee 200\epsilon. \end{cases}$$

We now define that

$$
\begin{aligned}
T_l(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon) &\coloneqq \frac{75}{16\eta N^2 \zeta^{(2N-2)/N}} \log \frac{|w^\star_{\max} - \alpha^N|}{\epsilon} + \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}} \\
&\quad + \frac{6}{\eta N^2 (w^\star_{\min})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon} \mathbb{1}(A), \\
T_u(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon) &\coloneqq \frac{5}{N(N-1)\eta\zeta} \left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}} \right) \\
&\quad + \frac{25}{N(N-1)\eta w^\star_{\min}} \left(\frac{1}{\alpha^{(N-2)/2}} - \frac{1}{\alpha^{(N-2)/4}} \right) \mathbb{1}(A).
\end{aligned}
\tag{A.13}
$$

The error bound (A.4) holds for any t such that

$$
T_l(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon) \leq t \leq T_u(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon).
$$

The equation (A.11) and (A.12) ensure that it is not a null set.

Thus, we finish generalizing Theorem 8 to general signals with an extra requirement $\Psi(w^\star_{\min}, N)$ on the initialization α.

For the case $k = 0$, i.e., $\mathbf{w}^\star = \mathbf{0}$, we set $w^\star_{\min} = 0$ and

$$
\alpha \leq \left(\frac{\epsilon}{p+1} \right)^{4/N}.
$$

Conditioning on $\|\mathbf{e}_t\|_\infty \leq \alpha^{N/4}$, we still have that

$$
\|\mathbf{b}_t\|_\infty + \alpha^{N/4} \leq p\alpha^{N/4} + \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty + \alpha^{N/4} \leq 2 \left(\left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty \vee \epsilon \right) \leq C_b \zeta \leq \frac{1}{40} \zeta.
$$

Therefore, by Lemma 8, for $\eta \leq \frac{1}{N(N-1)\zeta\alpha^{(N-2)/2}}$, we ensure $\|\mathbf{e}_t\|_\infty \leq \alpha^{N/4}$ up to

$$
\frac{5}{N(N-1)\eta\zeta} \left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{(N-2)/2}} \right),
$$

which agrees to the definition of $T_u(\mathbf{w}^\star, \alpha, N, \eta, \zeta, \epsilon)$ in this case. □

A.4.2 Proof of Corollary 1

Since $\boldsymbol{\xi}$ is made of independent σ^2-sub-Gaussian entries, by Lemma 6 with probability $1 - 1/(8p^3)$ we have

$$\left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \boldsymbol{\xi} \right\|_\infty \leq 2\sqrt{\frac{2\sigma^2 \log(2p)}{n}}.$$

Hence, letting $\epsilon = 2\sqrt{\frac{2\sigma^2 \log(2p)}{n}}$, we obtain that

$$\|\mathbf{w}_t - \mathbf{w}^\star\|_2^2 \lesssim \sum_{i \in S} \epsilon^2 + \sum_{i \notin S} \alpha^{N/2} \leq k\epsilon^2 + (p-k)\frac{\epsilon^2}{(p+1)^2} \lesssim \frac{k\sigma^2 \log p}{n}.$$

□

A.4.3 Proof of Theorem 2

We now state Theorem 2 formally as below.

Theorem 9. *Let T_1, T_2, T_3 and T_4 be the number of iterations defined in Proposition 1 and Proposition 2. Suppose $\zeta \geq 1$, $w^\star_{\max} \geq 1$ and the initialization $\alpha \leq \exp(-5/3)$, fixing α and η for all N, both $T_2 - T_1$ and $T_4 - T_3$ have a tight lower bound that is increasing as N increases ($N > 2$).*

Proof. We observe first that under the assumption $\zeta \geq 1$ and $w^\star_{\max} \geq 1$, $\frac{75}{16\eta N^2 \zeta^{(2N-2)/N}} \log \frac{|w^\star_{\max} - w_0|}{\epsilon}$ and $T_3 = \frac{6}{\eta N^2 (w^\star_{\max})^{(2N-2)/N}} \log \frac{w^\star_{\min}}{\epsilon}$ are decreasing as N increases.

For the rest part of $T_2 - T_1$, we will be showing that a lower bound of that is increasing as N increases. As $T_2 - T_1$ is by design a lower bound of the "true" early stopping window, the lower bound we get here is tight for $T_2 - T_1$ and is treated as equivalent to $T_2 - T_1$ to indicate the monotonicity of the "true" early stopping window.

$$\begin{aligned} &\frac{5}{N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}}\right) - \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}} \\ &\qquad \geq \frac{5}{4N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{4}{\alpha^{(N-2)/2}}\right) \end{aligned}$$

Denote

$$f(N) = \frac{1}{N(N-1)} \left(\frac{1}{\alpha^{(N-2)}} - \frac{4}{\alpha^{(N-2)/2}} \right).$$

Therefore,

$$\begin{aligned} f'(N) &= \frac{-(2N-1)}{N^2(N-1)^2} \left(\frac{1}{\alpha^{(N-2)}} - \frac{4}{\alpha^{(N-2)/2}} \right) \\ &+ \frac{1}{N(N-1)} (-\log \alpha) \left(\frac{1}{\alpha^{(N-2)}} - \frac{4}{2\alpha^{(N-2)/2}} \right) \\ &= \frac{-(2N-1) - (N-1)N \log \alpha}{2N^2(N-1)^2} \left(\frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{(N-2)/2}} \right) \\ &+ \frac{2N-1}{N^2(N-1)^2} \frac{2}{\alpha^{(N-2)/4}} \end{aligned}$$

Note that the second term is always positive, we just need to show the first term is positive.

$$\begin{aligned} -(2N-1) - (N-1)N \log \alpha \geq 0, \\ \frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{(N-2)/2}} \geq 0, \end{aligned}$$

which is satisfied when

$$\begin{aligned} \log \alpha \leq \min_{N \geq 3} \frac{(1-2N)}{N(N-1)} = \min_{N \geq 3} \left(\frac{1}{1-N} - \frac{1}{N} \right) = -\frac{5}{6} \\ \alpha^{(N-2)/2} \leq 1/2. \end{aligned}$$

We can further derive that when $\alpha \leq \exp(-5/6) \wedge 1/4$, we have a lower bound of $T_2 - T_1$ is increasing as N increases.

To show $T_4 - T_3$ is increasing as N increases, we just need to show T_4 is increasing. It follows a similar proof.

We can further derive that when $\alpha \leq \exp(-5/3) \wedge 2$, we have $T_4 - T_3$ is increasing as N increases.

□

A.4.4 Proof of Remark 2

The proof is indeed similar to that of Theorem 9. Fixing any $N > 2$ and step size η, we look at $T_2 - T_1$ and $T4 - T3$ and show that a tight lower bound of that is increasing as α decreases. We start with $T_2 - T_1$.

Recall that

$$\begin{aligned} T_2 - T_1 &= \frac{5}{N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}}\right) - \frac{15}{8N(N-2)\eta\zeta\alpha^{(N-2)}}) \\ &\quad - \frac{75}{16\eta N^2\zeta^{(2N-2)/N}}\log\frac{|w^\star_{\max} - \alpha^N|}{\epsilon} \\ &\geq \frac{5}{N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}}\right) - \frac{75}{16\eta N^2\zeta^{(2N-2)/N}}\log\frac{|w^\star_{\max}|}{\epsilon} \end{aligned}$$

Notice that the second term is not about α. We just need to show that $f(\alpha) = \frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}}$ is increasing as α decreases. With the general requirement of $\alpha < 1$, we have that

$$\begin{aligned} f'(\alpha) &= -\frac{(N-2)}{\alpha^{(N-1)}} + \frac{(N-2)/2}{\alpha^{N/2}} \\ &= (N-2)\left(\frac{1}{2\alpha^{N/2}} - \frac{1}{\alpha^{(N-1)}}\right) \\ &= (N-2)\frac{\alpha^{(N-2)/2} - 2}{2\alpha^{(N-1)}} < 0. \end{aligned}$$

For $T_4 - T_3$, it follows a similar proof.

A.5 Experiments on MNIST

The efficacy of different depth parameter N is shown in Figure 2.2 and Figure A.1 on both simulated data and real world datasets. The number of measurements is set as $n = 392$, where the dimension of the original image is $p = 784$. We use Rademacher sensing matrix. The MNIST

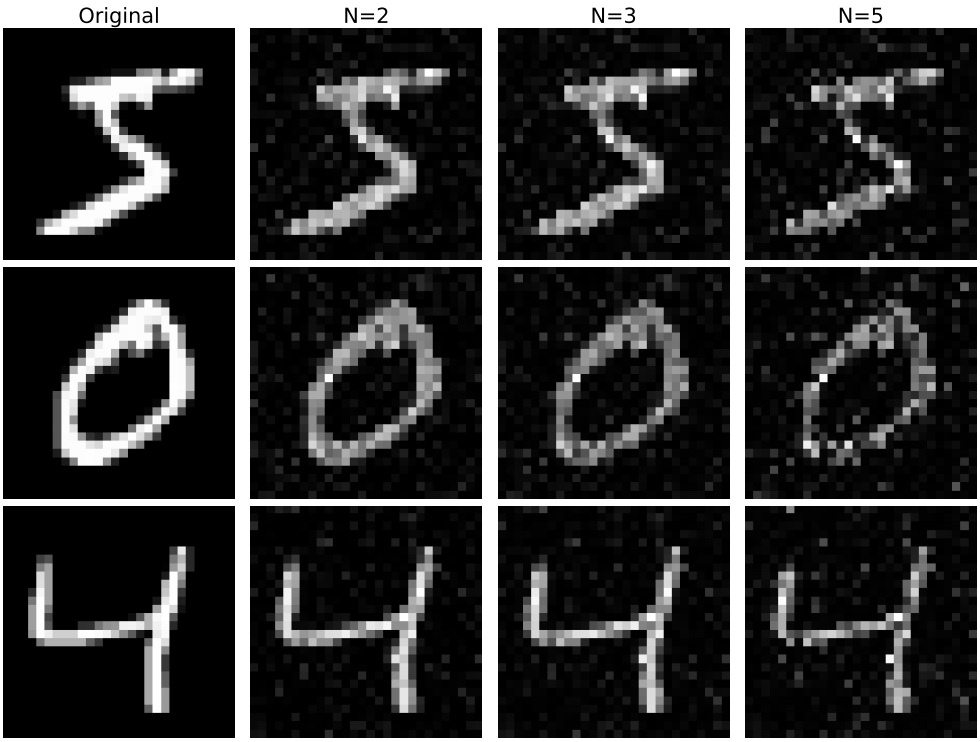


Figure A.1: Experiments with different choice depth parameter N. Reprinted with permission from [1].

examples are successfully recovered from Rademacher linear measurements using different deep parametrizations.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER III

B.1 Geometric properties of the parametrization

We start by calculating the vector field induced by the parameterization $G(\cdot)$.

$$\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) = 2u_{g(i)}v_i\mathbf{e}_{g(i)} + u_{g(i)}^2\mathbf{e}_{L+i},$$

where $\mathbf{e}_i \in \mathbb{R}^{L+p}$ is only 1 on i^{th} entry and 0 elsewhere, and

$$\nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top]) = 2v_i\mathbf{E}_{g(i),g(i)} + 2u_{g(i)}\mathbf{E}_{g(i),L+i} + 2u_{g(i)}\mathbf{E}_{L+i,g(i)},$$

where $\mathbf{E}_{i,j} \in \mathbb{R}^{(L+p)\times(L+p)}$ is the one-hot matrix for i^{th} row and j^{th} column. For $i \neq j$ s.t. $g(i) = g(j)$,

$$\begin{aligned}\nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_j([\mathbf{u}^\top, \mathbf{v}^\top]) &= (2v_i\mathbf{E}_{g(i),g(i)} + 2u_{g(i)}\mathbf{E}_{g(i),L+i} + 2u_{g(i)}\mathbf{E}_{L+i,g(i)}) \\ &\quad \cdot (2u_{g(j)}v_j\mathbf{e}_{g(j)} + u_{g(j)}^2\mathbf{e}_{L+j}) \\ &= 4u_{g(j)}v_iv_j\mathbf{e}_{g(i)} + 4u_{g(i)}u_{g(j)}v_j\mathbf{e}_{L+i} \\ &= 4u_{g(i)}v_iv_j\mathbf{e}_{g(i)} + 4u_{g(i)}^2v_j\mathbf{e}_{L+i},\end{aligned}$$

similarly,

$$\nabla^2 G_j([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) = 4u_{g(i)}v_iv_j\mathbf{e}_{g(i)} + 4u_{g(i)}^2v_i\mathbf{e}_{L+j}.$$

Proof for Lemma 1. For two indices within the same group, i.e, $i \neq j$ and $g(i) = g(j)$, we

obtain that

$$\begin{aligned}[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) =& \nabla^2 G_j([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) - \nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_j([\mathbf{u}^\top, \mathbf{v}^\top]) \\ =& 4u_{g(i)}^2 v_j \mathbf{e}_{L+i} - 4u_{g(i)}^2 v_i \mathbf{e}_{L+j},\end{aligned}$$

which is not always $\mathbf{0}$ when $v_i \neq v_j$. Therefore, $G(\cdot)$ is not commuting. □

Proof for Theorem 3. For $i \neq j$ and $g(i) \neq g(j)$, we have

$$[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) = \mathbf{0}.$$

For $i \neq j$ and $g(i) = g(j)$, we have that

$$[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) = v_j \nabla G_i - v_i \nabla G_j \in \operatorname{span}\{\nabla G_i\}_{i=1}^p.$$

By Corollary 4.13 in [97] and Lemma 1, we show that there exists and initialization and a time-dependent loss that the gradient flow can not be analyzed by mirror flow. □

Alternatively, we can show directly that the necessary condition in Theorem 4.10 in [97] is violated, i.e.,

$$\langle \nabla G_j, [\nabla G_i, [\nabla G_i, \nabla G_j]]\rangle([\mathbf{u}^\top, \mathbf{v}^\top]) \neq 0$$

for some $[\mathbf{u}^\top, \mathbf{v}^\top]$ in every open set M.

We first obtain that

$$\begin{aligned}\nabla[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) =& 8u_{g(i)} v_j \mathbf{E}_{L+i,g(i)} + 4u_{g(i)}^2 \mathbf{E}_{L+i,L+j} \\ & - 8u_{g(i)} v_i \mathbf{E}_{L+j,g(i)} - 4u_{g(i)}^2 \mathbf{E}_{L+j,L+i}.\end{aligned}$$

Therefore,

$$
\begin{aligned}
[\nabla G_i, [\nabla G_i, \nabla G_j]]([\mathbf{u}^\top, \mathbf{v}^\top]) &= \nabla[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top])\nabla G_i([\mathbf{u}^\top, \mathbf{v}^\top]) \\
&\quad - \nabla^2 G_i([\mathbf{u}^\top, \mathbf{v}^\top])[\nabla G_i, \nabla G_j]([\mathbf{u}^\top, \mathbf{v}^\top]) \\
&= (8u_{g(i)}v_j\mathbf{E}_{L+i,g(i)} + 4u_{g(i)}^2\mathbf{E}_{L+i,L+j} \\
&\quad - 8u_{g(i)}v_i\mathbf{E}_{L+j,g(i)} - 4u_{g(i)}^2\mathbf{E}_{L+j,L+i}) \\
&\quad \cdot (2u_{g(i)}v_i\mathbf{e}_{g(i)} + u_{g(i)}^2\mathbf{e}_{L+i}) \\
&\quad - (2v_i\mathbf{E}_{g(i),g(i)} + 2u_{g(i)}\mathbf{E}_{g(i),L+i} + 2u_{g(i)}\mathbf{E}_{L+i,g(i)}) \\
&\quad \cdot (4u_{g(i)}^2v_j\mathbf{e}_{L+i} - 4u_{g(i)}^2v_i\mathbf{e}_{L+j}) \\
&= 16u_{g(i)}^2v_iv_j\mathbf{e}_{L+i} - 16u_{g(i)}^2v_i^2\mathbf{e}_{L+j} - 4u_{g(i)}^4\mathbf{e}_{L+j} - 8u_{g(i)}^3v_j\mathbf{e}_{g(i)} \\
&= 16u_{g(i)}^2v_iv_j\mathbf{e}_{L+i} - (16u_{g(i)}^2v_i^2 + 4u_{g(i)}^4)\mathbf{e}_{L+j} - 8u_{g(i)}^3v_j\mathbf{e}_{g(i)}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\langle \nabla G_j, [\nabla G_i, [\nabla G_i, \nabla G_j]]\rangle([\mathbf{u}^\top, \mathbf{v}^\top]) \\
=&\langle 2u_{g(i)}v_j\mathbf{e}_{g(i)} + u_{g(i)}^2\mathbf{e}_{L+j}, 16u_{g(i)}^2v_iv_j\mathbf{e}_{L+i} - (16u_{g(i)}^2v_i^2 + 4u_{g(i)}^4)\mathbf{e}_{L+j} - 8u_{g(i)}^3v_j\mathbf{e}_{g(i)}\rangle \\
=&-16u_{g(i)}^4v_j^2 - 16u_{g(i)}^4v_i^2 - 4u_{g(i)}^6 < 0.
\end{aligned}
$$

By Theorem 4.10 in [97], there exists an initialization such that no Legendre function R is able to make the gradient flow be written as a mirror flow with respect to R.

B.2 Proof for Analysis of Gradient Flow

Proof for Lemma 2. Recall

$$
\frac{\partial \mathcal{L}}{\partial u_l} = -\frac{2}{n}u_l\mathbf{v}_l^\top\mathbf{X}_l^\top\mathbf{r}(t), \quad \frac{\partial \mathcal{L}}{\partial \mathbf{v}_l} = -\frac{1}{n}u_l^2\mathbf{X}_l^\top\mathbf{r}(t).
$$

Therefore, we obtain that

$$\begin{aligned}\frac{\partial \left\|\mathbf{v}_l(t)\right\|^2}{\partial t} &= 2\mathbf{v}_l^\top(t)\frac{\partial \mathbf{v}_l(t)}{\partial t} = 2\mathbf{v}_l^\top(t)\left(-\frac{\partial \mathcal{L}}{\partial \mathbf{v}_l}\right) \\ &= \frac{2}{n}u_l^2\mathbf{v}_l^\top(t)\mathbf{X}_l^\top\mathbf{r}(t) \\ &= u_l\left(-\frac{\partial \mathcal{L}}{\partial u_l}\right) = \frac{\partial \frac{1}{2}u_l^2(t)}{\partial t}.\end{aligned}$$

□

Proof for Lemma 3. We start with decomposing $\mathbf{v}_l(0)$

$$\begin{aligned}\mathbf{v}_l(0) = \eta\frac{1}{n}\mathbf{X}_l^\top\mathbf{y} &= \eta\mathbf{w}_l^\star + \eta\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X} - \mathbf{I}\right)\mathbf{w}_l^\star + \eta\sum_{l'\neq l}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}\mathbf{w}_{l'}^\star + \eta\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi} \\ &= \eta\mathbf{w}_l^\star + \eta\mathbf{b}_l.\end{aligned}$$

With this decomposition, we have that

$$\begin{aligned}\langle\mathbf{v}_l(0), \mathbf{v}_l^\star\rangle^2 &= \eta^2((u_l^\star)^2 + \langle\mathbf{b}_l, \mathbf{v}_l^\star\rangle)^2 \\ \left\|\mathbf{v}_l(0)\right\|_2^2 &= \eta^2((u_l^\star)^4 + 2\langle\mathbf{b}_l, \mathbf{w}_l^\star\rangle + \left\|\mathbf{b}_l\right\|_2^2).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\langle\mathbf{v}_l(0), \mathbf{v}_l^\star\rangle^2}{\left\|\mathbf{v}_l(0)\right\|_2^2} &= \frac{\eta^2((u_l^\star)^2 + \langle\mathbf{b}_l, \mathbf{v}_l^\star\rangle)^2}{\eta^2((u_l^\star)^4 + 2\langle\mathbf{b}_l, \mathbf{w}_l^\star\rangle + \left\|\mathbf{b}_l\right\|_2^2)} \\ &= 1 - \frac{\left\|\mathbf{b}_l\right\|_2^2 - \langle\mathbf{b}_l, \mathbf{v}_l^\star\rangle^2}{(u_l^\star)^4 + 2\langle\mathbf{b}_l, \mathbf{w}_l^\star\rangle + \left\|\mathbf{b}_l\right\|_2^2} \\ &= 1 - \frac{\left\|\mathbf{b}_l/(u_l^\star)^2\right\|_2^2 - \langle\mathbf{b}_l/(u_l^\star)^2, \mathbf{v}_l^\star\rangle^2}{1 + 2\langle\mathbf{b}_l/(u_l^\star)^2, \mathbf{v}_l^\star\rangle + \left\|\mathbf{b}_l/(u_l^\star)^2\right\|_2^2} \\ &= 1 - \frac{1 - \langle\mathbf{b}_l/\left\|\mathbf{b}_l\right\|, \mathbf{v}_l^\star\rangle^2}{1 + 2\left\|\mathbf{b}_l\right\|/(u_l^\star)^2\langle\mathbf{b}_l/\left\|\mathbf{b}_l\right\|, \mathbf{v}_l^\star\rangle + \left\|\mathbf{b}_l\right\|^2/(u_l^\star)^4}\left\|\mathbf{b}_l/(u_l^\star)^2\right\|^2 \\ &\geq 1 - \left\|\mathbf{b}_l/(u_l^\star)^2\right\|_2^2,\end{aligned}$$

where last inequality is from

$$\frac{1-\alpha^2}{\beta^2+2\alpha\beta+1} = \frac{1}{\frac{\beta^2+2\alpha\beta+1}{1-\alpha^2}} = \frac{1}{1+\frac{\beta^2+2\alpha\beta+\alpha^2}{1-\alpha^2}}$$
$$= \frac{1}{1+\frac{(\alpha+\beta)^2}{1-\alpha^2}} \leq 1,$$

for $0 \leq \alpha \leq 1$.

Since

$$\|\mathbf{b}_l\|_2 \leq \delta_{in}(u_l^\star)^2 + L\delta_{out}(u_l^\star)^2 + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2,$$

we obtain that

$$\left\langle \frac{\mathbf{v}_l(0)}{\|\mathbf{v}_l(0)\|}, \mathbf{v}_l^\star \right\rangle \geq 1 - \left(\delta_{in} + L\delta_{out} + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2 / (u_l^\star)^2\right)^2.$$

□

Lemma 20. *Consider a simplified case where* $\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_l = \mathbf{I}$, $\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_{l'} = \mathbf{O}, l \neq l'$, *if* $\mathbf{v}_l(0) = \eta\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$, *then*

$$\mathbf{v}_l(t) = c\frac{1}{n}\mathbf{X}_l^\top \mathbf{y},$$

for some constant c.

Proof. From the gradient on the directions, we have that

$$\frac{\partial \mathbf{v}_l(t)}{\partial t} = \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \mathbf{r}(t) = \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \mathbf{y} - \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \sum_{l'} \mathbf{X}_{l'} u_{l'}^2(t)\mathbf{v}_{l'}(t)$$
$$= \frac{1}{n}u_l^2(t)\mathbf{X}_l^\top \mathbf{y} - u_l^4(t)\mathbf{v}_l(t).$$

Since $\mathbf{v}_l(0)$ is with the same direction as $\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$ at the initialization. Therefore, $\frac{\partial \mathbf{v}_l(t)}{\partial t}$ has the same direction as $\mathbf{v}_l(t)$. We obtain that $\mathbf{v}_l(t) = c\frac{1}{n}\mathbf{X}_l^\top \mathbf{y}$ for some constant c. □

Lemma 21. *If the gradient flow satisfies*

$$\frac{1}{2}\frac{\partial u^2(t)}{\partial t} \leq u^6(t) + \sqrt{2}u^4(t)B$$

for some constant $B > 0$, then for any $t \leq T = \frac{\log \frac{1}{\theta}}{2\theta^2 + \theta\sqrt{2}B}$ we have $u(t) \leq \sqrt{\theta}$ with initialization $u(0) = \theta$.

Proof. We wanted to find some time T such that when $t \leq T$, $u(t) \leq \sqrt{\theta}$. Since the gradient is bounded from above, we obtain that

$$\begin{aligned}\frac{1}{2}u^2(T) &\leq \frac{1}{2}\theta^2 \cdot \exp\left(\int_0^T 2u^4(t) + \sqrt{2}u^2(t)Bdt\right) \\ &\leq \frac{1}{2}\theta^2 \cdot \exp\Big((2\theta^2 + \sqrt{2}\theta B)T\Big) \leq \frac{1}{2}\theta.\end{aligned}$$

This gives us

$$T \leq \frac{\log \frac{1}{\theta}}{2\theta^2 + \theta\sqrt{2}B}.$$

□

Lemma 22. *Fix any $\tau < \frac{1}{2}$. Consider the gradient flow*

$$\frac{1}{2}\frac{\partial u^2(t)}{\partial t} \geq (1-2B)\sqrt{2}u^3(t)(u^\star)^2 - u^6(t) - \sqrt{2}u^3(t)B(u^\star)^2$$

for some constant $0 < B < \frac{1}{10}$ with initialization $u(0) = \theta < \frac{1}{2}u^\star$, we have that

$$\left|\frac{1}{\sqrt{2}}u^3(t) - (u^\star)^2\right| < (1 - 3B - \tau)(u^\star)^2,$$

after

$$t \geq T = \frac{2^{1/3}(u^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u^\star)^2\theta} + \frac{2\log_2 \frac{1}{2\tau}}{3(u^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u^\star)^2\right)^{1/3}}.$$

Proof. For any $T \geq 0$, we have that

$$\frac{1}{2}u^2(T) \geq \frac{1}{2}\theta^2 \cdot \exp\left(\int_0^T (1-2B)2\sqrt{2}u(t)(u^\star)^2 - 2u^4(t) - 2\sqrt{2}u(t)B(u^\star)^2 dt\right).$$

When $u(t) < \frac{1}{2}u^\star$, we first aim to get T_1 such that $\frac{1}{\sqrt{2}}u^3(T_1) \geq \frac{1}{2}(u^\star)^2$. Therefore,

$$\begin{aligned}
&\frac{1}{2}\theta^2 \cdot \exp\left(\int_0^T (1-2B)2\sqrt{2}u(t)(u^\star)^2 - 2u^4(t) - 2\sqrt{2}u(t)B(u^\star)^2 dt\right) \\
&\geq \frac{1}{2}\theta^2 \cdot \exp\left(\left((1-2B)2\sqrt{2}(u^\star)^2 - \sqrt{2}(u^\star)^2 - 2\sqrt{2}B(u^\star)^2\right)\theta T_1\right) \\
&\geq \frac{1}{2}\left(\frac{\sqrt{2}}{2}(u^\star)^2\right)^{2/3}.
\end{aligned}$$

We obtain that

$$T \geq \frac{2^{1/3}(u^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u^\star)^2\theta}.$$

When $t \geq T_1$, we have that $\frac{1}{\sqrt{2}}u^3(t) \geq \frac{1}{2}(u^\star)^2$. Let us denote $\frac{1}{\sqrt{2}}u^3(0) = ((1-3B)-\eta)(u^\star)^2$, we wonder how many iterations T_d are needed to make $\frac{1}{\sqrt{2}}u^3(T_d) \geq \left((1-3B)-\frac{1}{2}\eta\right)(u^\star)^2$.

$$\begin{aligned}
&\frac{1}{2}\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{2/3} \cdot \exp\left(\int_0^T (1-2B)2\sqrt{2}u(t)(u^\star)^2 - 2u^4(t) - 2\sqrt{2}u(t)B(u^\star)^2 dt\right) \\
&\geq \frac{1}{2}\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{2/3} \cdot \exp\left(\left(\frac{1}{2}\eta(u^\star)^2\right)\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3} T_2\right) \\
&\geq \frac{1}{2}\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{2/3} \cdot \left(1 + \left(\frac{1}{2}\eta(u^\star)^2\right)\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3} T_2\right) \\
&\geq \frac{1}{2}\left(\sqrt{2}\left((1-3B)-\frac{1}{2}\eta\right)(u^\star)^2\right)^{2/3}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
T_2 &\geq \frac{\left((1-3B)-\frac{1}{2}\eta\right)^{2/3}-\left((1-3B)-\eta\right)^{2/3}}{\left((1-3B)-\eta\right)^{2/3}} \frac{1}{\frac{1}{2}\eta(u^\star)^2\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3}} \\
&\geq \frac{2}{3}\frac{\frac{1}{2}\eta}{\frac{1}{2}\eta(u^\star)^2\left((1-3B)-\eta\right)\left(\sqrt{2}\left((1-3B)-\eta\right)(u^\star)^2\right)^{1/3}} \\
&\geq \frac{2}{3(u^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u^\star)^2\right)^{1/3}}.
\end{aligned}$$

Overall, we obtain that

$$\left|\frac{1}{\sqrt{2}}u^3(t)-(u^\star)^2\right| < (1-3B-\epsilon)(u^\star)^2,$$

after

$$t \geq T = T_1 + T_2 \log_2 \frac{1}{2\tau}.$$

□

Proof of Theorem 4. Denote $\zeta = 100\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty$. For $l \in S$, the gradient flow can be simplied as

$$\begin{aligned}
\frac{1}{2}\frac{\partial u_l^2(t)}{\partial t} &= \frac{2}{n}\mathbf{w}_l^\top(t)\mathbf{X}_l^\top\mathbf{r}(t) \\
&= 2\mathbf{w}_l^\top(t)(\mathbf{w}_l^\star-\mathbf{w}_l(t)) + \frac{2}{n}\mathbf{w}_l^\top\mathbf{X}_l^\top\boldsymbol{\xi} \\
&\geq 2u_l^2(t)(u_l^\star)^2\langle\mathbf{v}_l(t),\mathbf{v}_l^\star\rangle - 2u_l^4(t)\left\|\mathbf{v}_l(t)\right\|_2^2 - 2u_l^2(t)\left\|\mathbf{v}_l(t)\right\|_2\left\|\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right\|_2.
\end{aligned}$$

Since the initialization is balanced $\frac{1}{2}u_l^2(0) = \left\|\mathbf{v}_l(0)\right\|_2^2$, we know that from the balancing result Lemma 2,

$$\frac{1}{2}u_l^2(t) = \left\|\mathbf{v}_l(t)\right\|_2^2.$$

Since the initialization of $\mathbf{v}_l(t)$ is aligned with direction $\frac{1}{n}\mathbf{X}_l^\top\mathbf{y}$, and with our assumption on orthogonal design, by Lemma 3 and Lemma 20, if $\left\|\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right\|_2 \leq B(u_l^\star)^2$, we can further simplify the

gradient flow as

$$\begin{aligned}\frac{1}{2}\frac{\partial u_l^2(t)}{\partial t} &\geq \sqrt{2}(1-2B^2)u_l^3(t)(u_l^\star)^2 - u_l^6(t) - \sqrt{2}u_l^3(t)B \\ &\geq \sqrt{2}(1-2B)u_l^3(t)(u_l^\star)^2 - u_l^6(t) - \sqrt{2}u_l^3(t)B,\end{aligned}$$

where the last inequality holds when $B < 1$. We will verify that $B < 1$ holds in the following analysis.

If $\zeta \geq (u_{max}^\star)^2$, then our desired inequality is achieved at the initialization.

If $(u_{min}^\star)^2 \leq \zeta \leq (u_{max}^\star)^2$, for these group that $\zeta \leq (u_l^\star)^2$, applying Lemma 22 with

$$B = \frac{\left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_2}{(u_l^\star)^2} \leq \frac{\left\|\frac{1}{n}\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty}{(u_l^\star)^2} \leq \frac{1}{100}, \quad \tau = \frac{\epsilon}{(u_l^\star)^2}$$

we obtain the convergence on magnitudes

$$\left|\, \|\mathbf{w}_l(t)\|_2 - \|\mathbf{w}_l^\star\|_2 \,\right| \leq (3B+\epsilon)\, \|\mathbf{w}_l^\star\|_2 \,,$$

after

$$\frac{2^{1/3}(u_l^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u_l^\star)^2\theta} + \frac{2\log_2 \frac{(u_l)^2}{2\epsilon}}{3(u_l^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u_l^\star)^2\right)^{1/3}}.$$

If $\zeta \leq (u_{min}^\star)^2$, similarly applying Lemma 22, the number of iterations needed for entries on the support to converge is

$$T_l = \frac{2^{1/3}(u_{max}^\star)^{4/3}}{\theta^2}\frac{1}{(1-6B)\sqrt{2}(u_{min}^\star)^2\theta} + \frac{2\log_2 \frac{(u_{max})^2}{2\epsilon}}{3(u_{min}^\star)^2(1/2-3B)\left(\sqrt{2}(1/2-3B)(u_{min}^\star)^2\right)^{1/3}}.$$

We now have that for $l \in S$,

$$\left|\, \|\mathbf{w}_l(t)\|_2 - \|\mathbf{w}_l^\star\|_2 \,\right| \leq (3B+\epsilon)\, \|\mathbf{w}_l^\star\|_2 \,,$$

where $B = \frac{\left\|\frac{1}{n}\mathbf{X}^\top \mathbf{y}\right\|_\infty}{(u^\star_{min})^2} \leq \frac{1}{100}, \forall l \in S$.

Recall that the direction is lower bounded by Lemma 3 and Lemma 24,

$$\left\langle \frac{\mathbf{w}_l(t)}{\|\mathbf{w}_l(t)\|_2}, \frac{\mathbf{w}^\star_l}{\|\mathbf{w}^\star_l\|_2} \right\rangle \geq 1 - B^2.$$

Therefore, the error bound on the support is as follows,

$$\begin{aligned}
\|\mathbf{w}_l(t) - \mathbf{w}^\star_l\|_\infty \leq \|\mathbf{w}_l(t) - \mathbf{w}^\star_l\|_2 &= \left\| \left(\|\mathbf{w}_l(t)\|_2 - (u^\star_l)^2\right) \frac{\mathbf{v}_l(t)}{\|\mathbf{v}_l(t)\|} + (u^\star_l)^2 \left\langle \frac{\mathbf{v}_l(t)}{\|\mathbf{v}_l(t)\|}, \mathbf{v}^\star_l \right\rangle \right\|_2 \\
&\leq (3B+\tau)(u^\star_l)^2 + (u^\star_l)^2 \sqrt{2 - 2\left\langle \frac{\mathbf{v}_l(t)}{\|\mathbf{v}_l(t)\|}, \mathbf{v}^\star_l \right\rangle} \\
&= (3B+\tau)(u^\star_l)^2 + (u^\star_l)^2 \sqrt{2} B \leq \left\| \frac{1}{n}\mathbf{X}^\top \mathbf{y} \right\|_\infty + \epsilon.
\end{aligned}$$

For $l \notin S$, we derive a lower bound on the growth rate

$$\begin{aligned}
\frac{1}{2}\frac{\partial u^2_l(t)}{\partial t} &= \frac{2}{n}\mathbf{w}_l^\top(t)\mathbf{X}_l^\top \mathbf{r}(t) \\
&= 2\|\mathbf{w}_l(t)\|_2^2 + \frac{2}{n}\mathbf{w}_l^\top \mathbf{X}_l^\top \boldsymbol{\xi} \\
&\leq u^6_l(t) + \sqrt{2} u^4_l(t) B.
\end{aligned}$$

By applying Lemma 21 with $B = \left\|\frac{1}{n}\mathbf{X}^\top \mathbf{y}\right\|_\infty$, we obtain that before

$$T_u = \frac{\log \frac{1}{\theta}}{2\theta^2 + \theta\sqrt{2}B}.$$

Since $\theta < \frac{\epsilon}{2(u_{max})^2}$, $T_l < T_u$ is ensured.

□

B.3 Analysis of gradient descent

B.3.1 Monotonic updates

Lemma 23. *With an initialization $u(0) < u^\star$ and step size $\gamma \leq \frac{1}{4(u^\star)^2}$, the updating sequence*

$$u(t) = u(t-1) + 2\gamma u(t-1)[(u^\star)^2 - u^2(t-1)],$$

is always bounded above by $u^\star$.

Proof. We prove it by contradiction. Assume there is a time t s.t.

$$u(t) \leq u^\star, u(t+1) > u^\star.$$

Therefore,

$$u(t) + 2\gamma u(t)[(u^\star)^2 - u^2(t)] > u^\star.$$

Denote $\lambda = u(t)/u^\star$, we have that

$$1 + 2\gamma(u^\star)^2(1-\lambda^2) - 1/\lambda > 0$$

for some $\lambda \in (0, 1]$.

Let $f(\lambda) = 1 + 2\gamma(u^\star)^2(1-\lambda^2) - 1/\lambda$, we obtain the derivative

$$f'(\lambda) = -4\gamma(u^\star)^2\lambda + \frac{1}{\lambda^2} > 0.$$

However, $f_{max}(\lambda) = f(1) = 0$, and $f(\lambda) \leq 0$ for all $\lambda \in (0, 1]$, which gives our desired contradiction. □

B.3.2 Updates with bounded perturbations

To study the general non-orthogonal and noisy case, we first extend the lemmas above to gradient dynamics with bounded perturbations.

Consider the update on $\mathbf{v}(t)$ with bounded perturbations

$$\begin{aligned}\mathbf{z}(t+1) &= \mathbf{v}(t) + \eta_t u^2(t)((u^\star)^2\mathbf{v}^\star - u^2(t)\mathbf{v}(t)) + \eta_t u^2(t)\mathbf{b}_t \\ \mathbf{v}(t+1) &= \frac{\mathbf{z}(t+1)}{\|\mathbf{z}(t+1)\|}.\end{aligned} \tag{B.1}$$

and the updates on $u(t)$

$$u(t+1) = u(t) + 2\gamma u(t)\mathbf{v}^\top(t+1)\{(u^\star)^2\mathbf{v}^\star - u^2(t)\mathbf{v}(t+1)\} + 2\gamma u(t)e_t, \tag{B.2}$$

Note that if we choose $\eta_t = \frac{1}{u^4(t)}$, Eq. (B.1) is recast as

$$\begin{aligned}\mathbf{z}(t+1) &= \frac{(u^\star)^2}{u^2(t)}\mathbf{v}^\star + \frac{1}{u^2(t)}\mathbf{b}_t \\ \mathbf{v}(t+1) &= \frac{\mathbf{z}(t+1)}{\|\mathbf{z}(t+1)\|}.\end{aligned} \tag{B.3}$$

Lemma 24. *Consider the update in Eq.* (B.3)*, if* $\|\mathbf{b}_t\| \leq B(u^\star)^2$ *for some constant* $0 < B < 1$*, we have that*

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq 1 - B^2.$$

Proof. We have that

$$\begin{aligned}\langle \mathbf{z}(t+1), \mathbf{v}^\star \rangle &= \frac{(u^\star)^2}{u^2(t)} + \frac{1}{u_l^2(t)}\langle \mathbf{b}_t, \mathbf{v}^\star \rangle \\ \|\mathbf{z}(t+1)\|^2 &= \frac{(u^\star)^4}{u^4(t)} + 2\frac{(u^\star)^2}{u^4(t)}\langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \frac{1}{u_l^4(t)}\|\mathbf{b}_t\|^2,\end{aligned}$$

therefore,

$$\begin{aligned}
\frac{\langle \mathbf{z}(t+1), \mathbf{v}^\star \rangle^2}{\|\mathbf{z}(t+1)\|^2} &= \frac{\frac{(u^\star)^4}{u^4(t)} + 2\frac{(u^\star)^2}{u^4(t)}\langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \frac{1}{u_l^4(t)}\langle \mathbf{b}_t, \mathbf{v}^\star \rangle^2}{\frac{(u^\star)^4}{u^4(t)} + 2\frac{(u^\star)^2}{u^4(t)}\langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \frac{1}{u_l^4(t)}\|\mathbf{b}_t\|^2} \\
&= 1 - \frac{\|\mathbf{b}_t\|^2 - \langle \mathbf{b}_t, \mathbf{v}^\star \rangle^2}{(u^\star)^4 + 2(u^\star)^2\langle \mathbf{b}_t, \mathbf{v}^\star \rangle + \|\mathbf{b}_t\|^2} \\
&= 1 - \frac{\|\mathbf{b}_t/(u^\star)^2\|^2 - \langle \mathbf{b}_t/(u^\star)^2, \mathbf{v}^\star \rangle^2}{1 + 2\langle \mathbf{b}_t/(u^\star)^2, \mathbf{v}^\star \rangle + \|\mathbf{b}_t/(u^\star)^2\|^2} \\
&= 1 - \frac{1 - \langle \mathbf{b}_t/\|\mathbf{b}_t\|, \mathbf{v}^\star \rangle^2}{1 + 2\|\mathbf{b}_t\|/(u^\star)^2\langle \mathbf{b}_t/\|\mathbf{b}_t\|, \mathbf{v}^\star \rangle + \|\mathbf{b}_t\|^2/(u^\star)^4} \left\|\mathbf{b}_t/(u^\star)^2\right\|^2 \\
&\geq 1 - \left\|\mathbf{b}_t/(u^\star)^2\right\|^2 \\
&\geq 1 - B^2.
\end{aligned}$$

Hence, we have that

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq \sqrt{1 - B^2} \geq 1 - B^2.$$

□

Lemma 25. *Consider the updates in Eq.* (B.2) *with* $|e_t| \leq B$*, if* $u^2(0) \leq (u^\star)^2$*, then* $u^2(t) \leq (u^\star)^2 + B$ *for all* t*. If* $u^2(0) \geq (u^\star)^2$ *and* $|\langle \mathbf{v}(t), \mathbf{b}_t \rangle| \leq B_2\tau(u^\star)^2$*, then* $u^2(t) \geq (1 - B_2)(u^\star)^2 - B$ *for all* t*.*

Proof. Proof by contradiction similarly to Lemma 23. □

Lemma 26. *Fix the step size* γ *for the update on* $u(t)$*, and choose* $u(0) = \alpha \leq \frac{1}{5}u^\star$*. Consider the updates in Eq.* (B.2) *and Eq.* (B.1) *with* $|\langle \mathbf{v}(t), \mathbf{b}_t \rangle| \leq \frac{1}{20}(u^\star)^2$ *and* $|e_t| \leq \frac{1}{20}(u^\star)^2$*, then* $T \geq \frac{\log \frac{(u^\star)^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(u^\star)^2)}$*, we have that* $u^2(T) \geq \frac{1}{2}(u^\star)^2$*.*

Proof. Apply Lemma 24 with $B = \frac{1}{20}$,

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq 1 - B^2 = 1 - \frac{1}{400} \geq \frac{4}{5}$$

Starting from $t = 1$, we have that

$$\mathbf{v}^\top(t)\{(u^\star)^2\mathbf{v}^\star - u^2(t)\mathbf{v}(t)\} \geq \frac{4}{5}(u^\star)^2 - u^2(t),$$

therefore, we obtain an lower bound of the growth rate on $u(t)$, which reads

$$\begin{aligned} u(t+1) &\geq u(t) + 2\gamma u(t)\left(\frac{4}{5}(u^\star)^2 - u^2(t) - \frac{1}{20}(u^\star)^2\right) \\ &= u(t)\left(1 + 2\gamma\left(\frac{3}{4}(u^\star)^2 - u^2(t)\right)\right) \\ &\geq u(t)\left(1 + \gamma\frac{1}{2}(u^\star)^2\right). \end{aligned}$$

Therefore, the requirement on the number of iterations is recast as

$$\begin{aligned} &\alpha^2\left(1 + \gamma\frac{1}{2}(u^\star)^2\right)^{2T} \geq \frac{1}{2}(u^\star)^2 \\ \Longleftrightarrow &2T \geq \frac{\log\frac{(u^\star)^2}{2\alpha^2}}{\log(1 + \gamma\frac{1}{2}(u^\star)^2)} \\ \Longleftrightarrow &T \geq \frac{\log\frac{(u^\star)^2}{2\alpha^2}}{2\log(1 + \gamma\frac{1}{2}(u^\star)^2)}. \end{aligned}$$

With these requirements, by Lemma 25, we also have that $u^2(t) \leq \frac{3}{2}(u^\star)^2, \forall t \geq 0$. □

Lemma 27. *Fix the step size γ for the update on $u(t)$, and choose the initialization $u(0)$ such that $|(u^\star)^2 - u^2(0)| \leq \tau(u^\star)^2$ where $0 < \tau \leq 1/2$. Consider the updates in Eq.* (B.2) *and Eq.* (B.1) *with $|\langle\mathbf{v}(t), \mathbf{b}_t\rangle| \leq \frac{1}{10}\tau(u^\star)^2$ and $|e_t| \leq \frac{1}{10}\tau(u^\star)^2$, then after $T \geq \frac{5}{2\gamma(u^\star)^2}$, we have that $\langle\mathbf{v}(t), \mathbf{v}^\star\rangle \geq 1 - \frac{1}{5}\tau^2$ for all $t \leq T$ and $|u^2(T) - (u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$.*

Proof. When $u^2(0) \leq (u^\star)^2$, by applying to Lemma 24, we have that

$$\langle\mathbf{v}(t+1), \mathbf{v}^\star\rangle \geq 1 - \left(\frac{1}{10}\tau\right)^2 \geq 1 - \frac{1}{5}\tau^2,$$

therefore,

$$\begin{aligned} u(t+1) &\geq u(t) + 2\gamma u(t)\left(\left(1-\frac{1}{5}\tau\right)(u^\star)^2 - u^2(t) - \frac{1}{10}\tau(u^\star)^2\right) \\ &= u(t)\left(1+2\gamma\left(\left(1-\frac{3}{10}\tau\right)(u^\star)^2 - u^2(t)\right)\right). \end{aligned}$$

Further, we want to find an lower bound requirement on T s.t.

$$\left((u^\star)^2 - \tau(u^\star)^2\right)\left(1+2\gamma\left(\left(1-\frac{3}{10}\tau\right)(u^\star)^2 - \left((u^\star)^2 - \frac{1}{2}\tau\right)(u^\star)^2\right)\right)^{2T} \geq (u^\star)^2 - \frac{1}{2}\tau(u^\star)^2,$$

which can be relaxed as

$$\begin{aligned} &\left((u^\star)^2 - \tau(u^\star)^2\right)\left(1+\frac{2}{5}\gamma T\tau(u^\star)^2\right) \geq (u^\star)^2 - \frac{1}{2}\tau(u^\star)^2 \\ \Longleftrightarrow &1+\frac{2}{5}\gamma T\tau(u^\star)^2 \geq \frac{(u^\star)^2 - \frac{1}{2}\tau(u^\star)^2}{(u^\star)^2 - \tau(u^\star)^2} \\ \Longleftrightarrow &\frac{2}{5}\gamma T\tau(u^\star)^2 \geq \frac{\frac{1}{2}\tau(u^\star)^2}{((u^\star)^2 - \tau(u^\star)^2)} \\ \Longleftrightarrow &T \geq \frac{5}{4\gamma(u^\star)^2(1-\tau)} \\ \Longrightarrow &T \geq \frac{5}{2\gamma(u^\star)^2}. \end{aligned}$$

When $u^2(0) > (u^\star)^2$, we have that

$$\begin{aligned} u(t+1) &\leq u(t) + 2\gamma u(t)\left((u^\star)^2 - u^2(t) + \frac{1}{10}\tau(u^\star)^2\right) \\ &= u(t)\left(1 + 2\gamma\left(\left(1 + \frac{1}{10}\tau\right)(u^\star)^2 - u^2(t)\right)\right). \\ &\leq u(t)\left(1 - \frac{4}{5}\gamma\tau(u^\star)^2\right). \end{aligned}$$

Similarly, we want to get

$$\begin{aligned} &(u^\star)^2 + \frac{1}{2}\tau(u^\star)^2 \geq \left((u^\star)^2 + \tau(u^\star)^2\right)\left(1 - \frac{4}{5}\gamma T\tau(u^\star)^2\right) \\ \Longleftrightarrow &\frac{(u^\star)^2 + \frac{1}{2}\tau(u^\star)^2}{(u^\star)^2 + \tau(u^\star)^2} \geq 1 - \frac{4}{5}\gamma T\tau(u^\star)^2 \\ \Longleftrightarrow &\frac{4}{5}\gamma T\tau(u^\star)^2 \geq \frac{\frac{1}{2}\tau(u^\star)^2}{(u^\star)^2 + \tau(u^\star)^2} \\ \Longleftrightarrow &T \geq \frac{5}{8\gamma(u^\star)^2(1+\tau)} \\ \Longrightarrow &T \geq \frac{5}{8\gamma(u^\star)^2}. \end{aligned}$$

If $u(0) <= u^\star$ and $u(t) > u^\star, t < T$, or $u(0) > u^\star$ and $u(t) \leq u^\star, t < T$, we have already have $|u^2(t) - u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$. By Lemma 25, $|u^2(T) - u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$ remains to hold.

Hence, after $T \geq \frac{5}{2\gamma(u^\star)^2}$, we have $|u^2(T) - u^\star)^2| \leq \frac{1}{2}\tau(u^\star)^2$. □

B.3.3 Analysis of perturbations

We decompose the updates into several terms for later investigation.

The gradient of $\mathcal{L}(\cdot)$ on each $\mathbf{v}_l$ is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{v}_l} &= -\frac{1}{n} u_l^2 \mathbf{X}_l^\top \left(\mathbf{y} - \sum_{l' \neq l} u_{l'}^2 \mathbf{X}_{l'} \mathbf{v}_{l'} \right) + \frac{1}{n} u_l^4 \mathbf{X}_l^\top \mathbf{X}_l \mathbf{v}_l \\ &= -\frac{1}{n} u_l^2 \mathbf{X}_l^\top \left(\mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2 \mathbf{X}_{l'} \mathbf{v}_{l'} \right)\end{aligned}$$

When $l \in S$, the gradient update on each $\mathbf{v}_l$ is

$$\begin{aligned}\mathbf{z}_l(t+1) &= \mathbf{v}_l(t) + \eta_{l,t} u_l^2(t) \frac{1}{n} \mathbf{X}_l^\top \left(\mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2(t) \mathbf{X}_{l'} \mathbf{v}_{l'}(t) \right) \\ &= \mathbf{v}_l(t) + \eta_{l,t} u_l^2(t) ((u_l^\star)^2 \mathbf{v}_l^\star - u_l^2(t) \mathbf{v}_l(t)) \\ &+ \eta_{l,t} u_l^2(t) \left(\frac{1}{n} \mathbf{X}_l^\intercal \mathbf{X}_l - \mathbf{I} \right) ((u_l^\star)^2 \mathbf{v}_l^\star - u_l^2(t) \mathbf{v}_l(t)) \\ &+ \eta_{l,t} u_l^2(t) \sum_{l' \neq l, l' \in S} \frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_{l'} ((u_{l'}^\star)^2 \mathbf{v}_{l'}^\star - u_{l'}^2(t) \mathbf{v}_{l'}(t)) \\ &- \eta_{l,t} u_l^2(t) \sum_{l' \in S^c} \frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_{l'} u_{l'}^2(t) \mathbf{v}_{l'}(t) \\ &+ \eta_{l,t} u_l^2(t) \frac{1}{n} \mathbf{X}_l^\top \boldsymbol{\xi}.\end{aligned}$$

The gradient of $\mathcal{L}(\cdot)$ on each u_l is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial u_l} &= -\frac{2}{n} u_l \left\langle \mathbf{X}_l \mathbf{v}_l, \mathbf{y} - \sum_{l' \neq l} u_{l'}^2 \mathbf{X}_{l'} \mathbf{v}_{l'} \right\rangle + \frac{2}{n} u_l^3 \left\| \mathbf{X}_l \mathbf{v}_l \right\|^2 \\ &= -\frac{2}{n} u_l \left\langle \mathbf{X}_l \mathbf{v}_l, \mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2 \mathbf{X}_{l'} \mathbf{v}_{l'} \right\rangle\end{aligned}$$

When $l \in S$, the gradient update on u_l reads

$$
\begin{aligned}
u_l(t+1) &= u_l(t) + \gamma\frac{2}{n}u_l(t)\left\langle \mathbf{X}_l\mathbf{v}_l(t+1), \mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2(t)\mathbf{X}_{l'}\mathbf{v}_{l'}(t+1)\right\rangle \\
&= u_l(t) + 2\gamma u_l(t)\mathbf{v}_l^\top(t+1)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t+1)) \\
&+ 2\gamma u_l(t)\mathbf{v}_l^\top(t+1)\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l - \mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t+1)) \\
&+ 2\gamma u_l(t)\mathbf{v}_l^\top(t+1)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\neq l, l'\in S}\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t+1)) \\
&- 2\gamma u_l(t)\mathbf{v}_l^\top(t+1)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\in S^c}\mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t+1) \\
&+ 2\gamma u_l(t)\frac{1}{n}\mathbf{v}_l^\top(t+1)\mathbf{X}_l^\top\boldsymbol{\xi}.
\end{aligned}
$$

We now rewrite the definition of bounded perturbation in Eq. (B.1, B.2), where the bounded perturbation $e_{l,t}$ on updates of $u_l(t)$ reads

$$
\begin{aligned}
e_{l,t} &= \mathbf{v}_l^\top(t+1)\left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l - \mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t+1)) \\
&+ \mathbf{v}_l^\top(t+1)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\neq l, l'\in S}\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t+1)) \\
&- \mathbf{v}_l^\top(t+1)\frac{1}{n}\mathbf{X}_l^\top\sum_{l'\in S^c}\mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t+1) \\
&+ \frac{1}{n}\mathbf{v}_l^\top(t+1)\mathbf{X}_l^\top\boldsymbol{\xi},
\end{aligned}
$$

and the bounded perturbation $\mathbf{b}_{l,t}$ on updates of $\mathbf{v}_l(t)$ reads

$$
\begin{aligned}
\mathbf{b}_{l,t} &= \left(\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_l - \mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)) \\
&+ \sum_{l'\neq l, l'\in S}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t)) \\
&- \sum_{l'\in S^c}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t) \\
&+ \frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}.
\end{aligned}
$$

We show in Lemma 27 that when the perturbations are bounded, the direction is roughly accurate ($\langle \mathbf{v}_l(t), \mathbf{v}^\star \rangle$ is large) and $u_l(t)$ converges exponentially. Now we show below that when the direction is roughly accurate and $u_l(t)$ is close to $u_l^\star$, the perturbations are bounded.

Lemma 28. *Assume* $\delta_{in} \le \frac{(u^\star_{min})^2}{120(u^\star_{max})^2}$ *and* $\delta_{out} \le \frac{(u^\star_{min})^2}{120s(u^\star_{max})^2}$, $\alpha < \frac{1}{2}\sqrt{\frac{\tau_0}{L}}u_l^\star$, $\left\|\frac{1}{n}\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty \le \frac{1}{80}\tau_0 (u_l^\star)^2$ *and* $|(u_l^\star)^2 - u_l^2(0)| \le \tau (u_l^\star)^2$ *for each* $l \in [L]$ *where* $0 < \tau_0 \le \tau \le 1/2$. *If* $\langle \mathbf{v}_l(t), \mathbf{v}_l^\star \rangle \ge 1 - \frac{1}{5}\tau^2$, *then* $|\langle \mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \le \frac{1}{10}\tau (u_l^\star)^2$ *and* $|e_{l,t}| \le \frac{1}{10}\tau (u_l^\star)^2$.

Proof. We first verify

$$\begin{aligned}
\left\|(u_l^\star)^2 \mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)\right\| &= \left\|\{(u_l^\star)^2 - u_l^2(t)\}\mathbf{v}_l^\star - u_l^2(t)\{\mathbf{v}_l(t) - \mathbf{v}_l^\star\}\right\| \\
&\le |(u_l^\star)^2 - u_l^2(t)| + u_l^2(t)\left\|\mathbf{v}_l(t) - \mathbf{v}_l^\star\right\| \\
&\le \tau (u_l^\star)^2 + u_l^2(t)\sqrt{2 - 2\langle \mathbf{v}_l(t), \mathbf{v}_l^\star\rangle} \\
&\le \tau (u_l^\star)^2 + \frac{3}{2}(u_l^\star)^2 \frac{\sqrt{2}}{\sqrt{5}}\tau \qquad \text{(B.4)} \\
&\le 3\tau (u_l^\star)^2.
\end{aligned}$$

By Assumption 1, we have that

$$\begin{aligned}
&\left|\mathbf{v}_l^\top(t)\left(\frac{1}{n}\mathbf{X}_l^\intercal \mathbf{X}_l - \mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)) + \mathbf{v}_l^\top(t)\sum_{l'\neq l, l'\in S}\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t))\right| \\
&\le 3\delta_{in}\tau (u^\star_{max})^2 + 3s\delta_{out}\tau (u^\star_{max})^2 \le \frac{1}{40}\tau (u_l^\star)^2 + \frac{1}{40}\tau (u_l^\star)^2 = \frac{1}{20}\tau (u_l^\star)^2.
\end{aligned}$$

For the other two terms, we have that

$$\left|\mathbf{v}_l^\top(t)\sum_{l'\in S^c}\frac{1}{n}\mathbf{X}_l^\top \mathbf{X}_{l'}u_{l'}^2(t)\mathbf{v}_{l'}(t)\right| \le \delta(L-s)\alpha^2 \le \frac{1}{80}\tau (u_l^\star)^2,$$

and

$$\begin{aligned}\left|\mathbf{v}_l^\top(t)\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right| &\leq \left\|\mathbf{v}_l^\top(t)\right\|_1 \left\|\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right\|_\infty \\ &\leq \left\|\mathbf{v}_l^\top(t)\right\|_2 \left\|\frac{1}{n}\mathbf{X}_l^\top\boldsymbol{\xi}\right\|_\infty \\ &\leq \frac{1}{80}\tau(u_l^\star)^2.\end{aligned}$$

Therefore,

$$|e_{l,t}| = |\langle\mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \leq \frac{1}{20}\tau(u_l^\star)^2 + \frac{1}{80}\tau(u_l^\star)^2 + \frac{1}{80}\tau(u_l^\star)^2 \leq \frac{1}{10}\tau(u_l^\star)^2.$$

□

Lemma 27 shows that when the upper bound of perturbation is fixed, $u_l(t)$ grows. Now we show that after $u_l(t)$ grows, the upper bound of perturbations will be decreased.

Lemma 29. *Assume* $\delta_{in} \leq \frac{(u_{min}^\star)^2}{120(u_{max}^\star)^2}$ *and* $\delta_{out} \leq \frac{(u_{min}^\star)^2}{120s(u_{max}^\star)^2}$, $\alpha < \frac{\sqrt{\tau_0}}{2\sqrt{L}}u_l^\star$, $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \leq \frac{1}{80}\tau_0(u_l^\star)^2$ *and* $\langle\mathbf{v}_l(t), \mathbf{v}_l^\star\rangle \geq 1 - \frac{1}{5}\tau^2$. *If we achieve that* $|(u_l^\star)^2 - u_l^2(0)| \leq \frac{1}{2}\tau(u_l^\star)^2$ *for each* $l \in [L]$ *where* $0 < \tau_0 \leq \tau \leq 1/2$, *then* $|\langle\mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \leq \frac{1}{20}\tau(u_l^\star)^2$ *and* $|e_{l,t}| \leq \frac{1}{20}\tau(u_l^\star)^2$.

Proof. Similarly to the proof of Lemma 27,

$$\begin{aligned}\left\|(u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)\right\| &\leq \frac{1}{2}\tau(u_l^\star)^2 + u_l^2(t)\sqrt{2 - 2\langle\mathbf{v}_l(t), \mathbf{v}_l^\star\rangle} \\ &\leq \frac{1}{2}\tau(u_l^\star)^2 + \frac{3}{2}(u_l^\star)^2\frac{1}{\sqrt{5}}\tau \\ &\leq \frac{3}{2}\tau(u_l^\star)^2.\end{aligned}$$

By Assumption 1, we have that

$$\left|\mathbf{v}_l^\top(t)\left(\frac{1}{n}\mathbf{X}_l^\intercal\mathbf{X}_l-\mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star-u_l^2(t)\mathbf{v}_l(t))+\mathbf{v}_l^\top(t)\sum_{l'\neq l,l'\in S}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star-u_{l'}^2(t)\mathbf{v}_{l'}(t))\right|$$
$$\leq\frac{3}{2}\delta_{in}\tau(u_{max}^\star)^2+\frac{3}{2}s\delta_{out}\tau(u_{max}^\star)^2\leq\frac{1}{40}\tau(u_l^\star)^2,$$

where $\delta\leq\frac{1}{60s}$. Similarly, we obtain that

$$|e_{l,t}|=|\langle\mathbf{v}_l(t),\mathbf{b}_{l,t}\rangle|\leq\frac{1}{40}\tau(u_l^\star)^2+\frac{1}{80}\tau(u_l^\star)^2+\frac{1}{80}\tau(u_l^\star)^2\leq\frac{1}{20}\tau(u_l^\star)^2.$$

□

By Lemma 26, we know that after certain iterations, we have that $|u^2(t)-(u^\star)^2|\leq\frac{1}{2}(u^\star)^2$. Starting from there, we will apply Lemma 27 and Lemma 28 iteratively until we have our desired accuracy.

We just need to verify when $\tau=\frac{1}{2}$, the condition of either Lemma 27 and Lemma 28 is satisfied. Note that the condition of Lemma 26 already satisfies the condition of Lemma 27 at $\tau=\frac{1}{2}$. Note the condition of Lemma 26 is satisfied when $\delta_{in}\leq\frac{(u_{min}^\star)^2}{120(u_{max}^\star)^2}$ and $\delta_{out}\leq\frac{(u_{min}^\star)^2}{120s(u_{max}^\star)^2}$, $\alpha\leq\frac{1}{4}(u_{min}^\star)^2$, $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty\leq\frac{1}{80}\tau_0(u_{min}^\star)^2$.

B.3.4 Error analysis outside the support

We only care about the growth rate of $u_l(t)$ when $l \notin S$. When $l \in S^c$, the gradient updates on u_l reads

$$\begin{aligned}
u_l(t+1) &= u_l(t) + \gamma \frac{2}{n} u_l(t) \left\langle \mathbf{X}_l \mathbf{v}_l(t), \mathbf{y} - \sum_{l'=1}^{L} u_{l'}^2(t) \mathbf{X}_{l'} \mathbf{v}_{l'}(t) \right\rangle \\
&= u_l(t) - 2\gamma u_l^3(t) \\
&- 2\gamma u_l^3(t) \mathbf{v}_l^\top(t) \left(\frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_l - \mathbf{I} \right) \mathbf{v}_l(t) \\
&+ 2\gamma u_l(t) \mathbf{v}_l^\top(t) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \in S} \mathbf{X}_{l'} ((u_{l'}^\star)^2 \mathbf{v}_{l'}^\star - u_{l'}^2(t) \mathbf{v}_{l'}(t)) \\
&- 2\gamma u_l(t) \mathbf{v}_l^\top(t) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \neq l, l' \in S^c} \mathbf{X}_{l'} u_{l'}^2(t) \mathbf{v}_{l'}(t) \\
&+ 2\gamma u_l(t) \frac{1}{n} \mathbf{v}_l(t) \mathbf{X}_l^\top \boldsymbol{\xi}.
\end{aligned}$$

Consider the initialization is $u_l(0) = \alpha$, we wonder the smallest number t of iterations that we can ensure $u_l(t) \leq \sqrt{\alpha}$. Denote

$$\begin{aligned}
e_{l,t} &= -u_l^2(t) - u_l^2(t) \mathbf{v}_l^\top(t) \left(\frac{1}{n} \mathbf{X}_l^\top \mathbf{X}_l - \mathbf{I} \right) \mathbf{v}_l(t) \\
&+ \mathbf{v}_l^\top(t) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \in S} \mathbf{X}_{l'} ((u_{l'}^\star)^2 \mathbf{v}_{l'}^\star - u_{l'}^2(t) \mathbf{v}_{l'}(t)) \\
&- \mathbf{v}_l^\top(t) \frac{1}{n} \mathbf{X}_l^\top \sum_{l' \neq l, l' \in S^c} \mathbf{X}_{l'} u_{l'}^2(t) \mathbf{v}_{l'}(t) \\
&+ \frac{1}{n} \mathbf{v}_l^\top(t) \mathbf{X}_l^\top \boldsymbol{\xi}.
\end{aligned}$$

We have that

$$|e_{l,t}| \leq \alpha + \alpha\delta_{in} + \alpha\delta_{out}(L-s) + \frac{3}{2}(u_{max}^\star)^2 \delta_{out} s + \left\| \frac{1}{n} \mathbf{X}_l^\top \boldsymbol{\xi} \right\|_\infty .$$

If $\alpha \leq \frac{1}{80L}(u^\star_{min})^2$, $\delta_{in} \leq \frac{(u^\star_{min})^2}{120(u^\star_{max})^2}$ and $\delta_{out} \leq \frac{(u^\star_{min})^2}{120s(u^\star_{max})^2}$, we have that

$$|e_{l,t}| \leq \frac{1}{20}(u^\star_{min})^2 + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_\infty. \tag{B.5}$$

Lemma 30. *Consider*

$$u(t+1) = u(t)(1 + 2\gamma e_t)$$

where $|e_t| \leq B$ *and* $u(0) = \alpha$. *Let the step size* $\gamma \leq \frac{1}{4B}$, *then for any* $t \leq T = \frac{1}{32\gamma B}\log\frac{1}{\alpha^4}$, *we have* $u(t) \leq \sqrt{u(0)}$.

Proof. We start by observing,

$$\begin{aligned}\sqrt{\alpha} \geq u(t) \geq \alpha(1+2\gamma B)^t \\ \iff t \leq \frac{\log\frac{1}{\sqrt{\alpha}}}{\log(1+2\gamma B)}.\end{aligned}$$

By using $\log x \leq x - 1$,

$$\frac{\log\frac{1}{\sqrt{\alpha}}}{\log(1+2\gamma B)} \geq \frac{1}{2\gamma B}\log\frac{1}{\sqrt{\alpha}} \geq \frac{1}{32\gamma B}\log\frac{1}{\alpha^4}.$$

□

B.4 Proof of Theorems in Chapter 3.4

B.4.1 Proof of Theorem 5

Proof. If $\zeta \geq (u^\star_{max})^2$, at the initialization, we already have for $\forall l \in [L]$

$$\begin{aligned}\left\|u_l^2(0)\mathbf{v}_l(0) - (u_l^\star)^2\mathbf{v}_l^\star\right\|_\infty &\leq u_l^2(0) + (u_l^\star)^2 \leq \alpha^2 + (u^\star_{\max})^2 \\ &\leq 2(u^\star_{\max})^2 \leq 2\zeta \\ &\leq 160\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \vee 160\epsilon.\end{aligned}$$

If $\zeta \leq (u^\star_{max})^2$, for those $l \in S$ such that $\zeta \leq (u^\star_l)^2$, we can apply Lemma 26. After

$$T_1 = \frac{\log \frac{(u^\star_l)^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(u^\star_l)^2)},$$

we obtain that $\frac{1}{2}(u^\star_l)^2 \leq u_l^2(T_1) \leq \frac{3}{2}(u^\star_l)^2$, where we also have that $\left\|\frac{1}{n}\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty \leq \frac{1}{80}(u^\star_l)^2$ for every l.

Let m_0 be the number s.t.

$$2^{-m_0-1}(u^\star_{max})^2 \leq \zeta \leq 2^{-m_0}(u^\star_{max})^2,$$

which can be written as $m_0 = \lfloor \log_2 \frac{(u^\star_{max})^2}{\zeta} \rfloor$. We can apply Lemma 27 and Lemma 28 together m_0 times. Then further after

$$T_2 = \lfloor \log_2 \frac{(u^\star_{max})^2}{\zeta} \rfloor \frac{5}{2\gamma(u^\star_l)^2},$$

we have that

$$\begin{aligned} |u_l^2(T_2) - (u^\star_l)^2| &\leq 2^{-m_0}(u^\star_{max})^2 \leq 2\zeta \\ \langle \mathbf{v}_l(T_2), \mathbf{v}^\star_l \rangle &\geq 1 - \frac{1}{5}2^{-2m_0}. \end{aligned}$$

Therefore,

$$\begin{aligned} \left\|u_l^2(T_2)\mathbf{v}_l(T_2) - (u^\star_l)^2\mathbf{v}^\star_l\right\|_\infty &\leq \left\|u_l^2(T_2)\mathbf{v}_l(T_2) - (u^\star_l)^2\mathbf{v}^\star_l\right\|_2 \\ &\leq \left\|(u_l^2(T_2) - (u^\star_l)^2)\mathbf{v}_l(T_2) - (u^\star_l)^2(\mathbf{v}^\star_l - \mathbf{v}_l(T_2))\right\|_2 \\ &\leq 2^{-m_0}(u^\star_{max})^2 + (u^\star_l)^2\sqrt{2 - 2\langle \mathbf{v}_l(T_2), \mathbf{v}^\star_l\rangle} \\ &\leq 2^{-m_0}(u^\star_{max})^2 + (u^\star_l)^2\frac{2}{5}2^{-m_0} \\ &\leq 2\zeta. \end{aligned} \tag{B.6}$$

Note that the above inequality holds for every $l \in S$ such that $(u^\star_l)^2 \geq \zeta$. For those l such that $\zeta \geq (u^\star_l)^2$, we are not able to recover the true signal $(u^\star_l)^2$. the gradient dynamics on this group

behaves as errors outside group, and bounded by Lemma 30.

For entries outside the support, we know that from Eq. (B.5),

$$B = \frac{1}{20}(u^\star_{min})^2 + \left\|\frac{1}{n}\mathbf{X}_l^\top \boldsymbol{\xi}\right\|_\infty \leq \frac{1}{10}(\zeta \vee (u^\star_{min})^2).$$

By Lemma 30, we have that before $T_3 \leq \frac{1}{32\gamma B}\log\frac{1}{\alpha^4}$, $u_l(T_3) \leq \sqrt{\alpha}$.

When $\zeta \leq (u^\star_{min})^2$, Eq. (B.6) holds for every $l \in S$. Therefore, a uniform number of iterations T_1 and T_2 for all groups is written as

$$T_1 = \frac{\log\frac{(u^\star_{max})^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(\zeta \vee (u^\star_{min})^2))},$$

and

$$T_2 = \lfloor \log_2 \frac{(u^\star_{max})^2}{\zeta} \rfloor \frac{5}{2\gamma(\zeta \vee (u^\star_{min})^2)}.$$

All we left is to show that $T_3 \geq T_1 + T_2$. We observe that

$$\begin{aligned} T_1 = \frac{\log\frac{(u^\star_{max})^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(\zeta \vee (u^\star_{min})^2))} &\leq \frac{1+\gamma\frac{1}{2}(\zeta \vee (u^\star_{min})^2))}{\gamma(\zeta \vee (u^\star_{min})^2))}\log\frac{(u^\star_{max})^2}{2\alpha^2} \\ &\leq \frac{2}{\gamma(\zeta \vee (u^\star_{min})^2)}\log\frac{(u^\star_{max})^2}{2\alpha^2} \end{aligned}$$

where the first inequality is by $\log x \geq \frac{x-1}{x}$.

With our choice of small initialization on α, we have $T_1 \leq \frac{1}{2}T_3$, due to $\alpha < \frac{1}{(u^\star_{max})^8}$. We have $T_2 \leq \frac{1}{2}T_3$, because of $\alpha < \frac{\zeta^4}{(u_{max})^8}$.

Hence, we obtain that after $T_l = T_1 + T_2 \geq \frac{\log\frac{(u^\star_{max})^2}{2\alpha^2}}{2\log(1+\gamma\frac{1}{2}(\zeta\vee(u^\star_{min})^2))} + \lfloor \log_2 \frac{(u^\star_{max})^2}{\zeta} \rfloor \frac{5}{2\gamma(\zeta\vee(u^\star_{min})^2)}$, and before $T_u = T_3 \leq \frac{5}{16\gamma(\zeta\vee(u^\star_{min})^2)}\log\frac{1}{\alpha^4}$,

$$\left\|u_l^2(t)\mathbf{v}_l(t) - (u_l^\star)^2\mathbf{v}_l^\star\right\|_\infty \lesssim \begin{cases} \left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \vee \epsilon, & \text{if } l \in S. \\ \alpha, & \text{if } l \notin S. \end{cases}$$

□

B.4.2 Proof for Corollary 2

Since $\boldsymbol{\xi}$ is made of independent σ^2-sub-Gaussian entries, by Lemma 6 with probability $1 - 1/(8p^3)$ we have

$$\left\| \frac{1}{n} \mathbf{X}^\top \boldsymbol{\xi} \right\|_\infty \leq 2\sqrt{\frac{2\sigma^2 \log(2p)}{n}}.$$

Hence, letting $\epsilon = 2\sqrt{\frac{2\sigma^2 \log(2p)}{n}}$, we obtain that

$$\left\| (\mathbf{D}\mathbf{u}(t))^2 \odot \mathbf{v}(t) - \mathbf{w}^\star \right\|_2^2 \lesssim \sum_{l \in S} \epsilon^2 + \sum_{l \notin S} \alpha \leq s\epsilon^2 + (L - s)\frac{\epsilon^2}{L^2} \lesssim \frac{s\sigma^2 \log p}{n}.$$

□

B.4.3 Convergence for algorithm 2

Lemma 31. *Consider the update in Eq.* (B.1)*, choose the step size* $\eta_t = \eta \leq \frac{4}{9(u^\star)^4}$*, if* $\langle \mathbf{v}(t), \mathbf{v}^\star \rangle \geq 1 - \frac{1}{5}\tau$*,* $|u^2(t) - (u^\star)^2| \leq \tau(u^\star)^2$ *and* $\|\mathbf{b}_t\| \leq \frac{1}{10}\tau(u^\star)^2$ *for some constant* $0 < \tau < \frac{1}{2}$*, we have that*

$$\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq 1 - \frac{1}{5}\tau.$$

Proof. We first rewrite $\mathbf{z}(t+1)$ as

$$\mathbf{z}(t+1) = \eta u^2(t)(u^\star)^2 \mathbf{v}^\star + (1 - \eta u^4(t))\mathbf{v}(t) + \eta u^2(t)\mathbf{b}_t.$$

Therefore,

$$\begin{aligned}
\langle \mathbf{z}(t+1), \mathbf{v}^\star \rangle &\geq \eta u^2(t)(u^\star)^2 + (1 - \eta u^4(t))\langle \mathbf{v}(t), \mathbf{v}^\star \rangle + \eta u^2(t)\langle \mathbf{b}_t, \mathbf{v}^\star \rangle \\
&\geq \eta u^2(t)(u^\star)^2 + (1 - \eta u^4(t))\left(1 - \frac{1}{5}\tau\right) - \eta u^2(t)\frac{1}{10}\tau(u^\star)^2 \\
\|\mathbf{z}(t+1)\| &\leq \eta u^2(t)(u^\star)^2 + (1 - \eta u^4(t)) + \eta u^2(t)\frac{1}{10}\tau(u^\star)^2.
\end{aligned}$$

We obtain that

$$\begin{aligned}\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle = \frac{\langle \mathbf{z}(t+1), \mathbf{v}^\star \rangle}{\|\mathbf{z}(t+1)\|} &\geq 1 - \frac{\frac{1}{5}\tau(1-\eta u^4(t)) + 2\eta u^2(t)\frac{1}{10}\tau(u^\star)^2}{\eta u^2(t)(u^\star)^2 + (1-\eta u^4(t)) + \eta u^2(t)\frac{1}{10}\tau(u^\star)^2} \\ &\geq 1 - \frac{1-\eta u^4(t)+\eta u^2(t)(u^\star)^2}{\eta u^2(t)(u^\star)^2 + (1-\eta u^4(t)) + \eta u^2(t)\frac{1}{10}\tau(u^\star)^2}\frac{1}{5}\tau \\ &\geq 1 - \frac{1}{5}\tau.\end{aligned}$$

□

Note that compared with Lemma 24, under the condition $\|\mathbf{b}_t\| \leq B(u^\star)^2$, we get $\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq 1 - B$ instead of $\langle \mathbf{v}(t+1), \mathbf{v}^\star \rangle \geq 1 - B^2$. Accordingly, we need to a new version for Lemma 28 with a smaller bound on δ to make up the loss in Lemma 31.

Lemma 32. *Assume* $\delta_{in} \leq \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120(u^\star_{max})^2}$ *and* $\delta_{out} \leq \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120s(u^\star_{max})^2}$, $\alpha < \frac{1}{2}\sqrt{\frac{\tau_0}{L}}u_l^\star$, $\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \leq \frac{1}{80}\tau_0(u_l^\star)^2$ *and* $|(u_l^\star)^2 - u_l^2(0)| \leq \tau(u_l^\star)^2$ *for each* $l \in [L]$ *where* $0 < \tau_0 \leq \tau \leq 1/2$. *If* $\langle \mathbf{v}_l(t), \mathbf{v}_l^\star \rangle \geq 1 - \frac{1}{5}\tau$, *then* $|\langle \mathbf{v}_l(t), \mathbf{b}_{l,t}\rangle| \leq \frac{1}{10}\tau(u_l^\star)^2$ *and* $|e_{l,t}| \leq \frac{1}{10}\tau(u_l^\star)^2$.

Proof. Similarly to Lemma 28, we have that

$$\begin{aligned}\left\|(u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)\right\| &\leq \tau(u_l^\star)^2 + u_l^2(t)\sqrt{2-2\langle \mathbf{v}_l(t), \mathbf{v}_l^\star\rangle} \\ &\leq \tau(u_l^\star)^2 + \frac{3}{2}(u_l^\star)^2\frac{\sqrt{2}}{\sqrt{5}}\sqrt{\tau} \\ &\leq \left(1 + 2\frac{1}{\sqrt{\tau_0}}\right)\tau(u_l^\star)^2.\end{aligned} \tag{B.7}$$

By Assumption 1, we have that

$$\begin{aligned}&\left|\mathbf{v}_l^\top(t)\left(\frac{1}{n}\mathbf{X}_l^\intercal\mathbf{X}_l - \mathbf{I}\right)((u_l^\star)^2\mathbf{v}_l^\star - u_l^2(t)\mathbf{v}_l(t)) + \mathbf{v}_l^\top(t)\sum_{l'\neq l, l'\in S}\frac{1}{n}\mathbf{X}_l^\top\mathbf{X}_{l'}((u_{l'}^\star)^2\mathbf{v}_{l'}^\star - u_{l'}^2(t)\mathbf{v}_{l'}(t))\right| \\ &\leq \left(1 + 2\frac{1}{\sqrt{\tau_0}}\right)\delta_{in}\tau(u^\star_{max})^2 + \left(1 + 2\frac{1}{\sqrt{\tau_0}}\right)s\delta_{out}\tau(u^\star_{max})^2 \leq \frac{1}{20}\tau(u_l^\star)^2,\end{aligned}$$

where $\delta \leq \frac{\sqrt{\tau_0}(u^\star_{min})^2}{60s(u^\star_{max})^2}$. The other two terms follows exactly what we did in Lemma 28. Therefore,

$$|e_{l,t}| = |\langle \mathbf{v}_l(t), \mathbf{b}_{l,t} \rangle| \leq \frac{1}{20}\tau(u^\star_l)^2 + \frac{1}{80}\tau(u^\star_l)^2 + \frac{1}{80}\tau(u^\star_l)^2 \leq \frac{1}{10}\tau(u^\star_l)^2.$$

□

Proof to Theorem 6. The proof is similar to that of Theorem 5. For the first stage, we apply Lemma 26, as nothing is changed from Theorem 5. For the second stage, instead of applying Lemma 27 and Lemma 28, we apply Lemma 31 and Lemma 32 iteratively. To apply these lemmas, we first observe that

$$\zeta \leq \tau_0(u^\star_{max})^2 \iff \frac{\zeta}{(u^\star_{max})^2} \leq \tau_0.$$

Therefore the requirement on δ's becomes $\delta_{in} \leq \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120(u^\star_{max})^3}$ and $\delta_{out} \leq \frac{\sqrt{\tau_0}(u^\star_{min})^2}{120s(u^\star_{max})^3}$. The number of iterations and convergence results follow from the proof of Theorem 5.

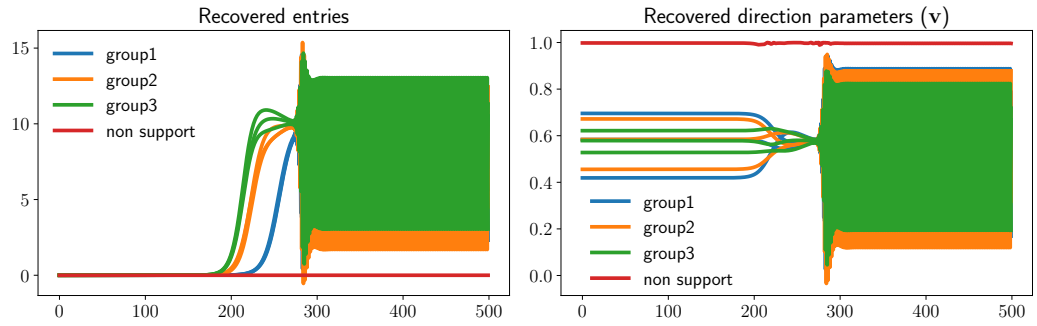
□

The criterion for switching time. We provide some motivation for the practical criterion. We first note that, the criterion in Theorem 6 actually indicates a lower bound of switching time. With more derivations, our results still hold if one choose to switch after the time when the criterion is first satisfied (instead of switching right at that time.) Let us focus on the entries on the support. In the proof of Theorem 5, one can also obtain the convergence on $u_l(t)$ as the positiveness of $u_l(t)$ can be ensured with a small step size γ (since the power-parametrization will recast the gradient updates into a multiplicative sequence). Therefore, with an appropriate choice of τ, the practical criterion $\max_{l \in S}\{|u_l(t+1) - u_l(t)|/|u_l(t) + \varepsilon|\} < \tau$ would imply the theoretical criterion $u_l(t)^2 \geq \frac{1}{2}u^\star_l(t)^2$ on the support, and therefore would indicate a possibly later switching time than what the theoretical criterion determines. For gradient updates outside the support, we observe slow growth rate and hence the practical rule is likely satisfied on the non-support entry, which we

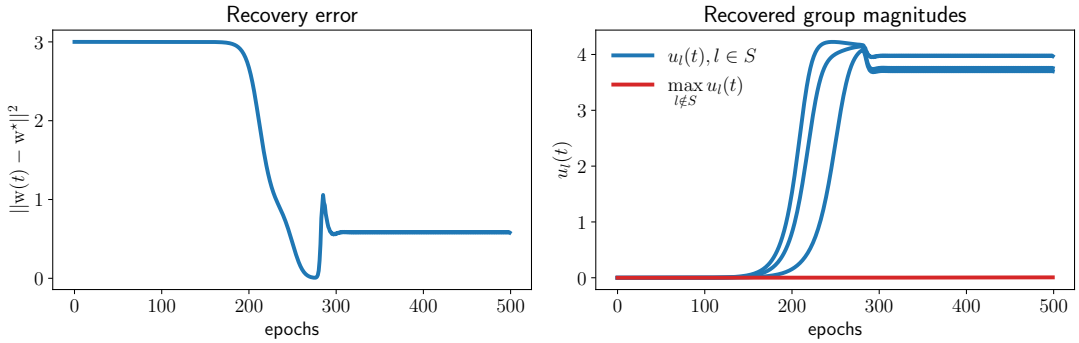
observe in the numerical experiments. Note that the switching only happens when both the support and non-support entries fulfill the criterion.

B.5 More numerical results

B.5.1 Stability issue of Algorithm 1 and standard GD



(a) Numerical instability in direction estimations.



(b) Parameter estimation error remains small.

Figure B.1: Numerical instability of algorithm 1. Reprinted with permission from [2].

Stability issue of Algorithm 1. Figure B.1 presents the recovered entries and direction parameters $\mathbf{v}(t)$ under the same setting as Figure 3.2. Because of the large learning rate on $\mathbf{v}$, the algorithm may not show a convergent result in the latter stage due to the irreducible error (perturbations). Although the parameter estimation is still reasonable with normalization on each $\mathbf{v}_l, l \in [L]$, we still aim to get a stable algorithm, which motivates our algorithm 2.

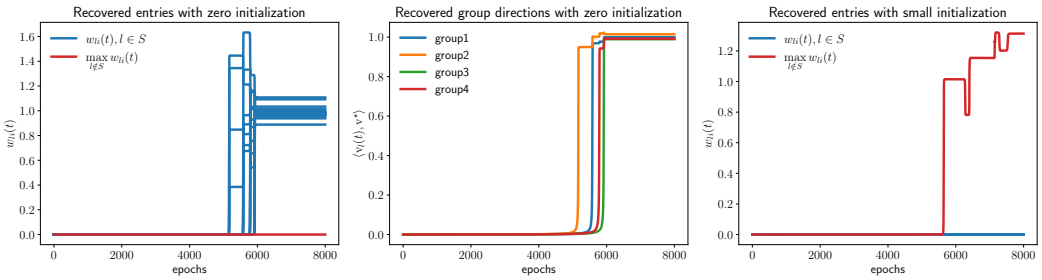


Figure B.2: Gradient descent without weight normalization. Reprinted with permission from [2].

Standard gradient descent. To further understand how weight normalization affects the gradient dynamics, we conduct experiments using standard gradient descent without weight normalization. For that, we use the same setting as in Figure 3.4 and show the result in Figure B.2. The left and middle figures are based on zero initialization on $\mathbf{v}$. We see a numerically convergent result, and the inner product between learned and true directions starts to grow from 0. As the directions guide the magnitude to grow, there is an extra stage for the directions to become roughly accurate. The choice of this initialization is necessary and subtle. The figure on the right is for small initialization 10^{-3}, where the entries outside support get significant magnitudes, and the algorithm fails.

B.5.2 Autoencoder with grouping layer

The grouping layers have been used in grouped CNN and grouped attention mechanisms [92, 91, 120], which usually leads to parameter efficiency and better accuracy. To demonstrate the practical value of such grouping layers, we conduct the following experiment about learning good representations on MNIST.

[121] proposed implicit rank-minimizing autoencoder (IRMAE), which is a deterministic autoencoder with implicit regularization. The idea is to apply more linear layers between encoder and decoder to penalize the rank of latent representation. A graphical illustration of the architecture is shown in Figure B.3, where we explicitly show the last convolution layer and the linear layers in the latent space, which are absorbed into the last layer of the encoder in practice. This

design is related to the power parametrization [84] trick to promote sparsity/low-rankness. One major advantage is that IRMAE produces a more interpretable latent representation, and the linear interpolation in the latent space gives a natural transition between two images.

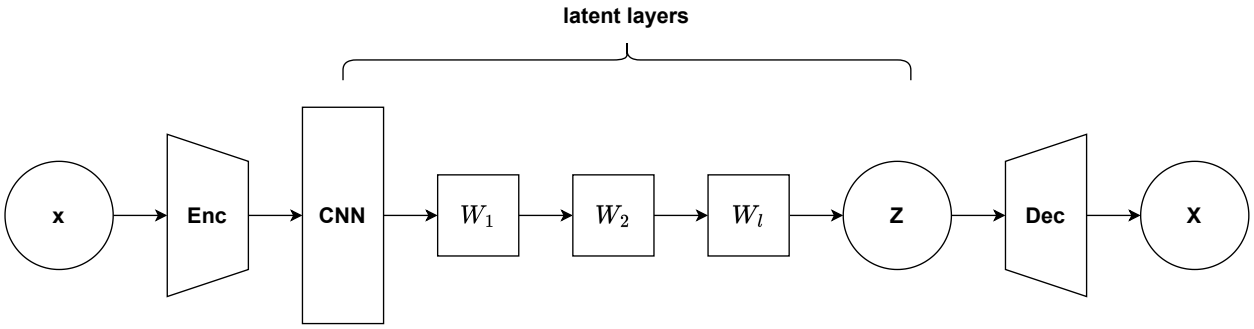


Figure B.3: Implicit rank-minimizing autoencoder. Reprinted with permission from [2].

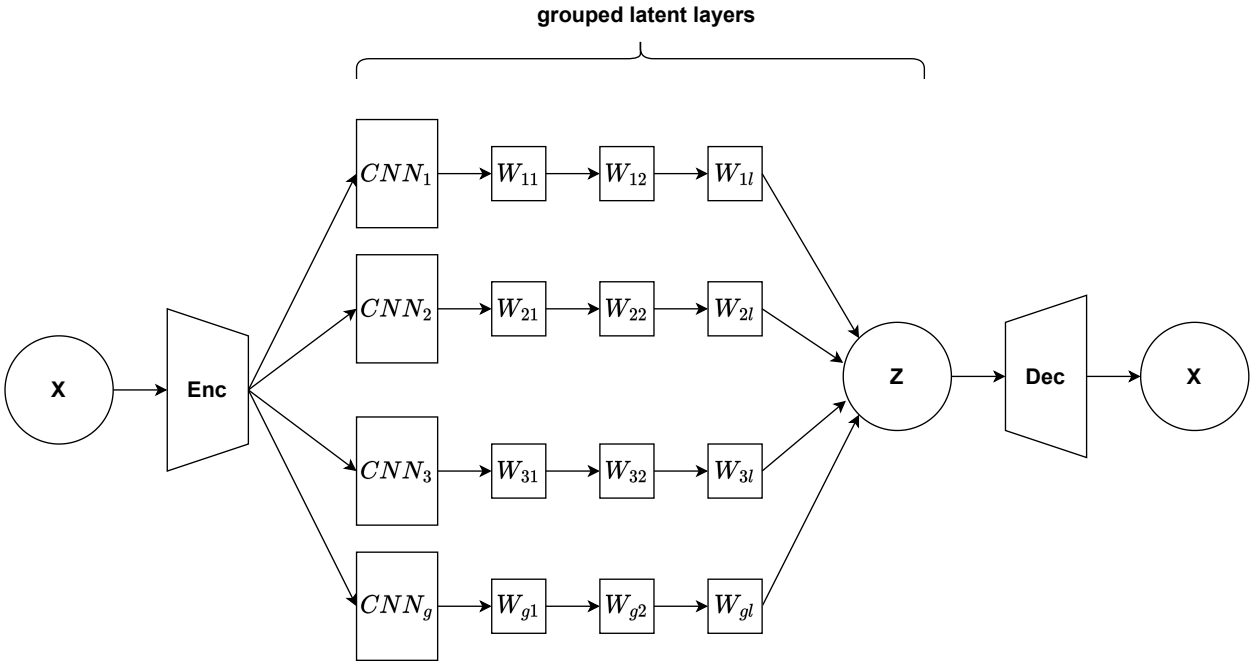


Figure B.4: Implicit rank-minimizing autoencoder with grouping layers. Reprinted with permission from [2].

Inspired by our DGLNN, we design a CNN analog of it, which we call grouped autoencoder (GAE). The architecture is shown in Figure B.4. The channels feed into the last convolutional layer of encoder is separable into g groups. The linear layers (power-parametrization) are applied within each group. Grouping channels of convolutional layers is a common practice to improve the parameter efficiency. With these grouping and power layers in the latent space, we expect it learns a better latent representation as IRMAE does.

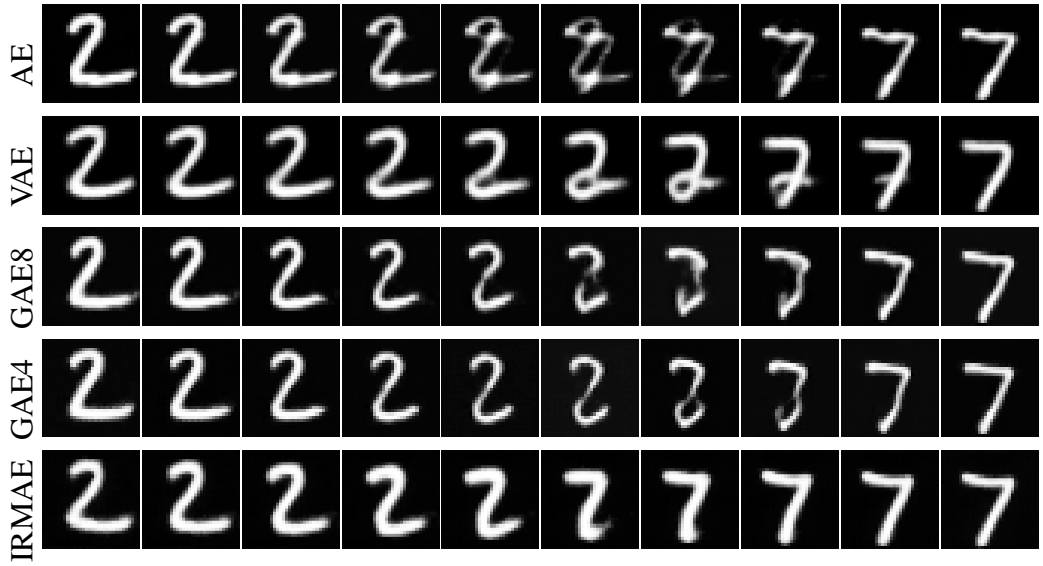


Figure B.5: Linear interpolations between data points on the MNIST dataset. GAE4/8 stands for grouped autoencoder with 4/8 groups. Reprinted with permission from [2].

The linear interpolations between data points in the latent space are shown in Figure B.5. We compare the grouped autoencoder (GAE) with autoencoder (AE), variantional autoencoder (VAE) and implicit rank-minimizing autoencoder (IRMAE). We see that GAE outperforms AE and VAE, and gives comparable results with IRMAE. However, GAE achieves a better parameter efficiency as shown in Table B.1.

	# of params
IRMAE	786K
GAE4	196K
GAE8	98K

Table B.1: Number of parameters of hidden layers in latent space. Reprinted with permission from [2].

B.5.3 Experiments with Gaussian measurements

Besides the numerical results shown in Section 3.5, we conduct the following experiments with sampling each entry of $\mathbf{X}$ from a standard normal distribution.

The effectiveness using Gaussian design. We follow the same setting with that Figure 3.3 except changing Rademacher random variables to Gaussian random variables. The convergence of Algorithm 2 is shown in Figure B.6. We see that the recovered entries, group magnitudes and directions successfully converge to the true ones.

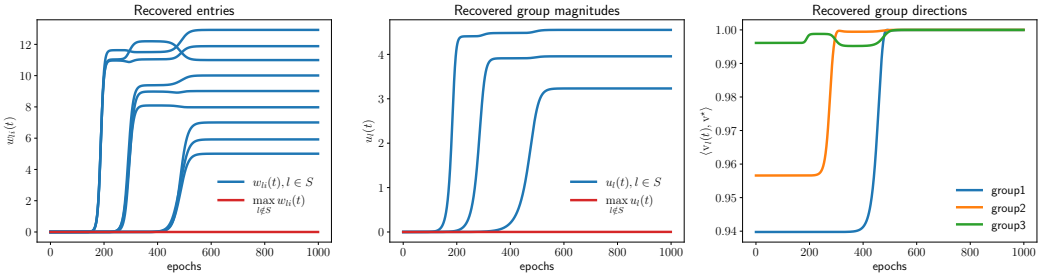


Figure B.6: Convergence of algorithm 2 with Gaussian measurements. Reprinted with permission from [2].

Comparisons with explicit regularization methods. We compare Algorithm 2 with proximal gradient descent implemented in [122] and primal-dual procedure [123]. Each entry of $\mathbf{X}$ is sampled from a standard Gaussian distribution. We set $n = 150$ and $p = 300$, and the number of non-zero entries is 10, divided into 3 groups with size 4. We vary the variance in the noise to achieve different signal-to-noise ratios (SNR). The experiment is repeated 30 times at each noise

level. The average and standard deviation of the estimation error are depicted in Figure B.7. Our algorithm is consistently better than explicit regularization methods, whereas the primal-dual procedure has a comparable performance when SNR is large.

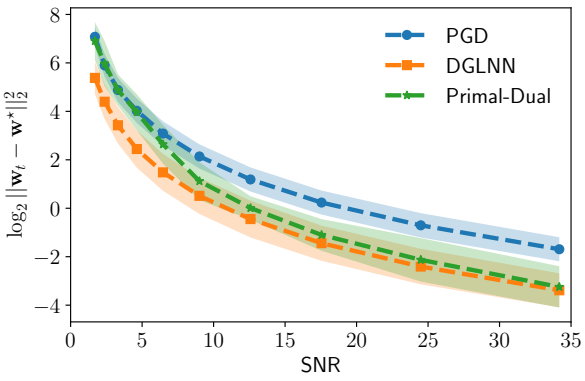


Figure B.7: Comparisons with proximal gradient descent and iterative regularization. Reprinted with permission from [2].

To further discover the potential applications of our findings, we use a gene expression dataset from the Microarray experiments of mammalian eye tissue samples [124]. The dataset consists of 120 samples with 100 predictors that are expanded from 20 genes using 5 basis B-splines, as described in [125]. The goal is to predict the gene expression level of TRIM32, which causes Bardet-Biedl syndrome. We randomly split the data equally, and use the validation dataset for hyperparameter tuning and early stopping. We compare our approach with the commonly used proximal gradient descent and a primal-dual approach. The result is shown in Table B.2. Our approach achieves the best performance among these three methods.

Test error	PGD	Primal-Dual	Our approach
MSE	0.03096	0.02868	0.02477

Table B.2: Comparisons of MSE (mean squared error) on test set. Reprinted with permission from [2].

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER IV

C.1 Proof of Theorem 7

Proof. It follows from the definition of the estimator $\widehat{\mathbf{A}}$ that

$$\ell(\widehat{\mathbf{A}}) + \lambda \left\|\widehat{\mathbf{A}}\right\|_\star \leq \ell(\mathbf{A}_\star) + \lambda \left\|\mathbf{A}_\star\right\|_\star,$$

equivalently

$$\ell(\widehat{\mathbf{A}}) - \ell(\mathbf{A}_\star) \leq \lambda \left(\left\|\mathbf{A}_\star\right\|_\star - \left\|\widehat{\mathbf{A}}\right\|_\star\right)$$

which implies

$$\langle \nabla \ell(\mathbf{A}_\star), \mathrm{vec}(\widehat{\mathbf{A}} - \mathbf{A}_\star)\rangle + \mathrm{vec}(\widehat{\mathbf{A}} - \mathbf{A}_\star)^\top \nabla^2 \ell(\widetilde{\mathbf{A}}) \mathrm{vec}(\widehat{\mathbf{A}} - \mathbf{A}_\star) \leq \lambda \left(\left\|\mathbf{A}_\star\right\|_\star - \left\|\widehat{\mathbf{A}}\right\|_\star\right).$$

where $\widetilde{\mathbf{A}} = t\mathbf{A}_\star + (1-t)\widehat{\mathbf{A}}$ for some $t \in [0,1]$.

Let's denote $\boldsymbol{\Delta} = \widehat{\mathbf{A}} - \mathbf{A}_\star$. We first observe that

$$\left\|\widehat{\mathbf{A}}\right\|_\star = \left\|\mathbf{A}_\star + \boldsymbol{\Delta}_{\overline{\mathcal{M}}} + \boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\right\|_\star \geq \left\|\mathbf{A}_\star + \boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\right\|_\star - \left\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}\right\|_\star = \left\|\mathbf{A}_\star\right\|_\star + \left\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\right\|_\star - \left\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}\right\|_\star.$$

By choosing $\lambda \geq 2\left\|\nabla \ell(\mathbf{A}_\star)\right\|$, we have

$$
\begin{aligned}
\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top\nabla^2\ell(\widetilde{\mathbf{A}})\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star) &\le \lambda\left(\|\mathbf{A}_\star\|_\star-\left\|\widehat{\mathbf{A}}\right\|\right)+\|\nabla\ell(\mathbf{A}_\star)\|\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_\star \\
&\le \lambda\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}\|_\star-\lambda\left\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\right\|_\star+\frac{\lambda}{2}\left\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}+\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\right\|_\star \\
&\le \frac{3\lambda}{2}\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}\|_\star-\frac{\lambda}{2}\left\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}\right\|_\star \\
&\le \frac{3\lambda}{2}\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}\|_\star \\
&\le \frac{3\lambda\psi(\overline{\mathcal{M}})}{2}\left\|\mathbf{A}_\star-\widehat{\mathbf{A}}_{\overline{\mathcal{M}}}\right\|_F\le\frac{3\lambda\psi(\overline{\mathcal{M}})}{2}\left\|\mathbf{A}_\star-\widehat{\mathbf{A}}\right\|_F \\
&\le \frac{3\lambda\sqrt{2r}}{2}\left\|\mathbf{A}_\star-\widehat{\mathbf{A}}\right\|_F.
\end{aligned}
$$

Case 1: when $\mathbb{E}\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top\nabla^2\mathcal{L}_n(\tilde{\mathbf{A}})\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)\le\sqrt{\frac{192(\log(d)+\log(n))}{\log(6/5)n}}$, from Lemma 39 we have that

$$
\kappa_M\frac{2}{\tau_l^2m_1m_2}\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_F^2\le\mathbb{E}\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top\nabla^2\mathcal{L}_n(\tilde{\mathbf{A}})\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)\le\sqrt{\frac{192(\log(d)+\log(n))}{\log(6/5)n}},
$$

i.e.

$$
\frac{1}{m_1m_2}\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_F^2\le c\frac{\tau_l^2}{2\kappa_M}\sqrt{\frac{\log(d)}{n}},
$$

where $c=\sqrt{\frac{192}{\log(6/5)}}$.

Case 2: when $\mathbb{E}\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top\nabla^2\mathcal{L}_n(\tilde{\mathbf{A}})\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)>c\sqrt{\frac{\log(d)+\log(n)}{n}}$, by Lemma 35, we have that

$$
\begin{aligned}
\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top\nabla^2\ell(\mathbf{A})\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)\ge\mathbb{E}\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top\nabla^2\ell(\mathbf{A})\operatorname{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top \\
-1024B^2r(2a)^2\tau_l^2m_1m_2/\kappa_M(\mathbb{E}\|\Sigma_{g,R}\|)^2
\end{aligned}
$$

where the constrain set is $\mathcal{C}(2r)$ and $\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_\infty \leq 2a$. Therefore, we have that

$$
\begin{aligned}
&\kappa_M \frac{2}{\tau_l^2 m_1 m_2}\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_F^2 - 1024B^2 r(2a)^2\tau_l^2 m_1 m_2/\kappa_M(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2 \\
\leq &\mathbb{E}\mathrm{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top \nabla^2\ell(\mathbf{A})\mathrm{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top - 1024B^2 r(2a)^2\tau_l^2 m_1 m_2/\kappa_M(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2 \\
\leq &\mathrm{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star)^\top \nabla^2\ell(\mathbf{A})\mathrm{vec}(\widehat{\mathbf{A}}-\mathbf{A}_\star) \\
\leq &\frac{3\lambda\sqrt{2r}}{2}\left\|\mathbf{A}_\star-\widehat{\mathbf{A}}\right\|_F.
\end{aligned}
$$

After some simplification,

$$
\begin{aligned}
\frac{2}{m_1 m_2}\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_F^2 &\leq \frac{3\tau_l^2\lambda\sqrt{2r}}{2\kappa_M}\left\|\mathbf{A}_\star-\widehat{\mathbf{A}}\right\|_F + 1024B^2 r(2a)^2\tau_l^4 m_1 m_2/\kappa_M^2(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2 \\
&\leq \frac{9\tau_l^4\lambda^2 r}{8\kappa_M^2} m_1 m_2 + \frac{1}{m_1 m_2}\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_F^2 \\
&\quad + 1024B^2 r(2a)^2\tau_l^4 m_1 m_2/\kappa_M^2(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{1}{m_1 m_2}\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_F^2 &\leq \frac{9\tau_l^4\lambda^2 r}{8\kappa_M^2} m_1 m_2 + 1024B^2 r(2a)^2\tau_l^4 m_1 m_2/\kappa_M^2(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2 \\
&\leq \frac{1}{\kappa_M^2} 4196 r\tau_l^4 m_1 m_2 \max\{\lambda^2, B^2 a^2(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2\}.
\end{aligned}
$$

Overall,

$$
\begin{aligned}
\frac{1}{m_1 m_2}\left\|\widehat{\mathbf{A}}-\mathbf{A}_\star\right\|_F^2 \leq \max\Bigg\{ &\frac{1}{\kappa_M^2} 4196 r\tau_l^4 m_1 m_2 \max\{\lambda^2, B^2 a^2(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2\}, \\
&c\frac{\tau_l^2 a^2}{\kappa_M}\sqrt{\frac{\log(d)}{n}}\Bigg\}.
\end{aligned}
$$

□

C.2 Proof of Corollary 7

Proof. By Lemma 34, there exist some constant $c > 0$, with probability $\frac{1}{d}$, when $\frac{3c}{128} Lm \log(d) \leq n \leq m_1 m_2$, we are able to choose

$$2 \left\| \nabla \ell(\mathbf{A}) \right\| \leq \lambda \leq cB \sqrt{L \frac{\log(d)}{mn}}.$$

By Lemma 38, we have that

$$\left(\mathbb{E} \left\| \Sigma_{g,R} \right\| \right)^2 \leq C^\star \frac{L \log(d)}{nm}.$$

Therefore, there exists some constant $c' > 0$ such that

$$\begin{aligned} \frac{1}{m_1 m_2} \left\| \widehat{\mathbf{A}} - \frac{\mathbf{A}_\star}{\phi} \right\|_F^2 \leq & c' \max \Bigg\{ \frac{r \tau_l^4 B^2 L}{\kappa_M^2} \frac{M \log(d)}{n} \max\{1, a^2\}, \\ & \frac{\tau_l^2 a^2}{\kappa_M} \sqrt{\frac{\log(d)}{n}} \Bigg\}. \end{aligned}$$

□

C.3 Useful lemmas

Lemma 33.

$$\mathbb{E}\{\nabla \ell(\mathbf{A})\} = 0$$

Proof. Denote

$$R_{kk'}(\mathbf{A}) = \exp(-(Y_k - Y_{k'}) \langle \boldsymbol{X}_k - \boldsymbol{X}_{k'}, \mathbf{A} \rangle).$$

Note that

$$\nabla \ell(\mathbf{A}) = -\binom{n}{2}^{-1} \sum_{1 \leq k < k' \leq n} \frac{R_{kk'}(\mathbf{A})}{1 + R_{kk'}(\mathbf{A})} (Y_k - Y_{k'})(\boldsymbol{X}_k - \boldsymbol{X}_{k'}).$$

Following the same procedure in [47], we are able to show the expectation of gradient is 0. □

Lemma 34. *We choose c s.t., $3c/512B^2 \geq 4$. When $n \geq \frac{3c}{128} Lm\log(d)$,*

$$\Pr\left\{\|\nabla \ell(\mathbf{A})\| \geq \sqrt{cLB^2 \frac{\log(d)+\log(n)}{mn}}\right\} \leq \frac{1}{d}.$$

Proof. We denote

$$\mathbf{L}_{kk'} = -\frac{R_{kk'}(\mathbf{A})}{1+R_{kk'}(\mathbf{A})}(Y_k - Y_{k'})(\boldsymbol{X}_k - \boldsymbol{X}_{k'}).$$

When n is even, by Lemma S.4 in [126], we are able to partition the collection $\mathcal{P} = \{(k,k') : 1 \leq j < j' \leq n\}$ into $n-1$ groups $G_1, \ldots, G_{n-1}$, s.t. $|G_i| = n/2$ and no individual occurs more than one time within a group.

$$\nabla \ell(\mathbf{A}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{G_i} \frac{2}{n} \mathbf{L}_{kk'}.$$

When n is odd, similarly, let's add one extra index. We are able to partition the collection $\mathcal{P} = \{(k,k') : 1 \leq j < j' \leq n\}$ into n groups $G_1, \ldots, G_{n-1}$, s.t. $|G_i| = (n+1)/2$ and no individual occurs more than one time within a group. In each group, the extra index only appears once, and we remove that pair. Therefore, we have n groups $G_1, \ldots, G_{n-1}$, s.t. $|G_i| = (n-1)/2$ and no individual occurs more than one time within a group.

$$\nabla \ell(\mathbf{A}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{G_i} \frac{2}{n-1} \mathbf{L}_{kk'}.$$

WLOG, we assume n is even. The case when n is odd directly follows. We apply the dilation from [127],

$$\mathscr{T}(\mathbf{L}_{kk'}) = \begin{bmatrix} 0 & \mathbf{L}_{kk'} \\ \mathbf{L}_{kk'}^\top & 0 \end{bmatrix}.$$

Since $|Y_k| \leq B$, we have that

$$\frac{2}{n} \|\mathscr{T}(\mathbf{L}_{kk'})\| \leq \frac{2}{n} 2\sqrt{2}B \leq \frac{4}{n}\sqrt{2}B = R.$$

Denote $(r(k), c(k))$ as the row-index and column-index of k. Before looking into the variance parameter, we look at the term

$$\begin{aligned}
&\left\|\mathbb{E}\mathbf{L}_{kk'}\mathbf{L}_{kk'}^\top\right\| \\
\leq& 4B^2 \left\|\mathbb{E}(\boldsymbol{X}_k - \boldsymbol{X}_{k'})(\boldsymbol{X}_k - \boldsymbol{X}_{k'})^\top\right\| \\
=& 4B^2 \left\| \sum_{kk'} \pi_k\pi_{k'}\mathbf{E}_{r(k)r(k)} + \sum_{kk'} \pi_k\pi_{k'}\mathbf{E}_{r(k')r(k')} - \sum_{kk'} \pi_k\pi_{k'}(\mathbf{E}_{r(k)r(k')} + \mathbf{E}_{r(k')r(k)})1(r(k) \neq r(k')) \right\| \\
=& 4B^2 \left\| 2\sum_{k} \pi_k\mathbf{E}_{r(k)r(k)} - \sum_{kk'} \pi_k\pi_{k'}(\mathbf{E}_{r(k)r(k')} + \mathbf{E}_{r(k')r(k)})1(r(k) \neq r(k')) \right\| \\
=& 8B^2 \left\| \sum_{r\in[m_1]} \pi_r\mathbf{E}_{rr} - \sum_{rr'} \frac{\pi_r\pi_{r'}}{2}(\mathbf{E}_{rr'} + \mathbf{E}_{r'r})1(r \neq r') \right\|.
\end{aligned}$$

Let's now consider a weighted graph such each row index pair has a weight

$$w_{rr'} = \frac{\pi_r\pi_{r'}}{2}, \quad r \neq r'.$$

Therefore,

$$d_{rr} = 2\sum_{r'\neq r} \frac{\pi_r\pi_{r'}}{2} = \pi_r(1-\pi_r).$$

We then consider the corresponding graph Laplacian matrix, by [128], we have that

$$\|D - A\| \leq 2\max_r \pi_r(1-\pi_r).$$

Therefore, we have that

$$
\begin{aligned}
&\left\|\mathbb{E}\mathbf{L}_{kk'}\mathbf{L}_{kk'}^\top\right\| \\
\leq& 8B^2\left\|\sum_{r\in[m_1]}\pi_r(1-\pi_r)\mathbf{E}_{rr}-\sum_{rr'}\frac{\pi_r\pi_{r'}}{2}(\mathbf{E}_{rr'}+\mathbf{E}_{r'r})1(r\neq r')+\sum_{r\in[m_1]}\pi_r^2\mathbf{E}_{rr}\right\| \\
\leq& 8B^2\left\|\sum_{r\in[m_1]}\pi_r(1-\pi_r)\mathbf{E}_{rr}-\sum_{rr'}\frac{\pi_r\pi_{r'}}{2}(\mathbf{E}_{rr'}+\mathbf{E}_{r'r})1(r\neq r')\right\|+8B^2\left\|\sum_{r\in[m_1]}\pi_r^2\mathbf{E}_{rr}\right\| \\
\leq& 16B^2\max_r\pi_r(1-\pi_r)+8B^2\max_r\pi_r^2 \\
\leq& 16B^2\frac{L}{m}+8B^2\frac{L^2}{m^2}.
\end{aligned}
$$

The variance parameter

$$
\begin{aligned}
\sigma^2 &= \max\left\{\sum\frac{4}{n^2}\left\|\mathbb{E}\mathbf{L}_{kk'}\mathbf{L}_{kk'}^\top\right\|,\sum\frac{4}{n^2}\left\|\mathbb{E}\mathbf{L}_{kk'}^\top\mathbf{L}_{kk'}\right\|\right\} \\
&= \frac{2}{n}\max\left\{\left\|\mathbb{E}\mathbf{L}_{kk'}\mathbf{L}_{kk'}^\top\right\|,\left\|\mathbb{E}\mathbf{L}_{kk'}^\top\mathbf{L}_{kk'}\right\|\right\}\leq\frac{2}{n}\left(16B^2\frac{L}{m}+8B^2\frac{L^2}{m^2}\right).
\end{aligned}
$$

Therefore, by Theorem 6.1 from [127], we have that the following holds for all $t\geq 0$,

$$
\begin{aligned}
\Pr\left\{\left\|\sum_{kk'\in G_i}\frac{2}{n}\mathbf{L}_{kk'}\right\|\geq t\right\} &\leq d\exp\left(\frac{-t^2/2}{\sigma^2+Rt}\right) \\
&\leq\begin{cases} d\exp(-3t^2/8\sigma^2) & \text{for } t\leq\sigma^2/R \\ d\exp(-3t/8R) & \text{for } t\geq\sigma^2/R\end{cases},
\end{aligned}
$$

where $R=\frac{4}{n}\sqrt{2}B$ and $\sigma^2=\frac{2}{n}\left(16B^2\frac{L}{m}+8B^2\frac{L}{m^2}\right)\leq\frac{64}{nm}B^2$. Choosing $t=\sqrt{cLB^2\frac{\log(d)+\log(n)}{mn}}$, when $n\leq\frac{c}{128}Lm\log(d)$, $t\geq\sigma^2/R$, we have that

$$
\Pr\left\{\|\nabla\ell(\mathbf{A})\|\geq\sqrt{cLB^2\frac{\log(d)+\log(n)}{mn}}\right\}\leq nd\exp(-3\sqrt{cn(\log(d)+\log(n))/m}/32\sqrt{2}B).
$$

When $n \geq \frac{3c}{128} Lm \log(d)$ (keeping $n \leq m_1 m_2$), $t \geq \sigma^2/R$, we have that

$$\begin{aligned}
\Pr\left\{ \|\nabla \ell(\mathbf{A})\| \geq \sqrt{cLB^2 \frac{\log(d)+\log(n)}{mn}} \right\} &\leq nd \exp(-3c(\log(d)+\log(n))/512B^2) \\
&\leq nd \exp(-3c\log(d)/512B^2 - \log(n)) \\
&\leq d \exp(-3c\log(d)/512B^2),
\end{aligned}$$

where the second inequality is from $\frac{3c}{512B^2} \geq 4$. We choose c s.t., $3c/512B^2 \geq 4$. Therefore, when $n \geq \frac{3c}{128} Lm \log(d)$,

$$\Pr\left\{ \|\nabla \ell(\mathbf{A})\| \geq \sqrt{cLB^2 \frac{\log(d)+\log(n)}{mn}} \right\} \leq \frac{1}{d}.$$

□

Recall the Hessian

$$\begin{aligned}
\nabla^2 \ell(\tilde{\mathbf{A}}) &= \frac{2}{n(n-1)} \sum_{1 \leq k < k' \leq n} \{\psi''(Y_{k\backslash k'} \langle \boldsymbol{X}_{k\backslash k'}, \tilde{\mathbf{A}} \rangle) Y_{k\backslash k'}^2 \mathrm{vec}(\boldsymbol{X}_{k\backslash k'})^{\otimes 2}\} \\
&= \frac{2}{n(n-1)} \sum_{1 \leq k < k' \leq n} (Y_k - Y_{k'})^2 \frac{\exp((Y_k - Y_{k'})\langle \boldsymbol{X}_k - \boldsymbol{X}_{k'}, \tilde{\mathbf{A}} \rangle)}{(1+\exp((Y_k - Y_{k'})\langle \boldsymbol{X}_k - \boldsymbol{X}_{k'}, \tilde{\mathbf{A}} \rangle))^2} \mathrm{vec}(\boldsymbol{X}_{k\backslash k'})^{\otimes 2} \\
&= \frac{1}{n-1} \sum_g \sum_{kk' \in g} \frac{2}{n} Z_{kk'}^2 \mathrm{vec}(\boldsymbol{X}_{k\backslash k'})^{\otimes 2},
\end{aligned}$$

where

$$Z_{kk'} = (Y_k - Y_{k'}) \frac{\exp((Y_k - Y_{k'})\langle \boldsymbol{X}_k - \boldsymbol{X}_{k'}, \tilde{\mathbf{A}} \rangle / 2)}{1+\exp((Y_k - Y_{k'})\langle \boldsymbol{X}_k - \boldsymbol{X}_{k'}, \tilde{\mathbf{A}} \rangle)}.$$

By the boundedness of Y_k, we have that $|Z_{kk'}| \leq 2B$. We also assume $\mathbb{E}(Z_{kk'}^2 | \boldsymbol{X}_k, \boldsymbol{X}_{k'}) \geq \kappa_M$.

We consider the following constrained set

$$\mathcal{C}(r) = \Bigg\{ \mathbf{U} \in \mathbb{R}^{m_1 \times m_2} \;\Bigg|\; \langle \mathbf{J}, \mathbf{U} \rangle = 0, \|\mathbf{U}\| = 1, \|\mathbf{U}\|_\star \le \sqrt{r}\, \|\mathbf{U}\|_F, \\ \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A})\mathbf{u} \ge c\sqrt{\frac{\log(d)+\log(n)}{n}} \Bigg\}.$$

For each $\mathbf{U} \in \mathbb{R}^{m_1 \times m_2}$, we denote $\mathbf{u} = \mathrm{vec}(\mathbf{U})$. Note that

$$\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A})\mathbf{u} = \frac{1}{n-1} \sum_g \sum_{kk' \in g} \frac{2}{n} Z_{kk'}^2 (u_k - u_{k'})^2,$$

and

$$\begin{aligned} \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A})\mathbf{u} &= \mathbb{E} Z_{kk'}^2 (u_k - u_{k'})^2 \ge \kappa_M \mathbb{E}(u_k - u_{k'})^2 \\ &\ge \kappa_M \frac{1}{\tau_l^2 m_1^2 m_2^2} \sum_{1 \le k,k' \le m_1 m_2} (u_k - u_{k'})^2 \\ &= \kappa_M \frac{2}{\tau_l^2 m_1 m_2} \|\mathbf{U}\|_F^2 . \end{aligned}$$

Denote

$$\Sigma_{g,R} = \sum_{kk' \in g} \varepsilon_{kk'} \frac{2}{n} Z_{kk'} (\boldsymbol{X}_k - \boldsymbol{X}_{k'})$$

where $\varepsilon_{kk'}$ are independent Radamacher variable, and define

$$\mathcal{E} = 512 B^2 r \tau_l^2 m_1 m_2 / \kappa_M (\mathbb{E} \|\Sigma_{g,R}\|)^2.$$

Lemma 35. *For all* $\mathbf{U} \in \mathcal{C}(r)$*,*

$$\mathbf{u}^\top \nabla^2 \ell(\mathbf{A})\mathbf{u} \ge \mathbb{E}\mathbf{u}^\top \nabla \ell(\mathbf{A})\mathbf{u} - 512 B^2 r \tau_l^2 m_1 m_2 / \kappa_M (\mathbb{E} \|\Sigma_{g,R}\|)^2$$

with probability at least $1 - \frac{2}{d}$*.*

Proof. We will show that the probability of the following bad event is small

$$\mathcal{B} = \left\{ \exists \mathbf{U} \in \mathcal{C}(r) \text{ such that } \left| \mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} - \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} \right| > \frac{1}{2} \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} + \mathcal{E}. \right\}$$

We use a standard peeling argument. Let $v = \sqrt{\frac{3(\log(d)+\log(n))}{2c\log(6/5)n}}$ and $\alpha = \frac{6}{5}$, where $c = \frac{1}{128}$. For $l \in \mathbb{N}$ set

$$\mathcal{S}_l = \left\{ \mathbf{U} \in \mathcal{C}(r) : \alpha^{l-1}\nu \leq \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} \leq \alpha^l \nu \right\}.$$

If the event $\mathcal{B}$ holds for some matrix $\mathbf{U} \in \mathcal{C}(r)$, then $\mathbf{U}$ belongs to some $\mathcal{S}_l$ and

$$\begin{aligned} \left| \mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} - \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} \right| &> \frac{1}{2} \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} + \mathcal{E} \\ &> \frac{1}{2} \alpha^{l-1} \nu + \mathcal{E} \\ &> \frac{5}{12} \alpha^l \nu + \mathcal{E}. \end{aligned}$$

For each $T > \nu$, we consider the following set of matrices

$$\mathcal{C}(r, T) = \{ \mathbf{U} \in \mathcal{C}(r) : \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} \leq T \},$$

and the following event

$$\mathcal{B}_l = \left\{ \exists \mathbf{U} \in \mathcal{C}(r, \alpha^l \nu) \text{ such that } \left| \mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} - \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} \right| > \frac{1}{2} \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} + \mathcal{E}. \right\}$$

Denote

$$W_T = \sup_{\mathbf{U} \in \mathcal{C}(r,T)} \left| \mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} - \mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A}) \mathbf{u} \right|,$$

and

$$Z_{g,T} = \sup_{\mathbf{U} \in \mathcal{C}(r,T)} \left| \sum_{kk' \in g} \frac{2}{n} Z_{kk'}^2 (u_k - u_{k'})^2 - \mathbb{E} \sum_{kk' \in g} \frac{2}{n} Z_{kk'}^2 (u_k - u_{k'})^2 \right|.$$

We have that

$$W_T \leq \frac{1}{n-1}\sum_g Z_{g,T}.$$

We start with $Z_{g,T}$ first. The standard symmetrization trick still applies here,

$$\mathbb{E}Z_{g,T} \leq 2\mathbb{E}\sup_{\mathbf{U}\in\mathcal{C}(r,T)}\left|\sum_{kk'\in g}\varepsilon_{kk'}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2\right|.$$

Since $|Z_{kk'}| \leq 2B$ and $\|\mathbf{U}\|_\infty = 1$, $Z_{kk'}^2(u_k-u_{k'})^2 \leq 16B^2$. Therefore, $\phi(u) = u^2$, $|\phi(u)-\phi(v)| \leq |u+v||u-v| \leq 8B|u-v|$. The contraction inequality yields

$$\begin{aligned}
\mathbb{E}Z_{g,T} &\leq 2\mathbb{E}\sup_{\mathbf{U}\in\mathcal{C}(r,T)}\left|\sum_{kk'\in g}\varepsilon_{kk'}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2\right| \\
&\leq 16B\mathbb{E}\sup_{\mathbf{U}\in\mathcal{C}(r,T)}\left|\sum_{kk'\in g}\varepsilon_{kk'}\frac{2}{n}Z_{kk'}(u_k-u_{k'})\right| \\
&\leq 16B\mathbb{E}\sup_{\mathbf{U}\in\mathcal{C}(r,T)}\left|\sum_{kk'\in g}\varepsilon_{kk'}\frac{2}{n}Z_{kk'}\langle \boldsymbol{X}_k-\boldsymbol{X}_{k'},\mathbf{U}\rangle\right| \\
&\leq 16B\mathbb{E}\sup_{\mathbf{U}\in\mathcal{C}(r,T)}\left|\left\langle\sum_{kk'\in g}\varepsilon_{kk'}\frac{2}{n}Z_{kk'}(\boldsymbol{X}_k-\boldsymbol{X}_{k'}),\mathbf{U}\right\rangle\right| \\
&\leq 16B\mathbb{E}\left\|\sum_{kk'\in g}\varepsilon_{kk'}\frac{2}{n}Z_{kk'}(\boldsymbol{X}_k-\boldsymbol{X}_{k'})\right\|\|\mathbf{U}\|_\star \\
&\leq 16B\sqrt{r}\,\|\mathbf{U}\|_F\,\mathbb{E}\left\|\sum_{kk'\in g}\varepsilon_{kk'}\frac{2}{n}Z_{kk'}(\boldsymbol{X}_k-\boldsymbol{X}_{k'})\right\| \\
&\leq 8\sqrt{2}B\sqrt{r}\tau_l\sqrt{m_1m_2T/\kappa_M}\mathbb{E}\,\|\Sigma_{g,R}\|\,.
\end{aligned}$$

Note that

$$\begin{aligned}
8\sqrt{2}B\sqrt{r}\tau_l\sqrt{m_1m_2T/\kappa_M}\mathbb{E}\,\|\Sigma_{g,R}\| &\leq \frac{1}{4}T + 512B^2r\tau_l^2m_1m_2/\kappa_M(\mathbb{E}\,\|\Sigma_{g,R}\|)^2 \\
&\leq \frac{8}{9}\frac{5}{12}T + 512B^2r\tau_l^2m_1m_2/\kappa_M(\mathbb{E}\,\|\Sigma_{g,R}\|)^2.
\end{aligned}$$

By lemma 36, with probability smaller than $n\exp(-cnT^2)$, where $c = \frac{1}{128}$, we have that

$$W_T \geq \frac{5}{12}T + 512B^2 r\tau_l^2 m_1 m_2/\kappa_M(\mathbb{E}\left\|\Sigma_{g,R}\right\|)^2 = \frac{5}{12}T + \mathcal{E}.$$

Therefore, we obtain that

$$\Pr(\mathcal{B}_l) \leq n\exp(-cn\alpha^{2l}v^2).$$

Using the union bound, we have

$$\begin{aligned}\Pr(\mathcal{B}) &\leq \sum_{l=1}^{\infty}\Pr(\mathcal{B}_l) \leq \sum_{l=1}^{\infty} n\exp(-cn\alpha^{2l}v^2)\\ &\leq \sum_{l=1}^{\infty} n\exp(-2cn\log(\alpha)v^2 l) \leq n\frac{\exp(-2cn\log(\alpha)v^2)}{1-\exp(-2cn\log(\alpha)v^2)}\\ &\leq \frac{2}{d}.\end{aligned}$$

where the last inequality is obtained by choosing $v = \sqrt{\frac{3(\log(d)+\log(n))}{2c\log(6/5)n}}$. □

We aim to show concentration on W_T.

Lemma 36.

$$\Pr\left(W_T \geq \mathbb{E}Z_{g,T} + \frac{1}{9}\left(\frac{5}{12}T\right)\right) \leq n\exp(-cnT^2).$$

Proof. Recall

$$\begin{aligned}W_T &= \sup_{\mathbf{U}\in\mathcal{C}(r,T)} \left|\mathbf{u}^\top\nabla^2\mathcal{L}_n(\mathbf{A})\mathbf{u} - \mathbb{E}\mathbf{u}^\top\nabla^2\mathcal{L}_n(\mathbf{A})\mathbf{u}\right|\\ &= \sup_{\mathbf{U}\in\mathcal{C}(r,T)} \left|\frac{1}{n-1}\sum_g\sum_{kk'\in g}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2 - \mathbb{E}\frac{1}{n-1}\sum_g\sum_{kk'\in g}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2\right|\\ &\leq \frac{1}{n-1}\sum_g \sup_{\mathbf{U}\in\mathcal{C}(r,T)}\left|\sum_{kk'\in g}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2 - \mathbb{E}\sum_{kk'\in g}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2\right|,\end{aligned}$$

and

$$Z_{g,T} = \sup_{\mathbf{U}\in\mathcal{C}(r,T)}\left|\sum_{kk'\in g}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2 - \mathbb{E}\sum_{kk'\in g}\frac{2}{n}Z_{kk'}^2(u_k-u_{k'})^2\right|.$$

Within each grouping g, by Massart's concentration inequality (e.g., Theorem 14.2 in [129]), we have that

$$\mathbb{P}\left(Z_{g,T} \geq \mathbb{E}(Z_{g,T}) + \frac{1}{9}\left(\frac{5}{12}T\right)\right) \leq \exp\left(-\frac{1}{128}nT^2\right).$$

Therefore, with a union bound argument, we have that

$$\mathbb{P}\left(W_T \geq \mathbb{E}(Z_{g,T}) + \frac{1}{9}\left(\frac{5}{12}T\right)\right) \leq n\exp\left(-\frac{1}{128}nT^2\right).$$

□

We aim to bound $\mathbb{E}\left\|\Sigma_{g,R}\right\|$ below. Denote

$$\mathbf{L}_{kk'} = \varepsilon_{kk'} Z_{kk'} (\boldsymbol{X}_k - \boldsymbol{X}_{k'}).$$

Lemma 37. *There exists some absolute constant $c^\star$, s.t., with probability at least $1 - \exp(-t)$, we have that*

$$\left\|\frac{2}{n}\sum_{kk'\in g}\mathbf{L}_{kk'}\right\| \leq C^\star \max\left\{B\sqrt{\frac{L(t+\log(d))}{mn}}, B\log(m)\frac{t+\log(d)}{n}\right\}.$$

Proof. We first observe that

$$\mathbb{E}\mathbf{L}_{kk'} = 0.$$

Similar to the proof of Lemma 34, by observing the pattern of $\boldsymbol{X}_k - \boldsymbol{X}_{k'}$, we have that

$$\left\|\varepsilon_{kk'} Z_{kk'} (\boldsymbol{X}_k - \boldsymbol{X}_{k'})\right\| \leq 2\sqrt{2}B = U.$$

We then verify the variance parameter,

$$\sigma_L^2 = \max\left\{\left\|\frac{2}{n}\sum_{kk'\in g}\mathbb{E}(\mathbf{L}_{kk'}\mathbf{L}_{kk'}^\top)\right\|, \left\|\frac{2}{n}\sum_{kk'\in g}\mathbb{E}(\mathbf{L}_{kk'}^\top\mathbf{L}_{kk'})\right\|\right\}.$$

Following the same argument in the proof of Lemma 34, we have that

$$\sigma_L^2 \leq 32B^2\frac{L}{m}.$$

Therefore, with probability at least $1-\exp(-t)$, we have that

$$\left\|\frac{2}{n}\sum_{kk'\in G_i}\mathbf{L}_{kk'}\right\| \leq c^\star \max\left\{\sigma_L\sqrt{\frac{t+\log(d)}{n}}, U\log\frac{U}{\sigma_L}\frac{t+\log(d)}{n}\right\},$$

for some absolute constant $c^\star$, which further simplifies to

$$\left\|\frac{2}{n}\sum_{kk'\in g}\mathbf{L}_{kk'}\right\| \leq C^\star \max\left\{B\sqrt{\frac{L(t+\log(d))}{mn}}, B\log(m)\frac{t+\log(d)}{n}\right\}.$$

□

Lemma 38. *Assume $\varepsilon_{kk'}$ are i.i.d. Rademacher random variables. When $n \geq m\log^3(d)/L$, these is some absolute constant $C^\star > 0$ s.t.*

$$\mathbb{E}\left\|\frac{2}{n}\sum_{kk'\in g}\varepsilon_{kk'}Z_{kk'}(\boldsymbol{X}_k - \boldsymbol{X}_{k'})\right\| \leq C^\star\sqrt{\frac{2eL\log(d)}{nm}}.$$

Proof. Follow Lemma 6 from [44] (or Lemma 7 in [130]). Find the critical $t^\star$ when the tail behavior changes, and apply Holder's inequality. □

Lemma 39. *Suppose $\langle \mathbf{J}, \mathbf{U}\rangle = 0$ and $\mathbb{E}(Z_{kk'}^2|\boldsymbol{X}_k, \boldsymbol{X}_{k'}) \geq \kappa_M$, we have that*

$$\mathbb{E}\mathbf{u}^\top\nabla^2\mathcal{L}_n(\mathbf{A})\mathbf{u} \geq \kappa_M\frac{2}{\tau_l^2 m_1 m_2}\|\mathbf{U}\|_F^2.$$

Proof.

$$\begin{aligned}
\mathbb{E}\mathbf{u}^\top \nabla^2 \mathcal{L}_n(\mathbf{A})\mathbf{u} &= \mathbb{E} Z_{kk'}^2 (u_k - u_{k'})^2 \geq \kappa_M \mathbb{E}(u_k - u_{k'})^2 \\
&\geq \kappa_M \frac{1}{\tau_l^2 m_1^2 m_2^2} \sum_{1 \leq k, k' \leq m_1 m_2} (u_k - u_{k'})^2 \\
&\geq \kappa_M \frac{1}{\tau_l^2 m_1^2 m_2^2} \sum_{1 \leq k, k' \leq m_1 m_2} u_k^2 + u_{k'}^2 - 2u_k u_{k'} \\
&= \kappa_M \frac{2}{\tau_l^2 m_1 m_2} \|\mathbf{U}\|_F^2 .
\end{aligned}$$

□